




**ADVERTIMENT.** L'accés als continguts d'aquesta tesi queda condicionat a l'acceptació de les condicions d'ús establertes per la següent llicència Creative Commons:  <https://creativecommons.org/licenses/?lang=ca>

**ADVERTENCIA.** El acceso a los contenidos de esta tesis queda condicionado a la aceptación de las condiciones de uso establecidas por la siguiente licencia Creative Commons:  <https://creativecommons.org/licenses/?lang=es>

**WARNING.** The access to the contents of this doctoral thesis it is limited to the acceptance of the use conditions set by the following Creative Commons license:  <https://creativecommons.org/licenses/?lang=en>

---

# Advanced computational strategies for metalloenzyme design

---

Thesis submitted by

**Laura Tiessler Sala**

for the degree of

Doctor in Bioinformatics

Doctorate Program in Bioinformatics

**Supervisors:**

Agustí Lledós Falcó and Jean-Didier Maréchal

**UAB**  
Universitat Autònoma  
de Barcelona



Recommended for acceptance by supervisors:

Agustí Lledós Falcó

Jean-Didier Maréchal



# *Abstract*

## **Advanced computational strategies for metalloenzyme design**

by Laura Tiessler Sala

Nature is abundant with proteins that contain metals, ranging from metalloenzymes serving crucial catalytic reactions to proteins that storage and transport metals. The molecular study of metalloproteins allows to find solutions to a wide range of chemical problems and provides insights into their recognition processes. Researchers have seized the opportunity to copy nature by create new catalysts, Artificial metalloenzymes (ArM). These ArMs are designed by combining natural protein scaffolds with metallic moieties, enabling them to carry out new reactions in nature. In the recent years, molecular modeling has become an essential tool in these fields, though certain challenges still need to be addressed.

This PhD thesis aims to apply molecular modeling to understand the behavior of metal containing proteins, focusing on natural metallic proteins and ArMs. The first part of this thesis encompasses the study of proteins containing the prototypical metallic cofactor heme. A molecular modeling protocol based on enhanced molecular dynamics techniques is applied to unravel different binding mechanism of heme on protein hemophore HasA. Furthermore, a software is developed for detecting heme binding sites based only on structural information. The second part of this thesis focuses on the computational-aided design of ArM. Integrative approaches combining various techniques, such as quantum mechanics, molecular docking, and classical molecular dynamics simulations are applied to obtain a deeper understanding of two different ArMs: one based on single or dual gold hydroamination and the other a Suzuki-Miyaura reaction involving palladium.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Metals in all life forms . . . . .	2
1.2	Metalloproteins . . . . .	6
1.2.1	Common metals in metalloproteins . . . . .	10
1.3	Artificial metalloenzymes (ArM) . . . . .	23
1.3.1	Overview of ArM history . . . . .	23
1.3.2	Development of ArM . . . . .	24
1.3.3	Discovery of ArM . . . . .	26
1.3.4	Optimization of ArM . . . . .	32
1.3.5	Reaction scope of ArM . . . . .	33
1.3.6	ArM containing porphyrin . . . . .	35
1.4	Molecular modeling for metalloproteins . . . . .	36
1.4.1	Overview of molecular modeling in metallic biomolecules .	37
1.4.2	Molecular modeling in ArM . . . . .	41
1.4.3	Predicting the binding of metals ions and heme to proteins .	44
<b>2</b>	<b>Methodology</b>	<b>45</b>
2.1	Quantum Mechanics (QM) . . . . .	45
2.1.1	Hartree Fock . . . . .	48
2.1.2	DFT . . . . .	49
2.1.3	Metals in QM . . . . .	54
2.2	Molecular Mechanics (MM) . . . . .	56
2.2.1	Force fields and metals for biomolecules . . . . .	60
2.2.2	Molecular Dynamics (MD) . . . . .	61
2.2.3	Enhanced sampling methods . . . . .	66

2.2.4	Analysis of molecular dynamics simulation . . . . .	69
2.2.5	Protein-ligand dockings . . . . .	71
2.2.6	Python language in science . . . . .	75
<b>3</b>	<b>Objectives</b>	<b>77</b>
<b>4</b>	<b>The relevance of heme binding processes and their prediction</b>	<b>79</b>
4.1	Exploring the molecular events of heme binding mechanisms . . .	80
4.1.1	Conformational variation study of heme binding sites . . .	81
4.1.2	Heme binding processes in hemophore HasA . . . . .	83
4.1.3	Methodology . . . . .	87
4.1.4	Results of Hemophore HasA <sub>yp</sub> . . . . .	89
4.1.5	Results of Hemophore HasA <sub>sm</sub> . . . . .	93
4.1.6	Conclusions: comparison of both systems . . . . .	100
4.2	Development of software for the identification of heme binding sites . . . . .	102
4.2.1	Conceptualization of the software . . . . .	103
4.2.2	Statistical analysis . . . . .	105
4.2.3	Workflow of software . . . . .	108
4.2.4	Scoring . . . . .	111
4.2.5	Benchmark . . . . .	113
4.2.6	Case study 1: Detection of natural heme binding sites . . . .	114
4.2.7	Case study 2: Application in design of ArM . . . . .	118
4.2.8	Conclusions . . . . .	121
<b>5</b>	<b>Applicative cases of computational aided design of ArM</b>	<b>123</b>
5.1	Overview . . . . .	123
5.2	Methodology and computational details . . . . .	125
5.3	Molecular modeling to optimize an Au-ArM for heterocyclization .	129
5.3.1	Context and experimental background . . . . .	129
5.3.2	Objectives and methodology . . . . .	131
5.3.3	Learning from the organometallic side: DFT calculations . .	134
5.3.4	Modeling of protein systems . . . . .	139
5.3.5	Molecular dockings and MD simulations of TSs . . . . .	141
5.3.6	Additional studies: Longer linker . . . . .	147
5.3.7	Conclusions . . . . .	149
5.4	Rationalization of a streptavidin based suzukiase . . . . .	150
5.4.1	Context and experimental background . . . . .	150
5.4.2	Objectives and methodology . . . . .	153
5.4.3	Studying the reaction with DFT . . . . .	155

5.4.4	Molecular modeling of the protein system . . . . .	160
5.4.5	Docking and MD simulation of TS or intermediates . . . . .	164
5.4.6	Dockings of OA intermediates . . . . .	165
5.4.7	Conclusions . . . . .	171
<b>6</b>	<b>Finding metal binding sites to design a new ArM</b>	<b>173</b>
6.1	Overview of BioMetAll . . . . .	174
6.2	Applicative case of design of ArM: $\alpha$ -Rep . . . . .	177
6.2.1	Initial screening of crystallographic monomeric structures .	178
6.2.2	Initial screening of crystallographic dimeric structures . . .	179
6.2.3	MD simulation of dimer A3_A3 . . . . .	181
6.2.4	Mutation calculations . . . . .	182
6.2.5	Docking calculations with metal . . . . .	183
6.3	Conclusions . . . . .	185
<b>7</b>	<b>Other works</b>	<b>187</b>
<b>8</b>	<b>Conclusions</b>	<b>189</b>
<b>A</b>	<b>Other works</b>	<b>193</b>
A.1	Free energy studies on metalloproteins . . . . .	194
A.2	Direct benzene hydroxylation with O <sub>2</sub> induced by copper complexes	197
A.3	Molecular modeling of an artificial Co-Hemoprotein for H <sub>2</sub> production . . . . .	200
<b>B</b>	<b>Chapter 4: Supplementary information</b>	<b>203</b>
B.1	Exploring the molecular events of heme binding mechanisms . . .	203
B.2	Development of software for the identification of heme binding sites	209
<b>C</b>	<b>Chapter 5: Supplementary information</b>	<b>213</b>
C.1	Molecular modeling to optimize an Au-ArM for heterocyclization .	213
C.2	Rationalization of a streptavidin based suzukiase . . . . .	218
<b>D</b>	<b>Chapter 6: Supplementary information</b>	<b>219</b>
D.1	Finding metal binding sites to design a new ArM . . . . .	219
	<b>List of publications</b>	<b>221</b>
	<b>Bibliography</b>	<b>223</b>



# CHAPTER 1

## Introduction

Metal ions are essential to maintain all types of living organisms, going from animals and plants to bacteria. Metals, when associated with biomolecules, can perform many fundamental biological processes thanks to their unique catalytic and structural properties. The relevance of the functions of metals is reflected in many organisms' genomes, in which metal-containing proteins represent more than one-third of the proteome.<sup>1-3</sup> Metallic ions are involved in a broad range of biological processes including structural, acid-base or redox catalysis, signalling, electron transfer or energy storage. In metalloproteins, metals are usually bound to amino acids from the protein, inorganic anions or organic cofactors. Depending on the nature of the metal, they tend to bind to different residues (Glu/Asp for hard metals and Cys/Met for soft metals).<sup>4</sup>

Because numerous biological functions depend on the presence of metal ions, any disruptions in their homeostasis can lead to a broad range of physiological disorders or even death. For example, the deficiency of Fe causes anaemia, while alterations the concentration of in Cu or Al can cause heart problems and brain diseases like Alzheimer's.<sup>5,6</sup> The presence of essential metals is fundamental for survival, inversely, the excess of these metals can also be toxic. A clear example is the excess of cobalt which can lead to lethal cardiomyopathies.<sup>7-9</sup>

In consequence, the homeostasis of metal ions is of importance, the concentration of metal ions has to be maintained strictly in cells and tissues to avoid either excess or deficiency of metals. Metals also have an important role in medicine, Li has been used to treat depression in the form of Lithium carbonate for more than 50 years. Au compounds are used to treat Rheumatoid Arthritis and several

anticancer drugs contain metals, like wide used Pt-cisplatin or potential future Ru-anticancer drugs.<sup>10-14</sup>

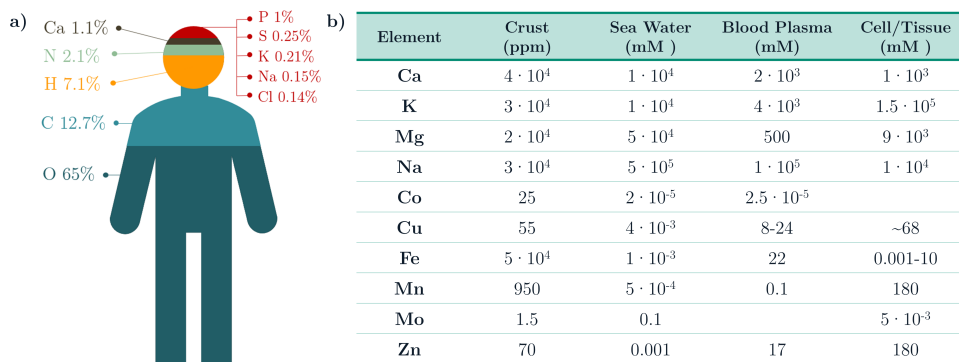
For the two last decades, an array of bioinspired metal-containing systems has emerged. Among these families of non-natural systems are Artificial Metalloenzymes (ArMs). Those systems are obtained by introducing a catalytic organometallic compound into a compatible biomolecular scaffold. This leads to enzyme displaying new-to-nature reactions as well as substrate selectivity and catalytic specificity. Amongst successful ArMs designs we can count on enzymes able to perform enantioselective cyclopropanation, Suzuki-Miyaura, imine reductions, or epoxidation.<sup>15</sup> In this chapter, we will get deeper into the properties of metals that make them indispensable for life and understand which types of reactions are carried out by the most common natural metalloenzymes. Then, we will look at non-natural ones by focusing on how ArM are designed, how they diverge from natural metalloenzymes and how molecular modeling can be used in this field.

## 1.1. Metals in all life forms

Currently, twenty elements are currently considered as essential for all living organisms on earth. First, essential elements should be defined as elements that are present on all tissues and crucial for survival. If an essential element is removed from the diet, its deficiency or absence causes irreversible damages, that can only be prevented or cured by supplementing the element itself.<sup>16</sup> Essential elements should be differentiated from beneficial elements, which serve limited functions or only provide some health benefits, as in the case of fluorine or silicon.<sup>17</sup> **Figure 1.1** highlights the twenty essential elements, five essential elements for several species, and lastly, Cr, which is still under consideration as an essential element.<sup>18</sup>

In terms of the human's body composition, the most abundant elements are C, H, O, N, which together with P and S, constitute the non-metallic bulk elements (**Figure 1.2a**). These six elements constitute all intermediates of metabolism and all essential building blocks to synthesize proteins, DNA, polysaccharide and membranes. However, living cells need additional functionalities beyond these building blocks. For example, DNA or protein charges need to be neutralized with metallic ions and metals are better suited to carry out redox or catalyze





**Figure 1.2:** a) Distribution of essential elements in human human body of 70kg. b) Concentration of essential metals in crust, sea water, plasma and cell/tissue. Adapted from [19, 22].

Six of the ten essential elements are in the top 10 regarding abundance in the earth's crust. A minimal concentration of an element on earth is always required, but the availability of the soluble form in soil/water is relevant for an organism's acquisition. Organisms have developed mechanisms to adsorb and maintain the concentration of inorganic elements constant inside cells, either to avoid toxic concentrations or because of its low abundance. This is illustrated by the difference of concentration between sea water and cells. Despite low concentration of certain metals in sea water, cells are able to maintain higher concentrations due to its uptake and regulating mechanisms.<sup>19,22</sup>

Lastly, the most important criteria for the essential character of a metal it is the function that it fulfills in an organism. The suitability of a metal for a certain function is determined by the properties of the metal. There is not a specific property that defines which metals are essential.<sup>19</sup> However, it has been found that one of the best ways to understand the relationship between function and properties of metals is through the *Hard and Soft Acid and Base* theory, which divides Lewis acids and bases as *Hard* and *Soft* categories depending on size, oxidation state and polarizability. *Hard* acids are characterized by their small size, high oxidation state and low polarizability, therefore, they tend to interact stronger with hard bases through ionic interactions. Contrarily, *Soft* acids have large size, low oxidation state and high polarizability, making them interact with soft bases. In between are borderline acids, which have intermediate properties.<sup>23</sup> Main properties and functions are summarized in **Table 1.1**.

Metal	Hard or Soft	Binding	Mobility	Function
Na <sup>+</sup> , K <sup>+</sup>	Hard	Weak	High	Charge carriers
Mg <sup>2+</sup> , Ca <sup>2+</sup>	Hard	Moderate	Semi-mobile	Structural
Zn <sup>2+</sup>	Borderline	Moderate\Strong	Intermediate	Lewis acid
Co, Cu, Fe, Mn, Mo	Borderline	Strong	Low	Redox
Pd <sup>2+</sup> , Cd <sup>2+</sup> , Hg <sup>2+</sup>	Soft	Strong	Low	Toxic

**Table 1.1:** General properties and functions of essential metal ions. Adapted from [19].

On the one hand, *Soft* metals are usually not essential to life, they are mostly toxic when present in the body, like Pd<sup>2+</sup>, Cd<sup>2+</sup> or Hg<sup>2+</sup>. As they interact more strongly with soft bases, in a biological context that means that they tend to break hydrogen or disulfide bonds or displace essential metals from their binding site. Even in low concentrations, they can disrupt the structure of crucial enzymes and, consequently, their function. The only exception in this group is Cu<sup>+</sup> that collaborates with Cu<sup>2+</sup> in redox reactions.<sup>16,24</sup> On the other hand, *Hard* metals like Na<sup>+</sup>, K<sup>+</sup>, Mg<sup>2+</sup> and Ca<sup>2+</sup> are essential to life. Due to its low charge and large size, K<sup>+</sup> and Na<sup>+</sup> have weak binding to organic ligands, resulting in a higher mobility. They generate gradients across membranes to maintain osmotic balance and transmit nerve signals. Additionally, they act as counter ions to neutralize negative charges from proteins. As Ca<sup>2+</sup> and Mg<sup>2+</sup> have a higher charge, their binding strength to organic ligands is slightly higher, but still low compared to borderline metals. Therefore, they play an important structural role, in special Ca<sup>2+</sup>, which binding can induce structural protein changes that act as triggers of signal transmission.<sup>16,19,24</sup>

In between *Hard* and *Soft* metals are borderline metals, which are mainly transition metals and many of them are essential. These have a higher binding strength for ligands than *Hard* metals and a lower mobility, making them good candidates to coordinate ligands, improve its reactivity or acidity.<sup>25</sup> Zn<sup>2+</sup> is the only one that has only one oxidation state and has the lowest binding affinity. In consequence, it also plays structural roles like Ca<sup>2+</sup> and Mg<sup>2+</sup>, however, it has a more important role as Lewis acid in catalytic reactions. The remaining transition metals (Co, Cu, Fe, Mn and Mo) have a stronger binding to organic ligands and have a wide range of oxidation states available due to its unfilled d orbitals. Consequently, these are involved in a wide range of enzymatic reactions, mainly redox reactions.<sup>16,19</sup>

In the following section the specific roles of each one of these transition metals will be explained in terms of metalloproteins in detail.

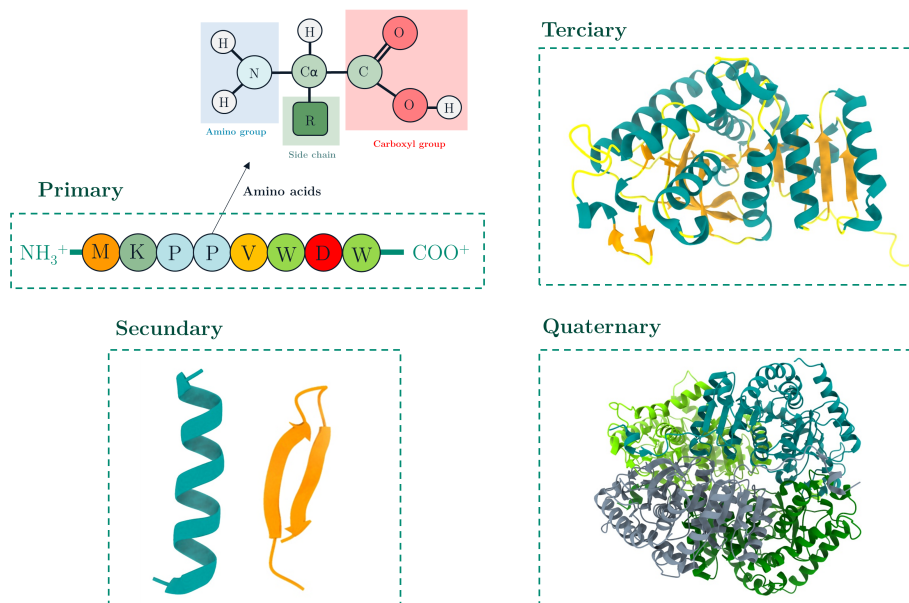
## 1.2. Metalloproteins

Proteins, the most abundant macromolecule in all living organisms, directly express genetic information. Their relevance relies on the fact that they have a very wide heterogeneous range of functions, from structural or regulation to being the catalysts of most biological reactions.<sup>26</sup>

From a structural point of view, proteins are built up from twenty canonical amino acids, which are joined through peptide bonds, constituting a unique polypeptide chains. Each amino acid is constituted by a  $\alpha$ -carbon connected to a hydrogen, a carbonyl group, an amino group and a variable side chain that defines its physicochemical properties.<sup>27,28</sup> The amino acid sequence can be arranged into different patterns,  $\alpha$ -helices,  $\beta$ -sheets or coiled-coils (secondary structure) due to local hydrogen bonds established between amino acids. As a result of the interactions between amino acids, these secondary structures are folded into three-dimensional structures (tertiary structures). These 3D conformations are mainly maintained through weak hydrophobic VdW interactions, but hydrogen bonds and ionic interactions also contribute. Some proteins contain more than one subunit, the arrangement between subunits constitutes the quaternary structure (**Figure 1.3**).<sup>27–29</sup>

X-ray crystallography, Cryogenic Electron microscopy (cryo-EM) or Nuclear Magnetic Resonance (NMR) are techniques that elucidate the 3D structure of proteins experimentally. These reveal structural features of a protein, such as shape, motifs or domains and possible binding sites of ligands, substrates or cofactors. The structure is directly correlated with the protein's function, which include a wide range of roles from structural, regulatory and hormonal to transport or storage of molecules. Most notably, proteins can also be enzymes that catalyze chemical reactions.<sup>28,29</sup>

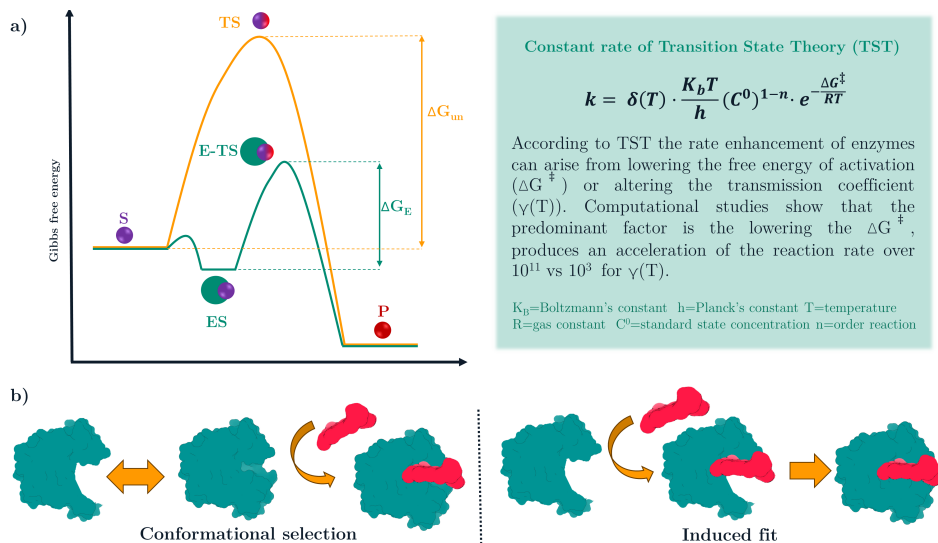
Enzymes are catalyst that accelerate reactions rates in order to occur in biologically relevant timescales without being consumed in the process.<sup>28</sup> Considering the Transition State Theory, the rate enhancement of enzymes principally arises from lowering the free energy of activation ( $\Delta G^\ddagger$ ) (**Figure 1.4a**).



**Figure 1.3:** Four levels of protein structure: primary, secondary, tertiary and quaternary.

The enzymatic reaction takes place in the active site, region containing a pre-organized environment that favors catalysis with groups that stabilize the transition state (TS), rather than the substrate, through non-covalent interactions.<sup>30,31</sup> However, how the dynamical nature of enzymes influence catalysis is still in debate.<sup>32–34</sup>

Initially, the lock-and-key model described the need of the substrate's shape to match the active site, assuming a rigid enzyme structure.<sup>35</sup> Currently, molecular recognition events are understood as dynamical processes in which proteins undergo conformational changes. Two possible binding mechanisms seem to exist (**Figure 1.4b**). Conformational selection is based on the assumption that several unbound states of the protein exist in equilibrium and the ligand binds preferably to one or several preorganized ones.<sup>36</sup> Induced fit implies that the conformational change of the receptor is a product of the entrance of the ligand.<sup>37</sup> Still, in many cases both mechanism may be occurring or one may prevail over the other depending on the concentration of the ligand and protein.<sup>38–42</sup>



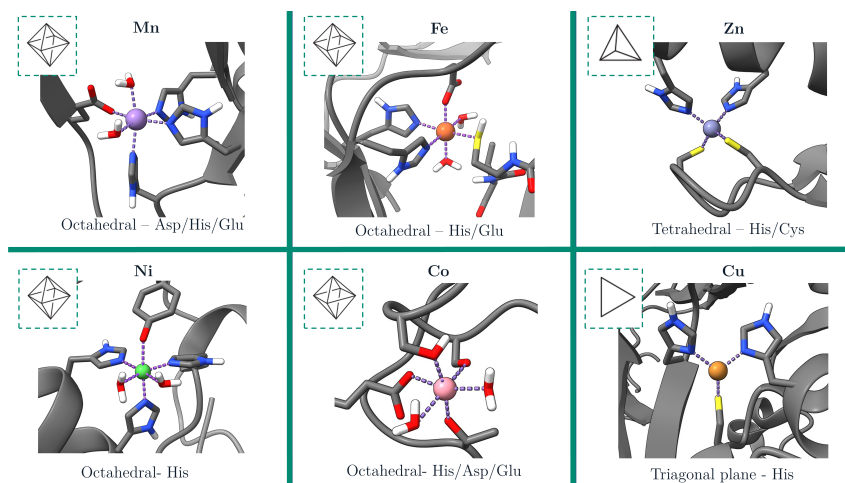
**Figure 1.4:** a) Free energy profile of catalyzed reaction by enzymes (green) and uncatalyzed reaction (orange). Constant rate defined by Transition State Theory (TST). b) Molecular recognition mechanisms: conformational selection and induced fit.

On most occasions residues of the active site alone are not able to catalyze all chemical reactions by themselves, requiring organic or metallic cofactors to carry out specific reactions. Proteins that incorporate at least one metallic cofactor, either as metal ion or as metal-containing cofactor, are known as metalloproteins.<sup>43</sup> When the protein does not contain the cofactor is known as the apo-protein, the bound form is known as holo-protein.<sup>43</sup> In coordination chemistry, a ligand is a group of atoms that surrounds a central metal ion via coordinate covalent bonds. The number of atoms involved is called coordination number and different geometrical geometries can be adopted depending on this number. The most common ligands in proteins are Asp, Glu, His, Trp, Cys, Met, Ser, Thr, Tyr, Asn, Gln and backbone, but artificial amino acids or exogenous ligands can also be ligands.<sup>44</sup> Each metal presents different amino acid and geometric propensities depending on its characteristics (**Figure 1.5a**).<sup>45</sup>

Commonly, the first coordination sphere of the metal is the group of ligands that are directly bonded to the metallic atom. The second coordination sphere refers to the ligands that are interacting with the first coordination sphere, like protein residues, molecules or ions. It seems surprising that residues far away from the metal can impact the catalysis, however, it has been demonstrated that this

second coordination sphere impact catalysis through non-covalent interactions. Consequently, the second coordination sphere can influence reactivity and selectivity by different ways: altering the redox potential of the metal, the pKa of certain residues or influencing the positioning of the substrate. The influence of the second coordination sphere has been studied in several types of metalloenzymes, containing Fe, Cu or heme for example.<sup>46</sup> Therefore, the range of the reactions that an metalloenzyme catalyses expands beyond the properties of the metal ion, as reactivity is not only defined by the metal.

Metal ions play essential roles in metalloproteins, with two principal functions: 1) enhance structural stability of the protein in order to reach a functional conformation 2) participate in catalysis either directly or as cofactors.<sup>45</sup> The latter function applies to metalloenzymes, which are present in the 76% subclasses of enzymes, indicating their ubiquitousness and involvement in nearly all catalytic reactions. According to database Metal-MaiCE, 30% of metalloenzymes are involved in redox reactions, where the metal can be directly intervening in the redox catalysis or assisting it. In metalloenzymes that are not redox, metals tend to act as activators of the reacting species or stabilize electrostatically the TS or intermediates (increasing electrophilicity or acidity).<sup>1</sup> As mentioned in the previous section, the metal homeostasis and concentration of metals need to be controlled to avoid toxicity or deficiency. Therefore, metalloproteins are also involved in metal transport, storage, detoxification and delivery.



**Figure 1.5:** Most common coordinating residues and coordination geometries of essential TM metals.

### 1.2.1. Common metals in metalloproteins

In this section, by studying the diverse roles of the most important metalloproteins, we aim to understand the significance and impact of metalloproteins on different biological systems.

**Zinc** is the second most abundant metal in enzymes. Biologically it has only one redox state, Zn (II), and due to its filled d-shell orbitals it is redox inert.<sup>47</sup> Zn catalytic sites rely on the ability of Zn to act as a Lewis acid and usually coordinates 3-4 residues and at least one water molecule. This water molecule is critical for catalysis and can be activated by ionization, polarization by surrounding protein residues or displacement by the substrate. This is the case of carbonic anhydrase (**Figure 1.6a**), carboxypeptidase A or alkaline phosphatase. Other Zn-enzymes like aminopeptidase or  $\beta$ -lactamase contain co-catalytic sites, in which several Zn ions are in close proximity (connected by amino acids), one is catalytic and the other enhances the catalysis by stabilizing the active site conformation directly or indirectly.<sup>48-50</sup>

Zn can also have a pure structural function by stabilizing the 3D structure of proteins and DNA. In these cases, Zn is coordinated to four residues in a tetrahedral geometry, usually Cys and His, unlike in catalytic roles where it coordinates water. A very common domain is zinc-finger, which is structured around one or two zinc ions bonded to His/Cys stabilizing the fold of the motif.<sup>51</sup> Initially, they were only considered DNA binding motif, but nowadays, they are also considered for proteins.<sup>48,50</sup> Zn-fingers are one of the most common DNA-binding motif found in transcriptional factors, such as TFIIIA.<sup>52</sup> Structural Zn can also be found in enzymes like alcohol dehydrogenase or endonucleases. Zn can also have a regulatory or signaling role like in glyceraldehyde 3-phosphatase or cathepsin B.<sup>53</sup>

**Nickel** is only required by a limited group of enzymes. However, they play important roles in bacteria and archaea and few higher eukaryotes.<sup>54,55</sup> Ni sites have flexible coordination geometry, and Ni exhibits versatile redox properties due to allowing several oxidation states.<sup>19,25</sup> The nine Ni enzymes known to date are divided into redox and non-redox, all related to energetic metabolism and production of gases (CO<sub>2</sub>, CH<sub>4</sub>, CO, H<sub>2</sub>) and oxygen, nitrogen and carbon cycles. These roles were crucial at the beginning of life, particularly in the pre-oxygen era, suggesting that nickel enzymes were essential then. However, as time

progressed, these roles have been acquired by much more bioavailable metals like Fe or Zn, which can carry out the same functions.<sup>55,56</sup> In bacteria and archaea, redox Ni-enzymes like hydrogenase, CO-dehydrogenase and acetyl-CoA synthase are more prominent. In eukaryote Ni can only be found in ureases, in which Ni acts as a Lewis acid catalyst (**Figure 1.6b**).<sup>56,57</sup>

While not used as widely as other metals, **cobalt** can be found in vitamin B12-containing enzymes. Vitamin B12 is an essential dietary intake for humans, however, only a fraction of bacteria can synthesize it and supply it to other organisms. Vitamin B12 contains a cobalt atom in a corrin macrocycle (derived tetrapyrrole), a nucleotide loop with a base (dimethylbenzimidazole) that coordinates with Co and with cyanide as sixth coordinating residue. In eukaryotes, the bioactive forms of B12 are Methylcobalamin, which contains a methyl group as a third component, and coenzyme B12 (adenosylcobalamin) which contains an adenosine.<sup>58</sup> All reactions are possible due to homolytic/heterolytic formation-cleavage of the (Co–C)-bond of B12 derivatives, which allows transfer of different functional groups. Coenzyme B12 participates in isomerase reactions, as in methylmalonyl CoA mutase (**Figure 1.6c**), while Methylcobalamin in methyl transferases. In prokaryotes, B12 enzyme are involved in a lot of reactions.<sup>59</sup> Co can also be found in non-corrin enzymes like Methionine aminopeptidase or nitrile hydratase. These have applications in industry.<sup>60,61</sup>

**Copper** enzymes are associated with important biological processes like electron transport, redox reactions, oxygen binding and transport and copper storage. Cu has two main oxidation states: Cu(I) and Cu(II). Cu(I) is *Soft* and Cu(II) is *Hard*, therefore, they have different properties and propensities for ligands. Cu enzymes are classified into three different types depending on their structural and spectroscopic properties.<sup>19</sup> Type I are characterized by having a single Cu center in distorted geometry with 2His and 1Cys in a trigonal planar geometry and weaker amino acid ligand. These are involved in reversible electron transfer and include enzymes like azurin, plastocyanin, nitrite reductase or laccase.<sup>62</sup> Type II have a square planar conformation with N or O ligands. This type of enzymes are able to bind O<sub>2</sub> and in consequence function as oxidases (amine oxidases or galactose oxidase) and oxidoreductases (monooxygenases and dioxygenases).<sup>62,63</sup> Type III enzymes contain dinuclear Cu centers, each coordinated to 3His, that bind oxygen. This group includes O<sub>2</sub> carriers like

hemocyanin (**Figure 1.6d**) or enzymes involved in oxidation like Tyrosinase or catechol oxidase. A part from the three types, binuclear CuA and CuB centers can be found in cytochrome C oxidase. The CuA site is in charge for electron transfer, while the CuB site binds oxygen and transforms it into water during cellular respiration.<sup>62</sup>

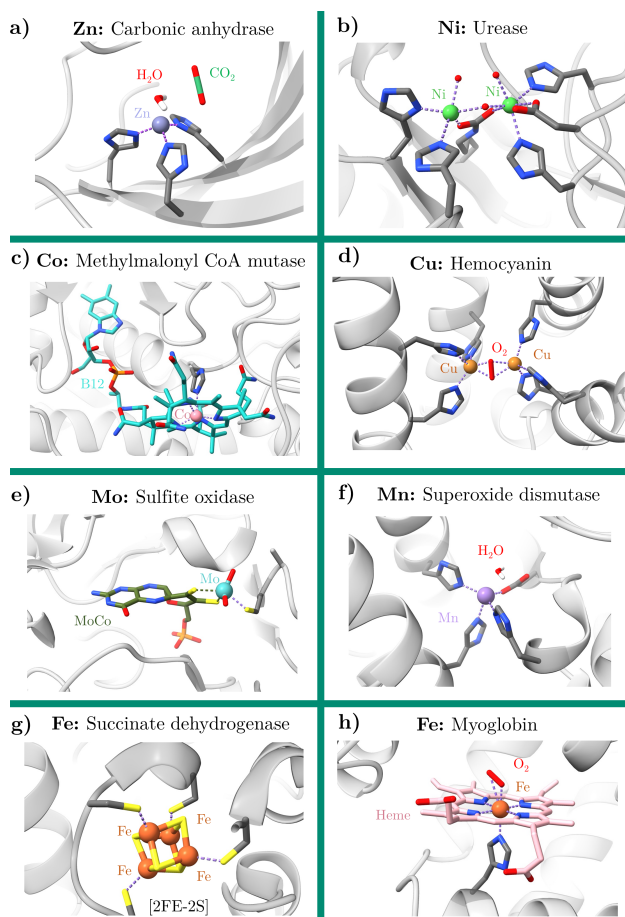
**Molybdenum** is the only second row transition metals essential for most living organism, except for thermophilic anaerobes and archaea, which contain tungsten. Mo and W enzymes do not incorporate the metal by itself, instead the catalytic cofactor molybdenum pyranopterindithiolate (MoCo) is required, except for the nitrogenase that contains the FeMo cofactor.<sup>19</sup> Its biological relevance is attributed to its range of oxidation states from IV to VI, with its intermediate state V that allows to catalyze several oxo-transfer reactions. Therefore, Mo and W enzymes catalyze oxidation and reduction of substrates involved in the nitrogen, carbon and sulfur cycles. Some of the most known enzymes are xanthine dehydrogenase or sulfite oxidase (**Figure 1.6e**).<sup>64</sup>

**Manganese** is employed by few enzymes, usually found in oxidated state Mn(II), either as a Lewis catalyst or as redox catalyst due to the possibility to oxidate to Mn(III) or Mn(IV). Mn enzymes are mainly involved in oxygenic photosynthesis in plants, algae and cyanobacteria. Specifically, the most studied one is the tetra-Mn cluster found in photosystem II that catalyzes the oxidation of water to dioxygen powered by solar energy.<sup>19</sup> Mn can also be found in superoxide dismutase, which is an antioxidating enzyme that catalyzes the dismutation of ROS into dioxygen (**Figure 1.6f**).<sup>65</sup>

**Iron** is indispensable for life, its fitness for several catalysis reactions is related to the variability of the redox potential between Fe(II) and Fe(III), that changes depending on the ligands bound. Fe(III) is considered a *Hard* metal, while Fe(II) is borderline, this difference implies that they have different ligands and geometries, which define the characteristics of the metal center. In addition, Fe is also involved in electron transfer, acid-base reactions, structural functions, oxygen binding, metal storage and transport.<sup>19</sup> One important class of Fe proteins are Fe-S cluster proteins, which contain iron atoms bound to sulphur from thiols from Cys or inorganic sulfide in different forms. These are known to be involved in electron transfer in crucial processes, like photosynthesis or mitochondrial respiration like NADH dehydrogenase or succinate

dehydrogenase (**Figure 1.6g**). Still, Fe-S clusters are also involved in assisting in Lewis acid reactions like aconitase or have regulatory roles like SoxR.<sup>19,66</sup>

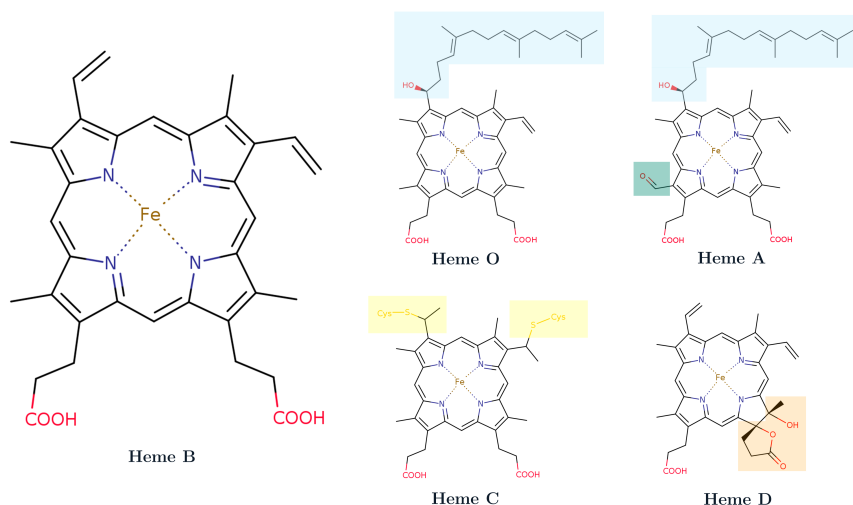
Another class of proteins is non-heme mononuclear or dinuclear iron proteins that catalyze a wide range of reactions, including dioxygenases, hydroxylases or oxidases. Lastly, there are the proteins in charge of the transport of Fe in blood and fluids like transferrin or ferritin that store Fe in the cells. One of the most prominent groups of Fe metalloproteins is heme containing proteins, like myoglobin or haemoglobin, involved in the binding of oxygen (**Figure 1.6h**).<sup>19,66</sup> As this thesis is heavily centered on heme, this is explained in more detail in the next section.



**Figure 1.6:** Active site of different metalloproteins.

### 1.2.1.1 Heme binding proteins

Heme is a prosthetic group that comprises two parts: a porphyrin and an iron atom. The porphyrin is composed of four pyrrole rings connected by methene bridges at the alpha carbons. The iron can exist in two states: ferrous ( $\text{Fe}^{2+}$ ) and ferric ( $\text{Fe}^{3+}$ ). The iron is ligated to four nitrogen atoms from the pyrrole rings and can coordinate to two more groups at its axial positions. The fifth coordination site, known as the proximal residue, is typically occupied by a His, Cys, or Met residue. A residue or a small molecule can occupy the sixth coordination site. Several important types of heme are defined by the substituents found on the pyrroles. The nature of the substituents determines specific chemical and physical properties that define each type of heme (**Figure 1.7**).<sup>67–69</sup>

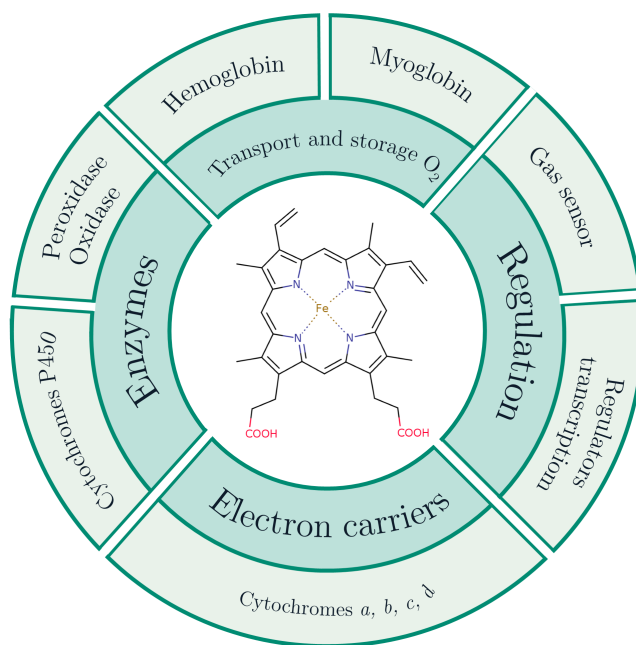


**Figure 1.7:** Chemical structure of different types of heme.

Biologically speaking, the most common type of naturally occurring heme systems are heme *b* and *c*. Heme *b* is the prototype and the precursor of all other types of heme. Heme *b* contains four methyl groups, two vinyl groups and two carboxylic sidechains and is usually non-covalently bound to proteins. Heme *c* is obtained by addition of two cysteine-sulphur groups to the 2, 4-vinyl groups from heme *b* and is covalently bound to proteins through these Cys residues as in cytochrome *c*. Compared to heme *b*, heme *c* has a wider range of redox potentials and suffers more distortions. Additionally, the covalent attachment impacts protein stability, vibrational coupling or electronic structure, and its function.<sup>70,71</sup>

Less common types of heme include heme *a*, *o* or *d*. Both heme *a* and *o* have a large isoprenoid chain attached to carbon 2, and heme *a* also has a formyl group on carbon 8 instead of a methyl. Finally, in the case of heme *d*, one pyrrole is reduced and hydroxylated, therefore it is less aromatic than others.<sup>72</sup> Despite its differences, all types of heme share common features. Heme molecules have an aromatic character and contain a very hydrophobic part that corresponds to porphyrin ring and its substituents. However, the propionates and the metal exhibit hydrophilic nature. Despite being a small molecule, heme presents these properties that contribute to a high chemical and biological versatility, making heme an excellent candidate to carry out a wide range of functions in the body.

Given the unique characteristics of heme, heme binding proteins can have four general functions as represented in **Figure 1.8**). The following section, we will delve into each function and provide the most relevant examples.



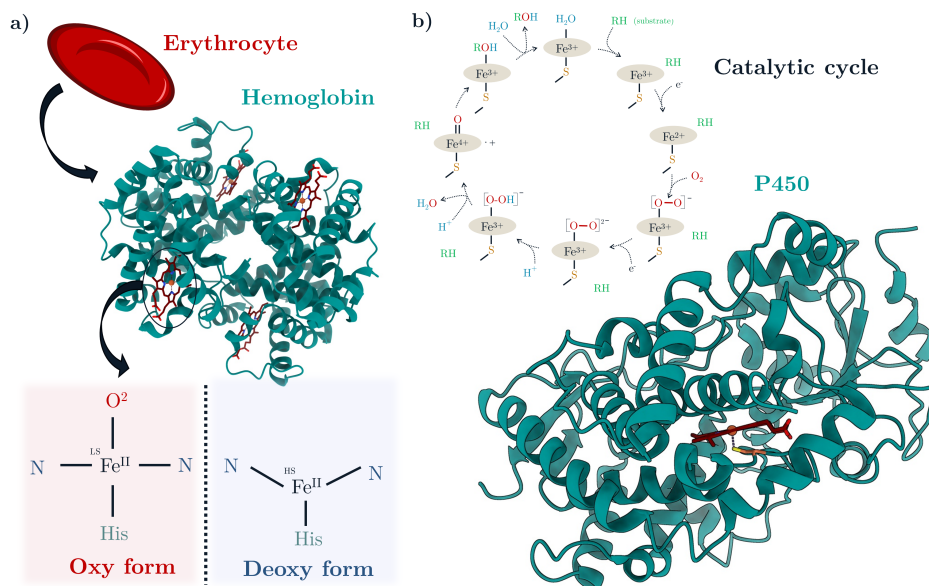
**Figure 1.8:** Summary of heme functions in proteins and examples of hemoproteins for each function.

On the one hand, the ability of heme to bind small molecules at its sixth coordinating position makes heme an excellent candidate for **transport, binding** and sensing of small molecules like  $O_2$  or NO. Usually, heme in a five-coordinate

state is involved in these functions, as one axial ligand is a residue and the vacant site is available for coordination of a small molecule. The prototypical example of these are hemoglobin and myoglobin. These globular proteins contain heme *b* in ferrous state bound to a His residue. Both proteins exist in two forms depending on if oxygen is serving as a sixth ligand or not. In the oxy form oxygen is bonded to iron, whereas in the deoxygenated form the sixth coordination site is vacant. Differences in heme planarity and spin state of iron between forms are depicted in **Figure 1.9a**. While both proteins are designed to bind oxygen, their functions are not the same. Hemoglobin transports oxygen from the lung to all the tissues of the body, whereas myoglobin takes oxygen from hemoglobin and stores it in the muscles, where it is used to produce ATP in the mitochondria.

Another crucial characteristic of heme is the ability of Fe to change its oxidation state between +2 and +3, which gives rise to two types of heme proteins: **1)** Enzymes that bind oxygen and are able to catalyze a wide range of reactions and **2)** Proteins that do not bind oxygen and function as electron carriers. The former group includes **enzymes** such as peroxidases and catalases, which use hydrogen peroxide as oxidizing agent. The former one catalyzes the oxidation of a wide range of substrates, whereas catalases neutralize hydrogen peroxide formed by other enzymes. Another class these enzymes are oxygenases, which catalyze the oxidation of substrates by transferring oxygen from O<sub>2</sub>. Depending on the number of oxygen inserted on the substrate, they are divided into two groups: monooxygenases and dioxygenases.

The most common monooxygenases are cytochromes P450 (CYP), which are one of the largest families of that can be found in all plants, mammals, bacteria or insects.<sup>73</sup> CYP are classified into two large classes based on their roles. The first class are enzymes involved in the metabolism of xenobiotics like drugs, pesticides, natural products or toxins, whereas the second are related to the regulation/biosynthesis of endogenous compounds like hormones, steroids or fatty acids. The main activity that is carried out by P450s is monooxygenase, however, they can also have an oxidase, reductase, desaturase or isomerase activity. P450s contain a heme-thiolate group and in the resting form Fe(III) is coordinated in a six way with a water molecule and a Cys or Met residue as are represented in **Figure 1.9b**.<sup>74</sup>

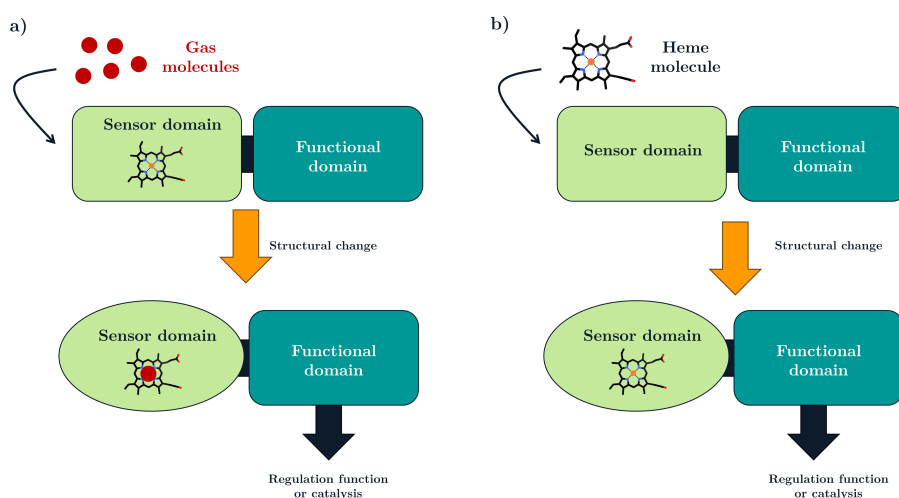


**Figure 1.9:** **a)** Erythrocytes contain hemoglobin to transport O<sub>2</sub>. Tetrameric structure of hemoglobin contains four heme molecules, which can have two forms. The oxy form is planar with low spin Fe(II) and O<sub>2</sub> bound to Fe. Deoxy form is non-planar with high-spin Fe(II) out of the plane. **b)** CYP450 structure contains one heme molecule and mostly function as monooxygenases (catalytic cycle).

**Electron carriers** correspond to cytochromes, which can transfer single electrons due to the oxidation state change from ferrous to ferric. Cytochromes vary depending on the heme molecule present (*a*, *b*, *c* or *d*). Cytochromes function as electron transporters across different cellular locations, from mitochondria, chloroplasts, endoplasmic reticulum or bacterial redox chains.<sup>19,73</sup>

Apart from these three prototypical and well-studied functions of heme, new roles have emerged in the field of **regulation**: **1)** sensor of gas molecules and **2)** regulation of transcription, translation and protein assembly.<sup>75</sup> The mechanism of action of this type of heme proteins involves binding the heme into the sensing domain of the protein, inducing a structural change that is transduced into the responsive/functional unit. This unit is the one that has a catalytic function or can regulate important physiological processes.<sup>75,76</sup> In the case of gas sensor heme proteins, it is the binding of gas molecule (NO, CO or O<sub>2</sub>) into a heme-containing sensor domain that starts the transduction process (**Figure 1.10a**). This would be the case of soluble guanylate cyclase, for which the binding of NO increases synthesis of cGMP, messenger involved in vasodilation or

platelet aggregation.<sup>76,77</sup> Regarding the other case, heme binding into the sensor domain starts the transduction process (**Figure 1.10b**). For example, heme binding to protein transcription regulators like Hap1 or NPAS2, promotes the binding to DNA and regulates transcription of respiration genes in yeast and circadian rhythm mammals, respectively. Other transcription factors like Bach1, p53 or Rev-erb $\alpha$  are also regulated by heme. Heme binding to sensor proteins like HRI regulates protein synthesis, whereas ALAS1 or Bach2 regulate protein degradation via ubiquitination.<sup>78</sup>



**Figure 1.10:** Mechanism of action of: **a)** Gas sensor **b)** heme regulated proteins. Adapted from [79].

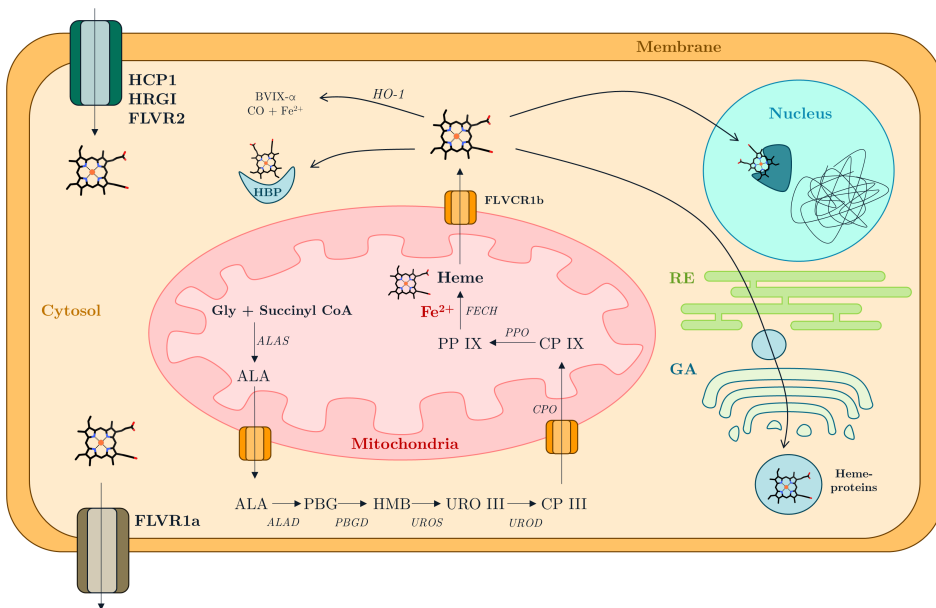
Aside from the functionality of heme, it is important to understand the complexity of its metabolism from synthesis, transport and degradation.<sup>68</sup> The relevance of this is evident by the severance of diseases that are caused by heme depletion like anemia or alterations of heme synthesis, such as porphyrias. Heme is mainly synthesized in erythrocytes and hepatocytes, but it can also be acquired through the diet. Although extensive research has been performed on the mechanism of iron acquisition, that is not the case of heme. Initially, a heme carrier known as HCP1 was identified in the small intestines. However, further investigation revealed that HCP1 is in fact a high affinity folate transporter. The study of heme binding on HCP1 will be examined in this thesis. Another potential candidate for heme acquisition is HRG1, a transmembrane transporter found in macrophages, which is also expressed in the small intestine.<sup>80</sup>

The *de novo* synthesis of heme consists of a series of highly conserved steps performed by eight enzymes. The synthesis begins at the mitochondrial matrix, where the precursor ALA is made via the C4 or C5 pathway. Subsequently, ALA molecules are transported to the cytosol, where they are condensed to form porphobilinogen (PBG). PBG is the building unit of tetrapyrrole, therefore, four PBG are assembled to form a lineal tetrapyrrole. The next enzymatic reaction involves the rearrangement and cyclization of this product, followed by a decarboxylation that leads to coproporphyrinogen III (CPP III). This product is transported again into the mitochondria, which undergoes aromatization to form protoporphyrin IX (PP IX). Finally, the incorporation of iron takes place to complete heme synthesis.<sup>73,80</sup> In **Figure 1.11** are represented all intermediates and enzymes involved in this process, which differ depending on the cell type.

Once heme is synthesized in the mitochondria it is transported to the cytosol through FLVCR1b. From the cytosol it can be transported to the nucleus to bind transcription factors or to Endoplasmic Reticulum (RE) and Golgi Apparatus (GA) to be incorporated into synthesized proteins (**Figure 1.11**). As mentioned previously, the concentration of metals has to be controlled; this is also the case of heme. Free heme can be toxic through two mechanisms: it can induce the production of ROS (reactive oxygen species) or it can induce cell lysis by inserting itself into the phospholipid bilayer due to its amphipathic property. Consequently, there are strategies in order to detoxify free heme:<sup>80,81</sup>

- Extracellular toxicity is solved by soluble scavengers like hemopexin or albumin, which bind free heme in the plasma, particularly heme that is released from the hemolysis of red cells.
- Extracellular heme can be imported into the cytosol through transporters HGR1 and FLVCR2. In the cytosol heme can be detoxified by two processes. Heme oxygenase (HO) degrades heme in order to reduce its toxicity. HO is an enzyme that degrades free heme by catalyzing the conversion of heme to biliverdin IX- $\alpha$  (BVIX- $\alpha$ ), iron and CO (**Figure 1.11**).
- Heme binding proteins (HBP) in the cytosol, like HBP22, fatty acid binding protein or glutathione S-transferase bind heme and contribute to its detoxification.<sup>82</sup>
- Antioxidant enzymes like catalases (CAT), peroxidases (POD) or superoxide dismutase (SOD) are also present to prevent oxidative stress.

**Figure 1.11:** Heme transport and synthesis inside mammalian cell. Enzymes are in *italics* and transporters in **bold**. Abbreviations: ALA,  $\delta$ -aminolevulinic acid; ALAD, ALA dehydratase; ALAS, ALA synthase; PBG, porphobilinogen; PBGD, PBG deaminase; HMB, hydroxymethylbilane; URO, uroporphyrinogen; UROS, URO III synthase; UROD, URO decarboxylase; CPP, coproporphyrinogen; CPO, CPP oxidase; PPO, protoporphyrinogen oxidase; PP IX, protoporphyrin IX; FECH, ferrochelatase; HO-1, heme oxygenase, BVIX- $\alpha$ , biliverdin IX- $\alpha$ ; CO, carbon monoxide, Endoplasmic Reticulum (RE) and Golgi Apparatus (GA).



### 1.2.1.2 Relevance of heme binding processes

In view of the relevance of heme and its regulation, several studies have been performed in the field of predicting the binding of ligands or heme itself and its structural consequences. As explained previously in section 1.2, it is important to emphasize that there are two possible molecular binding mechanisms: conformational selection and induced fit.

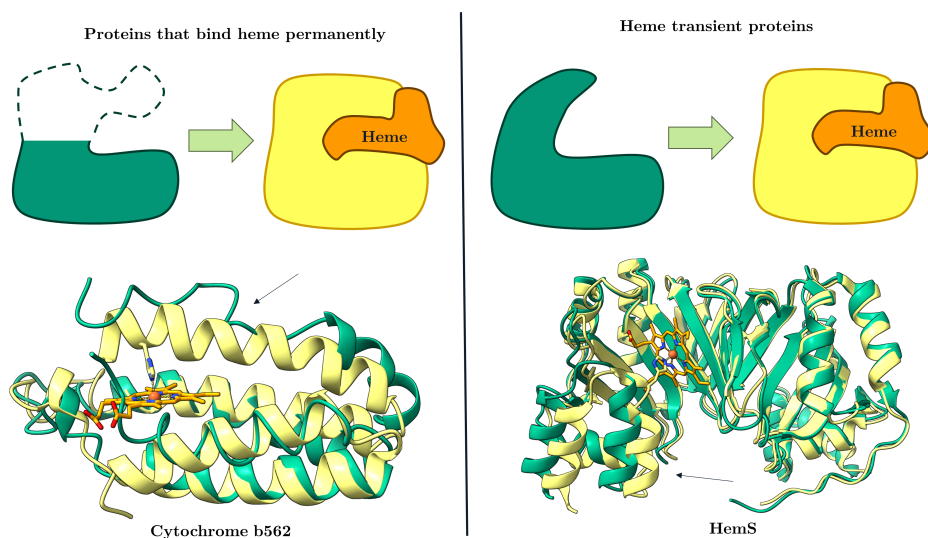
When focusing on heme proteins, the cooperative binding of oxygen and conformational changes of hemoglobin have been studied through the past 50 years.<sup>86–88</sup> At the beginning it was thought that hemoglobin followed either the induce fit or the conformational selection model. However, the current understanding is that both phenomena are likely to participate.<sup>89</sup> Overall, hemoglobin follows a conformational selection model. It has two states in equilibrium, the relaxed (R) with high affinity for dioxygen and the tense (T) with low affinity. In the deoxy form the five-coordinated iron is out of plane, however, when oxygen binds to heme iron acquires an in-plane disposition that causes the movement of a helix that results in breaking salt bridges between subunits and the consequent transition to the relaxed form.<sup>90</sup> The Tertiary Two State model states that the induced local transitions of the tertiary structure is what in reality is coupled to these quaternary transitions.<sup>91</sup>

Research of the different molecular binding mechanism to hemoproteins as CYP through kinetics or computational modeling has been performed over the years. The binding of substrates to P450s 2C8, 2D6, 3A4, 4A11, 17A1 and 21A2 was predicted to be conformational selection, whereas in the case of P450 3A4 it was assigned to induce fit.<sup>92–94</sup> Therefore, in the case of P450, for the current state of the art seems that conformational selection may prevail over induce fit, however, in some cases like P450 2E1 both mechanisms appear to be a good fit.<sup>95</sup>

Nonetheless, the molecular mechanism under which the heme binds to its receptor has yet been rarely explored. Spectroscopic and crystallographic data on heme proteins have shown that the tertiary structure of the apo and holo forms are in general similar.<sup>96</sup> Two different mechanisms seem to irrupt depending on the heme protein.<sup>97</sup> Systems that bind heme permanently like Myoglobine, Cytochrome b5 or Cytochrome b562, the binding of heme causes some secondary structure rearrangement and folding, commonly in the proximal region of the heme.<sup>98–101</sup> In the case of cytochrome *c*, the covalent attachment of heme to the

protein is what initiates the folding process.<sup>102</sup> In others cases where heme binds transiently, only subtle change in the relative orientation of the secondary structure are observed, those systems are heme-chaperones, like HemS systems or hemophores from the HasA family.<sup>103,104</sup> Recently, heme binding to transcription regulators has been studied by different experimental techniques.<sup>105–107</sup> Comparison of two heme binding mechanism with cytochrome b562 and HemS are represented in **Figure 1.12**.

Even though that there are some experimental studies for heme binding for the mentioned systems, there is no clear molecular description of heme uptake. Therefore, molecular modeling approaches would be ideal candidates to study in more depth the molecular processes that lead to conformational changes upon heme binding. In the case of ligand binding, there are computational studies that describe the molecular processes related to oxygen binding to hemoglobin or ligand binding to P450.<sup>93,108,109</sup> Still few studies have been performed related to heme binding, this is a field that is explored in this thesis.



**Figure 1.12:** Heme binding mechanism for a) proteins that bind heme permanently and b) Heme transient proteins. Adapted from [97].

### 1.3. Artificial metalloenzymes (ArM)

The preceding sections have established the relevance of metals in nature, their fundamental properties, and their role in providing natural enzymes with the ability to catalyze a wide range of chemical reactions. Transition metals are versatile species and their complex with organic moieties provide a wide spectrum of chemical reactions. Therefore, the combination of transition metals with proteins led to the field of Artificial Metalloenzymes (ArMs) able to catalyze reactions absent from nature.

ArM are hybrid catalysts that are obtained by integrating a catalytic organometallic compound into a protein scaffold.<sup>110</sup> Those biohybrids can catalyze new-to-nature reactions (defined by the metallic center) as well as substrate selectivity and catalytic specificity (controlled by the biological environment). This section of the manuscript will focus on the concept of ArMs, explore their design strategies, and their applications. First, a brief historical overview will be provided to understand how the field of ArM has been growing over the past few decades.

#### 1.3.1. Overview of ArM history

Considering a generic definition of ArM, the first occurrence was reported in 1956 by Fujii *et al.* They described how the adsorption of palladium chloride into silk fibers catalyzed the asymmetric hydrogenation of amino acid precursors.<sup>111</sup> In the 1960s, several studies focused on substituting natural metal ions with non-biological ones. Although the enzymatic reaction was preserved, exchange of natural metals led to changes of selectivity, alterations of activity or  $K_M$  of natural metalloenzymes. For example, exchange of Zn(II) for Co(II) in carbonic anhydrase enhanced/decreased certain activities, exchange of Ca(II) for Sr(II) in staphylococcal nuclease change the selectivity of the reaction and exchange of Ca(II) for Nd(II) accelerated the rate of activation.<sup>112–115</sup> It was not until mid-1970's that first ArMs based on the interaction of organometallic complexes with proteins with new-to-nature reactivity were reported.

In 1976, K. Yamamura and E.T. Kaiser, reported that the exchange of Zn(II) for Cu(II) in Carboxypeptidase A resulted in an alternative catalysis function, changing from native peptidase activity to oxidation of ascorbic acid.<sup>116</sup> This

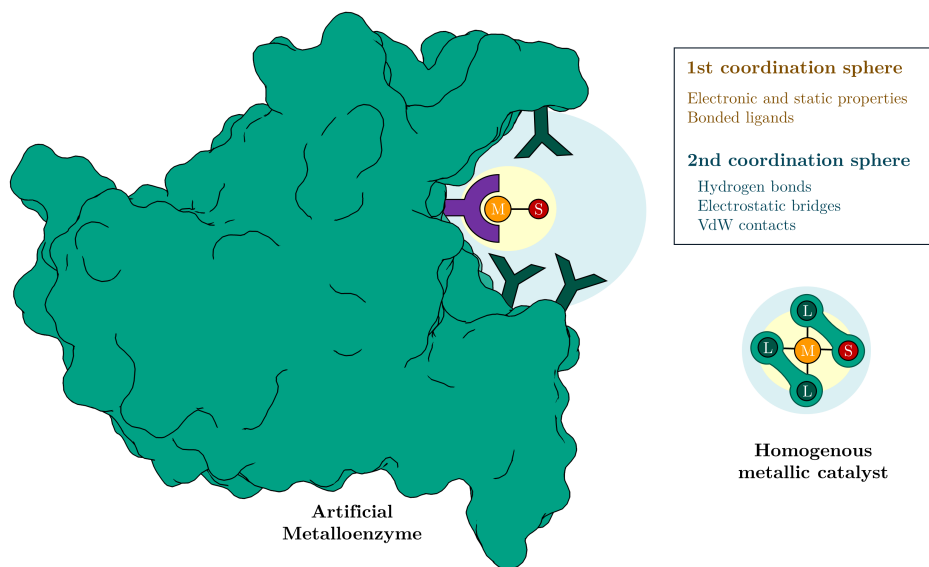
finding revealed the potential of metal substitution in enzymes to introduce new catalytic capabilities. Two years later, Wilson and Whitesides pioneered the strategy of anchoring a synthetic metallic cofactor to a protein scaffold. Using avidin's high affinity for biotin, researchers complexed avidin with a biotinylated Rhodium(I) homogeneous catalyst. This hybrid performed the enantioselective hydrogenation of  $\alpha$ -acetamidoacrylic acid.<sup>116</sup>

Despite these promising results, the true potential of ArM was not exploited until the beginning of the XXI<sup>st</sup> century. Initially, the focus was on developing non-natural metallic cofactors because it was easier to synthesize enantiopure metallic ligands than expressing, purifying, and engineering proteins. However, the field of ArM was further expanded thanks to the rise of structural biology and bioengineering including new technologies like directed evolution. These methods based on random mutagenesis allow for a more efficient optimization of ArM and reduce the screening efforts<sup>15</sup>

### 1.3.2. Development of ArM

The motivation for designing ArM is to combine the best of both homogeneous catalyzes and enzymology in one entity. The homogeneous catalyst provides the chemical reaction, while the biological scaffold provides selectivity under mild conditions. When working with a conventional transition-metal catalyst, the catalytic activity, specificity, as well as the selectivity, are all determined by the first coordination sphere. The challenge resides in designing ligands for the metal able to control the reaction profile and the selectivity.<sup>117,118</sup>

In ArM, only the reaction is controlled by the first coordination sphere of the metal; all other variables regarding activity, specificity, and selectivity are controlled mainly by the second sphere of coordination of the metal provided by the biological environment (**Figure 1.13**). This control can be handled with many tools available in bioengineering like the introduction of mutations in the protein that could impact on hydrogen bonds, hydrophobic contacts or electrostatic bridges between the biological host and the cofactor or the substrate.<sup>117</sup> Some of these modifications aim to accommodate better the catalytic cofactor, both kinetically and catalytically.<sup>119</sup>



**Figure 1.13:** 1st and second coordination sphere of ArM and conventional homogeneous catalyst.

The biomolecular scaffold defines a specific micro-environment, including certain chirality. As a result, by changing the second coordination sphere, certain regio- or enantiospecificity can be favored.<sup>120,121</sup> Furthermore, modifying the second coordination sphere can lead to kinetic changes, like the increase of the reaction rate, even in rare cases, reaching the rate of natural enzymes.<sup>122,123</sup>

The mentioned characteristics clearly showcase that the potential to optimize the reaction is increased in ArM compared to homogeneous catalysts.<sup>117</sup> Among other advantages of ArMs is the ability to work under mild conditions. Furthermore, enzymes have extended catalyst lifetime and higher *in vivo* compatibility, resulting in better candidates for biocompatible applications. Even though, compared to homogeneous catalysts, ArMs have narrow substrate scope and a limited choice in the first coordination sphere.<sup>118,124</sup>

The process of developing ArM can be divided into two key steps. The first step is the discovery of the ArM, where the focus is on identifying the adequate initial substrate, metal, and protein scaffold and how to combine them. Once a first ArM with the desired activity has been identified, the next step is its optimization, where the interplay between all components is improved (i.e. enantioselective profiles). The following two subsections will focus on each step.

### 1.3.3. Discovery of ArM

At the discovery stage, three essential elements need to be considered: the scaffold, the metallic cofactor and the strategy to assemble these two components. These play an essential role in determining the performance and catalytic capabilities of the ArM.

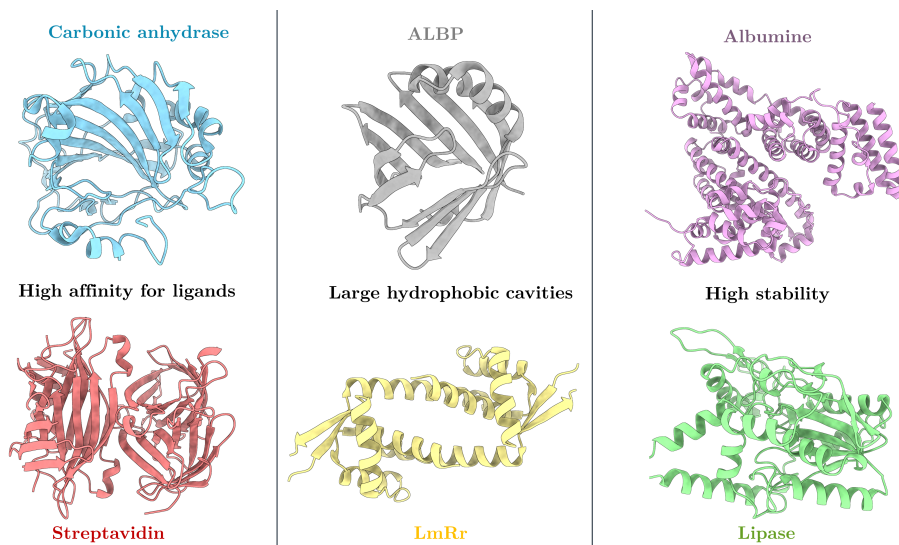
#### 1) Metallic cofactor

The selection of a specific metallic cofactor is crucial, and the choice usually depends on the desired reaction that is going to take place in the ArM. It is essential to understand the nature of the reaction to determine the optimal metallic cofactor to be employed. Still, there are some specific characteristics that all metallic cofactors should have. The most important requirement is to be orthogonal to the protein scaffold, meaning that it must be selective and avoid reacting with any of the functional groups of the protein. Furthermore, cofactors should exhibit tolerance to water as the experimental methodology of ArM involves working with aqueous solvent.<sup>119</sup>

#### 2) Scaffold

The selection of the protein scaffold is a critical aspect of the design of ArM, as protein residues serve as ligands and determine the catalytic micro-environment. According to a recent review of Thomas Ward, 83 different scaffolds have been used so far for the design of ArM.<sup>15</sup> They can be classified into two families depending on their biological scaffold: those with pre-existing protein folds and those with *de novo* (new-to-nature) scaffolds. By far, most of them are constructed on the first concept, using pre-existing protein scaffolds. The challenge resides in foreseeing a good match between the protein and the cofactor. In general, this is mainly performed by identifying cavities in each protein environment and assessing, sometimes with help of molecular modeling, the most interesting complementarities. **Figure 1.14** illustrates the most common protein scaffolds and their main characteristics that makes them suitable for ArM.

When designing an ArM using an **existing protein scaffold**, a relevant aspect to consider is the selection of the protein pocket for the binding of the metallic moiety. Two main approaches exist: using an already existing pocket or a new one. Using an existing pocket relies on the principle of taking advantage of the native high binding affinities of certain proteins for specific ligands and modify



**Figure 1.14:** Most common existing scaffold for the design of ArM.

them to incorporate metals. The advantage is that the second sphere is already adapted for binding the compound and can be further optimized. Examples of this family are streptavidin (Sav) or carbonic anhydrase, which are very common scaffolds. Another option is repurposing proteins that already harbor metal ions or complexes, which are good candidates for exchange metal methodology, like heme binding proteins.<sup>119,125</sup>

On the other hand, designing a new site in a natural protein, the ideal protein scaffold would need to naturally possess a big empty cavity that could be easily adapted to incorporate a metallic cofactor through bioconjugation or unnatural residues. Examples of this series are Adipocyte Lipid Binding Protein (ALBP), Sterol carrier protein type 2 (SCP-2L) or multidrug resistance regulator LmrR, which contain large hydrophobic cavities. Another approach would be to look for non-metallic proteins that contain pre-organized residues that could coordinate metals.<sup>119,125</sup> For instance, screening have been performed to look for possible binding sites of uranyl<sup>126</sup> or for facial two-histidines one-carboxylate (FTM) binding motif<sup>127</sup> in PDB to find new protein scaffolds. Applying metal binding predictors based on preorganization would also be handy for the search for new ArM scaffold. This is one of the software that has been developed in this thesis.

In all cases, there are some factors need to be considered when identifying the best possible scaffold. Protein stability is essential as it allows experiments in different conditions and allows mutations without compromising the function. Others to consider include the total charge, pH and temperature stability, availability in the market or resistance to organic solvent. As a result, protein scaffolds that fulfil these requirements, like lysozyme, bovine serum albumin (BSA) or lipase, have been widely employed for ArM's design. When extreme thermal stability is a requirement, proteins from thermophile organisms can be utilized, such as tHisF from *Thermotog maritima*.<sup>119,125</sup>

**De novo scaffolds** represent the most challenging option as it consists of building a polypeptide sequence from scratch without any natural protein equivalent. The difficulty of this approach relies on the fact that the designed protein should fold into a specific 3D structure capable of accommodating the metallic cofactor. Predicting the final folded structure and its stability presents a challenge.<sup>128</sup> De novo approaches involve building from scratch using naturally occurring scaffolds as 4-helical bundles, coiled coils or TIM barrels and incorporating a metallic cofactor into these frameworks.<sup>129–132</sup> Successful examples involve a combination of computation and experimental work.

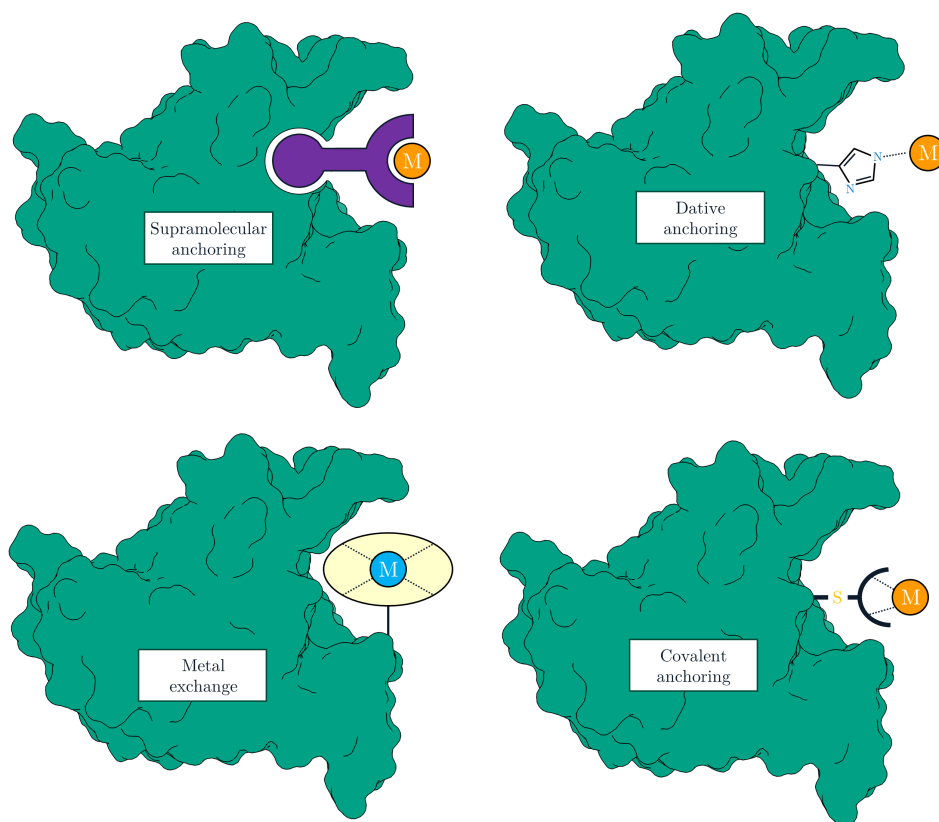
Several groups have focused on assembling 4-helical bundles (called 'maquettes') and placing residues at strategic positions to bind different metallic cofactors like heme, flavins, chlorins or iron-sulfur clusters. The presence of different cofactors dictated the different possible oxidoreductase functions, that can range from electron transfer, O<sub>2</sub> binding or peroxidase.<sup>133,134</sup> A very successful project involves the *de novo* design of 4-helix bundles that mimic natural di-Fe proteins, which were tuned to perform oxidase, oxygenase or ferroxidase activities. Using a computational approach, first the backbone is defined by the di-metallic center geometry, which then dictates the position of the anti-parallel four helices. Algorithms are then used to improve the design with second shell hydrogen bonds interactions and the addition of interhelical loops or a stabilizing apolar cores. The methods described were shown to be suitable to design highly stable and robust artificial catalysts.<sup>130</sup>

Other *de novo* scaffolds based on  $\alpha$ -helices are  $\alpha$ -Rep proteins, which derive from a sub-class of HEAT-like repeat thermostable proteins. An homodimeric form,  $\alpha$ -Rep A3, presents a wide cavity that has been used to covalently bound metal

complexes.<sup>135</sup> Ricoux *et al.* designed a Diels-Alderase by anchoring covalently Cu-phenanthroline or Cu-terpyridine to previously mutated Cys residues. Results show that the systems are able to catalyze Diels-Alder reaction of azachalcone with cyclopentadiene with up 52% enantiomeric excess, but low yield.<sup>136</sup>

### 3) Assembly

The last aspect that should be considered when designing ArM is how to assemble the metallic cofactor and the protein scaffold. There are four main approaches for assembling ArMs: **a)** metal exchange **b)** dative anchoring, **c)** covalent anchoring and **d)** supramolecular anchoring. In the following section, they will be explained, along with examples to illustrate the process. **Figure 1.15** represents different assembling strategies.



**Figure 1.15:** Four approaches for assembly of ArM: **a)** metal exchange **b)** dative, **c)** covalent and **d)** supramolecular anchoring. Adapted from [118].

The **metal exchange** approach is based on substituting the native metal ion or metal complex from a natural metalloprotein with a different metal ion or complex. As a result of this substitution, the ArM has a different catalytic function compared to the original natural form. The key advantages of this technique are that the non-natural metal provides a new reactivity, and the second coordination sphere is already customized for binding a metal. There are two distinct candidates in this approach: one would correspond to a system in which the metal is in the form of a metal ion, and it directly coordinates with the residues of the protein. while in the other case the metal can be part of a prosthetic group, like heme.<sup>118</sup> This method requires that the apo form of the protein is stable and exhibits promiscuity for other metals besides its natural one. Additionally, it is desirable that the non-natural metal shares some properties with the natural metal like the coordination geometry or its character (*Soft* or *Hard*) in order to interact as optimal as possible with the protein.<sup>125</sup>

The first ArM developed by this approach was reported in 1976 and over the next 50 years this approach has been used by several groups. For example, exchanging the native  $\text{Zn}^{2+}$  in carbonic anhydrase (CA) for  $\text{Mn}^{2+}$  resulted in peroxidase activity instead of hydrolase activity. Specifically, this ArM catalyzed an enantioselective epoxidation of styrene with a maximum of 67% ee, a value comparable to natural peroxidases.<sup>137</sup> Another example involves the enzyme laccase, in which the native  $\text{Cu}^{2+}$  was substituted by osmate. This metal exchange led to a new activity, the dihydroxylation of alkenes with a high enantiomeric excess of 92%. Additionally, the laccase was conjugated with poly(2-methyloxazoline), allowing the reaction to occur in organic solvent.<sup>138</sup> The most common examples include repurposing heme proteins; this will be treated in detail at section 1.3.6.

The **dative anchoring** method involves a dative bond between an unsaturated metal and a functional protein group, such as Cys, Glu, Asp or Ser. The dative interaction can also be established between the metal and an unnatural amino acid.<sup>119,124</sup> Using this methodology, Ward *et al.* were able to design an ArM for olefin dihydroxylation by combining Streptavidin (Sav) with  $\text{OsO}_4$ . Using directed site mutagenesis, they were able to increase selectivity and change its enantiopreference.<sup>139</sup> Roelfles *et al.* were able to create ArM by introducing *in vivo* BpyAla (unnatural amino acid) in protein LmrR. This system successfully catalyzes asymmetric Friedel–Crafts alkylation with 83% of enantioselectivity.<sup>140</sup>

The **covalent anchoring** approach is achieved through a process known as bioconjugation. This method involves an irreversible reaction between the metallic cofactor and a functional group the protein scaffold. In this process, it is required that the metallic cofactor contains a functional group compatible with condensation with the residue of choice. These modifications constrain the cofactor at a given location in the protein. However, if there are other residues with the same properties, they need to be removed to avoid several active metal species. Among the most common residues of choice is cysteine (Cys) due to its nucleophilic character, high reactivity and low abundance in proteins. Other residues employed are Lys and Ser, but less commonly due to the difficulties of derivatizing them. The main advantage of this approach is that a broader range of scaffolds can be used, as the only requirements are a big empty cavity and a single suitable residue for bioconjugation.<sup>118</sup>

The large protein cavity and the presence of a unique Cys makes adipocyte lipid binding protein (ALBP) the perfect candidate to attach cofactors covalently. In a pioneer study, DiStefano group took advantage of ALBP to covalently attach a phenanthroline-Cu(II) moiety. The resulting ArM catalyzed the hydrolysis of amino acids with a 86%ee.<sup>141</sup>  $\beta$ -barrels with large cavities, such as nitrobidin or FhuA are also good candidates for covalent anchoring. However, in these cases both require specific mutations to Cys before the bioconjugation.<sup>142,143</sup> ArM designed by anchoring to Ser or Lys has also been developed.<sup>144,145</sup>

The **supramolecular anchoring** approach takes advantage of the fact that certain proteins have a strong affinity for specific cofactors, inhibitors or substrates. By modifying these cofactors to incorporate metallic moieties, their properties are altered, enabling them to catalyze a desired reactions. The success of this approach relies on the robust non-covalent interactions between the protein and the metallic cofactor, which guarantee the specific location into an already known binding site. Consequently, this allows modification of the specific binding site and have more control over the reaction. Compared to covalent anchoring, the synthesis of the metallic cofactor is easier as it does not require specific reactive functional groups or bioconjugation. The main drawback of this approach is that the range of scaffolds is more limited as there are not many cases of these strong non-covalent interactions. One of the most reliable combinations for ArM design is the streptavidin-biotin systems, as the streptavidin-biotin interaction is one of the strongest interactions known.<sup>118,119</sup>

As mentioned in the introduction of this section, Whitesides was the first to use this approach to design a hydrogenase ArM using an avidin-biotin strategy. Inspired by this work, Ward's group reported an ArM based on biotin-streptavidin technology for the hydrogenation of amino acids with high substrate specificity and selectivity. Instead of avidin, streptavidin was used as the affinity with the [Rh(cod)-(biot-1)]<sup>+</sup> cofactor was higher. This system was optimized by combining genetic mutagenesis on position S112 with a collection of different metallic cofactors with two different substrates. From this study it was revealed that changes in the metallic cofactor contributed to more diversity than genetic changes, although different mutations at specific positions inverted selectivity.<sup>146</sup> This is one example of this methodology that has successfully employed the biton-Sav technology, but, from this point up until now, 12 ArMs based on this strategy have been successful to carry out a wide range of synthetically relevant reactions that include: hydrogenation, alcohol oxidation, sulfoxidation, dihydroxylation, hydroamination, allylic alkylation, transfer hydrogenation, Suzuki cross-coupling, C-H activation and metathesis.<sup>147</sup>

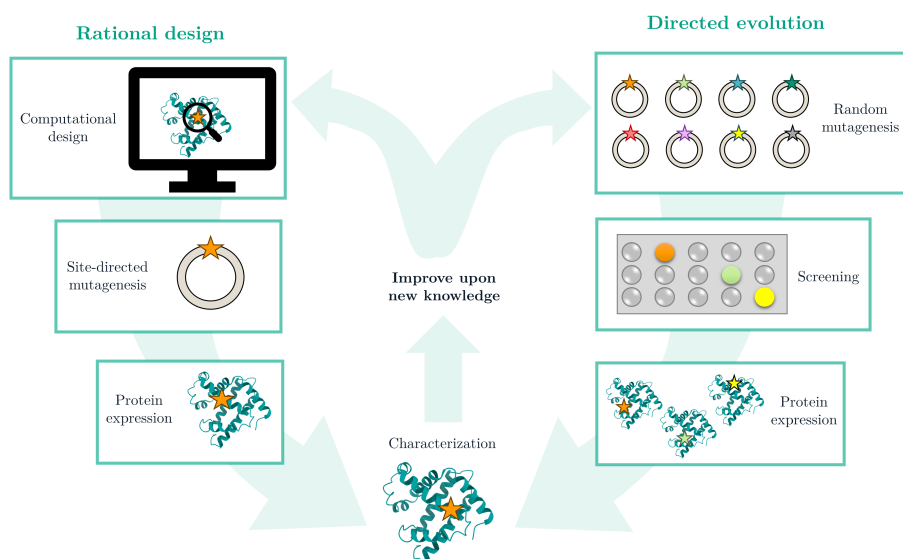
### 1.3.4. Optimization of ArM

In ArM, the scaffold and the metallic cofactor can be optimized independently or combined, which is considered a chemogenetic exploration. The metallic cofactor is normally optimized by rational design, by analyzing the structural data available in the ArM system. On the other hand, the scaffold can be optimized by two different techniques: rational design or directed evolution (**Figure 1.16**).

**Rational design** consists of optimizing the scaffold by consciously performing specific binding site changes. Structural or biochemical knowledge is needed to allow the binding site and detect which residues should be mutated. If an X-ray or NMR is available, visual inspection of the binding site-specific residues can be pinpointed for site-mutagenesis to alter the anchoring or activity of the metallic cofactor. In this approach, computational resources are usually involved as they can shed light on the most important interactions or dynamics of the protein, as the crystal is static. This approach can only be successful if there is enough knowledge of the system. However, when this is not the case, directed evolution can be of interest.<sup>119,148</sup>

**Directed evolution** is based on Darwin theory of how enzymes have evolved through the introduction of mutations. A library of ArM is obtained by the introduction of random and massive mutagenesis. Screening and selecting the most active variants are performed on several iterative processes until the desired activity or selectivity is achieved. This approach does not require previous structural knowledge, but an initial minimum activity is required<sup>119,148</sup>

The more recent studies on ArM design tend to incorporate both strategies, rational design and directed evolution, in an iterative way until the system has been optimized. Furthermore, both experimental and computational methods are performed to improve the optimization process.



**Figure 1.16:** Rational design strategy vs directed evolution approaches.

### 1.3.5. Reaction scope of ArM

Several reviews have tackled the reaction scope of ArM.<sup>15,118,149</sup> To see a general overview of all the reactions that ArM can carry out, **Table 1.2** summarizes most important reactions that have been carried out so far. For each ArM, the scaffold, metallic cofactor and assembly strategies are indicated to see the variety of techniques and systems that have been developed over the years.

Reaction type	Reaction	Metallic Cofactor	Scaffold	Assembly approach	Optimization strategy	ArM Reaction
Reduction	Hydrogenation	Biotin-Rh	Streptavidin	Supramolecular Anchor	Site mutations and cofactor optimization	Asymmetric hydrogenation of dehydroamino acid derivatives
		mALs-Ru/Rh	Papain	Supramolecular Anchor	Cofactor optimization	Hydrogenation of ketones
	Transfer hydrogenation ATHases	Biotin-Ir	Single chain dimeric Streptavidin	Supramolecular Anchor	Site mutations	Transfer hydrogenation of prochiral imines
		Benzene-Ru	$\beta$ -lactoglobulin	Dative	-	Transfer hydrogenation of aryl ketone
		Ir-catalyst	Azotochelin – CeuE siderophore	Redox Reversible anchor	Site mutations	Transfer hydrogenation of chiral imines
		Aryl-sulfonamide-Ir	Human carbonic anhydrase II (CA-II)	Dative	Computational redesign	Transfer hydrogenation of imines
Oxidation	Peroxidation	Fe-TPA	Steroid carrier protein 2L (SCP-2L)	Covalent	Computational rational design	Oxidation of the benzylic alcohol of lignin
	Dihydroxylation	OsO <sub>4</sub> -ester	Bovine Serum Albumin (BSA)	Dative	-	cis-Hydroxylation of alkenes
Hydration	Hydrolysis	Zn	Cyt <sub>562</sub>	Dative	Directed evolution	B-lactamase activity: Ampicillin hydrolysis
C-C bond	Suzuki coupling	Biotin-Pd	Streptavidin	Supramolecular Anchor	Directed evolution and cofactor optimization	Synthesis of biaryls
	C-H activation	Ir(Me)-PIX	P450	Metal exchange	Directed evolution	Intramolecular C–H bond amination of sulfonyl azides
	Cyclopropanation	Heme	Lactococcal multidrug resistance regulator (LmrR)	Dative	Site mutations and computational study	Cyclopropanation of styrene
		Rh <sub>2</sub> -catalyst	prolyl oligopeptidase (POP)	Covalent	Site mutations	Cyclopropanation of olefin
		IrMe-PIX	Myoglobin	Metal exchange	Directed evolution	C-H insertion and cyclopropanation of olefin
	Diers-alder	Cu-phenanthroline	$\alpha$ -Rep A3	Covalent	-	Diels-Alder cycloaddition
		Cu(II)	Nitrobindin-Pyr	Dative	Cofactor optimization	Diels-Alder between azachalcone and cyclopentadiene
	Fiedel-Crafts	Cu(II)	Different TetR proteins	Dative	Different proteins	Friedel-Crafts alkylation of indoles
		Binaphthyl derivative-Cu/Pd	Monoclonal antibodies (mAbs)	Supramolecular Anchor	-	Friedel-Crafts alkylation

Table 1.2: Reaction scope of ArM developed over the years.

### 1.3.6. ArM containing porphyrin

As introduced in section 1.2, heme can carry out a wide range of functions. Consequently, incorporating heme into proteins to design ArM expands the scope of reactions that the ArM can catalyze due to the versatility of heme. Over the last few years, directed evolution has been performed on heme-containing enzymes, mainly P450, cytochrome *c* or myoglobin to obtain evolved species with new activities, stereoselectivities or new-to-nature reactions.<sup>150</sup> Examples include enzymes that catalyze C-H, N-H or Si-H insertion<sup>151–153</sup>, alkene cyclopropanation<sup>154</sup> between others. These repurposed natural hemeproteins studies significantly impacted the design of ArM. Another common approach that provides an easy and reliable way to obtain new catalytic activities in heme-reconstituted ArM is the metal exchange. Heme enzymes' catalytic activities can be altered by modifying the heme moiety, either by substituting the type protoporphyrin or certain functional group, or by exchanging the metal ion.

For instance, Hartwig's group designed ArMs based on a CYP119 from a thermophile organism. The native Fe was exchanged for Ir, resulting in an ArM that catalyzed the enantioselective insertion of carbenes into C–H bond. To optimize the ArM performance, directed evolution was applied, which lead to an ArM with high selectivity (up to 98% enantiomeric excess) and high productivity (activities similar to natural enzymes).<sup>155</sup> Directed evolution with the same system has lead to different variants with other activities.<sup>156,157</sup> In the alternative approach of changing the scaffold of heme, there are several examples using a simple scaffold as myoglobin (Mb). Hayashi *et al.* replaced protoporphyrin IX for porphycene, maintaining Fe as the metal center and not introducing any mutation. This change allowed Mb to catalyze the cyclopropanation of sterene and demonstrated the impact of the heme scaffold to the enzyme activity.<sup>158</sup>

Another very popular approach with heme are ArM obtained by inclusion of heme to *de novo* scaffolds. The 'maquette' approach has been used to incorporate up to four heme molecules into 4-helical bundle proteins for O<sub>2</sub> binding or electron transfer.<sup>159,160</sup> This initial works led to nowadays the design of cytochrome *c* type maquettes with peroxidase, cyclopropanation activity or N-H insertion. C45 is produced *in vivo* and one heme molecule is covalently attached.<sup>161</sup> However, the first crystallographic evidence was obtained recently. Computational studies on 'maquette' PS1 led to a *de novo* ArM based in

Mn-porphyrin (MPP1), which catalyzes the oxidation of thioanisole<sup>162</sup> Ricoux's group has also contributed to the design ArM based on *de novo* scaffold with porphyrins. The ArM was obtained by covalently attaching MnTPP into  $\alpha$ -Rep A3 bidomain protein. The biohydrid did not display any catalytic activity by itself, but peroxidase and monooxygenase activities were detected in the presence of imidazole or certain protein domains (bA3-2 or His6-bA3-2). Best catalyst was obtained in presence of His6-bA3-2, as bA3-2 opens the bidomain and His-tag coordinates to metallic center favoring the reaction. Computational approaches were used to shed light on this ArM mechanism<sup>163</sup>

Heme-based ArM have also been obtained by embedding heme moieties into non-heme containing scaffolds. On the one hand, heme can be introduced directly into scaffold without any modifications, as in the case of LmrR, which has been shown to catalyze cyclopropanation reactions when loaded with heme.<sup>164</sup> On the other hand, heme can be modified to have higher affinity for a certain scaffold. This is the case of Ricoux study, in which heme was synthesized with testosterone to bind to neocarzinostatin protein to obtain an ArM that catalyzed oxidation reactions.<sup>165</sup>

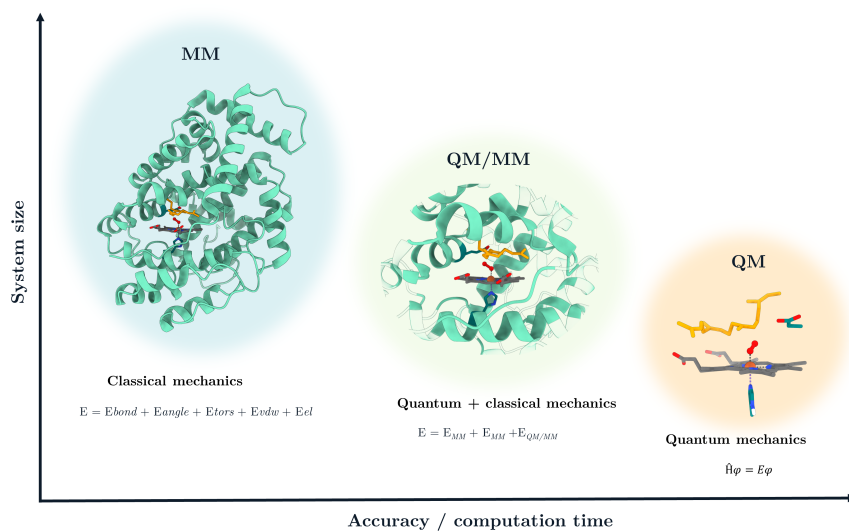
It can clearly be observed how the field of heme-based ArM has been exploited over the past year due to the catalytic versatility of heme. Still, there are few examples of heme-based ArM obtained using non heme scaffolds and the necessity of developing this type of ArM that can incorporate heme in novel scaffolds has not been resolved. One of the objectives of this thesis is to develop new computational tools that can be used to predict heme binding site and in order to design heme-based ArM.

## 1.4. Molecular modeling for metalloproteins

Molecular modeling is essential for understanding both structural and functional role of metalloproteins. In this section first we will focus on the different molecular modeling techniques that are available to study metalloproteins in general. Then, we will focus on two pivotal contributions that are important for ArM design: multiscale approaches and predicting metallic binding sites.

### 1.4.1. Overview of molecular modeling in metallic biomolecules

In a general way, molecular modeling comprises various computational methods that aim to mimic or simulate molecular systems. These methods rely on powerful computers to solve complex equations to study atoms and molecules' behavior. Molecular modeling aims to solve different kinds of problems that range from conformational exploration of molecules, energetic evaluation, reactivity or the interaction between proteins and ligands. Different levels of theory can be employed depending on the system's size and complexity, the desired accuracy and the research goal in question. These levels of theory are mostly either based on quantum mechanics (QM) or molecular mechanics (MM), each offering a different balance between accuracy and system size (**Figure 1.17**). Accurate methods as QM require more computational power and time, while low accuracy methods as MM are faster and require less resources.



**Figure 1.17:** Computational techniques depending on system's size and accuracy requirements.

**Molecular mechanics** describes molecular systems via classical Newton physics, where electrons are not treated explicitly. Instead, molecules are represented as collection of balls (atoms) connected by springs (bonds). The force field comprises a set mathematical functions and parameters that are used to compute the system's potential energy based only on the nuclear positions.<sup>166</sup> The energy functions of the force field consist of the sum of bonded terms (bonds, angle,

torsion) and non-bonded terms (VdW and electrostatics). Due to their intrinsic simplifications, this methodology cannot provide information about the electronic properties of the systems. Consequently, phenomena like bond breaking/formation or chemical reaction cannot be studied. However, these simplifications allow working with large systems (up to  $10^6$  atoms) at high computational speed, which makes it a very attractive method for simulating large time scale events in proteins or DNA. In general, MM provides accurate results regarding relative energies between conformers, interconversion pathways between them or predicting equilibrium geometries.<sup>167</sup>

Despite the advantages of MM mentioned above, this method depends on the ability of the force field to describe the parameters of the biomolecules. When working with metalloproteins the problem is that most common force fields do not include metal parameters. The parameterization of metals is not trivial due to their unique characteristics, such as the ability to change their coordination geometry during a dynamical event like a binding process or interaction with a substrate. In MM methods, the nature of the interaction of the metal ion and its surrounding ligands can be represented by different models: **1)** The non-bonded model only considers the non-bonded interactions, electrostatic and VdW terms, using the Coloumbic and the 12-6 Lennard-Jones (LJ) potential respectively. **2)** The bonded model adds a covalent bond between the metal and the surrounding ligands, considering bond, angle, dihedral, the electrostatic and VdW terms. **3)** The cationic dummy atom model involves placing cationic dummy atoms between the metal and the ligands in a specific initial geometry.<sup>168-170</sup>

Depending on the system's characteristics and the type of study to perform, the model of choice can be different. Furthermore, it must be considered that each model has disadvantages that need to be considered. The non-bonded models are the simplest approach because they do not consider any bonding terms. Due to their simplicity, the accuracy is higher for monovalent ions like alkali metals, but not for complex multinuclear centers or for transition metals.<sup>171-173</sup> The main problem of bonded model is that as the metals are covalently bonded and, in current state-of-the-art, it is therefore impossible to consider ligand exchange or geometry changes in the simulations. This approach requires careful parameterization, and the transferability of the parameters is difficult, not like in the case of non-bonded models. Still, this is an effective approach when the ligand exchange process does not occur at computational time scale and

nowadays parameterization can be speed up with scripts like *MCPB.py*.<sup>168</sup> Lastly, the main drawback of the dummy model is the requirement of an elaborated parameterization process, but, like non-bonded models, it allows transition between different geometries.<sup>174</sup>

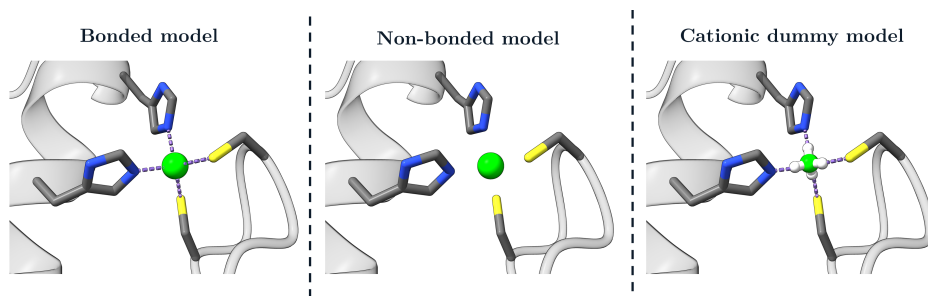


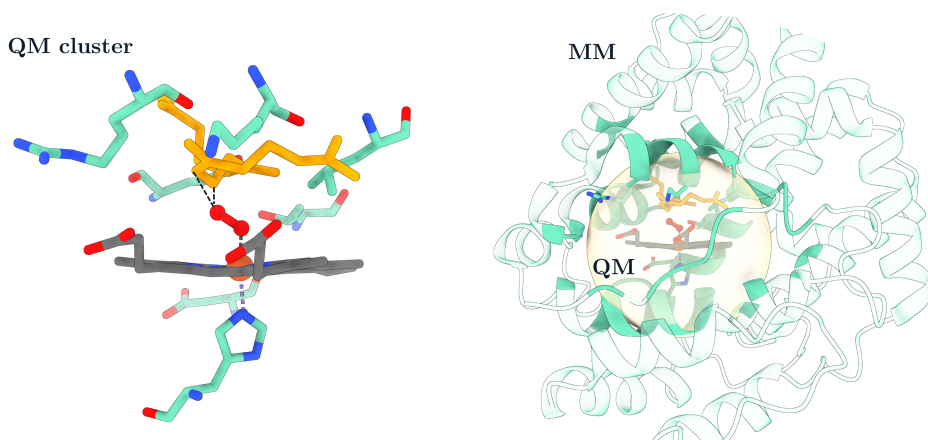
Figure 1.18: Approaches to model metals in molecular mechanics.

The other family of modeling approaches are based on **quantum mechanics (QM)**. QM methods aim to solve Schrodinger's equation to describe electron distribution in detail, considering both nuclei and electrons. However, in practice, approximations are needed as Schrodinger's equation cannot be solved for systems with more than one electron.<sup>175</sup> Compared to MM, QM methods are restricted to small systems (less than  $10^3$  atoms) and are computationally more expensive. Yet, they can provide accurate information regarding the chemical reactivity, potential energy surfaces, electronic excitation and charge transfers. Several approaches are included in this group, from very accurate and expensive post-Hartree Fock methods to fast and feasible DFT or even semi-empirical approaches.<sup>176</sup>

Because of their cost, *ab initio* methods and especially post-Hartree Fock ones, have been applied to biomolecules in a limited way. In this field, DFT and semi-empirical methods are prevalent. The former presents a better accuracy and require little or no parameterization when compared to the latter. When it comes to bioinorganics though, DFT is the most recommended approach since the presence of metal is almost inaccessible via semi-empirical methods.<sup>177,178</sup> However, with DFT is not possible to study the reaction mechanism of an enzyme and sample the conformational space of an entire protein at the same time. Between QM and MM are hybrid methods, which solve the aforementioned problems and are the method of choice for metalloenzymes.

**QM/MM approaches** combine the best qualities of computational techniques by dividing the system into two regions (**Figure 1.19a**). QM is applied to a group of selected residues from active site as well as the metal or small molecules (if any), while the rest of the enzyme that is not directly involved in the reaction is treated with MM.<sup>179,180</sup> These methods afford a good balance between accuracy and computational cost. However, its applicability depends on three crucial considerations: how to treat the interactions between the MM and the QM region or the covalent bonds that cross the QM/MM interface region and how to calculate the total energy of the system.<sup>181</sup> Additionally, the sampling of the protein's conformational space is limited too since the cost of the calculation of the energy of a unique structure still depends on the QM approach.

With metalloenzymes it is also frequent to perform QM calculations on cluster models, in which only the residues of active site and atoms involved in the reaction are considered in a QM level. This approach limits the calculation to a few hundred of atoms as the rest of the enzyme is not considered, but the surrounding protein environment is modeled by continuum solvation implicit model with specific dielectric constant. This method requires an X-ray structure, and the steric effects of the protein are usually simulated by constraining the atoms where the system has been truncated (**Figure 1.19b**). The accuracy of the results depends on the residues included and the choice of the dielectric constant. Studies to compare both QM/MM and QM cluster have been performed.<sup>182,183</sup>



**Figure 1.19:** Representation of hybrid methods **a)** QM/MM **b)** cluster QM.

### 1.4.2. Molecular modeling in ArM

Over the past few years, molecular modeling techniques have made a significant progress, leading to several examples of ArMs obtained with the help of computation. In this regard, there are two families of methodologies that seem to have erupted over the last decade: theozyme and multiscale approaches.

#### 1.4.2.1 Theozyme approaches

A theozyme is defined as a computational model of the transition state (TS) with a minimal active site of the enzyme containing specific functional groups. First, the theozyme is computed by QM and subsequently its docked it into an inert protein, which is optimized to favor the stability of the TS.<sup>184,185</sup> Most popular software is RosettaMatch, which identifies the adequate scaffold protein to match the theozyme. Finally, the active site is optimized with RosettaDesign (**Figure 1.20a**).<sup>186</sup> One group that has made remarkable contributions in this field is the Baker and coworkers. By combining computational approaches (Rosetta software) and directed evolution, they have successfully design of de novo ArM, including a Kemp eliminase and a Diels-Alderase.<sup>187,188</sup> However, design of de novo ArM represents a more difficult challenge and few examples can be found.

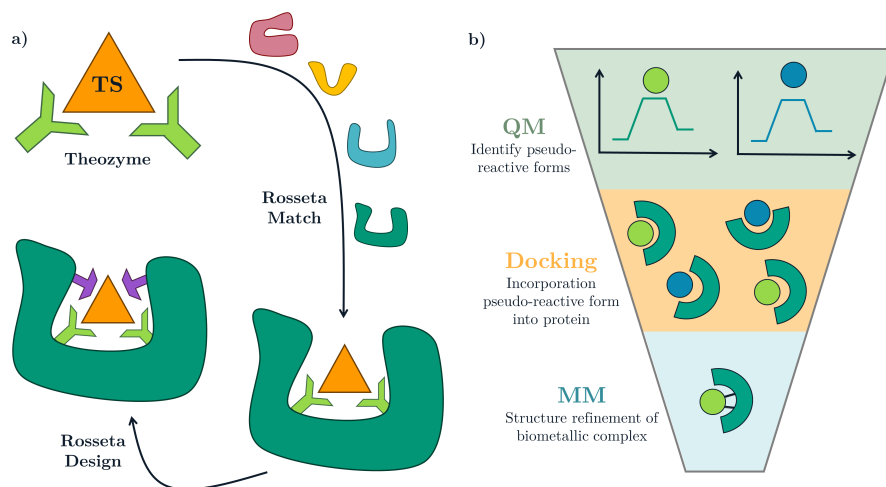
One of the few examples is the design a de novo metalloenzyme with unnatural amino acid (BpyA). In their approach, the theozyme (BpyA, Fe, substrate and two coordinating residues) was optimized by QM and then introduced to a protein scaffold using RosettaMatch, which found possible scaffolds that could match the theozyme. In the next step the second coordination sphere was optimized by inserting additional residues to improve the stability of the theozyme using RosettaDesign. After two round of this process, crystallographic evidence revealed structures bound to divalent metals with only small deviations from the designed models. Due to the high affinity for  $\text{Zn}^{2+}$ , this could be a starting point for future hydrolases and other related enzymes.<sup>189</sup> A very similar protocol using Rosetta software has also been used to redesign ArMs, such as the case of mononuclear Zn that catalyzed organophosphate hydrolysis.<sup>190</sup> Besides this applications, Rosetta design has also been employed to identify mutations to improve the activity and enantioselectivity of ArMs.<sup>191</sup> Despite the importance of the theozyme approach, limited success has been obtained for the design of ArM.

### 1.4.2.2 Multiscale approaches for ArMs

The molecular modeling of ArM is especially challenging because it needs to account for the scaffold's conformational sampling without overlooking the metallic center's electronics. By such, multiscale approaches are particularly relevant. These are based on the statement that efficient ArMs provide a better secondary coordination sphere environment to stabilize the TS structure. Because the interactions between the cofactor, the substrate and the scaffold are purely based on non-natural interactions (the triad complementarity is far from optimal since they lack years of evolution), the identification of such TS structures is very difficult. To do so, different variables need to be considered that range from protein-ligand interactions, dynamical effects by MM approaches and simulation of the catalytic mechanism by QM-based ones. This protocol has been reinforced, updated and adapted for computer aided design of ArM. It can be divided in the following steps (Figure 1.20b):

1. **Identification of the conformation of pseudo-reactive forms:** This first step stands on the initial identification of the most likely pseudo-reactive form, either an intermediate state or pseudo-TS of interest from the reaction with the cofactor, the substrate and eventually a few amino acids models isolated from the receptor (reminiscent to the theozyme ).
2. **Flexible docking into several protein conformations:** Then, through docking the pseudo-reactive are introduced into the binding site of the receptor considering its different conformations. The complexity relies in adequately describing the metallic species under the classical force fields that are the grounds of molecular mechanics.
3. **Structure refinement of pseudo-reactive forms:** Finally, the best possible structural candidates of these pseudo-reactive structures of the entire system is filtered then refined with QM/MM calculations or Molecular dynamics (MD) depending on the system of study. This involves a major *tour de force* in term of system preparation and conformational exploration since force field that could handle metal ions and coordination rules are needed.

These approaches have been applied very successfully to guide the design of ArM in different steps of the process, rationalize its selectivity or functionality. One of the objectives of this thesis is to apply multiscale approaches to rationalize and guide the design of different ArM.



**Figure 1.20:** Computational approaches for ArM design: **a)** Theozyme **b)** Multiscale.

A very successful case is the design of an enantioselective hydratase. The computational workflow developed in our group started with QM cluster of  $\text{Cu}(\text{II})$ –(2,2′-bipyridine) complex with substrate and an aspartate (acting as base) to favor the reaction. This was followed by dockings and MD simulations of the complex into LmrR proteins which revealed that there were no general bases in the vicinity to favor the reaction. Subsequently, upon inspection of the docking results, certain residues were identified for mutation to serve as bases. MD simulations we again employed to find which mutant displayed optimal distances and qualitative predictions about the enantioselectivity of the reaction. Experimental validation of the selected mutant agreed with computational predictions, showing the potential of multiscale approaches.<sup>192</sup>

In the study of an ArM POP-Rh<sub>2</sub> cyclopropanase, the inclusion of GPathFinder calculations as a last step of the multiscale workflow revealed how the entrance and diffusion of the substrates can also be the origin of the enantioselectivity of the ArM.<sup>193</sup> On other works, the combination of dockings and MD simulations have shed light on the importance of the dynamics of protein and how these can affect the activity of the ArM.<sup>194</sup> For example, MDs revealed that LmrR protein is able to tightly bound heme for cyclopropanation reactions. However, the flexibility of its helices allows to open up facilitating the entrance of substrates and reaching the pre-catalytic stage.<sup>164</sup>

### 1.4.3. Predicting the binding of metals ions and heme to proteins

Binding processes and interactions between proteins and metal moieties are essential for understanding physiological roles or for designing new ArMs. The prediction of binding of metal moieties to proteins is of high relevance, especially when there is not an available protein structure with the metal bound. Metal binding sites can be detected individually using experiment methods like X-ray crystallography, NMR or X-Ray Absorption Spectroscopy (XAS).<sup>195–198</sup> These techniques are time-consuming and have high costs associated with them. Therefore, computational tools are a good alternative for predicting metal binding sites and they offer better insight into protein-metal interactions.

Computational prediction of metal ions binding sites is based on either sequence or structural 3D information, sometimes even the combination of both. Sequence based methods include MetalDetector, which detects binding sites for transition metals that involve His and Cys using machine learning.<sup>199</sup> On the other hand, there are predictors only based on structure like TEMPS, that predicts Zn binding sites.<sup>200</sup> MetSite or MIB are metal site predictor that combines both sequence and structure information. MIB uses fragment transformation method, but recently included AlphaFold and (PS)<sup>2</sup> to improve the prediction.<sup>201–203</sup> Recently, predictor based solely on machine learning have been developed based on sequence like DeepMBS or based 3D structure like MetalSiteHunter.<sup>204,205</sup>

Regarding the prediction of metallic cofactors, the prototypical case is heme due to its biological relevance. Compared to the prediction of metal ions, less software has been developed for detecting heme binding sites. So far, most programs for heme rely principally on sequence information. SCMHBP is a heme predictor that is only based on sequence and is based on scoring card method (SCM), whereas HemeBIND prediction is based on structural attributes combined with sequence information using machine learning (SVM).<sup>206,207</sup> Even though that structure-based predictors for heme and metals ions have been developed, predictors that take only into account the geometrical predisposition of the binding sites and backbone atoms have not been produced yet. The advantage of this approach is that it could be applied to develop either new ArM based on heme or metal ions, because no sequence pattern would be needed, just the preorganization of the binding site and mutation of residues. This is the main objective of software development in this thesis.

## CHAPTER 2

# Methodology

This chapter aims to provide an overview of the theoretical background of the techniques used in this PhD thesis. Section 1.4.1 has already covered the general applicability, context, and purpose of the different techniques. This chapter aims to describe in more detail the key principles of each method focusing on its advantages or inconvenience, without doing an extensive explanation.

## 2.1. Quantum Mechanics (QM)

Quantum mechanics (QM) are based on the postulate that the wave function contains all the physical properties that characterize a system. The wave function is a probabilistic descriptor, with its square determining the probability of particle being at a certain region. QM employs mathematical operators that, when applied to the wave function, return observable properties. In most chemical applications, the aim of QM is to solve the time-independent Schrodinger's equation 1, where  $H$  is the Hamiltonian operator,  $\Psi$  is the wave function and  $E$  the energy.<sup>208</sup>

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

This equation is an eigenvalue equation, in which the Hamiltonian operator extracts all the energy (eigenvalue) of the system of interest from the wave function (eigenfunction). For a system of  $N$  electrons and  $M$  nuclei, the Hamiltonian is an energy operator that is composed of two parts, the kinetic ( $T$ ) and the potential energy ( $V$ ).

$$\hat{H} = \hat{T} + \hat{V} = \hat{T}_n + \hat{T}_e + \hat{V}_{en} + \hat{V}_{ee} + \hat{V}_{nn} \quad (2)$$

The kinetic energy comprises two terms, the kinetic energy of the nuclei ( $\hat{T}_N$ ) and the electrons ( $\hat{T}_e$ ), while the potential energy includes three terms: the coulomb attractive interactions between electrons and nuclei ( $\hat{V}_{en}$ ) and the coulomb repulsive interaction between electrons ( $\hat{V}_{ee}$ ) and between nuclei ( $\hat{V}_{nn}$ ).

Schrödinger's equation has a limitation, it can only be solved for systems with less than two electrons, as in the case of the simplest molecule  $H^2+$ . For other systems, the exact solution of the Schrödinger's can not be obtained, and the solution will only approximate the real solution. The problem arises from fact that equation 2 contains pairwise attraction and repulsions terms, thereby introducing a correlation between the motion of particles.<sup>208</sup> To address this problem, the Born-Oppenheimer approximation is applied; the motion of the electrons is separated from the motion of the nuclei. Since the mass of the nuclei is approximately 1800 higher than the electrons, it can be assumed that electrons adapt almost immediately to the movement of the nuclei. Therefore, this approximation assumes that electrons move in a field generated by a fixed nuclei because the motions of electrons and nuclei are decoupled.<sup>209</sup> This simplification leads to two principles:

1. The total wave function can be expressed as a product of electronic and nuclear wave functions, where  $r$  are electron coordinates and  $R$  are nuclei coordinates.<sup>176</sup>

$$\Psi_{tot}(r, R) = \Psi_e(r; R) \Psi_n(R) \quad (3)$$

2. Since the electronic wave function depends parametrically on the nuclear coordinates, the kinetic energy of the nuclei is neglected and the coulomb interaction between nuclei is a constant. The semi-colon indicates a parametric dependence, by varying the values of  $R$  different electronic equations are obtained. Therefore, the Schrödinger equation that describes the motions of electrons at fixed position of the nuclei can be expressed as in equation 4.<sup>176</sup>

$$\hat{H}_e(r; R) \Psi_e(r; R) = E_e(R) \Psi_e(r; R) \quad (4)$$

$$\hat{H}_e = +\hat{T}_e + \hat{V}_{en} + \hat{V}_{ee} \quad (5)$$

When the electronic Schrödinger equation is solved, the pure electronic energy ( $E_e$ ) of the system is obtained for a certain nuclear disposition of the nuclei  $R$ . This is related to the potential energy function ( $U_n$ ), which can be obtained by adding the constant  $V_{nn}$  term to the electronic energy.<sup>176</sup>

$$U_n = E_e + V_{nn} \quad (6)$$

Furthermore, when the electronic Schrödinger equation is solved for different nuclear coordinates, a set of values of  $U_n$  can be obtained. These values lead to the Potential Energy Surface (PES), which indicate the variation of the potential energy as the positions of the nuclei change. This is a crucial concept in computational chemistry for finding stationary points, especially minima and transition states. From the PES, effective potential energy functions for the nuclear motion can be derived and applied to the nuclear Schrödinger equation.<sup>7,167,176</sup> By solving this equation, it will lead to the vibrational, rotational and translational states of nuclei.<sup>167</sup>

$$[\hat{T}_n + \hat{U}_n]\Psi_n(R) = E_{tot}\Psi_n(R) \quad (7)$$

No analytical solution exists for the electronic equation 3, approximate mathematical techniques such as variational principle or perturbation are needed. The former states that the calculated approximated energy will equal or higher than the system's exact energy, while the latter is based on adding a small perturbation to a known solution.<sup>166</sup>

Moreover, more approximations are required to solve the electronic equation when dealing with many-body systems. This comes from the fact that the  $V_{ee}$  term implies estimating electron-electron interactions which is a multidimensional problem without a solution. These methods are divided into wave function-based and density functional theory-based (DFT) methods. The calculations performed in this thesis are based on DFT methods, consequently, this will be explained in more detail.

### 2.1.1. Hartree Fock

The main challenge of studying multielectronic systems using QM approaches is obtaining an approximated value for the  $V_{ee}$  term. The first methods are based on electronic wave function and are encompassed to the Hartree Fock (HF) formalism. The many-electron wave function of a system can be approximated as a product of individual one-electron wave functions, called orbitals.

$$\Psi = \psi(1)\psi(2)\psi(3)...\psi(n) \quad (8)$$

These orbitals are expressed as spin orbitals ( $\phi$ ), the product of spatial orbitals and spin function ( $\alpha$  or  $\beta$ ) to satisfy the Pauli principle. To fulfill the antisymmetric principle and the indistinguishability of electrons, the wave function is expressed as a unique determinant composed of one spin-orbital per electron, called Slater determinant. To describe the space for each electron of the systems, those spin-orbital can be written using either Slater functions (STO) or a combination of gaussian ones (GTO). The HF method determines variationally the lowest energy determinant in an iterative way (SCF).

$$\Psi(1,2...N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \dots & \phi_1(N) \\ \phi_2(1) & \phi_2(2) & \dots & \phi_2(N) \\ \dots & \dots & \dots & \dots \\ \phi_N(1) & \phi_N(2) & \dots & \phi_N(N) \end{vmatrix}$$

One of the main disadvantages of HF is that it does not consider electron correlation. This implies that it only accounts for the average electron repulsion, as it considers that each electron moves in an average electronic field generated by all other electrons. Therefore, the electron-electron repulsion is not treated explicitly, reducing the accuracy of geometry and energy calculation. Despite HF provides with a first approximation to deal with electron-electron interactions, the electron-cloud model is a rough solution that appears often as a limiter. On the ground of HF formalism, numerous approaches have been built to increase the quality. Some use perturbative approaches as Moller-Plesset (MP), variational ones or even coupled cluster CCSD(T). Still, post-HF calculations are expensive computationally and can only be applied to small systems.<sup>166</sup>

### 2.1.2. DFT

DFT is based on the premise that the energy can be calculated from the electron density rather than the wave function. This fundamental idea originated from two theorems enunciated by of Hohenberg and Kohn. Both theorems rely on the concept of a functional, which is a function that has a function as an argument.<sup>208</sup>

The **first theorem** defines that the ground state energy from Schrödinger equation of a system is a unique functional of the electron density  $\rho(r)$  that only depends on three spatial coordinates. Given the ground state electron density, the Hamiltonian operator can be determined and therefore the wave functions and all the properties of the system can be computed.<sup>166,210</sup>

$$E = E[\rho(r)] \quad (9)$$

The **second theorem** states that using a variational approach the functional will give the lowest energy of the ground state energy if the provided density is in fact the ground state density. If not, it will give an energy higher than the true value. The electron density that gives the minimum energy of the functional is the ground state electron density derived from the solution of Schrödinger's equation. If the true functional were known, the electron density could be adjusted until the energy reached its minimum value.<sup>166,210</sup> From the electronic Schrödinger equation, the energy functional can be written as the sum of three terms:

$$E[\rho] = T[\rho] + V_{ne}[\rho] + V_{ee}[\rho] \quad (10)$$

The second term  $V_{ne}[\rho]$  corresponds to the interaction between electrons and the external potential created by the nuclei, which functional can be defined. Contrarily, the energy functionals of the first and third term, the kinetic energy  $T[\rho]$  and electron-electron interaction energy  $V_{ee}[\rho]$  are not known. The success of DFT is based on the **Kohn-Sham theory (KS)**, which solves this problem. The premise is that a set of fictitious  $N$  non-interacting electrons can have the same ground state electron density that a set of  $N$  interacting electrons. The energy of the ground state is expressed as an energy functional of the density:<sup>209</sup>

$$E[\rho(r)] = T_S[\rho(r)] + V_{en}[\rho(r)] + V_H[\rho(r)] + E_{xc}[\rho(r)] \quad (11)$$

The first term of equation 12 represents the kinetic energy of the non-interacting system with the same density as the real system, which corresponds to the sum of the individual kinetic energies.

$$T_S[\rho(r)] = \sum_{i=1}^N \int \phi_i(r) \left( \frac{-\nabla^2}{2} \right) \phi_i(r) dr \quad (12)$$

The second term corresponds to the nuclear-electron interactions, while the third corresponds to the classical energetic repulsion between the electrons.

$$V_{en} = \sum_{A=1}^M \int \frac{Z_A}{|r - R_A|} \rho(r) dr \quad V_H = \frac{1}{2} \iint \frac{\rho(r_1)\rho(r_2)}{|r_1 - r_2|} dr_1 dr_2 \quad (13)$$

The last term is the exchange correlational energy functional, which accounts for all the energy contributions that were not accounted in the other terms. The  $E_{XC}$  includes all the corrections regarding the difference in kinetic energy between the non-interacting systems with the real interacting system, effect of quantum electron exchange and correlation and energy for electron self-interaction.<sup>209</sup>

The KS theory represents the electron density as the sum of the square of a set of one-electron orbitals (KS orbitals), which are described by a single Slater determinant. As in HF, Kohn-Sham orbitals are represented as linear combinations of basis sets.<sup>209,210</sup>

$$\rho(r) = \sum_{i=1}^N |\phi_i(r)|^2 \quad (14)$$

By applying the variational principle and density expression from equation 11 it leads to the Kohn-Sham equations.

$$\left\{ \frac{-\nabla^2}{2} - \left( \sum_{A=1}^M \int \frac{Z_A}{|r_1 A|} \right) + \int \frac{\rho(r_2)}{r_{12}} dr_2 + V_{XC}[r_1] \right\} \phi_i(r_1) = \epsilon_i \phi_i(r_1) \quad (15)$$

These equations are solved in a self-consistent approach by first making an initial guess of the density. Then, using this trial density the Kohn-Sham equations are solved, and a set of wave function orbitals are derived. The electron density is

calculated from these orbitals, which is used in a new iteration and so on. This is performed until convergence is reached.<sup>209,210</sup>

One of the most important points of DFT calculations is that they include correlation energy. The idea is that electrons move in a way that maximizes the attraction to the nuclei and minimizes the repulsion with other electrons. Therefore, there are regions around the electrons that are considered XC-holes, regions that other electrons are not allowed to penetrate. Exchange functionals can be divided into exchange and correlation, corresponding the exchange part to the Pauli principle and the correlation to the repulsion of electrons.<sup>176</sup>

The exchange correlation functional is defined as the functional derivative of the exchange-correlation energy as in equation 13.

$$V_{XC}(r) = \frac{\partial E_{XC}(r)}{\partial n(r)} \quad (16)$$

Each type of exchange functional defines this differently. The problem that we have is that to solve the Kohn-Sham equations the exchange-correlation functional has to be specified. If this functional was known, the Kohn-Sham equations would lead to the exact energy. Therefore, different classes of exchange-correlations functionals have been approximated over the years, some of them including empirical parameters and other experimental. Perdew stated that the different types of functionals can be classified in a ladder way (Jacob's ladder), if you move up the ladder it increases the chemical accuracy, however, it also increases the computational time **Figure 2.1**.<sup>211</sup>

### 2.1.2.1 Exchange correlation functionals

The **Local Density Approximation (LDA)** is based on the idea of a uniform electron gas, meaning that the functional only considers local values of density to describe exchange-correlation energy. LDA has to be replaced for LSDA (Local Spin Density Approximation) in systems in which there is spin polarization. Despite the simple idea behind this functional, it gives good results for calculating geometries or when working with metals, but it does not predict well bond energies or thermodynamics.

In the second step of the ladder, instead of only depending on the local values of density, it also depends on the derivatives of the density. These correspond to

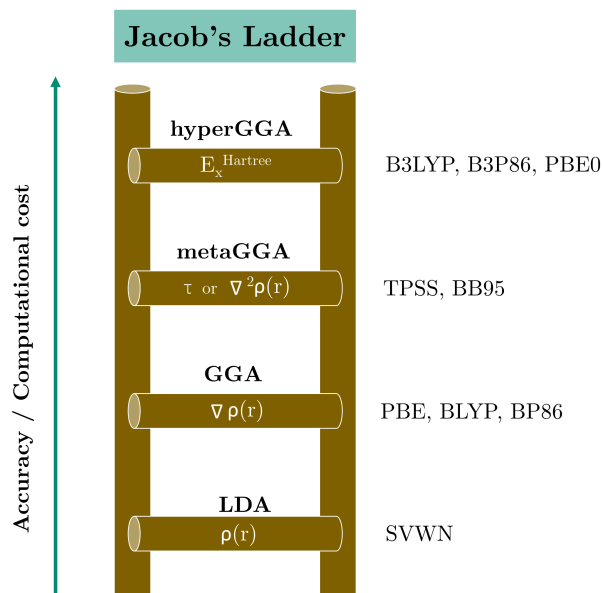


Figure 2.1: Types of functionals classified according to Jacob's ladder.

**Generalized Gradient Approximation (GGA)** functionals, which basically add the gradient of the local density, typically separated into exchange and correlation contributions. One of the earliest popular GGA exchange functionals is Becke's (B), which corrects LDA functional with a mathematical form and empirical parameters. On the other hand, Lee, Yang and Parr (LYP) correlation functional is also very popular, which is usually combined with Becke's and defined as BLYP.

The extension of GGA is **metaGGA**, which depends on the second derivatives of the electron density, the Laplacian. Additionally, the functional can also depend on the orbital kinetic energy density. Both carry the same information, but the latter is more numerically stable. Becke and Roussel were one of the first to propose metaGGA exchange functional (BR), which also depended on kinetic-energy density.

The last step of the ladder corresponds to the method used in this thesis, **hybrid methods**, which consists in combining the exchange energy from HF and the correlation energy from DFT. This approach is based on the Adiabatic Connection Formula (ACF), in which the exchange-correlation energy is coupled to a  $\lambda$  parameter that relates the fictitious non-interacting system with the real one.

$$E_{XC}(r) = \int_0^1 \langle \Psi(\lambda) | V_{XC} | \Psi(\lambda) \rangle d\lambda \quad (17)$$

On one end, we have the non-interacting fictitious system, where there is no correlation part because it's non-interacting. Therefore, since the wave function is a Slater determinant of the Kohn-Sham functions, the exchange energy corresponds to the exact exchange energy of HF. On the other end, we have the real interacting system and the exchange energy can be approximated by the LSDA functional, resulting in a half-and-half method. This idea was improved by expressing it as a linear combination of the exact exchange energy, exchange energy approximated by LSDA and exchange and gradient correction terms.

$$E_{XC}^{B3LYP} = (1 - a)E_x^{LSDA} + aE_x^{exact} + b\Delta E_x^{B88} + (1 - c)E_x^{LSDA} + cE_x^{LYP} \quad (18)$$

Expression 17 is the form that takes B3LYP, which is the most widely used hybrid functional. B3LYP combines 20% Hartree-Fock exchange with 80% DFT exchange-correlation functional. In the expression,  $a$ ,  $b$  and  $c$  are determined by fitting to experimental data. Other popular hybrid functionals include PBE0, HSE06 and M06.

### 2.1.2.2 Overview of DFT

One advantage of DFT is that the electron density is a function that only depends on three spatial variables, which is independent on the number of electrons and scales at  $N^3$ . On the other hand, in HF, the many-electron wave-function depends on  $4N$  variables, three spatial variables and one spin coordinate for each electron and scales at  $N^4$ . Consequently, DFT calculations are substantially faster than HF. More importantly, DFT solves one of the disadvantages of HF, at the same computational cost, in DFT electron correlation is included. The main difference with HF is that it incorporates correlation energy.

One of the main disadvantages of DFT is that it does not accurately model VdW interactions (London dispersion), which are long-range weak interactions between non-bonded atoms. These interactions are important in large systems, and it is not well treated especially in the semi-local functionals and hybrid functionals as they do not represent correctly the  $C_6/R_6$  dependence. Currently, there are several approaches that include dispersion for DFT, one group of them being the semiclassical corrections (DFT-D). The idea is that the dispersion correction energy is added to the DFT energy:  $E_{DFT+D} = E_{DFT} + E_{disp}$

The general form of DFT-D takes all atoms pairs in a system and applies London's formula in the following expression:

$$E_{disp}^{DFT-D2} = \frac{-1}{2} s_6 \sum_{A \neq B} \frac{C_6^{AB}}{R_{AB}^6} f_{damp}^{DFT-D2}(R_{AB}) \quad (19)$$

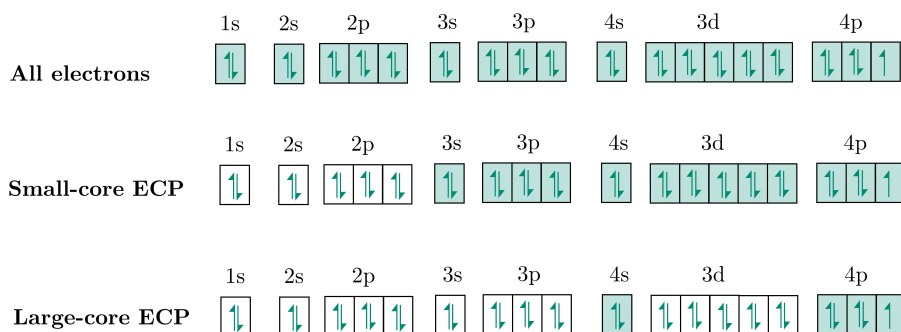
Where  $C_6^{AB}$  is dispersion coefficient for atom pair,  $s_6$  is a global scaling parameter and  $f_{damp}$  is a damping function. One of the most popular corrections are Grimme's D3, which is a refined version of DFT-D2 that is more accurate as it includes both two-body and three-body term in a more complex form.<sup>212</sup>

$$E_{disp}^{DFT-D2} = \frac{-1}{2} \sum_{i=1}^{N_{at}} \sum_{j=1}^{N_{at}} \sum_L \left( f_{d,6}(r_{ij,L}) \frac{C_{6ij}}{r_{ij,L}^6} + f_{d,8}(r_{ij,L}) \frac{C_{8ij}}{r_{ij,L}^8} \right) \quad (20)$$

### 2.1.3. Metals in QM

Molecular modeling of systems containing transition metals represent a challenge due to the difficulty of handling some of transition metals (TMs) characteristics in computation, like the presence of d or f orbitals, its multiple oxidation states or its complicated chemical bonding and multiple coordination numbers. When modeling TM, DFT are the methods of choice, as they consider electron density and dynamic correlation effectively, apart from its accuracy and high scalability for large systems compared to HF.<sup>213</sup> The choice of the functional depends a lot on the property and the system; when working with enzyme reactions containing TMs the functional of choice is usually hybrid functional B3LYP, as in general with TMs HF exchange can improve accuracy, but too much is not recommended.<sup>214,215</sup>

Heavy elements like TMs have many core electrons, which implies the use of a large number of basis sets and high computational cost. As core electrons usually do not have a significant impact in chemical behavior, Effective Core Potentials (ECP) are employed in order to reduce the computational costs of calculations. ECP are a set of potential functions that replace core electrons and their associated density. By replacing core electrons, only the valence electrons are considered explicitly in QM calculations. Consequently, the basis set size is decreased, and computation costs are reduced. Large-core ECP only consider explicitly valence electrons and Small-core ECP also consider the lower shell above the valence electrons (**Figure 2.2**). Apart from that, in heavy elements, relativistic effects are particularly significant. Therefore, Effect Core Potentials also solve this problem by including relativistic terms. The benefits of working with ECP increase as the we go towards the lower site of the periodic table.<sup>167</sup> Some frequently used ECP are LANL2 ECPs or Stuttgart-Dresden-Bonn ECPs.<sup>216,217</sup>

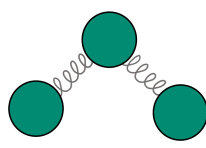


**Figure 2.2:** Small-core and large-core ECP exemplified for Br atom.

## 2.2. Molecular Mechanics (MM)

Quantum methods are suitable when dealing with electrons of small systems up to  $10^6$  atoms, but most biological macromolecular systems are unfortunately too large to be treated with full QM. In this context, interest is focused on Molecular Mechanics (MM) or force field methods, which ignore the electronic motions and the energy is only calculated as a function of the nuclear positions using empirically parameterized force fields. MM considers the system as a group of masses connected by springs and the different interactions between particles are treated through classical mechanics laws, specifically Newton's second law. MM methods are based on a series of assumptions.

As the Born-Oppenheimer is implicitly assumed, the energy of the system only depends on the position of the nuclei. In consequence, the potential energy of the force field is defined as a sum of bonded and non-bonded terms. All force fields contains the five invariable terms from equation 21, but they can contain more additional terms.


$$\mathbf{E}_{\text{tot}} = \underbrace{E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}}}_{E_{\text{covalent}}} + \underbrace{E_{\text{vdw}} + E_{\text{el}}}_{E_{\text{noncovalent}}} \quad (21)$$

In MM, atoms are considered particles defined by an atom type, which depends on the atomic number and contains information of the hybridization (type of chemical bond involved) and local environment of the atom. Each force field parameter used to calculate the potential energy equation is expressed in terms of the atom type. Consequently, every atom type has a set of specific force field parameters associated, which are defined by the force field of use. Transferability is a crucial aspect; the basis of MM is to derive the parameters from small molecules and then transfer them to large systems. The same parameters can be used to model different molecules, which expands the applicability of MM across different scales.<sup>167,209</sup>

The first term of equation 21,  $E_{str}$ , represents the energy function for elongating a bond. It models the interaction between two atoms that are bonded. The second term is  $E_{bend}$ , which represents the energy required to bend an angle between 3 atoms A-B-C (considering that A and C are bonded to B) as represented in **Figure 2.3**.  $E_{str}$  and  $E_{bend}$  are usually modeled using a harmonic potential or Hooke's Law. This function gives the variation of energy as the bond or angle displaces from its reference value. Therefore, each contribution is characterized by a force constant ( $K_b$  or  $K_\theta$ ) and a reference value ( $l_0/\theta_0$ ). The accuracy of the force field can be improved by introducing higher-order terms (Taylor expansion).<sup>166,209,218</sup>

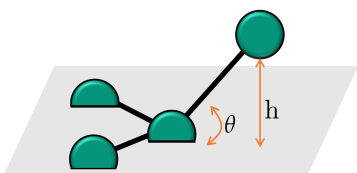
$$E_{str} = K_b(l - l_0)^2 \quad (22)$$

$$E_{bend} = K_\theta(\theta - \theta_0)^2 \quad (23)$$

$E_{tors}$ , is the energy associated to the rotation of a bond B-C found between four consecutive bonded atoms A-B-C-D. To correctly represent its periodicity,  $E_{tors}$  is expressed as a cosine series expansion, where  $\omega$  is the torsion angle **Figure 2.3**. For most cases, more terms are needed to correctly represent the rotation of a bond, when there are minima with different energy. In the expression,  $V_n$  constant determines the barrier height around the rotation,  $n$  represents the multiplicity ( $n=1$  periodicity every 360,  $n=2$  every 180...) and  $\lambda$  represents the phase factor (determines at which point is a minimum).<sup>166,209</sup>

$$E_{tors} = \sum_{N=0}^N V_n [1 + \cos(n\omega - \gamma)] \quad (24)$$

In the case of  $sp^2$ -hybridized atoms, an additional term called out-of-plane bending or improper torsion is added in most force fields to keep the  $sp^2$  atom and three bonded atoms in the same plane. This term can be incorporated as different functions. One approach is to consider an improper torsion between the four atoms and use a torsion potential to maintain the angle  $0^\circ$  or  $180^\circ$ . The other approach is to describe as a quadratic function depending on the angle or distance between the bond and the plane with other atoms are represented in equation 25.<sup>209</sup>

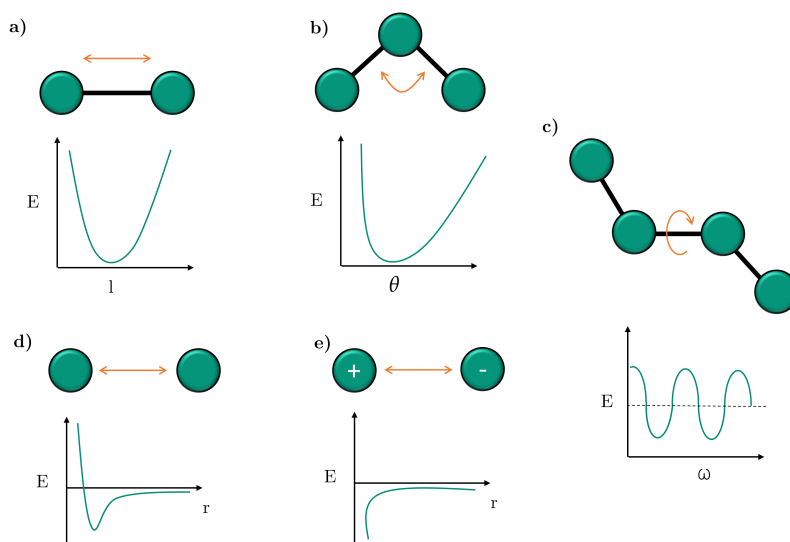
$$E_{imp} = \frac{K}{2}(\theta)^2 \quad E_{imp} = \frac{K}{2}(h)^2 \quad (25)$$


Non-bonded interactions include  $E_{vdw}$  and  $E_{el}$ , which play an important role in the final structure of the molecules. These interactions are between atoms that are not directly bonded and are usually represented as a function of the inverse of the distance.  $E_{vdw}$  is a function that represents the non-polar interaction (attraction or repulsion) between atoms that are not bonded. The  $E_{vdw}$  is zero at large distances, but as distance is reduced the energy decreases until a minimum, point at which it increases drastically due to repulsion. A function that describes this behavior is a Leonard-Jones 12-6 function, characterized by an attractive part that changes as  $r^{-6}$  and repulsive part that changes as  $r^{-12}$  **Figure 2.3**. This function depends on two parameters, the well-depth ( $\epsilon$ ) and the collision diameter ( $\sigma$ ), which is the distance it becomes 0.<sup>209</sup>

On the other hand,  $E_{el}$ , represents the electrostatic interactions between two molecules and these are best described by the sum of the interactions through Coloumb's Law **Figure 2.3**. This applied under the assumption that each atom has a partial charge assigned, represented by  $q_i$  and  $q_j$ , placed in a medium at certain dielectric constant  $\epsilon$ .<sup>209</sup> There are different approaches to calculate point charges, they can be assigned by fitting to empirical properties or by fitting the calculated electrostatic potential to the structure. A common practice of the latter is Restrained Electrostatic Potential (RESP). In this approach, the charges are obtained from fitting electrostatic potentials derived from QM. This method uses hyperbolic restrains on non-hydrogen atoms to reduces the charges on problematic non-surface atoms. This is used commonly used as it balances computational cost and accuracy.<sup>219</sup>

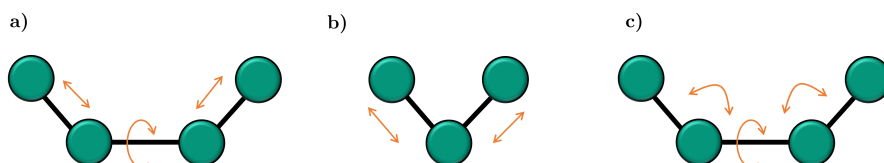
$$E_{vdw} = 4\epsilon \left[ \left( \frac{\theta}{r} \right)^{12} - \left( \frac{\theta}{r} \right)^6 \right] \quad (26)$$

$$E_{el} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (27)$$



**Figure 2.3:** Schematic and graphic representation of main five terms of force fields: **a)** stretch, **b)** bend, **c)** torsion, **d)** VdW and **e)** electronic.

Some force fields contain an additional the last term, cross-coupling, is only present in some force fields and it represents the coupling between two or more bonded terms.  $E_{stretch/bend}$  is the most important one and represent the coupling between bond stretching and the bending of an angle. For example, if an angle is decreased, the lengths of the bonds involved are increased. The  $E_{cross}$  term can include different corrections like  $E_{stretch/stretch}$ ,  $E_{stretch/tors}$ ,  $E_{bend/bend}$ , but distinct force fields usually include different types of cross-terms. In general, few cross-terms are considered necessary to reproduce structural properties accurately, except for when reproducing vibrational properties.<sup>167</sup> Some of the most relevant cross-terms are represented in **Figure 2.4**



**Figure 2.4:** Schematic representation of some cross-terms of a force field: **a)** stretch-torsion, **b)** stretch-stretch and **c)** bend-torsion.

### 2.2.1. Force fields and metals for biomolecules

For biomolecular systems like proteins, DNA, lipids or carbohydrates, several forcefields are available, being the most widely used AMBER, GROMOS and CHARMM. Both AMBER and CHARMM contain a general force field, GAFF and GenFF, respectively, that can be used to parameterize organic small ligands.<sup>166</sup> Still, when working with metals, AMBER or CHARMM only include parameters for common metallic cofactors as heme, but they lack to consider all possible coordination residues, geometries or different spin states related to the metal. Generally, most force fields do not include parameters for the interactions with metals, except for specialized force fields tailored to organometallic systems such as LFMM<sup>220</sup>, SIBFA<sup>221</sup> or VALBOND<sup>222,223</sup>. These force fields have been implemented successfully in some cases, metallic complexes containing Ga or Mn even Copper-proteins.<sup>224</sup> However, their application and availability in MD simulation packages is limited.<sup>215</sup>

As explained in section 1.4.1, there are three models to represent metals: bonded, non-bonded and dummy. The parameterization of metals is not trivial neither is the model choice. Within the scope of this thesis, the bonded model is the most adequate choice because the focus of interest is on the metallic interaction rather than the alteration of the coordination mode. Bonded metals parameters can be derived, either empirically, from NMR or X-ray values or computationally by quantum calculations.

Lin *et al* demonstrated that the force constants derived by Seminario method combined with RESP fitting for partial charges showed very good performance on Zn-proteins.<sup>225</sup> This initial work led to *MCPB.py*, which is program integrated in AMBER that facilitates obtaining metal force field parameters based on QM calculations. **MCPB.py** uses two different model schemes of the metal binding site to improve accuracy and speed. A smaller model is used to obtain the bonded parameters and a larger model to derive the partial charges. The small model is optimized through QM calculations and then the Seminario's Method uses the Cartesian Hessian matrix to calculate the force constants. In the large model only the hydrogens are optimized for speed and the RESP calculation is performed. All these parameters are extracted by *MCPB.py* and saved in *frmod* and *mol2* files to run in AMBER.

### 2.2.2. Molecular Dynamics (MD)

One of the major simulation techniques to obtain sampling of a system at specific temperature is Molecular Dynamics (MD). As MM ignores electrons motions and nuclei are large enough, it is considered that nuclei behave as classical particles. Therefore, MD simulate the behavior of a molecular system along time by solving the classical second law of Newton  $F = m \cdot a$ .<sup>209</sup>

The results of MD are trajectories that display how the position and velocities of particles of the system change over time at a finite trajectory. These trajectories are obtained by solving the differential form of Newton's equation 28.<sup>209</sup>

The premise is that the force is assumed to be constant at each time-step, then at each time step the total force is calculated as the potential by summing all the interactions. This is always the most demanding step in and MD simulation. From the force, the acceleration of the particles can be determined as in 29.<sup>209</sup>

$$\frac{-dV}{dr} = m \cdot \frac{d^2r}{dt^2} \quad (28)$$

$$a = \frac{F}{m} = -\frac{1}{m} \frac{\partial V}{\partial r} \quad (29)$$

Given the initial position of the system, which can be obtained from experimental data or from models, the initial velocities are assigned randomly from Maxwell-Boltzmann distribution. The force on each atom is calculated and the acceleration is obtained. Given the acceleration of the particles combined with the position and velocity of particles at time  $t$ , new velocities and positions can be determined at time  $t + \Delta t$ . The positions and velocity are updated, and new forces need to be calculated in these new positions. Finally, acceleration, positions and velocities are determined at new time, and this is repeated until simulation has finished.<sup>209</sup> A scheme of this process is represented in **Figure 2.5**.

All approaches consider that the position, velocity and acceleration of particles at certain time can be estimated by a Taylor Series expansion.

$$\begin{aligned} r(t + \Delta t) &= r(t) + v(t)(\Delta t) + \frac{1}{2}a(t)(\Delta^2 t) + \dots \\ v(t + \Delta t) &= v(t) + a(t)(\Delta t) + \frac{1}{2}b(t)(\Delta^2 t) + \dots \end{aligned} \quad (30)$$

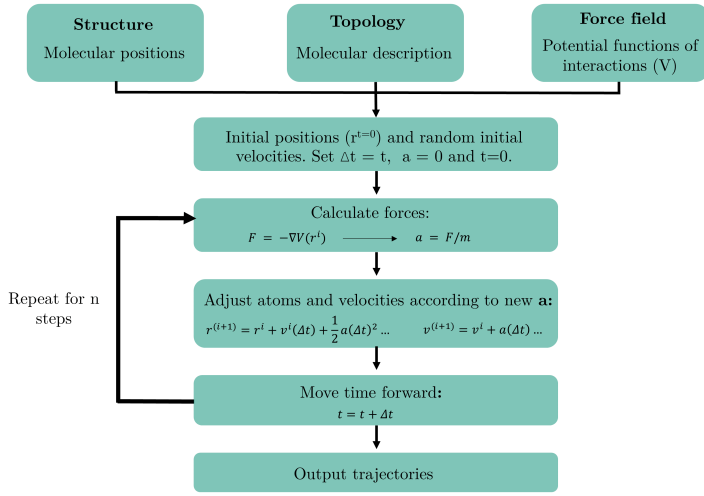


Figure 2.5: General scheme of MD simulations.

One of the most common finite methods is **Verlet algorithm**, which is able to predict the new positions at time  $t+\Delta t$  from the current position and acceleration at time  $t$  and the previous position at  $t-\Delta t$ . Given the positions at  $t+\Delta t$  and  $t-\Delta t$  as:

$$r(t + \Delta t) = r(t) + \frac{\partial r}{\partial t}(\Delta t) + \frac{1}{2} \frac{\partial^2 r}{\partial t^2}(\Delta t)^2 + \frac{1}{6} \frac{\partial^3 r}{\partial t^3}(\Delta t)^3 \dots \quad (31)$$

$$r(t - \Delta t) = r(t) - \frac{\partial r}{\partial t}(\Delta t) - \frac{1}{2} \frac{\partial^2 r}{\partial t^2}(\Delta t)^2 - \frac{1}{6} \frac{\partial^3 r}{\partial t^3}(\Delta t)^3 \dots \quad (32)$$

Verlet algorithm solves the equations numerically by adding these two terms:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)(\Delta t)^2 \quad (33)$$

In this equation the velocities do not appear explicitly, however, they can be calculated easily by either of these two ways:

$$v(t) = [r(t + \Delta t) - r(t - \Delta t)] / 2\Delta t \quad (34) \quad v(t + \frac{1}{2}\Delta t) = [r(t + \Delta t) - r(t)] / \Delta t \quad (35)$$

Verlet is not a self-starting algorithm as in the initial point the positions of the previous step are not available. Still, these can be solved by approximating the Taylor series equation until the first term as  $r_{(-\Delta t)} = r_0 - v_0\Delta t$ . Although this algorithm is straightforward, one of its main drawbacks is this lack of explicit velocities, as the velocities are not available until the positions are computed, which can be a problem when working with ensembles of constant temperature. Lastly, this algorithm has low precision as a very small term  $a(t)(\Delta t^2)$  is added to the difference of positions, which is a quite larger term  $2r(t) - r(t - \Delta t)$ .<sup>209</sup>

To improve the problems of the Verlet algorithm, variations have been developed. The **leap-frog algorithm** is based on the premise that the velocity and the positions are updated out of phase by half time-step.

$$r_{(t+\Delta t)} = r + v_{(t+\Delta t)}\Delta t \quad (36) \quad v_{(t+\frac{1}{2}\Delta t)} = v_{t-\frac{1}{2}\Delta t} + a_{(t)}\Delta t \quad (37)$$

The leap-frog algorithm improves the main two problems of Verlet, it includes the velocity implicitly and it has better numerical accuracy as it does not calculate large differences. Still, the main disadvantage of leap-frog is that the velocities and positions are not synchronized. The **velocity Verlet algorithm** resolves this last issue by calculating acceleration, velocity and positions all at the same time without compromising precision using equations 38 and 39.<sup>209</sup>

$$r_{(t+\Delta t)} = r_{(t)} + v_{(t)}\Delta t + \frac{1}{2}a_{(t)}\Delta t \quad (38)$$

$$v_{(t+\Delta t)} = v_{(t)} + \frac{1}{2} \left( a_{(t)} + a_{(t+\Delta t)} \right) \Delta t \quad (39)$$

In addition to the already mentioned algorithms, there are more complex integration schemes, like Beeman's algorithm. This one includes more accurate velocities and gives better energy conservation. When performing MD simulations, the choice of algorithm should factor in its computational efficiency and memory requirement. However, there are additional aspects that also need to be considered, which we explore in detail in the next section.<sup>176</sup>

### 2.2.2.1 Setting up MD simulations

Choosing the right time step is crucial when performing MD simulations. The smaller the time step the better or more realistic the trajectory will be. However, it comes with a high computational cost, and it will not cover a large proportion of the phase space. On the other hand, if the time-step is too large, it can cause problems with the integration algorithm. The maximum value of a time-step is determined by the fastest frequency oscillation. In general, the time-step should be one order of magnitude smaller than the fastest process. Therefore, time-steps are usually of the order of fs. In most systems, this process corresponds to the stretching vibration of hydrogen bonds. As these motions are not usually of interest, one approach is to freeze all the vibrations of hydrogen bonds using the SHAKE algorithm to use a larger time-step and obtain more simulation time at the same computational cost.<sup>167,209</sup>

MD simulations are usually performed under periodic boundary conditions (PBC) to reduce surface effects and avoid outer solvent molecules to flying off into space. The system that is being modeled is situated in the center of a solvent box that is duplicated in all directions in space forming a lattice. Under PBC if an atom or solvent molecule leaves the solvent box, its image will enter the box in the opposite site, with this way making sure that the system does not see vacuum (Figure 2.6).<sup>167,209</sup>

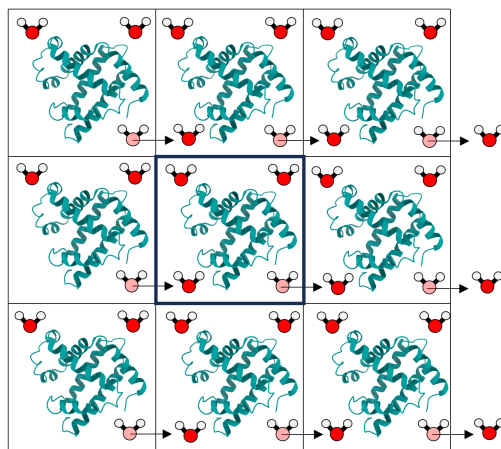


Figure 2.6: Schematic representation of periodic boundary conditions.

One of the main problems of PBC is that non-bonded interactions usually extend further than the simulation box. In the case of VdW, a cut-off is usually used because the Leonard-Jones potential decays with distance. The cut-off function can introduce some discontinuities, in most cases, a switching function is used. In both cases it requires updating a pair-list with all the atoms that should be accounted. In the case of coulombic electrostatics, a cut-off would introduce artifacts, consequently Ewald sum<sup>226</sup> or Particle Mesh Ewald<sup>227</sup> is normally used.

Usually starting structures are probably not realistic, consequently, before performing MD simulations, optimization algorithms are used to obtain minimized conformations. A minimization is performed to relax the system and avoid steric clashes that would lead to problems in the MD simulations. The main optimization algorithms are steepest descent and conjugate gradient. This is usually followed by an equilibration step in which the system is slowly heated until it reaches standard temperature (300K). Systems can be further equilibrated into different ensembles and finally the system runs under production state, usually under constant pressure, for the length that is desired depending on the objective MD simulations can be performed in different constant ensembles.<sup>167,209</sup>

Most common ensembles are NVT and NPT. The temperature of the system can be controlled to achieve certain constant temperature. As instant temperature depends on average kinetic energy, velocities can be modified at each time step to achieve certain temperature. Temperature can be modified either by scaling the velocities by a factor to achieve certain temperature or by coupling to an external a bath at certain temperature. The heat bath adds or removes energy by a thermostat couple to a term (Berendsen)<sup>228</sup>. As these methods introduce some fluctuations to the systems, the most widely used method is the Nosé-Hoover<sup>229,230</sup>, that is much more accurate because the bath is a part of the system. In NPT the pressure of the system is controlled by a pressure bath, in which instead of the velocities, the coordinates are scaled to modify the volume of the system. Usually, Nosé-Hoover is used to maintain the pressure.<sup>167,209</sup>

### 2.2.3. Enhanced sampling methods

As mentioned in the introduction, proteins do not have a unique structure, they exist in a dynamic ensemble and able to visit different conformations across the potential energy surface (PES). To understand the function of proteins it is essential to study the transitions between conformations of the landscape, which depend on the energy barrier between them. The problem is that if the energy barriers are large these transitions may happen at large time scales, around *ms*. Nowadays, even though GPU and parallelization have increased the speed of MD simulations, usually they are limited up to  $\mu s$ . One possible solution to this challenge is the application of enhanced sampling algorithms, which accelerate the dynamics and the sampling of rare conformations.<sup>231</sup>

The basis of these methods is to bias the PES of biomolecules and enhance the sampling of the conformational space by overcoming high energy barriers.<sup>232</sup> These methods are divided in two. Methods like umbrella sampling, metadynamics or adaptative biased force (ABF) use predefined collective variables (CVs) or reaction coordinates like RSMD (Root-mean-square deviation), distances or angles to guide the simulations. However, identification of the collective variables is not easy and the CVs can limit the sampling. Therefore, there are methods that do not use predefined reaction coordinates, like replica exchange molecular dynamics, temperature-accelerated dynamics or gaussian accelerated molecular dynamics.<sup>233</sup> This latter method is the one that has been used in some of the works of this thesis.

#### 2.2.3.1 Gaussian accelerated molecular dynamics (GaMD)

The predecessor of GaMD, are accelerated molecular dynamics (aMD), which were first implemented by Hamereld *et al.* on biomolecular systems.<sup>234</sup> aMD modifies the PES by applying a non-negative boost potential function to the true potential when the energy is lower than certain energy threshold, decreasing the energy barriers and accelerating the sampling of other conformations. Despite its advantages and showing good results for small systems, when applied to larger systems aMD reweighting has a lot of statistical noise. In GaMD, a harmonic boost that follows a Gaussian distribution is applied, which solves the reweighting problem because the original PES can be recovered by Gaussian approximation (cumulant expansion to the second order).<sup>235</sup>

The fundamental principle of GaMD is that when the energy potential of a system at positions  $\vec{r}$  is below a specific threshold energy  $E$ , the PES is smoothed by the application of a harmonic boost potential  $\Delta V_{(r)}$ . When the potential is higher than the threshold, the boost is set to zero,  $\Delta V_{(\vec{r})} = 0$ ), and the potential remains the same as the original.<sup>236</sup>

$$\begin{aligned} \text{When } V(\vec{r}) < E : \quad v^*(\vec{r}) &= v(\vec{r}) + \Delta V_{(r)} \\ \text{When } V(\vec{r}) > E : \quad v^*(\vec{r}) &= v(\vec{r}) \end{aligned} \quad (40)$$

The harmonic boost potential is defined as in equation 41, where  $K_0$  and  $E$  are determined automatically applying three criteria.

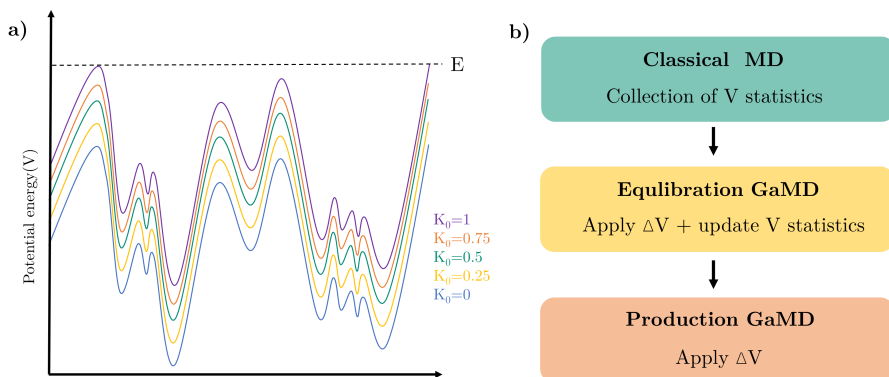
$$\Delta V_{(r)} = \frac{1}{2} k_0 \frac{1}{V_{max} - V_{min}} (E - V_{(\vec{r})})^2 \quad (41)$$

There are two possibilities for value reference  $E$ : either  $E$  can be set to the lower bound ( $E = V_{max}$ ) or to the upper bound ( $E = V_{min} + 1/k$ ), depending on the user input, iE=1 or iE=2, respectively. For each case, the value of  $K_0$  is obtained differently.

$$\text{When } E = V_{max} : \quad k_0 = \min(1, k'_0) = \min\left(1.0, \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{max} - V_{avg}}\right) \quad (42)$$

$$\text{When } E = V_{min} + 1/k : \quad k_0 = k''_0 \equiv \left(1 - \frac{\sigma_0}{\sigma_V} \cdot \frac{V_{max} - V_{min}}{V_{avg} - V_{min}}\right) \quad (43)$$

$V_{avg}$  is the average and  $\sigma_v$  is the standard deviation of the potential energy of the system ( $V$ ). The user specified the  $\sigma_0$ , for example, it can be  $10k_B T$  for accurate reweighting. The value of  $K_0$  should always be between 0 and 1. For equation 43, if  $K''_0$  is lower than 0 or higher than 1,  $K_0$  is calculated using 42. The  $K_0$  value is what determines how big the boost that is going to be applied. The higher the  $K_0$ , the higher is going to be the boost, as displayed in **Figure 2.7a**. Whenever the potential value is lower than  $E$ , the boost is applied according to the value of  $K_0$  in all the wells. In large systems, the bigger the standard deviation of  $\Delta V$  is, the smaller the  $K_0$  should be to obtain a correct reweighting.<sup>236</sup>



**Figure 2.7:** a) Representation of different PES obtained by adding a harmonic boost potential with different values of  $K_0$ . b) Steps of GaMD simulations.

GaMD simulations include the option to apply the boost potential to the total potential energy, to the dihedral energy, or to both simultaneously. The latter is known as dual-boost and provides higher accelerations. Recently, the option to boost the non-bonded potential alone or with dihedral energy has been added. Before GaMD simulation, energy minimization and a short cMD are performed to obtain initial well-balanced and minimized conformation.<sup>237</sup>

GaMD simulations are divided into three stages: **1) cMD** **2) GaMD equilibration** and **3) GaMD production (Figure 2.7b)**. The cMD comprises a preparatory stage in which the system is equilibrated and a second stage in which the potential statistics are collected ( $V_{max}$ ,  $V_{min}$ ,  $V_{avg}$  and  $\sigma_v$ ). These statistics are used to calculate the boost potential parameter, applied in the GaMD pre-equilibration stage. In the second stage of the GaMD equilibration, the boost potential is applied while at the same time the potential statistics are updated. After the GaMD equilibration it is assumed that the collected statistics accurately represent the PES. Consequently, the potential statistics are fixed and the boost potential is calculated. In the GaMD production, the boost potential is applied and the boost parameters remain constant without further updates.<sup>237</sup>

Based on GaMD, two other algorithms have been recently developed: LiGaMD and Pep-GaMD. These simulations are used to simulate binding and unbinding of ligands/peptides to proteins. This is achieved by applying boost on the non-bonded potential of the ligand alone or potential of the peptide, respectively, or this in combination with remaining total potential energy of system.<sup>238,239</sup>

### 2.2.4. Analysis of molecular dynamics simulation

As many programs have increased their user-friendliness and computer power has dramatically improved, performing a simulation has become a lot straight forward. Moreover, simulations lead to conformational spaces far wider than years ago. One of the challenges of these areas becomes therefore how to analyze and get relevant values for key questions on the systems. In this PhD, we took a particular concern to find practical solutions for extensive and relevant analysis.

One of the most crucial aspect to consider when performing a simulation is determining if a good conformational sampling has been achieved. It is essential to assess if the length of the simulation is enough to reached convergence. Convergence refers to the point in which enough data points have been collected that accurately sample the system phase space, where all states have been visited.<sup>208</sup> However, the concept of convergence is itself ambiguous, as MD simulations cannot be considered completely converged due to the intrinsically statistical uncertainty of them. Still, there are qualitative indicators that can address the quality of the convergence of simulations and more importantly, they indicate if all the conformational space has been sampled.<sup>240</sup>

Specifically, the methods should be able to ascertain whether the conformational space of the system has been completely explored in order to determine if the MD simulation has reached a length where statistical analysis can be extracted, or certain event can be observed. The set of tools used in this thesis including RMSD, all-to-all RMSD, counting clustering method, RMSF and PCA analysis. These methods have been tested to assess the convergence and conformational exploration of MD simulations in previous studies.<sup>194</sup> In **Figure 2.8** are represented all set of measures used in this thesis.

The most common strategy to analyze an MD trajectory is to perform RMSD analysis, which consist of computing the root mean square deviation (RMSD) of each frame against the initial reference structure.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_t - x_{ref})^2} \quad (44)$$

The presence of fluctuations in the RMSD indicates that the system has not converged, although its absence is not an indication of convergence. The main

application of RMSD is to know if the system is systematically changing. However, this method is limited because it does not provide information about the exploration of the conformational space during the simulation, or the states sampled during simulation. This limitation comes from the fact that the RMSD condenses a wide of structural information (3N dimensions of conformational space) to a unique number<sup>240</sup>

A more powerful indicator is all-to-all RMSD. This analysis consists of calculating the RMSD against all structures along the simulation instead of only the initial structure. The diagonal will always be zero, but low RMSD values along the diagonal indicate the occupation of a certain state. This is a better measure for the exploration of the system because low RMSD away from the diagonal indicate that the system is visiting an already sampled state.<sup>241</sup>

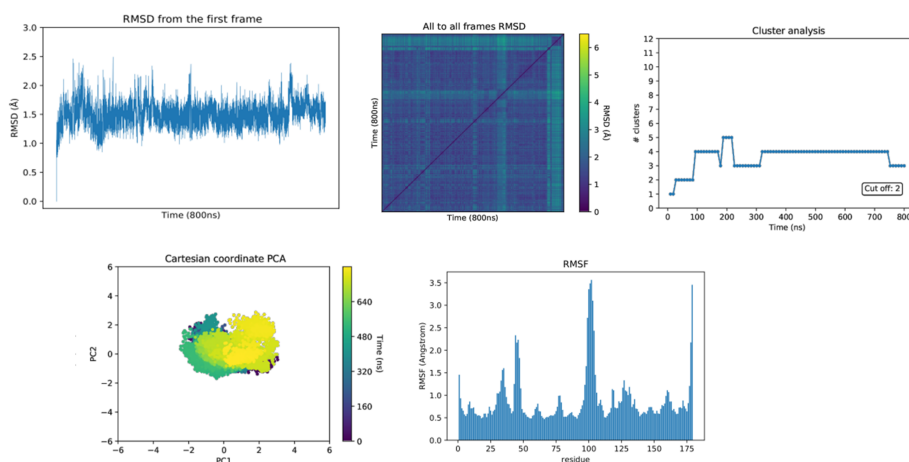
Another metric implemented by Daura *et al* is cluster counting.<sup>242</sup> A clustering algorithm groups data points based on a distance metric.<sup>241</sup> In the context of biomolecules, the clustering divides an ensemble of structures into sets of similar structures and the distance metric is RMSD.<sup>240</sup> For convergence analysis cluster counting is used to evaluate the rate of discovery of new clusters during the simulation using a RMSD cut-off set accordingly to each structure. When the rate of cluster counting is low it is assumed that the simulation has reached convergence.<sup>242</sup>

The PCA is another tool that has been used extensively for analysis of simulations. PCA analysis aims to extract the essential motions from a set of conformations by plotting the two principal PCA modes over time. This measure is an indication of the collective motions of the protein, but it is used to know how the system suffers a series of dynamic transitions through the simulation to visit different conformations. By convention, if the system returns to the starting point it means that the trajectory has converged.<sup>240,243</sup>

Finally, a similar measure to RMSD is RSMF (root mean square fluctuation), which indicates the fluctuation of an atom over a simulation. It can also be used to extrapolate the involvement of certain residues or regions in protein motions.<sup>244</sup>

$$RMSF = \sqrt{\langle (x_i - \langle x_i \rangle)^2 \rangle} \quad \langle x_i \rangle = \text{ensemble average position} \quad (45)$$

Related to convergence and conformational sampling, there is another controversial topic, performing replica simulations. Running multiple replicas in MD simulations is advised as it can provide better statistical sampling and improves the accuracy of results. More importantly, performing multiple replicas allows the exploration of different regions of the conformational space. Therefore, in cases in which new conformations want to be visited, replicas are highly advised.<sup>245</sup>



**Figure 2.8:** Examples of MD analysis: RMSD, all-to-all RMSD, counting clustering, PCA and RMSF.

## 2.2.5. Protein-ligand dockings

Protein-ligand dockings are a MM method that intends to predict the favored orientation and interactions of a ligand in a protein binding site. Docking algorithms are divided in two: **1)** A sampling algorithm performs an exploration of the conformational space and generates all possible orientations of the ligand in the protein binding site. **2)** All poses are evaluated using a scoring function, which approximates the binding energy of the complex.

### 2.2.5.1 Sampling algorithm

Docking a ligand into a protein binding site involves many degrees of freedom. On the one hand, there are 3 degrees of translational and 3 degrees of rotational freedom of one molecule relative to the other. On the other hand, there are the degrees of freedom of each molecule by itself, the ligand and the protein.

Dockings are classified depending on how many degrees of freedom are considered of each the protein and ligand. Initial docking programs considered both the protein and ligand as rigid entities, these are rigid docking, which only account for the six translational and rotational degrees of freedom, no conformational degrees of freedom are considered. However, through the years flexibility has been incorporated to either the ligand (flexible-ligand docking) or/and the protein (flexible-protein docking).

Three different sampling algorithms are able to identify different poses of a ligand around a chosen binding site:<sup>132,246,247</sup>

- **Shape matching** is one of the simplest methods as it is based on structural shape complementarity of the ligand and protein. A negative image of the binding site is constructed as a set of spheres the algorithm tries to match to the ligand atoms. This is usually performed as rigid docking. To contemplate flexibility a set pre-generated ligand conformations can be used as input. Examples of this include DOCK software.
- **Systematic search** algorithms consist of generating all possible conformations of the ligand by different methods and then each conformation is docked. Glide or Fred are algorithms that use exhaustive search method, in which all conformations are generated by rotating all bonds of the ligand. Ludi software uses fragmentation method as the ligand is divided in fragments, each fragment is place on the binding site and augmented gradually.
- **Stochastic methods** are based on generating random changes to the ligand, both in conformation and translation/rotation. In total four algorithms can be included in this category: Monte Carlo (MC), evolutionary algorithms (EA), Tabu search and swarm optimization (SO).

GOLD and GaudiMM, the programs used to perform docking on this thesis are based on EA, specifically on Genetic Algorithms (GA), which try to mimic the concept of evolution.<sup>248,249</sup> A random initial population of individuals is generated, each individual object characterized by a series of genes that describe the conformation of the ligand and protein. This set of individuals are allowed to recombine and mutate, specifically the crossover operator copies genes from the parents to a new child individuals, while the mutation operator mutates genes at randomly individuals. The best individuals will be selected according to the

scoring function and will propagate to the next generation and after a number of iterations the population will have evolved, surviving only the best.<sup>250</sup>

One of the main problems of dockings is that proteins have many degrees of freedom due to its large size. Current algorithms consider protein flexibility in different forms: **1)** flexibility is applied only on the side chains as means of a rotamer library **2)** using an ensemble of protein conformations or use Normal Modes calculations to account for different protein conformations. **3)** protein-ligand complexes can be minimized by MD or MC **4)** application of soft docking.<sup>132,246</sup>

### 2.2.5.2 Scoring functions

The scoring functions are a key element as they evaluate the docking poses, determining the quality and accuracy of the results. Scoring functions try to estimate the binding affinity between ligand and protein with a series of physicochemical parameters like intramolecular interactions, electrostatic effects... The accuracy of the scoring function increases with the number of parameters included, but at the cost of computational speed. Therefore, scoring functions must balance speed and accuracy to be reliable. The key benefit of scoring functions is their ability to evaluate rapidly and efficiently the energy, which is a mandatory requirement when dealing with a large conformation space to explore.

Scoring functions can be classified into three different groups:<sup>251,252</sup>

- **Force field:** estimate the energy of binding based only on the sum of non-bonded interactions derived from force-field parameters. The electrostatic terms are modelled by the Coulomb equation, while the VdW term is described by the Leonard-Jones potential. One challenge for FF functions is to incorporate solvent effects. To account for this some FF scoring functions incorporate distance-dependent dielectric constant. The main limitation of this is that it does not account for entropic effect. An example is Goldscore function, which accounts for the following terms:<sup>253</sup>

$$\text{Gold fitness} = S_{hb_{ext}} + S_{vdw_{ext}} + S_{hb_{int}} + S_{vdw_{int}} \quad (46)$$

- **Empirical:** energy is decomposed into weighted energetic terms: VdW, hydrogen bonds, electrostatics, entropy of solvation or hydrophobicity. A training set of protein-ligands complexes with known affinities is used to determine the weight of each term. Docking programs commonly use empirical functions because they are more computationally efficient than others due to their simple energetic terms. The function Chemscore is implemented in Gold as:<sup>254</sup>

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hb}S_{hb} + \Delta G_{met}S_{met} + \Delta G_{lipo}S_{lipo} + \Delta G_{rot}H_{rot} \quad (47)$$

- **Knowledge-based:** based on statistical potentials, specifically, potential mean force (PMF), directly derived from structural information. PMF is the inverse of the Boltzmann relation using the frequency of interactions in a protein-ligand complex.
- **Machine learning:** employ different machine learning algorithm like random forest, and vector machine to learn how to predict the binding affinity using a training data set. Some machine learning scoring functions have outperformed classical scoring functions, but as they are difficult to incorporate on docking software they have been used as rescoring functions.<sup>255</sup>

During the past few decades, a lot of effort has been put into scoring functions to consider ligand solvation and entropic changes, but still, protein flexibility is not fully considered in most cases, and when it is considered, it is at a high computational cost. There is still room for improvement in the field of molecular dockings.

### 2.2.5.3 Metals in dockings

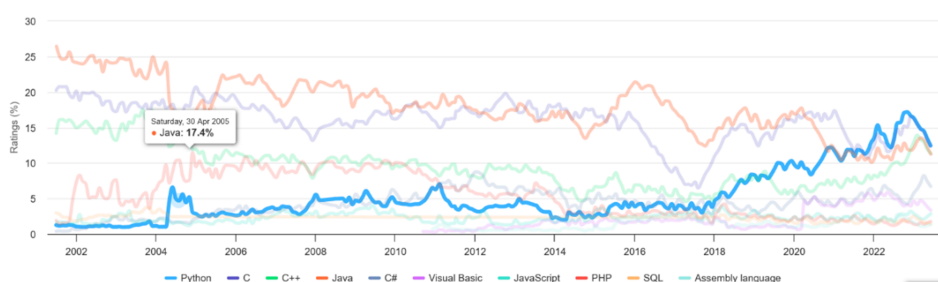
In most docking software, metal coordination is not considered; precise docking predictions, including metallic ligands, are still a work in progress, mainly due to the absence of metallic contribution in most scoring functions. This thesis implements the methodology used to consider metal coordination in metals in GOLD. The premise of this method is that, computationally, coordination bonds can be treated with hydrogen bond like functions.

In this model, the acceptor is a fictitious hydrogen that is added to the vacant sites of the metal to preserve the coordination geometry, and the electron donors are protein side-chains. A parameter file containing all pseudo-covalent interaction parameters for each metal and the metal interaction can be evaluated in the hbond intermolecular term in the Goldscore function.<sup>256</sup>

### 2.2.6. Python language in science

In this final section of the methods chapter, we will explore the development of new software for scientific applications and the primary languages and packages employed. This is particularly relevant, as two projects of this thesis involve the development of new software.

Until the mid-1980s, scientific programming was dominated by Fortran. However, nowadays, a wide range of programming languages and libraries are applied to different fields depending on their applications. In numerical computing, MATLAB dominates, whereas R is widely employed in statistics. Nevertheless, the use of Python has grown exponentially in computational chemistry and bioinformatics. According to the TIOBE index<sup>257</sup>, Python has grown in popularity in data science, surpassing C++ and Java over the last decades (Figure 2.9).



**Figure 2.9:** Popularity of programming languages over the years with Python highlighted. Obtained from: <https://www.tiobe.com/tiobe-index/>

Python was released in 1991 as an easy-to-use programming language. It is free and open source, it is a high-level and general purposed, object-oriented programming language. One of the greatest strengths of python is its flexibility thanks to its large set of packages and libraries. Its competitor, C++, is more

difficult to learn and has a slower learning curve. It is, required compilation and lacks built-in support and interactive execution. However, C++ is also widely used in software development as it is very fast. Therefore, in a lot of cases, C++ is combined with Python or sometimes Python includes libraries that are written in C++ to be faster.<sup>258</sup>

In scientific programming the packages that are commonly used include: NumPy (numerical arrays and mathematical functions), panda (tabular data), SciPy (high level numerical functions), matplotlib (visualization and plotting) or scikit-learn (machine learning). Apart from these general packages, several chemistry packages have been developed over the years: RDKit, pyChimera or pyPDB between others. In this thesis all software and scripts have been developed in Python due to all its advantages and the availability of packages dedicated to chemistry and biology.

## CHAPTER 3

# Objectives

Molecular modeling techniques play essential roles in exploring the conformations of molecular structures and understanding their behavior at the atomic level. The upstanding computational methodologies, especially multiscale ones, have matured enough to comprehensively study molecular entities containing metal moieties, including both natural and Artificial Metalloenzymes (ArM). Indeed, despite advances in the last decade, computational methods focused on metallic biomolecules still face some barriers and challenges.

This PhD aims to overcome some of the main computational challenges in metalloproteins, focusing on two particular fields of research: the study of heme-containing proteins and the design of ArM. This research is divided into two general objectives, each with its respective sub-objectives:

### 1) Study of heme binding proteins:

The main objective of Chapter 4 is to unravel the complex binding processes of heme and its corresponding protein counterpart. This is dealt in two ways:

1. Establish an integrative computational workflow tailored to decode heme-binding processes for hemoproteins. This work focuses on hemophore HasA from two species that present different heme binding mechanisms.
2. Development of software for identification of heme binding sites. Apply this software to detect natural heme binding sites in proteins and find new sites to design novel heme-ArM.

## 2) Computer-aided design of ArMs:

The second objective revolves around the study and design of new ArMs.

1. The objective of Chapter 5 is to apply updated integrative molecular modeling techniques to investigate and design ArM. The work is focused on guiding or rationalizing two ArM based on streptavidin. The first study encompasses optimizing a family of Au-containing ArMs for heterocyclization, while the second focuses on the rationalization of ArM catalyzing a Suzuki-Miyaura reaction.
2. The objective of Chapter 6 is to develop software for predicting metal binding sites focusing on its application on ArM design. As a case study, we employ  $\alpha$ -Reps proteins to validate the utility of this software.

## CHAPTER 4

# The relevance of heme binding processes and their prediction

Decoding the interactions between inorganic moieties and biological partners is a crucial aspect for gaining insights into the origin of life and exploring new biotechnological routes like designing ArMs. The study of metal-mediated binding processes is one of the most complex questions that could be addressed. The interplay between biology and inorganic chemistry surpasses standard knowledge of both chemical and biological sciences and assessing the relative influence of both partners in the recognition process is extremely challenging.

Heme *b* is one of the most ubiquitous metal containing ligands in nature and the introduction of this thesis has already highlighted its biological relevance in several biological functions, from transport of O<sub>2</sub>, redox catalysis to gene regulation. Therefore, heme *b* is a prototypical system to study the binding of metallic cofactors to proteins. In metalPDB, a curated structural database for biometallic species, in 2023 5.315 PDBs correspond to proteins that bind heme groups (9.24%). Despite the importance of heme, the binding of heme to proteins and its prediction has not been studied extensively.

In this chapter, we will study different aspects related to heme binding processes. The first section will focus on the preorganisation of heme binding sites and the mechanisms of heme binding of a family of heme proteins. The second section is dedicated to developing a new heme binding predictor for natural heme binding sites and exploring potential new heme binding locations for the design of new ArM based on heme.

## 4.1. Exploring the molecular events of heme binding mechanisms

In bioinorganic chemistry, heme binding mechanisms have gained a lot of attention. As introduced in section 1.2, these can be divided into two distinct modes: 1) Transient heme binding proteins require a fast heme binding and unbinding process to allow a rapid response as usually they are related to heme recruitment, transport or regulation.<sup>259</sup> Generally speaking, these systems contain a bidominal structure connected by a flexible loop that allows rapid association. This suggests that no big conformational changes should occur at a molecular level.<sup>68</sup> 2) Proteins that bind heme strongly and permanently correspond to catalytic, electron transfer, and any system requiring a strong metal-protein bond. In these cases, a stronger affinity is expected, and at the molecular level one would expect to observe folding induced by the binding of the heme.<sup>97</sup>

Data available at the structural level has revealed that proteins that bind heme permanently have apo structures similar to its holo form, but the heme binding region is disordered as in cytochrome b<sub>562</sub>.<sup>100</sup> In other cases as P450<sub>cam</sub>, the apo form exhibits a generally more disordered structure.<sup>260</sup> Contrarily, both apo and holo forms of transient heme proteins are completely folded, but there structural differences as in HemS or HasA.<sup>103,261</sup>

How the binding of metals impacts the apo protein remains a highly debated question. Nonetheless, in transient heme proteins, certain crystallographic structures highlight the ambiguity between the necessity of a conformational change or not. In fact, there are structures of transient systems that display significant variations between the apo and the holo form, as in the case of bacterial Hemophore HasA.<sup>261</sup> The main questions posed by these systems are whether these movements are necessary for binding or a consequence of the binding, and how they occur in certain species or under different circumstances.

Until now, computational studies regarding heme recognition processes by its binding partners have been limited. In most studies, simulations have been targeted in a way that the binding of the cofactor always induced the conformational change. These approaches do not allow us to see whether the apo form is naturally suitable for heme binding. It is impossible to determine

whether the heme binding mechanism follows an induced fit model, a conformational selection model, or a combination of both. The challenge here lies in the need for a theoretical protocol capable of studying the possible heme binding mechanisms (conformational selection or induced fit) and the potential preorganization or reorganization of the system. This is where the significance dockings and GaMD (Gaussian accelerated MD) combined comes into play.

Before exploring the heme binding mechanisms with GaMD and dockings, we explored the correlation between the conformational variation of the heme binding pocket and the ability to bind the heme with molecular dockings. At the same time, this study allows us to validate whether the molecular docking technique that we use can correctly capture metal interactions.

#### 4.1.1. Conformational variation study of heme binding sites

To test if our docking approaches are valid for studying heme binding complexes, ten pairs of apo-holo structures of transient heme proteins with different behaviors were chosen. Some have similar apo and holo structures, while other have completely different structures (though not losing their secondary structures). The objective of this work is also to use molecular dockings to get a first glimpse of the preorganization of heme binding sites by determining if heme is able to bind in the apo form and how different is this binding compared to the holo form. Performing these docking into apo-holo pairs that have different heme binding mechanism allow to validate if the preorganization of the apo binding site is present in all heme binding mechanism. At the same time, it could validate that the employed dockings have the capability to capture cofactor-protein interactions.

Rigid dockings were carried out on both the crystallized apo and holo forms of the heme proteins. Results are represented in **table 4.1**. The third and fourth column indicate the RMSD between apo-holo structure and whether there is a conformational change in the heme binding site. The fifth column indicates how many coordinating residues are detected for the heme. These results demonstrate the accuracy of our molecular docking approach, it is able to correctly detect all metal-protein interactions. In most cases heme coordination is observed through rigid docking, except for two cases where flexibility is needed in the coordinating residue.

Heme binding protein	Holo-Apo PDB	RMSD (Å)	Conformational change	Docking found coordination
Heme-binding protein <b>PhuS</b> ( <i>Pseudomonas aeruginosa</i> )	4mf9 - 4mgf	0.365	No	1 out of 1
Hemophore <b>HasA<sub>yp</sub></b> ( <i>Yersinia pestis</i> )	4jet - 4jer	0.536	No	1 out of 1
Anthrax hemophore <b>IsdX1</b> ( <i>Bacillus anthracis</i> )	3sik - 3sz6	0.675	No	1 out of 1
Heme oxygenase <b>HO-1</b> ( <i>Homo sapiens</i> )	1n45 - 1ni6	0.619	No	1 out of 1
Hemoglobin binding hemophore <b>Hbp2</b> ( <i>Listeria monocytogenes</i> )	4myp - 4nla	1.130	No	1 out of 1
Periplasmic Heme-Binding Protein <b>ShuT</b> ( <i>Shigella dysenteriae</i> )	2r7a - 2rg7	1.019	Yes	0 out of 1 *(1 out of 1)
Haem-chaperone Proteobacteria-protein <b>HemS</b> ( <i>Yersinia enterocolitica</i> )	2j0p - 2j0r	1.363	Yes	1 out of 1
latex clearing protein <b>LCP</b> ( <i>Streptomyces sp. K30</i> )	5o1m - 5o1l	1.534	Yes	1 out of 2
Heme-regulated transporter regulator <b>HrtR</b> ( <i>Lactococcus lactis subsp. lactis Il1403</i> )	3vp5 - 3vox	3.679	Yes	1 out of 2
Hemophore <b>HasA<sub>sm</sub></b> ( <i>Serratia marcescens</i> )	1dkh - 1ybj	5.612	Yes	0 out of 2 *(1 out of 2)

\* Result with flexibility on coordinating residues

**Table 4.1:** Dockings results for apo-holo pairs of heme proteins.

When the conformation of the apo form is essentially the same as the holo like HasA<sub>yp</sub>, HO-1, Hbp2, PhuS or IsdX1, the apo dockings reproduce the crystallographic holo structure. In all five cases, heme binds and coordinates with its corresponding coordinating residues. Therefore, these cases could correspond to rigid binding (lock and key) mechanisms with a slight conformational change needed for the binding of heme, either minor conformational selection or minor induce-fit mechanisms are needed.

On the other hand, there are cases in which the overall structure of the apo form is similar to the holo form except for some parts of the heme-binding site helices like in LCP, ShuT, HrtR, HemS or HasA<sub>sm</sub>. Docking calculations for most of these cases indicate that heme can bind to its binding site and establish coordination with at least one of its possible coordinating residues. In two cases (ShuT and HasA), rigid docking fails to find coordination, but coordination is achieved when flexibility is introduced to the coordinating residue. When the coordination of the second residue is not obtained in the docking's calculations, it is indicative that the second coordination is only possible after a conformational change. In some cases, the change needed is small, but in other cases, like HasA<sub>sm</sub> it is a considerably change. These cases could correspond to an induce-fit effect or to a conformational selection phenomenon.

From this initial study it can clearly be concluded that the docking approaches as they were improved in our lab are viable for heme binding complexes. Moreover, the study concludes that there is a preorganization of the heme binding site, even though in some cases there are conformational changes, heme is still able to bind in all cases with at least one coordinating residue. It suggests that in most cases, the apo form can be used to predict heme binding sites. This premise led to the idea of developing software for detecting heme binding sites based only on the structure and preorganization of the heme binding site, which will be further developed on the 4.2. section.

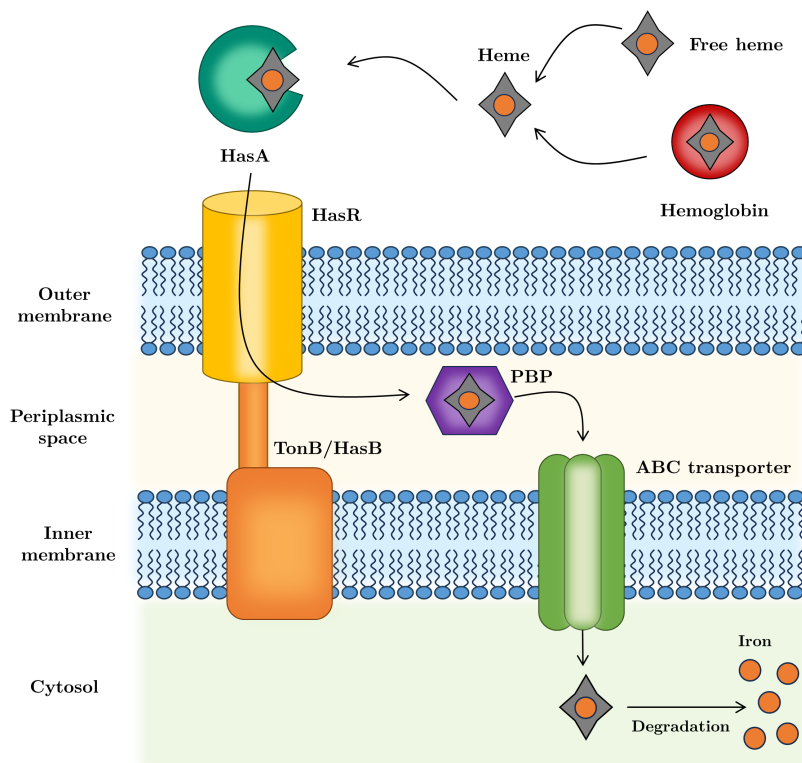
Based on these results though, it also appears that some amount of dynamic behavior could exist, depending on the cases. In those cases, docking calculations on apo structures may be unreliable and some amount of molecular motions need to be considered. In the following section we will study the heme binding mechanism, two opposite cases (HasA<sub>yp</sub> and HasA<sub>yp</sub>). The former corresponds to a slight conformational change upon heme binding and the latter to a major one. These are particularly interesting cases because they correspond to very close analogs from two pathogenic bacteria. The previous docking calculations suggest that this unique structural difference could describe very different heme binding modes.

As there is no clear molecular description of heme uptake and only a limited number of studies regarding heme binding mechanisms exist, this computational study pretends to give further insight by focusing on hemophores HasA. The challenge of this work lies in comprehending the precise nature of the interactions between heme and hemophore HasA and their dynamic nature. It requires combining different computational approaches, including enhanced MD simulations and molecular dockings.

#### **4.1.2. Heme binding processes in hemophore HasA**

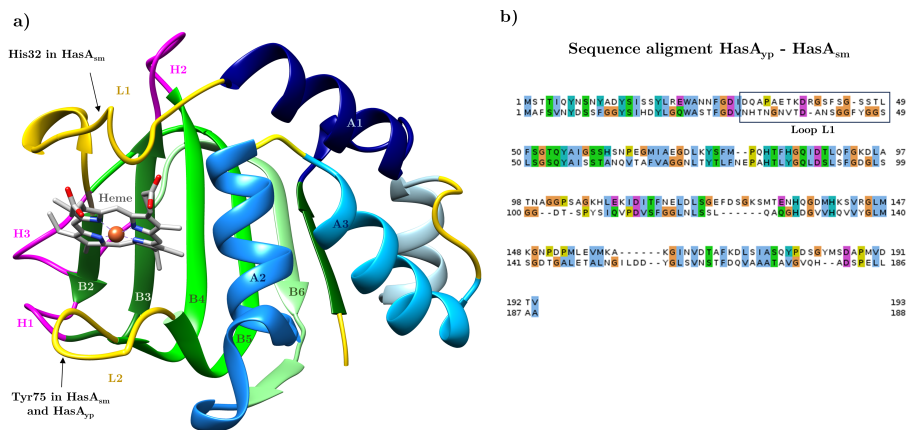
HasAs are extracellular heme-binding proteins that gram-negative bacteria use as heme uptake system. They can acquire free heme or extract heme from hemoglobin and deliver it to a specific receptor at the cell surface, HasR.<sup>262,263</sup> Heme is transported across the outer membrane to the periplasm by HasR, whose active transport is dependent on the energy provided by TonB or HasB complex.<sup>264</sup> Once in the periplasm, periplasmic binding proteins transfer heme

to an ABC transporter to be internalized to the cytosol, where the heme is degraded and used as an iron source (**Figure 4.1**).<sup>265</sup> HasA uptake systems have been identified in different negative gram bacteria organisms, including *S. marcescens*, *P. aeruginosa*, *P. fluorescens* and *Y. pestis*.<sup>266–268</sup>



**Figure 4.1:** HasA mechanism of heme transport into cytosol for gram negative bacteria.

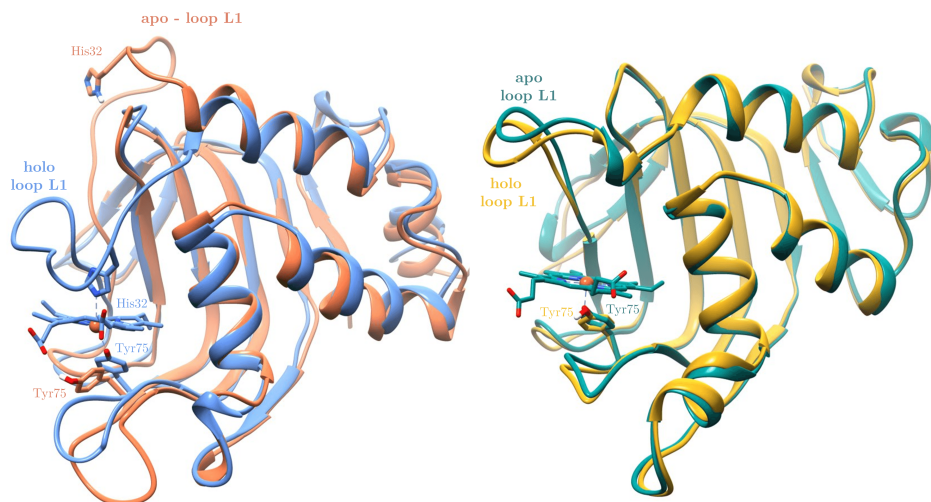
Interestingly, some hemophores from this family seem to have very distinctive heme binding mechanisms as already showcased by the previous docking studies, HasA from *Yersenia pestis* (HasA<sub>yp</sub>) and HasA from *Serratia marcescens* (HasA<sub>sm</sub>). The HasA family contains a  $\alpha + \beta$  fold structure, within which the heme is found between two large loops at the interface of the  $\alpha$  and the  $\beta$  domain: L1 and L2.<sup>261,269</sup> The overall structure of the heme-bound form of both HasA<sub>yp</sub> and HasA<sub>sm</sub> is similar, and the sequence identity is 31% (**Figure 4.2**), with the less conserved part being located at loop L1.<sup>270</sup>



**Figure 4.2:** a) Structure of HasA with secondary structure regions and heme coordinators indicated. b) Alignment sequences HasA<sub>yp</sub>-HasA<sub>sm</sub>. Reprinted from [271].

Despite this, experimental structures of both species' apo and holo forms show striking differences (**Figure 4.3**). In HasA<sub>yp</sub>, L1 loop position is almost invariant in apo and holo forms with a closed conformation, whereas in HasA<sub>sm</sub> a very large conformational change is observed between apo and holo form. The L1 loop displays an opened form in the apo structure and a closed form in the holo one, a motion of about 15 Å. The difference is the axial coordination of the iron in the two holo forms; the heme is bound to Tyr75 for the L2 loop in HasA<sub>yp</sub>, whereas in HasA<sub>sm</sub> it coordinates to both Tyr75 (L2) and His32 (L1).

Even though the spectroscopic data and the comparison between the apo and holo structures does not provide a clear molecular understanding, it can be deduced that very distinct heme binding mechanisms could occur. In this regard, computational tools could be a very valuable asset. Computational studies have been previously performed on HasA from *Pseudomonas aeruginosa* (HasA<sub>p</sub>), which displays similar geometrical features to HasA<sub>sm</sub>. Targeted molecular dynamics simulations (TMD) were applied to study the behavior of apo-HasA<sub>p</sub> upon heme binding. This led to the identification of a series of interactions and motions in helices  $\alpha 2$  and  $\alpha 3$  that could participate in the closing of loop L1.<sup>104</sup> Classical MD have not been able to simulate the transition from apo to the holo in HasA<sub>p</sub>, these have been used with mutant forms to pinpoint the residues that could be important for the closing of loop L1, such as Arg33.<sup>272</sup> However, Arg33 is not present in HasA<sub>sm</sub> and the residues involved in this species are not known.



**Figure 4.3:** Structure overlap between apo (1ybj) and holo (1dkh) forms of hemophore HasA from *Serratia marcescens* (left) and apo (4jes) and holo (4jet) form from *Yersinia pestis* (right). Shifts between apo and holo form in loop L1 are indicated. Reprinted from [271].

So far, the only computational study conducted on  $\text{HasA}_{\text{sm}}$  also involves TMD simulations. This study unravelled a funnel mechanism for the transfer of heme from HasA to HasR.<sup>273</sup> Regarding  $\text{HasA}_{\text{yp}}$ , no computational simulations have been reported. It is important to emphasize that TMD<sup>274</sup> simulations are informative for molecular mechanisms, but they are inherently biased as the steering forces are always defined to force the system towards a defined final. Such restraints restrict the ability to assess the conformational space that the protein explores. In this context, the evaluation of the conformational exploration of the apo form under standard conditions is biased by the restraints, making it difficult to determine the extend to which this exploration can be related to the heme binding mechanism.

The aim of this project is to apply a methodology that allows a wide conformational exploration of the apo-hemophore structures and long-range motions without imposing specific geometric restraints. The chosen method for this purpose is Gaussian accelerated MD (GaMD)<sup>235</sup>, methodology previously described in 2.2. Using a combination of GaMD simulations, the apo structures, intermediate heme-bound complexes, and the final complex were simulated to identify possible heme binding mechanisms, for both  $\text{HasA}_{\text{yp}}$  and  $\text{HasA}_{\text{sm}}$ . The

findings of this study give interesting insights into different heme binding mechanisms that could better envision protein engineering processes for heme containing enzymes and for artificial enzymes.

### 4.1.3. Methodology

The computational framework for this work combines two techniques: molecular dockings and GaMD simulations. The protocol starts with obtaining the crystallographic apo form available at the Protein Data Bank (PDB)<sup>275</sup>, which is used as starting point to run the initial GaMD simulation. The trajectories are analyzed, and the most representative cluster is used as input for dockings calculation of heme. From the docking results, GaMD simulations on heme-bound intermediates are performed to study the effect of heme binding to the protein conformation (Figure 4.4).

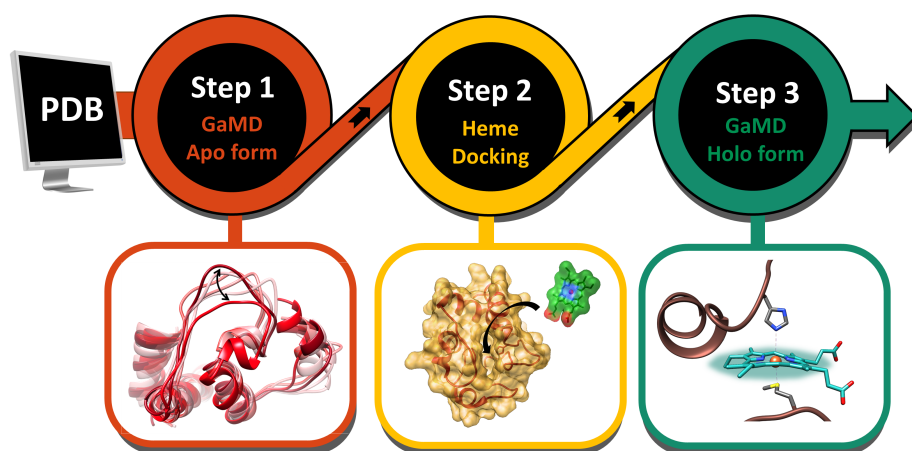


Figure 4.4: Multi-level computational protocol followed. Reprinted from [271].

**System Set up:** The X-ray structures of the apo forms of heme containing proteins were retrieved from the Protein Data Bank (PDB), 1ybj<sup>276</sup> was used for the apo form of HasA<sub>sm</sub> and 4jes<sup>269</sup> for HasA<sub>yp</sub>. Any crystallographic waters and small molecules present in structures were eliminated using UCSF Chimera<sup>277</sup>. Hydrogen atoms were added using Chimera. To ensure accurate prediction of protonation states of ionizable groups, the webserver H++ was employed to double-check.<sup>278</sup>

**Docking calculations:** Dockings were carried out with GOLD5.2<sup>254</sup> using a simulation box of 10-15Å, centered at the heme binding site. When necessary, side chain flexibility was applied on specific residues using the default rotamer library. The minimum number of operations was set to 100.000 and the number of the Genetic Algorithm (GA) runs to 50 for higher accuracy. For evaluating the solutions, the optimized version of Goldscore discussed in section 2.2, capable of predicting metal-protein interactions, was used.<sup>256</sup>

**GaMD simulations:** Unconstrained enhanced sampling was performed with the GaMD method.<sup>235</sup> As starting point for GaMD calculations, the coordinates derived from a classical MD after 10-20ns were used. In GaMD calculations, igamd was set to 3, applying force to dihedrals and to the total potential energy, while the threshold energy was set to the lower bound with a value of IE=1.

All systems were prepared with the xleap<sup>279</sup>. The force field ff14SB was applied for proteins, GAFF for non-standard residues, ions94.lid for ions and TIP3P for water.<sup>280</sup> For metalloproteins, metal parameters were obtained using the MCPB.py approach.<sup>168</sup> In MCPB.py, the charges were calculated using RESP<sup>281</sup> and the force constants and equilibrium parameters between the metal and the residues were calculated using the Seminario method.<sup>282</sup> Gaussian09<sup>283</sup> was used for optimization and frequency calculations at DFT level in water solvent (SMD continuum model).<sup>284</sup> The B3LYP hybrid functional, with Grimme's dispersion D3<sup>212</sup>, was employed with SDD+F (Fe) + 6-31G(d,p)<sup>285</sup>. Different Fe oxidation states (Fe<sup>2+</sup> and Fe<sup>3+</sup>) and multiplicity (low and high) were considered. Final calculations were performed with Fe<sup>3+</sup> and the multiplicity at sextet spin state.

For all GaMD simulations, the protein was solvated using an explicit solvent approach, in which the protein was embedded into a cubic box with a neutral charge. All simulations were performed using AMBER18<sup>279</sup> and periodic boundary conditions. Before the GaMD, preparatory classical MD simulations were performed. For these, an energy minimization was carried out to avoid steric clashes and relax the system, followed by equilibration steps that gradually heated the system from 100K to 300K. Finally, a production run of 10-20ns was performed. GaMD simulations started from coordinates of preceding classical MD. A GaMD equilibration of 50ns was performed followed by a production run of 800ns. In all cases, to ensure convergence and complete exploration of the conformational space, three replicas of each calculation were performed.

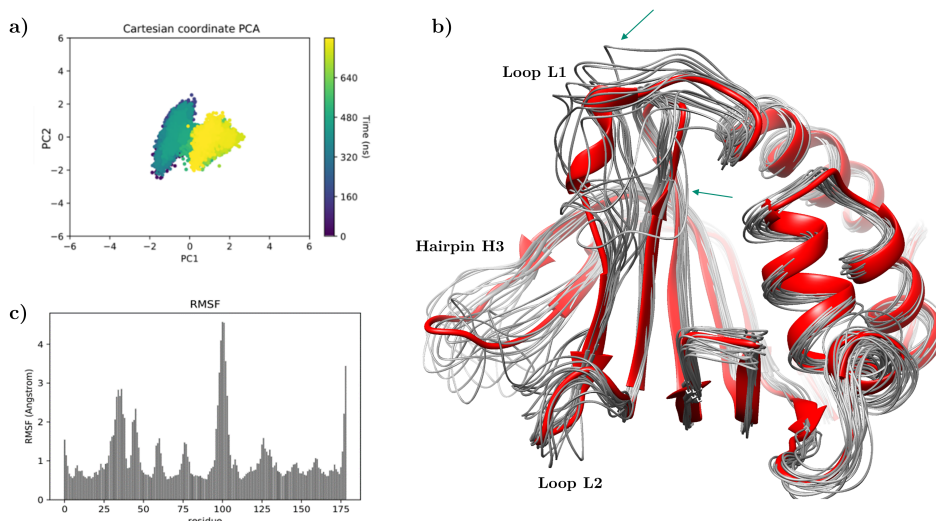
**Analysis of GaMDs:** All GaMD trajectories were processed using cpptraj implemented in Ambertools<sup>279</sup> and cluster analysis was performed with MDtraj.<sup>286</sup> The convergence of trajectories was assessed by evaluating the following factors: RMSD with respect to the initial structure, all-to-all RMSD, RSMF, PCA and a clustering counting method. These analyzes were performed using only the  $\alpha$ -carbons. For interaction analysis, the getContacts.py<sup>287</sup> script was used and modified accordingly to analyze all interaction types along the GaMDs simulations. The results from different replicas were combined to determine the mean frequency of contacts across all replicates. To study the differences between the apo and holo simulations, the variation of interactions was calculated and normalized. To calculate free energies, GaMD reweighting was performed with the PyReweighting toolkit and the Maclaurin method.<sup>288</sup>

In the following sections, HasA<sub>yp</sub> and HasA<sub>sm</sub> results are discussed separately and then a comparison between the two hemophores is presented.

#### 4.1.4. Results of Hemophore HasA<sub>yp</sub>

As represented before in (Figure 4.3b), the difference between the apo and the holo crystallographic forms of hemophore HasA<sub>yp</sub> is minimal, loop L1 remains in similar disposition in both forms. The objective of this part of the study was to assess the conformational space to see if changes in L1 could exist with respect to the X-ray structures. The main question to answer is: would it be possible that the loop could eventually reach conformation like HasA<sub>sm</sub> in the absence of heme? Three replicas of 800 ns GaMD simulations were performed to solve this question. The combination of five analysis tools revealed that after 800ns, all three replica simulations had converged, and the conformational space sampled minimal changes in the overall structure of the protein (Figure 4.5a and B.1).

The visual inspection of the trajectory and clusters of the GaMD simulations revealed how the tertiary structure of the protein, the  $\beta$ -sheet, and  $\alpha$ -helices displays very little conformational changes (Figure 4.5b). The loop and hairpin regions show some degree of flexibility, in particular hairpin H3 and loop L1 have an average RMSF of 2.86 and 1.87 Å respectively (Figure 4.5c). The latest oscillated during the simulation around its resting position although with only small changes in its rearrangement and could partially acquire the conformation of a small turn or helix.

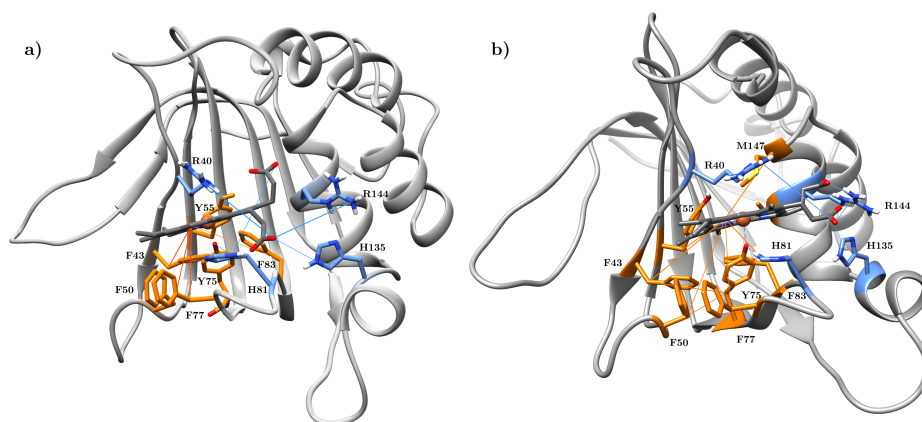


**Figure 4.5:** GaMD simulation of apo HasA<sub>yp</sub>. **a)** PCA extracted from one replica. **b)** Main cluster represented in red and remaining in gray. **c)** RMSF average of three replicas. Reprinted from [271].

Within the MD trajectory, only in punctual sections of the MD trajectory loop L1 reaches two slightly distinct conformations, highlighted in **Figure 4.5b**. The first conformation reflects a more closed arrangement, in which it could be expected that the heme could not bound. The second corresponds to a slightly more open disposition, but it never reaches the open conformation observed in the apo structure of HasA<sub>sm</sub>. To investigate the factors that influence this system to remain in general in a close conformation, an analysis on the interactions during GaMD simulations was performed. The aim of this analysis is to find which are the most relevant interactions that keep loop L1 in this “closed” arrangement. Hydrophobic interactions and hydrogen bonds were observed as preponderant during the whole simulation (>75%) in between the  $\alpha$ -helices and  $\beta$ -sheets, which keep the general structure invariable during GaMD simulation.

The heme binding regions is maintained in the same conformation due to a network of several  $\pi$ -contacts between aromatic residues of the pocket (Tyr55, Phe43, Phe83, Tyr75, Phe50, and Phe77) (**Figure 4.6a**). These interactions also contribute to keeping the stability of the overall protein structure in the absence of heme. Upon closer examination on loop L1, it was observed that several interactions could be responsible for maintaining it in the crystallographic arrangement and preventing it from reaching an open form. Throughout the

entire simulation, strong salt bridges and hydrogen bonds between residues Lys148 and Arg144 from helix- $\alpha$  A2 and residues Asp29 and 31 situated at the Nter of loop L1 were detected (**Figure 4.6a**). In addition, during 25–50% of the GaMD, hydrogen bonds between polar residues of loop L1 (Lys38, Arg40, Ser60), and the backbone or side chains of the same loop were found, which kept the small helix and loop conformation. Overall, the hairpin H3 exhibited highest flexibility due to the lack of strong interactions.



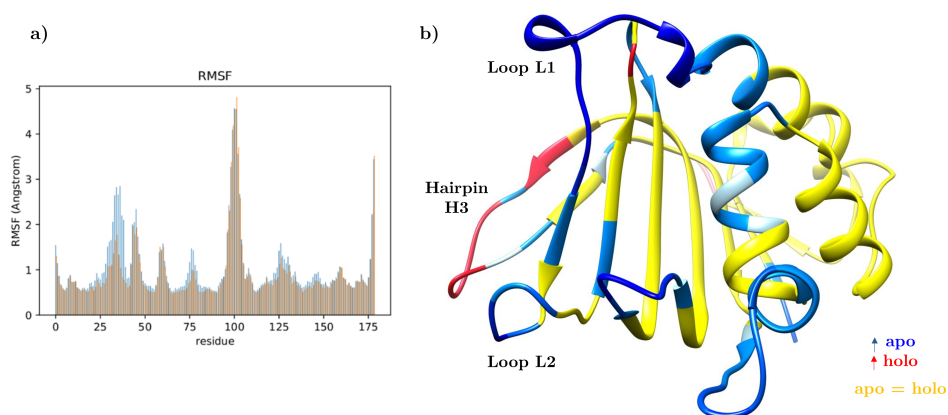
**Figure 4.6:** Interactions from HasA<sub>yp</sub> GaMD simulations. Hydrogen bonds in blue and hydrophobic contacts in orange. **a)** Interactions from GaMD apo. **b)** Interactions with heme from GaMD holo. Reprinted from [271].

To gain a deeper understanding into the heme binding process, protein–ligand docking of the heme cofactor was performed using the most representative protein conformations of the apo GaMD trajectory. Given the substantial resemblance between the apo form and the holo structures, it was not surprising that the dockings showed excellent scoring values (90 GoldScore units) for heme binding. The resulting complexes showed a structural arrangement similar to the experimental structure, featuring in all solutions the presence of a coordination bound between the metal and the Tyr75.

GaMD simulations with the heme bound were performed to validate the structural integrity and stability of the heme–HasA<sub>yp</sub> complexes predicted by the dockings. The hypothesis was that these complexes may differ from the X-ray structure and the aim of these simulations was to see how the presence of the heme affects the overall structure of the protein. Three replicas of GaMD

simulations starting from the heme-docking position were undertaken. No significant conformational changes were observed in the analysis of the GaMD trajectories of the holo forms of HasA<sub>yp</sub>. In the three replicas, the systems tend to reach convergence after 100 ns only. This is reflected in the stability of RMSD, cluster counting, and PCA analysis (**Figure B.2**). Neither the entire tertiary structure nor secondary motives, including loop L1, presented significant variations.

As in the apo GaMD, the core structure of the protein showed again low flexibility and high stability, differences could only be found in loops L1, L2 and helix  $\alpha$ -2. These appeared to have more flexibility in the apo than in the holo form, specifically loop L1, which showed the highest difference (average RMSF difference of 0.79 Å) (**Figure 4.7**). The only region that was slightly more flexible in the holo form was the hairpin H3 (average difference RMSF 0.13 Å). Therefore, it can be concluded that the presence of heme restricts the movement of loop L1.



**Figure 4.7:** RMSF difference between apo and holo from GaMD of HasA<sub>yp</sub>. Represented in blue are regions with RMSF higher in the apo form, and in red, regions with RMSF higher in the holo form. No significant RMSF differences in yellow. Reprinted from [271].

When comparing the network of interactions of the amino acids between the holo and apo forms, it was observed that most hydrogen bond interactions are preserved in both systems. In general, the holo form shows less interactions than the apo within the loop L1 and between the loop and the rest of protein, as these residues are now interacting with the heme. Despite the absence of direct coordination between heme and any residue of loop L1, strong interactions appear between the prosthetic group and the loop, reducing its flexibility.

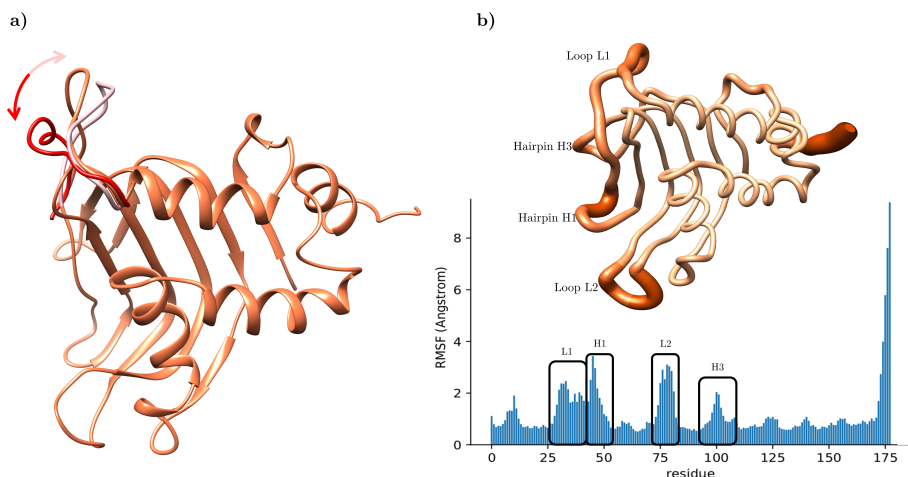
The heme moiety forms crucial hydrogen bond and salt bridge interactions between the propionates and residues Arg144, Arg40, and His135 (**Figure 4.6b**). These interactions are maintained throughout most of the simulations. Therefore, they may be responsible for maintaining loop L1 in a closed disposition even in the presence of heme bound. In addition, the hydrophobic interactions involving Phe43, Phe83, and Tyr55 observed in the apo system now are directly interacting with the heme instead of between themselves. Overall, the calculations revealed that the hydrophobic residues Tyr and Phe43,83,50,77 within the binding site serve as a platform for the binding of heme. On the other hand, positively charged residues Arg144,40 and His135 stabilize heme through salt bridges and also force the propionates to face the solvent.

#### 4.1.5. Results of Hemophore HasA<sub>sm</sub>

As introduced before, the apo and the holo form of HasA<sub>sm</sub> present different conformations regarding loop L1, the opposite case of HasA<sub>yp</sub>. The apo form of HasA<sub>sm</sub> presents an open conformation, while the heme-bound form adopts a close conformation (**Figure 4.3a**). The objective of this part of the study was to investigate the nature of the interactions occurring in both the holo and apo forms of HasA<sub>sm</sub> and understanding the differences for heme uptake with respect to the HasA<sub>yp</sub> system. Specifically, to see if loop L1 in the open apo form can shift towards the close conformation or if the binding of heme induces the conformational change of loop L1. Following the same protocol, GaMD simulations were carried out on the apo form of HasA<sub>sm</sub>. His32, which is the heme axial ligand, could be in different protonation states and influence the flexibility of loop L1. Therefore, simulations were performed with different protonation states of this residue, but no significant differences were observed. In this work, the system with His32 with monoprotonation at N $\delta$  will be described.

The analysis of PCA and clustering counting showed that the GaMDs reached convergence by the end of the 800ns simulations (**Figure B.3**). The  $\alpha$ -helix or  $\beta$ -sheets of the system did not show any major conformational changes. However, some variations were observed in the loop and hairpin regions (**Figure 4.8a**). The RMSF analysis showed the highest values for loop L1, L2 and hairpin H1, H3 (**Figure 4.8b**). Despite the considerable flexibility of loop L1 (average RMSF about 1.84 Å), no major movement of loop L1 was observed in any of the replicas. Throughout the entire simulations, loop L1 always remained close to

the initial open conformation without transitioning toward more closed conformations. Therefore, nothing in the three converged replica simulations of 800ns suggests that the apo form of HasA<sub>sm</sub> could naturally move toward a more closed conformation without the presence of the heme cofactor.



**Figure 4.8:** GaMD simulation from apo HasA<sub>sm</sub> **a)** Main cluster and two extreme positions of loop L1. **b)** RMSF extracted from GaMD with most relevant regions. Reprinted from [271].

Analysis of the interactions during the GaMD simulations was performed to better understand why L1 is not able to escape its open conformation and to reach the close heme-bound like states (**Figure 4.9a**). It seems that a network of hydrogen bonds between loop L1 and three other regions contribute on keeping the loop L1 fixed. The N-terminal part of loop L1 is fixed by interactions between Val30, Asn36, and Thr38 with hairpin H2 (Asn62 and Gln63). The central residues, Thr38 and Ser39, mainly interact with  $\beta$ -sheets B5-6 (Ser99 or Gln109), ensuring that the most flexible part of the loop is fixed. The C-terminal part presents interactions of the backbone 40–44 with  $\beta$ -sheet B3 and the side chains of Ser42 with Ser58–59. All these interactions seem to prevent loop L1 from transitioning into heme-bound conformations and keep it fixed to the initial close position. This is reinforced by a network of hydrogen bonds in the heme binding regions involving His83, 128, and 133.

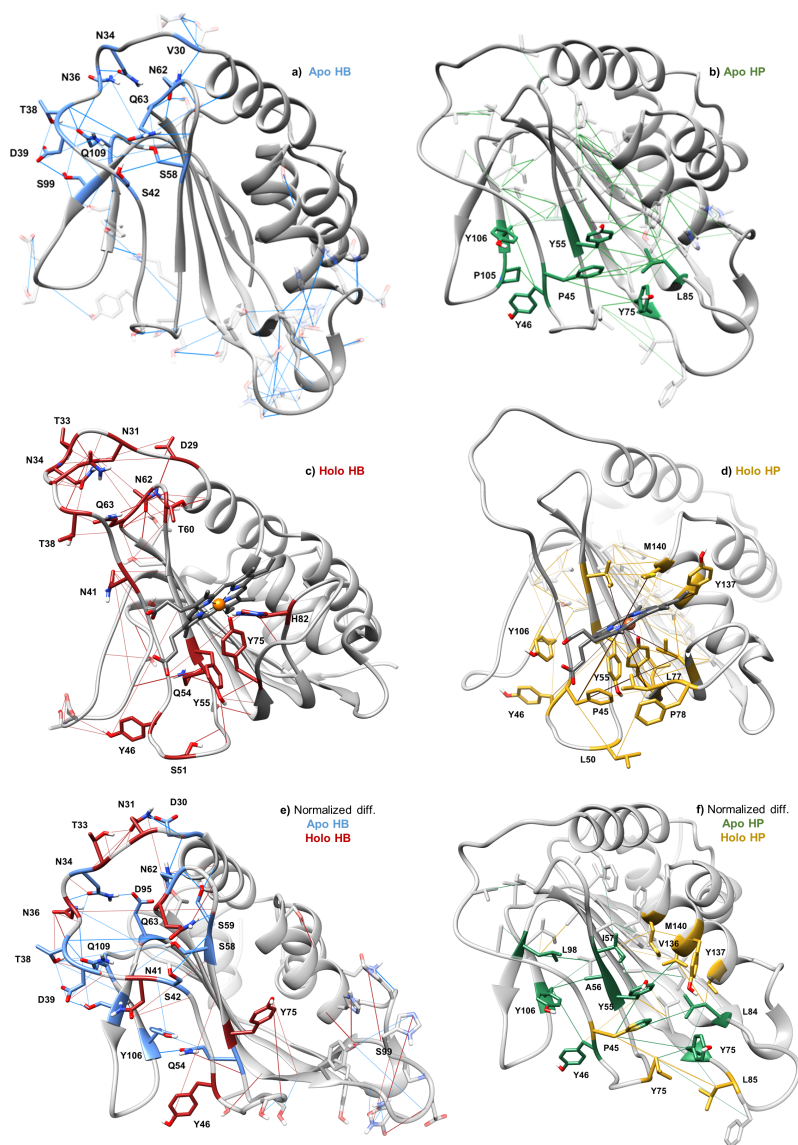
In terms of the hydrophobic interactions, hairpin H3 is involved in  $\pi$ -stacking interactions that stabilize close contacts with  $\beta$ -sheets B2–3 and loop L1. Key residues in these regions are Tyr46 and Ala56 from  $\beta$ -sheet B2, which interact

with hairpin H3 or  $\beta$ -sheets B5-6 with Pro105, Tyr106, and Leu98. However, it is important to highlight that the most relevant hydrophobic interactions are not related directly to loop L1 but rather to the heme binding region. There is a very robust network of several  $\pi$ -stacking contacts between the aromatic residues Phe45, Tyr75, Leu85, and Tyr55 (**Figure 4.9b**).

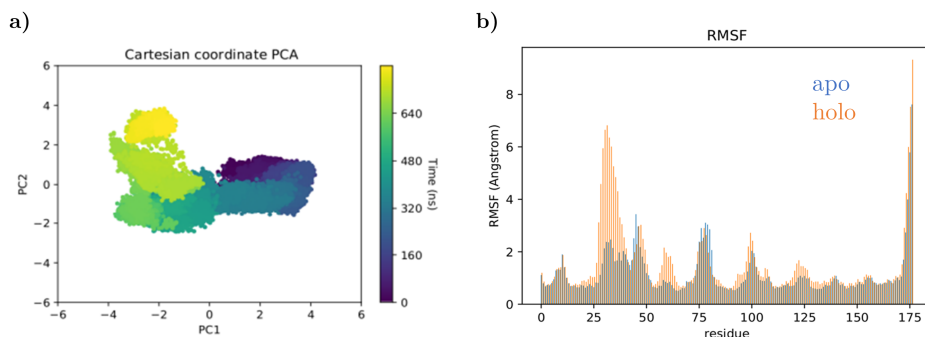
As the GaMD simulations of the apo-HasA<sub>sm</sub> did not show any conformational changes of loop L1 consistent with the heme-bound structure, the most plausible explanation would be that the conformational change is induced by the binding of the heme. Therefore, heme dockings followed by GaMD simulations with heme bound were performed.

Similar to HasA<sub>yp</sub>, dockings predicted good binding affinities (GoldScore values of 68) and similar binding poses as observed in the X-ray structure of the heme-bound form, with a coordination bond between the heme and the oxygen of the side chain of Tyr75. These findings suggest that binding of the heme to HasA<sub>sm</sub> could perfectly happen without requiring loop L1 to transit to the X-ray position and the His32 to coordinate the remaining axial site. Interestingly, the propionate groups from heme were facing the inner part of the protein in the best docking solutions. However, after 10-20ns of a classical MD simulation, the heme rotated toward the external part of the binding pocket and ended with a structure with excellent matching with the experimental holo form. These simulations were followed by three replicas of 800ns GaMD simulations.

The statistical analysis of the GaMD simulations (clustering, PCA, RMSD) revealed that convergence is reached after 100–400ns depending on the replica (**Figure B.4**). The system visited distinct conformations according to PCA analysis as depicted by the presence of several wells (**Figure 4.10a**). On the other hand, the RMSF analysis showed that the loops and hairpin regions were the most flexible part of the protein, specifically, it is in fact loop L1 the most flexible one. When comparing the RMSF with the apo form, once again the rigid core structure displayed very low flexibility and no changes in the tertiary structure were observed, except for loop L1 (RMSF difference of 4.1 Å) and the  $\beta$ -sheets B3 (RMSF difference 0.98 Å) (**Figure 4.10b**). In general, the apo form has less flexibility, except for loop L2, which is more flexible in the apo form.



**Figure 4.9:** Representation of hydrogen bonds (HB) and hydrophobic (HP) interactions during GaMD of HasA<sub>sm</sub> in apo (a,b) and heme-bound before loop closing (c,d). Normalized difference between apo and holo forms of both hydrogen bonds (e) and hydrophobic interactions (f) is represented. Reprinted from [271].

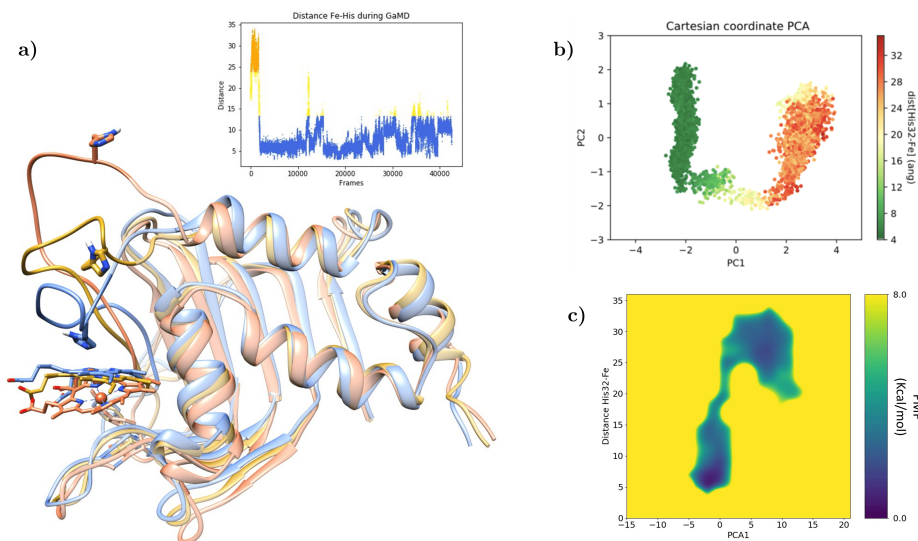


**Figure 4.10:** a) PCA of replica from holo GaMD of HasA<sub>sm</sub>. b) RMSF difference between apo in blue and holo in orange extracted from apo and holo GaMD of HasA<sub>sm</sub>. Reprinted from [271].

Upon visual inspection of the trajectory of the three GaMD replicas, a mechanism for loop L1 closing was observed. Detailed analysis revealed that at the beginning of the GaMD simulation, loop L1 tends to acquire an open conformation, separating from the region of hairpin H3 and  $\beta$ -sheets B5–6. Depending on the replica simulation, from 40ns up to 700ns, the loop adopts a turn or small helix conformation and moves toward the heme binding site, acquiring a more closed disposition (**Figure 4.11a**). This result support the notion that the binding of heme at its pocket, which is at the opposite site of loop L1, induces a conformational change at loop L1, initiating its closing mechanism.

In all GaMDs, it appears that once the heme has bound to the apo form, the system did not maintain a stable holo structure with His32 as the 6th ligand of the iron. Loop L1 oscillated between different states, at some points His32 was facing the heme at distances consistent with coordination to the metal, while in other instances His32 faces outside of the binding site preventing the coordination with the metal. The distance of the coordinating nitrogen of His32 to the iron fluctuates between 3 and 12 Å depending on the replica. This behavior arises from the fact that the force field for the metal is bonded and only allows one coordination state at a time. Since these simulations start with only Tyr75 bound and the corresponding parameters of a pentacoordinated first coordination sphere, no coordination with His32 is contemplated. Some Fe(III) simulations show shorter distances than in Fe(II) calculations, but the absence of an explicit Fe–His bond in the parameterization is the origin of the fluctuation. Additional GaMD simulations were performed to investigate further this point.

On one side, the holo X-ray structure (1DKH)<sup>270</sup> was subjected to GaMD simulations, with the same parameters, Fe–Tyr coordination and no coordination term for the Fe–His bond. In these simulations the His32 separates from heme binding pocket and fluctuates around 7.43 Å (going from 2.7 to 17.9 Å). However, loop L1 remains mostly in a closed geometry (**Figure 4.12**). Furthermore, GaMD simulations starting from the snapshot with the heme bound to Tyr75 with the smallest Fe–N<sub>His</sub> distance (2.9 Å) were performed. For these simulations parameters for the hexacoordinated metal with axial His–Fe–Tyr configuration were applied. In this case, loop L1 quickly reaches a conformation close to the experimental holo form (RMSD of 0.78 Å), and a stable His–Fe bond is observed during the simulation. This data emphasizes that reaching the final hexacoordinated structure requires the His coordination to be properly modeled.

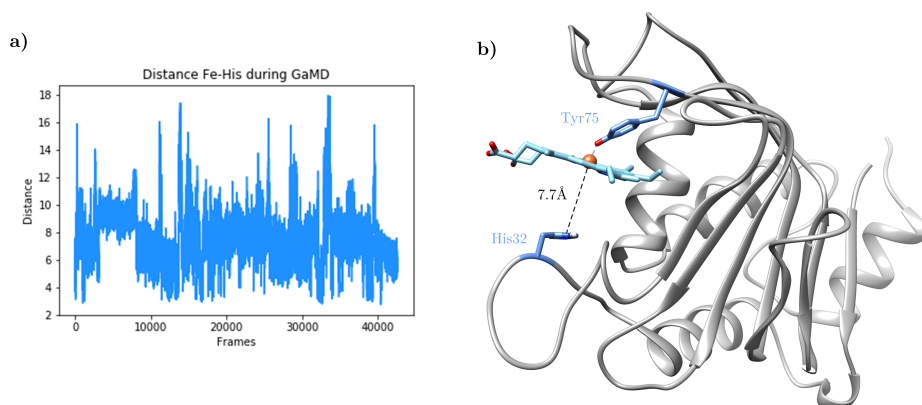


**Figure 4.11:** GaMD simulation of HasA<sub>sm</sub> with heme–Fe(III) bound. **(a)** Frames of GaMD showing the loop L1 closing process colored according to the distance between Fe and His32 during GaMD. Graphic of distance Fe–His32 represented. **(b)** PCA analysis colored according to the distance between Fe and His32 during GaMD. **(c)** Reweighted PMF calculations in front of PCA1 and distance Fe–His32. **(b,c)** are obtained using the fragment of the trajectory in which the loop is closing. Reprinted from [271].

The fluctuations of the protein following heme binding is the major determinant in the loop L1 closure motion. PCA analysis of fragment trajectory where loop L1 closes with respect to the Fe–His32 distance reveals how PC1 involves the movement of loop L1 (**Figure 4.11b**). Furthermore, reweighted PMF calculations based on the PCA1 distance Fe–His32 show that the barrier of the system

between the two different states is less than 8 kcal/mol (**Figure 4.11c** and **B.5**). This relatively low energetic barriers suggest that the transition between the open and close states is energetically feasible.

Examination of the interactions reveals the changes caused by the binding of the heme in the network of hydrophobic and stacking interactions (**Figure 4.10c,d**). Essentially, the binding of heme disrupts all the previously mentioned  $\pi$ -stacking and hydrophobic interactions within aromatic residues of the heme binding site (Tyr75, Tyr137, His83, Tyr55, and Phe45). These residues instead of interacting between themselves, now interact with the heme molecule itself. Moreover, the interaction between Tyr46 and Tyr106 weakens, resulting in an increased flexibility of hairpin H3. Consequently, hairpin H3 separates from  $\beta$ -sheets B2–3 and from loop L1, which in fact impacts loop L1 and its hydrogen bond interactions. In the initial stages of the simulations, all hydrogen bonds between L1 and hairpin H3 drastically weaken, making it more flexible. Instead, loop L1 establishes more hydrogen bonds with  $\beta$ -sheets B5–6 and hairpin H2, mainly with Thr60, Asn62, and Gln63.



**Figure 4.12:** GaMD simulation of X-ray holo (1dkh) from HasA<sub>sm</sub> with heme-Fe(III) bound and only Tyr coordination **a)** Distance between Fe and His32 during GaMD **b)** Most representative cluster of GaMD simulation with distance between Fe and His32 displayed. Reprinted from [271].

As previously mentioned, the binding of the heme decreases the  $\pi$ -stacking interactions in the heme binding site. For instance, interaction between Phe45 from loop L1 and the residues of the heme binding site decrease, consequently, Phe45 interacts with residues Leu50 and Leu77 from hairpin H1. This change of interactions triggers conformational changes in the region of hairpin H1 and

C-terminal of loop L1, potentially resulting in the formation of a turn or semi-helix in disposition on loop L1. As a results of this, the previously mentioned interactions with  $\beta$ -sheet B3 and H2 are replaced by a series of hydrogen bonds within the loop L1 itself. Consequently, this causes the separation of loop L1 from  $\beta$ -sheets B5-6, and loop L1 starts to move toward the heme binding site. Simultaneously,  $\alpha$ -helix A1 moves closer toward  $\alpha$ -helix A2. Once loop L1 adopts a closed disposition, a series of hydrogen bond interactions between Asn31 and Val30 from loop L1 and Ser141 from  $\alpha$ -helix A2 maintain loop L1 in a close conformation. More clear tendencies can be observed when comparing and normalizing the frequency of these interactions with the apo form (Figure 4.10e,f).

#### 4.1.6. Conclusions: comparison of both systems

In this study, the combination of GaMD simulations and docking highlights differing patterns of heme binding mechanisms between the hemophores HasA from *Y. pestis* and *S. marcescens*. Simulations on the HasA<sub>yp</sub> clearly demonstrate that the apo form remains in a geometry like that of the holo form. The transition between a close and open conformation resulting from the movement of loop L1, as in HasA<sub>sm</sub>, is never observed despite GaMD simulations allowing extensive conformational sampling. Calculations show that this is due to specific salt bridges and hydrogen bonds between negatively charged residues of loop L1 (Asp30 and Asp31) and positive residues from helix from the core (Lys148 and Arg144). It can be concluded that the mechanism of heme binding would correspond to a very light conformational selection as the binding of the heme only slightly impacts hydrophobic interactions, but not sufficiently to disrupt the geometry of loop L1.

Simulations performed on HasA<sub>sm</sub> show that both apo and holo structures exhibit remarkable stability and no structural rearrangement of loop L1 is observed in either case. Apo simulations demonstrate that the binding site of the heme is well pre-organized and that binding should occur naturally. Despite the open conformation of loop L1 and no coordination with His32, heme successfully binds to its cavity and coordinates with Tyr75 from loop L2. A crucial finding of this study is that the movements of loop L1 are only observed after the heme process occurs. Analysis of contacts shows that the binding of heme disrupts the network of hydrophobic and  $\pi$ -stacking interactions at the heme binding site and that this

information is propagated to the other extreme of the protein. As a consequence hairpin H3 acquires more flexibility and loses interactions with residues of loop L1, leading to a disruption of the interactions between the C-terminal part of loop L1. Both hydrogen bonds and hydrophobic contacts with hairpin H2 and  $\beta$ -sheets B5–6 are broken, which causes loop L1 to be more flexible and move up to the heme binding site.

From this work it can be concluded that in both cases, the apo forms are very stable and well pre-organized for the binding of the heme. In both cases, the binding of the heme occurs without any previous major conformational changes including the loop transition. These findings suggest that the binding of heme can be predicted in the apo forms of heme proteins. The main differences between both species come from the cascade of molecular events happening after the heme binds and the conformational changes associated to them. While very little perturbation of the overall map of contacts and interactions happens in *Y. pestis*, a series of changes take place in *S. marcescens*, which cause the closing of loop L1.

Compared to previous studies, here we can simulate the transition from apo to holo without forcing or constraining the system. Long GaMD simulations of 800 ns simulations have been performed, which have allowed all systems to converge and observe significant conformational changes. Furthermore, performing three replicas for each case assured that these events are not casual. This study completes our knowledge on heme binding complexity and how long-range interactions could be crucial for defining the heme-bound geometry and the mechanism of acquisition.

## 4.2. Development of software for the identification of heme binding sites

The prediction and identification of heme binding sites plays a crucial role in understanding better natural heme binding processes, but also provides insight in the design of ArM based on heme. Around 2010, several research groups started to dig into heme-protein interactions. Schneider *et al* identified the main structural motifs and interactions between heme and proteins.<sup>289</sup> Smith *et al* performed an in-depth study of binding sites of 34 heme-associated binding proteins belonging to different families and reported how specific structural characteristics of heme binding sites lead to different heme functions.<sup>97</sup> In 2010 Li *et al* conducted research focused on studying different aspects on heme binding site properties. The binding site environment of a non-redundant set of 125 heme binding was analyzed and revealed that heme binding pockets contain mainly residues with aromatic and non-polar properties. Interestingly, a visual comparison of ten apo-holo heme binding structures revealed that in 90% of the cases proteins suffer very small conformational changes upon heme binding, suggesting that in most cases, apo form can be used to predict heme binding.<sup>96</sup>

Based on these works and as experimental determination is time-consuming and expensive, development of computational methods for predicting heme binding sites have emerged over the past years. HemeBIND was the first to predict heme binding residues integrating structural attributes like solvent accessibility, depth, and protrusion with sequence information like residue evolutionary conservation. This software is based on the supervised machine learning method Support Vector Machine (SVM) that was trained on a dataset of 141 non-redundant proteins.<sup>207</sup> The same group developed a second version, HemeNET, which uses residue interaction network as input. The prediction of heme binding residues was improved by combining topological features of the residue network like degree or closeness with structure/sequence information.<sup>290</sup> HEMEsPred also combines structural and sequence information to predict specific heme B or heme C binding sites using an adaptive ensemble learning method to enhance the prediction.<sup>291</sup>

On the other hand, there are heme predictors based purely on sequence information, as in the case of SCMHP<sup>206</sup> or Xiong *et al* predictor.<sup>292</sup> The former exclusively relies on sequence information, while the latter combines sequence evolutionary and physicochemical properties. The most recent heme predictor algorithm SeqD-HBM is based on experimental analysis of heme-peptides that lead to a set of sequence features to heme predicting motif.<sup>259</sup> This algorithm is now implemented on HeMoQuest, a predictor that is aimed at transient heme binding sites.<sup>293</sup>

So far, heme predictors are sequence based or hybrid structure / sequence methods. Although these approaches are particularly relevant for natural enzymes, they can be a limitation for *de novo* systems which have not yet evolved for heme binding processes. That would be the case of heme-based artificial metalloenzymes. The aim of this work is precisely to provide a new piece of software for detecting heme binding sites based exclusively on the protein structure and the geometrical predisposition of heme binding sites. Furthermore, due to the principles of this software, a novel functionality is implemented, designing new heme binding site for ArM.

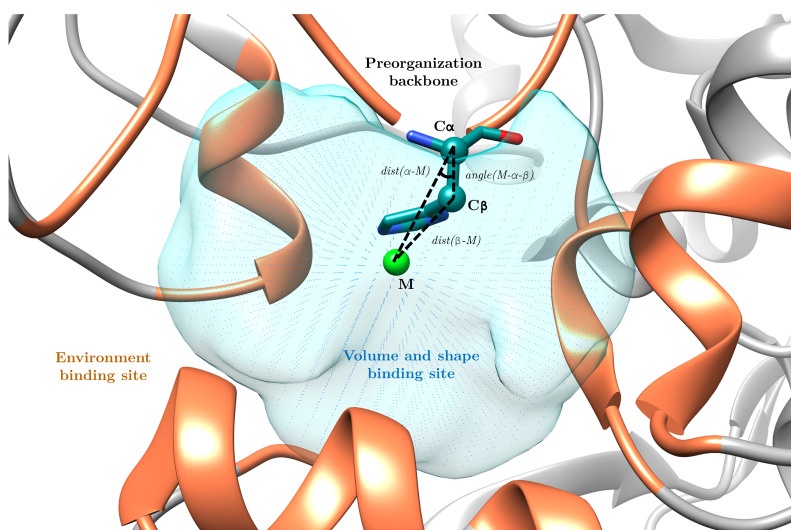
#### 4.2.1. Conceptualization of the software

This program was conceptualized from the idea of BioMetAll<sup>294</sup>, which is a predictor of metal binding sites based only on the preorganization of the backbone. It uses three geometrical criteria from the backbone to make the predictions: the distance from the metal to the  $\alpha$ -carbon  $\text{dist}(\alpha\text{-M})$ , the distance between the metal and the  $\beta$ -carbon  $\text{dist}(\beta\text{-M})$  and the angle between the metal, the  $\alpha$ -carbon, and the  $\beta$ -carbon  $\text{angle}(\alpha\text{-}\beta\text{-M})$ . Despite the success of BioMetAll in predicting metal binding sites, it is not entirely suitable for detecting heme binding sites because it does not consider the real dimension of the prosthetic group including the volume of the binding site, or the properties of the surrounding residues. Moreover, BioMetAll does not account for the possibility of axial coordination by either one or two amino acids.

Here we present a new program called HemeFinder that has the central philosophy of BioMetAll, but includes additional descriptors of heme binding (**Figure 4.13**). HemeFinder aims to predict heme binding sites considering only structural information in a fast approach relying on two assumptions:

1. All heme binding sites shared common geometrical properties for a well pre-organized binding site.
2. The coordination of the metal by axial amino acids is determined by the preorganization of the backbone of the same coordinating residues.

HemeFinder workflow is based on first detecting cavities within the protein, calculating its volume and storing them as grid points (probes) using a pyKVFinder.<sup>295</sup> Then, for each of the probes of the cavity it is determined which surrounding residues fulfill these three geometric criteria ( $\text{dist}(\alpha\text{-M})$ ,  $\text{dist}(\beta\text{-M})$  and  $\text{angle}(\alpha\text{-}\beta\text{-M})$ ) and could coordinate heme. At this point, feasible mutations can also be suggested. After calculating the centroid of all the coordinating probes, an ellipsoid is defined with all the probes at certain distance. The size of the ellipsoid and the properties of surrounding residues are calculated to assess the suitability of the binding site to bind a heme moiety.



**Figure 4.13:** Basis of HemeFinder prediction.

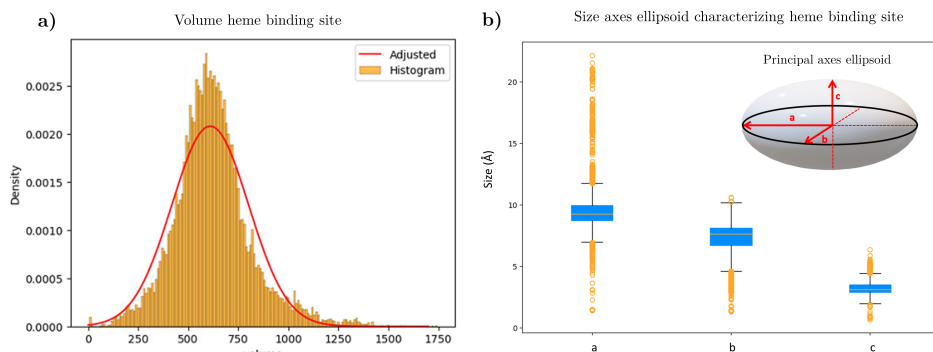
### 4.2.2. Statistical analysis

To establish the foundation of HemeFinder, an initial statistical analysis of all heme-containing protein structures from MetalPDB was conducted. It served to extract the principal properties associated with heme binding sites. The analysis focused on three main properties: volume and shape of the binding site and the types of residues surrounding the heme binding site. Additionally, the three geometrical descriptors from the backbone that define coordination and the proportion of Fe-coordinating residues were analyzed.

In 2022, MetalPDB reported a total of 5.164 PDB entries containing a total 13.133 heme binding sites. To perform the statistical analysis, 90% of these entries, were selected randomly using function *train\_test\_split* from *sklearn* and the other 10% of the entries were saved for the benchmark. However, due to poor annotations of the coordinating residues or heme name in MetalPDB or inconsistencies in PDB structures, the final number of entries analyzed was lower. Consequently, 3.812 PDB structures (9.229 heme sites) were used for the analysis of geometrical descriptors and the analysis of the properties of binding site was performed on 4.570 PDB structures (11.490 heme sites).

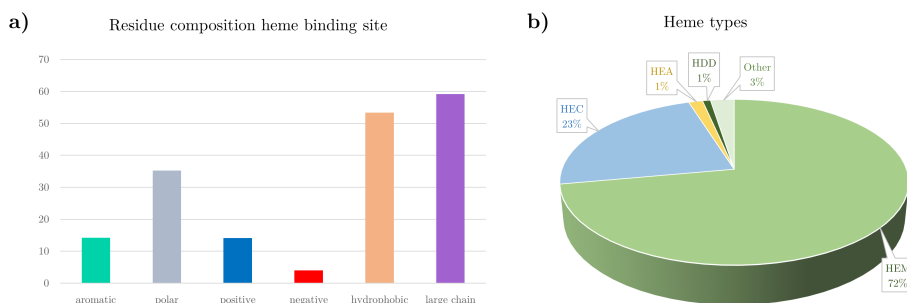
For the analysis of the main physical properties of the binding site, all PDB were downloaded, and each heme moiety was removed from the binding site. This analysis was performed using an in-house script including PyKVFinder<sup>295</sup> module to define the heme cavity, obtain the volume and save all cavities as a grid of probes. From the cavity files, the shape of the binding site was characterized using a series of functions as an ellipsoid with its principal axes. The environment of the binding site was defined by the different types of residues, all residues that were considered should be at a distance lower than 6.5 Å from carbon  $\alpha$  and 5 Å from carbon  $\beta$  from any point of the cavity. This analysis revealed how the distribution of the volume of an average heme binding follows a gaussian distribution with a mean of  $612.58 \pm 194.39$  Å<sup>3</sup> (Figure 4.14a). The shape of the principal axes of the ellipsoid shape are  $a = 9.26 \pm 1.5$  Å,  $b = 7.64 \pm 1.2$  Å and  $c = 3.12 \pm 0.52$  Å (Figure 4.14b).

When examining the residue environment of a heme binding site, the composition that was obtained is represented in Figure 4.15a. Most residues found in heme binding site are hydrophobic (57.05%) and contain large chains (60.16%), being Leu and Phe the most common residues. Additionally, polar



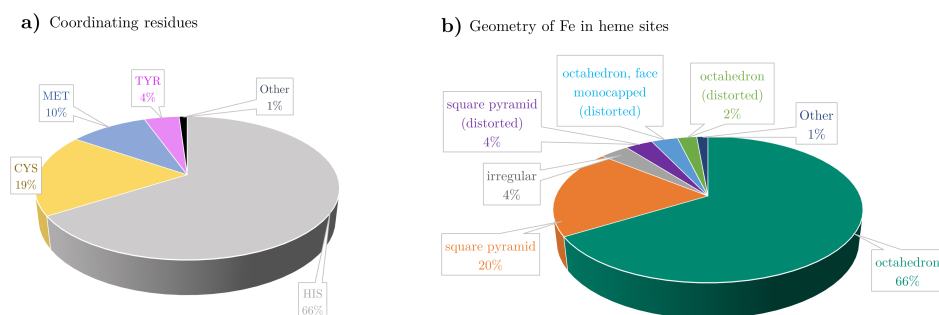
**Figure 4.14:** Statistical analysis of characteristics of heme binding site: **a)** Volume distribution with Gaussian fitting in red **b)** Sizes of three principal ellipsoid axes.

aromatic and positive residues are also quite common, 36.38% and 14.22% respectively. The presence of large hydrophobic residues can be explained by the apolar and aromatic character of the heme. The polar and negative nature of heme propionates can be related to the presence of polar and positive residues, especially Arg. Negative residues are almost not present in heme binding site (4%). It is important to emphasize that the classification of residues around the heme binding site is not exclusive, as some residues may belong to more than one category due to its properties (classification in **Figure B.6**). Regarding the types of heme present in nature, most heme binding sites correspond to heme B (23.72%) followed by heme C (22.96%) displayed in **Figure 4.15b**.



**Figure 4.15:** Composition of **a)** different types of residues in environment of heme binding site **b)** Most common types of heme.

A pychimera<sup>296</sup> script was used to go through all the entries of the XML file from MetalPDB and calculate the three geometric descriptors and the prevalence of each coordinating residue. Results of this analysis show that the average coordination number to the protein is 1.35 and the coordinating residues are His (65.95%), Cys (18.65%), Met (9.94%) and Tyr (4.43%), which represent the 99% of the whole database (**Figure 4.16a**). In consequence, as default, HemeFinder considers these four main residues as coordinating and coordination number as one. This analysis only accounts for protein residues, ligands are not accounted. If ligands are accounted, the predominant geometry of Fe is octahedral, followed by square pyramid (**Figure 4.16b**).

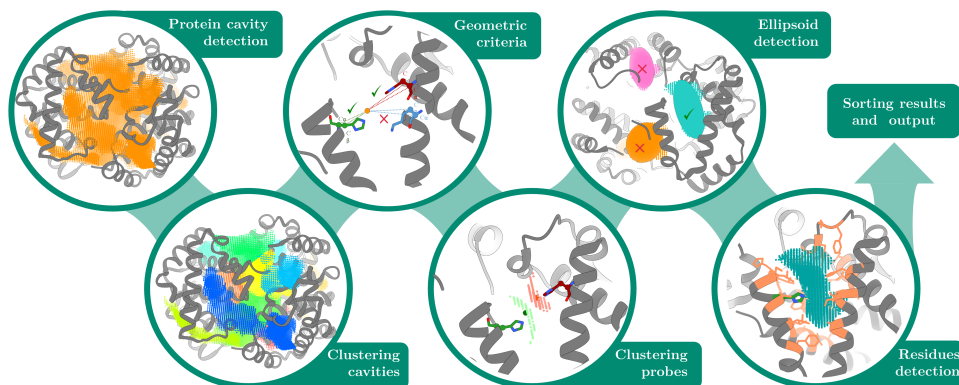


**Figure 4.16:** Percentage of **a)** Coordinating residues and **b)** Fe geometries.

All three geometric features are represented in **Figure B.7**. Distributions depend on the residue, but in all cases, individual residue distributions can be fitted to a bimodal function. An additional analysis was carried out for entries with two coordinating residues. Four descriptors were analyzed: the distances between  $\alpha$ -carbons/ $\beta$ -carbons of both coordinating residues, the angle between  $\alpha$ -carbons/ $\beta$ -carbons and the metal of both coordinating residues. Results reveal that distances follow a gaussian distribution (**Figure B.8a**), whereas angles follow a bimodal distribution. The most common type of coordination is His-His and His-Met (**Figure B.8b**).

### 4.2.3. Workflow of software

With all the descriptors fitted from the previous steps, Hemefinder was set up. It was entirely developed in Python3, containing as few dependencies as possible. The overall workflow of HemeFinder is divided into six different steps that are summarized graphically in **Figure 4.17**.



**Figure 4.17:** Main six steps of HemeFinder workflow.

**Protein cavity detection:** Input protein file must be provided in pdb format, or PDB ID can be specified, and its structure is directly downloaded from PDB database. The protein is parsed using pyKVFinder module *read\_pdb* and then all residues are parsed only to retain coordinates and residue names of C $\alpha$  and C $\beta$ . Protein cavities are detected using module *detect* from pyKVFinder and the volume is extracted with the *spatial* module. Default parameters for pyKVFinder calculations are established, but the user can modify the most decisive parameters, like grid spacing of probes or the radius of the probes from pyKVFinder that detect the cavity. Cavities smaller than the lower threshold of 2.5 standard deviation (SD) of our statistics study (130<sup>3</sup>) are discarded as they would not be able to fit a heme molecule.

**Clustering cavities:** Cavities that have a volume higher than the upper threshold of 2.5SD of our statistics study (1089<sup>3</sup>) are split into smaller cavities. In big heme binding proteins, cavities are connected by tunnels or may contain more than one heme moiety. Consequently, when the volume is higher than the threshold, clustering is performed to obtain the cavities of the mean size of heme. Kmeans<sup>297</sup> is used to perform the clustering with the number of clusters set to

the proportion between calculated volume of cavity and the statistical median value of a heme cavity. Each cavity is saved as a grid of probes separated by a grid spacing of 0.6 Å as default and they are saved in XYZ format. In cases in which there are two coordinating residues, the distance between them is very small and pyKVFinder leaves an empty hole where the iron is found. To avoid this problem, a function for detecting these regions based on the density of probes is applied and the mentioned holes are filled with probes.

**Geometric criteria:** For each probe of each cavity, the three geometric criteria are checked. For each probe the software checks that  $\text{dist}(\alpha\text{-M})$ ,  $\text{dist}(\beta\text{-M})$  and  $\text{angle}(\alpha\text{-}\beta\text{-M})$  are within the range that is established by our statistic study for all residues that could be coordinating. For the probes that fulfill all three criteria a scoring is assigned depending on how good the possible coordination is (explained in more details in following section). The output of this part is for each probe, the coordinates of the probe, the residue number of the possible coordinating residues and its score. At this point, if there are none possible coordinating residues in the protein, the user can also request for mutations, as the geometric criteria only depends on the backbone.

**Clustering of probes:** Once each probe has a score assigned and the possible coordinating residues, all the probes that have the same possible coordinating residues are grouped together and the scores are added. This provides a list with the possible coordinating residues associated with all the coordinating probes and a weighted geometric centroid with the scores.

**Detection of the ellipsoid:** From the centroid of the coordinating probes, all the probes around a sphere with size 7.63 Å (mean size of second ellipsoid axes from statistic study) are selected. The selected probes are modeled into an ellipsoid using a function, characterized by its three axes and center. If the three axes have a value lower than 2.5 standard deviation (SD) of the statistical study, this result is discarded. This means that in these cases there would be coordination, but the heme would not be able to bind in this binding site. For each ellipsoid, a score is assigned depending on the centrality of the centroid against the ellipsoid center (explained in scoring section).

**Residue environment:** To detect the residue composition of the cavity, a distance matrix is calculated between all the probes of the ellipsoid and the  $\alpha$ -carbon and  $\beta$ -carbon from all residues. Residues that its  $\alpha$ -carbon and  $\beta$ -carbon are within

a distance of 6.5 Å and 5.5 Å (respectively) from an ellipsoid probe are stored. The proportion of each type of residue is calculated and evaluated fitting it to the statistical values obtained (explained in scoring section). This is used to obtain a score that represents how adequate is the environment of the ellipsoid.

**Sorting and scoring:** Results from different clusters are integrated together. For results where there are two possible coordinating residues, an additional geometrical check is performed. The distances and angle between the two  $\alpha$ -carbon and the angle between the  $\beta$ -carbons and the centroid of the coordinating probes needs to be lower than the values obtained at the statistical analysis. If the values are higher, this result is discarded as the coordinating residues are also separated for coordination. Finally, each of the three scores (coordination, ellipsoid and residues) is normalized and the final score is obtained as the sum of the three normalized scores, which can be found between 0 and 3, being 3 the highest score possible.

**Output:** The output of the HemeFinder is a pdb file that contains the centroid of the coordinating probes, all the probes that make up the ellipsoid and the coordinating probes (Figure 4.18). This file can be visualized with any visualization program and each result is represented by different atom types (Centroid = He, ellipsoid = Xe and coordinating probes = Ne). The software also prints all the results and exports them into a json file, which contains all the possible heme binding sites sorted by score and with its corresponding coordinating residues.

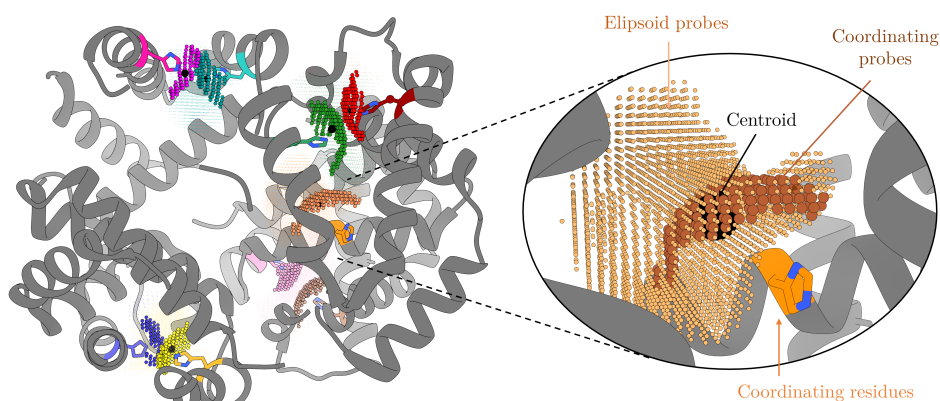


Figure 4.18: Example of output obtained by HemeFinder.

#### 4.2.4. Scoring

The scoring of HemeFinder aims to assess the suitability of a cavity for the binding heme considering three different items: geometrical criteria of backbone, residue composition of environment and metal centrality in ellipsoid. Geometrical criteria and residue composition scorings were developed considering the statistical study.

##### Geometrical criteria of backbone:

This fitness function measures how close the preorganization of a certain residues is to the ideal geometry required for binding of heme. To approximate these ideal values, the previous statistical analysis of three geometrical descriptors of coordinating residues:  $\text{dist}(\alpha\text{-M})$ ,  $\text{dist}(\beta\text{-M})$  and  $\text{angle}(\alpha\text{-}\beta\text{-M})$  was performed. The probability density function of the distribution of each parameter was approximated as a bimodal function, both for its mathematical simplicity and for the shape of the distributions observed. Thus, the likelihood that a given residue will be able to coordinate given the set of geometrical parameters  $x=\text{dist}(\alpha\text{-M})$ ,  $\text{dist}(\beta\text{-M})$ ,  $\text{angle}(\alpha\text{-}\beta\text{-M})$  is approximated as the union of the three distributions.

$$p(\text{coordination}|\text{residue}, x) = \frac{1}{N} \sum_{i=1}^{N=3} p_{\theta}(x_i|\text{residue}) \quad (1)$$

$$p_{\theta}(x_i|\text{residue}) = \lambda_1 \exp\left(-\frac{(d - \mu_1)^2}{2 \cdot \sigma_1^2}\right) + \lambda_2 \exp\left(-\frac{(d - \mu_2)^2}{2 \cdot \sigma_2^2}\right) \quad (2)$$

The score of the residue is defined by a bimodal distribution with a set of 6 parameters  $\theta = \lambda_1, \lambda_2, \mu_1, \mu_2, \sigma_1, \sigma_2$ .  $\mu_i$  are the centers of the distributions;  $\sigma_i$  the deviations; and  $\lambda_i$  a correction factor that accounts for the relative contributions of each component of the distribution. To find the most appropriate set of parameters  $\theta$ , the empirical data distributions were smoothen with a Kernel Density Estimation (KDE)<sup>298</sup> with gaussian kernel and then the curves  $p_{\theta}(x_i|\text{residue})$  were adjusted to the smooth distribution using the Levenberg-Marquadt algorithm<sup>299</sup> for non-linear least-squares adjustment.

As an example, in (Figure 4.19) are represented the distribution of the three geometrical descriptors of residue His, fitted to a bimodal function in orange. Distributions from Tyr, Lys, Met and Cys are represented in (Figure B.9).

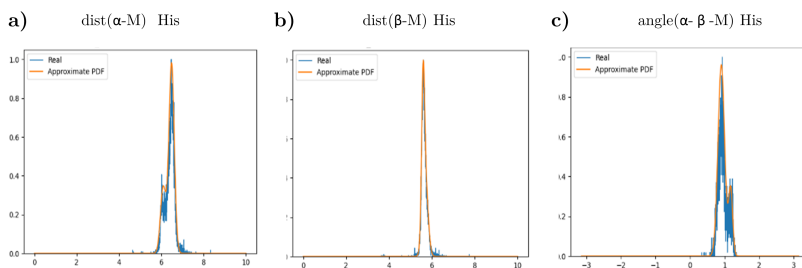


Figure 4.19: Bimodal distribution fitting for His. a)  $\text{dist}(\alpha\text{-M})$  b)  $\text{dist}(\beta\text{-M})$  c)  $\text{angle}(\alpha\text{-}\beta\text{-M})$ .

Furthermore, it is expected that some coordinating residues have more tendency to coordinate heme than others. Therefore, this tendency is modeled as the proportion of coordinations that each residue establishes with heme in the database ( $N_{\text{residue}}$ ) over the total number of coordinations documented for heme ( $N_{\text{heme}_{\text{coord}}}$ ). This allows coordinations involving a more common coordinating residue to be scored higher than coordinations involving fewer common residues. This is included as default, but can be excluded if the user desires.

$$p(\text{residue}) \simeq \frac{N_{\text{residue}}}{N_{\text{heme}_{\text{coord}}}} \quad (3)$$

Combining the two previous expressions, the final scoring function is obtained as following equation 4:

$$p(\text{coordination}) \simeq p(\text{coordination}|\text{residue}, x) \cdot p(\text{residue}) \quad (4)$$

### Residue composition:

This scoring function assesses the environment of the heme binding site. The statistical analysis of amino acid composition of heme binding sites was performed in order to determine what is the most suitable environment for stabilizing the heme group. The idea was to model the ideal environment composition. Residues were divided into 5 groups attending to their chemical nature as specified in (Figure B.6).

Then, the proportions of each group within each binding site in our training set was evaluated. The density function for each of the distributions generated was approximated with a bimodal function as in the geometric criteria. Thus, given a set  $c = \text{aromatic, polar, ...}$  of features of the chemical environment of a possible heme binding site. The probability of that environment being suitable for heme binding is approximated as:

$$p(c) = \sum_i^{N=5} \left( \lambda_1^i \exp \left( -\frac{c^i - \mu_1^i}{2 \cdot \sigma_{1,i}^2} \right) + \lambda_2^i \exp \left( -\frac{c^i - \mu_2^i}{2 \cdot \sigma_{2,i}^2} \right) \right) \quad (5)$$

where the individual score for each of chemical features  $i$  is defined by a bimodal distribution with 6 parameters  $\theta^i = \{\lambda_1^i, \lambda_2^i, \sigma_1^i, \sigma_2^i, \mu_1^i, \mu_2^i\}$ . These parameters are approximated as explained in the geometrical criteria. Distributions for each group of residues are represented in (Figure B.10).

#### Metal centrality in ellipsoid:

The aim of this score is to measure the centrality of the calculated centroid of the coordinating probes against the ellipsoid, which will give an indication of how deviated from the center of the ellipsoid is the centroid. This is modelled as the distance between the centroid of the coordinating probes ( $C_p$ ) and the center of the ellipsoid ( $C_e$ ) divided by the mean length of the three axes of the ellipsoid ( $a, b, c$ ).

$$\text{score ellipsoid} = \frac{\sqrt{(Xc_p - Xc_e)^2 + (Yc_p - Yc_e)^2 + (Zc_p - Zc_e)^2}}{\frac{a+b+c}{3}} \quad (6)$$

#### 4.2.5. Benchmark

To analyze how well is HemeFinder able to predict heme binding sites first we tested its performance by conducting two different benchmarks. Benchmark 1 contained the remaining 10% of PDB entries from the MetalPDB that were not used in the statistical study. Benchmark 2 includes all PDB entries that have been added to PDB over the course of 2023 and was cleaned manually to avoid redundancy. Both datasets were prepared by downloading all structures, removing all cofactors/solvent and selecting only monomeric structure. HemeFinder was applied to all systems with default parameters.

In **Table 4.2** are results for both benchmarks, with success rate referring to the % in which HemeFinder was able to detect at least one of the coordinating residue. In between brackets are results referring to the % for which HemeFinder was able to detect the exact solution. In both benchmarks HemeFinder can find the crystallographic heme in more than 90% of cases. In table 2 are also indicated the ranking of solutions, which % was found in the top 1, top 3, or top 10 solutions. For entries containing more than one binding site, the average was considered. In both benchmarks, solution can be found in the top 3 solutions in more than 80% of cases. In general, the average time per calculations was 73.63s per entry, 48.78s if an outlier (very large protein that took more than 3h) is not considered. The average rank position is between 2 and 3 in both benchmarks and in general between 10 and 20 solutions are obtain for each run.

	Success rate (%)	Top 1 (%)	Top 3 (%)	Top 10 (%)	Average rank	Average num. solutions
<b>Benchmark 1</b>	94.30 (91.35)	49.37 (39.03)	80.80 (73.21)	95.36 (90.93)	2.59 (3.47)	11.33
<b>Benchmark 2</b>	96.43 (91.07)	44.64 (32.14)	82.14 (67.86)	92.86 (89.29)	2.81 (5.69)	21.50

**Table 4.2:** Results obtained for Benchmark 1 and Benchmark 2.

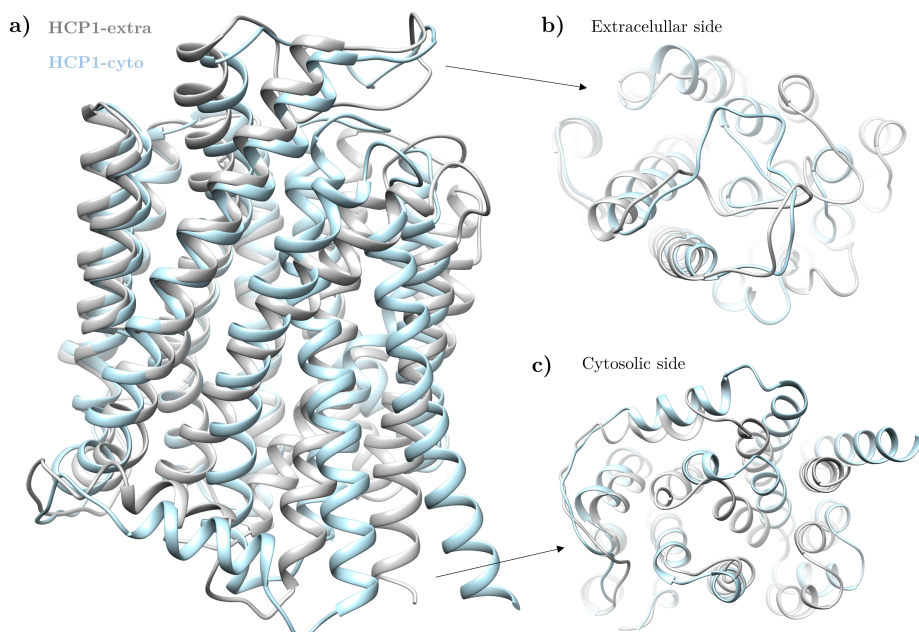
As the conceptualization of this software is different from the currently available ones, no comparison with others has been performed. Still, the success rate that we have shown is around current state-of-the-art programs. The idea of HemeFinder is not to only find heme binding sites, but to find additional heme binding sites that could make up possible heme pathways. It can also be used as a fast ‘docking’ approach to be performed on several structures to find new heme binding sites or design new ones. To show the uses of HemeFinder in the next sections we present two practical studies that show its potentiality.

#### 4.2.6. Case study 1: Detection of natural heme binding sites

The system of interest of this case study is Proton-Coupled Folate Transporter (PCFT) or Heme carrier protein 1 (HCP1). Initially, this transporter was identified as a low-affinity heme carrier found in the small intestine, specifically in the duodenum.<sup>300</sup> However, it is now established that this protein acts as a high-affinity proton-coupled folate transporter.<sup>301,302</sup> Specifically, it has a higher

affinity for folate ( $K_M = 1.67 \mu\text{M}$ ) when compared to heme ( $K_M = 125 \mu\text{M}$ ).<sup>303</sup> Currently, evidence suggest that PCFT/HCP1 acts primarily as a folate transporter, but it is also involved in heme transport. Still, more research has to be performed to understand its mechanism of heme uptake. The main missing piece of information is where and how the heme binds to the receptor.<sup>304,305</sup>

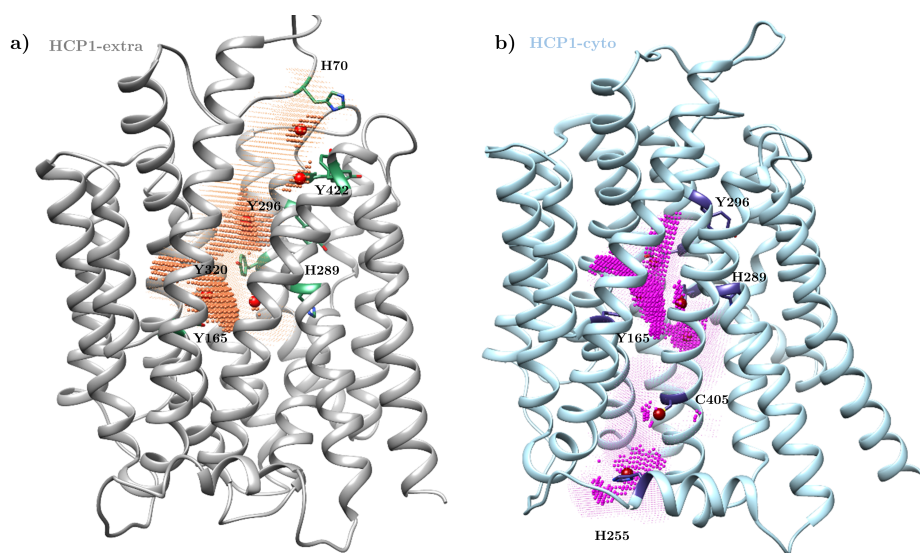
So far one structure has been determined for HCP1, which was obtained by cryo-EM from *Gallus gallus* (PDB 7BC7)<sup>306</sup>. This structure is in its apo form and in open conformation towards the extracellular site of the transporter (HCP1-extra). Additionally, an AlphaFold structure is available with open conformation towards the cytosolic site (HCP1-cyto). Comparison of both structures is depicted in **Figure 4.20**. The objective of this case study is to use HemeFinder to determinate where heme would bind, which would be the path of heme entrance in both protein conformations and with which residues would heme coordinate. The protein structures were downloaded from PDB and AlphaFold and prepared to contain only HCP1 structure. HemeFinder was applied at default, all four possible coordinating residues and a minimum of one coordinating residue.



**Figure 4.20:** Superposition of structures HCP1-extra (gray) and HCP1-cyto (blue) from **a)** a lateral view **b)** the extracellular side and **c)** the cytosolic side.

HemeFinder calculations reveal that in the open conformation toward the extracellular site there is one possible binding site in between His70 and Tyr422, whereas in the open conformation towards cytosol site there are two possible binding sites, His255 and Cys405 (**Figure 4.21**). These two heme binding regions would correspond to entrance or exit path for each site of the transporter. However, calculations in both conformations suggest the same heme binding regions on the center of the transporter.

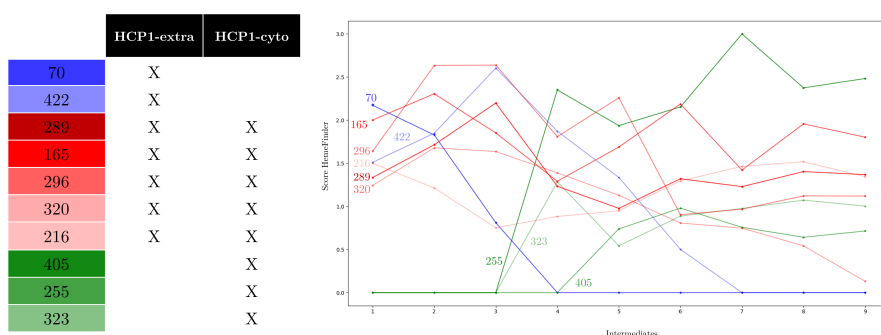
There are three main binding heme regions in the center of the helix barrel: **1)** region of Tyr165 and His289, best scored in both systems, **2)** region in which heme could coordinate with either Tyr296 or Tyr320, **3)** In HCP1-cyto, there is a site that has good score in conformation open towards cytosol, in which heme could coordinate with Tyr323. Results reveal how most heme coordinating residues would correspond to Tyr, which are known have low affinity for heme. This could explain the low affinity for heme of this receptor and higher affinity for folate. In figure 22a are represented all the residues that are involved in the possible heme binding pathway, indicating in which of the two forms they appear.



**Figure 4.21:** HemeFinder results obtained for **a)** HCP1-extra and **b)** HCP1-cyto.

Using Morphing tool implemented in UCSF Chimera, nine intermediate states between both conformations were obtained and HemeFinder was used to find

heme binding sites for all intermediates.<sup>277,307</sup> In **Figure 4.22** are represented the scoring for each binding site for all intermediates, number one representing the closest to HCP1-extra and number 9 to HCP1-cyto. In this representation it can be observed how His70 and Tyr422 disappear as the structures move towards the open cytosolic form (represented in blue), whereas His225, Tyr323 and Tyr405 start to appear (represented in green). The binding sites found in the center of the transporter (colored in red) are maintained in all states, but their scoring fluctuate in the different intermediate states. This clearly shows that this central region corresponds to the heme binding site and there are several possible coordinating residues and 70-422 and 255-405 are two possible pathways that lead there.



**Figure 4.22:** All possible heme coordinating residues obtained by HemeFinder in each intermediate with evolution of scoring over intermediates HCP1-extra and HCP1-cyto.

In order to validate these solutions, molecular dockings were performed to confirm that heme is able to bind in all these regions. Docking results reveal that in some cases Tyr coordination could not be observed, maybe due to rotamer restrictions. However, results show how both Tyr422 in the extracellular site and His255 in the cytosolic site would be able to bind heme and coordinate it.

To compare the performance with other programs, calculations were performed with the only two other heme predictors currently accessible, which are HEMEsPred<sup>291</sup> and HemoQuest.<sup>293</sup> In both cases, calculations took 14 and 30 minutes respectively, whereas HemeFinder took 20s. HEMEsPred was used with a threshold of 0.6 and 14 possible coordinating residues were found. As it uses sequence as input and generates the structure, structure obtained is no similar to crystallographic, consequently the reliability of the results can be questioned. Most results correspond to residues that usually do not coordinate heme, except

for M143, H289, C306 and C405. Met143 and C306 are found in the extracellular region and are completely exposed to the solvent. Two of them, H289 and C405 were also suggested by HemeFinder and are found in the central region and cytosolic part, respectively. HemoQuest looks for transient hemes and found 4 possible binding residues C23, H255, H289 and Y460. From these residues, the only one that in the crystal structure are H255 and H289, both were also suggested by HemeFinder, found in the cytosolic and central part respectively.

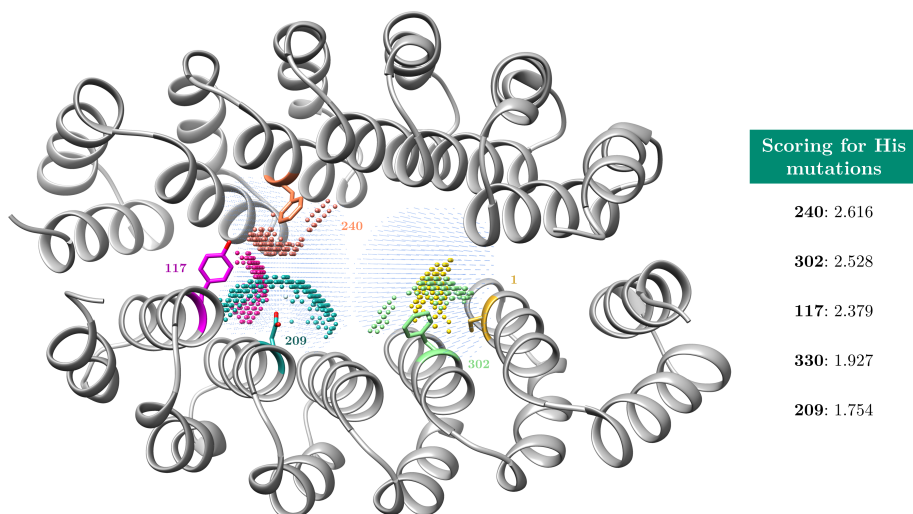
It can be concluded that HemeFinder is the fastest option available and as it is able to offer the whole heme binding pathway. All software suggested H289 as a possible heme binding site, which reinforces the validity of HemeFinder results. Still, all other software available depend on sequence, therefore the two different conformations of HCP1 can not be accounted for and neither HemoQuest nor HemePred was able to find heme binding pathways.

#### 4.2.7. Case study 2: Application in design of ArM

The system of interest of this study is a *de novo protein* that belongs to a family of artificial proteins that are derived from  $\alpha$ -helicoidal HEAT-like repeat protein scaffold, named  $\alpha$ -Rep. These systems contain variable positions that can be mutated without implying any structural changes.  $\alpha$ -Rep dimeric system has been used previously to design ArM with Mn-protoporphyrin through conjugation, as explained in section 1.3.6.<sup>163</sup> The aim of this part of the study is to assess if the heme would be able to bind inside  $\alpha$ -Rep directly and if not, which residues could be mutated to have an ArM.

HemeFinder detects 5 possible heme binding sites, but only two amino acids can bind heme: Tyr209 and Tyr24. Nonetheless, those residues are not the most prone to give hemoproteins with good oxidative profiles, one would expect His or Cys to provide better activity. The second part of the study was to assess if some positions in the cavity could be mutated to these amino acids. Two calculations were performed, using His or Cys as possible mutating residues. Regarding the results of the His calculation, HemeFinder finds 52 possible residues that could be mutated and coordinate with heme. From these, 41 correspond to variable positions from  $\alpha$ -Rep, residues that can be mutated without expecting structural changes.

On the other hand, 24 residues are suggested to be mutated for Cys, all of the belonging to position that are variable. The five residues with the highest scores are all in the middle of the helices, three of them in common between His and Cys. In **Figure 4.23** are represented the results obtained for His mutations.



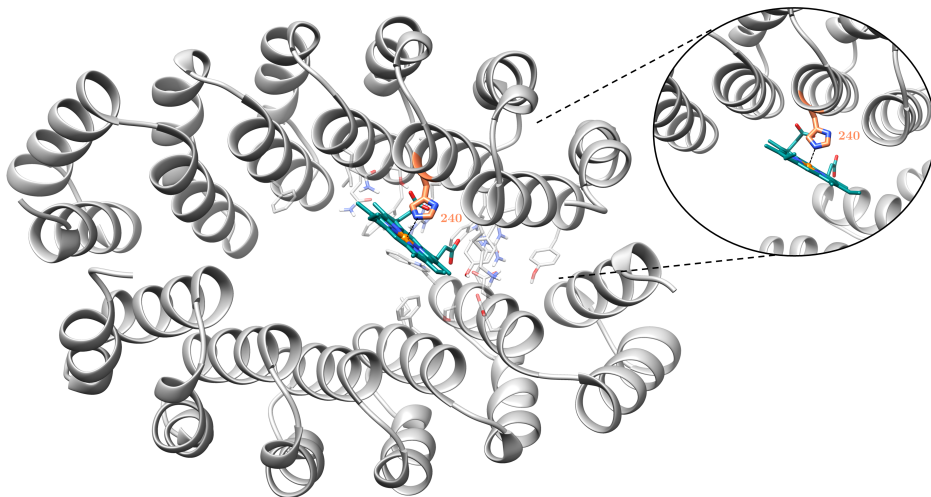
**Figure 4.23:** Results from HemeFinder with residues that could be mutated to His to coordinate heme and its respective scorings.

To check the validity of these results, docking calculations were performed. Residues suggested by HemeFinder were mutated to either His or Cys and dockings were performed with this residue with free rotation to allow coordination. Results for His and Cys are in table 3:

Residue	Scoring HemeFinder	Scoring Docking	Coord.	Residue	Scoring HemeFinder	Score Docking	Coord
H240	2.62	80.84	Yes	C240	1.58	68.61	No
H302	2.53	76.44	No	C55	2.17	69.66	No
H117	2.38	74.65	No	C302	2.14	79.57	No
H330	1.93	68.69	No	C117	2.13	78.52	No
H209	1.75	48.55	Yes	C236	1.83	68.94	No

**Table 4.3:** Results of dockings for His and Cys mutations.

Visually analyzing the results and considering the score from HemeFinder and dockings, the best solution appears to be the one with coordination at His240 (**Figure 4.24**). The docking results display very good coordination and several contacts between surrounding residues. Furthermore, there is an empty volume left next to the heme, which would be necessary for substrate binding.



**Figure 4.24:** Results of docking for mutation His at position 240.

If the same calculations are performed with HEMEsPred, some of the suggested residues are Glu252, Glu256 or Asp326. All these are found in the interface of the two monomers, however, the coordinators are Glu or Asp are not relevant in heme chemistry. HemoQuest suggests residues which are facing the external part of the protein, not favorable for the design of ArM. Neither of the two programs can suggest mutations as HemeFinder, therefore all binding sites suggested by other programs are not feasible for ArM design.

#### 4.2.8. Conclusions

Despite the importance of heme, the binding of heme to proteins and its prediction has not been widely studied. So far, most programs are based on sequence-based predictions which could be valuable for proteins from the natural realm but could be limitative for *de novo* ones. In this work we present, HemeFinder, a new program that allows detecting heme binding sites and designing new hemo-enzymes based only on the protein structure and the geometrical predisposition of the backbone.

HemeFinder has been able to predict more than 95% of crystallographic structures from benchmark, ranking possible heme binding sites considering how preorganized is the backbone of the coordinating residues, the characteristics of the environment and the shape of the binding site. HemeFinder shows great potential for detecting pathways of natural heme binding sites, but also for the design of new ArM. For example, due to the high speed of the program it can be used to screen possible scaffolds for the design of ArM. Compared to other software for heme prediction, HemeFinder offers a fast approach to detect heme binding site that can be further validated by molecular dockings or it could even be combined with other docking tools.



## CHAPTER 5

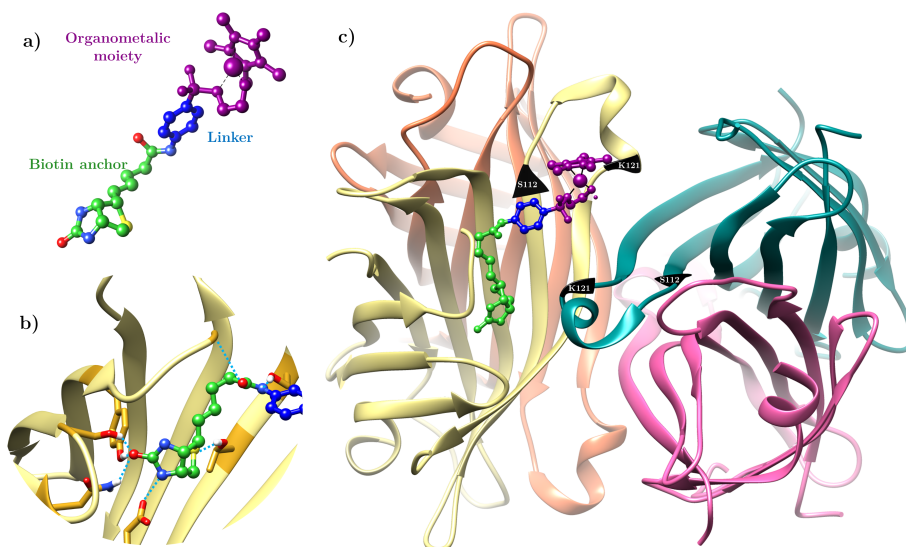
# Applicative cases of computational aided design of ArM

## 5.1. Overview

The aim of this chapter is to use molecular modelling tools to rationalize the design of two ArM. Both projects are based on experimental studies of T.R. Ward group, known for being the pioneer on biotin-streptavidin ArM technology.<sup>147,308</sup> This is an ArM supramolecular anchoring technique that relies on the high affinity between protein streptavidin (Sav) and its cofactor biotin. The dissociation constant between biotin and Sav is in the order of  $4 \cdot 10^{-14} \text{M}$ , which makes this interaction almost irreversible.<sup>309</sup> Sav is a homo-tetrameric protein composed of two dimers of dimers, each of the four monomers is an eight-stranded  $\beta$ -barrel that binds a biotin molecule, which interacts through strong hydrogen bonds and VdW interactions with residues from Sav. The loops connecting the  $\beta$ -stands have certain flexibility and some can acquire either an open or close disposition.<sup>310</sup>

The biotin-Sav strategy consists of incorporating a biotinylated cofactor, composed of a biotin an anchor, a linker, and a metallic moiety, into Sav (**Figure 5.1a**). This strategy is based on the ability to localize a biotinylated metallic cofactor on the biotin-binding site of Sav due to the presence of strong hydrogen bonds between the biotin anchor and surrounding Sav residues (**Figure 5.1b**). The optimization of this type of ArM is usually achieved through a chemogenetic strategy, which involves screening different variants of biotinylated metallic cofactors against several mutated variants of Sav, obtained either by rational

design or by directed evolution. Due to the rigidity of Sav it is difficult to optimize the binding site. However, two residues, Ser 112 and Lys 121, have been identified as potential positions to mutate due to the proximity to the biotin vestibule (**Figure 5.1c**).<sup>311</sup> The biotin-Sav strategy has been used to catalyze a wide range of chemical reactions over the years: hydrogenation, allylic alkylation, dihydroxylation, alcohol oxidation, transfer hydrogenation of ketones and imines, Suzuki cross-coupling or CH activation.<sup>15</sup> Very recently, the design and assembly of an artificial [Fe4S4]-containing Fischer-Tropschase relying on the biotin-Sav technology has been reported.<sup>312</sup>



**Figure 5.1:** a) Parts of biotinylated metallic cofactor b) Vestibule of biotin-binding site with hydrogen bonds with close-by residues. c) Sav tetrameric structure with biotinylated metallic cofactor with S112 and K121 highlighted in black.

The aim of the first study of this chapter is to guide with molecular modeling the design of an ArM based on Sav that catalyzes the hydroamination of alkynes by either single or dual gold catalysis. The premise here is to use computational tools to rationalize the experimental results of the different Sav mutated systems and guide the experimental design by proposing new mutations to improve the regioselectivity of the ArM. The second study is focused on an enantioselective ArM that embraces a palladium catalyst for the synthesis biaryls. The objective of this project is to use computational tools to rationalize how mutations and changes in the linker influence the enantioselectivity of the reaction.

## 5.2. Methodology and computational details

Both projects presented herein are based in a multilevel strategy built on integrating quantum mechanics (QM), molecular dockings and molecular dynamics (MD) approaches. This methodology enables to deal with different aspects of the ArM related to subtle electronic effects and the study of the impact of the protein environment on the embedded non-natural chemical reaction. The overall methodology for ArM design has been described in section 1.4, here we will proceed to explain the details for these two projects.

The general workflow was composed by four steps: **1)** characterization of the full reaction mechanism and identification of key intermediates and transition states (TS) using DFT, **2)** building protein models and exploration of the conformational landscape of the different systems through MD simulations, **3)** insertion of the different intermediates or TS structures into the proteins by protein-ligand approaches, **4)** conducting MD simulations to evaluate the overall interaction between the protein and the intermediates or TS leading to the different stereospecific products. The overall methodology is schematized in (Figure 5.2), followed by a detailed explanation of each step in both projects.

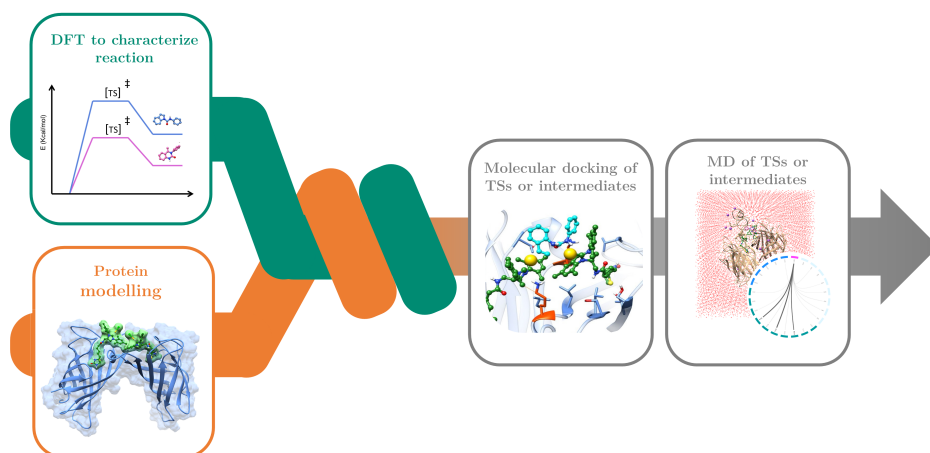


Figure 5.2: Multistep strategy followed for ArM design.

### 1) DFT calculations:

DFT calculations in water solvent were performed with biotinylated metallic cofactors and substrates to scrutinize the reaction mechanisms by computing reaction barriers and the relative stabilities of the possible reaction products. In this way, cofactor-optimized reaction intermediates or transition state (TS) structures were located and characterized by frequency calculations. These structures were then docked into the protein to assess the impact of the protein environment.

All geometry and frequency DFT calculations were performed with the Gaussian09<sup>283</sup> and Gaussian16<sup>313</sup> programs using B3LYP8 functional including Grimme's dispersion D3 (B3LYP-D3).<sup>212</sup> Calculations were carried out in water solvent (SMD continuum model)<sup>284</sup> with  $\epsilon = 78.35$ , except for calculations in which the effect of the medium polarity was tested using other solvents with different dielectric constant. The basis set 6-31G(d,p)<sup>314</sup> was used for non-metallic atoms and SDD for metals (including f polarization functions).<sup>285</sup> Energies in water were refined using an extended basis set, including Def2TZVP<sup>315</sup> for metals and Def2TZVP for non-metallic atoms. Gibbs energies in water solvent were calculated using equation 1. In this way, an additional correction of 1.9 kcal/mol was applied to the Gibbs energies to change the standard state from the gas phase (1 atm) to the condensed phase (1 M) at 298.15K ( $\Delta G^{1\text{atm} \rightarrow 1\text{M}}$ ).

$$G_{\text{water}} = E_{\text{water}}(\text{BS2}) + (G_{\text{water}}(\text{BS1}) - E_{\text{water}}(\text{BS1})) + \Delta G^{1\text{atm} \rightarrow 1\text{M}} \quad (1)$$

### 2) Protein modeling:

Starting from available X-ray structures or by building homology models, Sav systems were modeled in order to later perform the dockings of the relevant reaction intermediates or TSs. For the modeling of Sav, calculations were carried out using as initial structure the X-ray crystal structure of Sav available in PDB (3RY2 or 5CSE).<sup>310,316</sup> The systems were prepared by removing waters, ions and small ligands (except biotin) present in the X-ray structure. Duplicate conformers of amino acids were removed and subsequently hydrogen atoms were added using Chimera UCSF.<sup>277</sup> Mutated systems were obtained using the Dunbrack<sup>317</sup> rotamer library implemented in Chimera UCSF.

In cases in which the full structure was not be fully solved by X-ray, an homology model was using Modeller9.21 program.<sup>318</sup> Next, docking was performed to introduce the optimized biotinylated metallic cofactor into the modeled systems. After parameterization, MD simulations of Sav in its resting state, with the biotinylated cofactor, were carried out to explore the conformational space accessible (see subsection 4 for details of MD). Clustering analysis was performed to obtain the structure of the most populated cluster with a binding site wide enough to accommodate the intermediates or TSs in the dockings.

### 3) Protein-ligand docking:

Protein-ligand docking approaches allowed to incorporate QM-optimized reaction intermediates or TSs structures into the binding site of the different protein systems. These molecular dockings were performed in order to take into account the influence of the protein environment and to assess the effect of mutations. Docking simulations were performed following a covalent docking protocol in which the biotin moiety was fixed, and the rest of the intermediate or TS was covalently attached to this moiety. Two different software were used depending on the objective of the project:

- **GaudiMM17**<sup>249</sup>: the calculations were performed with a population of 100 individuals that evolved for 200 generations. Torsions (flexibility) were applied on the ligand (except for the biotin moiety). Solutions were evaluated considering scoring function Ligscore<sup>319</sup> and minimization of steric clashes. GaudiMM allows simultaneous dockings of two biotinylated entities with distance restriction, while GOLD does not.
- **GOLD5.8**<sup>320</sup>: the calculations were performed with an evaluation sphere of 15 Å and considering side-chain flexibility in key residues and the ligand (except the biotin bicyclic moiety). Genetic algorithm parameters were set to 50 GA runs, a minimum of 100.000 operations and the rest of the parameters were left to default. The Goldscore function was used as scoring function. All results were visualized and analyzed using GaudiView.<sup>249</sup>

### 4) Molecular Dynamics (MD) refinement and analysis:

Taking into account visual structural analysis and docking scores, the best results from docking were refined by MD simulations in order to assess the complementarity between the binding site of the Sav vestibule and the

biotinylated TSs or intermediates. Comparison of residue interactions with different TSs or intermediates was performed to identify the key residues for the catalysis and possible residue candidates for directed evolution.

The best docking results were employed as starting points for the MD simulations. All MD simulations were prepared with the xleap from AMBER18<sup>279</sup> or Amber20<sup>321</sup>. Each system was solvated using an explicit solvent approach by embedding it into a cubic box with a neutral charge (neutralization with Na<sup>+</sup> and Cl<sup>-</sup>). The AMBER14SB or AMBER19SB<sup>322</sup> force field was used for proteins, GAFF<sup>323</sup> for non-standard residues, ions94.lif for ions and TIP3P<sup>324</sup> for water. The parameters to characterize both *biot-Au-2* and TSs were calculated using *MCPB.py*.<sup>168</sup> Charges were calculated using RESP<sup>325</sup> (Restrained ElectroStatic Potential) model and force constants and equilibrium parameters between metal and residues were obtained through the Seminario<sup>282</sup> method. Basis sets and parameters employed for these DFT calculations were the same as specified previously in section 1).

The AMBER or OMMprotocol<sup>326</sup>, depending on the project, were used to perform the MD simulations following a standard simulation protocol. The Langevin integrator<sup>327,328</sup> was used with a time step of 1 fs with periodic box conditions. The simulation was performed at constant temperature and pressure by using a barostat coupled to a bath of 1.01325 bar. A cut-off of 1 nm was used for non-bonded interactions (short-range electrostatic and van der Waals interactions) and the PME<sup>227</sup> method was applied for long-range electrostatic interactions. Additionally, SHAKE<sup>329</sup> algorithm was used to constrain bonds that involve hydrogen and the rigid model was used to represent the water molecules. To avoid steric clashes and relax the system energy minimizations were performed, followed by several equilibration steps to heat the system from 100 K to 300 K in order to allow thermalization of water and side-chains. Finally, production runs of 200-300 ns were carried out.

All MD trajectories were processed using cpptraj implemented in Ambertools18<sup>279</sup> and MDtraj<sup>286</sup> was used to characterize the most populated structural clusters. The trajectory was considered converged when full exploration of the conformational space was reached according to a set of analysis as explained previously in section 2.2.5: RMSD, all-to-all RMSD, RMSE, PCA and clustering counting method.<sup>194</sup>

## 5.3. Molecular modeling to optimize an Au-ArM for heterocyclization

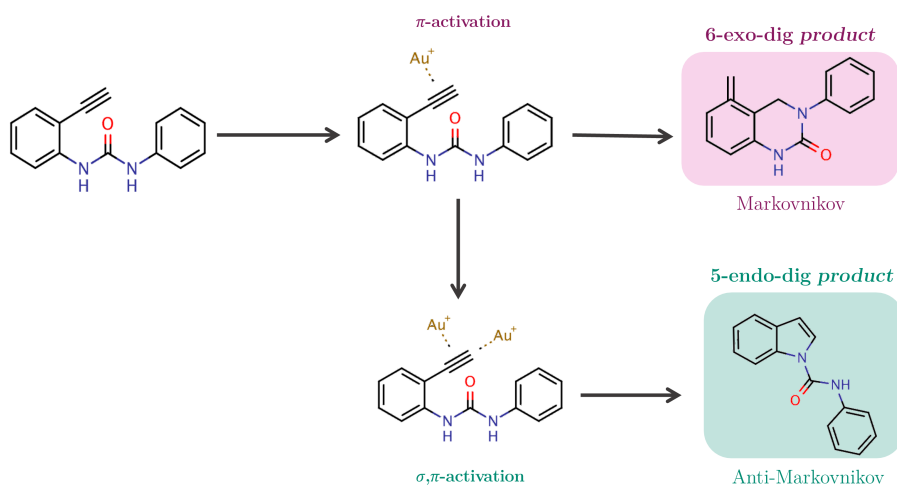
### 5.3.1. Context and experimental background

Biorthogonal ‘click’ reactions are a class of chemoselective, fast and high-yielding chemical reactions that occur in biological environments without any side reactions that might alter the endogenous biological system. These reactions have several fields of application, including biomedical bioimaging/labeling, polymer science or drug delivery.<sup>330–332</sup> The importance of this field is reflected on the fact that the Nobel Prize of 2022 was awarded to K. Barry Sharpless, Morten Meldal, and Carolyn R. Bertozzi for their contributions on biorthogonal click chemistry.

Among the classes of biorthogonal reactions are alkynes reactions. They can be catalyzed by copper in the case of azide-alkyne cycloadditions or catalyzed by ruthenium or palladium as in Suzuki-Miyaura cross-coupling reactions.<sup>330</sup> Yet, gold has recently risen as one of the most effective catalyst for the activation of alkynes under mild conditions.<sup>333–335</sup> Gold(I) complexes typically activate reactions through  $\pi$ -activation. However, dual gold activation through  $\sigma,\pi$ -activation has also been reported.<sup>336</sup> Gold precatalyst are not very reactive, they require prior transformation by chloride abstraction.<sup>335</sup> Recent studies have demonstrated that gold catalysis can be performed in biological environments, but no dual gold catalysis had been reported so far.<sup>337,338</sup> Upon this premise, T.R. Ward set up to design the first ArM that catalyses an hydromination reaction based on either single or dual gold activation of an alkyne.

The reaction of interest carried out by the ArM is the gold-catalyzed hydroamination of ethynylphenylurea. The regioselectivity of this reaction was studied previously in organic solvent by Asensio’s group, reporting that depending on the gold catalyst employed the reaction affords either the 6-exo-dig (quinazolinone) or 5-endo-dig (indole) product, which are the Markovnikov and Anti-Markovnikov products respectively.<sup>339</sup> Following studies revealed the competition between  $\pi$ - and dual  $\sigma,\pi$ -activation modes of gold, which depend on the gold complex of choice. Labeling experiments with deuterate terminal alkyne reported that the  $\pi$ -activation mode leads to the 6-exo-dig product, while the dual  $\sigma,\pi$ -activation mode lead to the 5-endo-dig

(Figure 5.3).<sup>340</sup> In the latter case, the initial  $\pi$ -activation mode increases the acidity of the alkyne's C-H, promoting its deprotonation and allowing the formation  $\sigma,\pi$ -gold complex.<sup>340</sup> Subsequent research by van der Vlugt's group demonstrated that a well-defined and preorganized dinuclear gold center enforces the  $\sigma,\pi$ -activation.<sup>341</sup> These precedents set the basis to design a dual gold hydroaminase ArM (HAMase).



**Figure 5.3:** Hydroamination of ethynylphenylurea with gold(I) can proceed through  $\pi$ -activation and afford the 6-exo-dig product (Markovnikov addition) or through  $\sigma,\pi$ -activation to provide the 5-endo-dig product (Anti-Markovnikov addition).

To perform the dual gold activation of alkynes, Sav was selected due to the possibility of placing of two gold catalyst at close distances. Five different gold complexes were synthesized and tested for the hydroamination of ethynylphenylurea, in both buffer solutions and in the presence of different variants of Sav. In the presence of Sav-wt the catalytic activity increased up to 12 TON, revealing that the ArM accelerates the catalytic activity compared to the cofactor alone. However, only the 6-exo-dig product was obtained.

Gold complex *biot-Au-2* was selected as the best cofactor and Sav mutants in positions K121 and S112 were screened. The different variants revealed that almost all mutations improved the TON, but only mutation in K121A produced TON of 5-endo-dig product. Still, even in this case the ratio between products was mainly displaced towards the 6-exo, indicating that two different modes of gold activation were occurring, but one was the predominant one (Figure 5.4).

With the aim of shielding the biotin-binding site, a Sav-K121A chimeric protein was engineered by inserting a two Greek key  $\beta$ -barrel domain referred as SOD from superoxide dismutase C. It was hypothesized that this lid may provide a hydrophobic environment and more stability, which was reflected in the increase of the TON of the 5-endo-dig product to similar ratios of the 6-exo-dig product. (Figure 5.4)

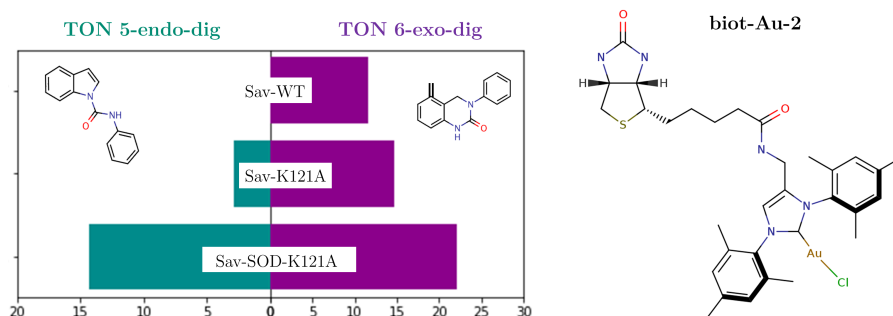


Figure 5.4: Experimental results with *biot-Au-2* biotinylated cofactor and three different Sav variants. Structure of biotinylated cofactor *biot-Au-2* employed.

### 5.3.2. Objectives and methodology

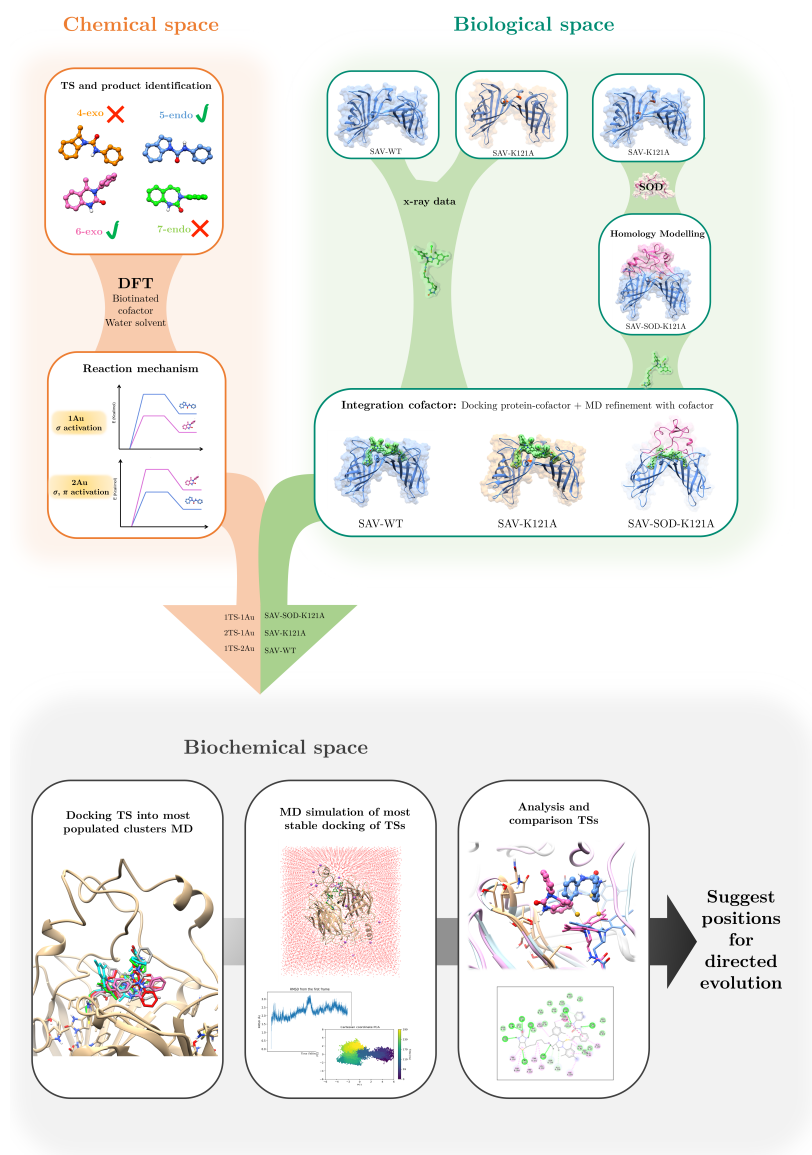
At this point, a multiscale computational study was carried out in order to gain more insight into the catalysis and guide the design of the ArM to afford the dual gold catalysis and its 5-endo-dig product. The objectives set for this part of the work were: **1)** establish most stable product in water solvent using DFT, **2)** Determine the behavior of each protein system using Docking and MD, **3)** Analysis of MD to suggest mutations to improve the ratio of the 5-endo-dig product. In order to fulfill these objectives, the workflow represented in (Figure 5.5) was employed. Detailed methodology was described previously, but briefly we will go through the specific steps of this project.

First, DFT calculations in water solvent were performed in order to establish the reaction mechanism for both  $\sigma$ - and  $\sigma,\pi$ -activation modes and to compute reaction barriers that lead to both reaction products (5-endo-dig and 6-exo-dig). The N-heterocyclic carbene gold(I) complex  $[(\text{IMes})\text{AuCl}]^+$  (IMes = 1,3-dimesitylimidazol-2-ylidene), its biotinylated version *biot-Au-2* and ethynilurea substrate were all employed in the DFT calculations.

Regarding the protein part, first the apo form of the protein was modeled. For Sav-WT and Sav-K121A, calculations were performed using X-ray structure (3RY2)<sup>310</sup>. Since the full structure of Sav-SOD-K121A could not be fully solved by X-ray, i.e. the SOD region was disordered, an homology model was built for Sav-SOD-K121A. The structure of the region of interest of SOD (Ala37-Lys71 from PDB:1PZS)<sup>14</sup> was positioned above the X-ray structure of Sav-SOD-K121A. Loop modeling was performed to link both protein ends and the best model was selected according to the energy score from Modeller<sup>318</sup> (DOPE). This structure was subjected to a MD simulation of 200 ns with Sav region constrained to allow SOD stabilization and accommodation into Sav. Once the apo form was modeled, the optimized biotinylated cofactors were introduced into the protein structure by molecular docking and refined by MD simulations of 300ns.

Molecular dockings were performed in order to take into account the influence of the protein environment and to assess the effect of mutation K121A and the presence of the SOD cap. DFT optimized structures of transition states 1TS-2Au ( $\sigma,\pi$ -activation) and 1TS-1Au ( $\pi$ -activation), leading to 5-endo-dig and 6-exo-dig products, respectively, were docked into most populated clusters from the previous MD simulations for all three systems (Sav-WT, Sav-K121A, Sav-SOD-K121A). MD simulations were performed in order to assess the complementarity between the binding site of the Sav vestibule and the number of *biot-Au-2* catalysts involved. Comparison of residue interactions with different TSs was performed to identify the key residues for the catalysis and possible residue candidates for directed evolution.

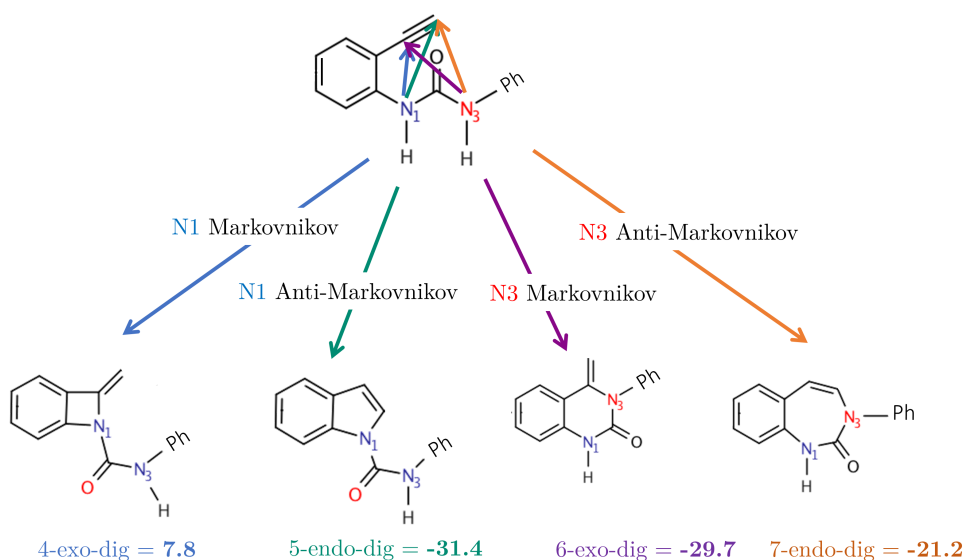
Residues contribution analysis was performed using StructureViz, a plug-in that links Cytoscape<sup>342</sup> and Chimera UCSF. Van der Waals (VdW) contacts and hydrogen bonds were analyzed using the Chimera default parameters every 200 ps of the trajectory. The contributions of the residues to the binding Gibbs energy were analyzed with the MMGBSA method implemented in the module MMPBSA.py<sup>343</sup> using  $igb = 0$  and an ionic strength = 0.1M. The calculations were carried out using the most stable fractions of the trajectory and the most relevant residues determined through network interaction analysis were considered.



**Figure 5.5:** Steps followed during molecular modeling workflow for optimization of HAMase.

### 5.3.3. Learning from the organometallic side: DFT calculations

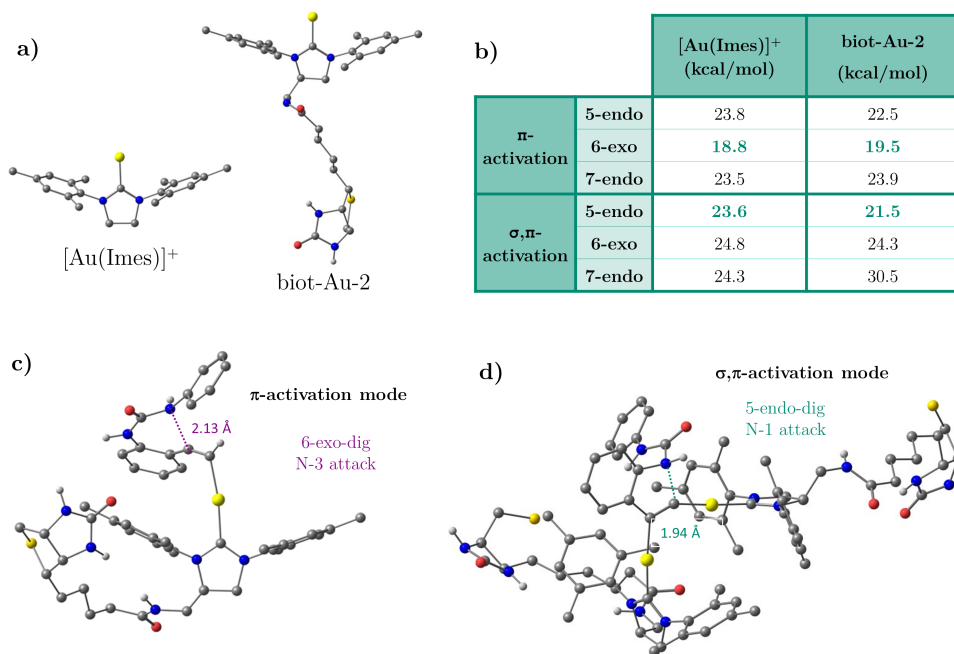
In a hydroamination reaction, a nucleophilic nitrogen is added to an unsaturated carbon center, resulting in a new unsaturated C-N bond. The focus of our study is the hydroamination reaction of ethynylphenylurea, which in principle can yield four different products. The outcome depends on which of the two N centers is added and to which sp carbon is added. Two of these products are Markovnikov (N is added to the most substituted carbon) and other two Anti-Markovnikov (N is added to the less substituted carbon). Initially, the relative stability of all four products was calculated (**Figure 5.6**). These results reveal the low stability of the 4-exo-dig product, leading us to exclude it from further calculations.



**Figure 5.6:** Four main possible products of intramolecular hydroamination of ethynylphenylurea and their calculated relative stability in Kcal/mol.

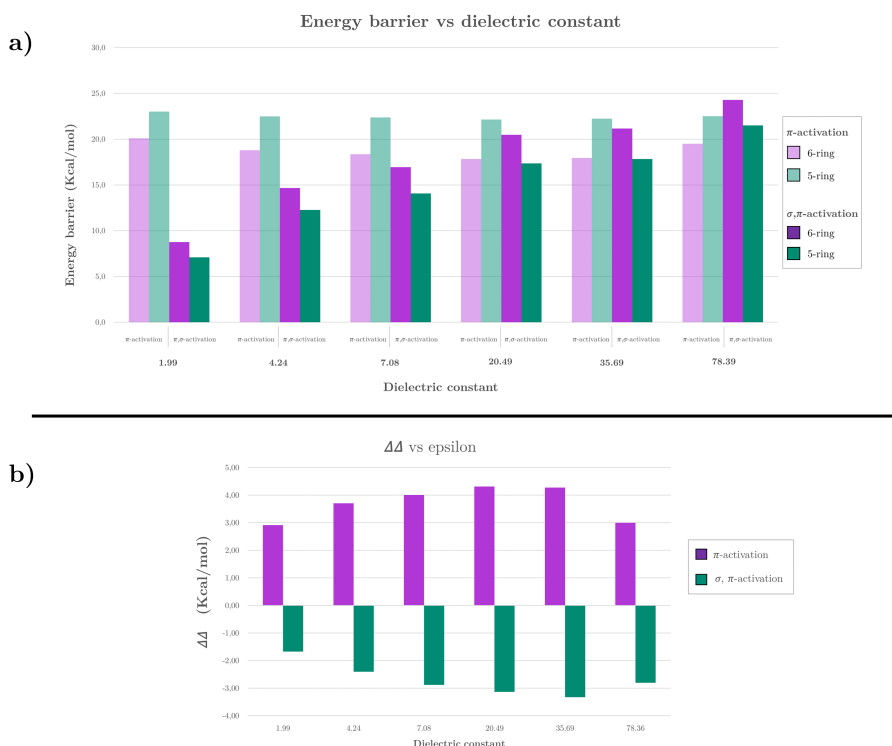
Then, DFT calculations were conducted in order to study the competition between the two mechanistic pathways: single gold  $\pi$ -activation and dual gold  $\sigma,\pi$ -activation modes, for 5-endo, 6-exo-dig and 7-endo products. The reaction pathways calculations were performed in water solvent considering two different systems as represented in **Figure 5.7a**: **a)** non-biotinylated IMes ligand **b)** catalytic cofactor *biot-Au-2*. The Gibbs energies of activation for both systems and both activation modes are represented in **Figure 5.7b**.

Regarding the biotinylated system, the 7-endo-dig product has the highest Gibbs energy barrier for both  $\pi$ - and  $\sigma,\pi$ -activation activating modes, explaining the absence of this product in the ArM experiment. With one gold(I) complex in the  $\pi$ -activation mode the formation of 6-exo-dig dig product is clearly favored, as the Gibbs energy barrier is 3 kcal/mol lower for 6-exo-dig than the 5-endo-dig product (19.5 vs 22.5). Contrarily, the 5-endo-dig product has a lower Gibbs barrier than the 6-exo-dig product (24.3 vs 21.5) in the  $\sigma,\pi$ -activation mode, when there are two gold(I) complexes interacting. Transition states for 6-exo-dig  $\pi$ -activated and 5-endo-dig  $\sigma,\pi$ -activated biotinylated complexes are represented in **Figure 5.7c,d**. Both calculations with non-biotinylated IMes and full *biot-Au-2* cofactor show the same tendency, indicating that the presence of biotin does not change the regioselectivity of the reaction. All calculated TS for biotinylated systems are represented in **Figure C.1**.



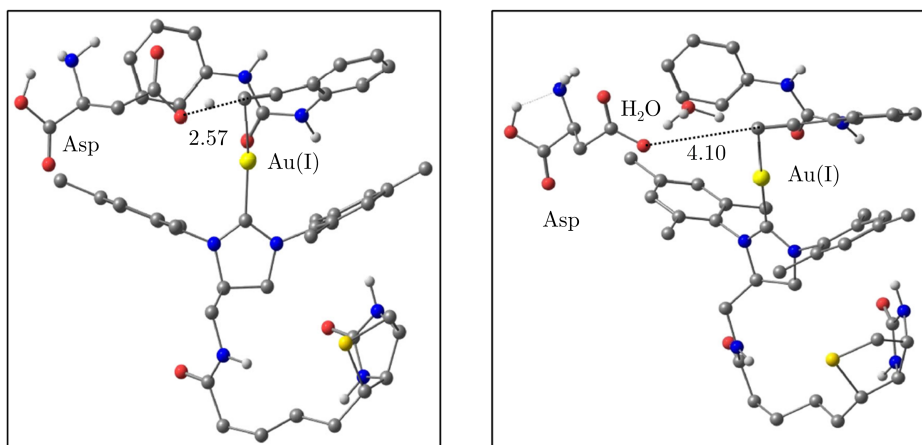
**Figure 5.7:** a) Structure of non-biotinylated IMes ligand and catalytic cofactor *biot-Au-2*. b) The Gibbs energies of activation of non-biotinylated IMes and *biot-Au-2* for both  $\pi$ - and  $\sigma,\pi$ -activation modes. c) Optimized structure of favored TS for 6-exo-dig product by  $\pi$ -activation mode. d) Optimized structure of favored TS for 5-endo-dig  $\sigma,\pi$ -activation mode. Reprinted from [344].

Overall, these results indicate that the competition between  $\pi$ - and dual  $\sigma, \pi$ -activation modes operates in water and the difference between the Gibbs energy barriers for both activation modes is about 2 kcal/mol (19.5 vs 21.5). This low difference suggests that subtle changes in the first or second coordination sphere of the metal in the protein environment may change the ratio between both products. Still, as the reaction takes place in a more hydrophobic environment in the protein, Gibbs energy barriers were calculated in solvents with different dielectric constants. From **Figure 5.8a** it can be clearly observed how the energy barriers of the  $\sigma, \pi$ -activation change considerably with the dielectric constant of the solvent, whereas that of the  $\pi$ -activation mode is less affected by the polarization of the medium. Interestingly, the tendency of the 6-exo-dig product being more favorable in the  $\pi$ -activation and the 5-endo-dig in  $\sigma, \pi$ -activation is not affected by the polarity of the medium (**Figure 5.8b**).



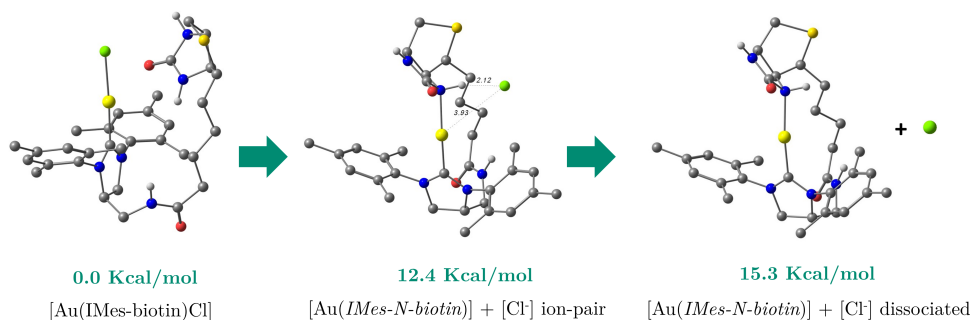
**Figure 5.8:** **a)** Gibbs energy barriers for the reactions yielding the 6-exo-dig and 5-endo-dig products by both activation modes in water ( $\epsilon = 78.39$ ) and in solvents of varying dielectric constant. **b)** Difference between the barriers ( $\Delta\Delta G^\ddagger$ ) for the 6-exo-dig and 5-endo-dig products for both activation modes in water ( $\epsilon = 78.39$ ) and in solvents of different dielectric constant. Reprinted from [344].

From these findings it can be hypothesized that the regioselectivity of the reaction is not considerably affected by polarity of the medium, but the protein environment or residues could induce a shift in the regioselectivity. Another aspect to consider is the protein environment is the alkyne's C-H deprotonation. As was mentioned in the introduction, a deprotonation of the alkyne's C-H proton is required to form the  $\sigma,\pi$ -activation mode. Therefore, the transition state and the Gibbs energy barrier was computed for the deprotonation of alkyne's C-H either by an aspartate, or through an activated water molecule activated by an aspartate (**Figure 5.7**). Both pathways have feasible Gibbs barriers of 5.7 and 9.5 kcal/mol respectively, meaning that in presence of a suitable base the deprotonation can happen in a protein environment. Once the  $\pi$ -activation complex has formed, the formation of the  $\sigma,\pi$ -activation is exergonic with a  $\Delta G$  of -7.3 kcal/mol.



**Figure 5.9:** DFT computed transition states in water for the alkyne's C-H deprotonation, either by **a)** an aspartate residue or **b)** through a water molecule activated by Asp. Reprinted from [344].

The final aspect from the reaction we investigated was the dissociation of the chloride ligand from metal complex *biot-Au-2*, which is the precursor step of the catalytic reaction. In water solvent, chloride dissociation can take place with a cost of about 15 kcal/mol. However it yields to species in which the biotin is coordinated to the gold center either through the O, N or S. These coordinated structures are 24 kcal/mol more stable than the linear form and they may explain the inactivity of the free form of the catalytic cofactor.



**Figure 5.10:** Extration of chloride from biotinylated cofactor *biot-Au-2* in water.

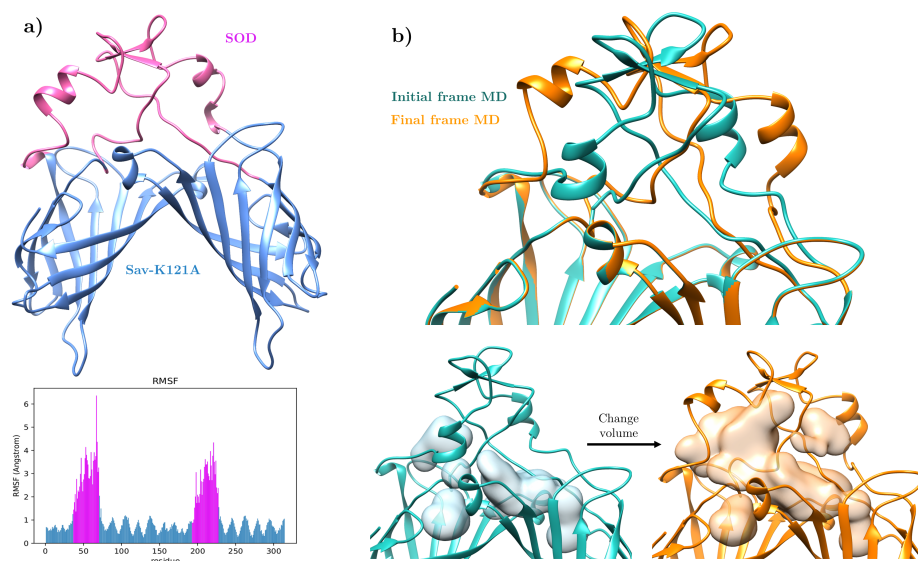
In a protein environment the biotin is fixed in the binding site adopting a linear conformation and consequently the previous coordinated species is not feasible. Calculations for the extraction of chloride were performed using a lineal biotinylated cofactor and the dielectric constant was set to four, which has been reported to be the approximate dielectric constant in Sav environment.<sup>345</sup> Results revealed that the extraction of chloride was only feasible when chloride is solvated by two water molecules, and it is substituted by urea. All other options represented at **Table 5.1** have higher Gibbs energies.

Different reactions	Elect BS1	Gibbs BS1
$\text{BTN\_AuCl} + \text{H}_2\text{O} \rightarrow \text{BTN\_Au} + \text{Cl\_H}_2\text{O}$	48.1	48.8
$\text{BTN\_AuCl} + 2\text{H}_2\text{O} \rightarrow \text{BTN\_Au} + \text{Cl\_2H}_2\text{O}$	48.7	49.0
$\text{BTN\_AuCl} + 2\text{H}_2\text{O} \rightarrow \text{BTN\_Au} + \text{Cl\_2H}_2\text{O}$	46.3	42.2
$\text{BTN\_AuCl\_2H}_2\text{O} \rightarrow \text{Cl\_2H}_2\text{O} + \text{BTN\_Au}$	57.7	42.7
$\text{BTN\_AuCl\_2H}_2\text{O} + \text{H}_2\text{O} \rightarrow \text{Cl\_2H}_2\text{O} + \text{BTN\_Au\_H}_2\text{O}$	16.3	11.6
$\text{BTN\_AuCl} + 2\text{H}_2\text{O} \rightarrow \text{Cl\_H}_2\text{O} + \text{BTN\_Au\_H}_2\text{O}$	13.3	15.0
$\text{BTN\_AuCl\_2H}_2\text{O} + \text{Urea} \rightarrow \text{Cl\_2H}_2\text{O} + \text{BTN\_Au\_Urea}$	<b>-1.3</b>	<b>-0.5</b>
$\text{BTN\_AuCl} + \text{Urea} + 2\text{H}_2\text{O} \rightarrow \text{Cl\_2H}_2\text{O} + \text{BTN\_Au\_Urea}$	-12.2	3.9

**Table 5.1:** Different reactions tested for extraction of chloride.

### 5.3.4. Modeling of protein systems

With no X-ray structure available for the Sav-SOD-K121A system, structural modeling was carried out followed by classical MD simulations constraining Sav to accommodate the SOD region. MD analysis reveals that convergence of the SOD regions is achieved after 200ns according to PCA and RMSD analysis (**Figure C.2**). The flexibility of the SOD region is high, with an average RMSF of 2.7 Å, but reaching 5-6 Å in the loops corresponding to most external part of the SOD region and the  $\alpha$ -helices (**Figure 5.11a**). The structural integrity of SOD regions remains intact during the MD simulation, with the upper two  $\beta$ -sheets interacting and maintaining the two subunits of SOD close. However, the  $\alpha$ -helices from each subunit separate, resulting in the expansion of the biotin-binding vestibule. This is clearly demonstrated by analyzing the volume of the binding site, which increases from 774.15 Å<sup>3</sup> in the initial frame to 1614.81 Å<sup>3</sup> at the last frame of the simulation (**Figure 5.11b**). The expansion of the biotin-binding vestibule allows the inclusion of cofactor *biot-Au-2*.



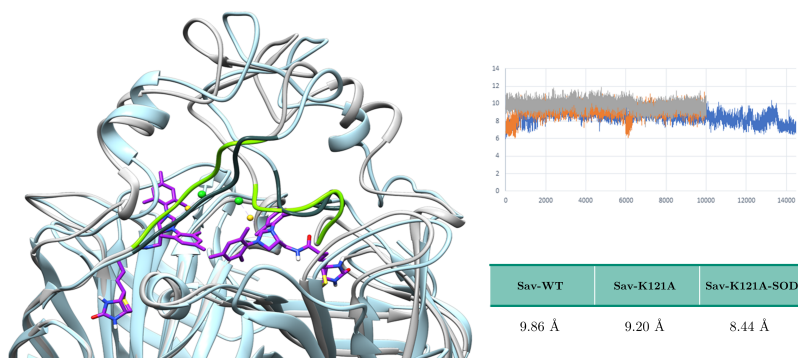
**Figure 5.11:** a) Structure of Sav-SOD-K121A and RMSF with blue representing Sav and pink the SOD region. b) Initial and final frame of MD with volume of biotin-binding vestibule.

Then, after inclusion of *biot-Au-2* into the protein vestibule by dockings, MD simulations on the three Sav scaffolds (Sav WT, Sav-K121A, Sav-SOD-K121A) in its ground state (bound to *biot-Au-2*) were performed to determine the

conformational space available for the TSs binding. In the case of Sav-WT and Sav-K121A, MD simulations converge only after 200ns, revealing the stability of the systems with the catalytic complexes bounded. The core of Sav is not flexible and the RMSF values are low, except for the loops that connect the  $\beta$ -barrels.

In the case of Sav-SOD-K121A, MD converged after 300ns, with a larger exploration of the conformational space compared to two previous systems according to PCA analysis (**Figure C.2**). The RSMF of the SOD region decreased to 1.94 Å from the previous MD simulation value (2.7 Å), indicating that the presence of the *biot-Au-2* lowers the flexibility of the SOD region. Superposition of most representative clusters of this MD with cofactors and without them highlights that the structure of the SOD region remains unchanged, but has moved slightly (**Figure 5.12a**). The SOD relevant loops are quite flexible and move significantly at the beginning of MD to adapt to the presence of *biot-Au-2*.

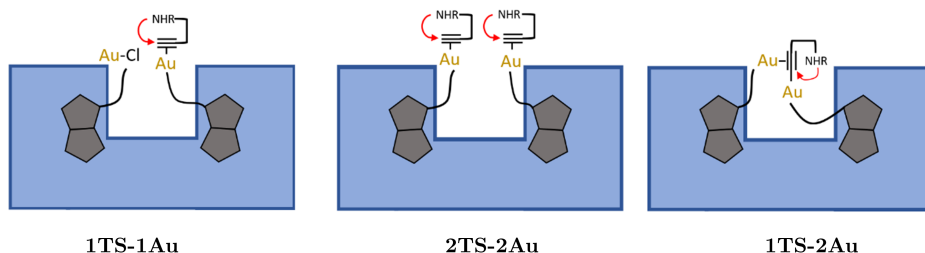
Analysis of the RMSD during MD of *biot-Au-2* reveals that in Sav-SOD-K121A both cofactors remain fluctuating in the same disposition in the binding site, while in the other Sav systems there are rearrangements, probably due to the exposition of *biot-Au-2* to the solvent. Analysis of the distances between two gold atoms from each binding site during the MD was performed to assess proximity of the two catalytic cofactors. From the three systems, the smallest distance corresponds to Sav-SOD-K121A, even at some points of this MD the distances reaches 6.5 Å. Sav-WT presents the largest gold distances, while Sav-K121A at some points reaches shorter distances (**Figure 5.12b**).



**Figure 5.12:** a) Superposition of most representative cluster from MD with cofactors (light blue) and without them (gray) with SOD relevant loops highlighted. b) Average values and distance between gold(I) centers during MD simulations of protein systems in their ground state. Reprinted from [344].

### 5.3.5. Molecular dockings and MD simulations of TSs

Clustering analysis was performed to obtain the structure of the most populated cluster with a binding site wide enough to accommodate the pseudo-TSs. Having identified by QM the TSs structures in water for the isolated cofactor and modeled the different protein systems, three different dockings approaches were considered to incorporate the different pseudo-TSs into the proteins: 1TS-1Au, 2TS-2Au and 1TS-2Au (**Figure 5.13**). 1TS-1Au contains only one  $\pi$ -activated transition state with one gold complex and one alkaline substrate occupying one biotin-binding vestibule and with a second *biot-Au-2* occupying the neighboring biotin-binding site. Similarly, 2TS-2Au contains two  $\pi$ -activated TS, which occupy both biotin-binding sites at the same time. Finally, 1TS-2Au corresponds the  $\sigma,\pi$ -activated transition state with two gold complexes, one in each biotin-binding vestibule interacting with only one alkyne substrate.



**Figure 5.13:** Schematic representation of dockings performed: 1TS-1Au, 2TS-2Au and 1TS-2Au.

Before performing the dockings, a DFT rotational study was performed to assess if certain bonds of the pseudo-TSs should rotate during the calculations. The amide bonds from the alkyne substrate have a rotational barrier of 8-18 kcal/mol, barrier that is feasible to overcome in a protein environment (**Figure C.3**). On the other hand, the bond between Au and the alkyne is completely rotatable as the barriers calculated are between 1-3 kcal/mol.

Molecular docking were performed for three different protein systems (Sav-WT, Sav-K121A and Sav-SOD-K121A) with TSs for both 5-endo-dig (TS5) and 6-exo-dig (TS6). GaudiMM scoring values obtained using scoring function Ligscore are summarized in **Table 5.2**, in which lower scores highlight a better complementarity. To compare scoring values for systems with one and two Au centers, the energy of interaction by gold complex are presented (bold values).

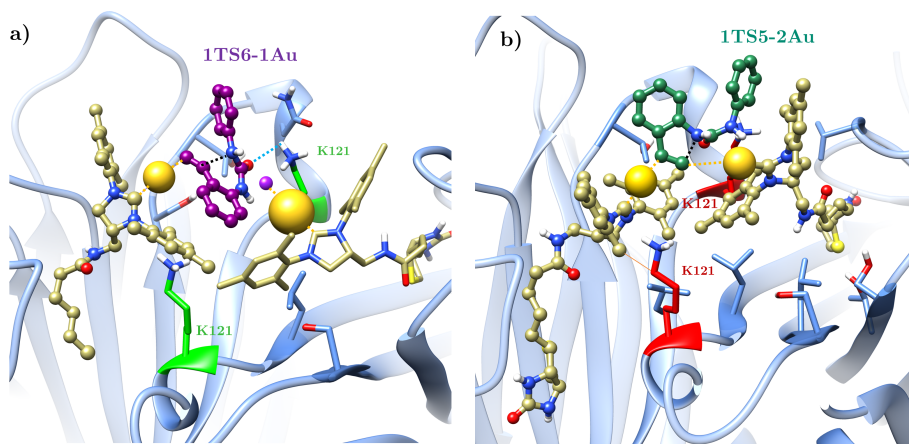
	WT		K121A		SAV-SOD-K121A	
	TS5	TS6	TS5	TS6	TS5	TS6
1TS-1Au2	-71.83	-72.22	-62.74	-61.17	-65.5	-62.54
2TS-1Au2	-118.04	-119.78	-94.99	-80.55	-31.44	-19.99
	<b>-59.02</b>	<b>-59.89</b>	<b>-47.50</b>	<b>-40.28</b>	<b>-15.72</b>	<b>-10.00</b>
TS-2Au2	-83.25	-82.01	-109.94	-81.68	-98.99	-60.70
	<b>-41.63</b>	<b>-41.00</b>	<b>-54.97</b>	<b>-40.84</b>	<b>-49.50</b>	<b>-30.35</b>

**Table 5.2:** GaudiMM scoring values (Ligscore) for TS5 and TS6. Three docking calculations (1TS-1Au, 2TS-2Au and 1TS-2Au) are performed for each Sav system. Energy of interaction by gold complex is presented in bold to compare different calculations. Adapted from [344].

Docking of **Sav-WT** display best complementarities for model 1TS-1Au, with docking scores similar for both 1TS5-1Au and 1TS6-1Au, but with different dispositions. In the best scored solutions, 1TS6-1Au is situated in one side of the binding site, interacting with residues Asn118 and Lys121 (**Figure 5.14a**), while 1TS5-1Au is situated in the center of the binding site. Worse complementarities are obtained for both 2TS-2Au and 1TS-2Au systems, specifically for the latter, as there are several clashes with Sav. Mainly there are clashes with K121, which is situated in the center of the binding site not allowing the two gold complexes to get close enough to make the  $\sigma, \pi$ -interaction **Figure 5.14b**.

From the docking calculations it can be determined that pseudo-TSs are completely exposed to the solvent, without making many interactions with the secondary coordination sphere. Docking scores show no significant differences and since it is completely exposed to the solvent, the Gibbs energy barriers are expected to those calculated for the free cofactor. Therefore, in the case of Sav-WT the 6-exo-dig product is favored instead of the 5-endo.

A MD simulation of 300ns was performed of 1TS6-1Au ( $\pi$ -activation mode) in one biotin vestibule and with *biot-Au-2* occupying the other to verify the stability of the docked pose. During the first 100ns the substrate part of 1TS6-1Au is interacting through a hydrogen bond with the backbone of K121. However, during the MD 1TS6-1Au rotates and establishes another hydrogen bond with the lateral chain



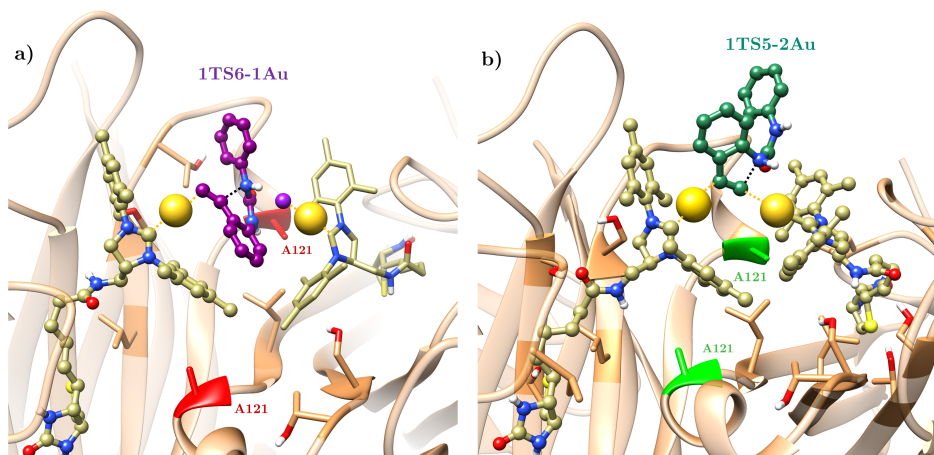
**Figure 5.14:** Computed TS docked within Sav-WT: a) 1TS6-1Au and b) 1TS5-2Au.

of K121. Due to this movement, the *biot-Au-2* from the other subunit rotates 90 degrees to accommodate this change. This interaction is maintained through the rest of the simulation revealing its strength.

In **Sav-K121A**, differences are observed due to the substitution of Lys (big and positively charged residue), for Ala (small and apolar) in position 121. Similar or worse fitting scores are obtained for this mutated system compared to Sav-WT for all pseudo-TSs, except for 1TS-2Au. In the case of 1TS6-1Au, the docking position closely resembles that of Sav-WT. However, the hydrogen bond and VdW interactions between K121 and the substrate are lost due to the mutation to Ala **Figure 5.15a**. This loss is reflected in the decrease of the docking scores. In contrast, for the  $\sigma,\pi$ -activated system for TS5 (1TS5-2Au), the docking scores obtained are higher compared to Sav-WT. This is attributed to the single mutation A121, which makes the binding site bigger due to the smaller size of Ala. In consequence, the two biotinylated gold complexes are able to approach each other and get close enough to make the  $\sigma,\pi$ -interaction (**Figure 5.15b**). There are less clashes between the gold complexes and residues from Sav. Docking scores are similar for 1TS5-2Au and 1TS6-2Au, which indicates that both activation modes are possible in this system.

From the DFT calculations it was determined that electronics favor the 5-endo-dig product in  $\sigma,\pi$ -activation mode and the dockings results support the idea that the disposition of the cofactors to achieve the this activation mode is better for the

TS5 than TS6. Still, as the biotinylated gold cofactors are completely exposed to the solvent and they are quite flexible, it can be hypothesized that the likelihood of the  $\sigma$ -interaction is not very high. This hypothesis aligns with the experimental results, which show a modest shift in favor of the 5-endo-dig product.

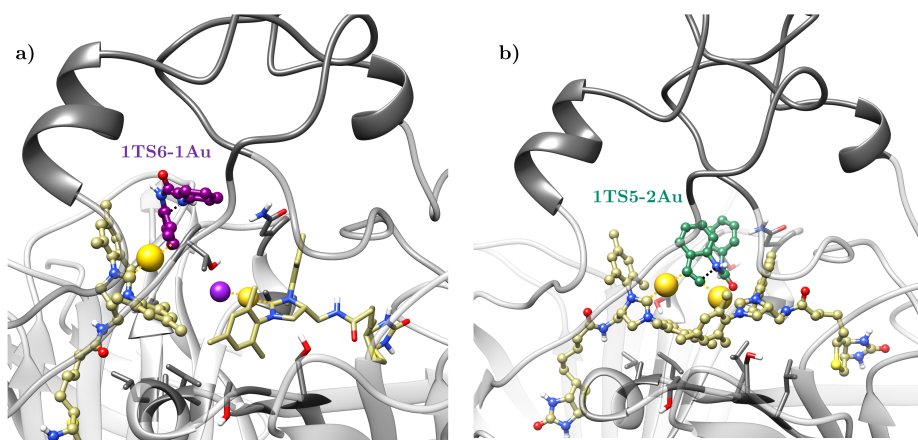


**Figure 5.15:** Computed TS docked within Sav-K121A: a) 1TS6-1Au and b) 1TS5-2Au.

MD simulations of Sav-K121A for both 1TS6-1Au and 1TS5-2Au were performed. In the former, the 1TS6-1Au moves away from A121 and positions at the center of the binding site. The TS is not directly interacting with any residues from Sav, but instead establishes  $\pi$ -stacking interactions with *biot-Au-2* that is occupying the other biotin-binding vestibule. In the case of the 1TS5-2Au, as the two gold cofactors are involved in the  $\sigma, \pi$ -interaction, the movement of this TS is restrained and remains in the center of the binding site.

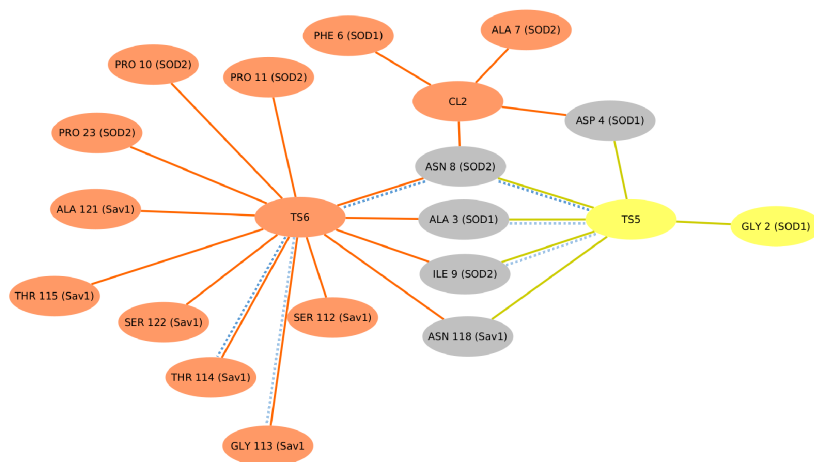
For **Sav-SOD-K121A** the best complementarities are obtained for both 1TS-1Au and 1TS-2Au. Scoring values for 2TS-2Au were extremely low as in the presence of the SOD lid there is limited space in the binding site for two pseudo-TSs. These results suggest that the possibility of a second substrate approaching 1TS-2Au to form 2TS-2Au is very unlikely in chimeric Sav-SOD-K121A. From this docking calculations what can be clearly determined is that the presence of SOD lid restricts the possible dispositions of the gold complexes. The docking solutions from Sav-WT and Sav-K121A were quite varied and very different dispositions of pseudo-TSs were obtained, whereas in the chimeric form of Sav-SOD-K121A one main clear disposition was obtained. For the  $\sigma, \pi$ -activation

mode with TS5, the mutation K121A allows the two gold complexes to interact and the presence of the SOD lid restricts the movement of the gold complexes, increasing the probability of the  $\sigma$ -interaction. The distance between Au and C from triple bond is the lowest compared to Sav-WT and Sav-K121A (2.91 Å). Depending on the TS, the substrate or substrates occupy different positions within the active site. The most notable difference is a 180-degree rotation of the substrate between 1TS6-1Au and 1TS5-2Au (**Figure 5.16**).



**Figure 5.16:** Computed TS docked within Sav-SOD-K121A: **a)** 1TS6-1Au and **b)** 1TS5-2Au.

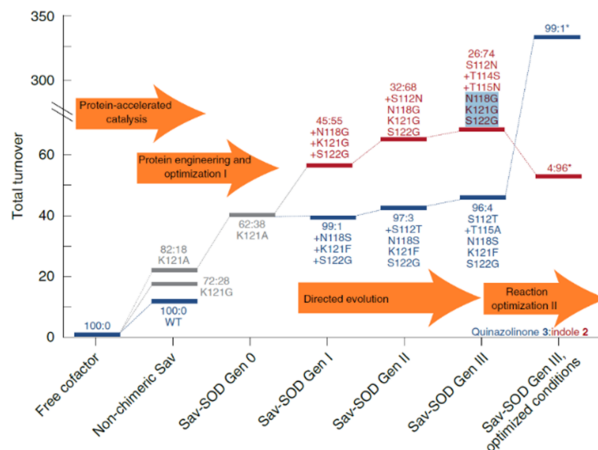
Docking calculations suggest that both 1TS6-1Au and 1TS5-2Au are feasible in Sav-SOD-K121A. In consequence, MD simulations of both systems were performed to decipher the most important interactions for each mode and determine which mutations can shift the regioselectivity towards the 5-endo-dig or 6-exo-dig product. MD simulations converge after 300ns (**Figure C.4**) and a qualitative analysis of the main interactions along the MD simulations was performed using Cytoscape. The interactions along the MD simulations are represented as interaction maps (**Figure C.5**). The node at the center of the map is the TS, the edges represent the interactions, and the nodes are the residues with whom its interacting. Dotted blue edges represent hydrogen bond interactions, whereas continuous edges are VdW interactions. The width of the edges is proportional to the presence of the interactions during MD simulation. Both interaction maps (1TS6-1Au and 1TS5-2AU) were combined in order to clearly determine which residues are common in both pseudo-transition states and which are critical for only one (**Figure 5.17**).



**Figure 5.17:** Residues contribution network extracted from MD simulations of Sav-SOD K121A comparing 1TS5-2Au and 1TS-1Au interactions. Common interacting residues colored in grey, 1TS-1Au interactions in orange and 1TS-2Au interactions in yellow. Reprinted from [344].

From the comparison map it can be clearly identified that the 1TS6-1Au has much more additional contacts: from Sav-S112 to S122, especially T114. The increased number of contacts is related to the higher level of flexibility of the 1TS-1Au versus the 1TS-2Au. As mentioned previously, the involvement of two gold cofactors in the 1TS-2Au restricts the flexibility of the cofactors. Both pseudo-TSs structures have few common interactions, only SOD-N8, SOD-I9, SOD-A3 and Sav-N118. Furthermore, most direct interactions of Au with close-lying amino acids are very weak, purely VdW contacts, with few residues. To assess quantitatively these interactions, MMGBSA was performed to obtain indicative energetic values (**Figure C.6**).

As the SOD lid is highly flexible (and disordered in the X-ray structure), we selected only close-lying residues belonging to Sav rather than the SOD region. Accordingly, the amino acids that had a different impact in either of the two TSs were selected for the directed evolution campaign: S112, T114, T115, N118, K121 and S122. Several rounds of directed evolution were performed on the five suggested positions until two chimeric Sav-SOD affording different regioisomers were obtained: favoring the 5-endo-dig product (4:96) and favoring the 6-exo-dig product (99:1) as reflected in **Figure 5.18**.

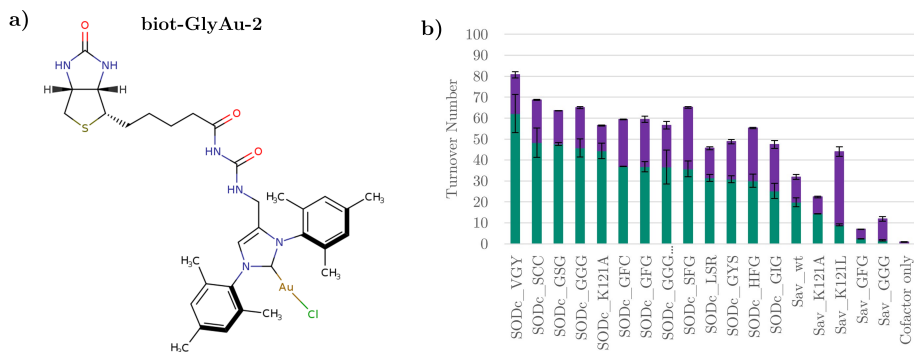


**Figure 5.18:** Directed evolution trajectory of ArM for hydromination reaction. Reprinted from [344].

### 5.3.6. Additional studies: Longer linker

Over the course of this study other approaches were tested in order to improve the ratio of the 5-endo-dig product. The most promising hypothesis was that biotinylated gold complexes with longer linker would be better for the  $\sigma, \pi$ -activation mode. Consequently, *biot-GlyAu-2* was synthesized by incorporating a glycine linker between the biotin and the NHC-gold moiety (**Figure 5.19a**). This resulted in a much higher selectivity for the 5-endo-dig product in most Sav systems (**Figure 5.19b**). We set up to study computationally if the change in regioselectivity towards the 5-endo-dig product was caused by the electronics or by the disposition of the cofactors in Sav.

DFT calculations with *biot-GlyAu-2* complex suggest that there some differences with *biot-Au-2* regarding the TSs and the calculated energy barriers. The Gibbs energy barrier for the 6-exo-dig  $\pi$ -activated system is slightly higher (20.9 Kcal/mol) than the barrier for the 5-endo-dig  $\sigma, \pi$ -activated system (20.6 Kcal/mol). These results indicate that the Gibbs energy barrier difference between two systems is 0.3Kcal/mol, instead of 2Kcal/mol as in *biot-Au-2* case. In terms of reactivity, this 0.3kcal/mol in favor of the 5-endo-digproduct are not significant enough to explain the experimental results. Therefore, dockings and MD simulations were performed to see if *biot-GlyAu-2* positions in a more favorable position in the binding pocket for the  $\sigma, \pi$ -interaction

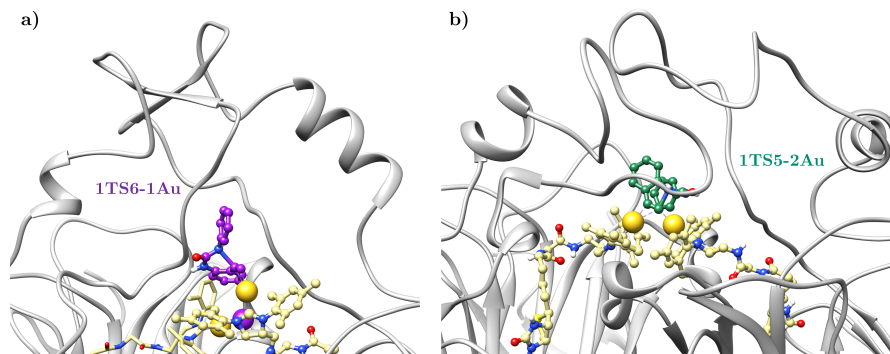


**Figure 5.19:** a) structure *biot-GlyAu-2* b) Experimental results with *biot-GlyAu-2*.

Initial MD simulation of SAV-SOD-K121A in its ground state (bound to *biot-GlyAu-2*) show a conformational rearrangement after the docking. However, the system is stable after 100ns and the overall structure of the protein is not affected drastically by the length of the linker. Docking of 1TS6-1Au show that there are no major changes compared to 1TS6-1Au with *biot-Au-2* cofactor, except that the new Gly part is able to establish more hydrogen bonds and the substrate is positioned in the upper part of the binding site (**Figure 5.20a**).

On the other hand, in 1TS5-2Au calculations, higher dockings score and less clashes are obtained with *biot-GlyAu-2*. The disposition of the cofactors for the  $\sigma, \pi$ -activation is better, there are less clashes with surrounding residues and the distance between Au and C from alkyne achieved in dockings is lower (**Figure 5.20b**). This results reveal that the cofactors can acquire a more favorable position in the binding pocket for the  $\sigma, \pi$ -interactions, which implies that this activation mode is more probable to happen due to the increased length of the linker of *biot-GlyAu-2*.

MD simulations were performed starting from these two dockings (1TS6-1Au and 1TS5-2Au). In the case of 1TS6-1Au, as the cofactor is longer the substrate is more exposed to the solvent and has more ability to move in the binding site. Regarding 1TS5-2Au, the simulations is very stable, 1TS5-2Au remains fixed at the same position in the center of the binding site during all MD, but the SOD region rearranges and moves towards one side.



**Figure 5.20:** Most populated cluster of TS within Sav-SOD-K121A with longer linker cofactor *biot-GlyAu-2*: **a)** 1TS6-1Au and **b)** 1TS5-2Au.

At the end, the route of the longer *biot-GlyAu-2* cofactor was not persuaded experimentally as it was discovered that it was not suitable for directed evolution experiments due to its low sensitivity to mutations. Still, computationally it was determined that the change in regioselectivity towards the 5-endo-dig product could be caused by the more favorable disposition of  $\sigma, \pi$ -interactions in the Sav cavity than the reactivity itself.

### 5.3.7. Conclusions

In this section we have presented how *in silico* modeling of the resulting chimeric HAMase provided insight into the two mechanistic manifolds and revealed close-lying amino acid residues to target by directed evolution, to favor the preferential formation of the anti-Markovnikov product (5-endo-dig) over the Markovnikov product (6-exo-dig).

Specifically, the DFT calculations provided an accurate and reliable insight into the two mechanistic pathways and its energies. Molecular dockings helped understand experimental results and the dual gold catalysis and the MD simulations revealed residue positions for directed evolution that favor the preferential formation of the anti-Markovnikov (5-endo) over the Markovnikov (6-exo) product.

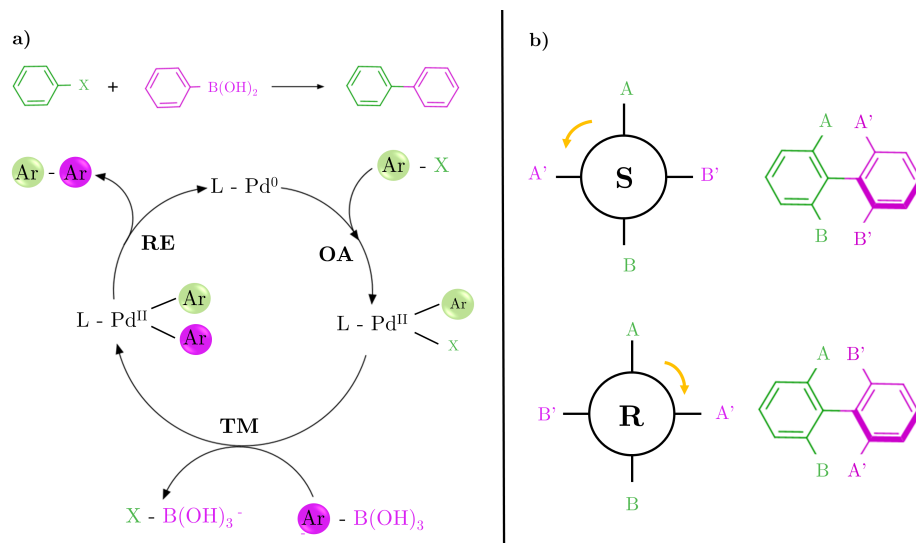
## 5.4. Rationalization of a streptavidin based suzukiase

### 5.4.1. Context and experimental background

Carbon-carbon bond formation reactions are essential for the synthesis of natural products, pharmacological active compounds and agrochemicals.<sup>346,347</sup> Among these are Suzuki–Miyaura cross-coupling (SMC) reactions, one of the most powerful reactions for the formation of C-C bonds that has been widely used to obtain biaryls and aromatic molecules.<sup>348</sup> The SMC reactions are catalyzed by palladium complexes and consist in the carbon-carbon formation between an aryl halide (organic electrophile) and an organoboron compound (organic nucleophile) in the presence of a base. The main advantage of this type of reactions is that they lead C-C coupled products in very high yields, all while operating under mild conditions.<sup>349</sup>

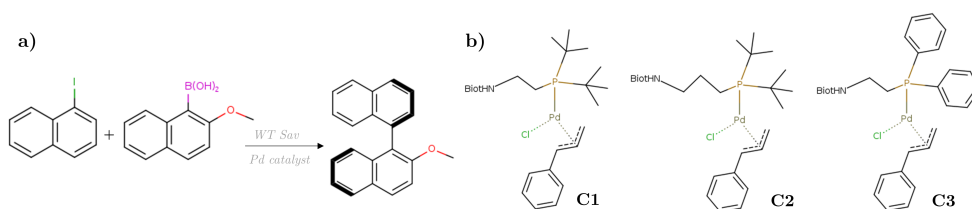
Mechanistic knowledge has been difficult to obtain by experimental means through the years, but computational chemistry has helped to elucidate a clear mechanism. The general mechanism of SMC reactions follows a catalytic cycle composed of three steps, as depicted in (Figure 5.21a). In the first step, the oxidative addition (OA), the organic halide is added to Pd<sup>0</sup>, oxidizing it to Pd<sup>II</sup> and forming the organopalladium intermediate, with Pd bound to both the halide and the organic group. The transmetalation (TM) step is the most characteristic step of the reaction, in which the organic group bound to the boron species is exchanged for the halide in the coordination sphere at the palladium. In this way, the second organic group is transferred to the Pd complex. The last step entails the coupling between the two organic groups and the reduction of the metal to Pd<sup>0</sup>, that why it is known as reductive elimination (RE).<sup>349,350</sup>

As mentioned previously, SMC are commonly used for the synthesis of biaryls, which are defined as two aromatic rings joined through a single C-C bond. Biaryl products possess the characteristic of exhibiting axial chirality in the single C-C bond due to the hindered rotation of this bond. The electronic or steric effects from the different substituents generate a significant rotational barrier around the  $\sigma$  bond, which enables the isolation of two distinct conformers. These are a particular case of enantiomers that receive the name of atropoisomers, as represented in (Figure 5.21b).<sup>351,352</sup>



**Figure 5.21:** a) Catalytic cycle of Suzuki-Miyaura reaction. b) Concept of atropoisomers for biaryls.

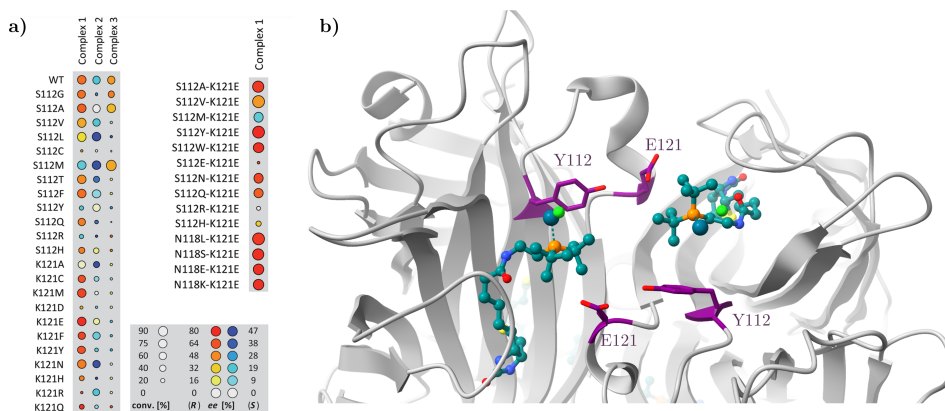
Asymmetric SMC catalysis in water has only been reported twice previously.<sup>353,354</sup> Due to this limited precedent, T. Ward speculated that ArM technology could provide a well-defined second coordination sphere for SMC catalysis. This study herein is based on biotin-sav system that is able to catalyze a Suzuki-Miyaura reaction, in their worlds, an artificial Suzukiase.<sup>355</sup> The design of the ArM consisted in anchoring a modified biotin cofactor with palladium into Sav or avidin scaffold. Specifically, the biotin cofactor is connected with a phosphine ligand of palladium by means of a linker. Five different biotinylated Pd-cofactors were synthesized and evaluated with either Sav or avidin for the synthesis of a biaryl compound, 2-methoxy-1,10-binaphthyl (**Figure 5.22a**). Screening results suggested that most promising combination was wt-Sav with catalyst C1, C2 or C3 from figure **Figure 5.22b**.



**Figure 5.22:** a) Synthesis of 2-methoxy-1,10-binaphthyl. b) Biotinylated cofactors tested.

Previous studies on biotin-Sav ArM suggested that position S112 and K122 could be mutated to optimize the reaction due to its proximity to the active site.<sup>356</sup> Accordingly, the three different biotinylated catalysts were screened against a library of Sav with mutations on S112 and K121. Results of this screening reported in **Figure 5.23a** show that higher conversions are obtained for t-But catalysts (C1 and C2) than Ph catalyst (C3). Focusing only on cofactor C1 and C2, an important feature was elucidated: variation of the length of the linker yields different enantiomeric products.

Mutated systems S112M and K121A with longer biotin-phosphine linker (C2) are the ones that afford the highest enantiomeric excess (ee) for S enantiomer, whereas system K121E with shorter linker (C1) affords the highest ee for R enantiomer. For this reason, mutation K121E was identified as a good candidate for the screening of double mutants with cofactor C1. The results indicate that double mutant S112Y-K121E yields the highest ee (%) and TON for R enantiomer, which was improved up to 90% ee by varying the experimental conditions. This double mutant was crystallized, and the disposition of the metallic cofactor and the mutations can be observed in **Figure 5.23b**



**Figure 5.23:** a) Screening results of cofactors C1,C2,C3 with different mutated Sav systems. Figure from [355]. b) Biotin-binding vestibule of X-ray structure of double mutant S112Y-K121E (PDB = 5cse).

### 5.4.2. Objectives and methodology

In the experimental study they were able to afford an enantioselective artificial Suzukiase for the synthesis of binaphthyls. However, it remains extremely challenging to rationalize or predict how the mutations and the length of the different cofactors leads the preferential synthesis of one atropoisomer over the other. The main objective of this work is to try to explain rationally and qualitatively how different enantioselectivities are obtained using an integrated computational approach. Since it is not feasible to study all mutated systems, this work focused on the three Sav systems with different behaviors: WT, S112M and S112Y-K121E. The experimental results of each Sav system for cofactors C1 and C2 are in **Table 5.3**.

Sav system	Cofactor C1	Cofactor C2
WT	58% R	10% S
S112M	14% S	44% S
S112Y-K121E	90% R	-

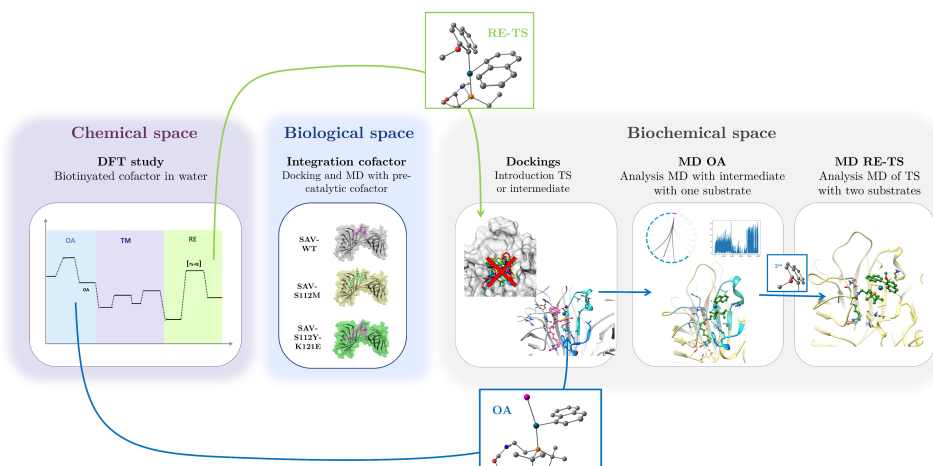
**Table 5.3:** Enantiomeric excess obtained depending on Sav system and cofactor.

The computational protocol that has been followed is consistent with the one described in the introduction of this chapter. Still, certain specifics details for this case and the differences will be further explained for a better understanding of the whole computational overflow. First, a mechanistic DFT study of the full catalytic cycle of the Suzuki-Miyaura reaction was performed in water with biotinylated cofactor C1 using as substrates 1-iodonaphtalene and 2-methoxy-1-naphthaleneboronic acid. Each intermediate and TS of the reaction step was characterized, and the Gibbs barriers were calculated to establish the rate determining step of the reaction. Different TSs leading to either the R or the S atropoisomer were also characterized. All functionals and bases used for calculations have already been described in section 5.2.

The next step was to model each Sav mutated system with precatalytic cofactors C1 or C2 bound, introduced previously by dockings using as input the structure from PDB 5CSE.<sup>316</sup> The structures were refined by performing MD simulations,

which also allowed to assess the stability of each system and the flexibility of each cofactor. At this point, docking calculations were performed to study the influence of the mutations on the rate-limiting TS, both proR-TS and proS-TS were used. Due to the complexity of Suzuki-Miyaura reaction and the great variance of dockings solutions it was not possible to draw any conclusion as will be explained in more details in the results section. As represented in **Figure 5.24**, instead of directly studying the TS of the determining step (RE-TS) (colored in green), another approach was followed (colored in blue). The protocol that was followed consisted on performing docking calculations and MD simulations with the first intermediate of the reaction (OA), which only contains one of the substrates. The intention was to study the behavior of the intermediate OA containing only the first substrate in both proR and proS conformation and then elucidate if the entrance of the second substrate would be feasible.

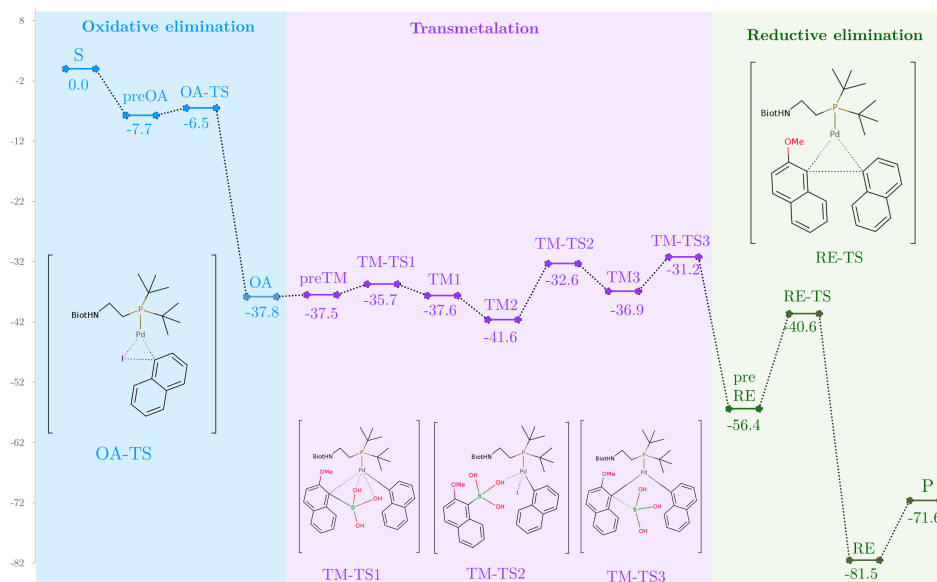
Clustering of intermediates OA was performed using cpptraj and the interactions between the intermediates and Sav were monitored during MD simulations using *getContacts.py* script. To assess the possibility of the entrance of the second substrate, an analysis of all the possible clashes between the second substrate and surrounding residues was performed using Chimera UCSF. This study allowed to discern if there are favored conformations for the entry of the second substrate. Finally, MD simulations of the RE-TS with the two substrates were performed.



**Figure 5.24:** Steps followed during molecular modeling workflow. The initial protocol is indicated in green, while the protocol that was ultimately pursued is delineated in blue.

### 5.4.3. Studying the reaction with DFT

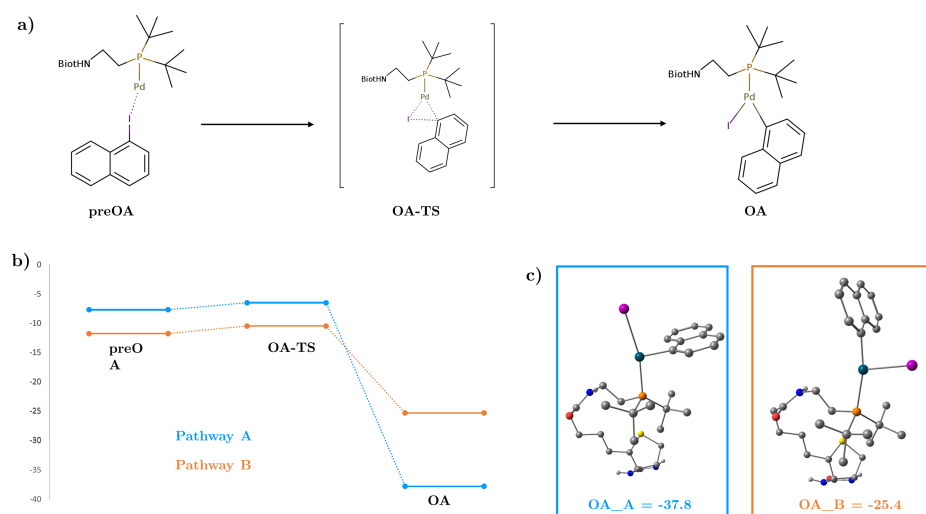
The complete Suzuki-Miyaura reaction was calculated in DFT using full biotinylated cofactor C1. The complete Gibbs energy profile is represented in **Figure 5.25**. Over this section, we will discuss in more detail the results of each of the three steps of the reaction and provide more insights with supplementary calculations that have been conducted for each step.



**Figure 5.25:** Suzuki-Miyaura profile obtained with cofactor C1.

In a Suzuki-Miyaura reaction the first step is the Oxidative Addition (OA). This first stage comprises the formation of the adduct between the aryl halide (1-iodonaphthalene in this case) and the biotinylated palladium catalyst (**Figure 5.26a**). This step leads to the OA intermediate, which contains a Pd in a tricoordinated form that has a T-shape. The addition of the 1-iodonaphthalene can take place via two different pathways depending on whether the iodine (pathway A) or the aryl (pathway B) is positioned in a trans conformation relative to the phosphine. Both pathways A and B exhibit nearly identical Gibbs energy barrier, measuring 1.2 kcal/mol and 1.3 kcal/mol, respectively (**Figure 5.26b**). However, when looking at the products of the two oxidative addition pathways, OA\_B is found to be 12.35 kcal/mol higher than OA\_A (**Figure 5.26c**).

In OA\_A the iodine is trans to the phosphine and the aryl is pointing toward the vacant position, while in OA\_B the aryl is trans to the phosphine. The high trans influence of the aryl destabilizes the OA\_B, as it shows on the Pd-P distance (2.32 Å in OA\_A and 2.43 Å in OA\_B). Therefore, due to the higher stability of OA\_A, for the rest of the reaction it was decided to consider the product of pathway A for the sake of simplicity. The results for this first step are in concordance with DFT study of Patel *et al*, in which they calculated the Gibbs energy profile for the SMC of tetra-ortho-substituted biaryls using different ligands. Their barrier for pathway A is 0.7 kcal/mol, the 0.5 kcal/mol difference with our study could be attributed to the use of different ligands or substrates.<sup>357</sup>



**Figure 5.26:** a) Oxidative addition steps. b) Energetic profile of two possible OA pathways: pathway A in blue and Pathway B in orange. c) Products and energy of OA for pathway A and pathway B.

In addition, the formation of OA\_A can occur via two pathways depending on orientation of the naphthalene group. The difference between the two possible products (OA\_A1 vs OA\_A2) is 0.4 kcal/mol (**Figure 5.27**). Since none of the products exhibit any unfavorable contacts, the energy gap between them is small, suggesting that both products are equally feasible in terms of energy. To assess the energetic cost of the transition between both conformations, the rotation around  $Pd - C_{aryl}$  bond was analyzed. Results of this rotation scan, depicted at figure **Figure 5.27**, reveal that the rotation of this bond has a barrier of 19.0 kcal/mol. Despite the minimal energetic difference between OA\_A1 and

OA\_A2, the transition between the two products is hindered by a barrier that is too high for reaction conditions. Therefore, it is assumed that once one conformation is adopted, there will be no transition towards the other conformation. This differentiation sets up the first layer for the enantioselectivity of the reaction. For simplicity, calculations for the rest of the reaction were performed for OA\_A1, which is slightly more stable, and it is assumed that values for the other pathway will be similar.

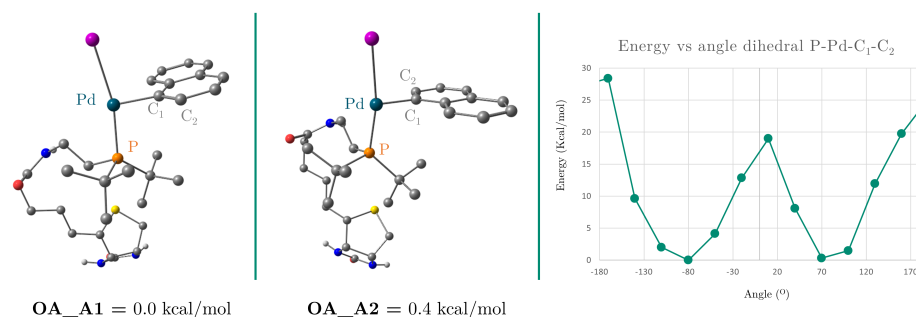


Figure 5.27: Torsion scan between product OA\_A1 and OA\_A2.

In the transmetalation (TM) step, the second organic substrate, containing a boronate group, is introduced and initially coordinates to the Pd through one hydroxyl. In a three-step process, the iodine and the boric acid are removed sequentially, while the bond between the entering substrate (2-methoxy-naphthalene) and the Pd are formed (**Figure 5.28**). Overall, the TM step involves three TSs, each characterized by Gibbs energy barriers of 1.8, 8.8 and 5.7 kcal/mol, respectively. The product of the TM has both organic substrates bound to the Pd catalyst (preRE).

It is important to note that in this step the entrance of the second substrate can occur in two different orientations, which will be referred as *cis* and *trans* from now on. In the *cis* conformation the two non-metallated rings of the naphthalene ligands are in the same side of the plane defined by the phosphine, palladium and metallic carbon of the methoxy-naphthalene. In the *trans* conformation they are in opposite sites. Both conformations in the preRE step are represented in **Figure 5.29** However, for simplicity, only the *cis* conformation was only considered in the energy profile of **Figure 5.25**. The stability of each conformation will be studied at the last step of the reaction (RE).

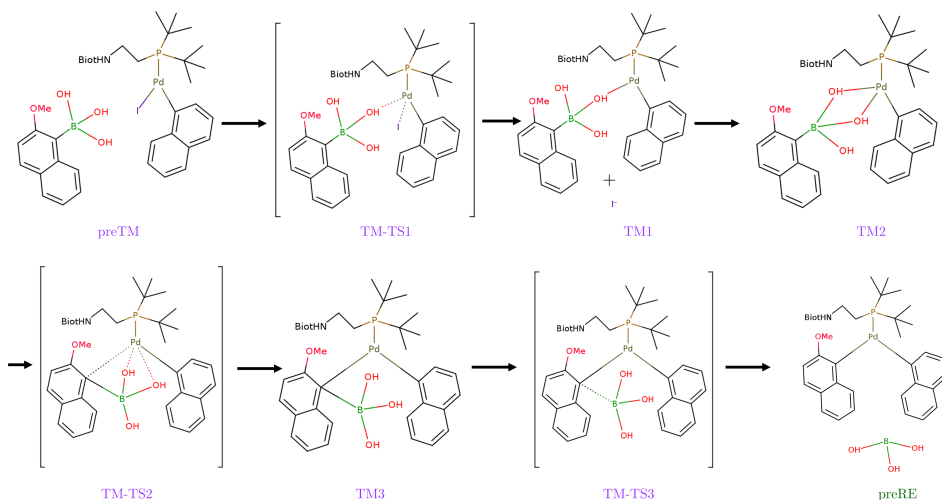


Figure 5.28: Transmetalation step: intermediates and TS.

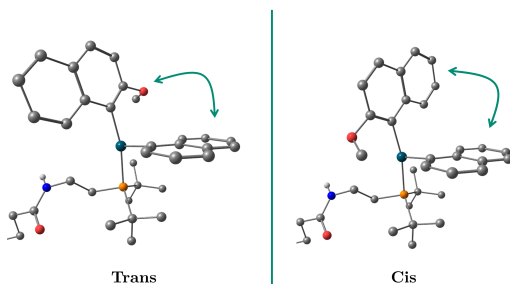
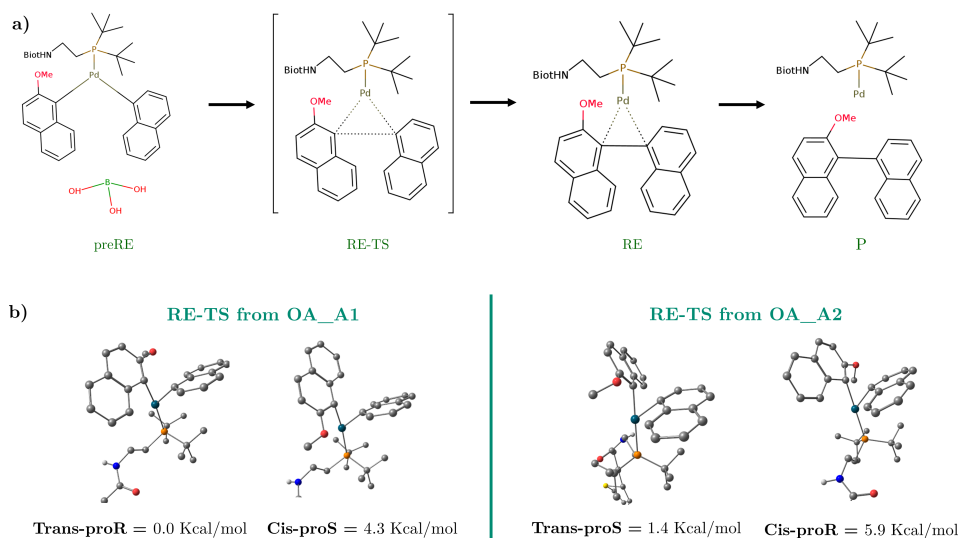


Figure 5.29: Representation of *cis* and *trans* conformations in preRE step.

The last step of the reaction is the reductive elimination (RE), in which the C-C bond between the two substrates is formed leading to the biaryl product (**Figure 5.30a**). The Gibbs barrier of this step is 15.8 kcal/mol, making it the highest barrier of the full reaction and therefore the rate-determining step (RDS). At the end of the cycle,  $\text{Pd}^{\text{II}}$  returns to initial state  $\text{Pd}^0$  and the final product of RE is the biaryl, which can exist as either R or S atropoisomer form. **Figure 5.25** represents the profile that leads to the R enantiomer, but additional TSs of possible enantiomers were studied in this step, including both *cis* and *trans* conformations.

All RE-TS determined are characterized by a T-shape arrangement between Pd and the two substrates. As discussed before (**Figure 5.27**) there are two possible



**Figure 5.30:** a) Reductive elimination steps ab) RE-TS in *cis* and *trans* conformation and their energies

orientations for the coordinated naphthyl (OA\_A1 and OA\_A2), each one can give rise to a *cis* and a *trans* conformation as hindered rotation between both aryls ligands is assumed. For both OA\_A1 and OA\_A2, the *cis* products are less stable, there is a difference of 4.3-4.6 kcal/mol compared to the *trans* product (**Figure 5.30b**). This difference is hypothesized to be caused by the steric hindrance between the naphthalene rings in the *cis* conformation. Considering only the *trans* conformations, from these calculations it can be established that the product from OA\_A1 leads preferentially to the R atropoisomer (proR), whereas the form A\_A2 leads preferentially to the S atropoisomer (proS). The energetic difference between the TSs in conformation proR and proS enantiomers is only 1.4 kcal/mol (**Table 5.4**). Calculations were performed for both catalytic cofactor C1 and C2, but the same tendencies were observed in both cases (**Table C.7**).

Cofactor C1

	RE-TS from OA_A1		RE-TS from OA_A2
Trans-proR	0.0	Trans-proS	1.4
Cis-proS	4.3	Cis-proR	5.9

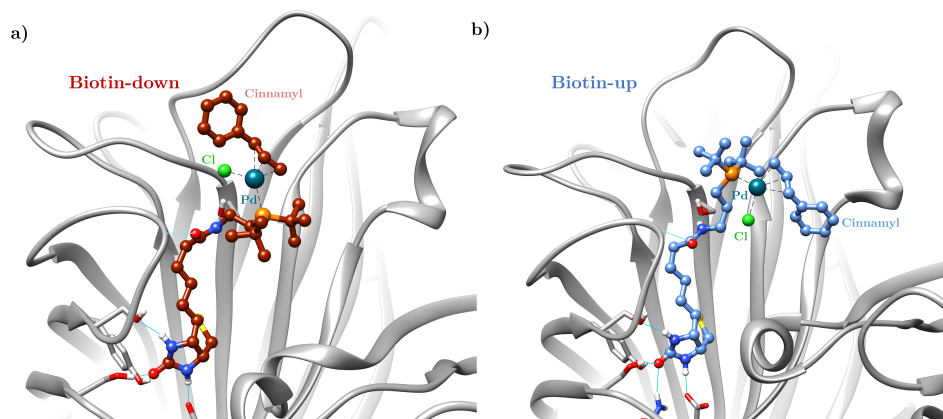
**Table 5.4:** Gibbs energies of RE-TS with cofactor C1.

#### 5.4.4. Molecular modeling of the protein system

Molecular docking simulations are conducted to introduce both precatalytic cofactors, C1 (short) and C2 (long), into the three different Sav systems (WT, S112M and S112Y-K121E). These precatalytic cofactors contain the biotinylated Pd catalyst saturated with chloride and a cinnamyl moiety, as previously depicted in (Figure 5.22). Dockings solutions reveal flexibility in the position of the cinnamyl and chloride moieties, attributed to the rotation of the bonds in the linker region, which is more pronounced when linker is longer in the C2 cofactor. Consequently, MD simulations are performed using as input the best docking solutions, which were selected according to scoring function and similarity to crystal structure.

The aim of the MD simulations is to explore the conformational space of the protein and assess the behavior of the two catalytic cofactors before the Suzuki-Miyaura reaction. Significant differences are observed among distinct Sav systems and cofactors. Considering the protein structure, it does not suffer significant changes across the MD simulations, except some mobility in the flanking loops between the  $\beta$ -barrels, which could be relevant for the catalysis. The biotin part of precatalytic cofactor is mainly fixed due to the presence of several hydrogen bond interactions with Sav residues. Contrarily, the linker part of the cofactor and the cinnamyl moiety display a remarkable flexibility by exploring the available conformational space extensively. This observation is supported by the x-ray structure of Sav-S112Y-K121E with the precatalytic cofactor, the full structure is resolved except for the cinnamyl region, which density is not determined probably due to its high flexibility.

From these simulations it can be determined that the cofactors in the resting state of the reaction can acquire two very different conformations, which will be referenced from now on as: *biotin-up* and *biotin-down* (Figure 5.31). In the *biotin-up* conformation, the  $P(t\text{-}But)_2$  of the biotinylated cofactor is in the upper region of the binding site with the chloride and cinnamyl moiety facing down towards the inside of the active site. Contrarily, in the *biotin-down* disposition the  $P(t\text{-}But)_2$  of the biotinylated cofactor is in the lower site of the binding site, under the Pd atom, and the chloride and cinnamyl moiety, which are facing the solvent. Depending on the system or the longitude of the cofactors the tendency to be on the *biotin-up* or the *biotin-down* disposition changes.



**Figure 5.31:** Conformations of precatalytic cofactor in MD: **a)** *biotin-down* **b)** *biotin-up*.

	WT	S112M	Double
C1-short	Down ↓	↑ Up (main) / Down ↓	Down ↓
C2-long	↑ Up (main) / Down ↓	Up ↑	-

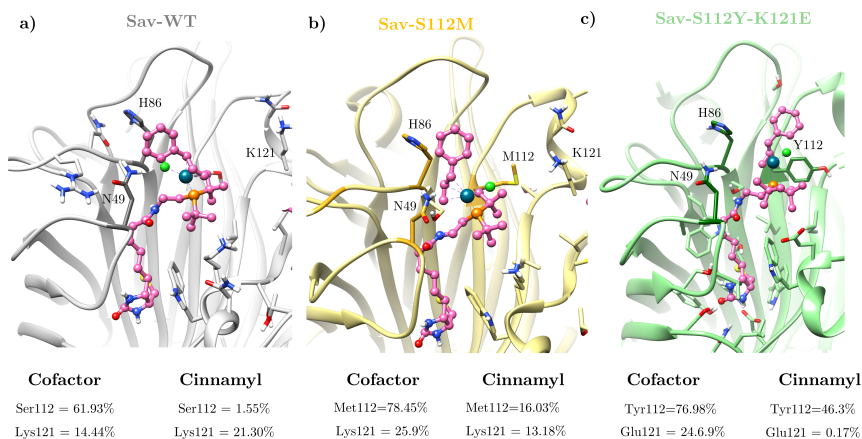
**Table 5.5:** Conformations of precatalytic cofactor observed in MD simulations for each Sav system.

MD simulations with the **C1 precatalytic cofactor** display a tendency for the *biotin-down* conformation, with the cinnamyl and the chloride moieties facing the solvent (**Table 5.5**). The short length of the linker seems to favor the *biotin-down* conformation, without the possibility of rotating towards the *biotin-up* conformation. However, differences are observed between the three systems related to the interactions of the cofactor and close-by mutated residues.

In the Sav-WT system, the *biotin-down* conformation is maintained during all MD simulations and the cinnamyl moiety swings between the two sides of Sav. It either interacts with loop containing Asn49 and His87 or Lys121 on the opposite site (**Figure 5.32a**). In Sav-S112M, the main disposition of cofactor C1 during the MD is *biotin-down*. However, at some points the biotinylated cofactor can rotate and acquire the *biotin-up* conformation. The capability to acquire the up conformation may be related to the mutation that has been introduced, close-by residue Ser112 has been mutated to Met, a much longer and hydrophobic

residue. Contacts during MD between the precatalytic cofactor C1 are higher with Met than in Sav-WT with Ser (**Figure 5.32a-b**). Met112 mainly interacts with the P(t-But)<sub>2</sub> from the cofactor and interacts less with cinnamyl. The rest of interactions of the cofactor and the swinging movement between regions is also observed in this system.

In the case of Sav-S112Y-K121E, only the *biotin-down* conformation is observed in the MD simulations, there is no rotation of the biotin toward the *biotin-up* conformation at any moment. The mutation of Ser112 to aromatic Tyr increases considerably the interaction with the cinnamyl, maintaining the precatalytic cofactor in the center of the biotin vestibule. Mutation Lys121 to Glu leads to a decrease of the interaction with cinnamyl moiety, which also causes cinnamyl to remain in the center of the binding site (**Figure 5.32c**). From these analysis it can be concluded that both mutations favor the biotin down conformation.

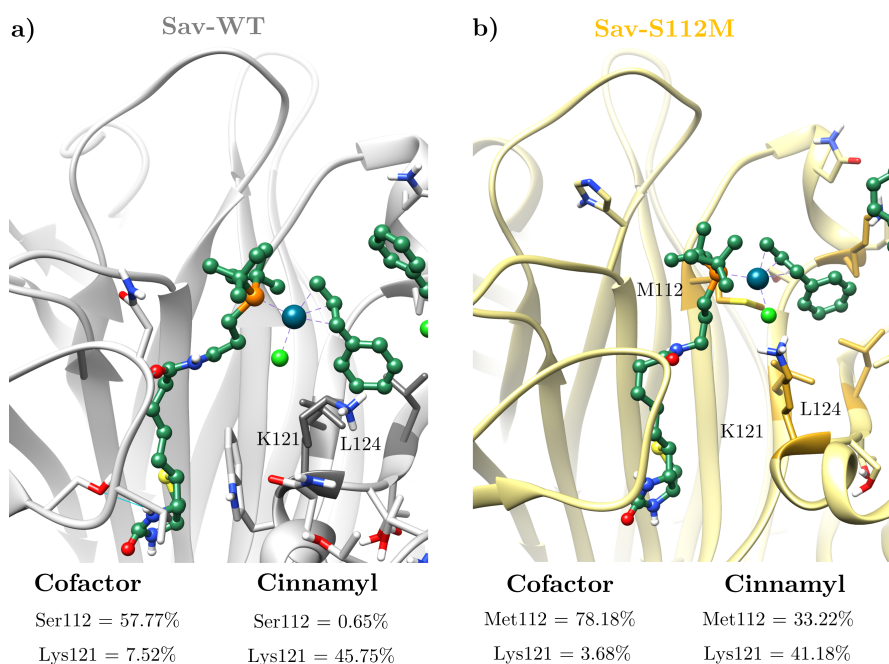


**Figure 5.32:** Most representative cluster and frequency of contacts during MD simulation with precatalytic cofactor C1 of three Sav systems: a) WT b) S112M and c) S112Y-K121E.

MD simulations with the longer **precatalytic cofactor C2** tend to acquire the *biotin-up* conformation, the cofactor can switch from the *biotin-down* to the *biotin-up* (**Table 4 5.4**). The additional carbon atom before the P(t-But)<sub>2</sub> group provides an increased capacity for the cofactor to rotate. However, there are some differences between Sav-WT and Sav-S112M due to the mutation introduced. In Sav-WT, the prevalent conformation is *biotin-up*, but in different monomers of Sav the transition from *biotin-up* to *biotin-down* and vice-versa can be observed, indicating that both conformations are inter-convertible. Contrarily,

in Sav-S112M, when the *biotin-up* conformation is acquired, it remains in this disposition and makes no transition to *biotin-down*. Once the cofactor has rotated toward the *biotin-up* conformation, Met112 remains under the cofactor interacting with cinnamyl hindering the transition to *biotin-down*. Interaction analysis values indicates that generally Met112 interacts more with the cofactor, especially with the cinnamyl compared to the WT (**Figure 5.33**). In both systems, when in the *biotin-up* disposition, the cinnamyl interacts with a hydrophobic patch found in the interface of the dimer containing Lys121, Val123 and Leu124 from both monomers.

This initial MD study reveals that the precatalytic cofactors can acquire two different dispositions that depend both on the length of the cofactor and the mutations introduced in the system. It can be concluded that the relevance of this study relies on the fact that the different dispositions of the cofactor can affect the entrance of the two substrates.



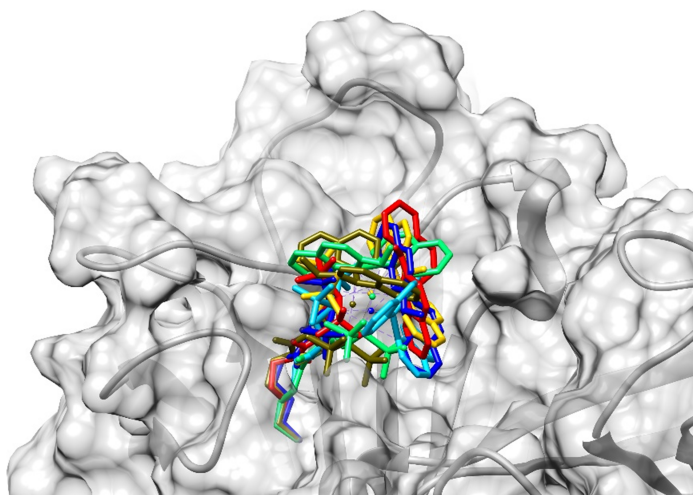
**Figure 5.33:** Most representative cluster and frequency of contacts during MD simulation with precatalytic cofactor C2 of a) WT and b) S112M-C2.

### 5.4.5. Docking and MD simulation of TS or intermediates

#### 5.4.5.1 Molecular dockings of TS-RE

The next step consisted in performing docking calculations of the pseudo-TS of the reductive elimination (RE) step for both proR and proS states. In studies concerning the reactivity of an asymmetric ArM, it is common that the pseudo-TS of choice is from the chirality determining step of the reaction. In this case, RE-TS was selected as the previous DFT study indicated that it is both the rate and chirality determining step. Therefore, our focus on studying RE-TS inside the protein scaffold. The objective here is to assess which residues stabilized each pseudo-TS using molecular dockings. proS and proR RE-TS characterized by DFT were docked in the three Sav systems using as input the most populated cluster of previous MD simulations.

Similarly to the observations from the docking of the resting state, these calculations display a lot of variability in the disposition of the pseudo-TSs. This does not allow to see differences between proR and proS configurations or between any of the Sav systems (**Figure 5.34**). RE-TS contains both substrates and are completely exposed to the solvent, which allows a lot of variation in the results. Still, it is clearly observed that when the cofactor is longer, there is even more variability due to the possibility of the *biotin-up* and *biotin-down* disposition.



**Figure 5.34:** Variability in docking solutions obtained by both proR and proS RE-TS.

Despite the group uses a quite standard multi-scale approach for modeling ArM, each system requires specific needs. For this study and at this point, another approach was taken, instead of studying RE-TS, we figured out that another step of the reaction could be vital: the oxidative addition (OA). The idea is to analyze how the location of the first substrate affects the binding of the second substrate.

#### 5.4.6. Dockings of OA intermediates

To recap, the DFT study determined that the OA intermediate can have two conformations only involving the first substrate of the SMC reaction, OA\_A1 or OA\_A2 (**Figure 5.27**). As discussed before, each of these OA intermediates will lead to a *trans* conformation (more stable than *cis*) of RE-TS where the two substrates are present, *trans*-proR and *trans*-proS depending if they come from pathway OA\_A1 or OA\_A2. Therefore, the OA intermediate is the first layer of enantioselectivity. First, molecular docking of two possible OA intermediates, proR-OA (OA\_A1) and proS-OA (OA\_A2) will be performed, followed by MD simulations. The localization of the intermediate OA (with only the first naphthalene substrate) will allow to later superpose RE-TS and study if the entrance of the second substrate would be possible. Instead of performing calculations for all five selected cases, simulations are performed for WT-C1, S112M-C2 and S112Y-121E, as they have higher ee differences.

Docking calculations are performed for both proR-OA and proS-OA for all Sav systems using as input the most populated cluster of MD simulations with the precatalytic cofactor. The conformation of the precatalytic cofactor is maintained fixed and the naphthalene substrate and iodine are docked, allowing rotation. In all cases, WT-C1, S112M-C2 and S112Y-K121E-C1, no relevant differences are found between proR and proS docking results, the scorings values and the disposition of the substrates are very similar (**Table C.8**). This means that the probability of the first substrate to acquire proR or proS conformation is the same at this stage of the reaction. Results of WT-C1 are represented in **Figure 5.35** as an example. The biotin vestibule has been divided in three regions that will be referred from now on as region A, B and C. Region A includes residues Ala86 and polar residues Asn49 or His87, region B includes mutant residue Ser112 and Lys121 and hydrophobic residues Leu110,124. Region C in the other monomer includes Trp241, Leu245,245 and mutant Lys242. In the three cases, the substrate is found interacting with region A of Sav, mainly with residue 112.

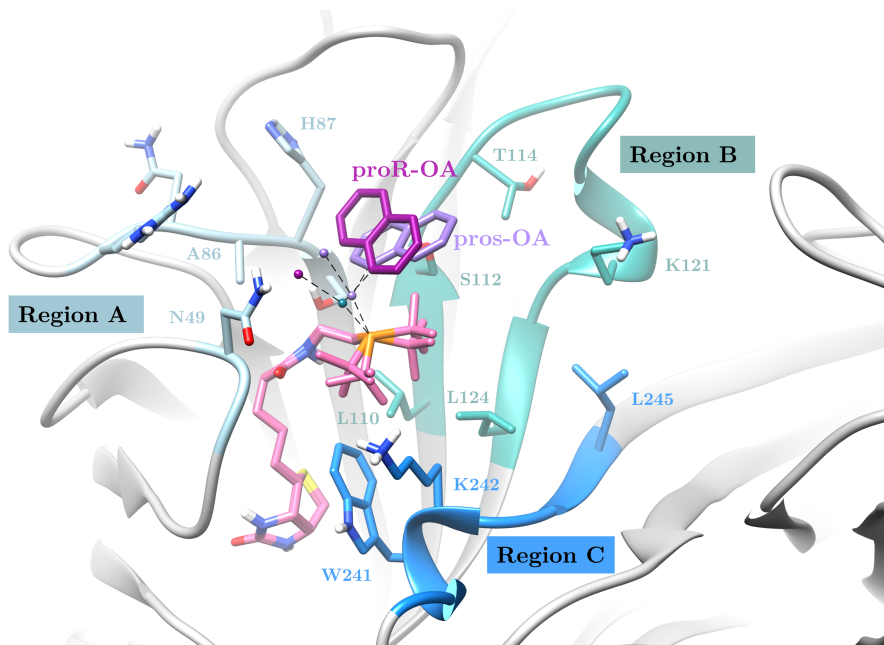


Figure 5.35: Sav biotin-binding vestibule and WT-C1 results for proS and proR OA.

#### 5.4.6.1 MD simulations of OA intermediates

MD simulations, initiated from the docking positions, are performed for the three Sav systems with proR and proS OA. Clustering of the intermediate OA is performed using cpptraj to find the most common conformation during the MD simulation. The prevalence of each cluster and the tendency of the proR-OA and proS-OA intermediates to acquire the *biotin-up* or *biotin-down* conformation are analyzed by cluster and summarized in **Table 5.6**.

An additional study is conducted to see if the entrance of the second substrate would be feasible by looking at the possible clashes that would happen between the second substrate and the protein if it were to bind. To achieve this, the RE-TS structure is superposed to the OA intermediate during the MD simulation to extrapolate the position of the second substrate and the clashes are monitored with Chimera UCSF. Mean value of clashes in each cluster are displayed in **Table 5.6**.

	proR				proS			
	% Cluster	Disposition	Region	Clashes	% Cluster	Disposition	Region	Clashes
S112M C2	41.8%	Down	B	5.8	87.2%	Up	B-C	2.29
	17.3%	Down	B		8.9%	Up	B-C	
	36%	Up	B-C	9.75				
WT C1	70%	Down	B	15.26	71.4%	Up	B-C	12.50
	28%	Down	A	5.2	23%	Down	A	16.84
S112Y- K121E C1	45%	Down	A	11.33	52.5%	Down	B	28.59
	35.8%	Down	B	15.4	27.4%	Down	A	
					15.9%	Up	B-C	6.7

**Table 5.6:** Analysis results of MD simulations with proS and proR OA.

An analysis from **Table 5.6** yields two observations: **1)** The proR-OA intermediate tends to acquire the *biotin-down* conformation preferably, whereas the proS-OA acquires the *biotin-up* disposition **2)** The entrance of the second substrate is more probable for the proR-OA in the *biotin-down* conformation and for the proS-OA in the *biotin-up* conformation. However, as we will explain in this section, these tendencies also depend on the mutations introduced on the system and the length of the cofactors. Now, we will comment a bit further these two aspects for each case and relate it to the experimental results.

In the case of **S112M-C2**, as seen in the MD simulations with the precatalytic cofactor, when the cofactor is longer the tendency to acquire *biotin-up* conformation increases probably due to the ability to rotate thanks to the extra carbon in the linker. Still, different behaviors are obtained for proR-OA and proS-OA systems. In the proR-OA, during 59.1% of the simulation the cofactor remains in the *biotin-down* conformation, interacting with Met112, Thr114 and Lys121 from region A (**Figure 5.36a**). However, the cofactor slightly rotates, and the substrate situates on top of Met112, acquiring the *biotin up* conformation during the rest of simulation (36%). On the other hand, in the case of proS-OA, the cofactor almost immediately acquires the *biotin-up* conformation and is maintained during all simulation. In the *biotin-down* conformation the substrate

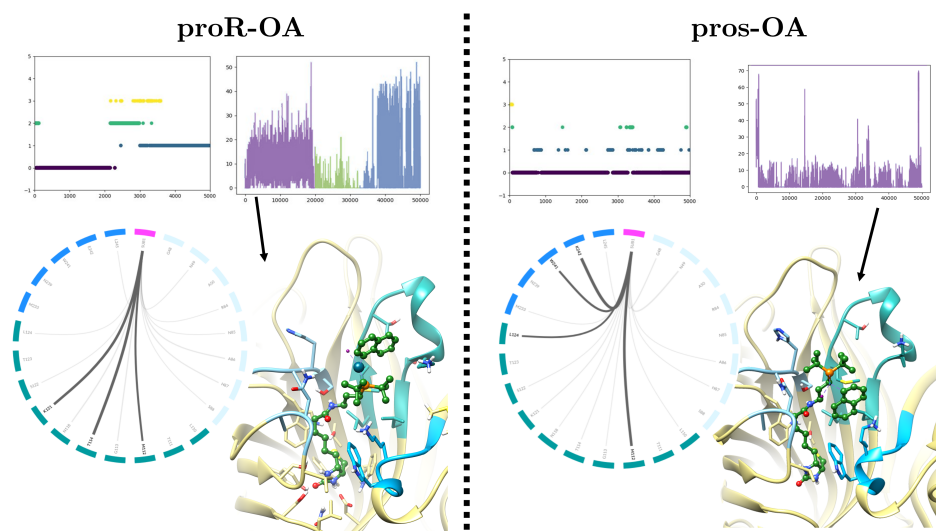
is interacting with region B-C, naphthalene is found between Met112 and Lys242 and surrounded by hydrophobic residues, Trp241, Leu110-124 and Lys121. The mutated Met112 changes its rotamer position slightly when acquiring the *biotin-up* conformation, which difficulties acquiring the *biotin-down* conformation again (Figure 5.36b).

The access of the second substrate is easier in the proS *biotin-up* conformation, with fewer clashes (2.3) with surrounding residues compared to the proR-OA (5.8-9.8), suggesting that more S product should be expected (Table 5.6). In the proR-OA, the entrance of the second substrate has less clashes in the *biotin-down* conformation and as the clashes are low in this conformation (5.8) for most of the simulation (59%), the R product should also be expected, but in lesser quantities than the S product. This is in concordance with the experimental results, 44% ee S, indicating the predominant formation of the S atropoisomer, though with a minor presence of the R atropoisomer. It can be hypothesized that both the long cofactor and mutation S112M favor the *biotin-up* conformation, which for the proS form the entrance of the second substrate is more feasible as there are less clashes. The length allows better rotation of the cofactors and Met112 mutation keeps the naphthalene substrate in the *biotin-up* form interacting with Lys121 and hydrophobic patch between monomers. Consequently, the entrance of the second substrate and reactivity is much more probable in the proS-OA form.

In **WT-C1** and **S112Y-K121E-C1**, when using the cofactor with the short linker, similarities can be observed in the behavior of the proR-OA form, but difference in the proS-OA. In both systems, for proR-OA, the *biotin-down* conformation is observed all through the MD simulation and, depending on the cluster, the naphthalene substrate is either interacting with region A or with region B.

For proS-OA, in the WT, the MD simulations proS-OA starts from the *biotin-down* conformation and remains in this disposition the first 23% of simulation, interacting with region A. However, it rotates and acquires the *biotin-up* conformation for remaining 71.4%, in which is interacting it is interacting with region B-C, mainly with Lys121, Leu110-124 and Ser112 (Figure C.9).

In S112Y-K121E, an opposite behavior is observed. ProS-OA immediately acquires the *biotin up* conformation, in which the naphthalene substrate is in the lower region of the binding site interacting with Trp241 and mutated Glu242 from region B-C. However, these interactions are not maintained as in the WT or



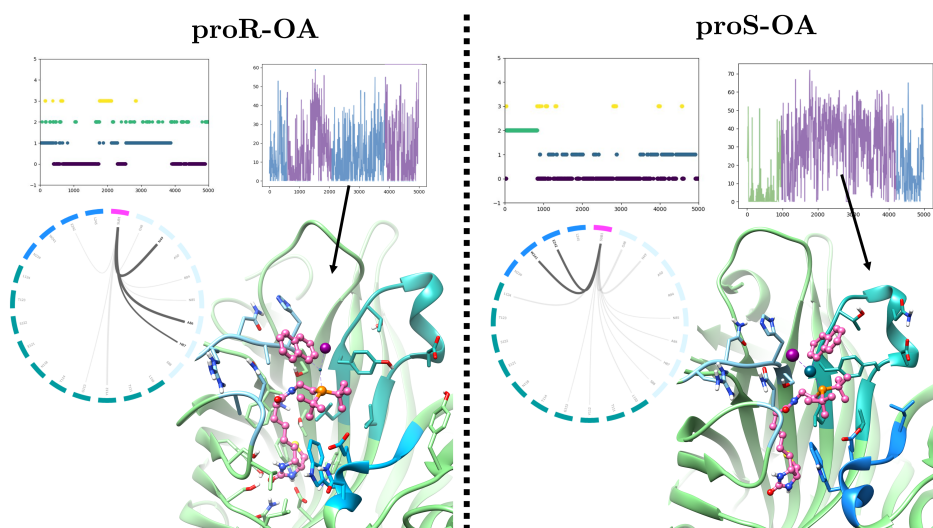
**Figure 5.36:** Results of Sav-S112M-C2 for proR and proS OA. Panels: 1) Cluster of OA during MD simulation 2) Clashes between Sav and second substrate during MD 3) Structure of OA cluster with less clashes and 4) Interaction map between OA and Sav regions.

S112M form, in which the naphthalene is interacting with Lys121 or Met112. Therefore, it rotates again and for the rest of the simulation the disposition is *biotin-down*, facing the solvent. In this *biotin-down* conformation there are two clusters. In the first cluster, the substrate is interacting with region B, while in the second cluster the substrate is interacting also with region A, mainly with Tyr112. Tyr112 is precisely one of the mutated residues, it is observed that this Tyr directly interacts with (t-But)<sub>2</sub> of cofactor. Probably, this interaction keeps the cofactor in the *biotin-down* conformation causing of the inability to rotate towards the *biotin-up* conformation. Therefore, mutations S112Y and K121E favor the *biotin-down* conformation.

Regarding the entrance of the second substrate for the WT, the proS-OA acquires the *biotin-up* conformation, but in this conformation there are more clashes (12.5) for the entrance of the second substrate than the proR in the *biotin-down* conformation, in which during 28% of the simulation there are 5.2 clashes. Therefore, in the WT the entrance of the second substrate is more probable in the proR-OA conformation, explaining the experimental 58% ee R.

In the case of S112Y-K121E it can be observed that the entrance of the second substrate is only possible in the *biotin-up* conformation as the clashes are low

(6.7), but this conformation is only observed in 16% of simulation. In the predominant *biotin-down* conformation there are too many several clashes with all surrounding residues (28.6). Experimentally, this system affords 90% ee R, which is demonstrated by these simulations in which only in the 16% of the simulations the substrate would be able to enter and obtain the S form. For the rest of the simulation of the proS the second substrate would not be able to enter and afford the S product due to the high clashes. It can clearly be stated that the mutation of Tyr112 and Glu121 do not favor OA in the *biotin-up* conformation, which is the one that is favored in the proS-OA.



**Figure 5.37:** Results of Sav-S112Y-K121E-C1 for proR and proS OA. Panels: 1) Cluster of OA during MD simulation 2) Clashes between Sav and second substrate during MD 3) Structure of OA cluster with less clashes and 4) Interaction map between OA and Sav regions.

#### 5.4.6.2 MD simulations with RE-TS

MD simulations of the three systems with TS-RE in proR and proS conformations were performed to see stability of final TS in protein environment. As input the cluster from previous MD simulations with less clashes is used and the second substrate is incorporated in the disposition of the TS-RE.

MD simulations with Sav-S112M show that this system favors the *biotin-up* conformation. During most of the simulation of the proR, it remains in the *biotin-down* disposition interacting region A, however, towards the end it

acquires the *biotin-up* conformation and interacts mainly with Lys121-242. This indicates that the longer cofactor and Met allows rotating and acquiring the other conformation. In the case of proS-TS the simulation already starts from *biotin-up* conformation and is also interacting with Lys121-242.

In the WT and Sav-S112Y-K12E, both TS, proR and proS remain in the same conformation that the MD simulation starts from, *biotin-down* for proR and *biotin-up* from proS. In both cases, the TS move from region B towards region A to interact mainly with Asn47. In proR from Sav-S112Y-K12E interacts with Tyr112, stabilizing the TS, while is the case of WT Ser does not interact. In the proS form from WT interaction with Lys242 is maintained during all simulations stabilizing the TS, whereas in the double mutant there is no interaction with E242. These differences in interactions may be related to the favor of R in double mutant, as proR is more stabilized by Tyr and proS in the WT.

#### 5.4.7. Conclusions

The findings of this study unravel that the enantiomeric differences found experimentally can be rationalized by the disposition of the first intermediate and the entrance of the second substrate.

DFT calculations reveal the overall Suzuki-Miyaura mechanism and indicate that the oxidative addition is the first layer for the enantiomerism and the reductive elimination is the rate and enantiomeric determining step. MD simulations of the oxidative addition step's with both proR and proS products shed light on how their dispositions determine the entrance of the second product and its regioselectivity.

From this work it can be concluded that the combination of DFT calculations, molecular dockings and MD simulations can be used to rationalize ArM. Furthermore it highlights that standard protocols for the design of ArM have to be adapted to the catalytic reaction in question to understand the whole concept.



## CHAPTER 6

# Finding metal binding sites to design a new ArM

Metal ions play essential roles in numerous proteins. Finding the location of metal binding sites is crucial. X-ray structures generally provide with a clear identification of the most stable metal binding site, but many of them are resolved without the cofactor. Assessing the position of the metal in the biological scaffolds with other experimental methods (EPR, MS-ESI, etc.) can be time-consuming, expensive, and often remain elusive.<sup>358</sup> Like in other fields, computational predictions could offer a valuable alternative. Metal binding predictors are rather rare in the current computational chemistry landscape. Few metal binding prediction tools have been developed over the years, based on sequence, structure patterns or a combination of both, such as IonCom, MIB2, MetalS2 and MetalPredator, between others.<sup>203,359–361</sup> In fact, the availability of accurate computational predictors for metal binding site in proteins extends beyond unraveling natural roles of metalloproteins. Indeed, one particularly promising application is the use of metal binding prediction tools for the design or redesign of de novo ArM.<sup>362</sup>

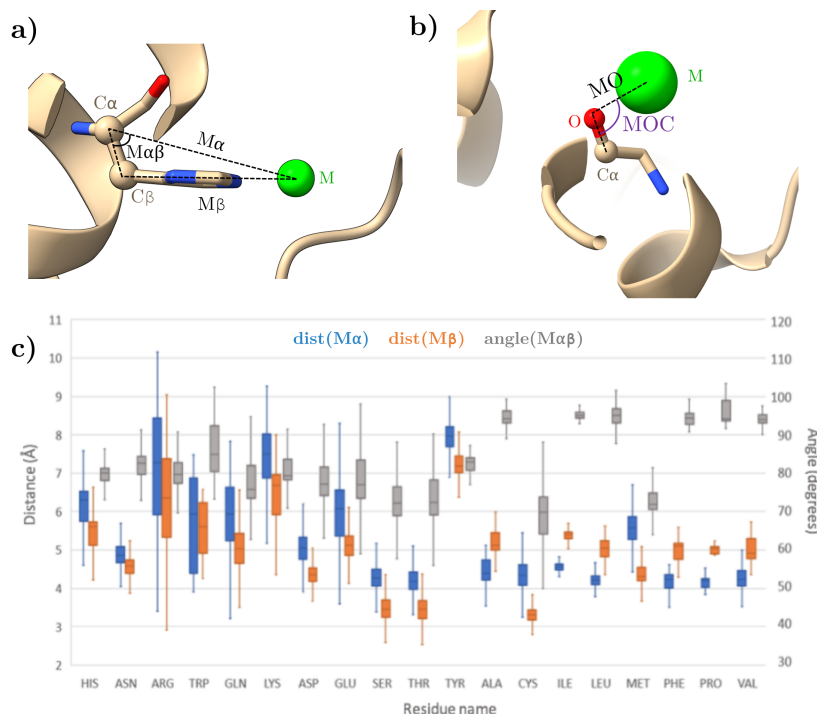
In this chapter we present BioMetAll<sup>294</sup>, a tool for predicting metal binding sites based only on structure information and backbone preorganization, that allows finding regions that could be mutated to become metal binding sites. Combined with molecular modeling tools, we used BioMetAll to engineer new ArM by finding regions to incorporate a metal ion and obtain novel functions. BioMetAll was briefly introduced in section 4.2, but now we will provide a broader overview, focusing on its possible applications to the design of novel ArM.

## 6.1. Overview of BioMetAll

BioMetAll is based on the premise that metal binding sites can be predicted taking into account only the preorganization of the backbone. This hypothesis is based upon a previous study in which a statistical analysis on 400 Fe-containing proteins revealed that a distance of 7Å between Fe and  $\alpha$ -carbons was an enough to correctly detected possible coordinating residues.<sup>363</sup> Similarly, a few years later, a script was developed to screen possible metal binding sites to narrow down the regions for molecular docking calculations. This script probed the protein space to look for regions in which the center of mass of the three possible coordinating residues was within a distance of 3.5Å.<sup>364</sup> Both works lead to the idea that instead of focusing on the side-chains disposition for coordination, backbone geometry could contain enough information to detect metal binding sites. Specifically, geometric descriptors related to backbone (distance and angles involving  $\alpha$ -carbon and  $\beta$ -carbon) can be good indicators of the backbone preorganization and can be used for the prediction of metal binding sites.

Assuming this hypothesis, a statistical analysis was carried out to obtain the geometric descriptors for all protein structures containing metals available in the database MetalPDB.<sup>365</sup> As explained before in section 4.2.1, the three geometric descriptors that were analyzed for side-chain coordination are: the distance from the metal to the  $\alpha$ -carbon ( $M\alpha$ ), the distance between the metal and the  $\beta$ -carbon ( $M\beta$ ), and the angle between the metal, the  $\alpha$ -carbon, and the  $\beta$ -carbon ( $M\alpha\beta$ ). For backbone coordination, only oxygen coordination was taken into account as nitrogen coordination was minimal in the statistical study. Two descriptors were considered for backbone coordination: the distance from the metal to the backbone oxygen ( $MO$ ) and the angle between the metal, the backbone oxygen, and the backbone carbon ( $MOC$ ). The two possible geometric features are represented in **Figure 6.1a,b**.

In total, 170.00 pdb entries were analyzed with a pyChimera<sup>366</sup> script, resulting in total of 500.000 metal binding site analysis. The results from the statistical analysis are summarized in **Figure 6.1c**.



**Figure 6.1:** Geometric features analyzed for all the structures of MetalPDB. **a)** Geometric features considered for coordination bonds with a side-chain donor. **b)** Geometric features considered for coordination bonds with a backbone oxygen donor. **c)** Bar plot for the different features ( $M\alpha$  in blue,  $M\beta$  in orange and  $M\alpha\beta$  in grey) studied in the statistical analysis. Represented values for each residue.

BioMetAll was developed in Python 3.7 language with only two required dependencies. The workflow is divided into four sequential steps (**Figure 6.1**).

First, the PDB of the input protein is parsed and only the coordinates of the  $\alpha$ -carbons,  $\beta$ -carbons, backbone oxygens and carbons are stored. Then, the protein is embedded in a sphere of equidistant probes. Probes at a distance lower than 1 Å from the backbone are removed. At the second step, for each of the probes the software calculates if any of the surrounding residues fulfill the three geometrical criteria established at the statistical analysis. At this point the user can specify which residues, motif, mutations should be considered and if the backbone atoms should be considered for coordination.

At the third step, each of the probes has assigned a list of possible coordinating residues. Each possible metal binding site is defined by all the probes that share

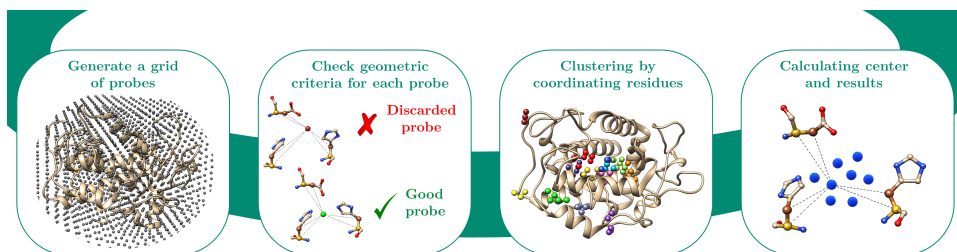


Figure 6.2: Schematic representation of BioMetAll workflow.

the same possible residue coordinators. Therefore, all the probes that share the same list of coordinating residues are grouped together constituting a binding site. Finally, the center of all the probes of the binding site is calculated, but in reality, the metal could bind to any of the probes of the binding site. Each binding site is ordered by the number of probes and results are saved in pdb format.

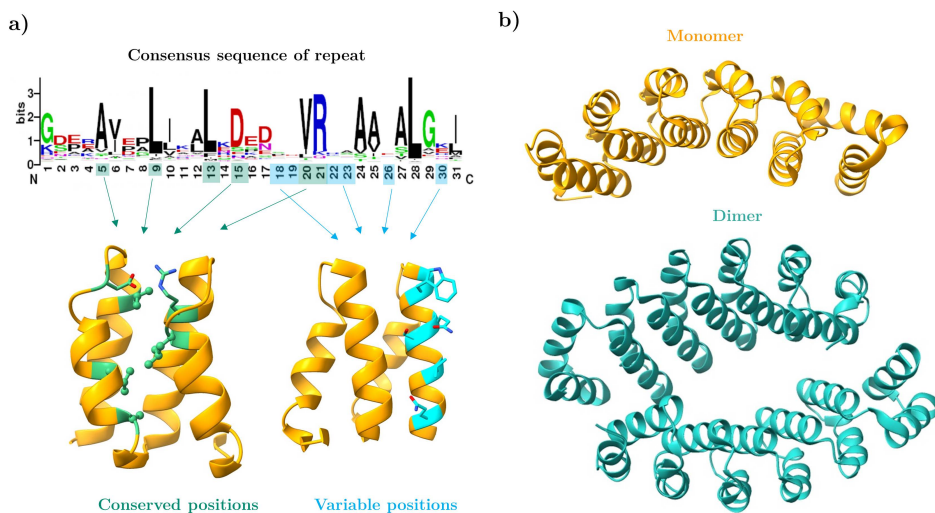
One of the unique characteristics of BioMetAll is its ability to predict potential mutations to complete a motif and generate a metal-binding site. This capability was tested as a tool for designing new ArM based on metals ions by revisiting a study of T. Ward *et. al.*<sup>367</sup> The protein 6-phosphogluconolactonase from *Mycobacterium smegmatis* (PDB 3OC6)<sup>368</sup> was repurposed from a hydrolase to peroxidase with  $\text{CuSO}_4$  and introduction of mutation Asn131Asp. However, in the X-ray structure, coordination was observed only with His67-His104.

BioMetAll was employed to find the most suitable mutation when looking for the motif His-His and mutation Asp or Glu. BioMetAll provided with a list of possible mutations and docking calculations were performed to assess the predictive accuracy of BioMetAll. Combination of BioMetAll and docking suggested mutation Tyr69/Asp-Glu as the best possible binding site. Low BioMetAll scoring was obtained for Asp131 and no docking calculation resulted in a good coordination with Asp13, as observed in the experimental structure. In conclusion, this study showcases the potential of combining BioMetAll with docking calculations to predict convenient side-chain rearrangements for mutations and design of ArM.

Upon these promising results, we set up to design a *de novo* ArM using BioMetAll to predict possible mutations, which could be then validated experimentally in collaboration with the group of Pierre Mahy.

## 6.2. Applicative case of design of ArM: $\alpha$ -Rep

The scaffold of choice for the ArM is *de novo* protein  $\alpha$ -Rep. Although this system was previously mentioned in section 4.2.7, its most relevant characteristics are emphasized again here.  $\alpha$ -Rep proteins are constituted by a variable number of  $\alpha$ -helical repeats that give rise a protein with a solenoid shape. The four-helix repeated motif derives from a sub-class of HEAT-like repeat thermostable proteins. The 31 residue repeated motif contains six hyper-variable positions that can be mutated without destabilizing the protein's structure. This feature is attributed to a series of conserved residues that maintain the proper folding of the protein (**Figure 6.3a**). The variable positions are located at the outside surface of the second helix of the motif.  $\alpha$ -Rep proteins can be found in the form of monomers or in its dimeric form (**Figure 6.3b**).<sup>135</sup>



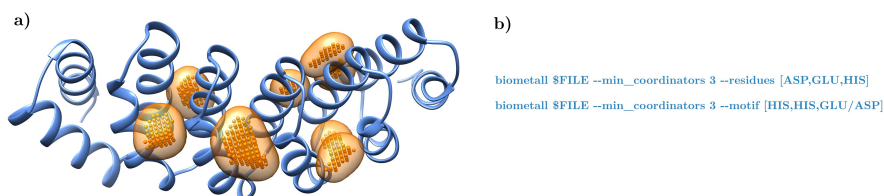
**Figure 6.3:** Structure  $\alpha$ -Rep repeats. Monomer and dimer conformation of  $\alpha$ -Rep.

These proteins are efficiently expressed, soluble, very stable and folded, all characteristics that are desired for the design of ArM. The possibility of mutations that do not alter the structure and all mentioned properties of  $\alpha$ -Rep make them very good candidates for ArM. In the following sections we will go through the process of the design of an ArM based on  $\alpha$ -Rep with the facial two-histidine one-carboxylate motif (FTM) using BioMetAll.

The FTM triad is a recurring motif within mononuclear iron enzymes that has been characterized structurally in more than 30 enzymes. The FTM motif consists two His and one carboxylate (either Asp or Glu), which are positioned in one triangular face of the metal octahedron. The other three opposite sites are available for binding substrates, dioxygen or cofactors. This arrangement provides the enzyme a flexibility which tunes the reactivity of the metal and allows to perform a wide range of oxidative reactions.<sup>369,370</sup>

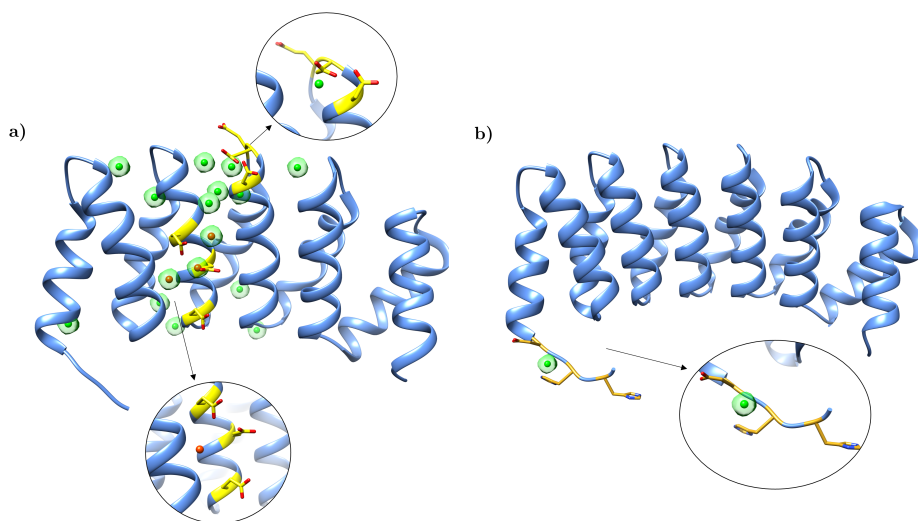
### 6.2.1. Initial screening of crystallographic monomeric structures

First, an initial search for natural metal binding sites in the monomer form of  $\alpha$ -Rep was performed. Monomeric structures were prepared with UCSF Chimera from the fourteen available PDB structures, from which nine were monomers and five were dimers (Table D.1). Initial volume analysis with pyKVFinder revealed small binding sites of an average volume of  $17.74 \text{ \AA}^3$  that could possibly bind metals (Figure 6.4a). In order to see if a natural metal binding site could be found in these regions, BioMetAll was applied using the two different modes with a minimum of three coordinators: searching for FTM residues = [HIS, GLU, ASP] or searching for the specific motif [HIS, HIS, ASP/GLU] (Figure 6.4b).



**Figure 6.4:** a) Structure of monomeric  $\alpha$ -Rep with volume in between repeats. b) Two search modes.

Regarding the first search mode, most structures have metal binding sites involving Glu or Asp residues, which are mainly found in the most external parts of the  $\alpha$ -helices, either in the superior part or the inferior part of the helices. In some structures, metal binding sites were found in center of the helices as represented in orange in Figure 6.5a. Fewer structures contained metal binding sites involving His. These could only be found in systems which contained a His-tag in their N-terminal or in C-terminal regions. When looking for the specific FTM motif, metal binding site were found only on two systems, which contain two His from the His-tag and a close-by Asp residues (Figure 6.5b).



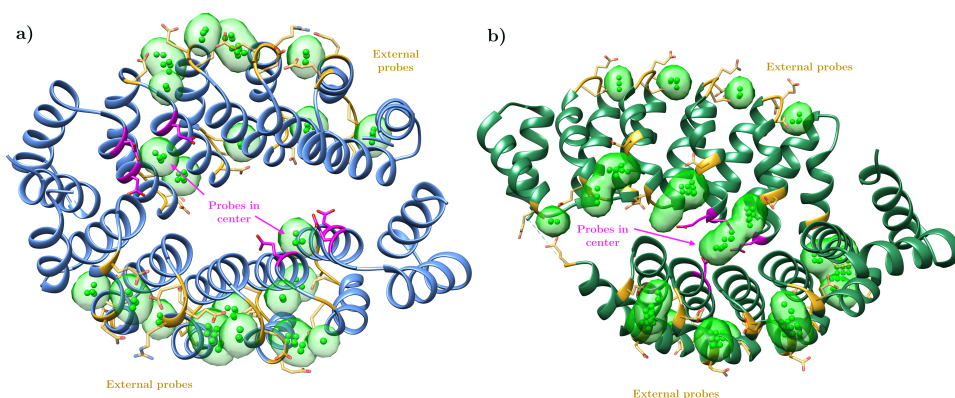
**Figure 6.5:** BioMetAll results for: a) PDB 4xpj in search mode 1 b) PDB 3ltj in search mode 2.

In all cases the metal binding sites are either completely exposed to the solvent or are found in loop regions, which usually implies a lot of not desired flexibility. Consequently, they would not be good candidates for binding a metal for the design of an ArM. No metal binding sites were found between the  $\alpha$ -helices, which as mentioned previously they would have been more appropriate. As all the mutable residues are facing the external face of the protein, no further calculations were performed with BioMetAll to find mutations that would fit better the FTM motif. In light of these results, we jumped to perform calculations with the dimeric form of  $\alpha$ -Rep.

### 6.2.2. Initial screening of crystallographic dimeric structures

As mentioned before, from the fourteen available structures, five correspond to dimeric structures. There are two different types of dimers involving  $\alpha$ -Rep: **1)** A3\_A3: two  $\alpha$ -Rep monomers parallel to each other and **2)** A3\_bGFPD:  $\alpha$ -Rep domain perpendicular to a bGFPD domain, which also contains  $\alpha$ -helices. The former one has a large cavity of  $2045.3 \text{ \AA}^3$ , while the latter has two large ones (with volumes of  $1129.25 \text{ \AA}^3$  and  $505.44 \text{ \AA}^3$ ). The same BioMetAll calculations were carried out for these systems. Results reveal regions in the interface between the two  $\alpha$ -Rep domains that could be metal-binding sites.

In the case of A3\_A3, most metal binding sites can be found either in the most external part of the  $\alpha$ -helices, which are not very promising (**Figure 6.6a** in yellow). However, there is a set of solutions involving two Glu and one Asp (138,165 and 169), which are found in the center of the interface between the two  $\alpha$ -Rep domains (**Figure 6.6a** in pink). These are not exposed to the solvent and are found in stable  $\alpha$ -helices, indicating that they could be good metal binding sites. Regarding the case of A3\_bGFPD, possible metal binding sites are found in the extremes of the  $\alpha$ -helices (**Figure 6.6b** in yellow). Due to the disposition of the two domains, in the interface between them, there is only one possible metal binding site that involves Asp and Glu residues from both  $\alpha$ -Rep domains, specifically, 53,83 and 275 (**Figure 6.6b** in yellow).



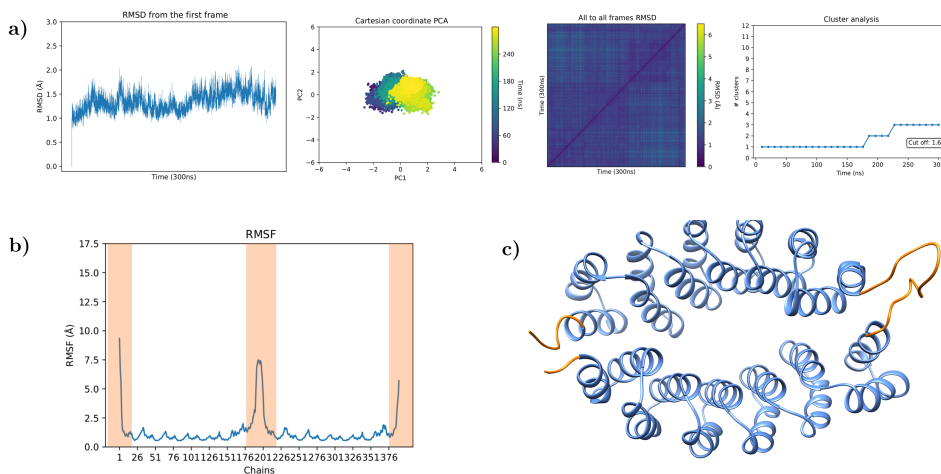
**Figure 6.6:** BioMetAll results for dimers **a)** A3\_A3 and **b)** A3\_bGFPD. BioMetAll probes colored in green. Possible coordinating residues found in the extremes of the  $\alpha$ -helices colored in red and coordinating residues found in the interface of the domain in pink.

From the two different types of systems, the best possible candidate for mutations and designing an ArM would be A3\_A3. The best metal binding sites involve residues that are found in the center of interface of the two  $\alpha$ -Rep domains, with enough space for possible substrates to interact. Furthermore, the three residues suggested by BioMetAll correspond to variable positions, which could be mutated to obtain the FTM motif. On the other hand, the best candidate from A3\_bGFPD does not have a clear binding site with residues around that could accommodate a substrate, it is very exposed to the solvent. The metal-binding sites from A3\_A3 is potentially good, but they do not fulfill the FTM motif.

Therefore, we set up to look for residues that could be mutated to HIS in BioMetAll and obtain the HIS, HIS, GLU/ASP. To perform the following calculations with the mutations, we decided to use A3\_A3 structure containing a loop connecting both subunits to ensure its stability, which had been previously modeled in the group for a previous work.<sup>163</sup>

### 6.2.3. MD simulation of dimer A3\_A3

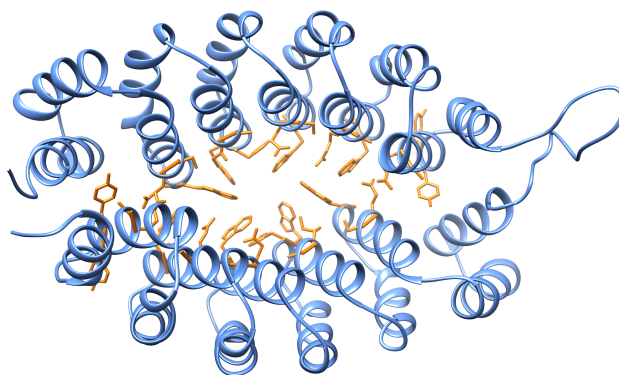
The original form of A3\_A3 that contains the connecting loop is covalently attached to a MnTPP, consequently this was removed and Cys26 was mutated to its original Tyr.<sup>163</sup> In order to have a reliable model, MD simulations of this system were performed prior to final BioMetAll mutation calculations. Classical MD simulation of 300ns was performed, with convergence indicators represented in (Figure 6.7) displaying the stability of the system. The overall structure of the system is maintained and only the C-terminal, C-terminal regions and the loop connecting both domains have a higher flexibility, which is expected for these loop regions. Clustering is performed to obtain the most representative cluster to use as input for following BioMetAll calculations



**Figure 6.7:** Analysis of MD simulations apo  $\alpha$ -Rep: rmsd, PCA, all-to-all RMSD, clustering, RMSF and most representative cluster.

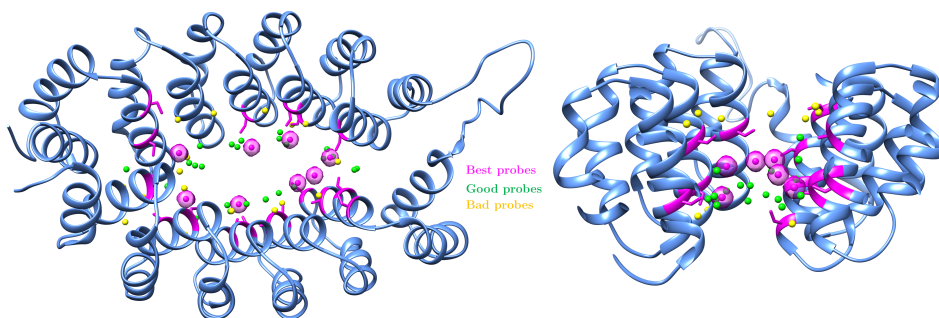
#### 6.2.4. Mutation calculations

To find a metal binding site that fulfills the FTM motif, two different strategies were used with the mutation tool from BioMetAll. The first strategy that was tested consisted of finding Glu/Asp residues and complete them with two His mutations. Once the calculations were performed, results were filtered in order to take into account only the hyper-variable from A3\_A3. In total, 17 solutions were obtained, but nine were discarded as they were not exactly in the cavity. This strategy presented the problem that the search was limited to the presence of Asp or Glu in the natural system. Therefore, a second strategy was carried out, all variable positions from A3\_A3 were mutated to Asp and all other residues were deleted. With this strategy all hyper-variable positions can be mutated to His and search with BioMetAll for the FTM motif (**Figure 6.8**).



**Figure 6.8:** Set-up for A3\_A3 BioMetAll calculations. In orange residues to be mutated to Asp and in blue residues to be deleted.

With this second strategy 56 possible binding sites were found by BioMetAll. The results are analyzed visually, taking into account the location of the probes and the coordinating residues in the cavity (**Figure 6.9**). Represented in yellow are the probes that are considered bad and were discarded as they would be very exposed to the solvent. In green are represented probes that would be potentially good, as they are found inside the binding site shielded from the solvent. In total 8 metal binding sites were selected as the best ones and are represented in pink. These metal binding sites are all found inside the cavity and in the center, which could facilitate the entrance of the substrate of the reaction. With the best 8 results, docking calculations were performed.



**Figure 6.9:** BioMetAll results for FTM motif in A3\_A3. Best probes represented in pink, good probes in green and bad probes in yellow.

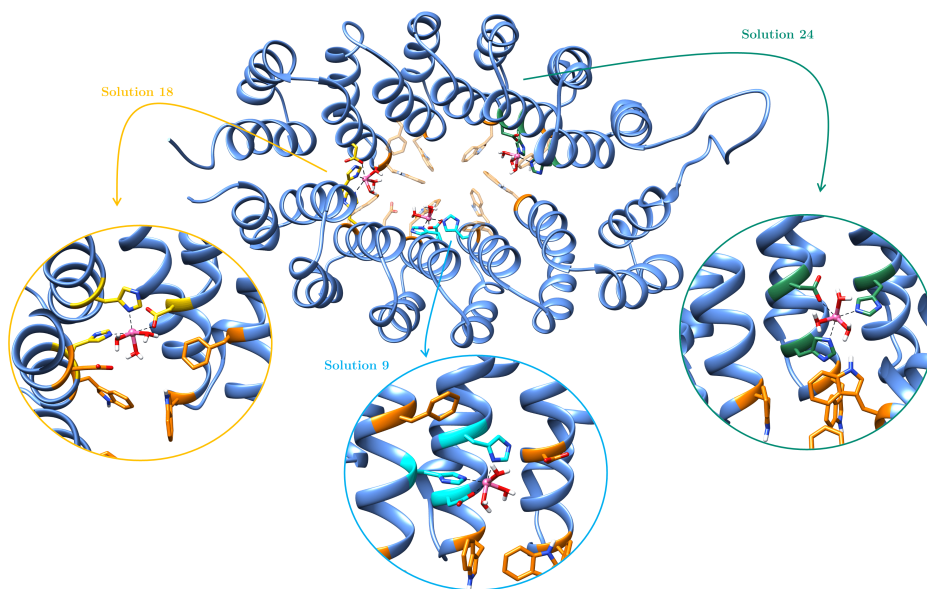
### 6.2.5. Docking calculations with metal

Docking calculations were performed with a Manganese ion with three water molecules in facial disposition and three vacant sites for coordination with the FTM motif. Residues involved in the FTM motif were mutated and calculations were performed with GOLD, using a radius of 8 Å and free rotation for possible coordinators. As scoring function an updated function for goldscore for metals is used, which considers coordination explicitly. Additional rotamers were also added in some cases for surrounding residues. Docking results are summarized in Table 6.1.

Biometall position	Score Docking	Asp residue	His mutations	Coordination
7	48.06	283	252,256	Yes
9	39.31	84	116,88	Yes
14	42.19	349	26,353	Yes
15	55.49	150	225,154	Yes
16	58.99	147	115,143	Yes
18	33.88	57	26,54	Yes
24	42.97	256	225,252	Yes
30	59.21	225	150,229	Yes

**Table 6.1:** Summary of proposed mutations and docking results.

In all cases, a complete coordination is observed between all three residues and high score are obtained. In general, very good complementarities are obtained between the metal and the coordinator residues of the FTM motif. The three solutions highlighted in Table 6.1 were selected as the best due to the characteristics that are beneficial for ArM design. All three metal binding sites are situated near a Trp patch, this region would block the entrance of possible substrates from that side of the system. This means that possible substrates would enter by the other site, which corresponds to the face of the metal with three vacant sites. This can be observed mainly in solution 24. On solutions number 9 and 18, there is also an Asp residue close-by, which could be involved in activating certain molecules and Phe residues which could also be of use for binding of certain aromatic substrates. Lastly, in case 18, coordinating residues involve both domains, which could be beneficial to maintain the fold of the protein. Other cases were situated at the opposite site of Trp patch, or the vacant site of metals were facing regions in which the binding of the substrate would not be possible. These three solutions were proposed to experimental group of Mahy and are being tested in the lab at the moment. Consequently, these results have not been yet validated experimentally at the moment.



**Figure 6.10:** Best solutions from docking for the design of ArM.

## 6.3. Conclusions

In this chapter we have showed an efficient predictor of metal-binding sites based on a few geometric descriptors of the conformation of the backbone. BioMetAll shows great potential to identify metal binding sites even with incomplete coordination spheres and allows to predict mutations that could be necessary to generate new metal-binding sites in a protein, for example, for building new ArM.

To showcase this ability of BioMetAll we have presented an applicative study based on *de novo*  $\alpha$ -Rep proteins. BioMetAll has been able to detect possible metal binding sites that fulfill the FTM motif that have been validated with molecular dockings. Short-term future goals include an experimental validation of the computational predictions of BioMetAll.



## CHAPTER 7

### Other works

In this chapter, we give a concise overview of different side projects related to metallic proteins which have not been included in the main body of the thesis. The three projects have been developed in collaboration with industrial companies or experimental research groups, where molecular modeling played a crucial role in the investigations. Detailed explanations are provided in Appendix A.

The first project was the core of a secondment at agrochemical company Syngenta (Jealott's Hill, UK, 2 months). Syngenta and our group are part of an H2020 EU consortium aiming to develop safe and selective pesticides (RISE project, CYPTOX). The objective of the secondment was to set a **Free Energy Perturbation** (FEP) protocol in a newly acquired software by the company to estimate the binding energies of herbicides toward the metalloprotein 4-Hydroxyphenylpyruvate Dioxygenase. Promising results were obtained with 12 ligands, resulting in a mean error of 0.61 kcal/mol and a correlation  $R^2$  of 0.38.

The second project is a collaborative effort with Perez's group from the Universidad de Huelva. They achieved **direct benzene hydroxylation** with  $O_2$  using copper complexes in the presence of ascorbic acid. By applying DFT, we identified the active species and elucidated the possible mechanism of action, revealing similarities with natural binuclear copper monooxygenases.

Finally, in collaboration with Mahy's group at the Université of Paris-Saclay, molecular modeling was applied to study an ArM based on cobalt hemoprotein to produce  $H_2$ . In the absence of X-ray data, molecular docking calculations combined with MD simulations shed light on the flexibility and positioning of the Co-cofactor while also allowing the detection of possible axial ligands.



## CHAPTER 8

# Conclusions

This PhD aimed to address significant challenges associated with the computational modeling of metalloproteins and apply newly acquired knowledge and skills to real problems in biomedicine and biotechnology. The primary focus of this work was placed on two fields of research: the study of heme-binding proteins and the design of Artificial Metalloenzymes (ArM). This main objective has been achieved by employing comprehensive multiscale modeling workflows and developing user-friendly software. These efforts led to the elucidation of metal recognition processes, encompassing both heme binding processes and rationalization of ArM. For each field of study, the specific conclusions that can be drawn are as follows:

**Heme binding processes:** A combination of molecular modeling tools and software developed during the PhD has been applied to study different aspects of the heme's binding mechanism to its receptors.

- A computational framework incorporating enhanced sampling technique GaMD and molecular docking optimized for metalloligands has been built to unravel and gain insight into intricate natural heme binding mechanisms. Its application proved valuable in elucidating the distinct heme mechanism of Hemophore HasA from *S. marcescens* and *Y. pestis*. This study emphasizes the importance of extensively exploring both apo and holo states of heme-binding proteins to understand the binding process but also underscores the relevance of the pre-organization within the heme sites. The latter set a precedent to develop software for predicting heme-binding sites based on the preorganization of the binding site.

- HemeFinder has been developed to predict natural heme-binding sites and explore the potential to design new ArM based on heme and porphyrin. This tool relies on structural and physico-chemical characteristics of heme sites, including shape, residue composition, and three geometric descriptors. The application of HemeFinder on heme carrier protein 1 (HCP1), a transmembrane protein involved in heme recruitment, has given the first structural evidence of the viability of Heme-HCP1 complexes and unveiled possible heme pathways. Furthermore, its application in designing ArMs based on de novo  $\alpha$ -Rep proteins appears promising with the identification of convenient mutation to create sites with additional histidine. The benchmark results show that HemeFinder's speed does not compromise its performance. In the future, its speed will allow for rapid screening of possible heme proteins, uncovering novel heme-binding proteins, or identifying new scaffolds for ArM design.

**Design of Artificial Metalloenzymes:** In the pursuit of designing ArM, our research has applied multiscale strategies combining quantum mechanics, molecular dockings, and MD simulations. This strategy has proved to give excellent results in rationalizing and guiding the design of two ArMs.

- In the case of an ArM for the hydroamination of alkynes, DFT and docking calculations have allowed us to rationalize how the catalytic mechanism could involve single or dual gold-based catalysis and its possible impact on regiospecificity. Integrating MD simulations with pseudo-transition states previously optimized by DFT was pivotal in unraveling the most critical residues for catalysis. These insights empowered experimentalists to perform directed evolution, culminating in improved regioselectivity of the ArM.
- In the context of the Suzukiase ArM, molecular modeling has helped rationalize the influence of mutations and linker cofactor length on the enantiomeric outcome. This work emphasizes the importance of adapting the computational protocol to the specific catalytic requirements of the ArM. In this ArM, the critical catalytic step is not the rate-determining step; the examination of the initial intermediate, containing only the first substrate in the protein scaffold, has shed light on how its disposition can influence the entrance of the second substrate.

Finally, the **BioMetAll** tool has been used to design a novel ArM based on *de novo* scaffold  $\alpha$ -Rep. Integrating BioMetAll and molecular dockings has led to the creation an ArM that could bind a metallic ion with the introduced catalytic FTM motif (His-His-Asp/Glu). This new ArM is being tested experimentally to see its ability to bind metallic ions and perform catalysis. Further work includes designing which catalytic activity and substrate could be used in this system.

In conclusion, this thesis brings novel insights to the bioinorganic field and lays the foundation for further advancements in our general understanding of metalloproteins, natural or artificial.



## APPENDIX A

### Other works

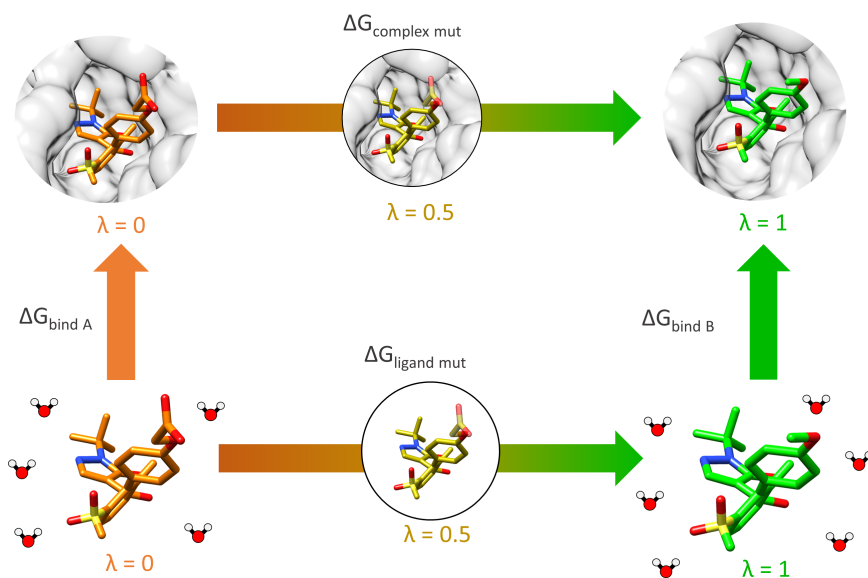
In this appendix, we explain other investigations developed during this thesis. The primary emphasis is put on our contributions, details of the experimental contributions are not explained, but short summaries are provided. Collaborations with industrial partners have confidentiality issues. Therefore, no structures or specific details are provided. In other cases, full articles provide all the details:

- Borrego, E, Tiessler-Sala, L, Lázaro, J.J., Caballero, A, Pérez, P.J. & Lledós, A. Direct Benzene Hydroxylation with Dioxygen Induced by Copper Complexes: Uncovering the Active Species by DFT Calculations. *Organometallics* **41** 1892–1904 (2022).
- Udry, G. A. O., Tiessler-Sala, L., Pugliese, E., Urvoas, A., Halime, Z., Maréchal, J.-D., Mahy, J.-P. & Ricoux, R. Photocatalytic Hydrogen Production and Carbon Dioxide Reduction Catalyzed by an Artificial Cobalt Hemoprotein *International Journal of Molecular Sciences* **23**, 14640 (2022).

## A.1. Free energy studies on metalloproteins

In the agrochemical industry, there is a keen interest in accelerating the design of new pesticides through computational tools. A promising approach is calculating the free energy binding to identify the most promising pesticide candidates and reduce the number that needs to be synthesized or tested. This work aims to perform Free Energy Perturbation (FEP) calculations to estimate the binding affinity of pesticides to a target of interest, all while testing the Flare software performance when applied metalloenzymes.

FEP calculates the free energy of binding of ligand A relative to the reference ligand B. This method is based on a thermodynamic cycle in which ligand A is transformed to ligand B, both in water and protein environment (**Figure A.1**). The alchemical transformation is performed through a set of  $\lambda$ -windows, in which non-physical intermediate ligands are simulated with MD simulations. The number of  $\lambda$ -windows has to be adequate to ensure a significant overlap of potential energy distributions between ligands. A mapping between the two ligands needs to be performed to establish which atoms should be mutated and which conserved.<sup>371</sup>



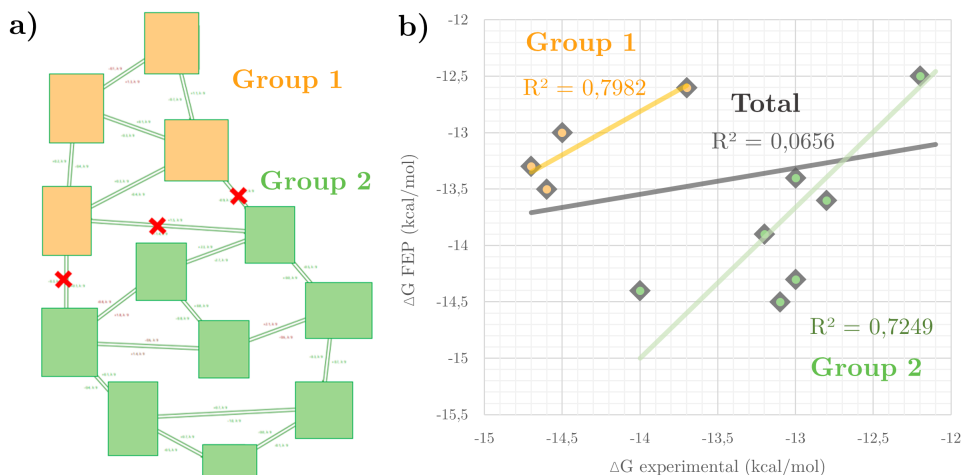
**Figure A.1:** a) The relative binding free energy between ligands A and B is calculated through a thermodynamics cycle using  $\lambda$  windows.

The system of interest of this study is 4-hydroxyphenylpyruvate dioxygenase (HPPD), an iron enzyme that catalyzes the oxygenation of HPPA to form HGA, a crucial precursor in plants. A class of herbicides has been designed to target HPPD, as its inhibition alters the synthesis of photoprotectant carotenoids, leading to intense bleaching and, ultimately death. The binding site of HPPD contains an octahedral  $\text{Fe}^{2+}$  with two His and one Glu, a water molecule and two additional sites, which can be either a ligand or two water molecules.<sup>372</sup>

MD simulations of the rat HPPD (PDB=1sqi)<sup>372</sup> were carried out to see the behavior of the metal parameters in Flare. MD results show that the distances between Fe and all six coordinators are maintained and Fe does not leave the binding site. MD simulation without the coordinating water molecule reveals that the ligand fluctuates, to complete the octahedral coordination, a Gln residue relocates to coordinate with the Fe. These results demonstrate that the force field correctly represents the metal coordination.

Similarity clustering was performed with 27 ligands from the same project to obtain those with the highest similarity. Dockings were conducted with the 12 selected ligands belonging to the same cluster. Analysis of different scoring functions reveals no correlation with binding energies, meaning that docking functions are not good approximates for  $\Delta G$ , at least for the HPPD system. To do the FEP calculations, a map was designed automatically by Flare software. However, this map was improved manually to include more redundancy. For FEP calculations, initial equilibration of 500ps was performed followed by MD of 4ns for each  $\lambda$ -windows, in average each connection had 10  $\lambda$ -windows.

Initial results of FEP show a mean unsigned error (MUE) of 0.88 kcal/mol and a correlation of  $R^2=0.07$ . The correlation between  $\Delta\Delta G_{\text{exp}}$  and  $\Delta\Delta G_{\text{FEP}}$  is 0.50 and its representation displays that three connections deviate from the general tendency. Closer inspection of the results shows there are two subsets of results which have very good correlation within them, meaning that the three connections are not able to link correctly the two subsets. These links correspond to connections involving changes in atom hybridization or transformation between different enantiomers. The FEP map obtained is represented in **Figure A.2a**, highlighting the two subgroups and problematic connections. If the correlation between  $\Delta G_{\text{exp}}$  and  $\Delta G_{\text{FEP}}$  is represented by subgroup, the correlations obtained are 10 times higher if compared to the total (**Figure A.2b**).



**Figure A.2:** a) FEP map with subsets and problematic links. b) Correlation between  $\Delta G_{\text{exp}}$  and  $\Delta G_{\text{FEP}}$ .

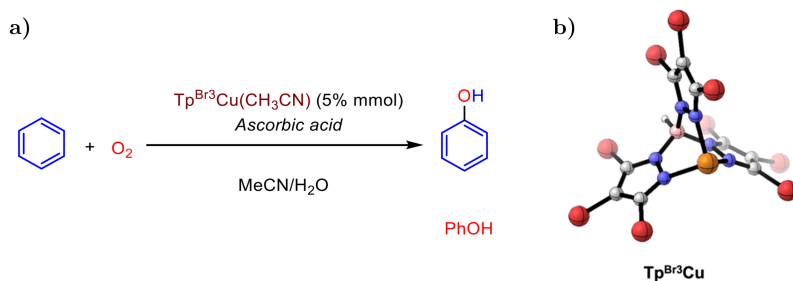
Therefore, FEP calculations were improved by removing the problematic connections and making new ones more conservative regarding the mapping. Additional FEP calculations, QM torsion scans, or short MD simulations were performed to assess specific substituents' rotation or establish the more active enantiomer. After studying in detail all these considerations, the results show a much better correlation,  $R^2=0.38$ , and a very low mean unsigned error (MUE) of 0.61kcal/mol. The error of FEP is usually between 1-1.5 kcal/mol, consequently these results are very promising.

This study showed us how FEP calculations on Flare with metallic proteins (HPPD) can correctly predict binding free energies with the error of 0.6 kcal/mol compared to experimental values. This work highlights how these calculations are very sensitive to initial configurations; each ligand's binding must be carefully studied, especially when interacting with metals. For future perspectives, this protocol can be implemented to work on other metalloenzymes to determine free energies of binding of other pesticides.

## A.2. Direct benzene hydroxylation with O<sub>2</sub> induced by copper complexes

Phenol has a pivotal role in the chemical industry due to its wide use for the production of bisphenol A or phenolic resins, which have applications in the automotive industry and construction sector.<sup>373</sup> Its predominant synthetic route, the cumene process, is a multi-step process that employs O<sub>2</sub> as the oxidant, but has very low yields.<sup>374</sup> An ideal approach would consist of direct hydroxylation using O<sub>2</sub> as an oxidant, which has been reported with heterogeneous catalysts and vanadium complexes as homogeneous catalysts.<sup>375,376</sup>

As copper is used widely in metalloenzymes to catalyze reactions, in this work the copper complex TpBr<sub>3</sub>Cu(NCMe) was used to perform the direct oxidation of benzene into phenol using O<sub>2</sub> in homogeneous phase and ascorbic acid as the source of protons and electrons (**Figure A.3**). Phenol was detected in a 60% yield when the reaction was performed at room temperature under 40 bars of O<sub>2</sub>. Prolonged reaction times led to mixtures of compounds, derived from overoxidation processes. The most relevant aspect is that ascorbic acid seems crucial as experiments employing other agents did not induce the oxidation of benzene to any extent. Therefore, this work aimed to perform DFT studies to propose the active species and mechanism for such oxidation.

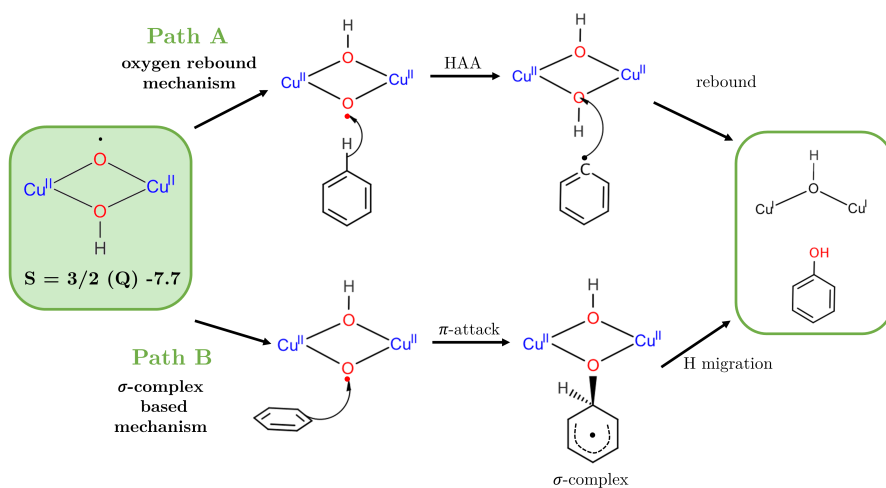


**Figure A.3:** a) Copper-mediated hydroxylation of benzene with O<sub>2</sub> in the presence of ascorbic acid. b) Structure of Co-complex. Reprinted from [377].

Eight mono- and binuclear copper–oxygen species were computationally tested as potential benzene oxidants. All monomeric and most dimeric copper–oxygen species were discarded because the products of the reactions, the TS or the species itself were more than 30–40 kcal/mol over the reactants and these

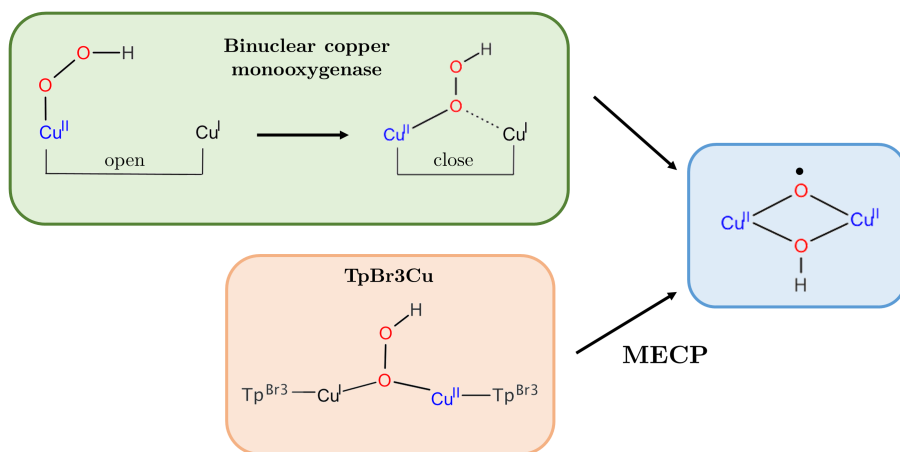
pathways are unfeasible. The only species that was able to perform H abstraction from benzene with a low barrier was the  $\text{Cu}^{\text{II}}(\mu\text{-O})\cdot(\mu\text{-OH})\text{Cu}^{\text{II}}$  complex in a quadruplet state (**Figure A.4**). The formation of this dimer form is a thermodynamically favored process and hydrogen-atom abstraction from ascorbic acid leads to the hydroperoxo species. Cleavage of the O–OH bond from the hydroperoxo leads to the active species, which is more stable in quartet state (has to go through doublet–quartet crossing).

The active species can lead to the phenol through two mechanisms (**Figure A.4**). The oxygen rebound mechanism is the most commonly accepted one, in which the active species abstracts the hydrogen from benzene, generating a phenyl radical and a hydroxide intermediate. In the subsequent rebound step, the phenyl radical attacks the Cu–OH center to give the phenol. The  $\sigma$ -complex-based mechanism has been proposed mainly for P450 proteins. It begins with an attack on the  $\pi$  system of the benzene by the active species to produce a  $\sigma$  complex. In the second step, a proton shuttle, mediated in P450 by the porphyrin ring, transfers the proton from the carbon to the oxygen, yielding the phenol. The whole reaction was calculated for the two different mechanisms, the latter entails an overall barrier of 16.0 kcal/mol, only slightly higher than that the former (14.3 kcal/mol). These calculations show that both mechanisms could be competitive for benzene hydroxylation by the active species.



**Figure A.4:** Possible mechanism for benzene hydroxylation by active species. Reprinted from [377].

Very recently, a computational study has revealed a similar mechanism for O<sub>2</sub> activation and substrate hydroxylation by binuclear copper monooxygenases.<sup>378</sup> They propose the same active species for substrate hydroxylation, even though these enzymes contain two copper(I) at a distance of 11 Å. The study reveals that in the presence of ascorbate, hydrogen-atom abstraction initially forms an inert Cu(II)–OOH intermediate. However, this intermediate transforms to a close conformation (distance of 5 Å), which through an oxygen rebound mechanism affords the same active species we found when working with TpBr<sub>3</sub>Cu(NCMe). **Figure A.5** compares the key steps for the formation of such species for binuclear copper monooxygenases with our results.



**Figure A.5:** Formation of Active Species in Binuclear Copper Monooxygenases and Cu-Complex. Reprinted from [377].

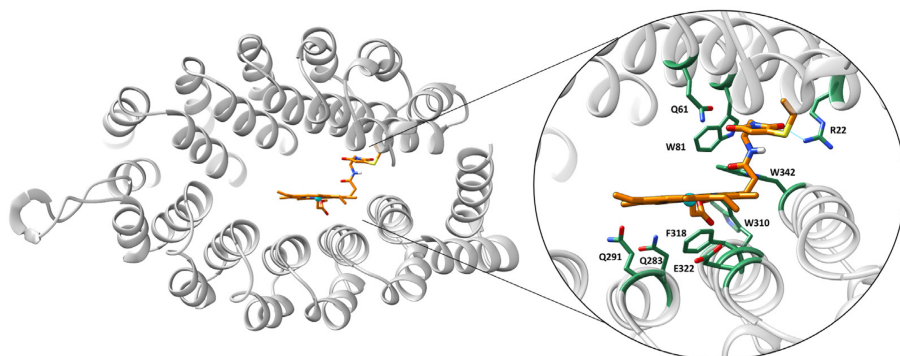
From this work it can be concluded that DFT calculations allowed us to propose the bimetallic Cu<sup>II</sup>(μ-O·)(μ-OH)Cu<sup>II</sup> as active species for the oxidation process of benzene. Ascorbic acid plays a crucial role for its formation through a hydrogen-atom abstraction from the O–H bond of ascorbic acid. While the oxygen rebound mechanisms is favored in this case, the σ-complex based mechanism would also be viable as there is only a 1.7 kcal/mol difference. Most importantly, this project has demonstrated that homogenous catalyst can carry out same mechanism as natural binuclear monooxygenases by employing the same active species.

### A.3. Molecular modeling of an artificial Co-Hemoprotein for H<sub>2</sub> production

In the quest to transition to more sustainable sources of energy, H<sub>2</sub> has emerged as a good candidate because its combustion only releases water. Due to its high demand, researchers have focused on new catalysts, such as Co-porphyrins. These show great potential for H<sub>2</sub> production, but they are limited by their low solubility and instability in water. To overcome these limitations ArM have been developed by incorporating the Co moieties into protein scaffolds. Other groups have reported several examples, but in this work Ricoux *et al* have engineered an ArM by covalent attachment of Co-porphyrin into  $\alpha$ -Rep, showing remarkable activity to catalyze the photoinduced production of H<sub>2</sub> and CO<sub>2</sub> reduction.

The cofactor Co(III)Mal-PPIXMME was synthesized in three steps and then was incubated with synthesized and purified (A3A3')Y26C. This biohybrid was characterized by MALDI-TOF, circular dichroism (CD) and UV-Visible studies. However, with no X-ray structure available, it was not possible to identify the cobalt axial ligand. Therefore, **molecular modeling** was performed to elucidate the most stable orientation of the non-natural porphyrin into (A3A3')Y26C and to assess the relevant interactions between both subsystems, including possible coordination bonds. For the protein structure, we used the representative geometry of the most populated cluster of conformations of a 300 ns molecular dynamics (MD) of A3A3' with Tyr26 was mutated to cysteine.

The artificial cofactor was docked to a rigid structure of modeled  $\alpha$ -Rep using the covalent protocol. The porphyrin stands very well into the inter-dominial pocket and displays good interactions with surrounding amino acids of the cavity such as Phe318. The porphyrin establishes a hydrogen bond with Arg22 or several contacts with hydrophobic patch constituted by Trp81, Trp310, and Trp342 (**Figure A.6**). Close-lying residues that could coordinate are Gln 61, 283, 291 or Asp322. Based on this structure, further dockings with flexibility on mentioned residues were carried out to see whether nearby lying residues could eventually interact with the metal. From those calculations, it appears that Glu322 and Gln291 may both reach convenient distances for coordination, though the former needs Phe318 to be out of the way to the iron. Additionally, these docking solutions show different degrees of rotation of the porphyrin around the linker.

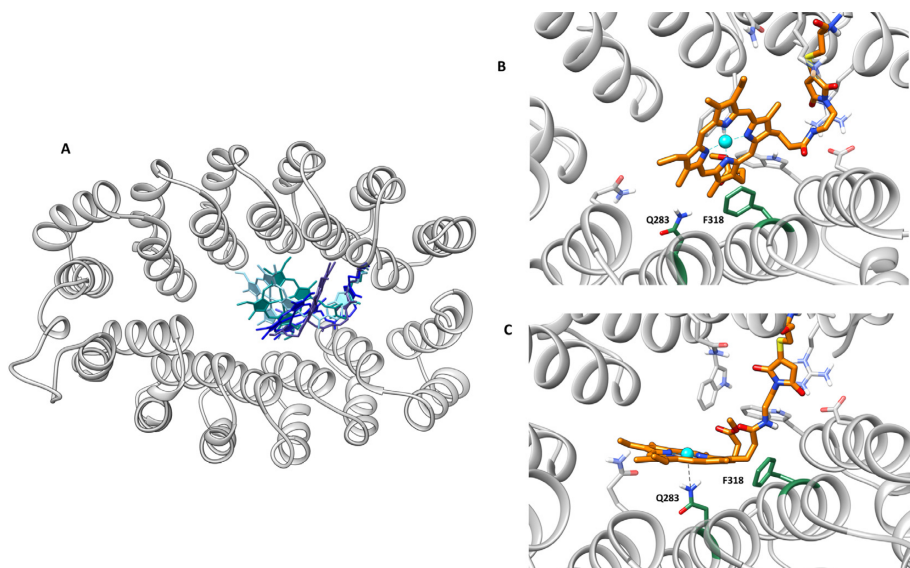


**Figure A.6:** Lowest energy docking solution of the (A3A3')Y26C-Co-cofactor biohybrid. Reprinted from [379].

To further analyze this hypothesis, classical MD was performed. The porphyrin linker was set up by defining a non-standard amino acid and by using of the MCPB.py algorithm for the parameterization of the metallic center and its first coordination sphere. The results of these calculations agreed with the flexible docking and showed a flexible cofactor in the binding site (**Figure A.7a**). However, some orientations clearly dominated.

The most populated one was similar to the best-docked structure described at the beginning of this section (**Figure A.7b**). One can observe a very stable  $\pi$ -stacking between Phe318 and the porphyrin. Interestingly, the Phe318 side chain placed most of the simulation just below the metal with a position that would be occupied by an axial ligand of the metal in natural hemoenzymes. This shows that only one face of the porphyrin is accessible for catalysis and has a well-defined and asymmetric distal environment. Alternative orientations correspond to geometries in which the interaction between Phe318 and Co-cofactor is lost. When such an interaction is lost, cofactor behaves very freely in the interface between the A3 and A3' subdomains.

Importantly, one of the most populated geometries of these alternative orientations showed short distances between Gln283 and the cobalt (**Figure A.7c**), some consistent with direct Co-Gln283 coordination, and others with a bridged water molecule between the amino acid and the metal. Such interaction is only possible because the absence of the interaction between Phe318 and Co-cofactor is associated with an increase in the accessibility of the metal.



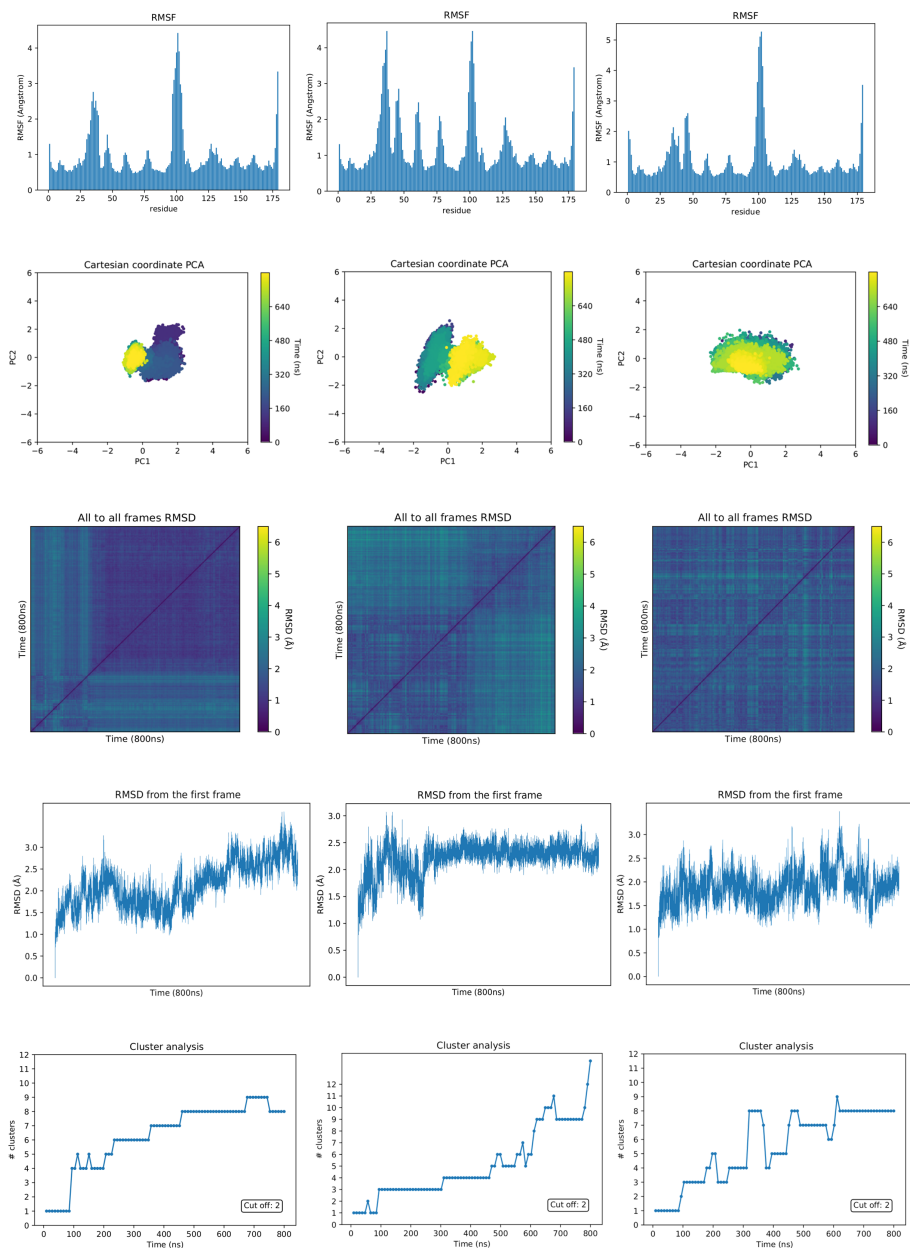
**Figure A.7:** **a)** Main clusters of cofactors extracted from MD simulation. **b)** Most populated cluster. **c)** MD Snapshot with the closest distance between the metal and Gln283. Reprinted from [379].

In conclusion, the molecular modeling study showed that the linkage of Co-cofactor to (A3A3/Y26C led to a macrocycle located at the interface between both subdomains and excluded from the solvent. Furthermore, the main orientation of Co-cofactor in A3A3/ presented a strong contact with Phe318 that aided in the packing of the porphyrin hydrophobically to one domain and led to an asymmetric environment on the distal side. Finally, the only possible coordination of the metal would only appear on transient structures with Gln283, but no possible coordination is observed to His or Cys residues.

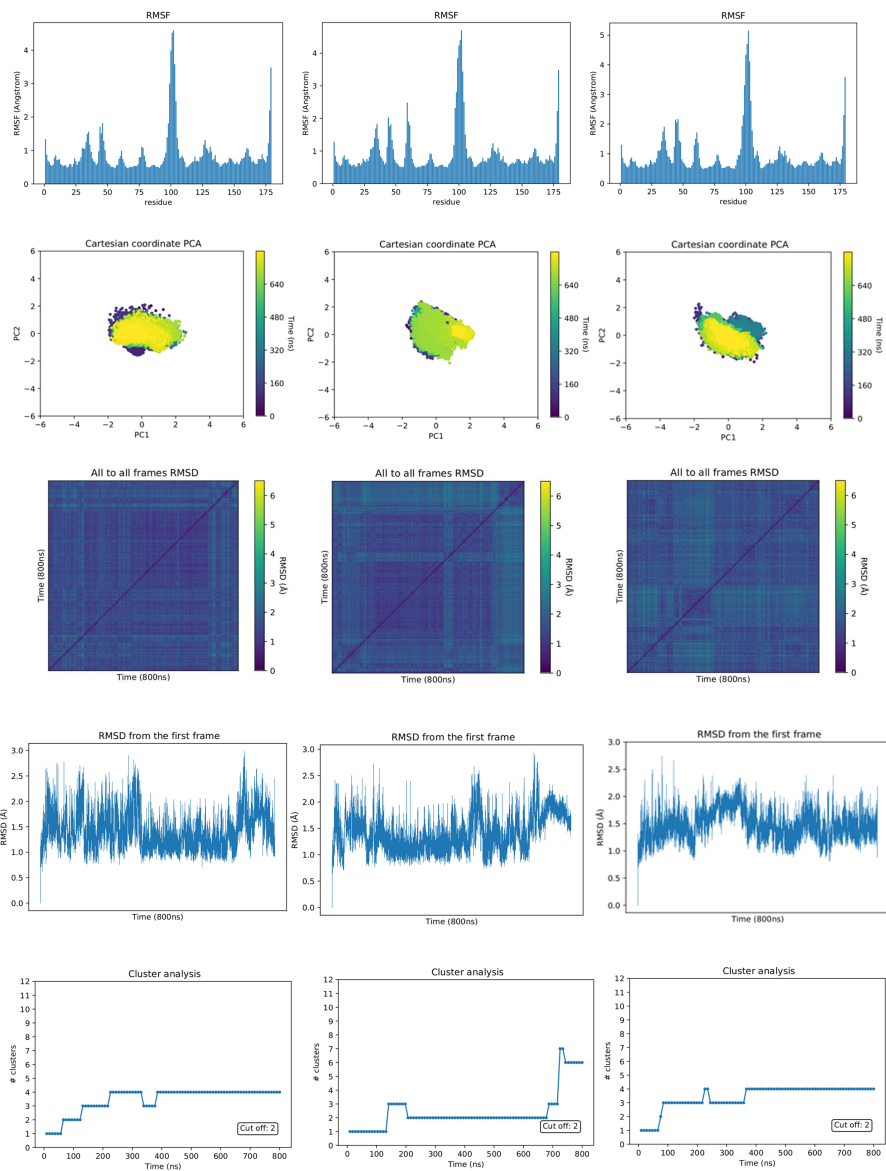
# APPENDIX **B**

## Chapter 4: Supplementary information

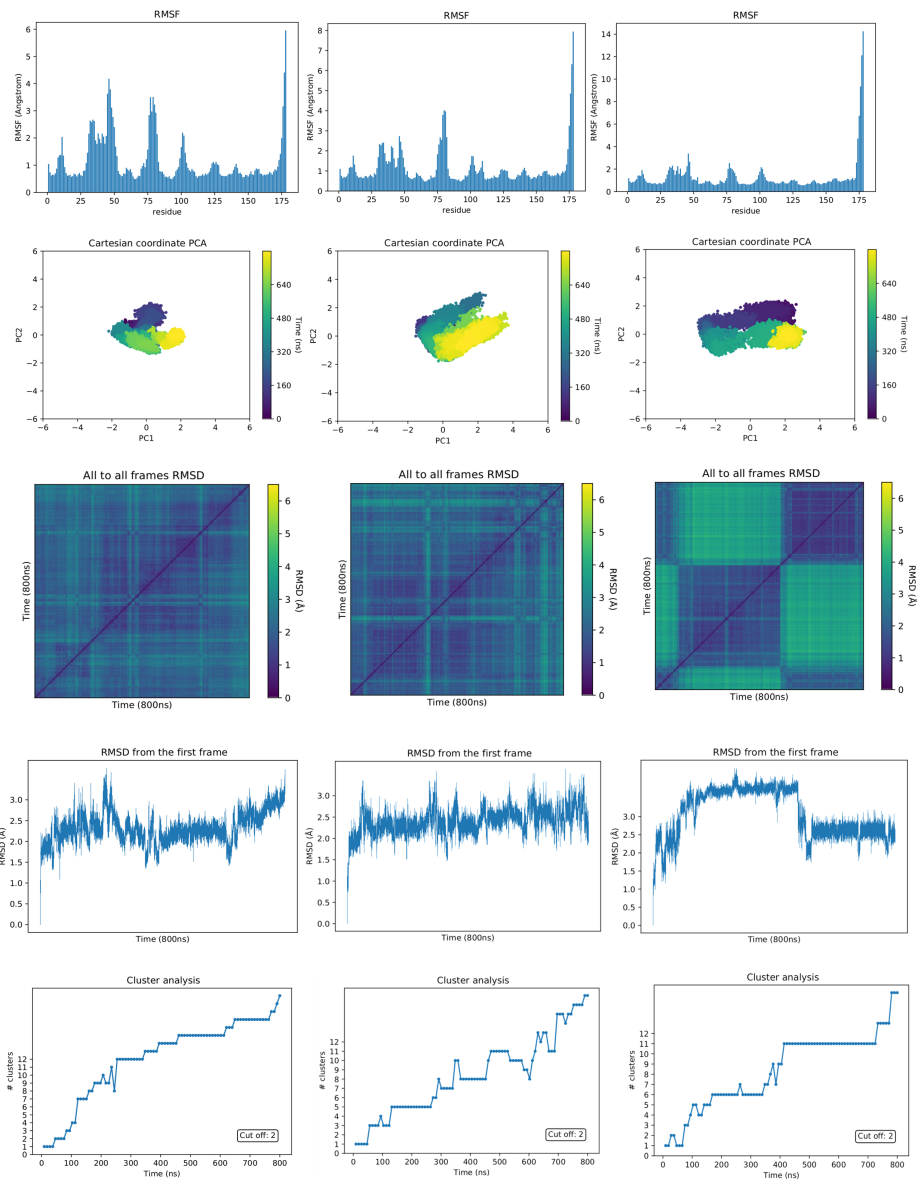
### **B.1.** Exploring the molecular events of heme binding mechanisms



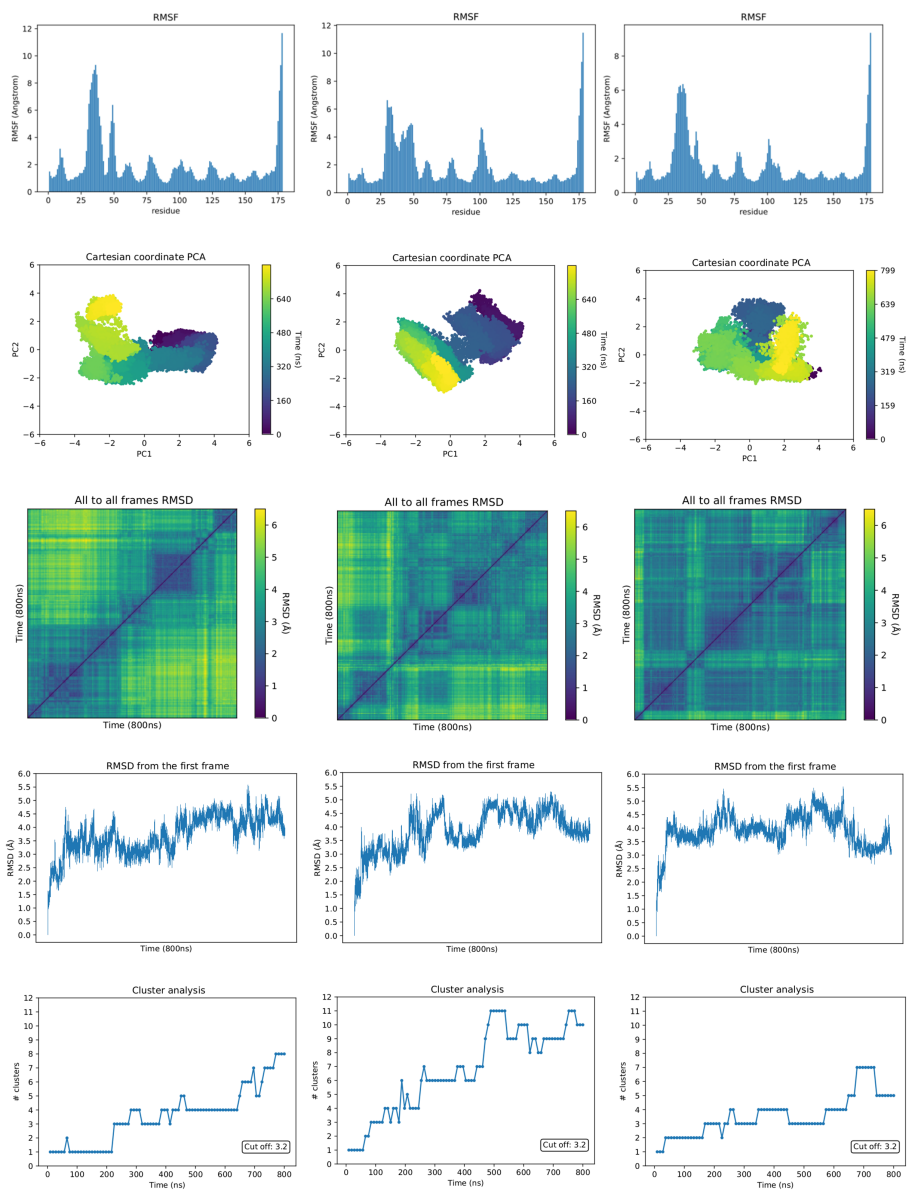
**Figure B.1:** GaMD convergence analysis of HasA apo form of *Yersinia pestis* (800ns – three replicas): RMSF, RMSD, PCA, all-to-all RMSD and cluster counting.



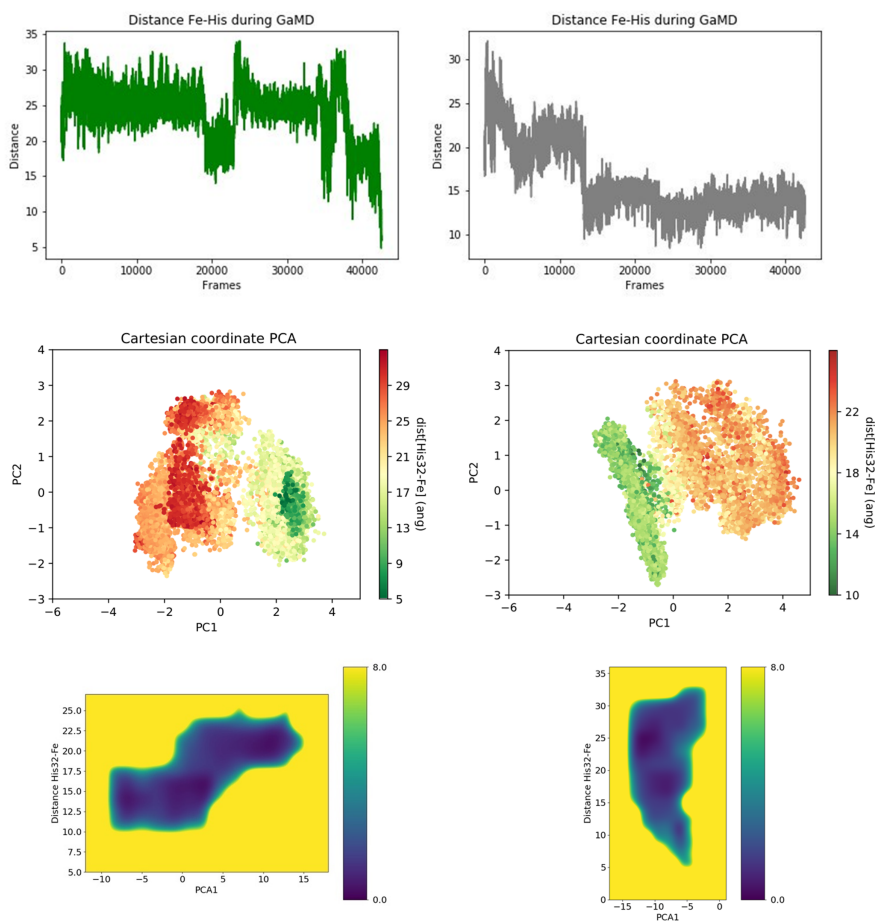
**Figure B.2:** GaMD convergence analysis of HasA holo form of *Yersenia pestis* (800ns – three replicas): RMSF, RMSD, PCA, all-to-all RMSD and cluster counting.



**Figure B.3:** GaMD convergence analysis of HasA apo form of *Serratia marcescens* (800ns – three replicas): RMSF, RMSD, PCA, all-to-all RMSD and cluster counting.



**Figure B.4:** GaMD convergence analysis of HasA holo form of *Serratia marcescens* (800ns – three replicas): RMSF, RMSD, PCA, all-to-all RMSD and cluster counting.



**Figure B.5:** GaMD simulation replicas of HasAsm with heme-Fe(III) bound. **a)** Distance between Fe and His32 during GaMD. **b)** Cartesian coordinate PCA analysis colored according to the distance between Fe and His32 during GaMD **c)** Reweighted PMF calculations in front of PCA1 and distance Fe-His32. b) and c) are obtained using the fragment of the trajectory in which the loop is closing

## B.2. Development of software for the identification of heme binding sites

**Aromatic:** PHE, TYR, TRP

**Polar:** TYR, THR, SER, CYS, MET, ASN, GLN, HIS,

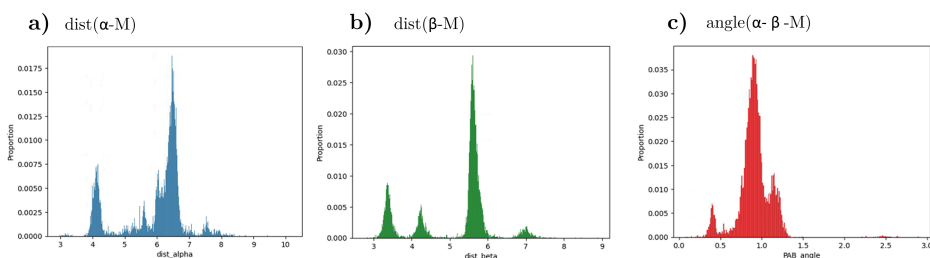
**Positive:** HIS, LYS, ARG,

**Negative:** ASP, GLU,

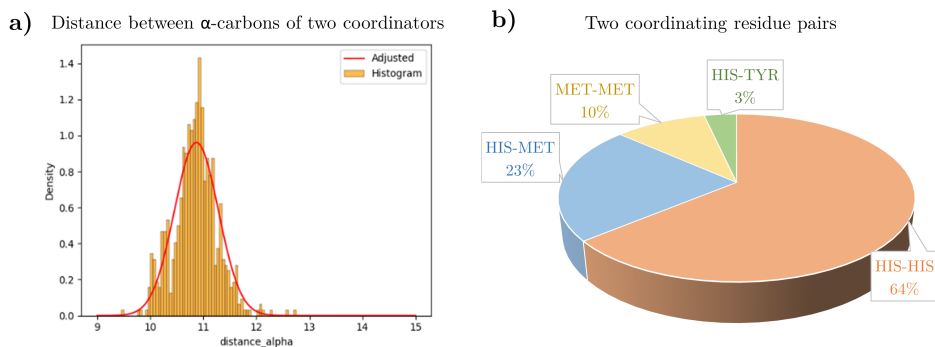
**Hydrophobic:** PHE, TRP, ALA, VAL, LEU, ILE, PRO,

**Large chain:** PHE, TYR, TRP, MET, GLN, GLU, ARG, LYS, HIS, LEU, ILE

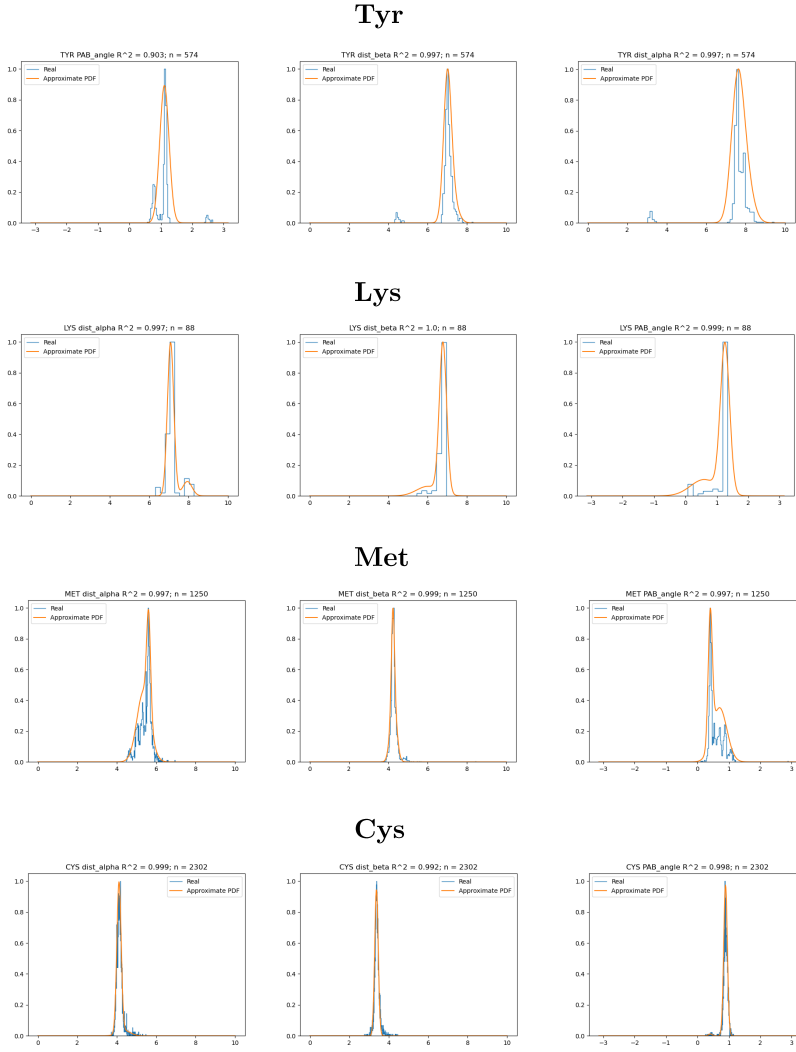
**Figure B.6:** Classification of residues for heme binding site characterization.



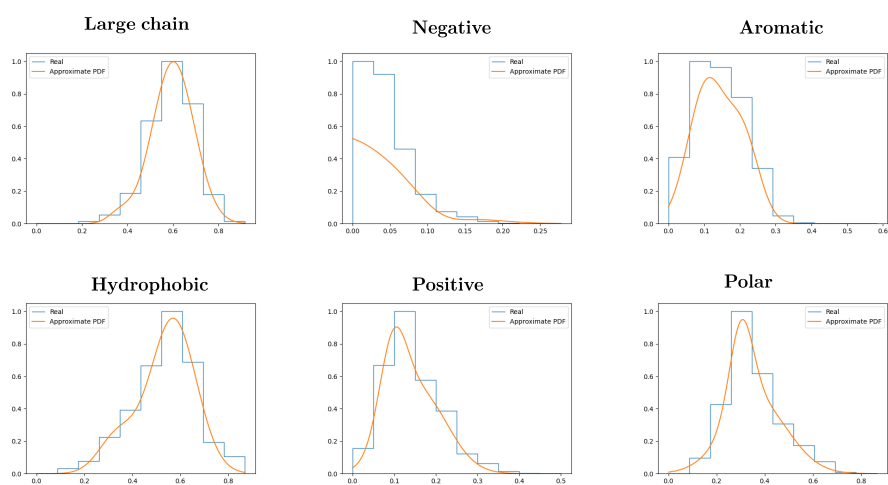
**Figure B.7:** a) Distribution of  $\alpha$ -carbon-Metal distances b) Distribution of  $\beta$ -carbon-Metal distances c) Distribution of  $\alpha$ -carbon- $\beta$ -carbon-metal angle



**Figure B.8:** a) Distances between C- $\alpha$  of two coordinating residues b) Proportion of two coordinating residues



**Figure B.9:** Bimodal distribution fitting for Tyr, Lys, Met and Cys for dist( $\alpha$ -M), dist( $\beta$ -M) and angle( $\alpha$ - $\beta$ -M).



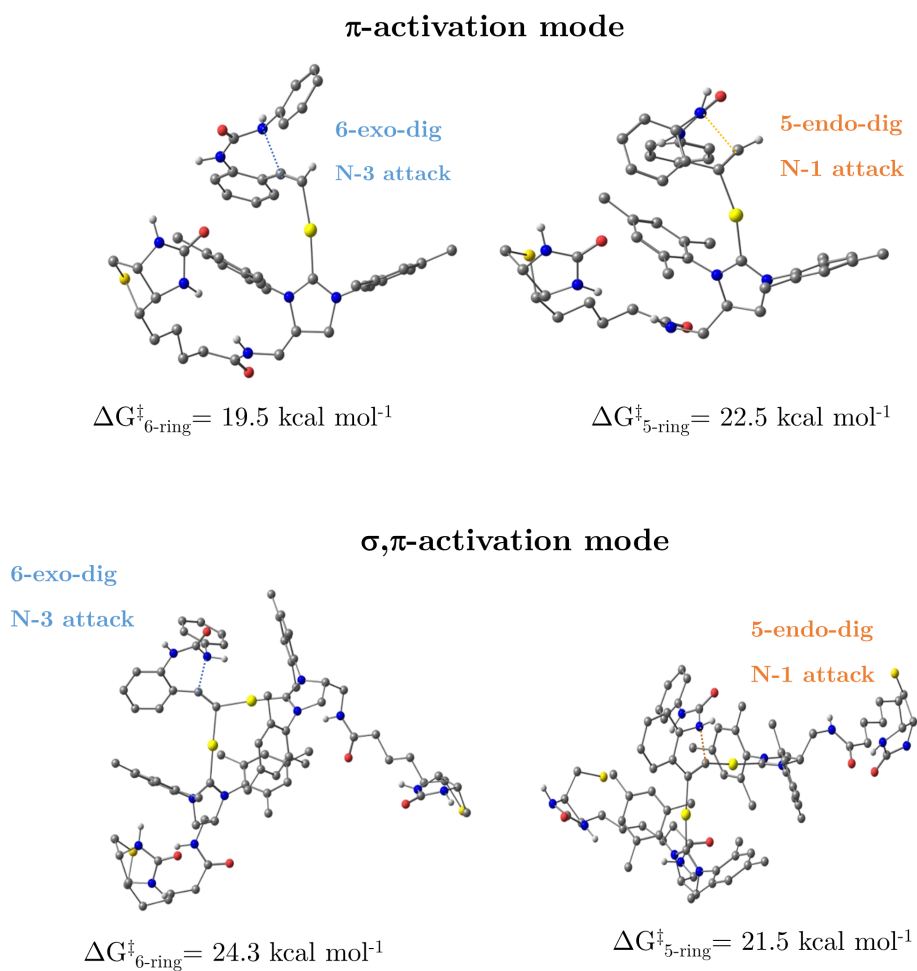
**Figure B.10:** Bimodal distribution fitting for types of residues.



# APPENDIX C

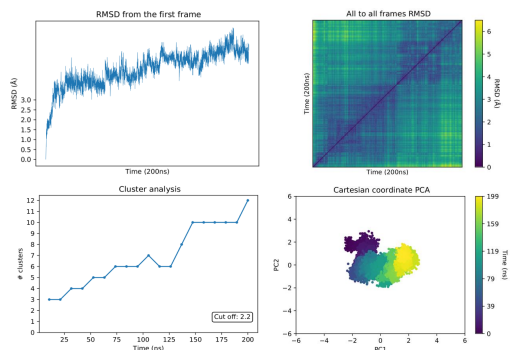
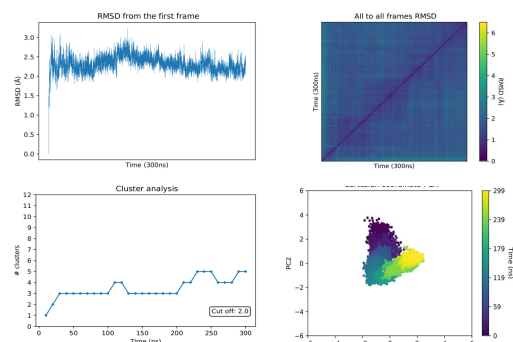
## Chapter 5: Supplementary information

### C.1. Molecular modeling to optimize an Au-ArM for heterocyclization

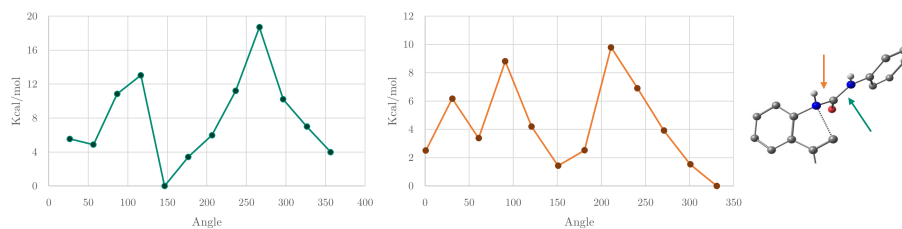


**Figure C.1:** Optimized structure of favored TS for 6-exo-dig product by  $\pi$ -activation mode and  $\sigma, \pi$ -activation mode.

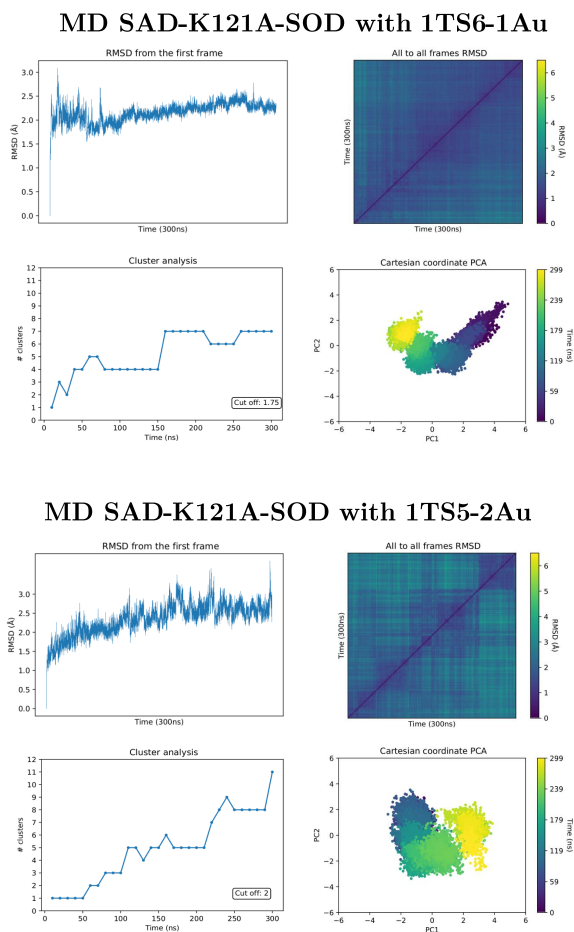
## MD SAD-K121A-SOD with Sav fixed

MD SAD-K121A-SOD with *biot*-Au-2

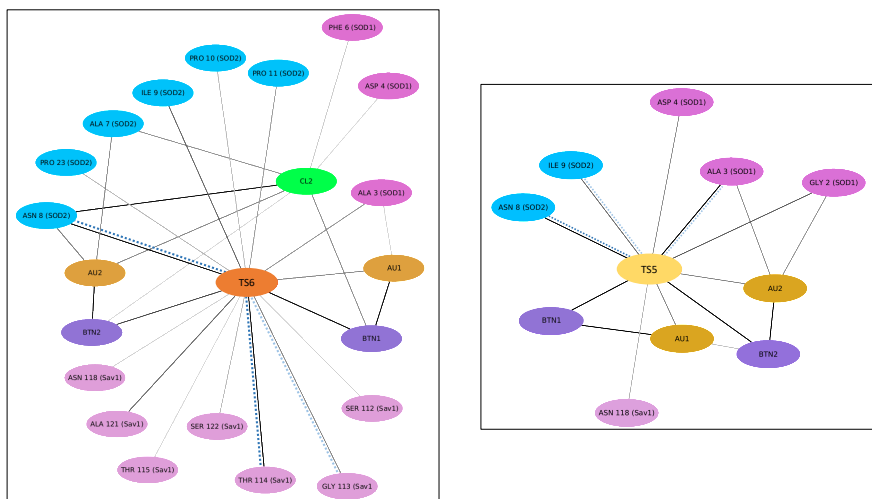
**Figure C.2:** MD convergence analysis (RMSD, all-to-all RMSD, PCA, cluster counting) of : a) SOD regions of initial MD of SAV-SOD (200 ns) b) MD of SAV-SOD with cofactors *biot*-Au 2 (300 ns)



**Figure C.3:** DFT rotational bond study of amide bonds from substrate ethynylphenylurea



**Figure C.4:** MD convergence analysis (RMSD, all-to-all RMSD, PCA, cluster counting) of SAV-K121-SOD with: **a)** 1TS6-1Au **b)** 1TS5-2Au.



**Figure C.5:** Residues contribution network extracted from MD simulations of Sav-SOD K121A with: **a)** 1TS6-1Au (6-exo-dig) and **b)** 1TS5-2Au. Network nodes represent protein residues, biotin (BTN), TS5/6 or atoms (Cl, Au) and edges represent VdW interactions (black-grey straight line). Edges are color-weighted by the number of interactions in each frame.

	1TS5-2Au	1TS6-1Au
Gly 2 (SOD1)	-1.887	-0.083
Ala 3 (SOD1)	-4.053	-1.671
Asp 4 (SOD1)	-0.026	0.237
Ala 7 (SOD2)	-0.102	-2.108
Asn 8 (SOD2)	-2.19	-4.930
Ile 9 (SOD2)	-1.045	-0.577
Pro 10 (SOD2)	-0.023	-0.099
Pro 11 (SOD2)	-0.014	-0.384
Pro 23 (SOD2)	-0.015	-0.369

	1TS5-2Au	1TS6-1Au
Ser 112 (Sav1)	-0.615	-1.001
Gly 113 (Sav1)	0.034	-0.011
Thr 114 (Sav1)	-0.329	-3.019
Thr 115 (Sav1)	-0.004	-0.369
Glu 116 (Sav1)	0.007	-0.020
Ala 117 (Sav1)	0.048	0.008
Asn 118 (Sav1)	-0.186	-0.380
Ala 119 (Sav1)	-0.121	-0.162
Trp120 (Sav1)	-2.482	-3.003
Ala 121 (Sav1)	-1.019	-0.937
Asn122 (Sav1)	-0.296	-0.383

**Figure C.6:** Decomposition of total energy (kcal/mol) for most relevant residues of SOD chains using MMGBSA and based on cytoscape analysis for encapsulated pseudo-transition states 1TS5-2Au ( $\sigma$ ,  $\pi$ -activation mode) and 1TS6-1Au ( $\pi$ -activation mode).

C.2. Rationalization of a streptavidin based suzukiase

Cofactor C2

	RE-TS from OA_A1		RE-TS from OA_A2
Trans-proR	0.0	Trans-proS	0.9
Cis-proS	3.4	Cis-proR	3.7

Figure C.7: Gibbs energies of RE-TS with cofactor C2.

	R	S
WT	18.58	21.96
S112M	6.9	17.62
Double	22.44	21.92

Figure C.8: Docking results for each Sav system for proR-OA and proS-OA using Gold and goldscore as scoring function.

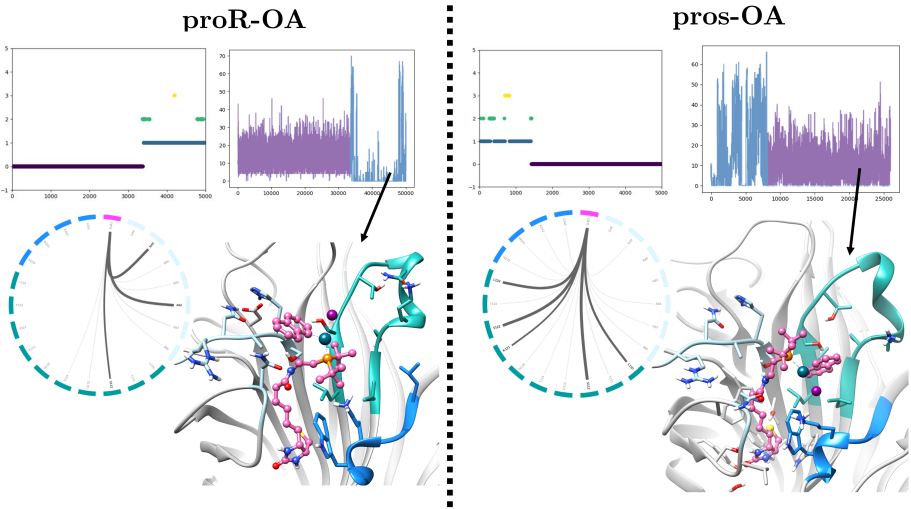


Figure C.9: Results of Sav-WT-C1 for proR and proS OA. Panels: 1) Cluster of OA during MD simulation 2) Clashes between Sav and second substrate during MD 3) Structure of OA cluster with less clashes and 4) Interaction map between OA and Sav regions.

APPENDIX

D

Chapter 6: Supplementary information

D.1. Finding metal binding sites to design a new ArM

Monomeric complex	PDB	Dimeric complex	PDB
tubulin : alphaRep-iE5	6GWC	A3 : A3 dimer	3LTJ
tubulin : alphaRep-iiH5	6GWC	A3 : A3 dimer : PEG	3LTM
tubulin : CopN : alphaRep-iiiA5	6GX7	A3_A3 bidomain	6FT5
octarellinV.1: alphaRep7	4ZV6	A3_bGFPD bidomain	6HWP
alphaRep2 : alphaRepA3	4JW2	A3_bGFPD bidomain	6FSQ
NCS3.24 : alphaRep	4JW3		
GFP : bGFPA	4XL5		
GFP : bGFPC	4XVP		
YabT : alphaRepE8	6G4J		
FNE : alphaRep	5DCQ		

Figure D.1: List of pdb structures with monomeric and dimeric  $\alpha$ -Reps.



# List of publications

- Yepes-Pérez, A. F., Herrera-Calderon, O., Sánchez-Aparicio, J.-E., Tiessler-Sala, L., Maréchal, J.-D., Cardona-G, W. Investigating Potential Inhibitory Effect of Uncaria Tomentosa (Cat's Claw) against the Main Protease 3CLpro of SARS-CoV-2 by Molecular Modeling. *Evidence-Based Complementary and Alternative Medicine* **2020**, 4932572 (2020).
- Sánchez-Aparicio, J.-E., Tiessler-Sala, L., Velasco-Carneros, L., Roldán-Martín, L., Sciortino, G., Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *Journal of Chemical Information and Modelling* **61**, 311–323 (2021)
- Christoffel, F., Igareta, N. V., Pellizzoni, M. M., Tiessler-Sala, L., Lozhkin, B., Spiess, D. C., Lledós, A., Maréchal, J.-D., Peterson, R. L., Ward, T. R. Design and Evolution of Chimeric Streptavidin for Protein-Enabled Dual Gold Catalysis. *Nature Catalysis* **4**, 643–653 (2021).
- Tiessler-Sala, L., Sciortino, G., Alonso-Cotchico, L., Masgrau, L., Lledós, A., Maréchal, J.-D. Getting Deeper into the Molecular Events of Heme Binding Mechanisms: A Comparative Multi-Level Computational Study of HasAsm and HasAyp Hemophores. *Inorganic Chemistry* **61**, 17068–17079 (2022).
- Borrego, E., Tiessler-Sala, L., Lázaro, J.J., Caballero, A., Pérez, P.J. & Lledós, A. Direct Benzene Hydroxylation with Dioxygen Induced by Copper Complexes: Uncovering the Active Species by DFT Calculations. *Organometallics* **41** 1892–1904 (2022).
- Udry, G. A. O., Tiessler-Sala, L., Pugliese, E., Urvoas, A., Halime, Z., Maréchal, J.-D., Mahy, J.-P. & Ricoux, R. Photocatalytic Hydrogen Production and Carbon Dioxide Reduction Catalyzed by an Artificial Cobalt Hemoprotein *International Journal of Molecular Sciences* **23**, 14640 (2022).



## Bibliography

1. Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. Metal ions in biological catalysis: from enzyme databases to general principles. *Journal of Biological Inorganic Chemistry* **13**, 1205–1218 (2008).
2. Waldron, K. J., Rutherford, J. C., Ford, D. & Robinson, N. J. Metalloproteins and metal sensing. *Nature* **460**, 823–830 (2009).
3. Azia, A., Levy, R., Unger, R., Edelman, M. & Sobolev, V. Genome-wide computational determination of the human metalloproteome. *Proteins: Structure, Function, and Bioinformatics* **83**, 931–939 (2015).
4. Bertini, I., Gray, H. B., Stiefel, E. I. & Silverstein Valentine, J. Biological inorganic chemistry: structure and reactivity (University Science Books, 2007).
5. Exley, C. The coordination chemistry of aluminium in neurodegenerative disease. *Coordination Chemistry Reviews* **256**, 2142–2146 (2012).
6. Kepp, K. P. Bioinorganic Chemistry of Alzheimer's Disease. *Chemical Reviews* **112**, 5193–5239 (2012).
7. Cruysen, J. R. W., Koper, M. C., Jelsma, J., Heymans, M., Heyligers, I. C., Grimm, B., Mathijssen, N. M. C. & Schotanus, M. G. M. Prosthetic hip-associated cobalt toxicity: a systematic review of case series and case reports. *EFORT Open Reviews* **7**, 188–199 (2022).
8. Leyssens, L., Vinck, B., Van Der Straeten, C., Wuyts, F. & Maes, L. Cobalt toxicity in humans. A review of the potential sources and systemic health effects. *Toxicology* **387**, 43–56 (2017).

9. Morin, Y. L., Foley, A. R., Martineau, G. & Roussel, J. Quebec beer-drinkers' cardiomyopathy: forty-eight cases. *Canadian Medical Association Journal* **97**, 881–883 (1967).
10. Mendels, J. Lithium in the treatment of depression. *The American Journal of Psychiatry* **133**, 373–378 (1976).
11. Noyes, R., Dempsey, G. M., Blum, A. & Cavanaugh, G. L. Lithium treatment of depression. *Comprehensive Psychiatry* **15**, 187–193 (1974).
12. Brown, D. H. & Smith, W. E. The chemistry of the gold drugs used in the treatment of rheumatoid arthritis. *Chemical Society Reviews* **9**, 217–240 (1980).
13. Rosenberg, B., Vancamp, L., Trosko, J. E. & Mansour, V. H. Platinum Compounds: a New Class of Potent Antitumour Agents. *Nature* **222**, 385–386 (1969).
14. Lee, S. Y., Kim, C. Y. & Nam, T.-G. Ruthenium Complexes as Anticancer Agents: A Brief History and Perspectives. *Drug Design, Development and Therapy* **14**, 5375–5392 (2020).
15. Schwizer, F., Okamoto, Y., Heinisch, T., Gu, Y., Pellizzoni, M. M., Lebrun, V., Reuter, R., Köhler, V., Lewis, J. C. & Ward, T. R. Artificial Metalloenzymes: Reaction Scope and Optimization Strategies. *Chemical Reviews* **118**, 142–231 (2018).
16. Zoroddu, M. A., Aaseth, J., Crisponi, G., Medici, S., Peana, M. & Nurchi, V. M. The essential metals for humans: a brief overview. *Journal of Inorganic Biochemistry* **195**, 120–129 (2019).
17. Michalak, I. & Chojnacka, K. Fluorine and Silicon as Essential and Toxic Trace Elements. in *Recent Advances in Trace Elements* 207–218 (John Wiley & Sons, Ltd, 2018).
18. Maret, W. The Metals in the Biological Periodic System of the Elements: Concepts and Conjectures. *International Journal of Molecular Sciences* **17**, 66 (2016).
19. Crichton, R. Biological Inorganic Chemistry: A New Introduction to Molecular Structure and Function 3rd ed. (Academic Press, 2019).
20. Vincent, J. B. New Evidence against Chromium as an Essential Trace Element. *The Journal of Nutrition* **147**, 2212–2219 (2017).
21. Di Bona, K. R., Love, S., Rhodes, N. R., McAdory, D., Sinha, S. H., Kern, N., Kent, J., Strickland, J., Wilson, A., Beaird, J., Ramage, J., Rasco, J. F. & Vincent, J. B. Chromium is not an essential trace element for mammals:

- effects of a “low-chromium” diet. *JBIC Journal of Biological Inorganic Chemistry* **16**, 381–390 (2011).
22. Freisinger, E. & Sigel, R. K. The Bioinorganic Periodic Table. *CHIMIA International Journal for Chemistry* **73**, 185–193 (2019).
  23. Pearson, R. G. Hard and Soft Acids and Bases. *Journal of the American Chemical Society* **85**, 3533–3539 (1963).
  24. Cammack, R. & Hughes, M. Considerations for the Specification of Enzyme Assays Involving Metal Ions. in: Proceedings of the 3rd Beilstein ESCEC Symposium “Experimental Standard Conditions Of Enzyme Characterization.” (2023).
  25. Brown, D. H. & Smith, W. E. Metal ions in biological systems. in *Enzyme Chemistry: Impact and applications* 162–195 (Springer Netherlands, 1984).
  26. O'Connor, C. & Adams, J. Essentials of Cell Biology (NPG Education, 2010).
  27. Buxbaum, E. Fundamentals of Protein Structure and Function (Springer International Publishing, 2015).
  28. Kessel, A. & Ben-Tal, N. Introduction to Proteins: Structure, Function, and Motion, Second Edition 2nd ed. (Chapman and Hall/CRC, 2018).
  29. Nelson, D. L. & Cox, M. M. Lehninger principles of biochemistry 4th ed. (W.H. Freeman, 2005).
  30. Garcia-Viloca, M., Gao, J., Karplus, M. & Truhlar, D. G. How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303**, 186–195 (2004).
  31. Gao, J., Ma, S., Major, D. T., Nam, K., Pu, J. & Truhlar, D. G. Mechanisms and Free Energies of Enzymatic Reactions. *Chemical reviews* **106**, 3188–3209 (2006).
  32. Tuñón, I., Laage, D. & Hynes, J. T. Are there dynamical effects in enzyme catalysis? Some thoughts concerning the enzymatic chemical step. *Archives of Biochemistry and Biophysics* **582**, 42–55 (2015).
  33. Zinovjev, K. & Tuñón, I. Quantifying the limits of transition state theory in enzymatic catalysis. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 12390–12395 (2017).
  34. Agarwal, P. K. Role of Protein Dynamics in Reaction Rate Enhancement by Enzymes. *Journal of the American Chemical Society* **127**, 15248–15256 (2005).
  35. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27**, 2985–2993 (1894).

36. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology* **12**, 88–118 (1965).
37. Koshland, D. E. Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America* **44**, 98–104 (1958).
38. Hammes, G. G., Chang, Y.-C. & Oas, T. G. Conformational selection or induced fit: A flux description of reaction mechanism. *Proceedings of the National Academy of Sciences* **106**, 13737–13741 (2009).
39. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chemical Biology* **5**, 789–796 (2009).
40. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in biochemical sciences* **35**, 539–546 (2010).
41. Vogt, A. D. & Di Cera, E. Conformational Selection or Induced Fit? A Critical Appraisal of the Kinetic Mechanism. *Biochemistry* **51**, 5894–5902 (2012).
42. Greives, N. & Zhou, H.-X. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. *Proceedings of the National Academy of Sciences* **111**, 10197–10202 (2014).
43. Robinson, P. K. Enzymes: principles and biotechnological applications. *Essays in Biochemistry* **59**, 1–41 (2015).
44. Glusker, J. P. Structural aspects of metal liganding to functional groups in proteins. *Advances in protein chemistry* **42**, 1–76 (1991).
45. Barber-Zucker, S., Shaanan, B. & Zarivach, R. Transition metal binding selectivity in proteins and its correlation with the phylogenomic classification of the cation diffusion facilitator protein family. *Scientific Reports* **7**, 16381 (2017).
46. De Visser, S. P. Second-Coordination Sphere Effects on Selectivity and Specificity of Heme and Nonheme Iron Enzymes. *Chemistry – A European Journal* **26**, 5308–5327 (2020).
47. Christianson, D. W. Structural Biology of Zinc. in *Advances in Protein Chemistry* 281–355 (Academic Press, 1991).
48. Auld, D. S. Zinc Enzymes. in *Encyclopedia of Inorganic and Bioinorganic Chemistry* 1–43 (John Wiley & Sons, Ltd, 2011).

49. Coleman, J. E. Zinc enzymes. *Current Opinion in Chemical Biology* **2**, 222–234 (1998).
50. McCall, K. A., Huang, C.-c. & Fierke, C. A. Function and Mechanism of Zinc Metalloenzymes. *The Journal of Nutrition* **130**, 1437S–1446S (2000).
51. Krishna, S. S., Majumdar, I. & Grishin, N. V. Structural classification of zinc fingers: survey and summary. *Nucleic Acids Research* **31**, 532–550 (2003).
52. Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. & Raschellà, G. Zinc-finger proteins in health and disease. *Cell Death Discovery* **3**, 1–12 (2017).
53. Maret, W. Zinc Biochemistry: From a Single Zinc Enzyme to a Key Element of Life. *Advances in Nutrition* **4**, 82–91 (2013).
54. Zhang, Y., Rodionov, D. A., Gelfand, M. S. & Gladyshev, V. N. Comparative genomic analyses of nickel, cobalt and vitamin B12 utilization. *BMC Genomics* **10**, 78 (2009).
55. Alfano, M. & Cavazza, C. Structure, function, and biosynthesis of nickel-dependent enzymes. *Protein Science* **29**, 1071–1089 (2020).
56. Ragsdale, S. W. Nickel-based Enzyme Systems. *The Journal of Biological Chemistry* **284**, 18571–18575 (2009).
57. Maroney, M. J. & Ciurli, S. Nonredox Nickel Enzymes. *Chemical reviews* **114**, 4206–4228 (2014).
58. Osman, D., Cooke, A., Young, T. R., Deery, E., Robinson, N. J. & Warren, M. J. The requirement for cobalt in vitamin B12: A paradigm for protein metalation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1868**, 118896 (2021).
59. Kräutler, B. Vitamin B12 and B12-Proteins. in *Encyclopedia of Biological Chemistry* 360–366 (Wiley-VCH, 1998).
60. Cheng, Z., Xia, Y. & Zhou, Z. Recent Advances and Promises in Nitrile Hydratase: From Mechanism to Industrial Applications. *Frontiers in Bioengineering and Biotechnology* **8**, 352 (2020).
61. Kobayashi, M. & Shimizu, S. Cobalt proteins. *European Journal of Biochemistry* **261**, 1–9 (1999).
62. Tishchenko, K., Beloglazkina, E., Mazhuga, A. & Zyk, N. Copper-containing enzymes: Site types and low-molecular-weight model compounds. *Review Journal of Chemistry* **6**, 49–82 (2016).
63. MacPherson, I. S. & Murphy, M. E. P. Type-2 copper-containing enzymes. *Cellular and molecular life sciences: CMLS* **64**, 2887–2899 (2007).

64. Fischer, M., Thöny, B. & Leimkühler, S. The Biosynthesis of Folate and Pterins and Their Enzymology. in *Comprehensive Natural Products II* 599–648 (Elsevier, **2010**).
65. Schmidt, S. B. & Husted, S. The Biochemical Properties of Manganese in Plants. *Plants* **8**, 381 (**2019**).
66. Kaim, W. Bioanorganische Chemie: An Introduction and Guide (John Wiley & Sons, Inc, **2013**).
67. Ortiz de Montellano, P. R. Hemes in Biology. in *Wiley Encyclopedia of Chemical Biology* 1–10 (John Wiley & Sons, Ltd, **2008**).
68. Zhang, L. Heme Biology: The Secret Life of Heme in Regulating Diverse Biological Processes (World Scientific Publishing Company, **2011**).
69. Baureder, M. & Hederstedt, L. Heme Proteins in Lactic Acid Bacteria. in *Advances in Microbial Physiology* 1–43 (Academic Press, **2013**).
70. Kleingardner, J. G. & Bren, K. L. Biological Significance and Applications of Heme c Proteins and Peptides. *Accounts of Chemical Research* **48**, 1845–1852 (**2015**).
71. Bowman, S. E. J. & Bren, K. L. The Chemistry and Biochemistry of Heme c: Functional Bases for Covalent Attachment. *Natural product reports* **25**, 1118–1130 (**2008**).
72. Lin, Y.-W. Structure and function of heme proteins regulated by diverse post-translational modifications. *Archives of Biochemistry and Biophysics* **641**, 1–30 (**2018**).
73. Everse, J. Heme Proteins. in *Encyclopedia of Biological Chemistry* 532–538 (Elsevier, **2013**).
74. Manikandan, P. & Nagini, S. Cytochrome P450 Structure, Function and Clinical Significance: A Review. *Current Drug Targets* **19**, 38–54 (**2017**).
75. Shimizu, T., Lengalova, A., Martínek, V. & Martínková, M. Heme: emergent roles of heme in signal transduction, functional regulation and as catalytic centres. *Chemical Society Reviews* **48**, 5624–5657 (**2019**).
76. Gondim, A. C. S., Guimarães, W. G. & Sousa, E. H. S. Heme-Based Gas Sensors in Nature and Their Chemical and Biotechnological Applications. *BioChem* **2**, 43–63 (**2022**).
77. Shimizu, T., Huang, D., Yan, F., Stranova, M., Bartosova, M., Fojtíková, V. & Martínková, M. Gaseous O<sub>2</sub>, NO, and CO in Signal Transduction: Structure and Function Relationships of Heme-Based Gas Sensors and Heme-Redox Sensors. *Chemical Reviews* **115**, 6491–6533 (**2015**).

78. Mense, S. M. & Zhang, L. Heme: a versatile signaling molecule controlling the activities of diverse regulators ranging from transcription factors to MAP kinases. *Cell Research* **16**, 681–692 (2006).
79. Igarashi, J., Kitanishi, K., Martinkova, M., Murase, M., Iizuka, A. & Shimizu, T. The Roles of Thiolate-Heme Proteins, Other Than the P450 Cytochromes, in the Regulation of Heme-Sensor Proteins. *Acta Chim. Slov.* **55** (2007).
80. Dutt, S., Hamza, I. & Bartnikas, T. B. Molecular Mechanisms of Iron and Heme Metabolism. *Annual review of nutrition* **42**, 311–335 (2022).
81. Chiabrando, D., Vinchi, F., Fiorito, V., Mercurio, S. & Tolosano, E. Heme in pathophysiology: a matter of scavenging, metabolism and trafficking across cell membranes. *Frontiers in Pharmacology* **5**, 61 (2014).
82. Donegan, R. K., Moore, C. M., Hanna, D. A. & Reddi, A. R. Handling heme: The mechanisms underlying the movement of heme within and between cells. *Free radical biology & medicine* **133**, 88–100 (2019).
83. Contreras, H., Chim, N., Credali, A. & Goulding, C. W. Heme uptake in bacterial pathogens. *Current opinion in chemical biology* **0**, 34–41 (2014).
84. Runyen-Janecky, L. Role and regulation of heme iron acquisition in gram-negative pathogens. *Frontiers in Cellular and Infection Microbiology* **3** (2013).
85. Richard, K. L., Kelley, B. R. & Johnson, J. G. Heme Uptake and Utilization by Gram-Negative Bacterial Pathogens. *Frontiers in Cellular and Infection Microbiology* **9** (2019).
86. Brunori, M. Variations on the theme: allosteric control in hemoglobin. *The FEBS Journal* **281**, 633–643 (2014).
87. Ciaccio, C., Coletta, A., De Sanctis, G., Marini, S. & Coletta, M. Cooperativity and allostery in haemoglobin function. *IUBMB Life* **60**, 112–123 (2008).
88. Eaton, W. A. Impact of Conformational Substates and Energy Landscapes on Understanding Hemoglobin Kinetics and Function. *Journal of Biological Physics* **47**, 337–353 (2021).
89. Changeux, J.-P. & Edelstein, S. Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biology Reports* **3**, 19 (2011).
90. Perutz, M. F., Wilkinson, A. J., Paoli, M. & Dodson, G. G. The Stereochemical Mechanism of the Cooperative Effects in Hemoglobin Revisited. *Annual Review of Biophysics and Biomolecular Structure* **27**, 1–34 (1998).
91. Henry, E. R., Bettati, S., Hofrichter, J. & Eaton, W. A. A tertiary two-state allosteric model for hemoglobin. *Biophysical Chemistry* **98**, 149–164 (2002).

92. Guengerich, F. P., Wilkey, C. J. & Phan, T. T. Human cytochrome P450 enzymes bind drugs and other substrates mainly through conformational-selection modes. *Journal of Biological Chemistry* **294**, 10928–10941 (2019).
93. Quiroga, I. & Scior, T. Induced fit for cytochrome P450 3A4 based on molecular dynamics. *ADMET & DMPK* **7**, 252–266 (2019).
94. Estrada, D. F., Skinner, A. L., Laurence, J. S. & Scott, E. E. Human cytochrome P450 17A1 conformational selection: modulation by ligand and cytochrome b5. *The Journal of Biological Chemistry* **289**, 14310–14320 (2014).
95. Guengerich, F. P., Wilkey, C. J. & Phan, T. T. N. Human cytochrome P450 enzymes bind drugs and other substrates mainly through conformational-selection modes. *The Journal of Biological Chemistry* **294**, 15875 (2019).
96. Li, T., Bonkovsky, H. L. & Guo, J.-t. Structural analysis of heme proteins: implications for design and prediction. *BMC Structural Biology* **11**, 13 (2011).
97. Smith, L. J., Kahraman, A. & Thornton, J. M. Heme proteins—Diversity in structural characteristics, function, and folding. *Proteins: Structure, Function, and Bioinformatics* **78**, 2349–2368 (2010).
98. Eliezer, D. & Wright, P. E. Is apomyoglobin a molten globule? Structural characterization by NMR. *Journal of Molecular Biology* **263**, 531–538 (1996).
99. Falzone, C. J., Mayer, M. R., Whiteman, E. L., Moore, C. D. & Lecomte, J. T. Design challenges for hemoproteins: the solution structure of apocytochrome b5. *Biochemistry* **35**, 6519–6526 (1996).
100. Feng, Y., Sligar, S. G. & Wand, A. J. Solution structure of apocytochrome B562. *Nature Structural Biology* **1**, 30–35 (1994).
101. Arnesano, F., Banci, L., Bertini, I., Faraone-Mennella, J., Rosato, A., Barker, P. D. & Fersht, A. R. The solution structure of oxidized *Escherichia coli* cytochrome b562. *Biochemistry* **38**, 8657–8670 (1999).
102. Wittung-Stafshede, P. Role of Cofactors in Protein Folding. *Accounts of Chemical Research* **35**, 201–208 (2002).
103. Schneider, S., Sharp, K. H., Barker, P. D. & Paoli, M. An Induced Fit Conformational Change Underlies the Binding Mechanism of the Heme Transport Proteobacteria-Protein HemS\*. *Journal of Biological Chemistry* **281**, 32606–32610 (2006).
104. Jepkorir, G., Rodríguez, J. C., Rui, H., Im, W., Lovell, S., Battaile, K. P., Alontaga, A. Y., Yukl, E. T., Moënné-Loccoz, P. & Rivera, M. Structural,

- NMR Spectroscopic, and Computational Investigation of Hemin Loading in the Hemophore HasAp from *Pseudomonas aeruginosa*. *Journal of the American Chemical Society* **132**, 9857–9872 (2010).
105. Freeman, S. L., Kwon, H., Portolano, N., Parkin, G., Venkatraman Girija, U., Basran, J., Fielding, A. J., Fairall, L., Svistunenko, D. A., Moody, P. C. E., Schwabe, J. W. R., Kyriacou, C. P. & Raven, E. L. Heme binding to human CLOCK affects interactions with the E-box. *Proceedings of the National Academy of Sciences of the United States of America* **116**, 19911–19916 (2019).
106. Mosure, S. A., Strutzenberg, T. S., Shang, J., Munoz-Tello, P., Solt, L. A., Griffin, P. R. & Kojetin, D. J. Structural basis for heme-dependent NCoR binding to the transcriptional repressor REV-ERB $\beta$ . *Science Advances* **7**, eabc6479 (2021).
107. Nam, D., Matsumoto, Y., Uchida, T., O'Brian, M. R. & Ishimori, K. Mechanistic insights into heme-mediated transcriptional regulation via a bacterial manganese-binding iron regulator, iron response regulator (Irr). *The Journal of Biological Chemistry* **295**, 11316–11325 (2020).
108. Fischer, S., Olsen, K. W., Nam, K. & Karplus, M. Unsuspected pathway of the allosteric transition in hemoglobin. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5608–5613 (2011).
109. Bringas, M., Petruk, A. A., Estrin, D. A., Capece, L. & Martí, M. A. Tertiary and quaternary structural basis of oxygen affinity in human hemoglobin as revealed by multiscale simulations. *Scientific Reports* **7**, 10926 (2017).
110. Jeschek, M., Panke, S. & Ward, T. R. Artificial Metalloenzymes on the Verge of New-to-Nature Metabolism. *Trends in Biotechnology* **36**, 60–72 (2018).
111. Akabori, S., Sakurai, S., Izumi, Y. & Fujii, Y. An Asymmetric Catalyst. *Nature* **178**, 323–324 (1956).
112. Coleman, J. E. Metal Ion Dependent Binding of Sulphonamide to Carbonic Anhydrase. *Nature* **214**, 193–194 (1967).
113. Cuatrecasas, P., Fuchs, S. & Anfinsen, C. B. Catalytic properties and specificity of the extracellular nuclease of *Staphylococcus aureus*. *The Journal of Biological Chemistry* **242**, 1541–1547 (1967).
114. Darnall, D. W. & Birnbaum, E. R. Rare earth metal ions as probes of calcium ion binding sites in proteins. Neodymium(3) acceleration of the activation of trypsinogen. *The Journal of Biological Chemistry* **245**, 6484–6486 (1970).

115. Gomez, J. E., Birnbaum, E. R. & Darnall, D. W. The metal ion acceleration of the conversion of trypsinogen to trypsin. Lanthanide ions as calcium ion substitutes. *Biochemistry* **13**, 3745–3750 (1974).
116. Yamamura, K. & Kaiser, E. T. Studies on the oxidase activity of copper(II) carboxypeptidase A. *Journal of the Chemical Society, Chemical Communications*, 830–831 (1976).
117. Rodríguez-Guerra, J., Alonso-Cotchico, L., Sciortino, G., Lledós, A. & Maréchal, J.-D. Computational Studies of Artificial Metalloenzymes: From Methods and Models to Design and Optimization. in *Artificial Metalloenzymes and MetalloDNAzymes in Catalysis* 99–136 (John Wiley & Sons, Ltd, 2018).
118. Davis, H. J. & Ward, T. R. Artificial Metalloenzymes: Challenges and Opportunities. *ACS Central Science* **5**, 1120–1136 (2019).
119. Rosati, F. & Roelfes, G. Artificial Metalloenzymes. *ChemCatChem* **2**, 916–927 (2010).
120. Sreenilayam, G., Moore, E. J., Steck, V. & Fasan, R. Stereoselective Olefin Cyclopropanation under Aerobic Conditions with an Artificial Enzyme Incorporating an Iron-Chlorin e6 Cofactor. *ACS Catalysis* **7**, 7629–7633 (2017).
121. Pordea, A., Creus, M., Panek, J., Duboc, C., Mathis, D., Novic, M. & Ward, T. R. Artificial Metalloenzyme for Enantioselective Sulfoxidation Based on Vanadyl-Loaded Streptavidin. *Journal of the American Chemical Society* **130**, 8085–8088 (2008).
122. Yu, Y., Cui, C., Liu, X., Petrik, I. D., Wang, J. & Lu, Y. A Designed Metalloenzyme Achieving the Catalytic Rate of a Native Enzyme. *Journal of the American Chemical Society* **137**, 11570–11573 (2015).
123. Hassan, I. S., Fuller, J. T., Dippon, V. N., Ta, A. N., Danneman, M. W., McNaughton, B. R., Alexandrova, A. N. & Rovis, T. Tuning through-space interactions via the secondary coordination sphere of an artificial metalloenzyme leads to enhanced Rh(III)-catalysis. *Chemical Science* **13**, 9220–9224 (2022).
124. Hamels, D. R. & Ward, T. R. Biomacromolecules as Ligands for Artificial Metalloenzymes. in *Comprehensive Inorganic Chemistry II* 2nd ed., 737–761 (Elsevier, 2013).

125. Jeong, W. J., Yu, J. & Song, W. J. Proteins as diverse, efficient, and evolvable scaffolds for artificial metalloenzymes. *Chemical Communications* **56**, 9586–9599 (2020).
126. Zhou, L., Bosscher, M., Zhang, C., Özçubukçu, S., Zhang, L., Zhang, W., Li, C. J., Liu, J., Jensen, M. P., Lai, L. & He, C. A protein engineered to bind uranyl selectively and with femtomolar affinity. *Nature Chemistry* **6**, 236–241 (2014).
127. Amrein, B., Schmid, M., Collet, G., Cuniasse, P., Gilardoni, F., Seebeck, F. P. & Ward, T. R. Identification of two-histidines one-carboxylate binding motifs in proteins amenable to facial coordination to metals. *Metallomics: Integrated Biometal Science* **4**, 379–388 (2012).
128. Kerns, S. A., Biswas, A., Minnetian, N. M. & Borovik, A. S. Artificial Metalloproteins: At the Interface between Biology and Chemistry. *JACS Au* **2**, 1252–1265 (2022).
129. Peacock, A. F. Recent advances in designed coiled coils and helical bundles with inorganic prosthetic groups—from structural to functional applications. *Current Opinion in Chemical Biology* **31**, 160–165 (2016).
130. Lombardi, A., Pirro, F., Maglio, O., Chino, M. & DeGrado, W. F. De Novo Design of Four-Helix Bundle Metalloproteins: One Scaffold, Diverse Reactivities. *Accounts of Chemical Research* **52**, 1148–1159 (2019).
131. Caldwell, S. J., Haydon, I. C., Piperidou, N., Huang, P.-S., Bick, M. J., Sjöström, H. S., Hilvert, D., Baker, D. & Zeymer, C. Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion. *Proceedings of the National Academy of Sciences* **117**, 30362–30369 (2020).
132. Huang, S.-Y. & Zou, X. Advances and Challenges in Protein-Ligand Docking. *International Journal of Molecular Sciences* **11**, 3016–3034 (2010).
133. Farid, T. A., Kodali, G., Solomon, L. A., Lichtenstein, B. R., Sheehan, M. M., Fry, B. A., Bialas, C., Ennist, N. M., Siedlecki, J. A., Zhao, Z., Stetz, M. A., Valentine, K. G., Anderson, J. L. R., Wand, A. J., Discher, B. M., Moser, C. C. & Dutton, P. L. Elementary tetrahelical protein design for diverse oxidoreductase functions. *Nature chemical biology* **9**, 826–833 (2013).
134. Koder, R. L., Ross Anderson, J. L., Solomon, L. A., Reddy, K. S., Moser, C. C. & Dutton, P. L. Design and engineering of an O(2) transport protein. *Nature* **458**, 305–309 (2009).

135. Di Meo, T., Ghattas, W., Herrero, C., Velours, C., Minard, P., Mahy, J.-P., Ricoux, R. & Urvoas, A.  $\alpha$ Rep A3: A Versatile Artificial Scaffold for Metalloenzyme Design. *Chemistry – A European Journal* **23**, 10156–10166 (2017).
136. Di Meo, T., Kariyawasam, K., Ghattas, W., Valerio-Lepiniec, M., Sciortino, G., Maréchal, J.-D., Minard, P., Mahy, J.-P., Urvoas, A. & Ricoux, R. Functionalized Artificial Bidomain Proteins Based on an  $\alpha$ -Solenoid Protein Repeat Scaffold: A New Class of Artificial Diels–Alderases. *ACS Omega* **4**, 4437–4447 (2019).
137. Fernández-Gacio, A., Codina, A., Fastrez, J., Riant, O. & Soumillion, P. Transforming Carbonic Anhydrase into Epoxide Synthase by Metal Exchange. *ChemBioChem* **7**, 1013–1016 (2006).
138. Konieczny, S., Leurs, M. & Tiller, J. C. Polymer Enzyme Conjugates as Chiral Ligands for Sharpless Dihydroxylation of Alkenes in Organic Solvents. *ChemBioChem* **16**, 83–90 (2015).
139. Köhler, V., Mao, J., Heinisch, T., Pordea, A., Sardo, A., Wilson, Y. M., Knörr, L., Creus, M., Prost, J.-C., Schirmer, T. & Ward, T. R. OsO<sub>4</sub>-Streptavidin: A Tunable Hybrid Catalyst for the Enantioselective cis-Dihydroxylation of Olefins. *Angewandte Chemie International Edition* **50**, 10863–10866 (2011).
140. Drienovská, I., Rioz-Martínez, A., Draksharapu, A. & Roelfes, G. Novel artificial metalloenzymes by in vivo incorporation of metal-binding unnatural amino acids. *Chemical Science* **6**, 770–776 (2014).
141. Davies, R. R. & Distefano, M. D. A Semisynthetic Metalloenzyme Based on a Protein Cavity That Catalyzes the Enantioselective Hydrolysis of Ester and Amide Substrates. *Journal of the American Chemical Society* **119**, 11643–11652 (1997).
142. Onoda, A., Fukumoto, K., Arlt, M., Bocola, M., Schwaneberg, U. & Hayashi, T. A rhodium complex-linked  $\beta$ -barrel protein as a hybrid biocatalyst for phenylacetylene polymerization. *Chemical Communications* **48**, 9756–9758 (2012).
143. Philippart, F., Arlt, M., Gotzen, S., Tenne, S.-J., Bocola, M., Chen, H.-H., Zhu, L., Schwaneberg, U. & Okuda, J. A hybrid ring-opening metathesis polymerization catalyst based on an engineered variant of the  $\beta$ -barrel protein FhuA. *Chemistry – A European Journal* **19**, 13865–13871 (2013).

144. Wu, Z. P. & Hilvert, D. Conversion of a protease into an acyl transferase: selenolsubtilisin. *Journal of the American Chemical Society* **111**, 4513–4514 (1989).
145. Nicholas, K. M., Wentworth, P., Harwig, C. W., Wentworth, A. D., Shafton, A. & Janda, K. D. A cofactor approach to copper-dependent catalytic antibodies. *Proceedings of the National Academy of Sciences* **99**, 2648–2653 (2002).
146. Klein, G., Humbert, N., Gradinaru, J., Ivanova, A., Gilardoni, F., Rusbandi, U. E. & Ward, T. R. Tailoring the active site of chemzymes by using a chemogenetic-optimization procedure: towards substrate-specific artificial hydrogenases based on the biotin-avidin technology. *Angewandte Chemie International Edition* **44**, 7764–7767 (2005).
147. Liang, A. D., Serrano-Plana, J., Peterson, R. L. & Ward, T. R. Artificial Metalloenzymes Based on the Biotin–Streptavidin Technology: Enzymatic Cascades and Directed Evolution. *Accounts of Chemical Research* **52**, 585–595 (2019).
148. Petrik, I. D., Liu, J. & Lu, Y. Metalloenzyme Design and Engineering through Strategic Modifications of Native Protein Scaffolds. *Current opinion in chemical biology* **0**, 67–75 (2014).
149. Himiyama, T. & Okamoto, Y. Artificial Metalloenzymes: From Selective Chemical Transformations to Biochemical Applications. *Molecules* **25**, 2989 (2020).
150. Vong, K., Nasibullin, I. & Tanaka, K. Exploring and Adapting the Molecular Selectivity of Artificial Metalloenzymes. *Bulletin of the Chemical Society of Japan* **94**, 382–396 (2021).
151. Zhang, J., Huang, X., Zhang, R. K. & Arnold, F. H. Enantiodivergent  $\alpha$ -Amino C–H Fluoroalkylation Catalyzed by Engineered Cytochrome P450s. *Journal of the American Chemical Society* **141**, 9798–9802 (2019).
152. Wang, Z. J., Peck, N. E., Renata, H. & Arnold, F. H. Cytochrome P450-Catalyzed Insertion of Carbenoids into N–H Bonds. *Chemical Science* **5**, 598–601 (2014).
153. Kan, S. B. J., Lewis, R. D., Chen, K. & Arnold, F. H. Directed Evolution of Cytochrome c for Carbon–Silicon Bond Formation: Bringing Silicon to Life. *Science* **354**, 1048–1051 (2016).
154. Knight, A. M., Kan, S. B. J., Lewis, R. D., Brandenburg, O. F., Chen, K. & Arnold, F. H. Diverse Engineered Heme Proteins Enable Stereodivergent

- Cyclopropanation of Unactivated Alkenes. *ACS Central Science* **4**, 372–377 (2018).
155. Dydio, P., Key, H. M., Nazarenko, A., Rha, J. Y.-E., Seyedkazemi, V., Clark, D. S. & Hartwig, J. F. An artificial metalloenzyme with the kinetics of native enzymes. *Science* **354**, 102–106 (2016).
156. Dydio, P., Key, H. M., Hayashi, H., Clark, D. S. & Hartwig, J. F. Chemoselective, Enzymatic C–H Bond Amination Catalyzed by a Cytochrome P450 Containing an Ir(Me)-PIX Cofactor. *Journal of the American Chemical Society* **139**, 1750–1753 (2017).
157. Gu, Y., Natoli, S. N., Liu, Z., Clark, D. S. & Hartwig, J. F. Site-Selective Functionalization of (sp<sup>3</sup>)C–H Bonds Catalyzed by Artificial Metalloenzymes Containing an Iridium-Porphyrin Cofactor. *Angewandte Chemie International Edition* **58**, 13954–13960 (2019).
158. Oohora, K., Meichin, H., Zhao, L., Wolf, M. W., Nakayama, A., Hasegawa, J.-y., Lehnert, N. & Hayashi, T. Catalytic Cyclopropanation by Myoglobin Reconstituted with Iron Porphycene: Acceleration of Catalysis due to Rapid Formation of the Carbene Species. *Journal of the American Chemical Society* **139**, 17265–17268 (2017).
159. Robertson, D. E., Farid, R. S., Moser, C. C., Urbauer, J. L., Mulholland, S. E., Pidikiti, R., Lear, J. D., Wand, A. J., DeGrado, W. F. & Dutton, P. L. Design and synthesis of multi-haem proteins. *Nature* **368**, 425–432 (1994).
160. Sharp, R. E., Moser, C. C., Rabanal, F. & Dutton, P. L. Design, synthesis, and characterization of a photoactivatable flavocytochrome molecular maquette. *Proceedings of the National Academy of Sciences* **95**, 10465–10470 (1998).
161. Stenner, R., Steventon, J. W., Seddon, A. & Anderson, J. L. R. A de novo peroxidase is also a promiscuous yet stereoselective carbene transferase. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 1419–1428 (2020).
162. Mann, S. I., Nayak, A., Gassner, G. T., Therien, M. J. & DeGrado, W. F. De Novo Design, Solution Characterization, and Crystallographic Structure of an Abiological MnPorphyrinBinding Protein Capable of Stabilizing a Mn(V) Species. *Journal of the American Chemical Society* **143**, 252–259 (2021).
163. Kariyawasam, K., Di Meo, T., Hammerer, F., Valerio-Lepiniec, M., Sciortino, G., Maréchal, J.-D., Minard, P., Mahy, J.-P., Urvoas, A. & Ricoux, R. An Artificial Hemoprotein with Inducible Peroxidase-and

- Monooxygenase-Like Activities. *Chemistry–A European Journal* **26**, 14929–14937 (2020).
164. Villarino, L., Splan, K. E., Reddem, E., Alonso-Cotchico, L., Gutiérrez de Souza, C., Lledós, A., Maréchal, J.-D., Thunnissen, A.-M. W. H. & Roelfes, G. An Artificial Heme Enzyme for Cyclopropanation Reactions. *Angewandte Chemie International Edition* **57**, 7785–7789 (2018).
165. Sansiaume-Dagousset, E., Urvoas, A., Chelly, K., Ghattas, W., Maréchal, J.-D., Mahy, J.-P. & Ricoux, R. Neocarzinostatin-based hybrid biocatalysts for oxidation reactions. *Dalton Transactions* **43**, 8344–8354 (2014).
166. Lewars, E. G. *Computational Chemistry* (Springer Netherlands, 2011).
167. Jensen, F. *Introduction to Computational Chemistry* 3rd ed. (Wiley, 2007).
168. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *Journal of Chemical Information and Modeling* **56**, 599–604 (2016).
169. Hu, L. & Ryde, U. Comparison of Methods to Obtain Force-Field Parameters for Metal Sites. *Journal of Chemical Theory and Computation* **7**, 2452–2463 (2011).
170. Li, P. & Merz, K. M. J. Metal Ion Modeling Using Classical Mechanics. *Chemical Reviews* **117**, 1564–1686 (2017).
171. Li, P., Song, L. F. & Merz, K. M. J. Systematic Parameterization of Monovalent Ions Employing the Nonbonded Model. *Journal of Chemical Theory and Computation* **11**, 1645–1657 (2015).
172. Li, P., Roberts, B. P., Chakravorty, D. K. & Merz, K. M. J. Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *Journal of Chemical Theory and Computation* **9**, 2733–2748 (2013).
173. Li, P., Song, L. F. & Merz, K. M. J. Parameterization of Highly Charged Metal Ions Using the 12-6-4 LJ-Type Nonbonded Model in Explicit Water. *The Journal of Physical Chemistry B* **119**, 883–895 (2015).
174. Duarte, F., Bauer, P., Barrozo, A., Amrein, B. A., Purg, M., Åqvist, J. & Kamerlin, S. C. L. Force Field Independent Metal Parameters Using a Nonbonded Dummy Model. *The Journal of Physical Chemistry B* **118**, 4351–4362 (2014).
175. Ramachandran, K., Deepa, G. & Namboori, K. *Computational Chemistry and Molecular Modeling: Principles and Applications* - K. I.

- Ramachandran, Gopakumar Deepa, Krishnan Namboori - Google Libros Springer (2008).
176. Kedziera, D. & Kaczmarek-Kedziera, A. Remarks on Wave Function Theory and Methods. in *Handbook of Computational Chemistry* 55–93 (Springer Netherlands, 2012).
177. Ban, F., Rankin, K. N., Gauld, J. W. & Boyd, R. J. Recent applications of density functional theory calculations to biomolecules. *Theoretical Chemistry Accounts* **108**, 1–11 (2002).
178. Salahub, D. R., de la Lande, A., Goursot, A., Zhang, R. & Zhang, Y. Recent Progress in Density Functional Methodology for Biomolecular Modeling. in *Applications of Density Functional Theory to Biological and Bioinorganic Chemistry* 1–64 (Springer, 2013).
179. Karabencheva-Christova, T. Combined Quantum Mechanical and Molecular Mechanical Modelling of Biomolecular Interactions (Academic Press, 2015).
180. Aminpour, M., Montemagno, C. & Tuszynski, J. A. An Overview of Molecular Modeling for Drug Discovery with Specific Illustrative Examples of Applications. *Molecules* **24**, 1693 (2019).
181. Ahmadi, S., Barrios Herrera, L., Chehelamirani, M., Hostaš, J., Jalife, S. & Salahub, D. R. Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. *International Journal of Quantum Chemistry* **118**, e25558 (2018).
182. Sumner, S., Söderhjelm, P. & Ryde, U. Effect of Geometry Optimizations on QM-Cluster and QM/MM Studies of Reaction Energies in Proteins. *Journal of Chemical Theory and Computation* **9**, 4205–4214 (2013).
183. Prejanò, M., Marino, T. & Russo, N. QM Cluster or QM/MM in Computational Enzymology: The Test Case of LigW-Decarboxylase. *Frontiers in Chemistry* **6**, 249 (2018).
184. Hanreich, S., Bonandi, E. & Drienovská, I. Design of Artificial Enzymes: Insights into Protein Scaffolds. *ChemBioChem* **24**, e202200566 (2023).
185. Kries, H., Blomberg, R. & Hilvert, D. De novo enzymes by computational design. *Current Opinion in Chemical Biology* **17**, 221–228 (2013).
186. Zanghellini, A., Jiang, L., Wollacott, A. M., Cheng, G., Meiler, J., Althoff, E. A., Röthlisberger, D. & Baker, D. New algorithms and an in silico benchmark for computational enzyme design. *Protein Science* **15**, 2785–2794 (2006).

187. Röthlisberger, D., Khersonsky, O., Wollacott, A. M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J. L., Althoff, E. A., Zanghellini, A., Dym, O., Albeck, S., Houk, K. N., Tawfik, D. S. & Baker, D. Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
188. Siegel, J. B., Zanghellini, A., Lovick, H. M., Kiss, G., Lambert, A. R., St.Clair, J. L., Gallaher, J. L., Hilvert, D., Gelb, M. H., Stoddard, B. L., Houk, K. N., Michael, F. E. & Baker, D. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **329**, 309–313 (2010).
189. Mills, J. H., Khare, S. D., Bolduc, J. M., Forouhar, F., Mulligan, V. K., Lew, S., Seetharaman, J., Tong, L., Stoddard, B. L. & Baker, D. Computational Design of an Unnatural Amino Acid Dependent Metalloprotein with Atomic Level Accuracy. *Journal of the American Chemical Society* **135**, 10.1021/ja403503m (2013).
190. Khare, S. D., Kipnis, Y., Greisen, P. J., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y., Gallaher, J. L., Silman, I., Leader, H., Sussman, J. L., Stoddard, B. L., Tawfik, D. S. & Baker, D. Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nature Chemical Biology* **8**, 294–300 (2012).
191. Heinisch, T., Pellizzoni, M., Dürrenberger, M., Tinberg, C. E., Köhler, V., Klehr, J., Häussinger, D., Baker, D. & Ward, T. R. Improving the Catalytic Performance of an Artificial Metalloenzyme by Computational Design. *Journal of the American Chemical Society* **137**, 10414–10419 (2015).
192. Drienovská, I., Alonso-Cotchico, L., Vidossich, P., Lledós, A., Maréchal, J.-D. & Roelfes, G. Design of an enantioselective artificial metallo-hydratase enzyme containing an unnatural metal-binding amino acid. *Chemical Science* **8**, 7228–7235 (2017).
193. Sánchez-Aparicio, J.-E., Sciortino, G., Mates-Torres, E., Lledós, A. & Maréchal, J.-D. Successes and challenges in multiscale modelling of artificial metalloenzymes: the case study of POP-Rh2 cyclopropanase. *Faraday Discussions* **234**, 349–366 (2022).
194. Alonso-Cotchico, L., Rodríguez-Guerra Pedregal, J., Lledós, A. & Maréchal, J.-D. The Effect of Cofactor Binding on the Conformational Plasticity of the Biological Receptors in Artificial Metalloenzymes: The Case Study of LmrR. *Frontiers in Chemistry* **7**, 211 (2019).

195. George, G. N. & Pickering, I. J. X-Ray Absorption Spectroscopy of Metals in Biology. in *Encyclopedia of Biophysics* 2762–2767 (Springer, **2013**).
196. Shi, W., Punta, M., Bohon, J., Sauder, J. M., D’Mello, R., Sullivan, M., Toomey, J., Abel, D., Lippi, M., Passerini, A., Frasconi, P., Burley, S. K., Rost, B. & Chance, M. R. Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy. *Genome Research* **21**, 898–907 (**2011**).
197. Handing, K. B., Niedzialkowska, E., Shabalin, I. G., Kuhn, M. L., Zheng, H. & Minor, W. Characterizing metal binding sites in proteins with X-ray crystallography. *Nature protocols* **13**, 1062–1090 (**2018**).
198. Li, H. & Sun, H. NMR studies of metalloproteins. *Topics in Current Chemistry* **326**, 69–98 (**2012**).
199. Passerini, A., Lippi, M. & Frasconi, P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Research* **39**, W288–W292 (suppl\_2 **2011**).
200. Zhao, W., Xu, M., Liang, Z., Ding, B., Niu, L., Liu, H. & Teng, M. Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics* **27**, 1262–1268 (**2011**).
201. Sodhi, J. S., Bryson, K., McGuffin, L. J., Ward, J. J., Wernisch, L. & Jones, D. T. Predicting Metal-binding Site Residues in Low-resolution Structural Models. *Journal of Molecular Biology* **342**, 307–320 (**2004**).
202. Lin, Y.-F., Cheng, C.-W., Shih, C.-S., Hwang, J.-K., Yu, C.-S. & Lu, C.-H. MIB: Metal Ion-Binding Site Prediction and Docking Server. *Journal of Chemical Information and Modeling* **56**, 2287–2291 (**2016**).
203. Lu, C.-H., Chen, C.-C., Yu, C.-S., Liu, Y.-Y., Liu, J.-J., Wei, S.-T. & Lin, Y.-F. MIB2: metal ion-binding site prediction and modeling server. *Bioinformatics* **38**, 4428–4429 (**2022**).
204. Haberal, İ. & Oğul, H. Prediction of Protein Metal Binding Sites Using Deep Neural Networks. *Molecular Informatics* **38**, 1800169 (**2019**).
205. Mohamadi, A., Cheng, T., Jin, L., Wang, J., Sun, H. & Koohi-Moghadam, M. An ensemble 3D deep-learning model to predict protein metal-binding site. *Cell Reports Physical Science* **3**, 101046 (**2022**).
206. Liou, Y.-F., Charoenkwan, P., Srinivasulu, Y. S., Vasylenko, T., Lai, S.-C., Lee, H.-C., Chen, Y.-H., Huang, H.-L. & Ho, S.-Y. SCMHP: prediction and analysis of heme binding proteins using propensity scores of dipeptides. *BMC Bioinformatics* **15**, S4 (**2014**).

207. Liu, R. & Hu, J. HemeBIND: a novel method for heme binding residue prediction by combining structural and sequence information. *BMC Bioinformatics* **12**, 207 (2011).
208. Cramer, C. J. Essentials of Computational Chemistry: Theories and Models, 2nd ed. (Wiley, 2002).
209. Leach, A. R. Molecular modelling: principles and applications 2nd ed (Prentice Hall, 2001).
210. Sholl, D. S. & Steckel, J. A. Density Functional Theory: A Practical Introduction (2009).
211. Perdew, J. P. & Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings* **577**, 1–20 (2001).
212. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *The Journal of chemical physics* **132**, 154104 (2010).
213. Cramer, C. J. & Truhlar, D. G. Density functional theory for transition metals and transition metal chemistry. *Physical Chemistry Chemical Physics* **11**, 10757 (2009).
214. Siegbahn, P. E. M. The performance of hybrid DFT for mechanisms involving transition metal complexes in enzymes. *JBIC Journal of Biological Inorganic Chemistry* **11**, 695–701 (2006).
215. Pengfei Li and Kenneth M. Merz Jr. Metal Ion Modeling Using Classical Mechanics. *Chemical Reviews* **117**, 1564–1686 (2017).
216. Dolg, M., Wedig, U., Stoll, H. & Preuss, H. Energy-adjusted ab initio pseudopotentials for the first row transition elements. *The Journal of Chemical Physics* **86**, 866–872 (1987).
217. Hay, P. J. & Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *The Journal of Chemical Physics* **82**, 299–310 (1985).
218. Vanommeslaeghe, K., Guvench, O. & MacKerell, A. D. Molecular Mechanics. *Current Pharmaceutical Design* **20**, 3281–3292 (2014).
219. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **97**, 10269–10280 (1993).

220. Burton, V. J., Deeth, R. J., Kemp, C. M. & Gilbert, P. J. Molecular Mechanics for Coordination Complexes: The Impact of Adding d-Electron Stabilization Energies. *Journal of the American Chemical Society* **117**, 8407–8415 (1995).
221. Gresh, N., Cisneros, G. A., Darden, T. A. & Piquemal, J.-P. Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand-Macromolecule Complexes. A Bottom-Up Strategy. *Journal of Chemical Theory and Computation* **3**, 1960–1986 (2007).
222. Root, D. M., Landis, C. R. & Cleveland, T. Valence bond concepts applied to the molecular mechanics description of molecular shapes. 1. Application to nonhypervalent molecules of the P-block. *Journal of the American Chemical Society* **115**, 4201–4209 (1993).
223. Huang, J., Devereux, M., Hofmann, F. & Meuwly, M. Computational Organometallic Chemistry with Force Fields. in *Computational Organometallic Chemistry* 19–46 (Springer, 2012).
224. Deeth, R. J. Molecular Mechanics for Transition Metal Centers: From Coordination Complexes To Metalloproteins. in *Advances in Inorganic Chemistry* 1–39 (Academic Press, 2010).
225. Lin, F. & Wang, R. Systematic Derivation of AMBER Force Field Parameters Applicable to Zinc-Containing Systems. *Journal of Chemical Theory and Computation* **6**, 1852–1870 (2010).
226. Ewald, P. Die Berechnung Optischer und Elektrostatischer Gitterpotentiale. *Annalen der Physik* **369**, 253–287 (1921).
227. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577–8593 (1995).
228. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).
229. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics* **81**, 511–519 (1984).
230. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **31**, 1695–1697 (1985).
231. Allison, J. R. Computational methods for exploring protein conformations. *Biochemical Society Transactions* **48**, 1707–1724 (2020).

232. Kamenik, A. S., Linker, S. M. & Riniker, S. Enhanced sampling without borders: on global biasing functions and how to reweight them. *Physical Chemistry Chemical Physics* **24**, 1225–1236 (2022).
233. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *The Journal of Chemical Physics* **151**, 070902 (2019).
234. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *The Journal of Chemical Physics* **120**, 11919–11929 (2004).
235. Miao, Y., Feher, V. A. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation* **11**, 3584–3595 (2015).
236. Wang, J., Arantes, P. R., Bhattarai, A., Hsu, R. V., Pawnikar, S., Huang, Y.-M. M., Palermo, G. & Miao, Y. Gaussian accelerated molecular dynamics (GaMD): principles and applications. *Wiley Interdisciplinary Reviews. Computational Molecular Science* **11**, e1521 (2021).
237. Miao, Y. & McCammon, J. A. Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. *Annual reports in computational chemistry* **13**, 231–278 (2017).
238. Miao, Y., Bhattarai, A. & Wang, J. Ligand Gaussian Accelerated Molecular Dynamics (LiGaMD): Characterization of Ligand Binding Thermodynamics and Kinetics. *Journal of Chemical Theory and Computation* **16**, 5526–5547 (2020).
239. Wang, J. & Miao, Y. Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): Enhanced sampling and free energy and kinetics calculations of peptide binding. *The Journal of Chemical Physics* **153**, 154109 (2020).
240. Grossfield, A. & Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual reports in computational chemistry* **5**, 23–48 (2009).
241. Grossfield, A., Patrone, P. N., Roe, D. R., Schultz, A. J., Siderius, D. & Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science* **1** (2019).
242. Smith, L. J., Daura, X. & van Gunsteren, W. F. Assessing equilibration and convergence in biomolecular simulations. *Proteins: Structure, Function, and Bioinformatics* **48**, 487–496 (2002).

243. David, C. C. & Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods in molecular biology* **1084**, 193–226 (2014).
244. Martínez, L. Automatic Identification of Mobile and Rigid Substructures in Molecular Dynamics Simulations and Fractional Structural Fluctuation Analysis. *PLoS ONE* **10**, e0119264 (2015).
245. Knapp, B., Ospina, L. & Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *Journal of Chemical Theory and Computation* **14**, 6127–6138 (2018).
246. Maden, S., Selin, S. & Acuner, S. Fundamentals of Molecular Docking and Comparative Analysis of Protein–Small-Molecule Docking Approaches. in (2022).
247. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular Docking: A powerful approach for structure-based drug discovery. *Current computer-aided drug design* **7**, 146–157 (2011).
248. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–623 (2003).
249. Rodríguez-Guerra Pedregal, J., Sciortino, G., Guasp, J., Municoy, M. & Maréchal, J.-D. GaudiMM: A modular multi-objective platform for molecular modeling. *Journal of Computational Chemistry* **38**, 2118–2126 (2017).
250. Clark, D. E. & Westhead, D. R. Evolutionary algorithms in computer-aided molecular design. *Journal of Computer-Aided Molecular Design* **10**, 337–358 (1996).
251. Liu, J. & Wang, R. Classification of Current Scoring Functions. *Journal of Chemical Information and Modeling* **55**, 475–482 (2015).
252. Li, J., Fu, A. & Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences* **11**, 320–328 (2019).
253. Sapundzhi, F., Prodanova, K. & Lazarova, M. Survey of the scoring functions for protein-ligand docking. *AIP Conference Proceedings* **2172**, 100008 (2019).
254. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W. & Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **52**, 609–623 (2003).
255. Yang, C., Chen, E. A. & Zhang, Y. Protein–Ligand Docking in the Machine-Learning Era. *Molecules* **27**, 4568 (2022).

256. 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. *Journal of computational chemistry* **39**, 42–51 (2018).
257. TIOBE Index. <https://www.tiobe.com/tiobe-index/>. (2023).
258. Rassokhin, D. The C programming language in cheminformatics and computational chemistry. *Journal of Cheminformatics* **12**, 10 (2020).
259. Wißbrock, A., Paul George, A. A., Brewitz, H. H., Köhl, T. & Imhof, D. The molecular basis of transient heme-protein interactions: analysis, concept and implementation. *Bioscience Reports* **39**, BSR20181940 (2019).
260. Pfeil, W., Nölting, B. O. & Jung, C. Apocytochrome P450cam is a native protein with some intermediate-like properties. *Biochemistry* **32**, 8856–8862 (1993).
261. Arnoux, P., Haser, R., Izadi, N., Lecroisey, A., Delepierre, M., Wandersman, C. & Czjzek, M. The crystal structure of HasA, a hemophore secreted by *Serratia marcescens*. *Nature Structural Biology* **6**, 516–520 (1999).
262. Létoffé, S., Ghigo, J. M. & Wandersman, C. Iron acquisition from heme and hemoglobin by a *Serratia marcescens* extracellular protein. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 9876–9880 (1994).
263. Ghigo, J. M., Létoffé, S. & Wandersman, C. A new type of hemophore-dependent heme acquisition system of *Serratia marcescens* reconstituted in *Escherichia coli*. *Journal of Bacteriology* **179**, 3572–3579 (1997).
264. Paquelin, A., Ghigo, J. M., Bertin, S. & Wandersman, C. Characterization of HasB, a *Serratia marcescens* TonB-like protein specifically involved in the haemophore-dependent haem acquisition system. *Molecular Microbiology* **42**, 995–1005 (2001).
265. Parrow, N. L., Fleming, R. E. & Minnick, M. F. Sequestration and Scavenging of Iron in Infection. *Infection and Immunity* **81**, 3503–3514 (2013).
266. Létoffé, S., Redeker, V. & Wandersman, C. Isolation and characterization of an extracellular haem-binding protein from *Pseudomonas aeruginosa* that shares function and sequence similarities with the *Serratia marcescens* HasA haemophore. *Molecular Microbiology* **28**, 1223–1234 (1998).
267. Idei, A., Kawai, E., Akatsuka, H. & Omori, K. Cloning and Characterization of the *Pseudomonas fluorescens* ATP-Binding Cassette

- Exporter, HasDEF, for the Heme Acquisition Protein HasA. *Journal of Bacteriology* **181**, 7545–7551 (1999).
268. Rossi, M. S., Fetherston, J. D., Létoffé, S., Carniel, E., Perry, R. D. & Ghigo, J. M. Identification and characterization of the hemophore-dependent heme acquisition system of *Yersinia pestis*. *Infection and Immunity* **69**, 6707–6717 (2001).
269. Kumar, R., Lovell, S., Matsumura, H., Battaile, K. P., Moënné-Loccoz, P. & Rivera, M. The hemophore HasA from *Yersinia pestis* (HasAyp) coordinates hemin with a single residue, Tyr75, and with minimal conformational change. *Biochemistry* **52**, 2705–2707 (2013).
270. Arnoux, P., Haser, R., Izadi-Pruneyre, N., Lecroisey, A. & Czjzek, M. Functional aspects of the heme bound hemophore HasA by structural analysis of various crystal forms. *Proteins: Structure, Function, and Bioinformatics* **41**, 202–210 (2000).
271. Tiessler-Sala, L., Sciortino, G., Alonso-Cotchico, L., Masgrau, L., Lledós, A. & Maréchal, J.-D. Getting Deeper into the Molecular Events of Heme Binding Mechanisms: A Comparative Multi-level Computational Study of HasAsm and HasAyp Hemophores. *Inorganic Chemistry* **61**, 17068–17079 (2022).
272. Kumar, R., Qi, Y., Matsumura, H., Lovell, S., Yao, H., Battaile, K. P., Im, W., Moënné-Loccoz, P. & Rivera, M. Replacing Arginine 33 for Alanine in the Hemophore HasA from *Pseudomonas aeruginosa* Causes Closure of the H32 Loop in the Apo-Protein. *Biochemistry* **55**, 2622–2631 (2016).
273. Exner, T. E., Becker, S., Becker, S., Boniface-Guiraud, A., Delepelaire, P., Diederichs, K. & Welte, W. Binding of HasA by its transmembrane receptor HasR follows a conformational funnel mechanism. *European biophysics journal: EBJ* **49**, 39–57 (2020).
274. Schlitter, J., Engels, M. & Krüger, P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics* **12**, 84–89 (1994).
275. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
276. Wolff, N., Izadi-Pruneyre, N., Couprie, J., Habeck, M., Linge, J., Rieping, W., Wandersman, C., Nilges, M., Delepierre, M. & Lecroisey, A. Comparative analysis of structural and dynamic properties of the loaded and unloaded

- hemophore HasA: functional implications. *Journal of Molecular Biology* **376**, 517–525 (2008).
277. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C. & Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).
278. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Research* **40**, 537–541 (W1 2012).
279. Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., III, Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Ghoreishi, D., Gilson, M. K., Gohlke, H., Goetz, A. W., Greene, D., Harris, R., Homeyer, N., Huang, Y., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T. S., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., D.J. Mermelstein, Merz, K. M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., R. Qi, Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., Smith, J., SalomonFerrer, R., Swails, J., Walker, R. C., Wang, J., Wei, H., Wolf, R. M., Wu, X., Xiao, L., York, D. M. & Kollman, P. A. AMBER 2018. **2018**.
280. Bayly, C. I., Merz, K. M., Ferguson, D. M., Cornell, W. D., Fox, T., Caldwell, J. W., Kollman, P. A., Cieplak, P., Gould, I. R. & Spellmeyer, D. C. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **117**, 5179–5197 (1995).
281. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. ACS Publications. (2023).
282. Seminario, J. M. Calculation of intramolecular force fields from second-derivative tensors. *International Journal of Quantum Chemistry* **60**, 1271–1277 (1996).
283. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O.,

- Nakai, H., Vreven, T., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, J. M., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J. & Fox, D. J. Gaussian 09. **2009**.
284. Marenich, A. V., Cramer, C. J. & Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *Journal of Physical Chemistry B* **113**, 6378–6396 (**2009**).
285. Ehlers, A. W., Böhme, M., Dapprich, S., Gobbi, A., Höllwarth, A., Jonas, V., Köhler, K. F., Stegmann, R., Veldkamp, A. & Frenking, G. A set of f-polarization functions for pseudo-potential basis sets of the transition metals ScCu, YAg and LaAu. *Chemical Physics Letters* **208**, 111–114 (**1993**).
286. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L. P., Lane, T. J. & Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (**2015**).
287. Venkatakrishnan, A. J., Fonseca, R., Ma, A. K., Hollingsworth, S. A., Chemparathy, A., Hilger, D., Kooistra, A. J., Ahmari, R., Madan, M., 6, B., Kobilka, B. K. & Dror, R. O. Uncovering patterns of atomic interactions in static and dynamic structures of proteins. *bioRxiv*, 840694 (**2019**).
288. Miao, Y., Sinko, W., Pierce, L., Bucher, D., Walker, R. C. & McCammon, J. A. Improved Reweighting of Accelerated Molecular Dynamics Simulations for Free Energy Calculation. *Journal of Chemical Theory and Computation* **10**, 2677–2689 (**2014**).
289. Schneider, S., Marles-Wright, J., Sharp, K. H. & Paoli, M. Diversity and conservation of interactions for binding heme in b-type heme proteins. *Natural Product Reports* **24**, 621–630 (**2007**).
290. Liu, R. & Hu, J. Computational Prediction of Heme-Binding Residues by Exploiting Residue Interaction Network. *PLOS ONE* **6**, e25560 (**2011**).

291. Zhang, J., Chai, H., Gao, B., Yang, G. & Ma, Z. HEMEsPred: Structure-Based Ligand-Specific Heme Binding Residues Prediction by Using Fast-Adaptive Ensemble Learning Scheme. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **15**, 147–156 (2018).
292. Xiong, Y., Liu, J., Zhang, W. & Zeng, T. Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome Science* **10**, S20 (Suppl 1 2012).
293. Paul George, A. A., Lacerda, M., Syllwasschy, B. F., Hopp, M.-T., Wißbrock, A. & Imhof, D. HeMoQuest: a webserver for qualitative prediction of transient heme binding to protein motifs. *BMC Bioinformatics* **21**, 124 (2020).
294. Sánchez-Aparicio, J.-E., Tiessler-Sala, L., Velasco-Carneros, L., Roldán-Martín, L., Sciortino, G. & Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *Journal of Chemical Information and Modeling* **61**, 311–323 (2021).
295. Guerra, J. V. d. S., Ribeiro-Filho, H. V., Jara, G. E., Bortot, L. O., Pereira, J. G. d. C. & Lopes-de-Oliveira, P. S. pyKVFinder: an efficient and integrable Python package for biomolecular cavity detection and characterization in data science. *BMC Bioinformatics* **22**, 607 (2021).
296. Rodríguez-Guerra Pedregal, J. & Maréchal, J.-D. PyChimera: use UCSF Chimera modules in any Python 2.7 project. *Bioinformatics (Oxford, England)* **34**, 1784–1785 (2018).
297. Lloyd, S. Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129–137 (1982).
298. Scott, D. W. Kernel Density Estimation. in *Wiley StatsRef: Statistics Reference Online* 1–7 (John Wiley & Sons, Ltd, 2018).
299. Moré, J. J. The Levenberg-Marquardt algorithm: Implementation and theory. in *Numerical Analysis* (ed Watson, G. A.) (Springer, 1978), 105–116.
300. Shayeghi, M., Latunde-Dada, G. O., Oakhill, J. S., Laftah, A. H., Takeuchi, K., Halliday, N., Khan, Y., Warley, A., McCann, F. E., Hider, R. C., Frazer, D. M., Anderson, G. J., Vulpe, C. D., Simpson, R. J. & McKie, A. T. Identification of an Intestinal Heme Transporter. *Cell* **122**, 789–801 (2005).
301. Qiu, A., Jansen, M., Sakaris, A., Min, S. H., Chattopadhyay, S., Tsai, E., Sandoval, C., Zhao, R., Akabas, M. H. & Goldman, I. D. Identification of an Intestinal Folate Transporter and the Molecular Basis for Hereditary Folate Malabsorption. *Cell* **127**, 917–928 (2006).

302. Inoue, K., Nakai, Y., Ueda, S., Kamigaso, S., Ohta, K.-y., Hatakeyama, M., Hayashi, Y., Otagiri, M. & Yuasa, H. Functional characterization of PCFT/HCP1 as the molecular entity of the carrier-mediated intestinal folate transport system in the rat model. *American Journal of Physiology-Gastrointestinal and Liver Physiology* **294**, G660–G668 (2008).
303. Yanatori, I., Tabuchi, M., Kawai, Y., Yasui, Y., Akagi, R. & Kishi, F. Heme and non-heme iron transporters in non-polarized and polarized cells. *BMC Cell Biology* **11**, 39 (2010).
304. Laftah, A. H., Latunde-Dada, G. O., Fakih, S., Hider, R. C., Simpson, R. J. & McKie, A. T. Haem and folate transport by proton-coupled folate transporter/haem carrier protein 1 (SLC46A1). *British Journal of Nutrition* **101**, 1150–1156 (2008).
305. Le Blanc, S., Garrick, M. D. & Arredondo, M. Heme carrier protein 1 transports heme and is involved in heme-Fe metabolism. *American Journal of Physiology-Cell Physiology* **302**, C1780–C1785 (2012).
306. Parker, J. L., Deme, J. C., Kuteyi, G., Wu, Z., Huo, J., Goldman, I. D., Owens, R. J., Biggin, P. C., Lea, S. M. & Newstead, S. Structural basis of antifolate recognition and transport by PCFT. *Nature* **595**, 130–134 (2021).
307. Krebs, W. G. & Gerstein, M. Survey and summary: The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Research* **28**, 1665–1675 (2000).
308. Mallin, H., Hesticová, M., Reuter, R. & Ward, T. R. Library design and screening protocol for artificial metalloenzymes based on the biotin-streptavidin technology. *Nature Protocols* **11**, 835–852 (2016).
309. Michael Green, N. Avidin and streptavidin. in *Methods in Enzymology* (eds Wilchek, M. & Bayer, E. A.) 51–67 (Academic Press, 1990).
310. Le Trong, I., Wang, Z., Hyre, D. E., Lybrand, T. P., Stayton, P. S. & Stenkamp, R. E. Streptavidin and its biotin complex at atomic resolution. *Biological Crystallography* **67**, 813–821 (Pt 9 2011).
311. Heinisch, T. & Ward, T. R. Artificial Metalloenzymes Based on the Biotin–Streptavidin Technology: Challenges and Opportunities. *Accounts of Chemical Research* **49**, 1711–1721 (2016).
312. Waser, V., Mukherjee, M., Tachibana, R., Igareta, N. V. & Ward, T. R. An Artificial [Fe4S4]-Containing Metalloenzyme for the Reduction of CO<sub>2</sub> to

- Hydrocarbons. *Journal of the American Chemical Society* **145**, 14823–14830 (2023).
313. Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Petersson, G. A., Nakatsuji, H., Li, X., Caricato, M., Marenich, A. V., Bloino, J., Janesko, B. G., Gomperts, R., Mennucci, B., Hratchian, H. P., Ortiz, J. V., Izmaylov, A. F., Sonnenberg, J. L., Williams-Young, D., Ding, F., Lipparini, F., Egidi, F., Goings, J., Peng, B., Petrone, A., Henderson, T., Ranasinghe, D., Zakrzewski, V. G., Gao, J., Rega, N., Zheng, G., Liang, W., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Throssell, K., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M. J., Heyd, J. J., Brothers, E. N., Kudin, K. N., Staroverov, V. N., Keith, T. A., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A. P., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Millam, J. M., Klene, M., Adamo, C., Cammi, R., Ochterski, J. W., Martin, R. L., Morokuma, K., Farkas, O., Foresman, J. B. & Fox, D. J. AMBER 2018. **2016**.
314. Petersson, G. A., Bennett, A., Tensfeldt, T. G., Al-Laham, M. A., Shirley, W. A. & Mantzaris, J. A complete basis set model chemistry. I. The total energies of closed-shell atoms and hydrides of the first-row elements. *The Journal of Chemical Physics* **89**, 2193–2218 (1988).
315. Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics* **8**, 1057–1065 (2006).
316. Chatterjee, A., Mallin, H., Klehr, J., Vallapurackal, J., Finke, A. D., Vera, L., Marsh, M. & Ward, T. R. An enantioselective artificial Suzukiase based on the biotin-streptavidin technology. *Chemical Science* **7**, 673–677 (2016).
317. Shapovalov, M. V. & Dunbrack, R. L. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure (London, England : 1993)* **19**, 844–858 (2011).
318. Webb Benjamin, A. S. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*. **15**, 5.6.1–5.6.30 (2016).
319. Krammer, A., Kirchhoff, P. D., Jiang, X., Venkatachalam, C. M. & Waldman, M. LigScore: A novel scoring function for predicting binding affinities. *Journal of Molecular Graphics and Modelling* **23**, 395–407 (2005).

320. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**, 727–748 (2002).
321. Case, D. A., Belfon, K., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham, T. E., III, Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Giambasu, G., Gilson, M. K., Gohlke, H., Harris, A. W. G. a., Izadi, S., Izmailov, S. A., Kasavajhala, K., Kovalenko, A., Krasny, R., Kurtzman, T., Lee, T. S., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Man, V., Merz, K. M., Miao, Y., Mikhailovskii, O., Monard, G., Nguyen, H., Onufriev, A., F. Pan, Pantano, S., Qi, R., Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., N.R. Skrynnikov, Smith, J., Swails, J., Walker, R. C., Wang, J., Wilson, L., Wolf, R. M., Wu, X., Xiong, Y., Xue, Y., York, D. M. & Kollman, P. A. AMBER 2020. **2020**.
322. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **11**, 3696–3713 (2015).
323. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *Journal of Computational Chemistry* **25**, 1157–1174 (2004).
324. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **79**, 926–935 (1983).
325. Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. *Journal of Physical Chemistry* **97**, 10269–10280 (1993).
326. Pedregal, J. R. G., Alonso-Cotchico, L., Velasco-Carneros, L. & Maréchal, J. D. OMMProtocol: A command line application to launch molecular dynamics simulations with OpenMM. *ChemRxiv* (2018).
327. Brünger, A., Brooks, C. L. & Karplus, M. Stochastic boundary conditions for molecular dynamics simulations of ST2 water. *Chemical Physics Letters* **105**, 495–500 (1984).
328. Schneider, T. & Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Physical Review B* **17**, 1302–1322 (1978).

329. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* **23**, 327–341 (1977).
330. Scinto, S. L., Bilodeau, D. A., Hincapie, R., Lee, W., Nguyen, S. S., Xu, M., am Ende, C. W., Finn, M. G., Lang, K., Lin, Q., Pezacki, J. P., Prescher, J. A., Robillard, M. S. & Fox, J. M. Bioorthogonal chemistry. *Nature Reviews Methods Primers* **1**, 1–23 (2021).
331. Smeenk, M. L. W. J., Agramunt, J. & Bongers, K. M. Recent developments in bioorthogonal chemistry and the orthogonality within. *Current Opinion in Chemical Biology* **60**, 79–88 (2021).
332. Bird, R. E., Lemmel, S. A., Yu, X. & Zhou, Q. A. Bioorthogonal Chemistry and Its Applications. *Bioconjugate Chemistry* **32**, 2457–2479 (2021).
333. Dorel, R. & Echavarren, A. M. Gold(I)-Catalyzed Activation of Alkynes for the Construction of Molecular Complexity. *Chemical Reviews* **115**, 9028–9072 (2015).
334. Zuccarello, G., Escofet, I., Caniparoli, U. & Echavarren, A. M. New-Generation Ligand Design for the Gold-Catalyzed Asymmetric Activation of Alkynes. *ChemPlusChem* **86**, 1283–1296 (2021).
335. Marín-Luna, M., Nieto Faza, O. & Silva López, C. Gold-Catalyzed Homogeneous (Cyclo)Isomerization Reactions. *Frontiers in Chemistry* **7** (2019).
336. Zhao, X., Rudolph, M. & Hashmi, A. S. K. Dual gold catalysis – an update. *Chemical Communications* **55**, 12127–12135 (2019).
337. Chang, T.-C., Vong, K., Yamamoto, T. & Tanaka, K. Prodrug Activation by Gold Artificial Metalloenzyme-Catalyzed Synthesis of Phenanthridinium Derivatives via Hydroamination. *Angewandte Chemie International Edition* **60**, 12446–12454 (2021).
338. Vidal, C., Tomás-Gamasa, M., Destito, P., López, F. & Mascareñas, J. L. Concurrent and orthogonal gold(I) and ruthenium(II) catalysis inside living cells. *Nature Communications* **9**, 1913 (2018).
339. Gimeno, A., Medio-Simón, M., de Arellano, C. R., Asensio, G. & Cuenca, A. B. NHC-Stabilized Gold(I) Complexes: Suitable Catalysts for 6-exo-dig Heterocyclization of 1-(o-Ethynylaryl)ureas. *Organic Letters* **12**, 1900–1903 (2010).
340. Gimeno, A., Cuenca, A. B., Suárez-Pantiga, S., de Arellano, C. R., Medio-Simón, M. & Asensio, G. Competitive Gold-Activation Modes in Terminal

- Alkynes: An Experimental and Mechanistic Study. *Chemistry – A European Journal* **20**, 683–688 (2014).
341. Vreeken, V., Broere, D. L. J., Jans, A. C. H., Lankelma, M., Reek, J. N. H., Siegler, M. A. & van der Vlugt, J. I. Well-Defined Dinuclear Gold Complexes for Preorganization-Induced Selective Dual Gold Catalysis. *Angewandte Chemie International Edition* **55**, 10042–10046 (2016).
342. Morris, J. H., Huang, C. C., Babbitt, P. C. & Ferrin, T. E. StructureViz: Linking Cytoscape and UCSF Chimera. *Bioinformatics* **23**, 2345–2347 (2007).
343. Miller, B. R., McGee, T. D., Swails, J. M., Homeyer, N., Gohlke, H. & Roitberg, A. E. MMPBSA.py: An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation* **8**, 3314–3321 (2012).
344. Christoffel, F., Igareta, N. V., Pellizzoni, M. M., Tiessler-Sala, L., Lozhkin, B., Spiess, D. C., Lledós, A., Maréchal, J.-D., Peterson, R. L. & Ward, T. R. Design and evolution of chimeric streptavidin for protein-enabled dual gold catalysis. *Nature Catalysis* **4**, 643–653 (2021).
345. Genheden, S. & Ryde, U. Comparison of end-point continuum-solvation methods for the calculation of protein–ligand binding free energies. *Proteins: Structure, Function, and Bioinformatics* **80**, 1326–1342 (2012).
346. Buskes, M. J. & Blanco, M.-J. Impact of Cross-Coupling Reactions in Drug Discovery and Development. *Molecules* **25**, 3493 (2020).
347. Taheri Kal Koshvandi, A., Heravi, M. M. & Momeni, T. Current Applications of Suzuki–Miyaura Coupling Reaction in The Total Synthesis of Natural Products: An update. *Applied Organometallic Chemistry* **32**, e4210 (2018).
348. Suzuki, A. Cross-Coupling Reactions Of Organoboranes: An Easy Way To Construct C-C Bonds (Nobel Lecture). *Angewandte Chemie International Edition* **50**, 6722–6737 (2011).
349. D’Alterio, M. C., Casals-Cruaños, È., Tzouras, N. V., Talarico, G., Nolan, S. P. & Poater, A. Mechanistic Aspects of the Palladium-Catalyzed Suzuki–Miyaura Cross-Coupling Reaction. *Chemistry – A European Journal* **27**, 13481–13493 (2021).
350. García-Melchor, M., Braga, A. A. C., Lledós, A., Ujaque, G. & Maseras, F. Computational Perspective on Pd-Catalyzed C–C Cross-Coupling Reaction Mechanisms. *Accounts of Chemical Research* **46**, 2626–2634 (2013).

351. Cheng, J. K., Xiang, S.-H., Li, S., Ye, L. & Tan, B. Recent Advances in Catalytic Asymmetric Construction of Atropisomers. *Chemical Reviews* **121**, 4805–4902 (2021).
352. Hedouin, G., Hazra, S., Gallou, F. & Handa, S. The Catalytic Formation of Atropisomers and Stereocenters via Asymmetric Suzuki–Miyaura Couplings. *ACS Catalysis* **12**, 4918–4937 (2022).
353. Uozumi, Y., Matsuura, Y., Arakawa, T. & Yamada, Y. M. A. Asymmetric Suzuki–Miyaura coupling in water with a chiral palladium catalyst supported on an amphiphilic resin. *Angewandte Chemie (International Ed. in English)* **48**, 2708–2710 (2009).
354. Benhamou, L., Besnard, C. & Kündig, E. P. Chiral PEPPSI Complexes: Synthesis, Characterization, and Application in Asymmetric Suzuki–Miyaura Coupling Reactions (2013).
355. Chatterjee, A., Mallin, H., Klehr, J., Vallapurackal, J., Finke, A. D., Vera, L., Marsh, M. & Ward, T. R. An enantioselective artificial Suzukiase based on the biotin–streptavidin technology. *Chemical Science* **7**, 673–677 (2015).
356. Creus, M., Pordea, A., Rossel, T., Sardo, A., Letondor, C., Ivanova, A., Letrong, I., Stenkamp, R. E. & Ward, T. R. X-ray structure and designed evolution of an artificial transfer hydrogenase. *Angewandte Chemie (International Ed. in English)* **47**, 1400–1404 (2008).
357. Patel, N. D., Sieber, J. D., Tcyrulnikov, S., Simmons, B. J., Rivalti, D., Duvvuri, K., Zhang, Y., Gao, D. A., Fandrick, K. R., Haddad, N., Lao, K. S., Mangunuru, H. P. R., Biswas, S., Qu, B., Grinberg, N., Pennino, S., Lee, H., Song, J. J., Gupton, B. F., Garg, N. K., Kozlowski, M. C. & Senanayake, C. H. Computationally Assisted Mechanistic Investigation and Development of Pd-Catalyzed Asymmetric Suzuki–Miyaura and Negishi Cross-Coupling Reactions for Tetra-ortho-Substituted Biaryl Synthesis. *ACS catalysis* **8**, 10190–10209 (2018).
358. Ye, N., Zhou, F., Liang, X., Chai, H., Fan, J., Li, B. & Zhang, J. A Comprehensive Review of Computation-Based Metal-Binding Prediction Approaches at the Residue Level. *BioMed Research International* **2022**, e8965712 (2022).
359. Hu, X., Dong, Q., Yang, J. & Zhang, Y. Recognizing metal and acid radical ion-binding sites by integrating *ab initio* modeling with template-based transfers. *Bioinformatics* **32**, 3260–3269 (2016).

360. Andreini, C., Cavallaro, G., Rosato, A. & Valasatava, Y. MetalS2: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids. *Journal of Chemical Information and Modeling* **53**, 3064–3075 (2013).
361. Valasatava, Y., Rosato, A., Banci, L. & Andreini, C. MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. *Bioinformatics* **32**, 2850–2852 (2016).
362. Akcapinar, G. B. & Sezerman, O. U. Computational approaches for de novo design and redesign of metal-binding sites on proteins. *Bioscience Reports* **37**, BSR20160179 (2017).
363. Robles, V. M., Ortega-Carrasco, E., Fuentes, E. G., Lledós, A. & Maréchal, J.-D. What can molecular modelling bring to the design of artificial inorganic cofactors? *Faraday Discussions* **148**, 137–159 (2010).
364. Sciortino, G., Garribba, E., Rodríguez-Guerra Pedregal, J. & Maréchal, J.-D. Simple Coordination Geometry Descriptors Allow to Accurately Predict Metal-Binding Sites in Proteins. *ACS Omega* **4**, 3726–3731 (2019).
365. Putignano, V., Rosato, A., Banci, L. & Andreini, C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Research* **46**, D459–D464 (D1 2018).
366. PyChimera: use UCSF Chimera modules in any Python 2.7 project | Bioinformatics | Oxford Academic. (2023).
367. Fujieda, N., Schätti, J., Stüttfeld, E., Ohkubo, K., Maier, T., Fukuzumi, S. & Ward, T. R. Enzyme repurposing of a hydrolase as an emergent peroxidase upon metal binding. *Chemical Science* **6**, 4060–4065 (2015).
368. Baugh, L., Phan, I., Begley, D. W., Clifton, M. C., Armour, B., Dranow, D. M., Taylor, B. M., Muruthi, M. M., Abendroth, J., Fairman, J. W., Fox, D., Dieterich, S. H., Staker, B. L., Gardberg, A. S., Choi, R., Hewitt, S. N., Napuli, A. J., Myers, J., Barrett, L. K., Zhang, Y., Ferrell, M., Mundt, E., Thompkins, K., Tran, N., Lyons-Abbott, S., Abramov, A., Sekar, A., Serbzhinskiy, D., Lorimer, D., Buchko, G. W., Stacy, R., Stewart, L. J., Edwards, T. E., Van Voorhis, W. C. & Myler, P. J. Increasing the structural coverage of tuberculosis drug targets. *Tuberculosis (Edinburgh, Scotland)* **95**, 142–148 (2015).
369. Bruijninx, P. C. A., Koten, G. v. & Gebbink, R. J. M. K. Mononuclear non-heme iron enzymes with the 2-His-1-carboxylate facial triad: recent

- developments in enzymology and modeling studies. *Chemical Society Reviews* **37**, 2716–2744 (2008).
370. Koehntop, K. D., Emerson, J. P. & Que, L. The 2-His-1-carboxylate facial triad: a versatile platform for dioxygen activation by mononuclear non-heme iron(II) enzymes. *JBIC Journal of Biological Inorganic Chemistry* **10**, 87–93 (2005).
371. Kuhn, M., Firth-Clark, S., Tosco, P., Mey, A. S. J. S., Mackey, M. & Michel, J. Assessment of Binding Affinity via Alchemical Free-Energy Calculations. *Journal of Chemical Information and Modeling* **60**, 3120–3130 (2020).
372. Yang, C., Pflugrath, J. W., Camper, D. L., Foster, M. L., Pernich, D. J. & Walsh, T. A. Structural Basis for Herbicidal Inhibitor Selectivity Revealed by Comparison of Crystal Structures of Plant and Mammalian 4-Hydroxyphenylpyruvate Dioxygenases. *Biochemistry* **43**, 10414–10423 (2004).
373. Weber, M., Weber, M. & Weber, V. Phenol. in *Ullmann's Encyclopedia of Industrial Chemistry* 1–20 (John Wiley & Sons, Ltd, 2020).
374. Fortuin, J. P. & Waterman, H. I. Production of phenol from cumene. *Chemical Engineering Science* **2**, 182–192 (1953).
375. Rahmani, N., Amiri, A., Ziarani, G. M. & Badiei, A. Review of some transition metal-based mesoporous catalysts for the direct hydroxylation of benzene to phenol (DHBP). *Molecular Catalysis* **515**, 111873 (2021).
376. Langeslay, R. R., Kaphan, D. M., Marshall, C. L., Stair, P. C., Sattelberger, A. P. & Delferro, M. Catalytic Applications of Vanadium: A Mechanistic Perspective. *Chemical Reviews* **119**, 2128–2191 (2019).
377. Borrego, E., Tiessler-Sala, L., Lázaro, J. J., Caballero, A., Pérez, P. J. & Lledós, A. Direct Benzene Hydroxylation with Dioxygen Induced by Copper Complexes: Uncovering the Active Species by DFT Calculations. *Organometallics* **41**, 1892–1904 (2022).
378. Wu, P., Fan, F., Song, J., Peng, W., Liu, J., Li, C., Cao, Z. & Wang, B. Theory Demonstrated a “Coupled” Mechanism for O<sub>2</sub> Activation and Substrate Hydroxylation by Binuclear Copper Monooxygenases. *Journal of the American Chemical Society* **141**, 19776–19789 (2019).
379. Udry, G. A. O., Tiessler-Sala, L., Pugliese, E., Urvoas, A., Halime, Z., Maréchal, J.-D., Mahy, J.-P. & Ricoux, R. Photocatalytic Hydrogen Production and Carbon Dioxide Reduction Catalyzed by an Artificial

Cobalt Hemoprotein. *International Journal of Molecular Sciences* **23**, 14640 (2022).