Rational design of multiple active sites in an unspecific peroxygenase

Marc Domingo Cabasés

Supervisor: Prof. Victor Guallar Tasies
Dr. Gerard Santiago Morcillo

Department of Life Science
Barcelona Supercomputing Center

This dissertation is submitted for the degree of
Master in Bioinformatics

Universitat Autònoma de
Barcelona

March 2025

Student:
Marc Domingo

Academic tutor:
Prof. Laura Masgrau

Essentially, all models are wrong, but some are useful.

GEORGE BOX

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 15,000 words including appendices, bibliography, footnotes, tables and equations.

Marc Domingo Cabasés
March 2025

Acknowledgements

I would like to acknowledge to my colleagues in Barcelona Supercomputing Center during these six months: Alberto, Alexis, Carles, Jefferson, Joan, Martí, Natàlia, Oriol, Pep, Rubén and Sergi, my supervsior Gerard and the group leader Víctor for this scientific experience.

And also to my family to give me support in my studies and to my friends.

Abstract

This work is centered on the rational design of unspecific peroxygenases (UPOs), specifically, one called PaDa-I. The project is based on the recent idea of PluriZymes, enzymes with multiple active sites (artificially introduced) on the same scaffold. It is divided into two big sections: the design of a catalytic triad and of a metal site. With the use of novel methods like Protein Energy Landscape Exploration and molecular dynamics, the stability and functionality of the designed variants will be studied. The aim of this work is also to develop a pipeline in order to design PluriZymes.

Table of contents

List of figures

List of tables

Nomenclature

Greek Symbols

δ	Torsion angle
ϵ	Electric permitivity
π	$\simeq 3.14\ldots$
ρ	Normalized distribution density
σ	Distance at which the inter-particle potential is zero
θ	Bending angle
ε	Depth of the potential well

Other Symbols

$\Delta G^{\ddagger}$	Gibbs energy of activation
k_B	$\simeq 1.381 \cdot 10^{-23} J \cdot K^{-1}$, Boltzmann constant
e	$\simeq 1.602 \cdot 10^{-19} C$, elementary charge
Å	Angstrom

Acronyms / Abbreviations

AdnMD Automated *de novo* Metalloprotein Design

ANM Anisotropic Network Model

BLOSUM62 BLOcks SUbstitution Matrix

BSC Barcelona Supercomputing Center

C/Cys Cysteine residue

D/Asp aspartic residue

GPU Graphic Processing Unit

H/His Histidine residue

MD Molecular dynamics

NADPH Nicotinamide adenine dinucleotide phosphate

PELE Protein Energy Landscape Exploration

R/Arg Arginine residue

RMSD Root Mean Square Deviation

S/Ser Serine residue

SASA Solvent Accessible Surface Area

SGBNP Surface Generalized Born + Non-Polar

SOD Superoxide Dismutase

T/Thr Threonine residue

T Temperature

UPO Unspecific peroxygenase

V/Val Valine residue

Chapter 1

Introduction

This work is centered in the study of a type of enzymes called unspecific peroxygenases (UPOs). First of all, it is convenient to define what is an enzyme. An enzyme is a biological system formed by smaller sub-units called amino acids which with a dehydration reaction are attached together forming a peptidic bond. At the same time each amino acid is constituted by an amino, a carboxyl group and a side chain that varies among the different amino acids. The atoms that constitute amino acids are C, N, O, H, S, Se. The number of amino acids that form enzymes in nature are 21 and they can be classified by their polarity in charged, polar or hydrophobic depending on the side chain it has. The synthesis of enzymes is called translation and it consists in the addition of amino acids one by one forming a peptidic chain. The order of addition of amino acids is read through the genetic code of an mRNA, which is a copy from one of the genes of an organism. A sequence of amino acids will lead to a concrete structure of an enzyme, which will have a specific function.

The function of enzymes is to catalyze chemical reactions that occur in living beings, so they are able to decrease the Gibbs energy of activation ($\Delta G^{\ddagger}$), which is the difference between the transition state of a reaction and the ground state of the reactants. Consequently the reaction rate increases a lot. While specific details vary between enzymes, in general, the specific disposition of residues in space and the consequent charge distribution are able to accommodate the substrate in a site of the enzyme and to stabilize the transition state of the process. At the end of a catalytic cycle, the enzyme remains unaltered so the chemical reaction can be performed again.

1.1 Unspecific peroxygenases

As it was said before, UPOs are the working system on this project, so let's give a brief introduction about them. UPOs are enzymes that can selectively catalyze a hydroxylation reaction with only the activation of H_2O_2 (which needs to be "externally" supplied). UPOs are formed by an heme-thiol group on their center and a cavity were an organic substrate can enter to get hydroxylated.

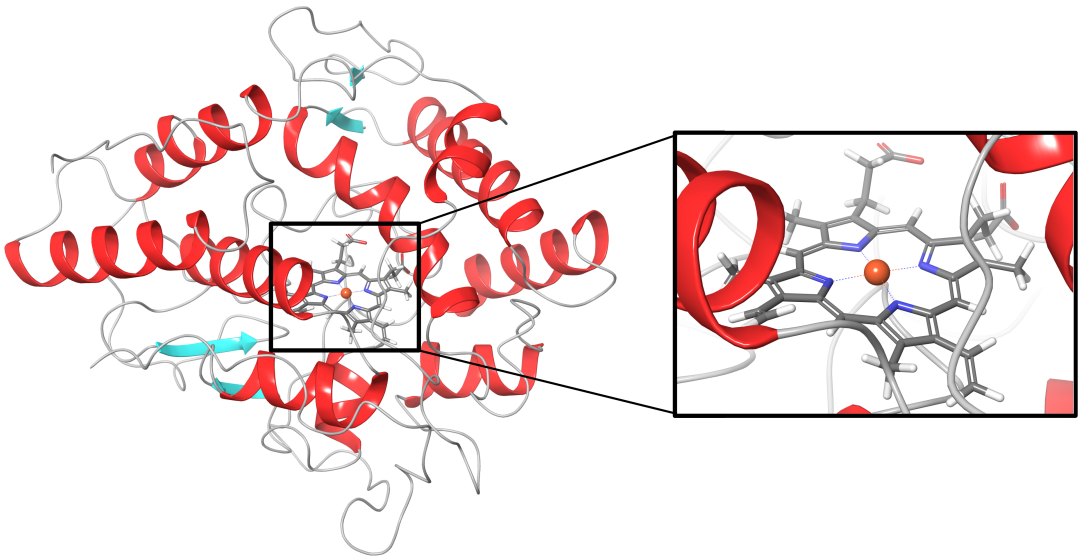


Fig. 1.1 PaDa-I crystal structure, which is the unspecific peroxygenase used in this work. Zoomed in to the heme group of the enzyme.

In Figure 1.1 it is shown the secondary structure of PaDa-I, which is the UPO modeled in this work and has been developed experimentally by Prof. Miguel Alcalde *et al.* [1]. PaDa-I is a mutant of an UPO found in *Agrocybe aegerita*. It is constituted by 327 amino acids, nine α-helices and a heme group inside it (an iron it is at the center of the heme with an octahedral coordination). The general reaction occurring in UPOs is shown in equation 1.1. An organic molecule gets hydroxylated and water is the only byproduct, so it is an environmentally friendly process.

$$\mathrm{RH} + \mathrm{H_2O_2} \longrightarrow \mathrm{ROH} + \mathrm{H_2O} \tag{1.1}$$

The catalytic cycle of UPOs is shown in Figure 1.2. In the first step H_2O_2 gets coordinated to metal the center, then a water molecule is lost and it is formed the

putative active species, the Compound-I, forming a radical cation. After, Compound-I attacks the substrate and the hydroxylation is done. The hydroxylated molecule leaves and a water molecule binds the ferric ion and the cycle is completed.

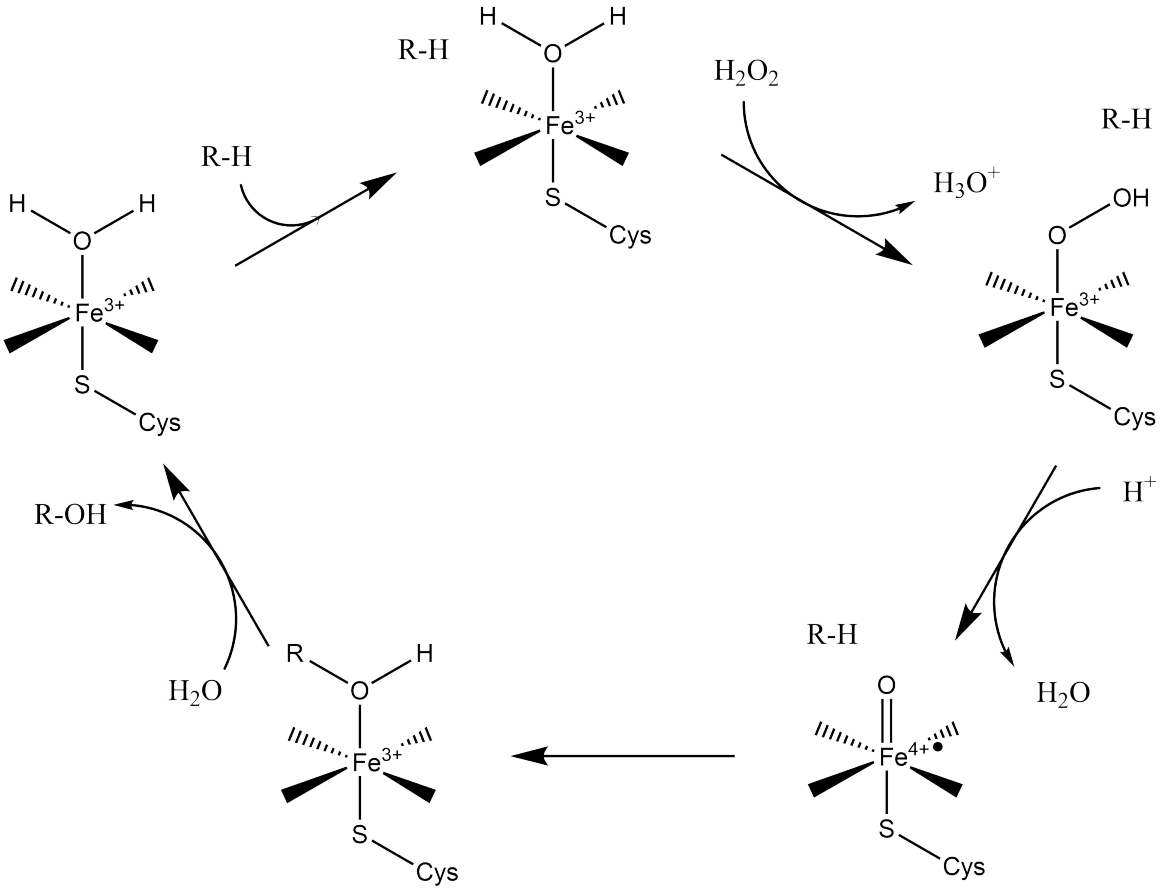


Fig. 1.2 General catalytic cycle of an unspecific peroxygenase leading to substrate (R-H) hydroxylation. Figure done with ChemDraw Professional.

UPOs are enzymes that are often compared with another variety called Cytochromes P450, which also can perform a hydroxylation reaction but in a different way. P450s require a protein partner to deliver electrons to reduce the iron and let the catalytic cycle work. Typically, electrons are transferred from NADPH via cythochromes P450 reductases or other reductases, which are the enzymes required to let the electron transfer happen. The general reaction ocurring in P450s is shown in equation 1.2.

$$\mathrm{RH} + \mathrm{O_2} + \mathrm{NADPH} + \mathrm{H^+} \longrightarrow \mathrm{ROH} + \mathrm{H_2O} + \mathrm{NADP^+} \tag{1.2}$$

The difference with UPOs is precisely that, UPOs do not need a coenzyme (NADPH) and another protein as a source of electrons, H_2O_2 is the electron source which is much

cheaper. For this reason UPOs have a great biotechnological potential because they can be used on the synthesis of many drugs and other chemicals as an alternative to common chemical synthesis, which usually has poor regioselectivity, a high cost and releases harmful products. Some examples are the hydroxylation of vitamins D_2 and D_3 [2] which are useful compounds for human health and animal feeding, the synthesis of 1-naphtol [3] which is a compound needed to produce herbicides, insecticides, pharmaceuticals and dye precursors and also for the synthesis of 5'-hydroxypropranolol [4] which is a metabolite of propanolol, a β-adrenergic receptor antagonist.

1.2 PluriZymes

In this project we want to develop a PluriZyme, which is an enzyme with an original active site and another one, artificially added, that is designed to achieve a specific goal. The first reference about rational engineering of multiple active sites is about an ester hydrolase in which another ester hydrolase active site like was designed a part from the original one [5]. The experiments showed clearly the existence of two actives sites, although with low activity for the designed site. However, an improved refinement modeling, involving the stabilization of the catalytic triad, significantly increase the artificial site performance. This fact reminds us that it is vital to do debugging and refinement when developing the enzyme.

This work is based on this idea and it is going to be discussed and developed. The modelization of a PluriZyme generally requires specific mutations to add a new functionality. The stability (global and local) of the new variants can be computationally tested by several techniques such as molecular dynamics. In any case, since the consequences of the mutations on the final folding of the enzyme are difficult to infer (with simulations) and because of the multiple approximations done in the modeling techniques, experimental validation is always mandatory. In Figure 1.3 it is shown the concept of a PluriZyme in which a new functionality is added to an enzyme.

1.2.1 Design of a metal binding site

We aim to model PaDa-I to improve its auto efficiency. With the design of the PluriZyme we want to go a step further: the goal is to make UPO to produce its own peroxide hydrogen in order to auto feed itself, increasing in this manner, its biotechnological potential. Thus, the approach consists in adding a new functionality to PaDa-I, which will be able to produce the necessary H_2O_2 quantity to activate itself.

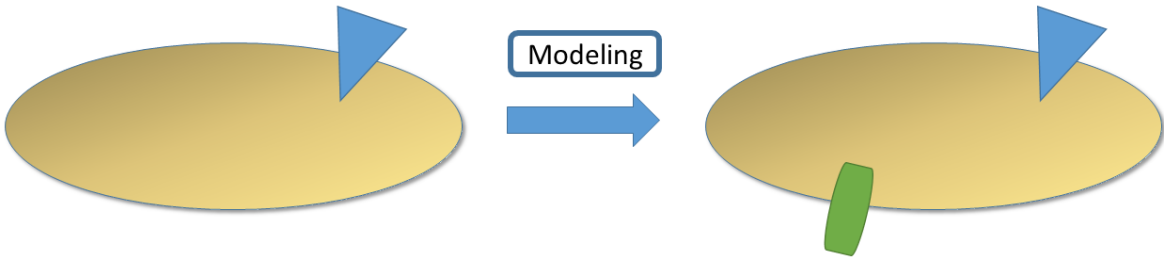


Fig. 1.3 Concept of PluriZyme. After a modelization work that consists of several steps, an enzyme with multiple active sites is obtained. The blue and green objects represent substrates binding into the enzyme.

In nature we find enzymes, called superoxide dismutases (SODs), that catalyze the dismutation of superoxide anion (that is mostly formed in mithocondrial respiratory chain in living beings) into H_2O_2 and O_2. SODs are metaloenzymes (i.e. enzymes which contain a metal as a cofactor). The global dismutation reaction is shown in equation 1.3 and each semi-reaction is described in equations 1.4 and 1.5. In semi-reaction 1.4 superoxide is oxidized to oxygen while the metal attached to SOD is reduced and in semi-reaction 1.5 superoxide is reduced to hydrogen peroxide while the metal is oxidized.

$$2\,O_2{}^{\bullet-} + 2\,H^+ \longrightarrow O_2 + H_2O_2 \tag{1.3}$$

$$SOD{-}M^{(n+1)+} + O_2{}^{\bullet-} \longrightarrow SOD{-}M^{n+} + O_2 \tag{1.4}$$

$$SOD{-}M^{n+} + O_2{}^{\bullet-} + 2\,H^+ \longrightarrow SOD{-}M^{(n+1)+} + H_2O_2 \tag{1.5}$$

There are several types of SODs in nature depending on the organism, but all of them have a metal binding site where the reaction takes place. The most common are Cu-Zn-SOD, Mn/Fe-SOD and Ni-SOD [6]. The idea is to design a metal binding site in PaDa-I trying to mimic a SOD active site, so the UPO can produce its own H_2O_2. The mimicked SOD in this work is a SOD5 (PDB ID: 4N3T), which has a similar architecture compared to Cu-Zn-SODs with the difference that SOD5 only has copper as metal cofactor. SOD5 is naturally expressed in *Candida albicans*, a human fungal pathogen.

The system consists of four coordinating histidines, three of them are always coordinating in a planotrigonal geometry and the other one is in an axial disposition and it is not always coordinating the copper. In Figure 1.4 it is shown the catalytic cycle of SOD5. The first step consists in the entrance of superoxide radical, assisted by an arginine residue, in the metal active site. Once it is coordinated to the metal an electron transfer to the metal happens, which leads to generation of O_2 and to

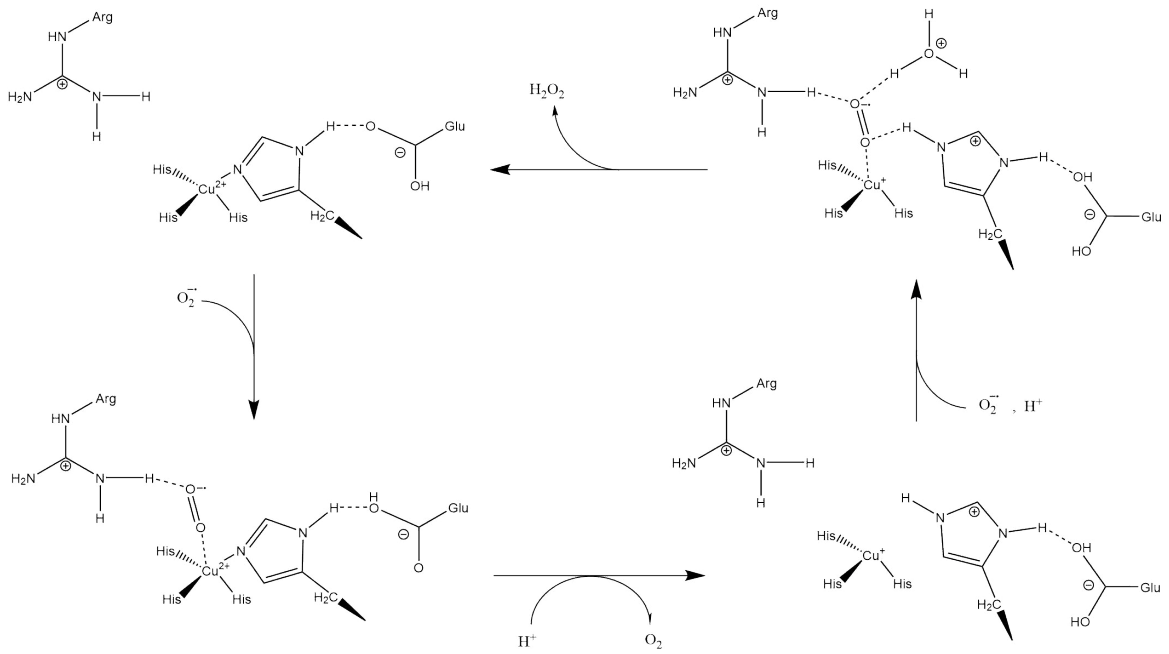


Fig. 1.4 Catalytic cycle of SOD5. Figure done with ChemDraw Professional

the histidine-metal breakage bond. Next, another superoxide molecule enters the metal center, an electron transfer occurs from Cu^+ to $O_2^{\cdot -}$ and the peroxide anion gets protonated by the histidine and by protons of the medium, so H_2O_2 is formed.

The copper coordination varies from pseudotetrahedral to pseudoplanotrigonal and depending on the catalytic cycle step.

Scientific viability

In previous works on UPOs the concentration of H_2O_2 employed oscillate between 0.5 mM, hydroxylation of vitamin D [2], and 1-2 mM, in the synthesis of 1-Naphtol [3]. Thus, the designed UPO should be capable if producing a similar quantity of hydrogen peroxide. Consequently, even if the SOD metal binding site is perfectly mimicked in PaDa-I, the PluriZyme is not going to work depending on the conditions of the medium. To achieve our goal, the superoxide concentration has to be controlled in some way. In addition, there are no references about SODs and UPOs working together in the way that SODs generate enough H_2O_2 to make UPOs work. However, as a PluriZyme is being developed much less quantity (undetermined quantity) of H_2O_2 would be needed because theoretically it is continuously generated on the enzyme itself (cascade reaction). Thus, the predictions whether our PluriZyme could work or not are not that easy.

To study the reaction rates of SODs the experimental techniques used are based on the controlled production of superoxide anion by electrochemical, chemical, photochemical, and photocatalytic methods described in Hayyan *et al.* [7] so in the experimental validation of the PluriZyme these methods should be considered to be used.

One inconsistency could be that $O_2^{\cdot -}$ rapidly disproportionates in water, so it is not necessary to design a metal site in PaDa-I. Here the motivation is that SODs increases the dismutation rate of superoxide. For superoxide $k_{SOD5}(pH = 7) \approx 10^9 M^{-1}s^{-1}$ [8] and the decay of superoxide in water $k_{H_2O_2}(pH = 7) \approx 5 \cdot 10^5 M^{-1}s^{-1}$ [9].

Economic viability

As the principal goal of the project is to make the whole process to be lower in costs it is necessary to study the economic viability of it.

Apparently the unique difference is that the current methodology uses H_2O_2 as substrate. This compound needs to be bought, so it represents an extra cost respect this project, in which H_2O_2 is produced by the UPO itself. However, in this work an additional element is used, the copper, so this fact should be considered.

In references [4, 3] the H_2O_2 concentration used is in a range of (0.5-2) mM, and the enzyme concentration is (0.0066-0.03) μM. As the dissociation constant for a generic metalloenzyme is lower than $K_D \approx 10^{-5}$ defined as $K_D = \frac{[M][P]}{[MP]}$ [10], it is considered that experimentally it is needed an estechiometric amount of a copper salt ($CuCl_2$). Price for $CuCl_2 \cdot 2H_2O$ (Product ID: 459097) and H_2O_2 (Product ID: 216763) are extracted from Sigma-Aldrich. Equations 1.6 and 1.7 show the calculation of costs in 1 liter of reaction mixture for either using a copper salt or H_2O_2.

$$\frac{0.03\ \mu mol\ enzyme}{1L\ disolution} \cdot \frac{10^{-6}\ mol}{1\ \mu mol} \cdot \frac{1\ mol\ Cu^{2+}}{1 mol\ enzyme} \frac{1\ mol\ CuCl_2 \cdot 2H_2O}{1 mol\ Cu^{2+}} \cdot \frac{170.489\ g\ CuCl_2 \cdot 2H_2O}{1\ mol\ CuCl_2 \cdot 2H_2O} \cdot \frac{45.10\ €}{5\ g\ CuCl_2 \cdot 2H_2O} = \frac{4.6 \cdot 10^{-5}\ €}{1L\ disolution} \tag{1.6}$$

$$\frac{1\ mmol\ H_2O_2}{1L\ disolution} \cdot \frac{10^{-3}\ mol\ H_2O_2}{1\ mmol\ H_2O_2} \cdot \frac{34.01\ g\ H_2O_2}{1\ mol\ H_2O_2} \cdot \frac{1\ g\ dissolution\ 30\%}{0.3\ g\ H_2O_2} \cdot \frac{1\ cm^3\ 30\%\ disolution}{1.11\ g\ 30\%\ disolution} \cdot \frac{53.40\ €}{500\ cm^3\ 30\%\ disolution} = \frac{0.011\ €}{1L\ disolution} \tag{1.7}$$

From calculations it is derived that using the PluriZyme instead of the original method (adding H_2O_2) is much cheaper per liter of reactant mixture.

However, the production of superoxide radical would represent a cost difficult to predict depending on the used method. The production of new PaDa-I variants will also imply an additional (low) cost.

1.2.2 Design of a catalytic triad

In addition, we aim to design an "esterase" catalytic triad, containing a Ser, a His and an Asp, on the surface of the protein. It could be useful in some cases where we want to perform a consecutive hydroxylation and ester hydrolysis of a certain compound or to attach a (metal) cofactor able to perform a certain functionality.

A catalytic triad is a set of three residues that is generally found in hydrolases and transferases enzymes. An acid (Asp) - base (His) - nucleophile (Ser) triad is a common motif for generating a nucleophilic residue for covalent catalysis. These residues form a network to polarise and activate the nucleophile (usually a serine), which attacks the substrate, forming a covalent intermediate which is then hydrolysed to release the product and regenerate free enzyme. Equation 1.8 shows the general reaction performed by esterases which consists in the hydrolysis of an ester into an alcohol and a carboxylic acid.

Figure 1.5 shows the catalytic cycle of an ester hydrolase. In the first step the histidine deprotonates the serine. The nucleophylic power of the serine is incremented, attacks the carbonylic carbon of the ester and a tetrahedral intermediate is formed. Then, an acyl group is formed, which will be attacked by a water molecule and a second tetrahedral intermediate is formed. Finally, the carboxylic acid is released and the cycle is completed.

$$R_1COOR_2 + H_2O \longrightarrow R_1COOH + R_2OH \tag{1.8}$$

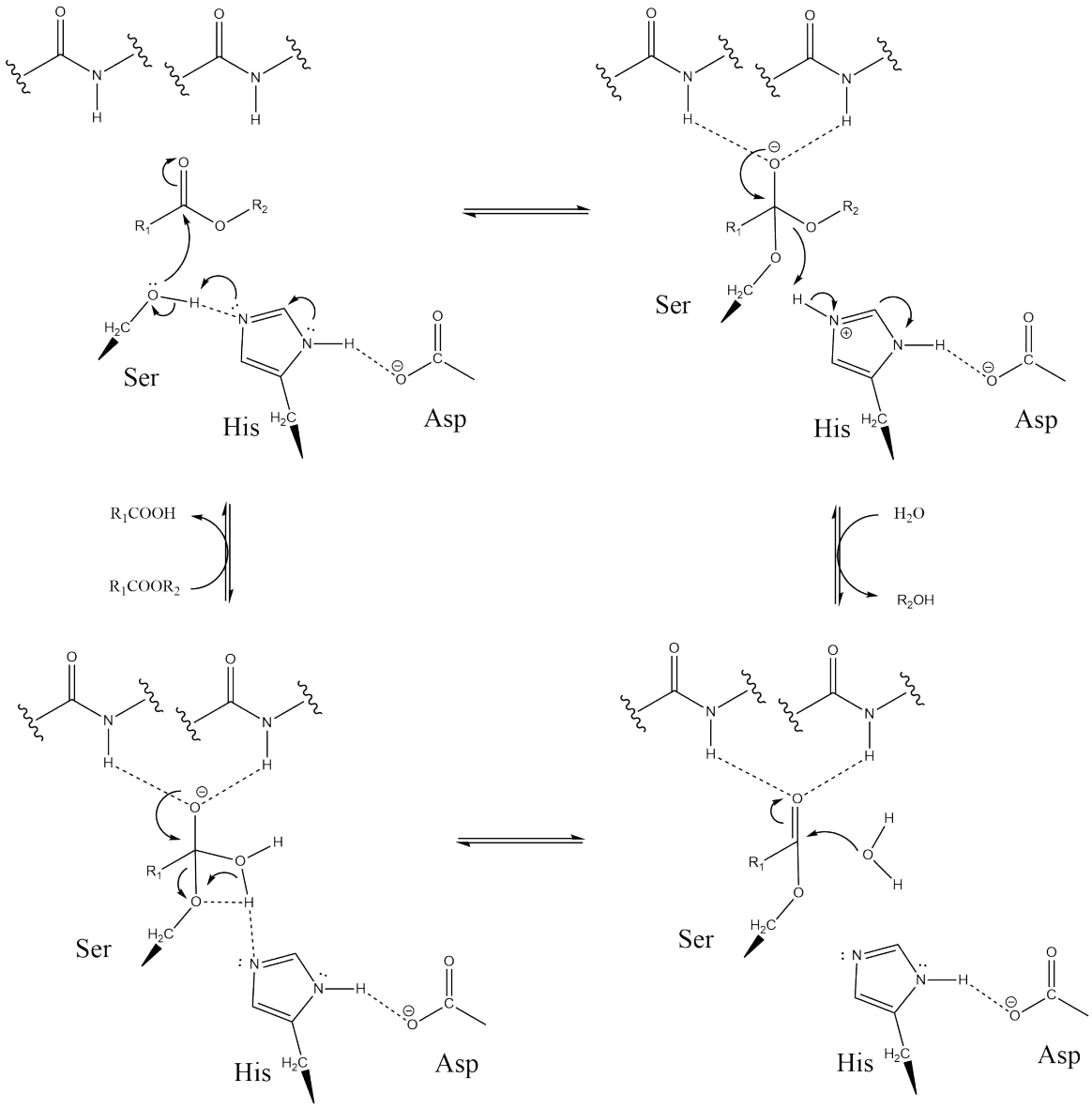


Fig. 1.5 General catalytic cycle for an ester hydrolase. Figure done with ChemDraw Professional.

1.3 Computers and science

It is obvious, but it should be remembered, that computers are an indispensable tool in science research, specially in the field of structural bioinformatics. In order to study biological systems like proteins, computational tools are a perfect complement, capable of performing simulations that help us in understanding the structure-function relationship. The use of computers is due to several reasons, first of all, the equations to be solved in the simulations require thousands of billions of calculations to be made due to the high number of interacting atoms, so machines can do it for us. In addition, the visualization of biochemical systems is essential to understand the types of interactions that occur and also to study the behaviour and evolution of the systems, therefore the use of graphic software is very necessary.

Molecular modeling allows us to understand at atomic scale what is observed in the macroscopic world and the simulations can give a reasonable and mechanistic explanation for physicochemical phenomena. It is also useful for designing experiments as simulations can guide us procedure that we must follow if we want to achieve a certain objective.

This work has been developed in the installations of Barcelona Supercomputing Center (BSC) taking advantage of their computing power required to perform the needed simulations.

Chapter 2

Objectives

This work is divided in two types of modelization, the design of a metal active site and the design of a catalytic triad. The objectives are the following:

- Propose a suitable pipeline to develop PluriZymes, either for a metal site and a catalytic triad.
- Design of a metal active site in PaDa-I in order to obtain an auto-feeding UPO.
- Design of a catalytic triad in PaDa-I to add a new functionality that could be used in applications.
- Develop a tool (AdnMD) to help in the design of metaloPluriZymes.

Chapter 3

Methodology

3.1 Atomistic model and force field

Before doing any type of simulation it is needed a model to describe the reality of biological systems like enzymes. As has been explained in Chapter 1, enzymes are a set of atoms and the way atoms are treated will define the model. If electrons are necessary to describe the process a quantum mechanical approach is followed and if not a classical physics approach is used in which electrons are taken into account implicitly.

In this study, the interactions between atoms are treated classically and they are described by a force field. There are many force fields in literature but in this work OPLS-2005 [11] is used for all types of simulations. The potential energy of the system is given by the sum of bonded and non-bonded terms. Bonded terms are formed by covalent bonds, angle bendings (treated as harmonic potentials) and torsions (treated as Fourier expansions). Non-bonded interactions are divided in electrostatic and Van der Waals (treated as a Lennard-Jones potential) potential terms . The global potential expression for OPLS-2005 is given by equation 3.1:

$$V_{MM} = \underbrace{\sum_{bonds} K_r \left(r - r_0\right)^2 + \sum_{angles} K_\theta \left(\theta - \theta_0\right)^2 + \sum_{dihedrals,n} K_{\phi,n} \left[1 + cos\left(n\phi - \delta_n\right)\right]}_{bonded} + \underbrace{\sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} f_{ij} + \sum_{i<j} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] f_{ij}}_{non-bonded} \tag{3.1}$$

where K_r, K_θ, $K_{\phi,n}$ parametres are the force constants for distances, angles and dihedrals, respectively. Bond distances correspond to r and angles to θ and the ones with subscript 0 correspond to equilibrium distances. n and δ_n are the periodicity and phase of ϕ dihedral angle. Factor f_{ij} for non-bonded terms is 0 if i-th and j-th atoms are separated by one or two bonds, 0.5 if they are separated by three bonds and 1 otherwise. Regarding the electrostatic term q_i and q_j are the charges of i-th and j-th atoms, ϵ_0 the electric permitivity in void. For Lennard-Jones potential σ_{ij} and ε_{ij} are parameters that characterize this term.

3.2 PELE

Protein Energy Landscape Exploration (PELE) is a Monte Carlo based technique designed to model protein-ligand interactions. PELE method consists in three main steps: an initial perturbation, a side chain prediction and a final minimization. The used force field is OPLS-2005 coupled to an implicit water model Surface Generalized Born + Non-Polar (SGBNP) [12]. Figure 3.1 shows the general workflow to prepare a system and the steps done by PELE software. First of all the protein is prepared with the Schrödinger Protein Preparation Wizard [13] tool and then, PlopRotTemp which is an in-house program that parametrizes the ligand and prepares its template, which includes the topology, the atom charge and the force field parameters (OPLS-2005).

In order to control the different parameters of the steps done with PELE, a control file is meticulously configured to achieve a certain goal and to extract metrics that could be useful for the analysis. In Local Perturbation, Side Chain Sampling and Minimization subsections it is explained the idea of each step.

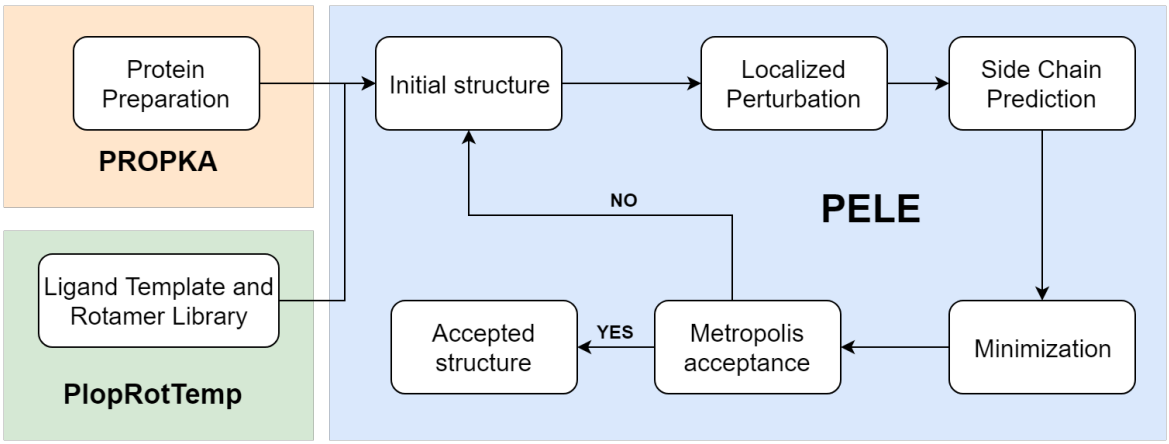


Fig. 3.1 System preparation and PELE workflow.

Local Perturbation

The first step consists in the generation of a perturbation in the ligand and a normal mode perturbation in the protein. The ligand perturbation algorithm consists in a different trials of (enforced) random translations and rotations, selecting the one conformation with the lowest energy. As ligands are not rigid, besides global rotational and translational, the internal dihedral angles are also perturbed. However, the bending angles and bond distances are kept invariant.

Regarding the protein, it is perturbed by a normal mode procedure. The protein is described by the anisotropic network model (ANM) which consists in an elastic network constituted by several nodes which are all the α-carbons of the protein. Neighbouring atoms (12 Å cutoff) are connected by harmonic potentials calculated by a power-based function. The frequencies of the normal modes are computed diagonalizing the Hessian matrix. The Hessian is a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field (the force field in this case). It describes the local curvature of a function of many variables. The six lowest frequency normal modes are selected and the protein is perturbed in a random combination of these six modes. This procedure is followed because it allows to explore global movements of the system.

Side Chain Sampling

As in the previous step the structure has been perturbed, steric clashes have to be removed. The next step consists by placing all side chains that are near to the ligand with a rotamer library side chain. The used algorithm is zhexin [14] which tries to find the most likely rotamer in a heuristic manner. To do so, it clusters the rotamers and estimates the partition function. It chooses the rotamer that has the greatest contribution. It is performed several times, until convergence is reached. In this work we used a resolution of 30^{o}.

Minimizaton

The next step involves the minimization of a user-defined region (typically it is included the full protein) with a Truncated Newton minimizer. The minimization is performed until a minimization criterion is reached.

Metropolis algorithm

The last step involves the acceptance or rejection of the obtained minimum. The new structure is accepted if it has less energy than the initial one or by the Boltzmann factor

criteria, $e^{-\frac{E_f - E_0}{k_B T}} > Rand(0,1)$, in which the structure is accepted if the inequality is satisfied. $Rand(0,1)$ is a random number between 0 and 1, E_f is the energy of the final structure and E_0 the one for the initial, and k_B the Boltzmann constant.

3.3 Molecular Dynamics

MD is a simulation technique that has the goal to reproduce the real dynamics of a system at atomistic scale. It consists in applying the classical mechanic's laws (Newton's equations) on the set of interacting atoms that define the system. Newton's law is a differential equation that describes the particle movement, this is, the evolution of an atom of the system is given by 3.2:

$$m_i \frac{\mathrm{d}\vec{v}_i}{\mathrm{d}t} = \sum_{j \neq i} \vec{F}_{j \to i} + \vec{F}_{external} = -\frac{\partial}{\partial \vec{r}_i} V_{MM}\left(\vec{r}_1, ..., \vec{r}_N\right) + \vec{F}_{external} \tag{3.2}$$

where m_i is the mass of atom i, $\vec{v}_i$ its velocity and V_{MM} the potential energy that depends on the position of all the atoms in the system. In Equation 3.1 it is detailed how these terms are calculated. Regarding $\vec{F}_{external}$ term it is introduced when a thermostat, barostat or an external field is used.

All MD programs have in common several steps: different input parameters are specified (like the temperature, pressure, trajectory time), the initialization of the trajectory, calculation of interactions between atoms and finally the integration of movement equations.

The used program in this work is Desmond [15] using OPLS-2005 as force field. All the simulations presented here have been prepared in an orthorombic box within 10 Å from the edge of the box and the most near atom of the protein. The system was neutralized with NaCl and an ion concentration of 0.15M was used. The water model used was single point charge (SPC) [16] which is basically a rigid triangle formed by one oxygen and two hydrogens.

All MD simulations were 100 ns long in NPT ensemble (at constant pressure, temperature and number of particles) at T=300 K and P=1 atm. In order to control the temperature the Nosé-Hoover chain [17] thermostat was used which is a thermostat that is deterministic, time-reversible and reproduces the canonical ensemble. The pressure is controlled by Martyna-Tobias-Klein barostat [18]. The time step used to discretize Newton's equation is 2 fs, bonds containing hydrogens were freezed and the electrostatic cutoff used is 9 Å, which means that the electrostatic interaction between two particles are not considered if its distance is higher than the cutoff.

Before performing the long production in MD the default procedure of Maestro is followed. It consists in 100 ps Brownian dynamics simulation (T=10K), a 12 ps NVT ensemble (T=10K) simulation, a 12 ps NPT ensemble (T=10K) simulation, a 24 ps NVT ensemble (T=300K) simulation and finally a 24 ps NPT ensemble (T=300K) simulation.

MD simulations in this work have been launched in MinoTauro supercomputer located at BSC, which has the technology NVIDIA GPUS.

3.4 Pipeline for this work

Designing PluriZymes is a recent topic so there is not a specific pipeline to develop these systems. Sections 3.4.1 and 3.4.2 show the workflow of the different techniques used for each of the research lines.

3.4.1 Design of a metal active site

In order to design a metalloprotein the workflow shown in Figure 3.2 is used. The first step consists in the visual exploration of the enzyme trying to find suitable places to mimic a certain metal binding site. Next, the modelization is done by mutating specific residues, trying different rotamers and doing minimizations to avoid clashes created in this process. In order to mimic a specific metal binding site not only the first coordination sphere of the metal is important but the second one; it is not a trivial process. Afterwards, a quantum optimization of the designed site is performed with a QM/MM approach using QSite. LACVP* basis set was used (the potential from the core electrons is treated with a parametrized potential). The used QM method is DFT using the M06-2X exchange-correlation functional. The QM region included the copper, the arginine and the four coordinating histidines capped by C-α, C-β bond (using a hydrogen cap). With the obtained structure MD and PELE simulations are performed following the workflow explained in sections 3.2 and 3.3. Finally QM/MM calculations could be performed to analyze the reaction profile and to study the electron transfer, however they have not been performed in this work.

In a parallel way and to establish a comparison for the obtained results, MD and PELE simulations have been performed for superoxide dismutase.

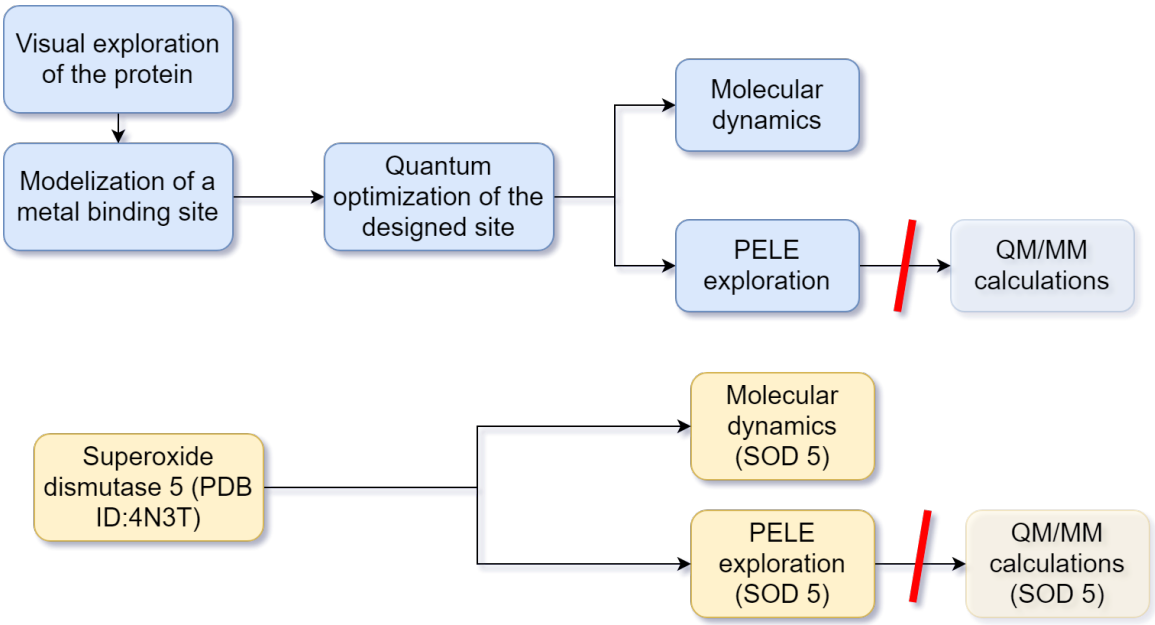


Fig. 3.2 Design of a metal binding site workflow.

3.4.2 Design of a catalytic triad

The designed workflow is similar to the previous section but simpler because it involves less number of residues and less geometric features. The first step consists in running a global PELE exploration with an ester as ligand to search different local binding minima. Next, the modelization is done by mutating specific residues, trying different rotamers and doing minimizations to obtain well-aligned catalytic triads. Finally, MD and PELE simulations are performed to study the stability of the designed catalytic center. The workflow for this section is shown in figure 3.3.

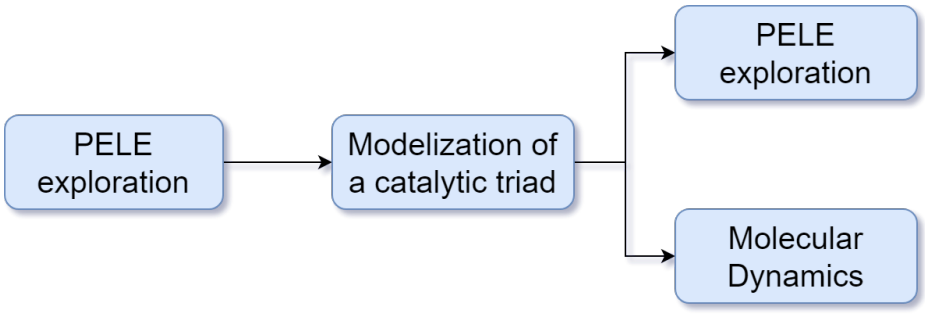


Fig. 3.3 Design of a catalytic triad workflow.

3.5 Development of AdnMD

Besides the explained methodology it has also been developed a method using Python that can serve as a tool to accelerate the modelization of PluriZymes that involve metals. The workflow of Automated *de novo* Metalloprotein Design (AdnMD) is shown in Figure 3.4.

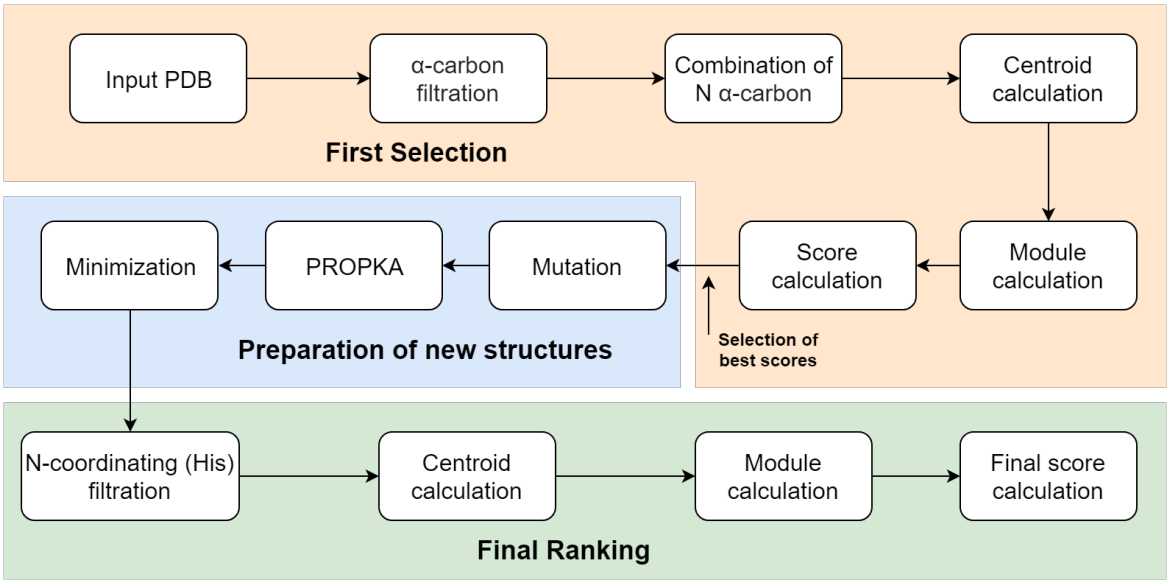


Fig. 3.4 AdnMD workflow.

The workflow works as follows: α-carbons are extracted from the PDB file of the enzyme it is wanted to create the metal site. Then all possible combinations of N α-carbons (N = 3 or 4) are made. The centroid is calculated for all of these combinations as $\frac{1}{N}\sum_{i=1}^{N} r_i$ where r_i are the coordinates of an α-carbon and then the distance between each vertice and the centroid is calculated. The same procedure is followed for the metal site which that is wanted to be mimicked (α-carbons from residues that coordinate the metal). A score based on the distance deviation of α-carbons and the centroid between all the combinations of N-α-carbons and the α-carbons that contain the metal site is computed using formula 3.3.

$$\min_{j \in 1,\ldots,N!} \left(\sum_{i=1}^{N} \sqrt{(m_{i,j} - m_{0,i})^2} \right) \tag{3.3}$$

where $m_{0,i}$ are the distances between the centroid and each i-α-carbon of the metal site it wanted to mimic, $m_{i,j}$ are the distances between the centroid and each of the different combinations of the selected α-carbon of the enzyme where we want to add a metal site.

The best scores are selected (in this work, 1000 when N=3 and 5000 when N=4) and the involved residues of the enzyme are mutated to form a new structure. These structures are then prepared using PROPKA and minimized using OPLS2005 forcefield. These new structures are reevaluated again taking into account the mutated histidines. Coordinating nitrogens (non protonated) are filtered from new PDBs and the same procedure as before is followed. The centroid for N nitrogens and the module from this centroid to the different vertices is calculated. A final score is computed based on the deviation of nitrogen's disposition respect the ones we want to reproduce computed as well with equation 3.3.

Designing functional metal binding sites is still a hard task in biomolecular design. It is not trivial to understand the relationships between protein structure and metal coordination. The first thing that should be done is to find sequence/structure motifs to do *de novo* metalloenzymes: pre-existing site in the interior of a tertiary or quaternary structure donated by the host scaffold, or derived from sequence similarity to natural scaffolds. Moreover, cysteine-derived disulfide bonds have been widely exploited to stabilize pre-existing protein architectures. Typically, histidines should be in a robust zone such as an α-helix to make the metal binding site stable but it is not a prerequisite. A recent article [19] has successfully implemented an approach called MASCoT with these ideas. AdnMD workflow shares similarities with a paper [20] in which they produce a metal binding site in a $\beta\alpha\beta$ structure. Other articles [21–23] try to design metalloproteins in a more manual way. The Python code is published on GitHub at https://github.com/marcdomingo/AdnMD.

Limitations of AdnMD

AdnMD pretends to be a tool to help the modelization of PluriZymes that involves metal. After obtaining the results, the modeler should check the stability of mutations with different techniques (such as MD) and to add the metal to continue the modelization. GaudiMM [24, 25] could be used to continue a more accurately modelization. GaudiMM is capable of finding metal binding sites considering the geometric particularities of the first coordination sphere of the metal in a docking process.

However there are some limitations and procedures that could be improved. The orientation of the side chain is not considered because only α-carbons are taken (only side chains that point to approximately the same point should be considered in the first selection to obtain faster/better results). Orientation should be treated with vectors and score should include a term for the distance deviations and another for angle deviations. It does not take into account the second coordination sphere, which

is important in metaloenzymes. Finally, the mutated histidines protonation should be checked again by the user because they can affect each other.

Chapter 4

Results

In this chapter the results are going to be shown and discussed. First for the rational design of the metal binding site, then for the catalytic triad and finally for AdnMD.

Interpretation of PELE plots

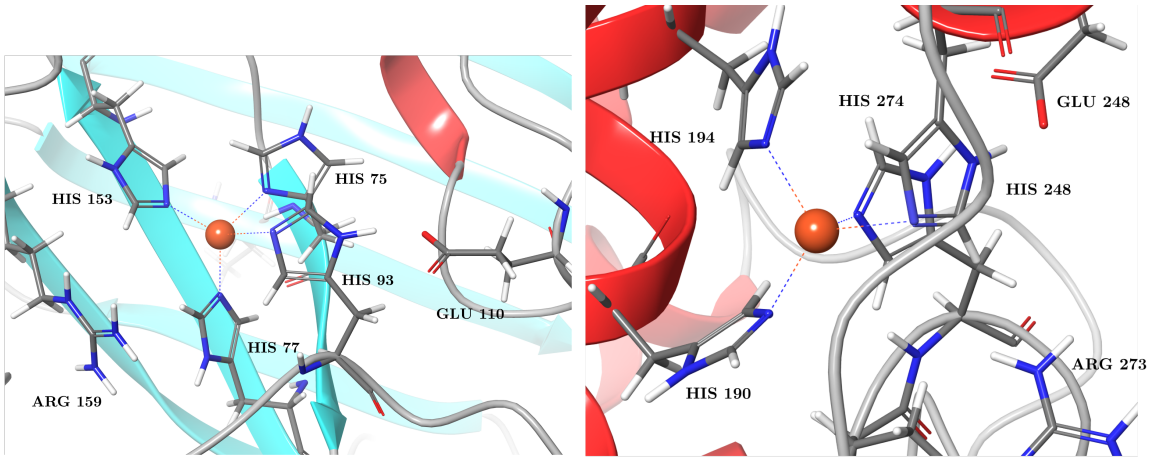
In this Chapter there are represented many plots coming from PELE simulations so it is convenient to explain them. Typically in X-axis it is represented a crucial metric (for example the distance from a catalytic residue to the ligand), Y-axis represents the protein-ligand interaction energy and the color plot represents the SASA (Solvent Accessible Surface Area) which is the surface area of a biomolecule that is accessible to a solvent. In addition, each point in the plot represents a pose of the enzyme + ligand complex extracted from the whole PELE simulation. Consequently the represented points that have better (lower) binding energy will be the poses in which a certain ligand is more accommodated to the protein.

4.1 Design of a metal active site

In section 4.1.1 we show the best mutations to design the metal active site, in section 4.1.2 the results for PELE simulations and in section 4.1.3 the results for molecular dynamics. Supporting data for the graphs is shown in Appendix A.1.

4.1.1 Proposed mutation

Figure 4.1 shows the metal site in the superoxide dismutase (4.1a) and the designed metal site in PaDa-I (4.1b). A total of five mutations were introduced in PaDa-I, precisely F190H, Y194H, F274H, V248H, I273R. Analysing these mutations from



(a) Metal site n SOD5. (b) Designed metal site in PaDa-I

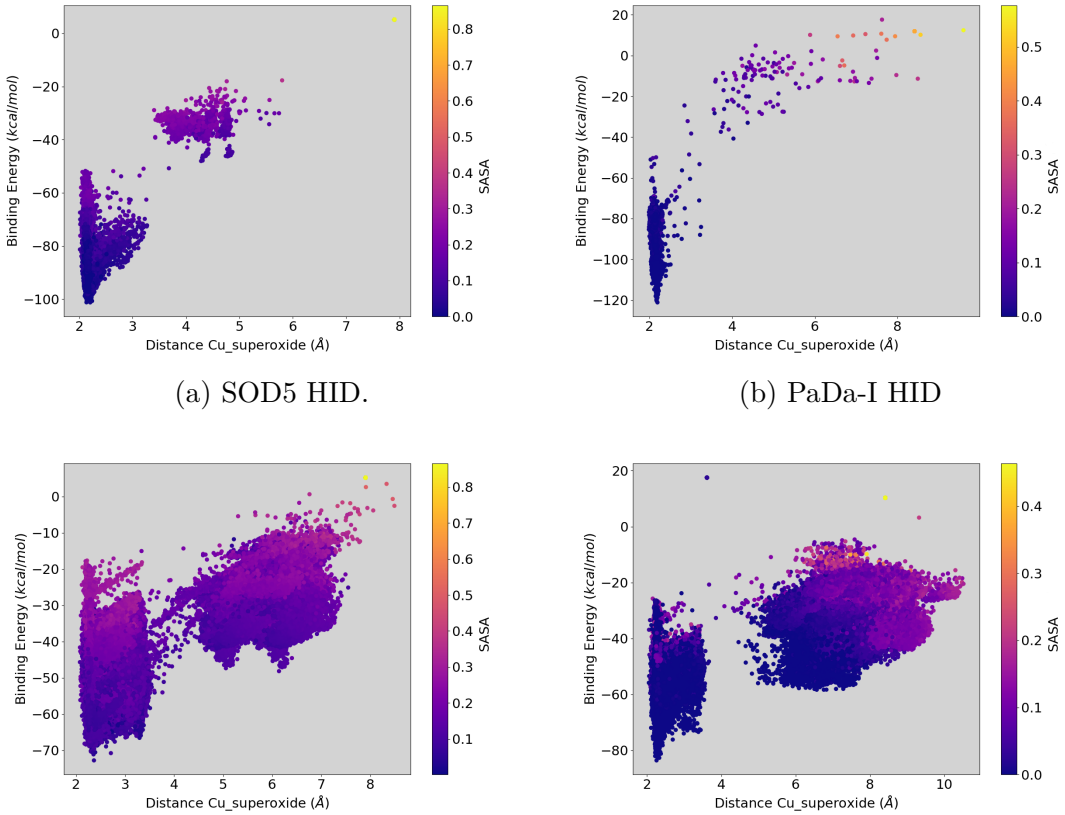
Fig. 4.1 Metal active sites

an alignment point of view with BLOSUM62 (BLOcks SUbstitution Matrix) matrix, the total score obtained is -6 ($F190H_{BLOSUM62} = -1$, $Y194H_{BLOSUM62} = 2$, $F274H_{BLOSUM62} = -1$, $V248H_{BLOSUM62} = -3$, $I273R_{BLOSUM62} = -3$). The most disadvantageous mutations are V248H and I273R because the hydrophobicity lost in the new variant.

Both systems can coordinate copper by four histidines. For the SOD5 system His77, His93 and His153 are δ-protonated and His75 is ε-protonated. However, for PaDa-I all histidines are δ-protonated. From a structural point of view in SOD5 three of the histidines (His75, His77 and His153) are in a β-sheet secondary structure and His93 (the histidine that gets coordinated and discoordinated) is part of a loop. For PaDa-I His190, His194, His248 (the one that is more labile) are in an α-helix and His273 in a loop. An arginine residue is on the surface of the protein, which could help the entrance of superoxide into the metal center. Moreover, a crucial glutamic residue is also on both systems. It has been proved [8] that this residue helps minimize alkaline pH inhibition of the H_2O_2 formation semi reaction (equation 1.5) described in the Introduction and ensures maximal activity in physiological pH up to 7.5. The δ nitrogen pKa of His93 is high enough so when it is not coordinating it is always protonated and H_2O_2 can be formed.

4.1.2 PELE analysis

Once the metal site has been optimized PELE simulations are performed with superoxide as ligand. A superoxide molecule ($O_2^{\cdot -}$) is modeled as $-\frac{1}{2}$ charge in each oxygen



(a) SOD5 HID.

(b) PaDa-I HID

(c) SOD5 HIP

(d) PaDa-I HIP

Fig. 4.2 PELE simulations on the metal binding site with superoxide radical anion.

atom and it is situated $\sim$ 8 Å to the copper site for all the simulations. The goal of running this simulation is to observe if both metal sites have similar behaviour and if superoxide is going to bind the metal and with which energy. Two different conformations for both systems have been studied representing the two different steps on the catalytic cycle shown in Figure 1.4. In Figure 4.2 we show the different PELE output plots; where the X-axis represents the $Cu - O_2^{\cdot -}$ distance.

The (a) and (b) plots present the simulations for the four coordinating histidine systems in which copper is modeled with a +2 charge. In both simulations superoxide anion gets coordinated at distance near 2 Å with an energy of 100 $kcal \cdot mol^{-1}$ for SOD5 and 120 $kcal \cdot mol^{-1}$ for the designed PaDa-I, the main interaction comes from the electrostatic term between copper and superoxide. Regarding (c) and (d) plots they present the simulation for three histidines coordinating systems (one of the histidines is now protonated and not coordinating) in which copper is modeled with a +1 charge. In both simulations superoxide anion gets coordinated at distance near 2 Å (as in (a)

and (b) simulations) with an energy of 70 $kcal \cdot mol^{-1}$ for SOD5 and 80 $kcal \cdot mol^{-1}$ for the designed PaDa-I. The binding energy has decreased respect previous simulations due to the decrease in copper charge.

4.1.3 MD analysis

Regarding molecular dynamics simulations, here we show the analysis of the simulations performed. RMSD, distances and angles observables are analyzed.

Simulation without the copper

Figure 4.3 shows different results for 100 ns MD simulations for SOD5 and PaDa-I systems in the apo form. In (a) and (b) panels the three histidines that always coordinate the copper are displayed in a random snapshot since their position is conserved throughout. Regarding the remaining histidine (H93 for SOD5 and H248 for the designed PaDa-I) they are represented in a time dependent way (from red to blue representing the time evolution). The backbone of H93 for SOD5 does not fluctuate so much, however for H248 in PaDa-I the backbone is not so stable and at the end of the simulation it is very far from the other histidines. The distance distribution (panel c) shows similar distribution for two distances but the H248 distribution is much more diffused than the H93 one. Local RMSD plots in (d) clearly show a better stability for SOD5 with a value less than 1 Å while in PaDa-I RMSD increases until a value of 5 Å due to H248.

These results indicate that the designed metal site in PaDa-I is not as stable as the original site in SOD5, the main problem being the large conformational changes seen in in H248.

Simulation with the copper

Figure 4.4 shows the results for 100 ns MD simulations for SOD5 and PaDa-I system in the holo form. The simulation was performed adding bond and angle constraints to the copper and each coordinating atom. (a) plot shows that the distance distribution from copper to nitrogen is similar for both systems and it is conserved and centred in an equilibrium position, obviously because a restraint is applied. In (b), values for angles are smaller in PaDa-I because the coordination tends to be more tetrahedral than planotrigonal. In (c) distance distribution for His-Glu is plotted. This is an important interaction and both distributions are quite similar. In (d) a plot of Glutamic and Histidine for SOD5 is shown.

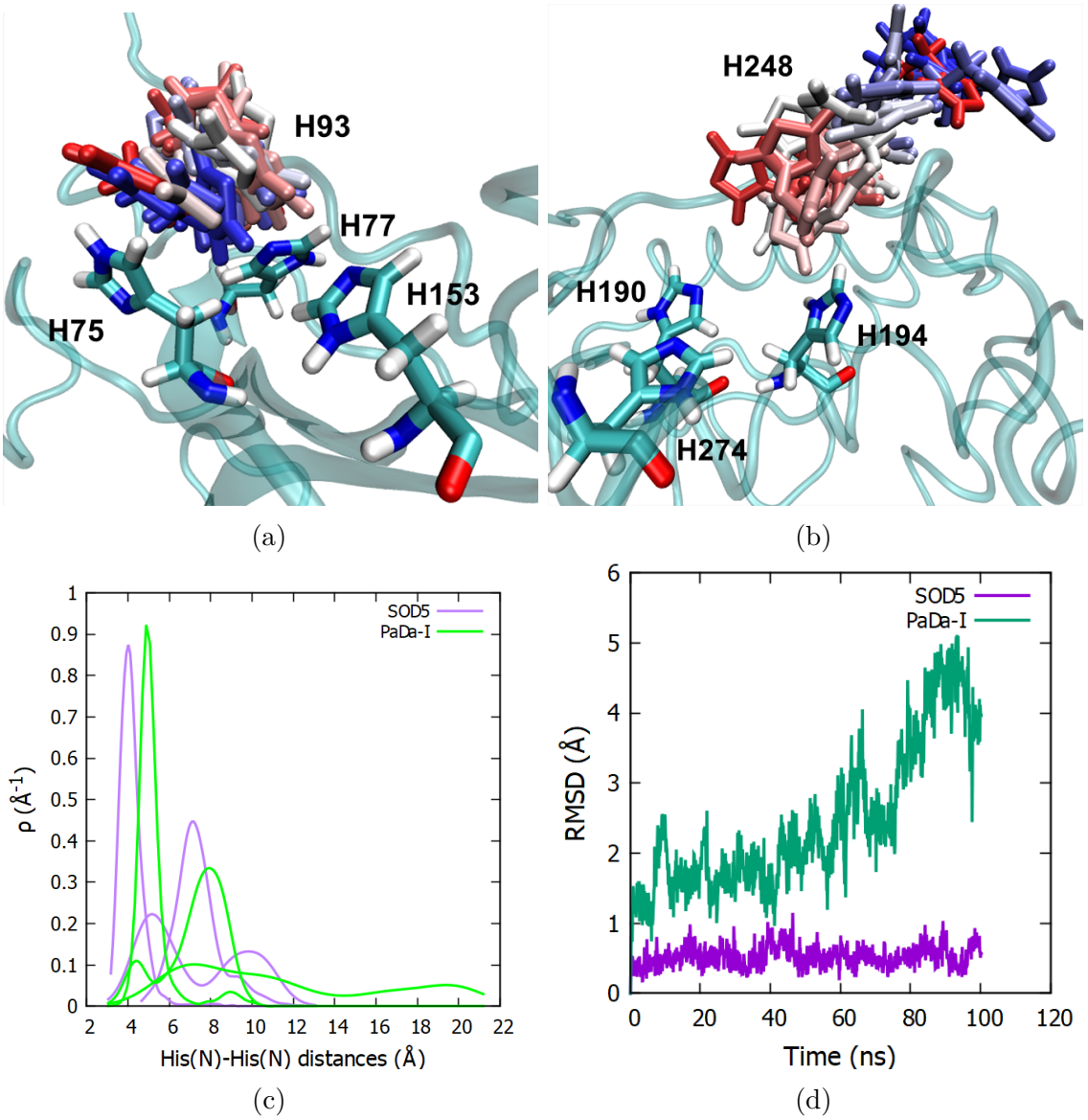


Fig. 4.3 Molecular dynamics simulations without the copper. (a): snapshot of MD simulation for SOD5 system, (b): snapshot of MD simulation for the designed PaDa-I system, (c): distance distribution between N-coordinating histidines and (d): local RMSD of the four histidines.

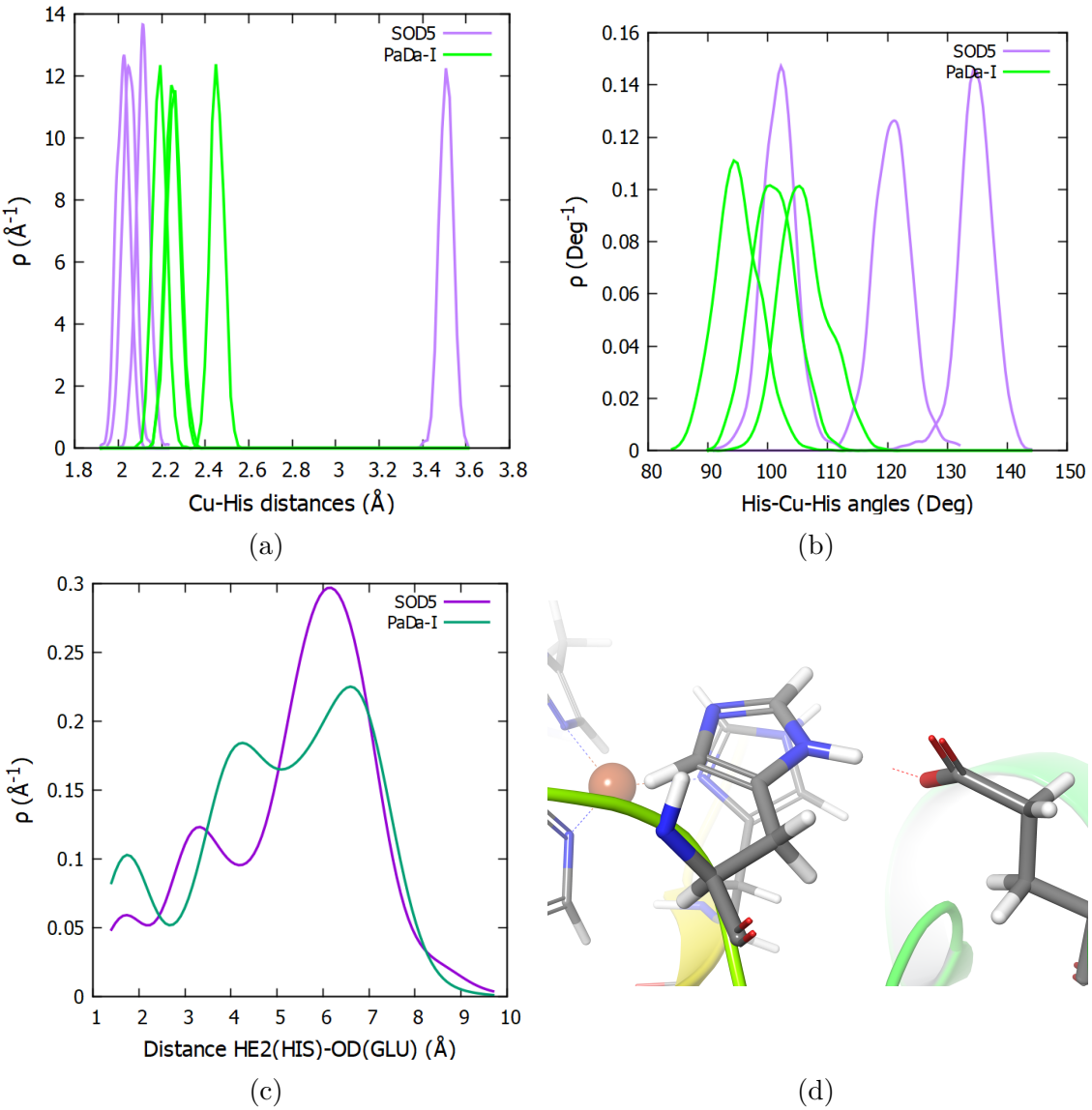


Fig. 4.4 Molecular dynamics simulations with copper. (a): distance distribution between N-coordinating histidines and copper, (b): angle distribution (N-Cu-N) N-coordinating histidines and copper, (c): distance distribution between histidine and glutamate and (d): His and Glu forming a hydrogen bond for SOD5 system.

4.2 Design of a catalytic triad

In section 4.2.1 the different mutations are explained and in sections 4.2.2 and 4.2.3 PELE results and MD are shown, respectively. Supporting data for the graphs is shown in Appendix A.2.

4.2.1 Proposed mutations

Following the workflow described in Figure 3.3 we performed a global PELE exploration with a glyceryltripropionate ligand. Seven different initial structures with the ligand in different positions around the enzyme were prepared.

In Figure 4.5 we observe that the most favorable poses are the ones in which the ligand entered the original cavity where the heme group is. Mutations can not be done there because it would interfere with the oxygenase activity of the enzyme. However there are interesting poses at 15-20 Å (boxed poses) in the plot. This site has been explored visually to see if it has a proper site to design a catalytic triad and three different mutations have been performed and studied with PELE and MD simulations.

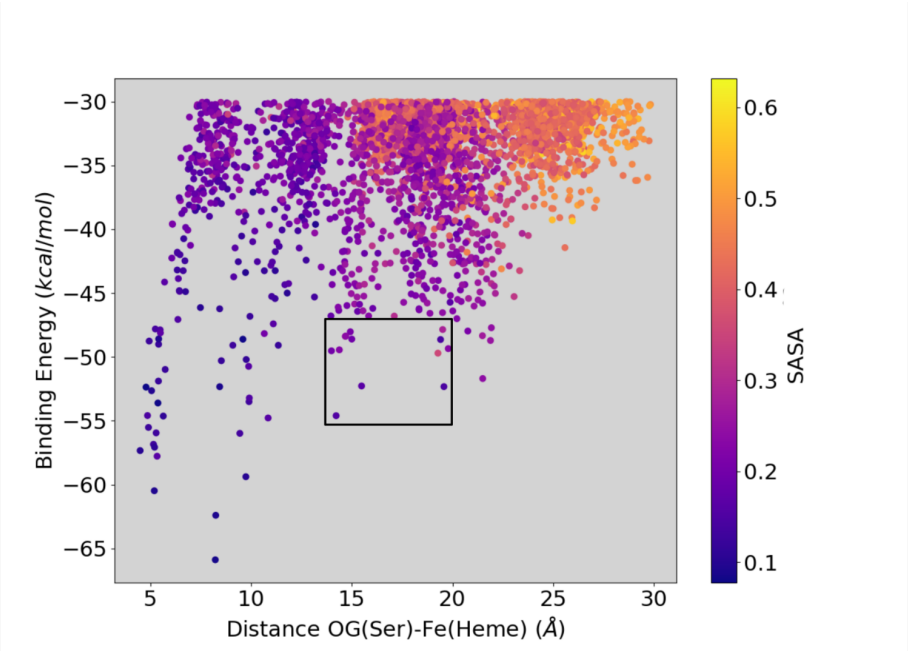


Fig. 4.5 PELE exploration for glyceryltriacetate ligand in PaDa-I. Binding energy of the ligand depending on Fe(HEME)-C(ligand) distance

The three designed catalytic triads are shown in Table 4.1 with their corresponding mutations. *A* triad consists of three mutations while *B* and *C* consists of only one

mutation. Analysing these mutations from an alignment point of view with BLOSUM62 matrix, the score for A is -5 ($R257D_{BLOSUM62} = -2$, $D268H_{BLOSUM62} = -1$, $T270S_{BLOSUM62} = -2$), -3 for B ($V186H_{BLOSUM62} = -3$) and -1 for C ($D124H_{BLOSUM62} = -1$). With this information one can expect that C mutation has less negative impact on the enzyme structure.

Table 4.1 Proposed catalytic triads and its labels

Mutations	Catalytic triad	Label
R257D/D268H/T270S	D257/H268/S270	A
V186H	D184/H186/S187	B
D124H	D131/H124/S184	C

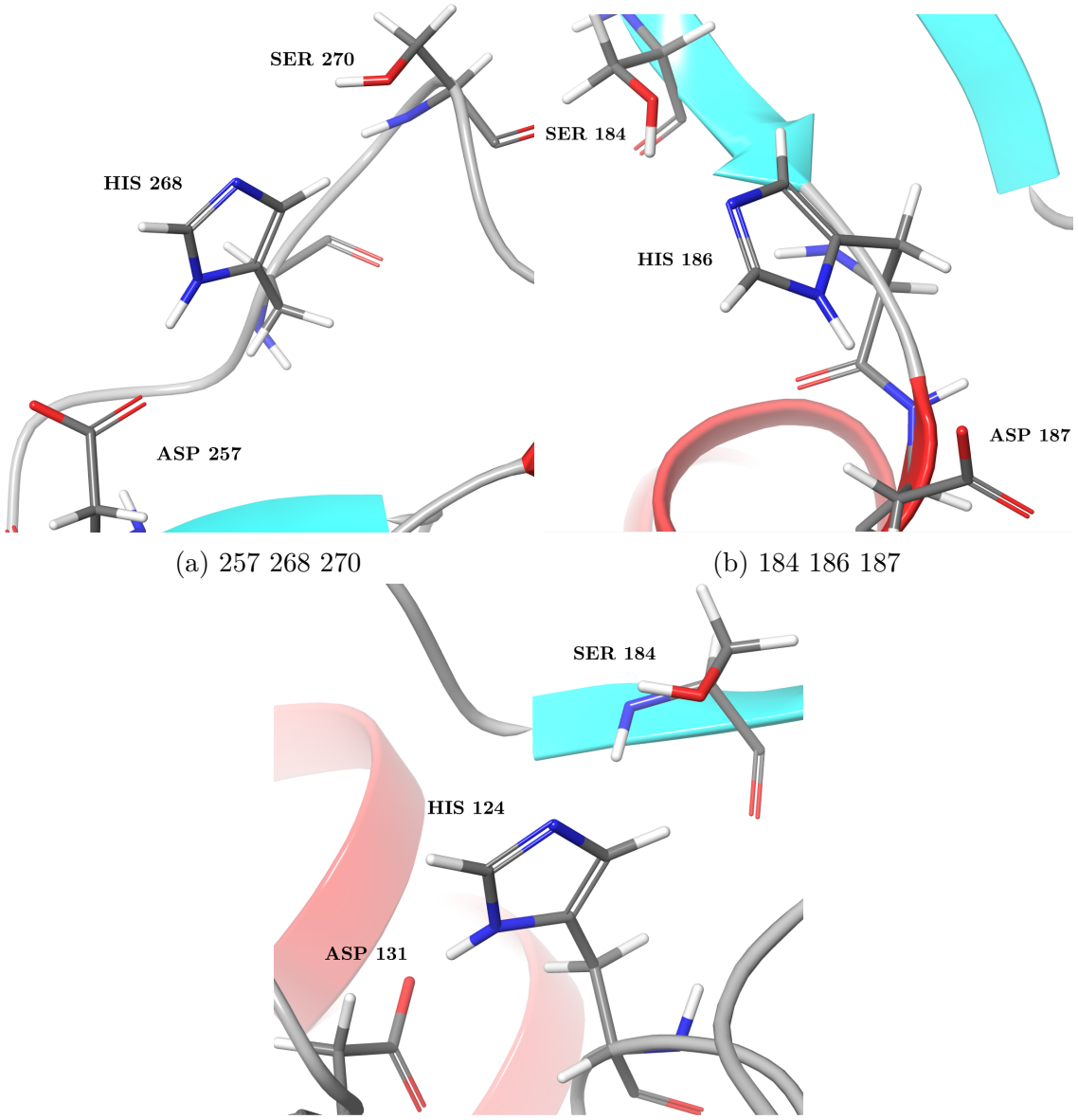
In Figure 4.6 the three catalytic triads are displayed. A triad is located in a loop and it is really exposed to the medium so before running a simulation it can be expected that it is not going to be stable. B and C triads are more buried but also on the surface so better results are expected.

4.2.2 PELE analysis

In order to understand the designed variants here we analyze the results for PELE explorations in the site of the created catalytic triad. In Figure 4.7 we show some PELE plots for the three variants and different ligands.

In Figure 4.7a the mutation site is explored and a minimum structure is found at 3.84 Å Ser-ligand distance. For (b) to (e) plots besides the local exploration on the designed triad we also explore the PaDa-I original hole so the minimum poses appearing at 15-20 Å correspond to structures in which the ligand has entered that hole. For 4.7b and 4.7c a minimum is observed in the catalytic triad site but the Ser-ligand distance is large. From 4.7d we observe that for a 'big' ligand, the ligand does not enter that much on UPO hole as in previous cases and its binding energy is similar to the one in the catalytic triad site. This is an interesting result because if the goal is to insert a covalent inhibitor into the protein a big molecule could be more suitable than a smaller one. For 4.7e a local minimum is observed at a distance of 3.5 Å.

In Figure 4.8 we show two plots for the distribution of catalytic distances in PELE simulations with glyceryltriacetate for B and C variants and glycerylpropionate for A. Regarding Asp-His distances, they are maintained at $\sim$1.5 Å despite A and B triads



(a) 257 268 270

(b) 184 186 187

(c) 124 131 184.

Fig. 4.6 Designed catalytic triads. (a): *A* catalytic triad, (b): *B* catalytic triad, (c): *C* catalytic triad

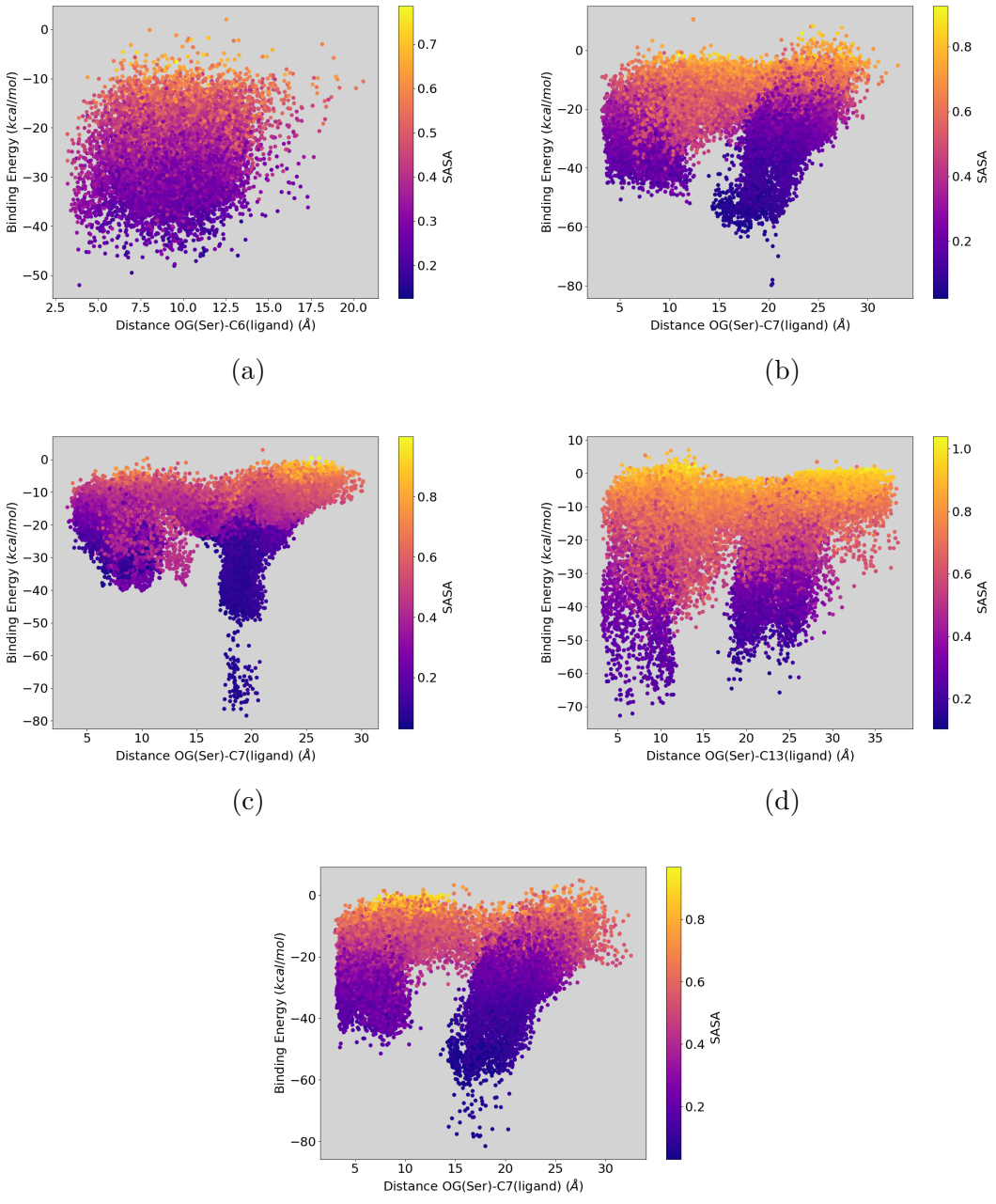
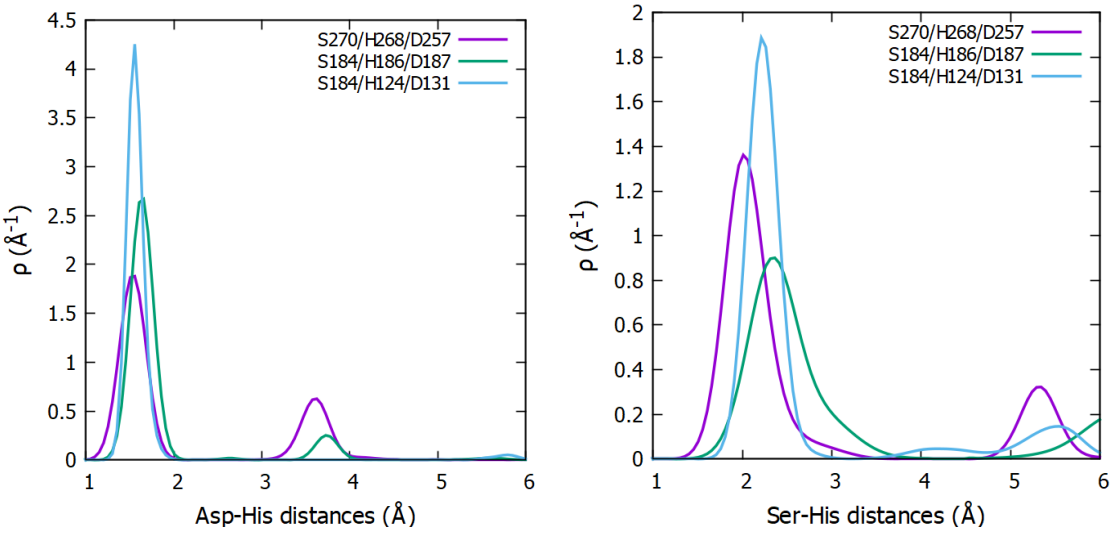


Fig. 4.7 PELE simulations for the different proposed mutations with different ligands. (a): mutation *A*, glyceryltripropionate ligand, (b): mutation *B*, glyceryltriacetate ligand, (c): mutation *B*, phenylacetate ligand, (a): mutation *B*, α-glucosepentacetate ligand, (a): mutation *C*, glyceryltriacetate ligand.



(a) Asp-His. (b) Ser-His

Fig. 4.8 PELE distribution distances for the different variants. (a): Distance distribution of OD(Asp)-HD1(His), (b): Distance distribution of NE(His)-HG(Ser).

have fewer populations at ∼3.5 Å. For 4.8b Ser-His distances are between ∼ 2-2.5 Å in most poses.

To sum up, catalytic distances between residues are maintained, a good signal for the designed mutations. However, a PELE distance analysis must be taken carefully because the obtained structures are minimized in each step and a bias can be created.

4.2.3 MD analysis

In this section we address the (100 ns) MD simulations performed to complement PELE's studies. They are separated in two sections: with and without ligand.

Simulation without a ligand

Distribution of catalytic distances and global RMSD are plotted in Figure 4.9 for a MD simulation without a ligand. In (a) it is observed that mutation B and C are the ones in which aspartic - histidine distances are more conserved with a value of ∼ 2 Å and ∼ 4.5 Å, respectively. However, for A mutation the distance is not stable. Regarding serine - histidine distances (b) B and C mutation have a maximum at ∼ 5 Å and ∼ 4 Å respectively. A has the maximum at 2.5 Å but there is also a big contribution for higher distances. All RMSD values are stabilized around a value of 2 Å so the specific mutations do not alter significantly the whole structure.

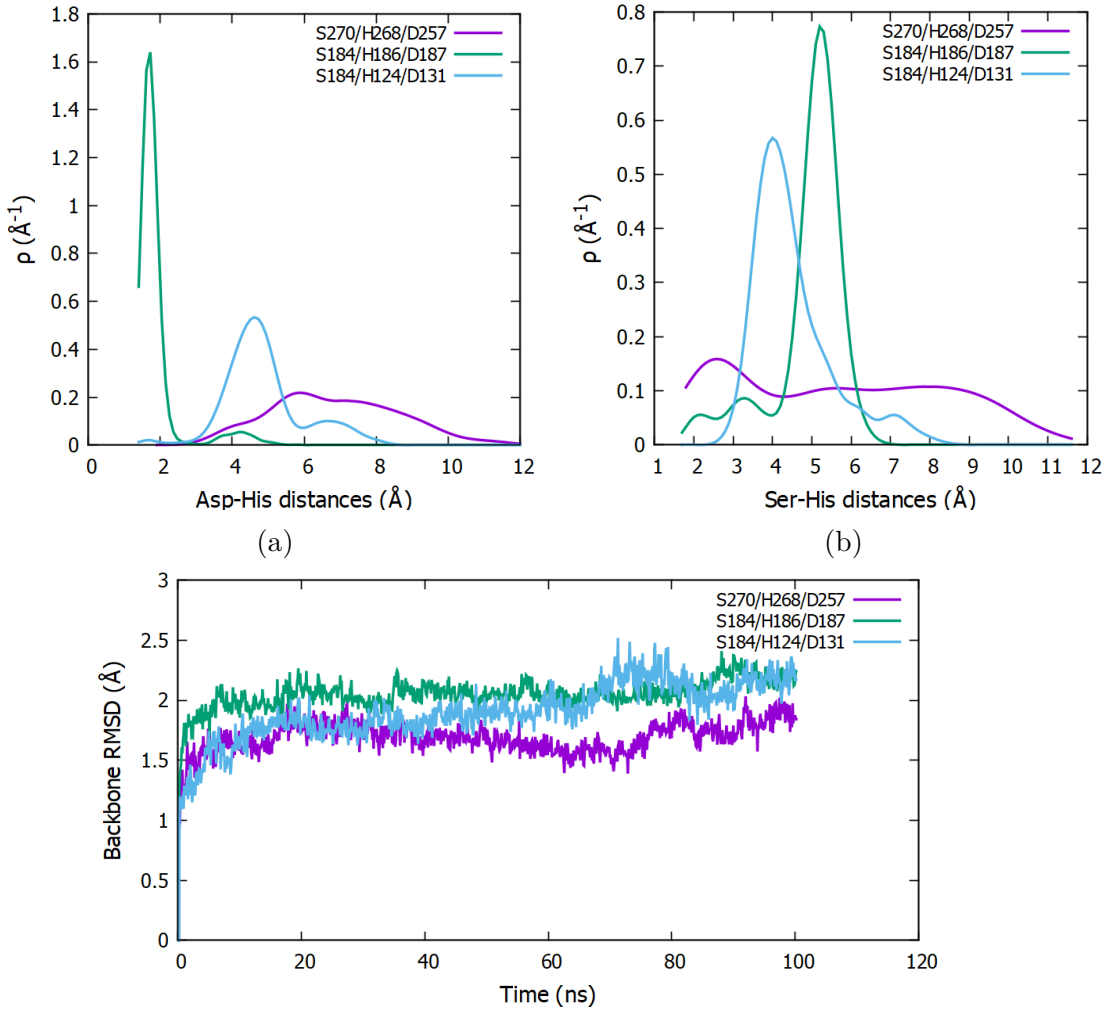


Fig. 4.9 Parameters along a 100 ns molecular dynamics simulations for the designed catalytic triads without a ligand. (a): Distance distribution of OD(Asp)-HD1(His), (b): Distance distribution of NE(His)-HG(Ser), (c): Global backbone RMSD. (a) and (b) plots are smoothed with a kernel gaussian function.

Simulation with a ligand

Distribution of catalytic distances and local RMSD are plotted in Figure 4.10 for MD simulations with a ligand. Results are similar from the ones in Figure 4.9 so the presence of the ligand does not affect significantly the catalytic distances. Regarding the Ser-ligand distribution distances (c) for A and C mutations the ligand goes away the catalytic site during the simulation, however for B one, the ligand remains near the serine residue. Finally, the local RMSD (d) (including the three catalytic residues) shows that for mutation A there is a bigger variation than the other ones. This fact is comprehensible because catalytic distances were not conserved.

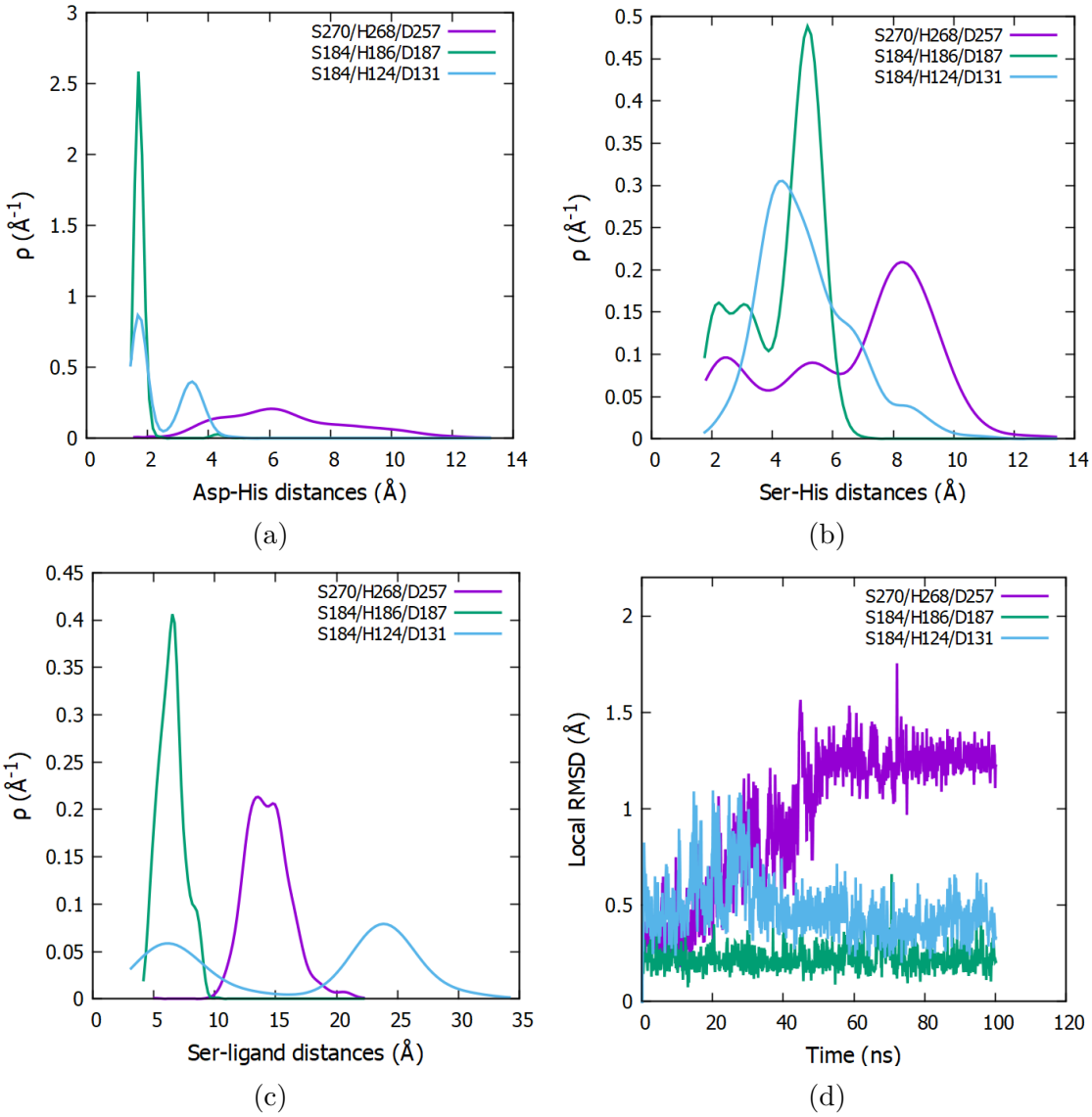


Fig. 4.10 Parameters along a 100 ns molecular dynamics simulations for the designed catalytic triads with the glyceryltriacetate ligand. (a): Distance distribution OD(Asp)-HD1(His), (b): Distance distribution NE(His)-HG(Ser), (c) Distance distribution of OG(Ser)-C(ligand) where 'C(ligand)' is the ester carbonylic carbon more close to OG(Ser) at each MD step, (d): Local backbone RMSD. (a), (b) and (c) plots are smoothed with a kernel gaussian function.

4.3 Results for AdnMD

We discuss here the results for running AdnMD with PaDa-I PDB as input and the coordinates of the coordinating histidines in SOD5 enzyme. Residues Cys36, Glu122, Gly123 and Ser126 were not considered due to the fact that they participate in the coordination of metals. Tables with best scores and their corresponding Figures are found in Appendix B.

4.3.1 3-His coordination

The best structure found is to mutate residues 23, 25 and 50 to histidines with a score of 1.43 Å. However, these results must be taken carefully because the first selection should more proper or increase the number of selected scores (1000 in this case). The bottleneck of the procedure is that most of the chosen scores at the first selection do not have proper orientation. Despite the disposition of α-carbons is very similar the three side chains do not point to the same place.

Regarding the 3-coordinating histidines designed in section 4.1.1 (H190/H194/H274) they are in position 10746 over 5564321 combinations with a score of 1.18 Å. As only the first 1000 scores are selected this combination did not pass to the next step, so this fact shows the facts commented before.

4.3.2 4-His coordination

The best structure found is to mutate residues 88, 293, 303 and 304 to histidnes with a score of 1.94 Å. The same analysis like in the previous section can be done. The modelization of the metal site should continue in performing simulations over the best obtained structures. Moreover, the metal have to be inserted.

However, it is interesting to infer that the region involving residues 88, 298, 293, 300, 303 and 304 appear three times in the best eight scores so this region of the protein seems to be an excellent place to study the modelization in more detail.

Regarding the 4-coordinating histidines designed in section 4.1.1 (H190/H194/H248/ H274) they are in position 228320 over 445145680 combinations with a score of 2.61 Å. As only the first 5000 scores are selected this combination did not pass to the next step.

Chapter 5

Conclusions

In this work, with the use of PELE and MD simulations, some interesting conclusions have been reached. Two different types of active site have been modeled, a copper metal site and a catalytic triad in an unspecific peroxygenase called PaDa-I.

Regarding the design of a metal binding site, it consisted in mimicking a superoxide dismutase in PaDa-I. PELE simulations showed that superoxide binds copper with similar energy in both systems. However, the results are not as good as they should because MD simulations for the apo form of the enzymes showed that the stability of histidine residues (concretely H248) in PaDa-I is not as good as the ones in SOD5.

Regarding the design of the catalytic triad, B and C mutations shown the most promising results due to the conservation of the catalytic distances and RMSD values in MD simulations.

For future perspectives in the designing of metaloPluriZymes and catalytic triad an automated procedure should be followed to gain time and accuracy. The developed Python script could be improved to find better solutions to design a specific enzyme.

Finally, the proposed mutations should be sent to experimental validation to check the activity of the designed site, plus studying the economic viability as the main application is the reduction of industrial production of drugs, apart from the scientific interest.

References

[1] P. Molina-Espeja, E. Garcia-Ruiz, D. Gonzalez-Perez, R. Ullrich, M. Hofrichter, and M. Alcalde. Directed evolution of unspecific peroxygenase from Agrocybe aegerita. *Appl. Environ. Microbiol.*, 80(11):3496–3507, Jun 2014.

[2] Fátima Lucas, Esteban D. Babot, Marina Cañellas, José C. del Río, Lisbeth Kalum, René Ullrich, Martin Hofrichter, Victor Guallar, Angel T. Martínez, and Ana Gutiérrez. Molecular determinants for selective c25-hydroxylation of vitamins d2 and d3 by fungal peroxygenases. *Catal. Sci. Technol.*, 6:288–295, 2016.

[3] P. Molina-Espeja, M. Canellas, F. J. Plou, M. Hofrichter, F. Lucas, V. Guallar, and M. Alcalde. Synthesis of 1-Naphthol by a Natural Peroxygenase Engineered by Directed Evolution. *Chembiochem*, 17(4):341–349, Feb 2016.

[4] Patricia Gomez de Santos, Marina Cañellas, Florian Tieves, Sabry H. H. Younes, Patricia Molina-Espeja, Martin Hofrichter, Frank Hollmann, Victor Guallar, and Miguel Alcalde. Selective synthesis of the human drug metabolite 5´-hydroxypropranolol by an evolved self-sufficient peroxygenase. *ACS Catalysis*, 8(6):4789–4799, 2018.

[5] Gerard Santiago, Mónica Martínez-Martínez, Sandra Alonso, Rafael Bargiela, Cristina Coscolín, Peter N. Golyshin, Víctor Guallar, and Manuel Ferrer. Rational engineering of multiple active sites in an ester hydrolase. *Biochemistry*, 57(15):2245–2255, 2018. PMID: 29600855.

[6] Yuewei Sheng, Isabel A. Abreu, Diane E. Cabelli, Michael J. Maroney, Anne-Frances Miller, Miguel Teixeira, and Joan Selverstone Valentine. Superoxide Dismutases and Superoxide Reductases. *Chemical Reviews*, 2014.

[7] Maan Hayyan, Mohd Ali Hashim, and Inas M. AlNashef. Superoxide ion: Generation and chemical implications. *Chemical Reviews*, 116(5):3029–3085, 2016. PMID: 26875845.

[8] Ryan L. Peterson, Ahmad Galaleldeen, Johanna Villarreal, Alexander B. Taylor, Diane E. Cabelli, P. John Hart, and Valeria C. Culotta. The phylogeny and active site design of eukaryotic copper-only superoxide dismutases. *Journal of Biological Chemistry*, 2016.

[9] Manabu Fujii and Erika Otani. Photochemical generation and decay kinetics of superoxide and hydrogen peroxide in the presence of standard humic and fulvic acids. *Water Research*, 2017.

[10] Zhiguang Xiao and Anthony G. Wedd. The challenges of determining metal-protein affinities, 2010.

[11] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 1996.

[12] Donald Bashford and David A Case. Generalized Born Models of Macromolecular Solvation Effects. *Annual Review of Physical Chemistry*, 51(1):129–152, 2000.

[13] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 2013.

[14] Zhexin Xiang, Peter J. Steinbach, Matthew P. Jacobson, Richard A. Friesner, and Barry Honig. Prediction of side-chain conformations on protein surfaces. *Proteins: Structure, Function and Genetics*, 2007.

[15] Molecular dynamics—Scalable algorithms for molecular dynamics simulations on commodity clusters. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC '06*, 2006.

[16] H. J.C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *Journal of Physical Chemistry*, 1987.

[17] Glenn J. Martyna, Michael L. Klein, and Mark Tuckerman. Nosé-Hoover chains: The canonical ensemble via continuous dynamics. *The Journal of Chemical Physics*, 1992.

[18] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189, sep 1994.

[19] Jonathan Rittle, Mackenzie J. Field, Michael T. Green, and F. Akif Tezcan. An efficient, step-economical strategy for the design of functional metalloproteins. *Nature Chemistry*, 2019.

[20] Cheng Zhu, Changsheng Zhang, Huanhuan Liang, and Luhua Lai. Engineering a zinc binding site into the de novo designed protein DS119 with a $\beta\alpha\beta$ structure. *Protein and Cell*, 2011.

[21] Yi Lu, Natasha Yeung, Nathan Sieracki, and Nicholas M. Marshall. Design of functional metalloproteins, 2009.

[22] Fangting Yu, Virginia M. Cangelosi, Melissa L. Zastrow, Matteo Tegoni, Jefferson S. Plegaria, Alison G. Tebo, Catherine S. Mocny, Leela Ruckthong, Hira Qayyum, and Vincent L. Pecoraro. Protein design: Toward functional metalloenzymes, 2014.

[23] Jefferson S. Plegaria and Vincent L. Pecoraro. De novo design of Metalloproteins and metalloenzymes in a Three-Helix bundle. In *Methods in Molecular Biology*. 2016.

[24] Jaime Rodríguez-Guerra Pedregal, Giuseppe Sciortino, Jordi Guasp, Martí Municoy, and Jean-Didier Maréchal. GaudiMM: A modular multi-objective platform for molecular modeling. *Journal of Computational Chemistry*, 2017.

[25] Giuseppe Sciortino, Eugenio Garribba, Jaime Rodríguez-Guerra Pedregal, and Jean Didier Maréchal. Simple Coordination Geometry Descriptors Allow to Accurately Predict Metal-Binding Sites in Proteins. *ACS Omega*, 2019.

Appendix A

Distance and RMSD data

A.1 Design of a metal site

A.1.1 Simulation without the copper

Table A.1 Distances between histidine's coordinating nitrogens in Å

Enzyme	Distance 1	Distance 2	Distance 3	RMSD
SOD5	4.17 ± 0.60	7.37 ± 1.14	7.2 ± 2.39	0.52 ± 0.16
PaDa-I	5.23 ± 0.99	7.23 ± 1.49	11.59 ± 4.95	2.45 ± 1.04

A.1.2 Simulation with the copper

Table A.2 Copper-histidine(N) distances in Å

Enzyme	Distance 1	Distance 2	Distance 3	Distance 4
SOD5	2.119 ± 0.031	2.056 ± 0.030	2.032 ± 0.030	3.512 ± 0.031
PaDa-I	2.195 ± 0.031	2.254 ± 0.032	2.261 ± 0.034	2.460 ± 0.030

Table A.3 Histidine(N)-copper-histidine(N) angles in Degrees

Enzyme	Angle 1	Angle 2	Angle 3
SOD5	102.4 ± 2.8	121.45 ± 3.1	135.2 ± 3.0
PaDa-I	95.5 ± 3.7	101.4 ± 3.6	106.6 ±4.0

Table A.4 His(HD1)-Glu(OD) distances in Å

Enzyme	Distance
SOD5	5.4 ± 1.7
PaDa-I	5.1 ± 1.8

A.2 Design of a catalytic triad

A.2.1 PELE results

Table A.5 PELE distances in Å

Catalytic triad	d(Ser(HG)-His(NE2))	d(His(HD1)-Asp (OD1))
A	2.7 ± 1.3	3.12 ± 0.95 9
B	3.5 ± 1.8	3.63 ± 0.76
C	2.7 ± 1.2	3.47 ± 0.51
	d(His(HD1)-Asp(OD2))	
A	2.13 ± 0.99	
B	1.93 ± 0.88	
C	1.65 ± 0.59	

A.2.2 MD results

Table A.6 MD without ligand, distances and RMSD in Å

Catalytic triad	d(Ser(HG)-His(NE2))	d(His(HD1)-Asp (OD))	Global RMSD
A	5.7 ± 2.7	6.9 ± 1.8	1.69 ± 0.12
B	4.95 ± 0.91	1.94 ± 0.66	2.05 ± 0.12
C	4.6 ± 1.0	4.9 ± 1.1	1.91 ± 0.23

Table A.7 MD with ligand, distances and RMSD in Å

Catalytic triad	d(Ser(HG)-His(NE2))	d(His(HD1)-Asp (OD))
A	6.6 ± 2.6	6.6 ± 2.2
B	4.33 ± 1.24	1.77 ± 0.36
C	5.1 ± 1.5	2.45 ± 0.90
	d(Ser(OG)-ligand(C_{min}))	Local RMSD
A	14.4 ± 1.8	0.96 ± 0.37
B	6.5 ± 1.0	0.223 ± 0.061
C	16.5 ± 8.9	0.48 ± 0.15

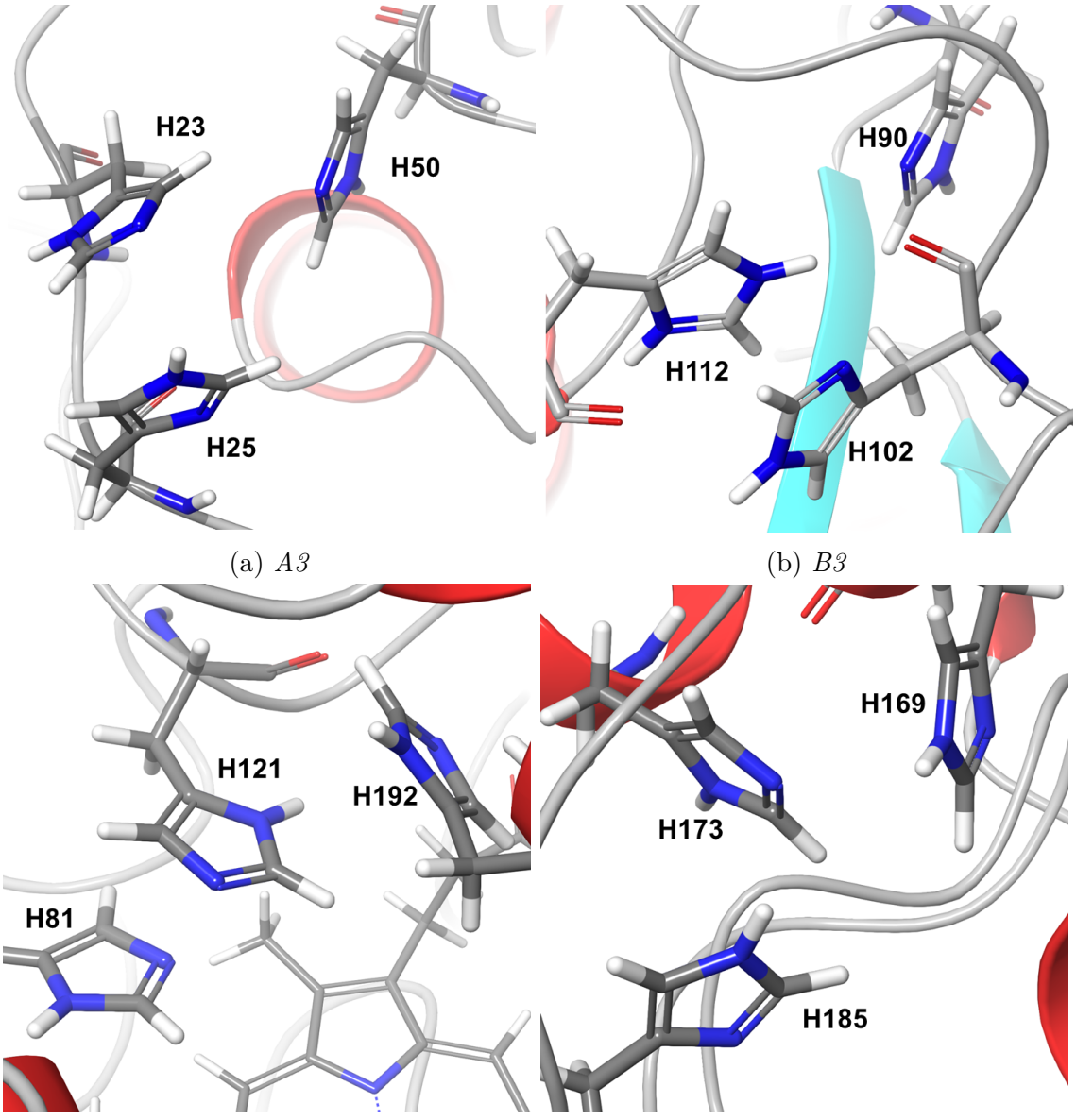
Appendix B

Results for AdnMD

B.1 3-His coordination

Table B.1 Best results obtained with AdnMD doing three mutations to histidines

Mutation	1st selection		Final score	
	Position	Score (Å)	Position	Score (Å)
A3: H23/H25/H50	961	0.49	1	1.43
B3: H90/H102/H112	702	0.44	2	1.83
C3: H81/H121/H192	113	0.23	3	1.99
D3: H169/H173/H185	236	0.31	4	2.12
E3: H3/H56/H79	656	0.43	5	2.35
F3: H80/H280/H284	354	0.35	6	2.45

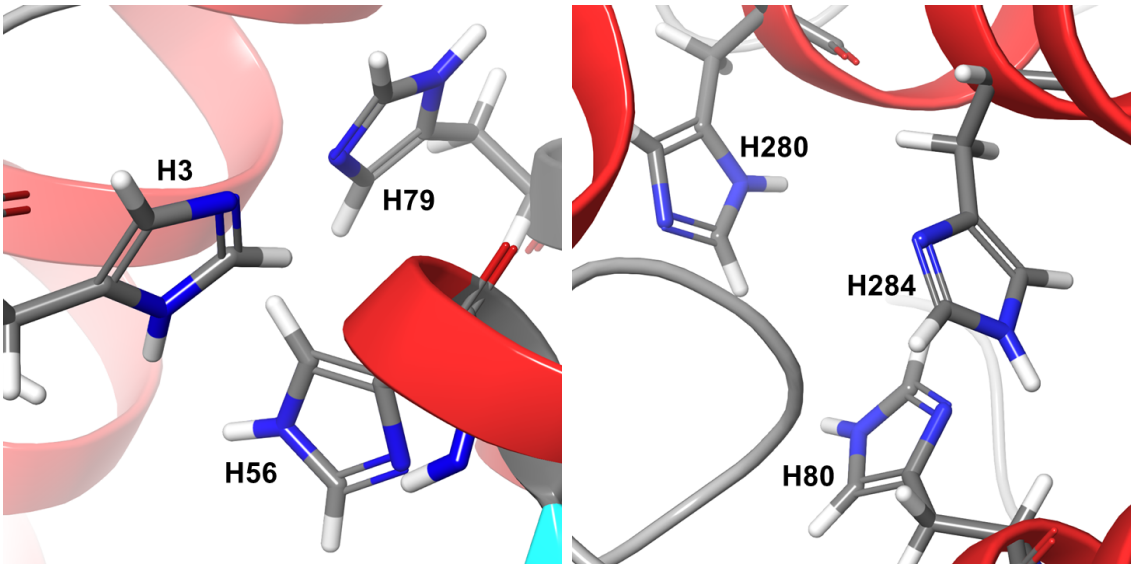


(a) *A3*

(b) *B3*

(c) *C3*

(d) *D3*



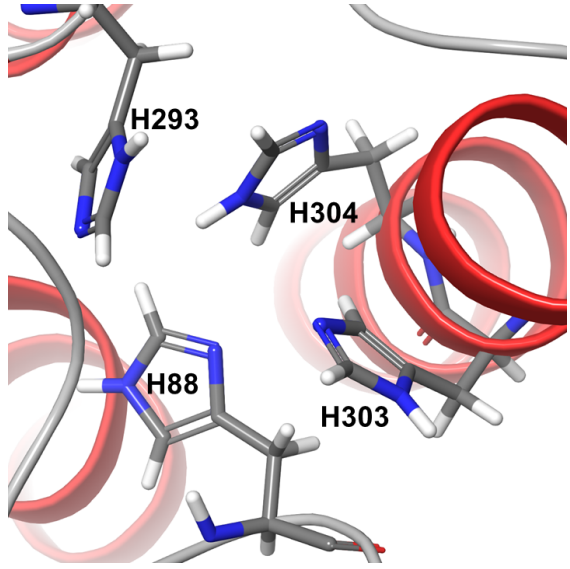
(e) *E3* (f) *F3*

Fig. B.1 Structures with best scores with 3-His coordination.

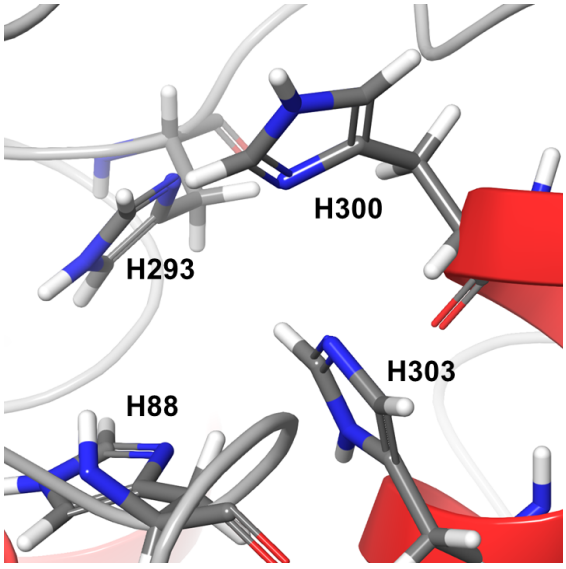
B.2 4-His coordination

Table B.2 Best results obtained with AdnMD doing four mutations to histidines

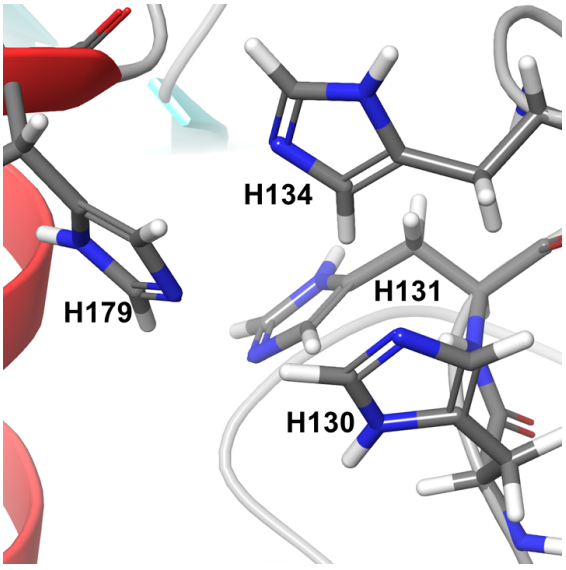
Mutation	1st selection		Final score	
	Position	Score (Å)	Position	Score (Å)
A4: H88/H293/H303/H304	1478	0.66	1	1.94
B4: H88/H293/H300/H303	1196	0.62	2	2.08
C4: H130/H131/H134/H179	2352	0.74	3	2.29
D4: H77/H195/H240/H241	4265	0.85	4	3.13
E4: H169/H193/H194/H253	2095	0.72	5	3.19
F4: H48/H59/H62/H95	2344	0.74	6	3.23
G4: H148/H152/H160/H200	4305	0.86	7	3.39
H4: H88/H289/H303/H304	1710	0.68	8	3.47



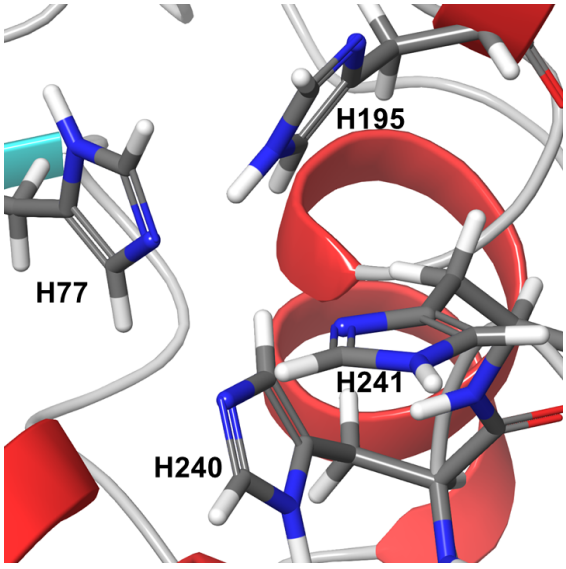
(a) *A4*



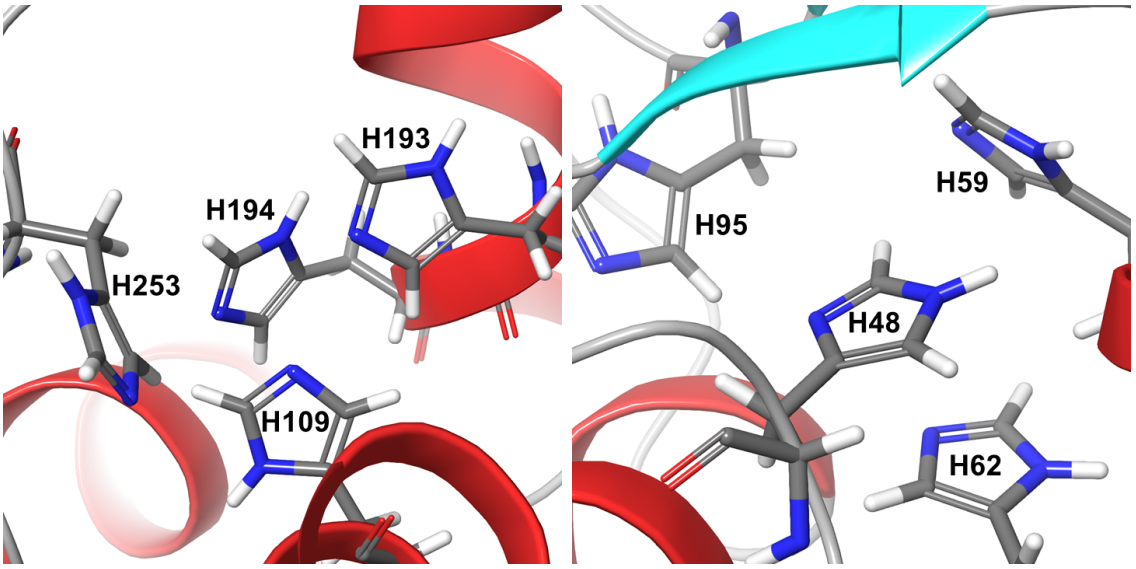
(b) *B4*



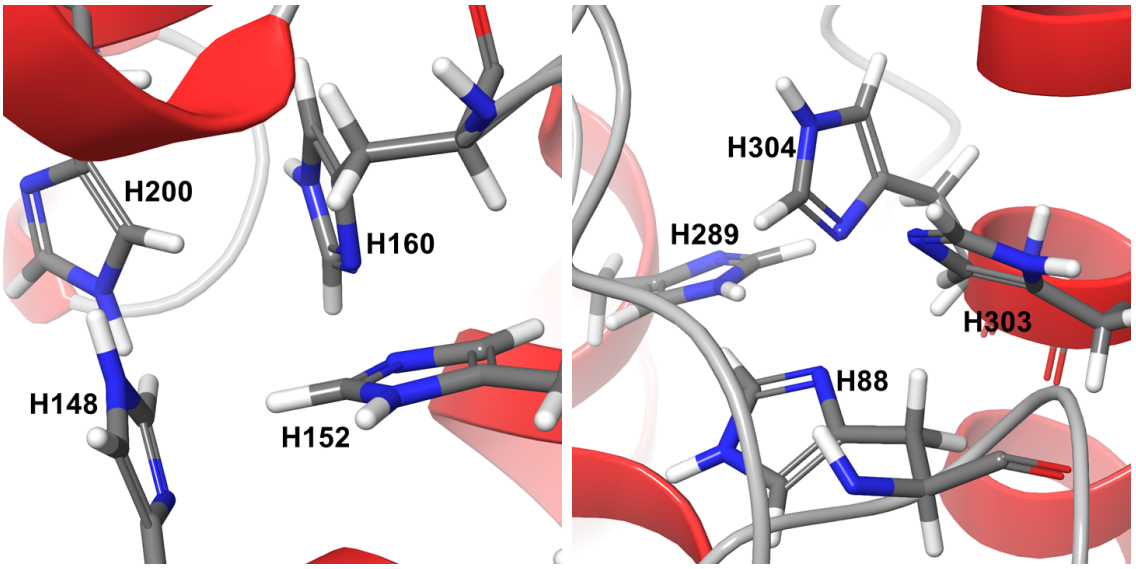
(c) *C4*



(d) *D4*



(e) *E4* (f) *F4*



(g) *G4* (h) *H4*

Fig. B.2 Structures with best scores with 4-His coordination.