



Development and Applications of Molecular Modelling Techniques for the Design and Optimization of Artificial Metalloenzymes

Victor Muñoz Robles

Ph. D. Thesis

Ph. D. in Biotechnology

Supervisors

Jean-Didier Maréchal

Agustí Lledós Falcó

Departament de Genètica i Microbiologia

Facultat de Ciències

2014



Universitat Autònoma de Barcelona
Departament de Química
Unitat de Química Física

Memòria presentada per aspirar al Grau de Doctor per Victor
Muñoz Robles

Victor Muñoz Robles

Vist i plau,

Jean-Didier Maréchal

Agustí Lledós Falcó

Bellaterra, 5 de Febrer de 2014

*A mi familia.
A mis amigos.
A mi amor.
A todos aquellos
gracias a los que hoy
soy como soy.*

Official Acknowledgements

I would like to thank the Spanish *Ministerio de Economía y Competitividad* (MINECO) for the financial support in the framework of the project CTQ2008-06886-C02-01 and for the FPI Fellowship (reference BES-2009-015849), which has made possible this Ph. D. thesis. I am also thankful to the MINECO for the financial assistance that allowed me to stay for three months in the group of Thomas Ferrin at the UCSF (San Francisco, USA) and another three months in the group of Thomas R. Ward at the University of Basel (Basel, Switzerland).

Unofficial Acknowledgments

Bueno, aquí estamos por fin! No se cuantas veces he dicho que ya era el final.... pero parece ser que esta vez sí!! Ya sólo falta escribir los agradecimientos y podré encuadernar la tesis!! He dejado esta parte para el final pensando que sería la mas fácil de escribir, pero nada mas lejos de la realidad. Sentado aquí, en la mesa dónde he pasado los últimos cuatro años y mirando la caja de cartón donde esta la primera impresión de mi tesis, me doy cuenta de que calificarla como “*mi*” tesis tal vez sea demasiado egocéntrico. Llegar hasta este punto no habría sido posible sin la ayuda y el apoyo que he recibido por parte de muchas personas, y no sólo durante mi doctorado, sino desde mucho antes. Por este motivo, quiero dedicar esta disertación a todas ellas.

A mis padres y a mi hermana. Por haberme apoyado desde siempre de forma incondicional. Porque aunque nunca entendisteis del todo qué es eso de la biotecnología o qué significa una tesis doctoral (seamos sinceros, os habría gustado mil veces mas qué hubiera hecho medicina), siempre confiasteis en que había tomado la decisión correcta porque había elegido hacer aquello que realmente me apasionaba.

A mis amigos de Mollet. A Edu, Enric, Ricard, etc. Por todo lo que nos ha unido durante todos estos años. Por haber estado siempre ahí, tanto en los momentos buenos como en los no tan buenos y por haber tenido tanta paciencia con mis (no pocas) liadas. Puede que pasé por muchas versiones diferentes, pero en el fondo siempre seré vuestro Viti.

A mis amigos de las Magics. A David, Oriol, Roc. Por esas noches del miércoles y por todos esos buenos ratos jugando a nuestro juego preferido de cartas. ¡Rayo a la cabeza!

A mis amigos de la carrera. A Aida, Ángela, Anna Puiggros, Cristina, Hernando, Núria, María, Roger, etc. Por todos aquellos cortos pero maravillosos años que fueron nuestra etapa universitaria y para que sigamos haciendo nuestras (aunque escasas) quedadas para recordar viejos tiempos.

A Jordi Corominas. A ti te reservo un apartado especial. Quién sabe dónde estaría sin nuestras sesiones de estudio, sin nuestros tutes y sin mi vaso personalizado anti-Viti. Por todas esas reuniones de “*sabis*” y por todo el apoyo a nivel personal que siempre me has dado.

A las Leonesas. A Coral, Paula, etc. Por todos esos congresos en los que aprendimos muchísimo asistiendo a todas las ponencias (guiño, guiño). Aunque ya se nos ha pasado la época de volver a ir, espero que nunca perdamos el contacto y podamos seguir haciendo esas escapaditas (y, no nos engañemos, me sigáis trayendo esa cecina tan rica de León).

A mis compañeros de trabajo. Por todos esos buenos ratos, esos coffees, esos congresos y esas cenacas que nos hemos pegado. A Eli, Rosa, Manu y Laia, los *joves* de transmet. Por todas esas quedadas entre semana, cenas (maldito Jose Cuervo) y fiestas que hemos disfrutado. Eli, merci por todas nuestras charlas y desahogos de frustración. Rosa, por esa chispa y por esa iniciativa que tanto te caracterizan. Manu, por mantener tu mesa ordenada (a pesar de tenerme al lado) y por aguantar mis bio-rollos (sin quedarte dormido). Laia, por esas risas y esa espontaneidad tan naturales que te definen. A Sergi y, Max, los "*veteran*" de transmet. Sergi, por todas esas risas que nos hemos echado con tus "*bacalladas*" y por quitarme el primer puesto de "*pissarras*". Max, mi compi del SAF, sin ti nunca hubiese sacado mi faceta "*f*cker*". Gracias por esos consejos y por haber sabido ponerme en mi lugar (😊). Y no me quiero olvidar tampoco del resto de gente que también han aportado su granito de arena: A Luca, Pietro, Gavor, Carles y Almudena. Merci por todo.

A Jean-Didier. El "*Jefe*". Merci por todo tu apoyo y por todos tus consejos. Por esa confianza ciega que depositaste en mi prácticamente desde el primer día y por esas charlas que hemos tenido y que han trascendido más allá de lo estrictamente laboral.

A Agustí y Gregori. Por haber hecho que este biotecnólogo se sienta como uno más de la familia de Transmet y por toda vuestra ayuda y paciencia.

To my colleagues abroad. Thanks Eric, Greg, Elaine, Conrad, etc. for treating me like I was one more of the team. Thanks Thomas for giving me the opportunity to learn so much in your group at Basel. Thanks Konrad for patiently answering all my mails (which were not just a few ones BTW).

To my foreign friends. Thanks Livia for taking care of me and having so much patient with a rusty biotechnologist. Thanks Julian for all those good moments we spend waiting for our E. coli cultures to grow. I hope you still remember my Spanish lessons!! And, definitely to all those who have help me getting over my homesick by making me feel like home.

Al Manolo y Loles. Por despertarme esta vocación científica ya desde mi juventud. Sin vuestro apoyo y ánimos y sin vuestros consejos (cuánta razón tenias en aquellas charlas de ciencia vs humanidades Manolo) jamás habría descubierto lo que ahora es más que una mera vocación.

A mi familia política. Porque aún que hace poco que os conozco, me habéis dado todo vuestro apoyo y tratado cómo a uno mas de los vuestros. Por que habéis hecho que el adjetivo “política” no sea mas que una simple formalidad.

Y, finalmente, a Soraya. Porque aunque eres una de las más recientes de esta lista, eres la que ha tenido un mayor impacto. Porque siempre has estado ahí, apoyándome en aquellas de mis decisiones que eran correctas y haciéndome ver aquellas que (por muy cabezón que me pusiera) no lo eran. Por aguantar todas mis (no pocas) manías e inseguridades y por saberme sacar una sonrisa cuándo más lo necesitaba. Por haber evitado que me convierta en el tipo de persona que nunca había querido ser. En definitiva, por todo aquello que te hace ser tan especial y que me hacen quererte tanto.

A todos vosotros, muchas gracias de todo corazón.

*“I am part of all that I have met;
Yet all experience is an arch wherethro'
Gleams that untravell'd world, whose margin fades
For ever and for ever when I move.”*

Lord Alfred Tennyson

Contents

Chapter 1 - General Concepts	1
I - Enzymes	3
I.I - Introduction to catalysis	3
I.II - Natural enzymes.....	4
I.III - Biocatalysis	5
I.IV - Man tailored enzymes.....	6
I.V - Artificial Enzymes	7
II - Catalysis with transition metals	9
II.I - General properties and basic concepts	9
II.II - Organometallic chemistry of transition metals	10
III - Proteins and metals	11
III.I - Life and metals: a love and hate relationship.....	11
III.II - Catalytic metalloproteins: Metalloenzymes	12
IV - Artificial Metalloenzymes	14
IV.I - Principles of artificial enzymes and metalloenzymes.....	14
IV.I.I - Protein based artificial metalloenzymes	15
IV.I.II - DNA based artificial metalloenzymes.....	17
V - Computation for enzyme design	19
V.I - <i>In silico</i> methods in biocatalysis	19
V.II - Computer aided de novo design	21
V.III - Challenges on artificial metalloenzymes	24
Chapter 2 - Computational Methods	27
I - Molecular Modelling	29
I.I - Introduction to the art of modelling.....	29
II - Quantum Mechanics	30
II.I - Basic Concepts.....	30

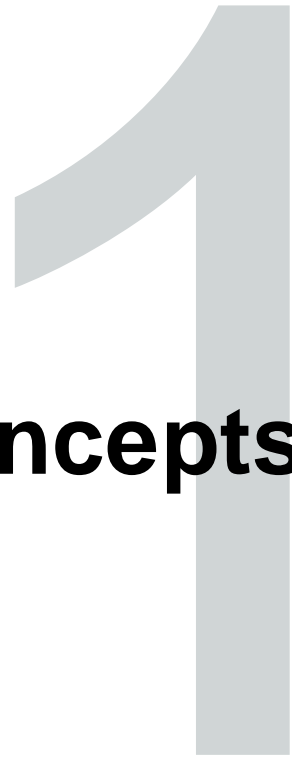
II.II - Density Functional Theory.....	31
III - Classical mechanics in molecular modelling.....	33
III.I - Molecular Mechanics – Basic Concepts.....	33
III.I.I - Force Field	33
III.II - Variety of MM techniques.....	34
III.II.I - Molecular Dynamics	34
III.II.II - Normal Modes Analysis	36
III.II.III - Protein-ligand Docking	39
IV - Potential Energy Surface	41
V - Integrative approaches.....	42
V.I - Hybrid QM/MM approach	42
VI - Integrative platforms.....	43
Chapter 3 - Objectives	47
I - Objectives	49
Chapter 4 - Prediction of the Binding of Homogeneous Catalysts to Proteins	53
I - Introduction – General concepts and challenges.....	55
II - Heme oxygenase based artificial enzyme	55
II.I - Computational tools: Integrative approaches.....	55
II.II - Benchmarking the protocol: Fe(Schiff base) \subset cdHO.....	56
II.III - Presentation of the protocol	60
II.IV - Analysis of the conformational space of the organometallic cofactor.....	63
II.V - Generation of the Fe(Schiff base) \subset cdHO model structures	64
II.VI - Analyzing the metal environment.....	68
II.VII - QM/MM refinement	70
II.VIII - Comparison with the experimental system	73

II.IX - Conclusions.....	74
III - Binding of porphyrin-like catalysts to monoclonal antibodies: Abzymes	76
III.I - Computational protocol	79
III.II - Refinement of the crystallographic structures by protein-ligand dockings.....	79
III.III - Structural basis for the optimization of porphyrinic cofactors.....	82
III.IV - Rationalization of the experimental catalytic profiles	84
III.V - Conclusions	87
IV - Decoding biotin-Streptavidin embedded systems	88
IV.I - Experimental behavior of the artificial hydrogenases.....	90
IV.II - Molecular Modelling of the structure of the pre-catalyst $[\text{Cp}^*\text{Ir}(\text{Biot-p-L})\text{Cl}] \subset \text{S112A}$ and $[\text{Cp}^*\text{Ir}(\text{Biot-p-L})\text{Cl}] \subset \text{S112K}$	91
IV.III - Computational Protocol	92
IV.IV - Identification of the preferred metal configuration for the S112A and S112K mutants	94
IV.IV.I - Rationalization of the binding process in the S112A mutant.....	94
IV.IV.II - Rationalization of the binding process in the S112K mutant	97
IV.V - Structural consequences of increasing the Ir/Sav ratio in $[\text{Cp}^*\text{Ir}(\text{Biot-p-L})\text{Cl}] \subset \text{S112A}$ and S112K	100
IV.V.I - Structural insights on the S112A fully loaded dimer	100
IV.V.II - Structural insights on the S112K fully loaded dimer	103
IV.VI - Rationalizing the kinetic and enantioselective profiles of S112A and S112K ATHase. 104	
IV.VI.I - $[\text{Cp}^*\text{Ir}(\text{Biot-p-L})\text{Cl}] \subset \text{S112A}$	105
IV.VI.II - $[\text{Cp}^*\text{Ir}(\text{Biot-p-L})\text{Cl}] \subset \text{S112K}$	106
IV.VII - Conclusions	107
Chapter 5 - Catalytic Mechanisms of Artificial Metalloenzymes.....	109
I - General introduction	111
II - Enantioselective study through binding modes analysis	112
II.I - Computational methods.....	114

II.II - Insertion of the cofactors inside the xylanase scaffold	116
II.III - Analysis of the substrate binding modes in the modeled artificial metalloenzyme	118
II.IV - Conclusions	120
III - Effects of the protein environment in the catalytic mechanism	122
III.I - Presentation of the protocol	124
III.II - Computational Methods	126
III.III - Unveiling the ATH mechanism for cyclic imines in a small cluster model	127
III.IV - Modelling the pseudo-transition state structures in the protein	133
III.V - Finding the real transition state structures	137
III.VI - Conclusions	140
Chapter 6 - Towards a Novel Integrative Computational Platform	143
I - Introduction	145
I.I - An integrative platform made for everyone	145
II - Molecular Dynamics simulation interface	147
II.I - MD visualization	151
III - Normal Modes Analysis interface	151
III.I - NMA visualization	156
IV - Novel interfaces for Gaussian and GOLD	158
IV.I - Gold visualization interface	158
IV.II - Gaussian visualization and ONIOM inputs set up interfaces	159
V - Perspectives	161
Chapter 7 - General Conclusions	163
I - Conclusions	165
Bibliography	171

Annex A: Computational Methods	A1
I - Quantum Mechanics	A3
I.I - Schrödinger equation resolution: An approximate approach	A3
I.I.I - Born-Oppenheimer approximation	A4
I.I.II - Hartree-Fock Self Consistent Field (HF-SCF)	A5
II - Force Field	A7
II.I - Bonding terms	A7
II.II - Non-Bonding terms	A8
III - Molecular Mechanics	A9
III.I - Docking Limitations	A9
III.II - MD algorithms	A10
III.III - How to run a MD simulation	A12
IV - QM/MM hybrid approaches: ONIOM	A13
IV.I - Defining the layers	A14
IV.II - The link atoms	A15
IV.III - Calculating the energy	A16
Annex B: Integrative Interface Examples	A19
I - Programs used on the Integrative Interface	A21
I.I - UCSF Chimera	A21
I.II - Molecular Modelling Toolkit	A21
I.III - Python	A22
II - MD example	A23
II.I - Preparing the PDB file	A23
II.II - Setting up the MD simulation	A24
II.III - Running the MD	A25
II.IV - Analyzing the results	A25
III - NMA example	A27

Annex C: Articles.....A31



General Concepts

I - Enzymes

I.1 - Introduction to catalysis

The term *catalysis* was first introduced by the Swedish chemist Jöns J. Berzelius in 1835.¹ On an annual report for the Stockholm academy of science he wrote:

“It is, then, proved that several simple or compound bodies, soluble and insoluble, have the property of exercising on other bodies an action very different from chemical affinity. By means of this action they produce, in these bodies, decompositions of their elements and different recombinations of these same elements to which they remain indifferent.”

He proposed to call this new force “*catalytic force*” and he called “*catalysis*” the decomposition of bodies by this force. Nowadays, the International Union of Pure and Applied Chemistry (IUPAC) defines catalysis as: “*The action of a substance (i.e. catalyst) that increases the rate of a reaction without modifying the overall standard Gibbs energy change in the reaction*”. This definition also introduces the term “*catalyst*”, which is the element that performs the catalysis. It can be considered both as a reactant and as a product of the reaction as it remains unaltered during the overall process. According to the IUPAC definition, the equilibrium of a reaction is not altered by the addition of a catalyst. If a given reaction is thermodynamically unfavorable, it will not become favorable and vice versa. In other words a catalyst does not affect the Gibbs term (ΔG_r) of a reaction, but rather the Gibbs activation energy (ΔG^\ddagger) by lowering the energy barrier of the transition state (Figure 1.1)

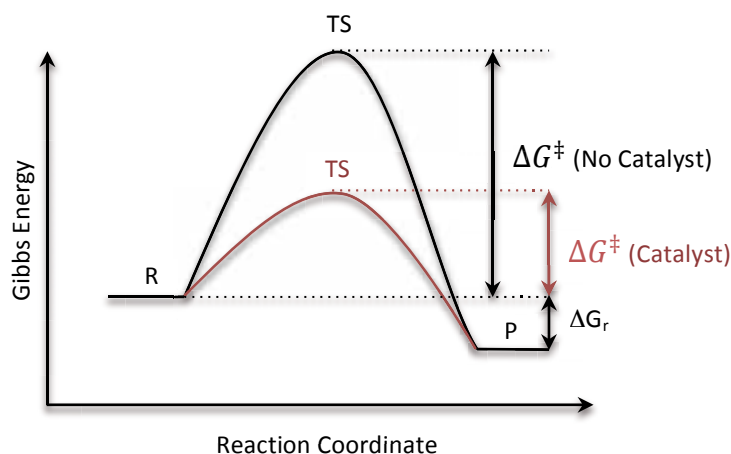


Figure 1.1 - General profile of an arbitrary reaction going from reactants (R) to products (P). Both catalyzed (red curve) and uncatalyzed (black curve) pathways are represented.

I.II - Natural enzymes

A reaction needing two hours to complete has no metabolic sense for a living cell as it needs to respond immediately to any stimulus from its surroundings in order to survive. Therefore, all reactions occurring inside it have to be catalyzed by some kind of catalyst whose activity needs to be modulated according to the cell necessities. Evolution has tailored enzymes for that purpose giving them special properties that are not common in classical chemical catalysts. They have a great specificity and display three major types of selectivity (regio-, chemo- and stereoselectivity): they will only catalyze the reactions they are specifically designed for without giving any unnecessary byproduct, which could be not only a waste of vital resources but also harmful.² Altogether, these features are the reason why many fields, such as pharmacology or organic chemistry, have set their eyes on them.

Enzymes can bind one or more substrate molecules in a region known as the active site. Some residues of this area provide with the physicochemical complementary necessary to guide the binding while others participate in the catalytic mechanism. Hypothesis on how those events take place were soon postulated in 1894 by Emil Fischer. In his model, known as the *lock and key*³ theory (Figure 1.2), an enzyme binds its substrate the same way a key can enter its lock. This theory could explain the specificity of the enzymes but not its catalytic features as no structural changes are originated neither on the enzyme nor on the substrate. It was in 1958 when Daniel Koshland refined Emil's idea and proposed the *induced fit* theory (Figure 1.2).⁴ He proposed that the binding of the substrate is inducing conformational changes in the protein. These will eventually lead to a conformational change in the active site, inducing the substrate to adopt a similar structure to that of the transition state, thus lowering the energy barrier and favoring the reaction. The latest theory was postulated in 1965 by Jean-Pierre Changeaux and coworkers; the *conformational selection*.⁵ This theory is based on the idea that enzymes are not actually static structures but an ensemble of different populated conformations in equilibrium. The substrate binds one of those conformations, thus selecting the active form of the enzyme. It is important to notice that this specific conformation does not have to be the most populated (Figure 1.2). Nowadays, the topic of how enzymes bind their substrates is still a matter of active debate,⁶⁻⁸ although there seems to be an increasing preference for the conformational selection hypothesis.^{8,9}

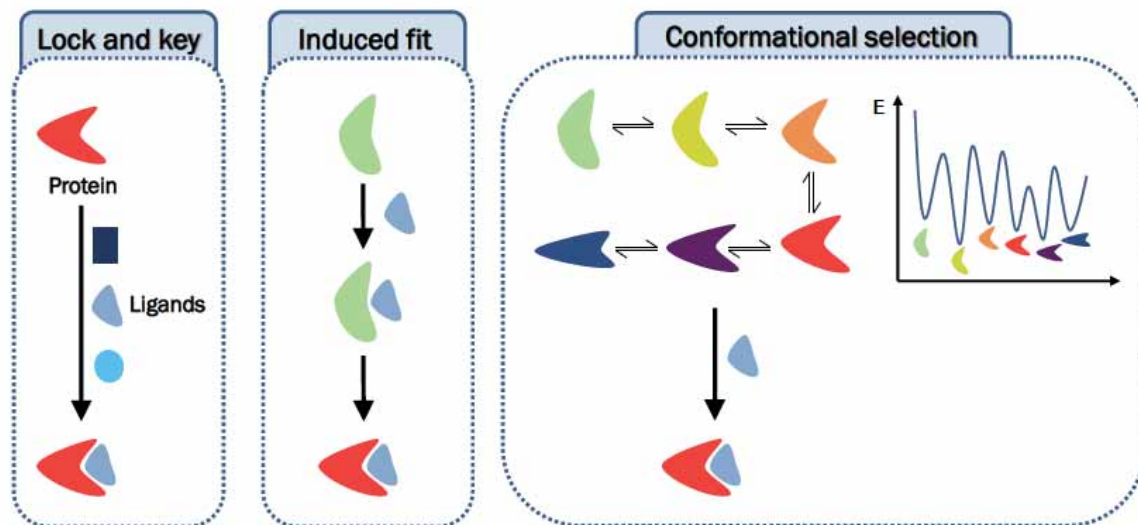


Figure 1.2 - Schematic representation of the three main theories explaining enzymatic activity.

But enzymes are not just providing an environment complementary to the transition state, in many cases they can directly participate in the reaction. Some residues with basic or acid properties are strategically positioned in the active site to have a major role in the catalytic mechanism. However, not every reaction can be performed with the limited set of functional groups that are available in the natural amino acids. In those cases the enzyme needs the aid of external molecules, non-protein chemical compounds known as *cofactors*. They bind near the active site of the enzyme and aid in the catalytic mechanism. Cofactors could be small metallic ions or complex organic or inorganic compounds. As in the catalysts case, cofactors are not irreversibly changed during the catalysis: at the end of the reaction they are regenerated so they can perform another round. In some cases there are metabolic processes associated to the enzyme reaction just to regenerate the cofactor (i.e. the case of the NADH/NAD⁺). A special mention is needed for those proteins that use cofactors containing a metal center. This family of enzymes is known as *metalloenzymes* and will be further explained in section III.II of this chapter.²

I.III - Biocatalysis

Using biological catalysts to perform chemical transformations on organic compounds is known as *biocatalysis*. This process can be done either by isolated enzymes or by living cells in culture.

The initial applications of biocatalysis could be considered more an art than a technology. Since ancient times, mankind has used biocatalytic process to craft many different alimentary products, like beer, wine or cheese. It has not been until the 19th century when

biocatalysis was first used as a new technology developed from scientific knowledge in the hydrolysis of starch using the enzyme Diastase.¹⁰ Since then, it has emerged as a potential alternative in many chemical processes as biocatalysts have some properties difficult to find in traditional chemical catalysts. First, they can display great substrate specificity and selectivity (including chemo-, regio- and stereoselectivity), something very important in many industrial processes seeking the synthesis of enantiopure products. Secondly, biocatalysts work under mild conditions, which is very demanded nowadays with many industries are searching for more environmental-friendly processes. But this does not imply in any case that biocatalysis is only restricted to polar molecules. In fact, recent advances have also achieved biocatalytic transformations of non-polar molecules under organic solvents. This could imply a huge expansion in the use of biocatalysts in many industrial processes regarding the synthesis of enantiopure organic products.¹¹

I.IV - Man tailored enzymes

Given the wide range of benefits that enzymes can offer in many chemical processes, it is not surprising that many fields, like organic and fine chemistry, have set their eyes on them. However, although there is a large and extended natural pool of different enzymes, they all have been designed by evolution to efficiently work under certain physiological conditions and on a narrow scope of substrates. For this reason, suitable use of enzymes as biocatalysts often requires further tailoring or redesign of the enzyme itself to improve their stability and specificity, and in some cases even provide them with new activities.¹² But tricking enzymes to behave as one desires is far from trivial and requires large amount of efforts.

Two different approaches have been used in the last decades for the tailoring process of enzymes: *rational design* or *directed evolution*.¹³ The former introduces changes in the system on structural based criteria in order to obtain a predicted effect on the activity. The later is a two-steps process where no structural knowledge of the system is needed. On the first step a random mutagenesis approach is used to obtain several different mutants of the enzyme. On the second step those mutants will go through a selective pressure (the desired reaction) and only the most effective ones will be selected. These mutants with improved activity can be used as the starting point of a new iteration in the directed evolution scheme for further optimization.

But it is no gold everything that shines and both rational design and directed evolution have their own good and bad points. Rational design is much straight forward than directed evolution, but a deep knowledge of the structural insights of the system is

needed and this is not always available. Moreover, the outcome of the applied modification can have an undesired effect due to the intrinsic complexity of proteins, making a prediction of the final activity a particularly difficult task. Meanwhile, directed evolution does not need any structural information, but the chances to find an improved catalyst will strongly depend on the efforts invested on the mutagenesis and screening processes. The libraries generated with all the obtained mutants could be significantly large and the screening could take an important amount of time and resources, making this step the bottleneck of the process.¹³

By combining both directed evolution and rational design the obtained libraries can be significantly shortened. Statistically speaking, the most relevant residues for the catalytic mechanism represent only the 10% of the protein sequence and can be located in a range of 10Å from the active site.^{14,15} With the sufficient structural knowledge, the random mutagenesis procedure can be confined to only these positions, thus dramatically decreasing the size of the libraries generated and the corresponding time needed for their screening. Moreover, the introduction of this probabilistic element allows the generation of enzymes with multiple synergetic mutations, something really hard to obtain through a simple rational design scheme.¹³ This new approach is referred as *semi-rational design* and was first used by Reetz and coworkers to successfully improve the enantioselectivity of an epoxide hydrolase.¹⁶ However, this approach also present several limitations.¹⁷ First, predicting which residues to target may not be a simple task since *in silico* algorithms are still at an early stage of development. Second, targeting the residues that are near the active site can exclude from the selection some important ones that, despite not being close to it, can still have an important role in the catalysis. Finally, this approach is not suitable if the aim of the study is to improve the stability or the solvent tolerance of the enzyme, since this would imply far more residues than only those adjacent to the binding site. Therefore, an approach involving the whole protein would be much more suitable in those cases.¹²

I.V - Artificial Enzymes

Far more ambitious than just tailoring natural enzymes is the idea to create from scratch new proteins with a predefined catalytic activity and enantioselective pattern. This *de novo* design would challenge all the available knowledge on proteins and could be considered the “*Holy Grail*” of enzymatic engineering.¹² First attempts tried to *in silico* generate a *de novo* scaffold displaying a pre-defined activity.¹⁸ However, these studies are far too ambitious as predicting the folding of a protein just from the aminoacid sequence is still unaffordable with current state-of-the-art computational techniques. In fact, for just

a 100 residues length protein there can be 20^{100} possible sequences, thus *in silico* screening which should be the correct folding pattern becomes an almost impossible task.^{19,20}

Instead of attempting to create new enzymes from scratch, some more feasible approaches have been reported to obtain artificial enzymes with different degrees of success. One of them was first reported by Peter G. Schultz in 1986 and consisted in generating antibodies against a transition state analog of a given reaction.²¹ If successful, the resulting antibodies presented certain complementarity with this transition state structure and could aid in the catalysis. These antibodies with catalytic properties were denominated Abzymes. However, studies on those artificial systems have demonstrated that just accommodating the transition state is not enough to efficiently catalyze the reaction. The created active site could not account for all the fine electronic effects present on the real transition state. For this reason, abzymes usually need several optimization rounds, which is a rather complicated task. They are produced by mammal cells (usually mice), which are very delicate and grow slow in comparison with other cell types. Altogether, these difficulties ended any option for abzymes to have a real applicability in the industry.

Another scheme to obtain artificial enzymes consists in the insertion of an homogeneous catalyst inside an existing protein scaffold. This way, the receptor provides a chiral protein environment and protects the organometallic compound, which confers the reactivity.²² This methodology will be extensively reviewed in section IV of this chapter.

Finally, the last attempt (and the most ambitious one) to create artificial enzymes has been extensively used by the group of David Baker and the group of Stephen L. Mayo, with particularly successful results.²³⁻³⁰ The basic idea is to find an already existing scaffold that could stabilize the transition state of a given reaction. However, each group has optimized their own way to search and optimize this putative protein. Both strategies will be explained in detail in section V.II of this chapter.

The *de novo* design of artificial enzymes opens the door to a world full of possibilities where all desired chemical reactions could be performed with the same efficiency as natural enzymes. However, our understanding on the structure-function relation is still not sufficient to compute for all the different variables that can come up during a full rational design. Deeper knowledge on this relationship and the optimization of already

existing algorithms or the creation of new ones will be vital for these approaches to be successful.¹²

II - Catalysis with transition metals

II.I - General properties and basic concepts

The term *transition metal* or *transition element* was first used in its modern electronic sense by the English chemist Charles Bury in 1921.³¹ He postulated that except for the helium, the most external shell (n -shell) of a noble gas always contained 8 electrons. Starting with period 4, once this shell has become part of the atomic core of the following period, the number of electrons it could allocate increases from 8 to 18 in the $n-1$ shell and from 18 to 32 in the $n-2$ shell. Bury referred to these groups as the 8-18 and 18-32 transition series and used them to rationalize the corresponding electronic structures. The term "*transition*" was used to indicate that these atoms were transient in their $n-1$ and $n-2$ shell layer occupancy from 8 to 18 in the former and from 18 to 32 in the later. In 1932 Lewis referred to the Bury's series as the "*transition elements*", which later became the current d -block and f -block in the periodic table.

Transition metals are often treated separately from the main groups of the periodic table (groups 1 to 2 and 12 to 18). They have unique chemistry properties mostly due to the presence of the d -electrons in their valence shell. Some of that unique features are their broad variety of stable oxidation states, the existence of paramagnetic states, the variable coordination number and, in some of them, their coloration.³²

Nowadays the IUPAC define as transition metals all those elements whose atom has an incomplete d sub-shell, or which can give rise to cations with an incomplete d sub-shell (Figure 1.3). This definition encompasses elements from group 3 to 11 belonging to periods 4 to 7. Elements from group 12 in the periodic table (Zn, Cd and Hg) are traditionally considered as transition metals because of their position on the d -block. However, these elements have a complete $(n-1)d^{10}$ shell in both their neutral and in higher oxidation states^a. Their d -orbitals do not have any influence in their chemical behavior so technically they should not be considered as transition metals.

^aThese elements are found in both I and II oxidation states, being the latter the most common. Therefore, the oxidized electrons come from the valence ns shell, and not from the complete $(n-1)d$ shell.

1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	57-70 La	71 Hf	72 Ta	73 W	74 Re	75 Os	76 Ir	77 Pt	78 Au	79 Hg	80 Tl	81 Pb	82 Bi	83 Po	84 At	85 Rn
86 Fr	87 Ra	88-101 Ac	102 Rf	103 Db	104 Sg	105 Bh	106 Hs	107 Mt	108 Uun	109 Uuu	110 Uub	111 Uut					

Transition Metals
 Key elements of life
 Non-metal trace elements
 Metal trace-elements
★ Observed metals in enzyme active sites

* Lanthanides	57 Ce	58 Pr	59 Nd	60 Pm	61 Sm	62 Eu	63 Gd	64 Tb	65 Dy	66 Ho	67 Er	68 Tm	69 Yb	70 Lu
^y Actinides	88 Th	89 Pa	90 U	91 Np	92 Pu	93 Am	94 Cm	95 Bk	96 Cf	97 Es	98 Fm	99 Md	100 No	101 Lr

Figure 1.3 - Periodic Table of the elements. Main elements found in living organisms are highlighted depending on its rarity (key elements or trace elements) and its type (metal and non-metal). Those metal elements that can be found at the active site of some enzymes are marked with a star.

II.II - Organometallic chemistry of transition metals

Life as we know is based on carbon. This element has some unique features that make carbon chemistry the chemistry of life. One of the most important ones is their ability to form stable bonds while retaining some reactivity. This is important in all life-supported process (i.e. all metabolism cycles) as carbon-based molecules are not only stable but can also be rapidly and readily transformed into other metabolites. Furthermore, carbon is one of the few elements able to bind itself forming arbitrarily large chains or polymers such as nucleic acids or proteins.³³ Because of that preeminent paper in life, chemistry has been traditionally separated into two different areas; organic chemistry that accounts for the carbon chemistry, and inorganic chemistry encompassing the rest of the periodic table.

However, as it usually happens in science, the boundaries between organic and inorganic chemistry are usually mixed. One clear example are carbon-based molecules containing a metal center. Those compounds are known as organometallic compounds and are characterized by a direct bond between a metal and a carbon.³² That bond does not have a covalent behavior, neither an ionic one due to their middle-of-the-road electronegativity. Additionally, the higher *d* orbitals of the metal atom could alter the electronic properties of the ligand by back donation. Those features allow organometallic

compounds to be relatively stable in water solution but at the same time remain ionic enough to undergo reactions. Ligands can be activated by an electronic alteration of the metal and σ and π bonds can be made, weakened or broken in between the same or different ligands. This rich pattern of different activities typical from transition metals combined with the broad scope that organic chemistry has is one of the main assets of organometallic chemistry.³²

III - Proteins and metals

III.I - Life and metals: a love and hate relationship

Almost one third of known elements are transition metals and each one of them has its own special properties. Life has taken advantage of this and have included them in many biological processes (Figure 1.3 and Table 1.1).² They can stabilize macromolecular complexes, participate in cross-linking processes and induce important conformational changes upon binding that can enhance or inhibit the enzyme activity (Table 1.1). But the most important features are the catalytic properties conferred by the metal. Metal containing enzymes (*metalloenzymes*) are able to perform some of the most difficult and complicated reactions in the natural world, including C-H and C-C bond activation or O₂ fixation for its transport (Table 1.1). But the use of transition metals in biological process is a double-edged sword as they can be very toxic. What will define the threshold concentration at which they become harmful are their own chemical properties, such as their redox potential, their acid/base properties and their ligand coordination chemistry. For example, copper is the third most abundant transition metal found in living organism, at a concentration of about 2 p.p.m. However, above that concentration copper becomes extremely toxic as it starts binding indiscriminately to all available thiol moieties (e.g. cysteine residues) and producing hydroxyl radicals through Fenton-type reactions, which can severely damage the cell.³⁴

Metal	Enzyme	Function of the metal
Fe	Cytochrome oxidase	Oxidation-Reduction
Cu	Ascorbic acid oxidase	Oxidation-Reduction
Zn	Alcohol dehydrogenase	Helps bind NAD ⁺
Mn	Histidine ammonia-lyase	Aids in catalysis by electron withdrawal
Co	Glutamate mutase	Co is part of cobalamin coenzyme
Ni	Urease	Catalytic site
Mo	Xanthine oxidase	Oxidation-Reduction
Fe	Cytochrome P450	C-H bond activation
Fe	Enolate reductase	C=C bond activation
V	Nitrate reductase	Oxidation-Reduction

Table 1.1 - Some examples of the different roles metals can bestow on metalloenzymes.

III.II - Catalytic metalloproteins: Metalloenzymes

Amongst the very wide family of natural enzymes, metalloenzymes are those that need the presence of a metal ion, whether alone or bound to an organic group, in order to maintain their functionality. Metalloenzymes are able to catalyze some of the most complex and important processes in nature, such as photosynthesis and water oxidation, thus being essential in many biological processes. The presence of the metal ion allows it to perform many essential reactions, such as oxydo-reduction, electron transfer or C-H bond activation (Table 1.1). Many of those reactions would become impossible to perform without the presence of the metal due to the limited set of functional groups found in natural amino acids.² The unique properties conferred by the metal and the selective features from its proteinic nature converts metalloenzymes in powerful tools, with many applications in many chemistry fields.

One particularly extended group of metalloenzymes are the hemoenzymes, a group of enzymes which gain their activity from a heme group present in their binding site (Figure 1.4). They are present in many major metalloenzymes families like catalases, cytochromes C, peroxidases and cytochromes P450. In fact, approximately the 75% of the enzymes involved in the metabolization of drugs belongs to this latest family.² A typical cytochrome P450 catalytic cycle is depicted in Scheme 1.1.

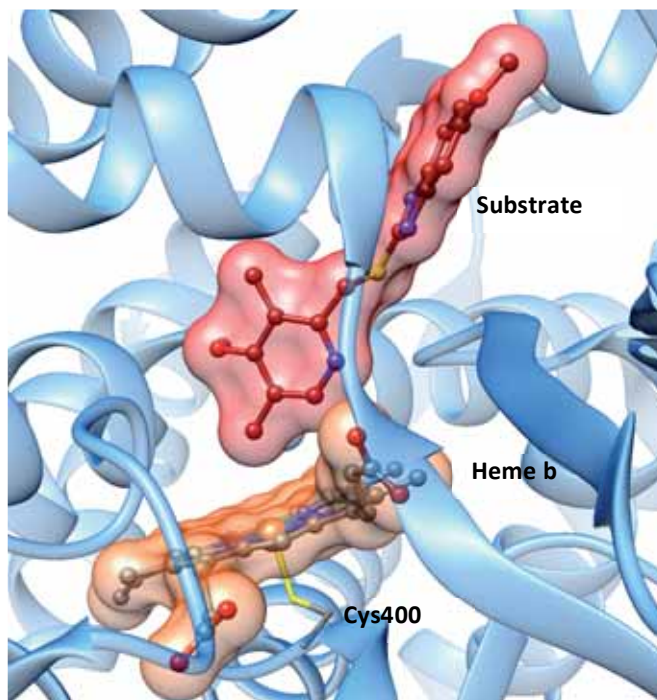
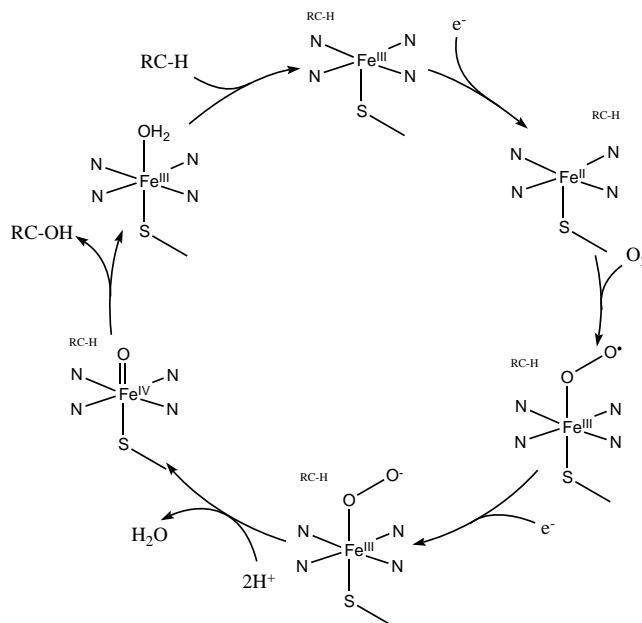


Figure 1.4 - Schematic representation of the binding site of a cytochrome P450 (PDB code: 4KEY).³⁵ The substrate (red surface) is placed next to the heme b group (orange surface). The iron atom is chelated by the Cys400 residue.



Scheme 1.1 - Representation of the cytochrome P450 catalytic cycle. The substrate (RC-H) is placed next to the Fe center of the heme b group. Next the iron is oxidized by an external cytochrome reductase (from Fe(III) to Fe(II)) and binds an oxygen molecule to later form the highly reactive oxo-Fe(IV) complex. Now, the oxygen can be transferred to the substrate and the Fe binds a water molecule to return to its resting state ready to start a new cycle.

IV - Artificial Metalloenzymes

IV.I - Principles of artificial enzymes and metalloenzymes

Up to date, homogeneous, heterogeneous and enzymatic catalysis have been the main actors in the search for enantiopure catalyzed reactions. Many efforts have been dedicated to the optimization of these species aiming for a better rate and enantioselectivity. In this context, artificial metalloenzymes have appeared as a promising alternative approach merging two different worlds; the intrinsic selectivity of the enzymes and the rich chemistry of the organometallic catalysts.³⁶ Indeed, these entities are based on the incorporation of a metal or a metal-containing cofactor inside a macromolecular receptor (e.g. protein scaffold or DNA strand). The metal bestows the reactivity on the system while the scaffold provides a chiral environment by favoring the binding of the precursor to only one pro-chiral face, protects the organometallic cofactor (if any) and stabilizes the transition state of the reaction.

Ideally, we should be able to create artificial metalloenzymes from scratch: using the 20 natural amino acids (for protein-based) or the 4 nucleobases (for DNA-based) a new receptor is created with a well-defined tertiary structure able to chelate metals or allocate metallic cofactors at its binding site. Using this approach some results have already been reported by inserting heme-group like catalysts on synthetically created α -helix.³⁷ However, this scheme represents an extremely complex task as it implies foreseeing the complementarity between exceptionally broad biological and chemical spaces. In this context, computational tools have demonstrated to be a preeminent ally. However, despite the significant advances on the *in silico* design of new artificial enzymes recently reported by the group of David Baker^{28,29,38} and by the group of Stephen L. Mayo,³⁹ our understanding of protein folding and the accuracy of current state-of-the-art computational methods is still far from sufficient to allow this kind of *de novo* design.²⁴

Due to the difficulty of creating *de novo* macromolecular receptors, the major part of the reported artificial metalloenzymes consists in the insertion of an homogeneous catalyst inside an already existing scaffold. One of the first attempts using this approach was reported by Kaiser and Whitesides in the late 1970s,⁴⁰ but it was rather misunderstood at that time due to the popular misbelief that organometallic moieties and proteins were incompatible. Nowadays, several studies have demonstrated that the creation of artificial metalloenzymes using this approach can yield to promising enantioselective entities, becoming a growing field of investigation with a tremendous repercussion.⁴¹⁻⁴⁸

First thing to take into account when selecting the appropriate scaffold to shelter the homogeneous catalyst is its chemical properties - overall charge, the pH and temperature stability or tolerance to organic solvents - as they can have a strong influence in the desired reaction. Another consideration is whether to choose a nucleotide or a protein scaffold. Both of them have been extensively used obtaining particularly successful results.⁴¹⁻⁵¹ Interestingly, the catalytic scope of the reactions performed with them are mainly complementary.⁵² In fact, the choice of the scaffold will depend on the transition metal catalyst and the desired activity. For example, DNA based scaffolds are not suitable for reactions implying catalytic oxidation, as they tend to undergo oxidative DNA strand scission.³⁶

IV.I.I - Protein based artificial metalloenzymes

Proteins are the most used scaffolds to design artificial metalloenzymes using an organometallic cofactor. Ideally, they should have a well-defined cavity with enough space to accommodate both the catalyst and the reactants and they should sustain multiple mutations on the active site without losing stability to allow further optimization.⁵² Three different approaches can be used to insert the inorganic catalyst inside the protein: (i) the dative, (ii) the covalent and (iii) the supramolecular anchorings.⁵³

Dative anchoring

The dative anchoring approach was first used by Kaiser and coworkers to obtain an artificial metalloenzyme by substituting the zinc atom in carboxypeptidase A by a copper, which conferred the system with new oxidase activity.⁵⁴ This approach aims at creating a chelation bond between the metal and some protein residues to ensure the localization of the inorganic moiety in the protein cavity.⁵⁵ Further interactions between the organometallic compound and the protein environment improve the stability of the structure (Figure 1.5). Additionally, using solid-phase synthesis or genetic encoding using the amber codon in *E. Coli* non-proteogenic amino acids can be introduced in the protein to bind the metal with higher efficiency.⁵⁶⁻⁵⁸

Covalent anchoring

Another anchoring strategy consists in covalently binding the cofactor to the protein active site. An accessible and reactive residue of the protein (normally a serine or a cysteine) should be chosen to covalently bind the inorganic moiety (Figure 1.6). This approach can be used to incorporate the catalyst into the protein scaffold when minimal structural information about the protein is known as the binding is guided by the covalent

attachment. Many different strategies for covalently binding inorganic moieties into protein scaffolds have been reported,^{59,60} although many of them resulted in an inactive hybrid.^{61,62} In order to improve the binding and the catalytic properties of the complex more than one covalent bond can be used to attach the organometallic compound.⁶⁰ However, a note of caution is needed in this approach. Nonetheless, the covalent union between the protein and the inorganic compound requires certain chemical modifications that might end up negatively affecting the cofactor itself. Moreover, this additional step of the covalent linkage in the design scheme might be a burden in the optimization of the biometallic complex.

Supramolecular anchoring

The supramolecular anchoring approach, also known as trojan horse strategy, takes advantage of strong non-covalent interactions between a protein receptor and its natural ligands (anchors). The homogeneous catalyst is covalently linked to this ligand, which will guide the binding of the whole compound (Figure 1.7). Whitesides and coworkers were the first to introduce this assembly strategy using the avidin-biotin system.⁴⁰ This duo displays the highest known non-covalent affinity in the biological world ($K_M \approx 10^{-13}$ M), so by linking the inorganic moiety to the biotin they could place it inside the protein active site. Since then, this technology has been one of the most fruitful strategies used in the design of artificial metalloenzymes.^{43-45,63,64}

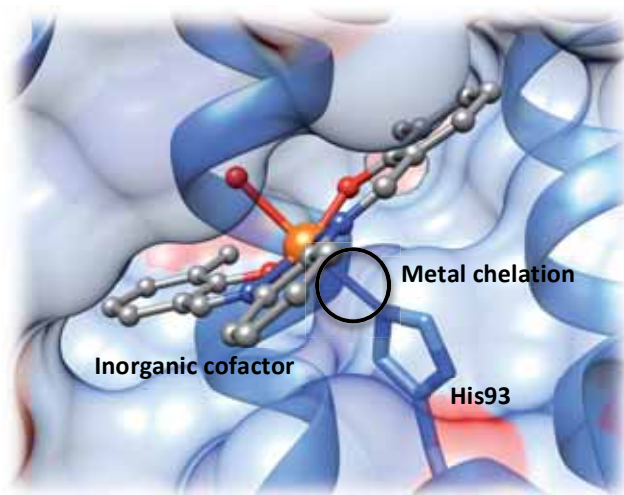


Figure 1.5 - X-ray structure of an artificial metalloprotein obtained by Watanabe and coworkers. The inorganic moiety is inserted in the myoglobin protein receptor using a dative anchoring strategy by coordinating the Fe center to residue His93 (PDB code: 1V9Q).⁵⁷

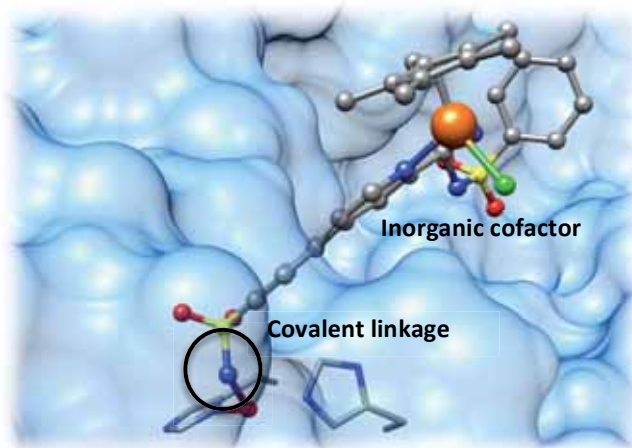


Figure 1.6 - Example of an inorganic moiety inserted into the protein scaffold by a covalent anchoring approach. To the best of our knowledge, there is no reported X-ray structure for this kind of artificial metalloenzyme. The image below is therefore a fiction model.

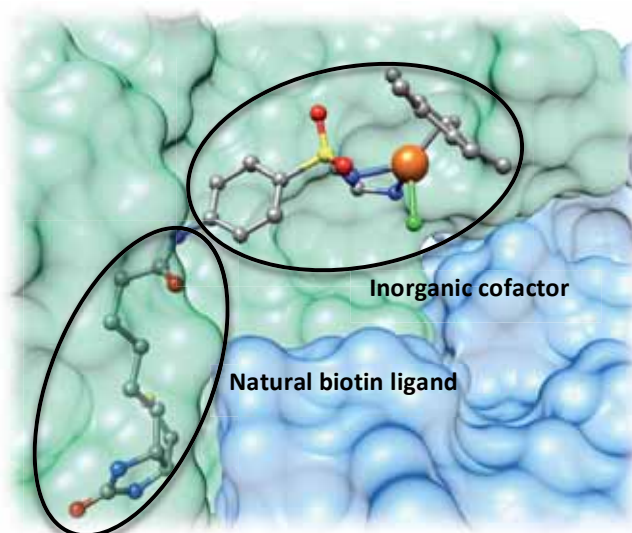
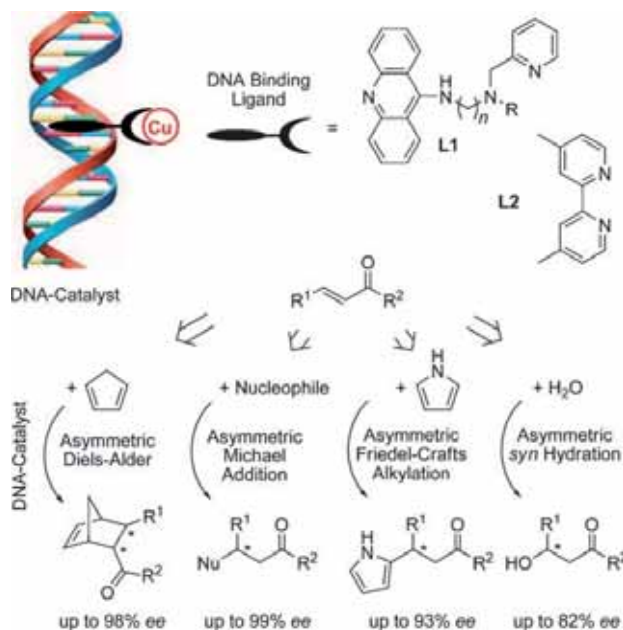


Figure 1.7 - Depiction of the artificial hydrogenase obtained by Ward and coworkers (PDB code 3PK2).⁴⁵ The homogeneous catalyst is covalently bound to a biotin molecule, the natural ligand of the streptavidin protein. In the image the two streptavidin monomers are depicted in green and blue respectively.

IV.1.II - DNA based artificial metalloenzymes

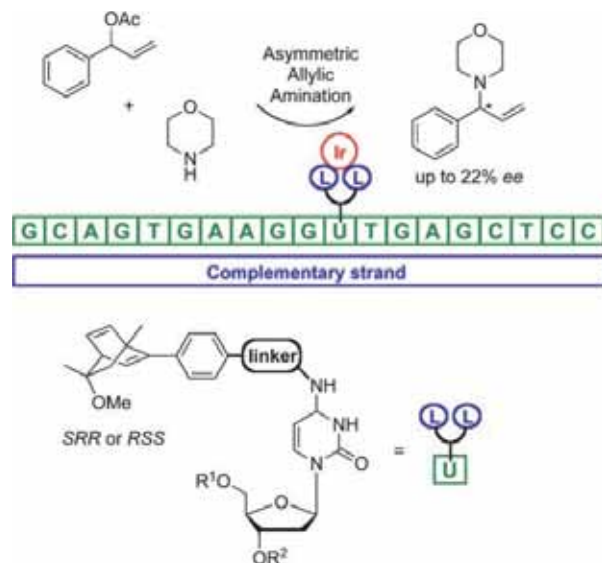
Discovered in 1869 by Friedrich Miescher, DNA is one of the cornerstones of life as it encodes all the genetic information of the cell. But DNA also presents other interesting features and many efforts are being made to see how they can be exploited for industrial applications. One of the most promising attempts was reported by Roelfes and coworkers, the first ones to use DNA to shelter a copper-based homogeneous catalyst to create an artificial metalloenzyme able to perform asymmetric Diels-Alder reaction.⁵⁰



Scheme 1.2 - Selection of different ligands used by Roelfes and coworkers to obtain artificial DNA based metalloenzymes to perform a wide range of different copper-catalyzed asymmetric reactions.^{49-51,65-69} The selectivity of the system was proved to be sequence dependent.⁶⁹ Using the **L1** ($n=2$, $R=3,5$ -dimethoxybenzene) ligand resulted in high enantioselectivity using alternating guanine-cytosine (GC) sequences while the **L2** resulted more efficient if using short guanidine (GGG) sequences.

DNA represents an ideal scaffold to embed the inorganic moieties due to its high affinity to bind small molecules through intercalations or groove binding (Scheme 1.2).⁵⁰ Additionally, the DNA sequence has a dramatic impact on the resulting *ee*, thus suggesting that the catalysis should take place near the double helix. This influence of the ATGC sequence in the catalysis opens the path to a vast landscape of possible optimization schemes than can be rapidly explored thanks to the self-assembly of the metal moiety into the DNA environment. However, several difficulties may arise during this process. The binding of the inorganic catalyst is not specific to only one of the grooves of the DNA, thus it could create numerous microenvironments and lead to a highly heterogeneous system.⁶⁹

To guide the binding event the organometallic moiety can be covalently attached to the DNA using a modified oligonucleotide (Scheme 1.3). Following this idea, Roelfes and coworkers covalently introduced a bipyridine-copper ligand in a DNA strand, obtaining 93 % *ee* for the Diels-Alder reaction between azachalcone and cyclopentadiene.⁷⁰ However, this approach does not allow a straightforward optimization due to the additional step of introducing the modified nucleobase.



Scheme 1.3 - Representation of the strategy used to covalently link an inorganic ligand to a DNA strand by A. Jäschke and coworkers.⁷¹

V - Computation for enzyme design

V.I - *In silico* methods in biocatalysis

Nature provides a large pool of different enzymes that can be used as biocatalysts but in many cases they must undergo a heavy optimization process before reaching the high catalytic standards needed for industrial process. Rational design can offer great perspectives in this process as it can significantly reduce the amount of time and resources needed in comparison with other *in vitro* approaches (e.g. directed evolution). Nonetheless, there is a vivid community constantly pushing the limits of computational methods to surpass the limitations of actual state-of-the-art techniques.⁷²

Optimization of a biological system through rational design involves using many different *Molecular Modelling* approaches. The aim of this area is to aid in the experimental part by generating an accurate model of the real system in which to perform different *in silico* experiments. These approaches can offer valuable information impossible to obtain otherwise (i. e. structural information of a transition state) and perform experiments much more efficiently than using an experimental approach (i. e. evaluating the binding of thousands of compounds).

Molecular modelling is composed of two main families: (i) *Quantum Mechanics* (QM) and (ii) *Molecular Mechanics* (MM).^b QM approaches tend to be more accurate than MM, but much more computational intensive, reason why QM is more often used to study small chemical systems whereas MM is more suitable for bigger systems (e.g. big biological complexes). In this context, main methodologies reported so far for the optimization of enzymatic systems involves *protein-ligand dockings* and *molecular dynamics* (MD) techniques.⁷³

Docking has been extensively used in the prediction of the binding of different molecules to protein receptors. These molecules could range from other proteins (protein-protein dockings) to small ligands (protein-ligand dockings). This last approach can give key molecular data of the main residues involved in the formation of the substrate-enzyme complex and aid in the characterization of the main factors dictating both the specificity and the selectivity of the protein.^{3,4} Therefore, it is natural that protein-ligand dockings are being widely used in elucidating main mutation candidates for the optimization of the active site. Using this approach Liu et. al. managed to create a variant of the *Candida Antarctica* lipase B with enhanced activity for acrylation reactions.⁷⁴

MD techniques have been also widely used to shed light on some biophysical and biochemistry features of enzymes such as their high thermostability,⁷⁵ their temperature optimum,⁷⁶ or their activity, specificity and selectivity.^{77,78} All these studies have given important structural information, enabling the design of optimization schemes. Additionally, in those cases where the enzymatic selectivity could be determined by residues far from the active site^{79,80} or by indirect, long-range effects^{25,81,82} MD becomes the only approach able to both predict and explain these phenomena.

It is important to notice that MD and docking approaches cannot be used for the study of those molecular events that need an accurate electronic representation of the system (e.g. the characterization of the transition state). In those cases we need to recall to QM approaches, but the size of the system becomes a limiting factor as the computational cost exponentially increases with it. To surpass these limitation the hybrid QM/MM scheme was created.⁸³⁻⁸⁵ This approach splits the system in two different parts: those atoms directly participating on the catalysis are treated using QM approaches while the rest of the system uses a MM representation. This significantly reduces the computational resources needed for the calculation thus enabling the QM study of big complexes,

^b Further explanation of these two families will be provided in Chapter 2 -

something otherwise impossible. The high accuracy of this QM/MM methodology can provide with a more quantitative description of the system in comparison with other computational techniques. Indeed, some promising results have already been reported for the case of a *Candida Antarctica* lipase A.⁸⁶ However, there are still some limitations regarding QM/MM schemes: (i) the size of the QM part is the limiting factor in those calculations, thus it is still high computational intensive if compared with other MM approaches and (ii) cannot explore large conformational spaces, thus the quality of the initial model has a dramatic impact on the final QM/MM prediction.

Although computational methods have proven to be a successful way to interpret experimental results and guide in the rational design/optimization of biological entities, the overall process is still mainly experimentally driven.²⁴ However, we are close to reverse this situation as improvements in computer efficiencies and molecular modelling algorithms are increasing the accuracy of the simulations while decreasing the computational cost. The works by Baker and Mayo groups, respectively, are reliable proves of that landscape shift, by *de novo* designing artificial enzymes mainly using computational approaches to guide the overall creation process.^{28,39,87}

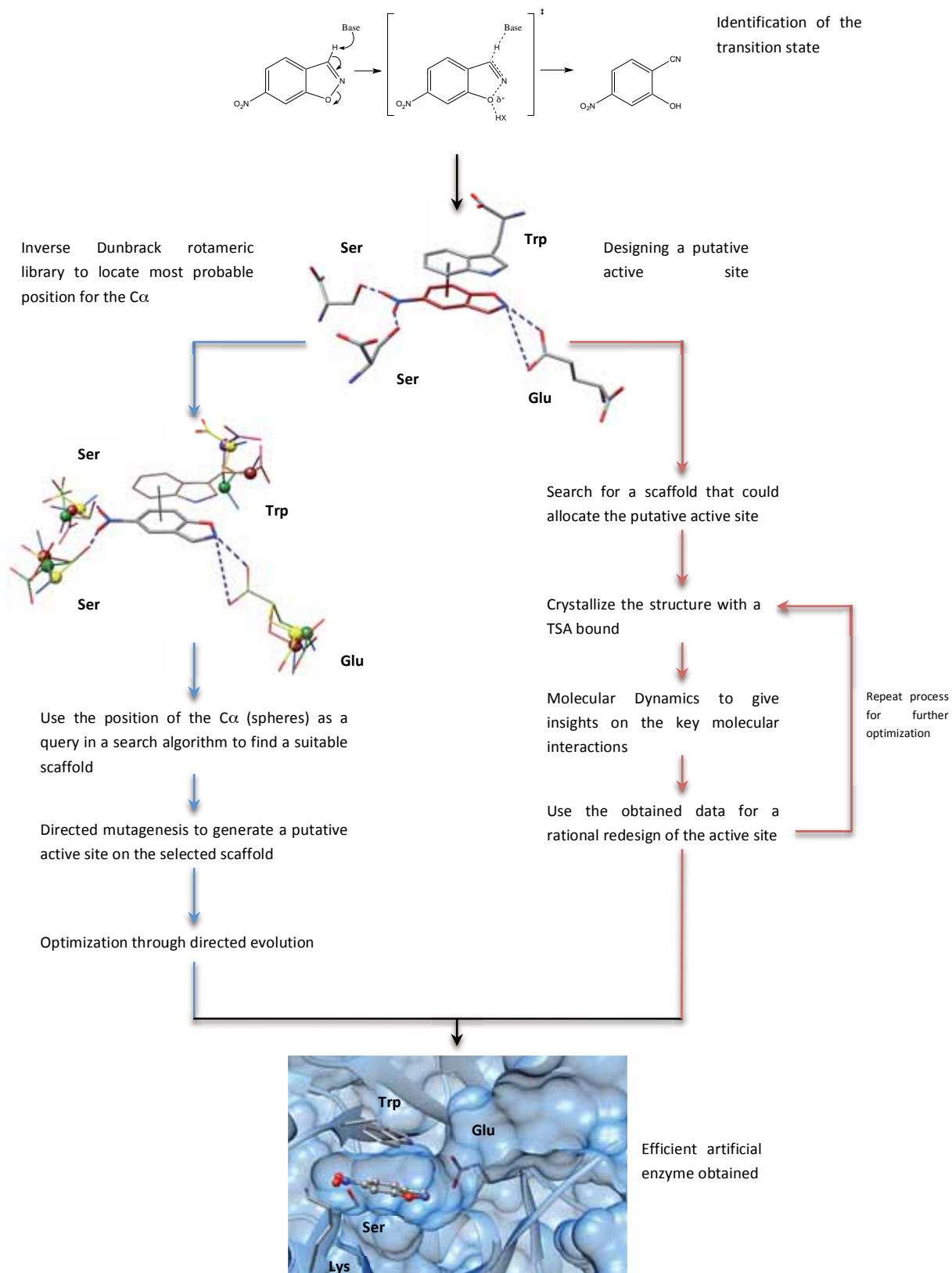
V.II - Computer aided *de novo* design

Ideally, designing a new enzyme from scratch would imply predicting the folding of a given amino acids sequence leading to a highly efficient active structure with enantioselective properties. However, dealing with the folding problem is something still far from the current the state-of-the-art computational approaches, as previously explained in section I.V of this chapter. For this reason, up to date the most successful way to *de novo* design artificial enzymes consists in optimizing the active site of an already existing protein to allocate a given transition state. This strategy has been exploited using different approaches by the group of David Baker^{28,87} and the group of Stephen L. Mayo.³⁹ Both of them share the same starting point: after the characterization of the transition state, they model an ideal binding site that could allocate and stabilize it. However, they differ in how they manage to build that active site (Scheme 1.4). On one hand, Baker and coworkers use an inverse Rotameric library^c to locate the most probable position of the C α of the residues involved in the catalysis and a search algorithm to find which protein structures available on the PDB could satisfy these restraints (Scheme 1.4). Not all the predicted

^c This library consist of the Dunbrack rotameric library,¹²⁰ but it uses the atoms interacting with the transition state to find the most probable positions the C α could have in the scaffold.

residues are necessary to be in the protein scaffold, they could be introduced using directed mutagenesis afterwards. The target system is finally optimized through several directed evolution cycles. Mayo, on the other hand, uses an iterative approach to obtain the artificial enzyme (Scheme 1.4). He also starts seeking a protein scaffold that could mimic the previously characterized transition state. Once the target is found, a round of directed mutagenesis is performed to enhance the stabilization of the transition state by the protein environment. The modified enzyme is crystallized with a transition state analog (TSA) bound and the obtained structure is then analyzed through a MD approach. The gathered data shed light on the main residues involved in the stabilization of the transition state and the impact of the previous mutations. This information is then used to perform another directed mutagenesis round, thus starting a new iteration. The overall process is repeated until obtaining an efficient artificial enzyme in terms of enantioselectivity and yield rate (Scheme 1.4).

These two approaches have highlighted the possibility to design enzymes from scratch using *in silico* driven experiments. However, they are still at an early stage and in many cases several experimental steps are still needed in the overall optimization process.



Scheme 1.4 - Scheme of the protocol used by Baker (Blue) and Mayo (Red) to successfully generate artificial enzymes.

V.III - Challenges on artificial metalloenzymes

As previously explained in section III.II, metalloenzymes can become a powerful tool in many chemistry fields due to the catalytic properties provided by the metal moiety. However, this is precisely the reason why actual approaches to *in silico* design *de novo* artificial enzymes cannot be applied. The search algorithms used by Baker or Mayo cannot take into account how the metal-containing cofactor and the protein would adapt to each other during the binding event. For this purpose, we need QM approaches for an accurate electronic representation of the metal environment, but exploring large conformational space with these methodologies is forbidden.

A promising way to create artificial metalloenzymes is through the insertion of an homogeneous catalyst inside a protein scaffold.^d In this framework, computational techniques could give insights on the partnership between the inorganic moiety and the protein receptor, the first step towards their optimization. However, this still represents a major challenge for current molecular modelling approaches. In fact, the flexibility of the cofactor and the presence of the metal represent a vast conformational space to explore during the binding analysis. Additionally, QM approaches are needed in order to accurately treat the first coordination sphere of the metal, thus exponentially increasing the computational time of the process. To overcome these limitations, some approaches combining several molecular modelling techniques have appeared aimed at surpassing the limitations of each individual method and yielding to promising results.^{72,88} Unfortunately, most of the available modelling softwares were not meant to be used as a part of a combined protocol. In fact, each of them has its own way to treat the flow of molecular data, thus several formatting issues appear and some information may be lost when trying to pass it from one program to the other. For this reason, a universal platform to perform several different molecular modelling techniques in tandem with a full compatibility between them would exponentially increase the efficiency of these integrative approaches.

^d For more information please refer to section IV of this chapter.

Computational Methods



I - Molecular Modelling

I.I - Introduction to the art of modelling

Molecular modelling encompasses all theoretical methods aimed at creating mathematical models to explain the behavior of molecules. It includes a wide panel of computational techniques that generally go from quantum chemistry and force field approaches to structural bioinformatics (Figure 2.1). In this part of the manuscript, we will give a general overview of most of the computational tools available nowadays.^a In particular, we will focus on how to deal with complex biometallic systems using integrative approaches encompassing several different molecular modelling approaches.

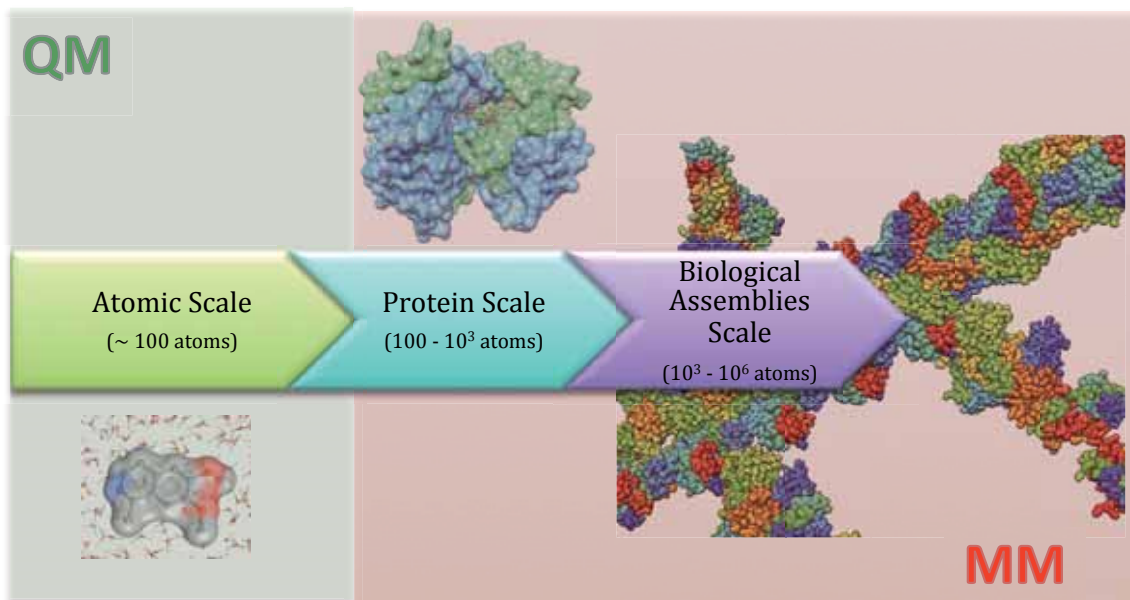


Figure 2.1 – Diagram displaying the most appropriate molecular modelling methodology to choose depending on the size of the system.

Molecular modelling tools can be mainly split in two different categories: *Molecular Mechanics* (MM) and *Quantum Mechanics* (QM). In MM simulations, atoms are treated as spheres with some physical properties (e.g. charges, Van der Waals radius, etc.) connected by springs (bonds) that follow Classical Physics laws. On the other hand, QM models are more realistic, taking into account both the nuclei and the electrons of the atoms and calculating the properties of the molecules using Quantum Physics theorems. As a result,

^a A more extensive description of each treated methodology in this chapter can be found in the Annex A.

models based on QM are much more accurate than those based on MM but also much more computational demanding. That is the reason why QM is often restricted to small models (up to hundreds of atoms) while MM is used for bigger systems (up to millions of atoms) (Figure 2.1). QM methodologies can be used in the study of catalytic reactions but cannot explore large conformational spaces due to the high computational cost. On the contrary, MM approaches allows exploration of large conformational spaces but does not allow the study of chemical transformations and are not as accurate as QM. Therefore, one must be cautious on which methodology to use considering both the particular molecular problem and the advantages/limitations of each individual technique. In fact, prior to any study we must define a threshold of accuracy in accordance with the aim of this study. There is no meaning in performing *in silico* simulations if it would take fewer resources to obtain the same information through the classic experimental procedure. If done properly, the modelling process offers a good representation of reality and should be able to explain observable data and predict new behaviors with the ultimate goal of guiding the design/optimization of molecular systems.

II - Quantum Mechanics

II.1 - Basic Concepts

Quantum mechanics is a fair complicated discipline and a deep explanation of it is far beyond the scope of this work, however a basic introduction is considered helpful for the understanding of the methodologies used in this thesis. For more detailed information, please refer to the corresponding section in the Annex A.

From a computational chemist point of view, QM is the main tool to describe a chemical system. Through the resolution of the *time-independent Schrödinger* equation, an accurate representation of the energy of the system can be obtained. But finding the exact solution with the current-state-of-the-art methodologies can only be done for a simple hydrogen atom. In fact, several approximations must be taken into account to simplify its resolution and wide its range of applicability to other chemical systems. Amongst all the efforts reported to simplify the Schrödinger equations, a special mention is needed for the *Born-Oppenheimer* approximation and for the *Hartree-Fock* equations. The first one decouples electronic and nuclear movement, considering that the electrons will automatically adapt to any displacements of the nuclei. This approximation is only

possible because of the high electrons/nuclei mass ratio.^b The second approximation is based on the idea that the N -electron wave function can be considered as the antisymmetrized^c product of N one-electron wave functions. This product is also known as the *Slater Determinant*. This determinant is then optimized through the *variational principle* to find the lowest energy solution. Main problem in the Hartree-Fock equations is that they not take into account the *electronic correlation* (e.g. electron-electron repulsion). For this reason the obtained energy of the system is always larger (less negative) than the real energy of the system. To deal with this problem, the *Post-Hartree-Fock* methods were developed, which are aimed at refining the Hartree-Fock equations by incorporating the missing electronic correlation term.^d

An alternative approach to obtain the Hamiltonian (energy) of the system is the *Density Functional Theory* (DFT). It offers one of the best *quality of the calculation/resource consuming* ratio and has been proved to be one of the most reliable approaches when dealing with transition state metals. Altogether, these reasons and the vast experience of our group in using DFT methods persuaded us to use it in all the QM and QM related calculations performed in this Ph. D. thesis.

II.II - Density Functional Theory

Since the early 1990's DFT has been one of the main methodologies used in computational chemistry to study organometallic compounds, specially those involving transition metals. It postulates that the molecular electronic features of a given system, including the ground-state energy and the wave function, can be determined solely by the *electronic probability density* (ρ). Hohenberg and Kohn demonstrated in 1964 that this approximation was true for systems with non-degenerated ground states.⁸⁹ They proposed the existence of a functional which, when applied to the electronic probability density, will return the exact energy of the fundamental state (Eq 2.1).^e

^b For the lightest nuclei, a proton (H^+), this ratio is already $\frac{m_{H^+}}{m_{e^-}} = 1836$.

^c The wave function describing the electrons behavior must be keep antisymmetric to meet with the Pauli exclusion principle.

^d For more detailed information about all these approximations to solve the Schrödinger equation, please refer to Annex A.

^e The demonstration of the Hohenberg and Kohn approximation is out of the introductory nature of this thesis and can be found in any general Quantum Mechanical manual.¹⁹⁴

$$E = E[\rho] = \min \left(F[\rho] + \int \rho(\vec{r}) V_{Ne} d\vec{r} \right) \quad \text{Eq 2.1}$$

The determination of the electron density is much simpler than the determination of the actual wave function with a Hartree-Fock or any other derived method. The reason for this is that the electron density is a function with only three spatial variables, whereas the wave function has $3N$ variables, where N is the number of electrons if using the Born-Oppenheimer approximation. Furthermore, applying this unknown functional will return the exact value of the energy of the system, including the electronic correlation (which was not included in the initial Hartree-Fock methods and is computationally very intensive in other Post-Hartree-Fock methodologies). However one problem arises in Eq 2.1; it is impossible to determine the exact value of the *universal functional* ($F[\rho]$).

The universal functional can be decomposed as follows (Eq 2.2):

$$F[\rho(\vec{r})] = T[\rho(\vec{r})] + J[\rho(\vec{r})] + E_{ncl}[\rho(\vec{r})] \quad \text{Eq 2.2}$$

While the $J[\rho(\vec{r})]$ term is known (describes the classical Coulomb interactions), the $T[\rho(\vec{r})]$ (contribution of the kinetic energy) and the $E_{ncl}[\rho(\vec{r})]$ (non-classical portion due to the self-interaction correction, exchange (i.e. antisymmetry) and electron correlation effect) remain a mystery. It was in 1965 when Walter Kohn and Lu Jeu Sham reported a way to approximate the kinetic term of the universal functional.⁹⁰ Instead of trying to find a way to accurately calculate it, they realized that it was much more efficient to compute only the known part of it (e.g. the one that can be computed exactly) and deal with the remainder in an approximate way. To do so, they introduced the concept of a non-interacting reference system built from a set of orbitals (e.g. one electron functions). The remainder is then fused with the $E_{ncl}[\rho(\vec{r})]$ term, which is also unknown but usually fairly small. Many methods have been reported to approximate this part of the universal functional using another functional. These are the *exchange-correlation functionals*, also known as *DFT functionals*.

Nowadays there are several different DFT functionals available and each one of them has their own advantage/limitations. It is therefore the user responsibility to decide which of them is the most appropriate choice to tackle a given molecular problem.

III - Classical mechanics in molecular modelling

Any *in silico* study has to either provide molecular information that cannot be obtained otherwise or predict molecular behavior using fewer resources than performing the actual experiment. For this reason we sometimes need to set up an accuracy threshold on the computational experiment, especially in those involving large systems. For example, even though QM techniques present the best accuracy level, they are too much computational intensive to perform certain analysis in a reasonable timespan. For these situations, MM approaches can be used as they dramatically decrease the level of computation needed for the simulation while still maintaining an acceptable level of accuracy.

III.I - Molecular Mechanics – Basic Concepts

MM embraces an ensemble of widely used methodologies that share the same approximation: they consider atoms as solid spheres with physical properties interacting with each other via harmonic forces and linked by springs (bonds) that are not allowed to break or form. To calculate the energy of the system MM resort to classical physics laws. Despite all these approximations to represent the system, this approximation has proven to be remarkably accurate to calculate geometrical and energetic features.⁹¹

III.I.I - Force Field

One of the most used ways to calculate the potential energy of the system in MM approaches is through the use of force fields (ff). They are a set of functions derived from some classical physics theorems (e.g. Hooke, Newton or Coulomb laws), which can either describe the bonded and the non-bonded interactions of the system (Eq 2.3).

$$U = U_{bond} + U_{non-bonded} \quad \text{Eq 2.3}$$

The bonded term is associated to those atoms that are covalently bound and can be decomposed in various terms depending on the number of bonds between them. This way, if the atoms are linked by one bond we have the bond-stretching term (i), if there are two bonds we have the bond-bending term (ii) and if there are three bonds we can have either the dihedral motion term (iii) or the out-of-plane angle potential (iv) (Figure 2.2). The non-bonded term accounts for the interactions between atoms that are not directly linked. This term is the sum of the van der Waals and the electrostatic interactions (Figure

2.2). As there can potentially be much more non-bonded than bonded interactions, the former is more computational intensive.^f

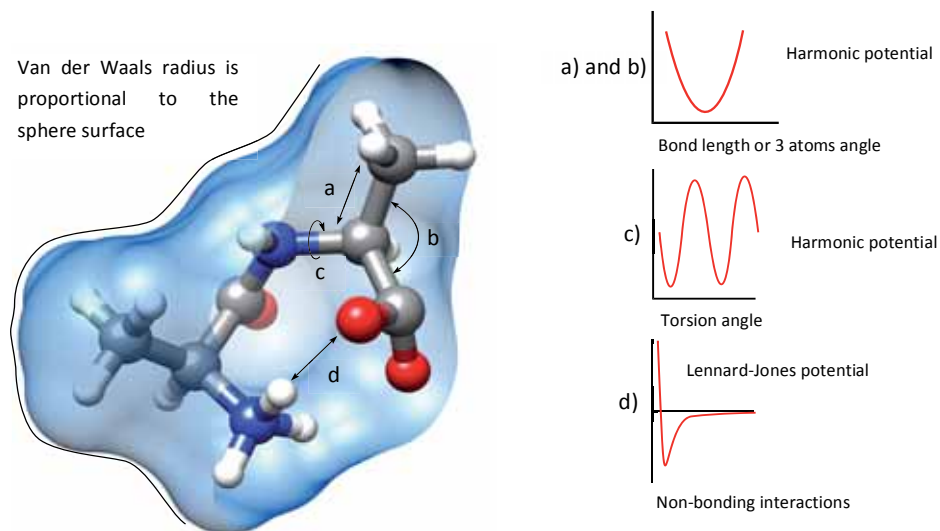


Figure 2.2 - Schematic representation of the ff terms and their associated potentials.

Each one of the ff terms takes into account several parameters (e.g. charges, van der Waals radius, bond length, etc.) associated to each atom that can either be found experimentally or using high-end QM calculations. In an ideal situation, each atom of the system should have its own parameters describing its unique features. However, as there are potentially infinite chemical systems, this would lead to an unpractical situation in which we should have to parametrize every atom prior to any simulation. To avoid this scenario, the atoms of some of the most common functional groups have been parametrized defining an *atom-type*. This way, every time we want to perform a MM calculation, we can describe the atoms of the system by assigning them the parameters of their most similar atom-type.

III.II - Variety of MM techniques

MM offers a wide variety of different methodologies that can be used for many different purposes. This section pretends to be a brief introduction of the approaches used in this work for a better understanding of the results obtained.

III.II.I - Molecular Dynamics

Molecular Dynamic (MD) simulations compute the motions of individual molecules. It computes how molecules dart to and fro, twisting, turning, colliding with one another,

^f A detailed explanation of each term of the ff can be found in the corresponding part of the Annex A.

and, perhaps, colliding with their container to describe how their positions, velocities and orientations evolve through time. These molecules are interacting under a defined potential and moving according to the Newton's equations of motion.[§] MD allows the prediction of static and dynamic properties directly from the underlying interactions between the molecules and allows obtaining mechanistic insights into experimentally observable processes.

The MD simulation is a deterministic methodology as the set of initial variables will define how the system evolves. In fact, in a MD run the future stages of the system will depend on how the previous one evolved. For example, for each MD step the coordinates and velocities of the previous step are needed to compute how the system evolves. The duration of the simulation is also an important factor. Its length must be related to the molecular event we want to study (Table 2.1). By choosing an inferior time the results will not be representative and a superior one will be just a waste of time and resources.

Molecular event	Time scale (s)
Local motions	10^{-15} - 10^{-1}
Rigid body motions	10^{-9} - 1
Large scale motions	10^{-7} - 10^4

Table 2.1 - Time scale of various molecular events

One important choice prior to launch the MD simulation is how to represent the solvent in which the system is confined (if any). There are two main methodologies for this purpose: (i) *implicit* or (ii) *explicit* solvent. In the former case the dielectric constant of the solvent is added to the electrostatic interactions term of the force field while in the later the system is surrounded by model molecules of the solvent. This last methodology is much more accurate than the implicit representation but, due to the increase in the number of non-bonded interactions to account for the ones between the system and the solvent, it is also much more computational intensive.

In the case of the explicit representation of the solvent we need to define the boundaries of the system wither defining (i) *stochastic boundary conditions* or (ii) *periodic boundary conditions*. In the stochastic boundary a radial gradient is used centering it on a certain area of interest (e.g. the active center). The solvent molecules are then placed on that

[§] A detailed explanation on how the Newton's equation of motion is used on a MD simulation is given in Annex A.

gradient, which will prevent them from escaping outside the designated region (Figure 2.3). This approach is an efficient way to study localized molecular events as the number of solvent molecules and the size of the system reduced considerably. The periodic boundary conditions consist in placing the system and the solvent on a unit cell or box. The edges of the box are the boundaries of the system, but they have special properties. If an atom goes through one of those edges, it appears at the opposite side of the box with the exact same velocity (Figure 2.3). This way, the number of atoms is maintained constant through the simulation without putting using an actual physical barrier or a gradient.

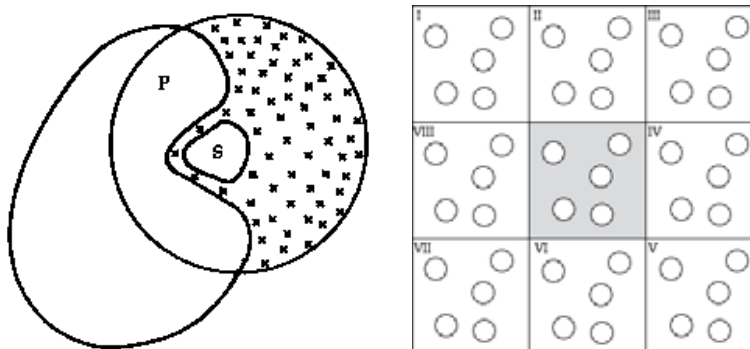


Figure 2.3 - Representation of the different ways to represent the boundaries of the system when using an explicit representation of the solvent. **Left:** Stochastic boundary conditions. **Right:** Periodic boundary conditions.

III.II.II - Normal Modes Analysis

MD is an efficient way to study the motions of biological macromolecules, but it is not the only one. Normal Mode Analysis (NMA) could also be used for that purpose as it calculates the vectors representing the motions of a protein, which are obtained from the analytical study of the harmonic potential wells. NMA is the only approach able to identify and characterize the slowest motions present in a macromolecular system, but the study is restricted to the potential energy well where the system is placed. Therefore, NMA is not well suited for the study of conformational transitions as this implies jumping through different potential wells. Fortunately, this limitation can be solved through the combination of NMA with other techniques like MD simulations.⁹²

The potential energy landscape of a protein is defined by numerous potential wells, each of them representing a particular stable conformation derived from big rearrangements of the backbone. The surface of this well is not smooth, instead it has several local minima separated by small energy barriers (Figure 2.4). Each of those local minima represents a

protein conformational substate⁹³⁻⁹⁵ derived from small rearrangements of the side chains of the protein. The starting points for the NMA is precisely the minimum of one of those potential wells. The structure of this stable conformation is obtained through X-ray crystallography and minimized to localize the global minimum of the well near this structure. Then, the harmonic approximation of the potential well is constructed around this conformation (Figure 2.4).

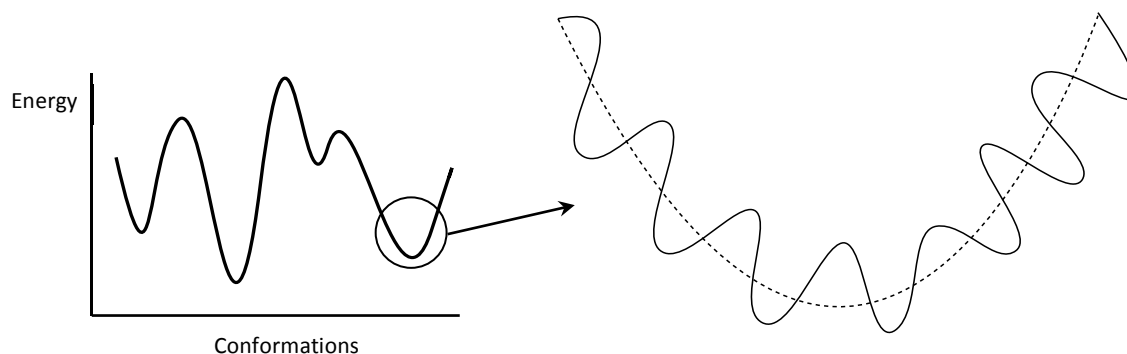


Figure 2.4 - A schematic one-dimensional view of a protein potential energy surface (left) and a close view of one of the potential wells (right). The dashed lines represent the harmonic approximation of the smooth potential well, while the plain line corresponds to the local minimum harmonic approximation defining the substates of this particular conformation.

A harmonic potential well has the form:

$$U(r) = \frac{1}{2}(r - R)K(R)(r - R) \quad \text{Eq 2.4}$$

Where R is a $3N$ -dimensional vector (N is the number of atoms) describing the structure taken as the starting point (located at the center of the well) and r is an equally $3-N$ vector describing the current conformation. K is a matrix describing the shape of the potential well, and can be described as the second derivative of the potential:

$$K_{ij} = \left[\frac{\partial^2 U}{\partial r_i \partial r_j} \right]_{r=R_{min}} \quad \text{Eq 2.5}$$

Where R_{min} corresponds to the protein structure at the minima of the potential well. Analytically solving that matrix, the natural movements of the protein (the normal modes) and its associated frequencies can be obtained. Each frequency is correlated with the energetic cost of that movement, thus the deformations of the protein can be classified according to it. The low-frequency modes correspond to the collective or delocalized

deformations, while the high-frequency are local displacements. Therefore, the low-frequency modes are the most relevant ones as they dictate the large-scale collective movements of the protein. It is important to notice that the 6 first modes are not giving any relevant information, as they are associated to the six rigid-body movements of the protein (translation and rotation).

The harmonic model obtained through NMA is an approximation to a conformational substate and its valid for very small motions around the local minima. However, the use of NMA has been extensively applied for larger amplitude motions, like the opening/closing motions of the active site of an enzyme. Although this usage is beyond the theoretical limit of applicability for a conformational substate model it can be justified empirically: NMA usually yields results that are in very good agreement with experimental observations. In fact, several studies have demonstrated that it is possible to study biomolecular functions by filtering the high-frequency modes and using the low-frequency ones (Figure 2.5).^{96,97} NMA is reflexing the intrinsic flexibility present in biological complexes, which probably evolved to take advantage of this to develop their biological function.

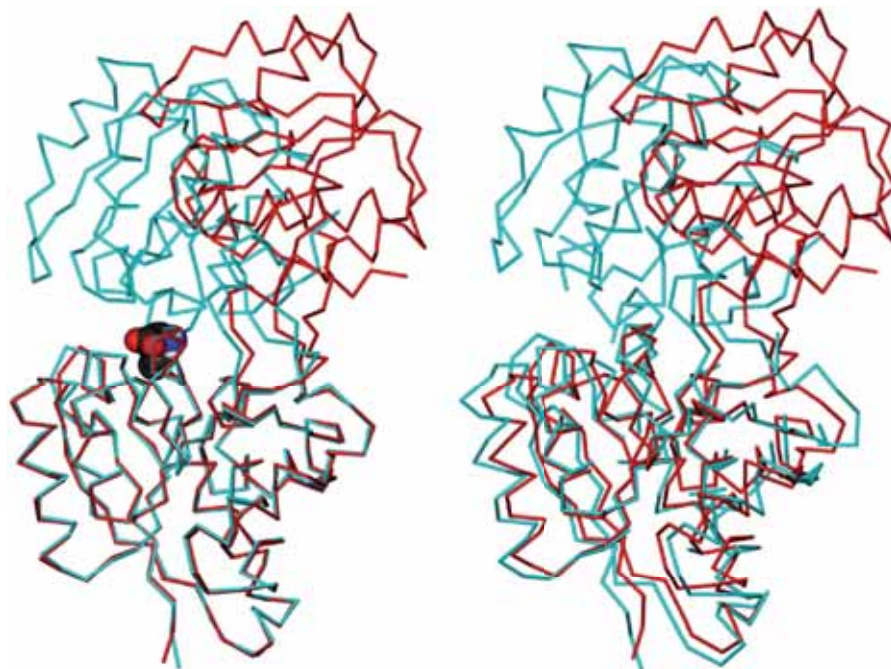


Figure 2.5 - Schematic representation of the harmonic modal displacement for LivJ protein of Escherichia Coli.⁹⁸ On the **left** side, the X-ray structure in open (red) and close (light blue) conformations are superimposed on the N-terminal globular domain. The movement from the open to the close conformation is originated by the binding of the isoleucine ligand in the cleft between the two domains. On the **right** side, the same X-ray open structure (red) is superimposed to the generated close model (light blue) obtained by walking along the lowest-frequency mode.

III.II.III - Protein-ligand Docking

Protein-ligand docking (hereafter referred as dockings) aims to predict the posing^h of a small ligand within a protein scaffold. This way, we can obtain a model of the substrate-enzyme complex structure, which can be used for the prediction of the biological activity of the system (Eq 2.6 and Figure 2.6).⁹⁹ The binding affinity between these two species is related to the free energy (ΔG) associated to that process (Eq 2.7 and Eq 2.8).

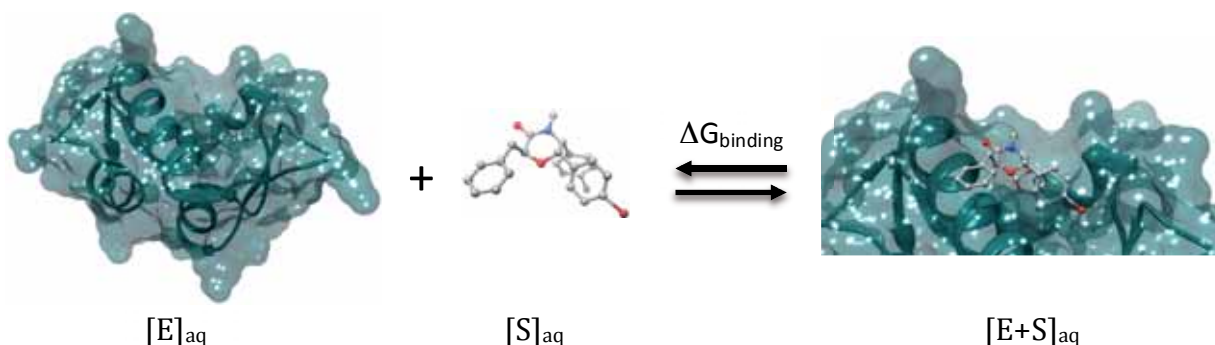


Figure 2.6 - Illustration of the protein MDM2 in complex with its inhibitor (2*S*,5*R*,6*S*)-2-benzyl-5,6-bis(4-bromophenyl)-4-methylmorpholin-3-one (PDB code 4JV7).¹⁰⁰

$$\Delta G = -RT \ln K_A \quad \text{Eq 2.7}$$

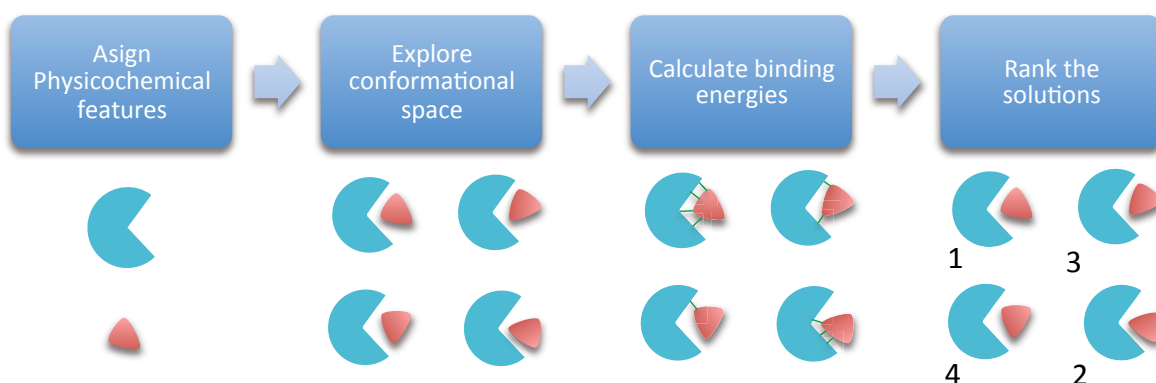
$$K_A = \frac{[ES]}{[E][S]} \quad \text{Eq 2.8}$$

A general docking simulation needs two different algorithms: (i) the *search* and (ii) the *ranking* algorithms.⁹⁹ The first one aims to predict the correct [E+S] structure. This is a rather complex task even for small organic molecules due to the high degrees of freedom it possesses. This algorithm must be fast to explore all the available conformational space while retaining enough accuracy to find the conformations that best fits into the receptor. The ranking algorithm evaluates the quality of the poses (ΔG binding, Eq 2.7) coming from the search algorithm by evaluating several biochemical properties of the [ES] complex (steric effects, electrostatic interactions, hydrogen bonds, ligand and enzyme strains, etc.). All this information is then passed onto a *Scoring Function*, which will rank the solutions

^h The posing is the process to determine whether a certain conformation and orientation of a ligand can fit into the binding site of a protein.

assigning a score to each one of them. Nowadays, there are several different scoring functions available, and in general each docking program has its own custom ones.

A docking simulation can be divided into four different processes (Scheme 2.1): (i) identify the physicochemical features of both the ligand and the receptor, (ii) do a sampling through all the conformational space available in the binding site and generate the ligand-receptor complex, (iii) calculate the binding energies between the ligand and the receptor and (iv) rank the solutions using a scoring function. Step ii) is done by the search algorithm, and steps iii) and iv) by the ranking algorithm. In most dockings programs steps (ii) and (iii) are done simultaneously: the interaction energy is calculated while the structures are being generated to guide the search towards a better conformational minima.



Scheme 2.1 – Schematic representation of the standard procedure of a docking simulation

During each docking run a vast number of different poses is generated. Even if we are using MM approaches, using standard ff to evaluate all of them could be particularly computational intensive. For this reason, scoring functions are one of the main assets of docking simulations. They use a much more simplified scheme to evaluate the energy of the complex, thus dramatically decreasing the computational burden. It is not surprising therefore that docking techniques are the preferred choice in *High Throughput Screening* studies, in which thousands of chemical compounds are tested on thousands of target proteins.

In the last years, docking approaches have achieved several improvements in both search algorithms and scoring functions,¹⁰¹ but it still encounters several limitations, including the entropic effects of the binding, the flexibility of the receptor, the solvation/desolvation

effects and the difficulty to distinguish “true” ligands from false positive bindings.^{99,i} However, docking calculations have proved to be of vital importance in the prediction of ligand-protein complexes and its use have been growing exponentially in the last years.¹⁰¹ A better understanding of the ligand-protein binding event to design better search and scoring algorithms will be of vital importance in the next years for the docking field to reach its zenith.

IV - Potential Energy Surface

The *potential energy surface* (PES) is one of the most used approaches to study a chemical system by computational means. It describes the energy of a given system in function of its nuclear coordinates. A complete analysis of the PES will unveil the energy of each molecular conformation and how it changes depending on the relative position of its atoms. Many different approaches exist to calculate this energy, ranging from very simple and fast methods based on empirical information (MM) to accurate *ab initio* models (QM). Selecting the most suitable methodology taking into account the accuracy/resources ratio is responsibility of the scientist. This energy will be used to determine the most probable path between reactants and products, the *reaction profile*.

In theory, the energy of every geometry that could be adopted by the system can be analyzed in a PES. However, the amount of possible conformations that can exist makes it an infeasible task. The PES has $3N-6$ dimensions, where N is the number of atoms of the chemical complex (the six rigid-body motion corresponding to the translation and the rotation are not taken into account). To make the PES analysis feasible, only the stationary points are taken into account; the minima and the saddle points. The minima points correspond to the stable geometries of the profile, which can be the reactants, the products or the intermediates (Figure 2.7). Each saddle point corresponds to a maximum that links two minima points, but it is a minima in all the other dimensions (Figure 2.7). This point is the lowest energy possible geometry that the system has to undergo between to different consecutive stable structures and corresponds to the transition state.

ⁱ For more information on those limitations, please refer to Annex A.

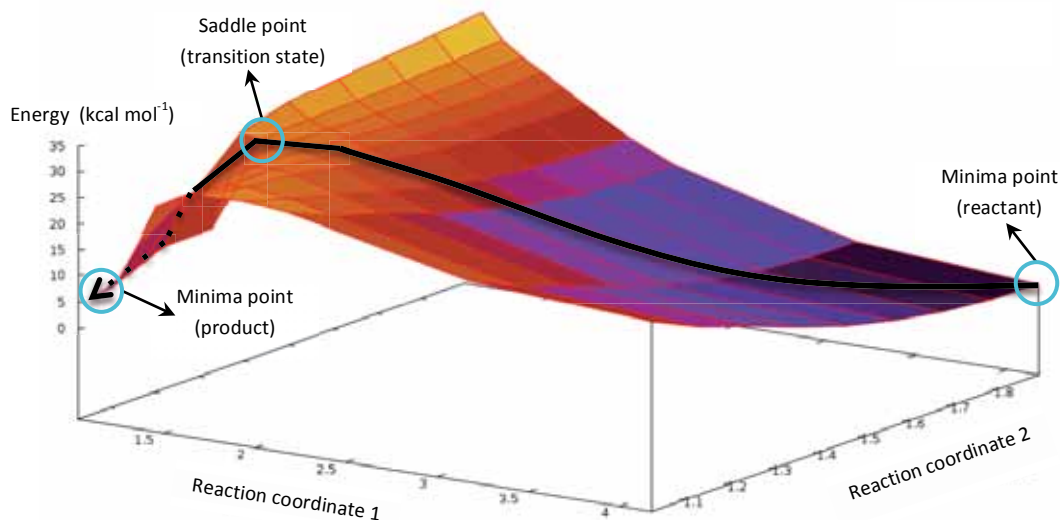


Figure 2.7 - Schematic representation of a three-dimensional PES. The most probable reaction pathway is marked with a black line. All the relevant points (reactant, transition state and product) are highlighted.

V - Integrative approaches

All the current state-of-the-art molecular modelling approaches have limitations narrowing the scope of their applicability. But these limitations can be overcome by combining some of them into new integrative approaches. A clear example of this strategy is the hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) scheme, which enables the study of catalytic mechanisms in big molecular systems (e.g. enzymes).

V.I - Hybrid QM/MM approach

The general idea behind the QM/MM methodology consists in creating different partitions of the system and treating each of them with different levels of theory. A typical example is the study of the catalytic cycle of an enzyme. The region involved in the reaction is usually small with just a few residues interacting with the substrate, while the rest of the system just have a structural or recognition role. Therefore, the system can be split in two, a small region containing the main components of the catalysis (treated with QM) and the rest of the system (treated with MM).

The Hamiltonian or energy of a system treated with a QM/MM approach can be written as follows:

$$\hat{H} = \hat{H}_{QM} + \hat{H}_{MM} + \hat{H}_{QM/MM} \quad \text{Eq 2.9}$$

The \hat{H}_{QM} term corresponds to the Hamiltonian describing the QM region, the \hat{H}_{MM} describes the MM region and the $\hat{H}_{QM/MM}$ describes the non-bonding interactions

between the QM and the MM region. This last term is directly affecting the QM region, as the MM charges can polarize the atoms included in this region.

Splitting the system in two distinctive parts will almost always lead to the scission of some covalent bonds where each one of the bonded atoms will end up in a different QM/MM region. This covalent frontier represents the most difficult part to model in this kind of simulations and has to meet three different criteria to be valid:

- The charge polarization in the frontier has to be the same as if the whole system was treated with a QM approach.
- The frontier atoms must adopt a valid geometry conformation.
- The energy profile of all the terms associated to the frontier bond (rotation, bending, etc.) must be consistent with the profiles obtained using either MM or QM approaches.

The breakdown of the energy terms and the partition of the system summarizes why this approach is so successful. We are not just treating the system independently with two different approaches but the two parts are entwined through the boundary regions and the molecular information can be transferred from the MM region to the QM region and vice versa.

There are several existing methods QM/MM approaches. In this thesis we selected the ONIOM¹⁰² approach as implemented in Gaussian.^j

VI - Integrative platforms

Applying molecular modelling techniques to the study of complex molecular processes still remains a challenging task. But they can be dealt with a multilevel approach consisting of using several molecular modelling techniques in tandem: the whole molecular process can be rendered as different consecutive molecular events, each of which can be described by a particular molecular modelling technique. As seen with the QM/MM hybrid approach, the combination of these methodologies can help to overcome the limitations encountered in each one of them. However this is not an easy task and many difficulties arise when trying to combine several techniques in an integrative protocol. The main problem is the transference of molecular information between them. In fact, in this integrative protocol each individual technique will be performed using different softwares,

^j For more information about the ONIOM methodology, please refer to Annex A.

and each one of them have their own way to store the molecular data. The molecular information has to be translated from one software to another, which is not an easy process and in many cases valuable molecular information could be lost. To avoid these problems and optimize the process, a unique platform encompassing all the different molecular modelling approaches and enabling the free flow of data between them is needed.

There are some commercially available interfaces integrating several molecular modelling techniques.^{103–105} However, these suites present some general disadvantages. First, they are rather expensive, thus limiting the number of groups in the modellers community that could make use of them. Second, they are not open-code and function with a black-box approach. The user can only define the input and wait to obtain the results with no control over what is happening inside the core algorithms of the program. This is rather inconvenient as in many cases the user would want to implement new functions to the original code to adapt it to its own necessities.

In this thesis we created an integrative interface combining several molecular modelling techniques.^k This new interface is free, as it is based on freely available software, and is open-code, so the user will be able to freely modify it to adjust his needs. Two main components will form this integrative interface. The Molecular Modelling Toolkit (MMTK)¹⁰⁶ and the UCSF Chimera.¹⁰⁷ The former is a set of algorithms designed to perform several different MM techniques, such as NMA or MD simulations. But it lacks an interface and preparing the inputs of the calculations is a tedious process. The later is a visualization suite with a highly user-friendly interface that has several tools to analyze molecular data, but cannot perform MM simulations on its own a part from minimizations using the MMTK minimizer. The two softwares are freely available, open code and are based on the same programming language (python). Combining them will result in a highly intuitive interface where several MM simulations can be performed under the same platform, thus the molecular information can be freely transferred from one procedure to the other.

^k The integrative interface created is thoroughly explained in chapter 6.

3

Objectives

I - Objectives

This thesis has been centered on two main entwined objectives. The first one is to **aid in the rational design of artificial metalloenzymes** resulting from the insertion of homogeneous catalysts inside biological scaffolds using molecular modelling approaches. To provide the necessary molecular information needed towards this goal, we aimed at the characterization of the main stages of their development, including:

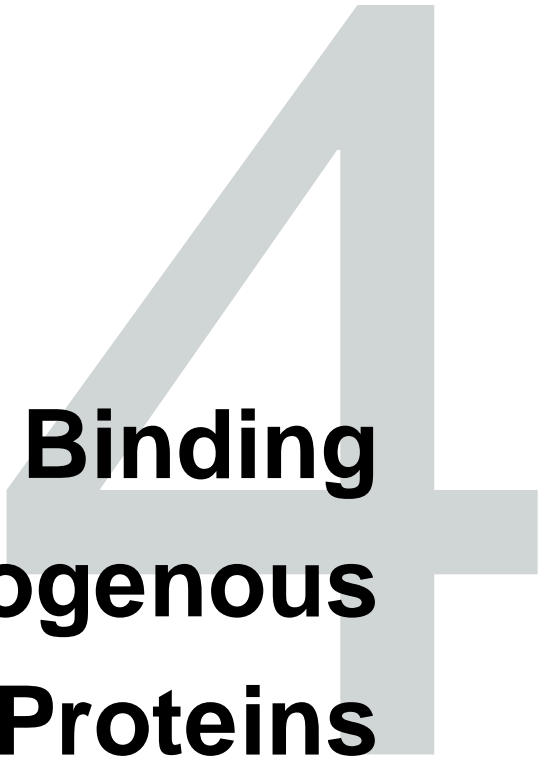
- The binding of the inorganic cofactor in the protein scaffold
- The binding of the substrate in the biometallic complex
- The interactions appearing between the protein/cofactor/substrate triad (characterization of the most likely transition states) for reactivity to occur

With this objective in mind we performed several different computational studies on some of the most relevant artificial metalloenzymes reported so far. The first ones were addressed at the study of the binding of different homogeneous catalysts to their respective biological scaffolds. Those included an artificial oxidase created by Ueno and coworkers,¹⁰⁸ an artificial oxygenase developed by the group of Jean-Pierre Mahy¹⁰⁹⁻¹¹² and some artificial hydrogenases designed by the group of Thomas R. Ward.⁴⁵ Another molecular modelling study was aimed at shedding light on the enantioselective mechanism of a biometallic hybrid constructed by Mahy and coworkers able to perform an epoxidation on several different styrenes.⁴⁷ Finally, we applied our gained *savoir-faire* to decipher the most probable pathways leading to the different enantiomeric products on the most efficient hydrogenase reported by the group of Thomas R. Ward.⁴⁵ Novel mutations/chemical modifications can be proposed for the rational optimization of those hybrids based on the data gathered from our models.

Giving the high complexity of all the events needed to consider in the study of artificial metalloenzymes we needed to combine several molecular modelling techniques, including different levels of theory (QM and MM), to deal with them. This kind of integrative approaches has proven to be useful in the theoretical study of complex chemical/biological systems and is becoming a standard in this field.^{72,84,88,92} However, these tools are very heterogeneous in terms of how the molecular data is processed. For this reason, our second objective focused on the **development of an integrative molecular modelling platform** that enabled the free-flow of data between the different theoretical methodologies implemented. Our developed platform should have the following features:

- Easy transfer of molecular information from one molecular technique to the other
- Open-code, user-friendly and highly customizable
- Free-of-charge
- Set up efficient interfaces for both QM and QM/MM, Normal Modes, protein-ligand docking and Molecular Dynamics
- All under a high standard visualization framework

Developing such platform from scratch would imply an amount of work far from the scope of a single Ph.D. thesis. For this reason, we used already existing molecular modelling tools as the cornerstones. In particular, we focused on the UCSF Chimera¹⁰⁷ and the Molecular Modelling ToolKit (MMTK)¹⁰⁶. Both packages were free-of-charge, python based and open-coded, which perfectly served our purposes. However, their respective roles in the platform were utterly different. While the UCSF Chimera provided with a highly user-friendly environment and a high level of computer graphics, the MMTK was in charge of all the calculation algorithms (including those needed for the molecular dynamics and the normal modes analysis). Additional interfaces were also generated for the well-spread Gaussian (QM) and GOLD (docking) softwares. This way, our integrative platform would also have access to all the molecular data generated by these two programs.



**Prediction of the Binding
of Homogenous
Catalysts to Proteins**

I - Introduction – General concepts and challenges

Artificial metalloenzymes have demonstrated to be of great interest in many fields of the industry, including fine and green chemistry and biotechnology.¹¹³ One of the most successful ways to obtain these entities is the insertion of a homogeneous catalyst inside a protein scaffold.²² However, the proteic receptor has not been optimized during the evolution to bind such synthetic species. Finding a match between homogenous catalysts and a macromolecular receptor is therefore highly challenging. For this reason, the characterization of efficient hybrids requires numerous screening and optimization steps before reaching an acceptable activity or enantioselectivity profile.

The first step along the development of artificial metalloenzymes with inserted synthetic cofactors is to identify systems with a correct complementarity between the homogeneous catalyst and the host protein. This generally requires the exploration of wide chemical and biological spaces. If performed by experimental means this screening could result extremely expensive and time consuming. Therefore, computational strategies and particularly molecular modelling tools could result as important assets in this field.

The studies reported in this chapter focus on this step of the development of artificial metalloenzymes. They include the design of an integrative computational approach able to reproduce all the key molecular events that drive the binding of organometallic catalyst with proteins. This first experience allowed us to further investigate on other systems built on this strategy. We expect that this protocol will provide vital structural information needed for further optimization and design of bioorganometallic hybrids.

II - Heme oxygenase based artificial enzyme

II.I - Computational tools: Integrative approaches

Molecular modelling techniques could help to elucidate the molecular features involved in the binding of an inorganic cofactor inside a protein scaffold. Nowadays, those techniques are broadly used to understand and predict the atomic behavior of chemical or biological systems and its molecular properties. In fact, molecular modelling has proven to be useful in both *de novo* design of proteins¹¹⁴ and in the study of metal-binding proteins.^{113,115,116} However, their application in the study of complex bioinorganic systems is still challenging because several different levels of theory are needed.

To predict the binding of organometallic systems to protein hosts two different processes must be taken into account. First, we need to explore the large conformational space involved in the binding process, which requires empirical calculations of the interaction energies (MM approaches and protein-ligand dockings). Second, we have to accurately represent the electronic properties of the metal to reproduce the influence of its first coordination sphere in the binding process. In this case, QM based methodologies and more particularly QM/MM approaches are required. Allying those two kinds of technologies for the design of artificial enzymes has never been, to the best of our knowledge, reported.

For the study of artificial metalloenzymes an integrative protocol has been designed including different state-of-the-art molecular modelling methodologies. It is mainly divided in two different stages. In the first one, protein-ligand dockings are used to generate an initial low energy ensemble of binding orientations of the inorganic moiety in the protein host. Then, these solutions are then refined through QM/MM calculations taking into account the possible effects that the protein environment could have on the first coordination sphere of the metal. This way we deal with the vast conformational space available in the first stage and model the chelation of the metal in the second one. It is important to take into account the electronic properties of the metal at this point as changes its coordination can by-and-large dictate the configuration of the overall cofactor.

II.II - Benchmarking the protocol: Fe(Schiff base)₂cdHO

Before applying our designed protocol to real systems, we have to benchmark it by *in-silico* reproducing a well-characterized system. Unfortunately, the number of artificial metalloenzymes with their X-ray structures released in the Protein Data Bank (PDB) is still quite small. Amongst all the available structures, we needed one crystallizing many of the challenging issues that could be encountered in the binding studies of homogeneous catalysts in biological scaffolds: (i) an active role of the metal center in the binding, (ii) a significant deformation of the bound cofactor and (iii) a highly flexible receptor. After an exhaustive search, we found an X-ray structure of an artificial metalloenzyme designed by Ueno et. al in 2006 meeting all these conditions.¹⁰⁸ We therefore focused in applying our integrative protocol to obtain an accurate model structure of the system reproducing all these features.

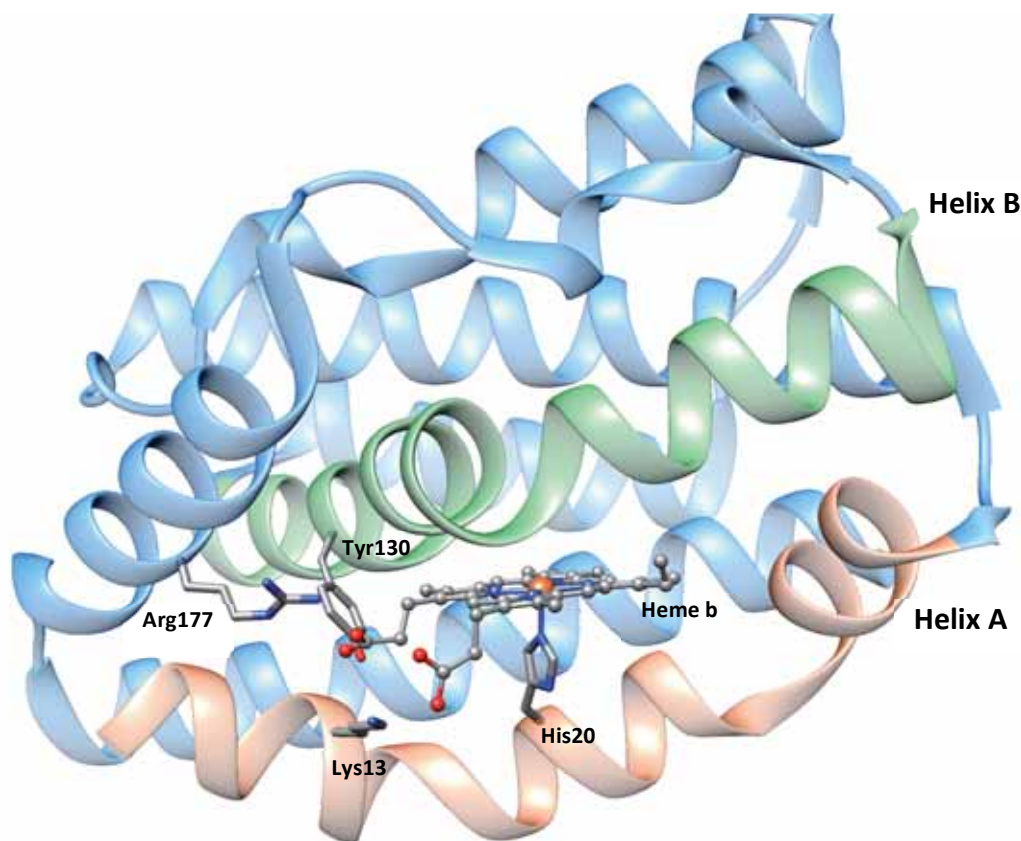
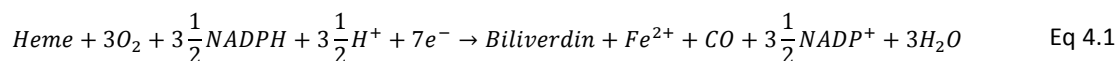


Figure 4.1 - Cartoon representation of the human HO complexed with the heme b group (PDB code 1IW1)¹¹⁷. The inorganic cofactor is located between the α helix A (red) and B (green) and bound to the protein through the chelation of the iron by His20. The polar path constituted by Lys13, Tyr130 and Arg177 interacts with the propionate group of the heme b to further stabilize and guide the binding.

Ueno and coworkers selected a Heme oxygenase (HO) of *Corynebacterium diphtheriae* (*cdHO*) as the host protein for the inorganic receptor (Figure 4.1).¹⁰⁸ This protein is an all- α enzyme responsible for the degradation of the heme group into biliverdin (Eq 4.1):



In the first step of the catalytic mechanism the heme group binds to the cavity of the apo HO between the α helix A and B. Afterwards, the NADPH cofactor provides with the electron necessary to reduce the iron of the heme from Fe(III) to Fe(II). It is only in this oxidation state that the iron is able to bind a molecule of oxygen, which will be transferred to one of the methylene bridges of the heme. As a result the porphyrinic macro-ring opens and the iron is released, obtaining the final biliverdin product (Figure 4.2).¹¹⁸

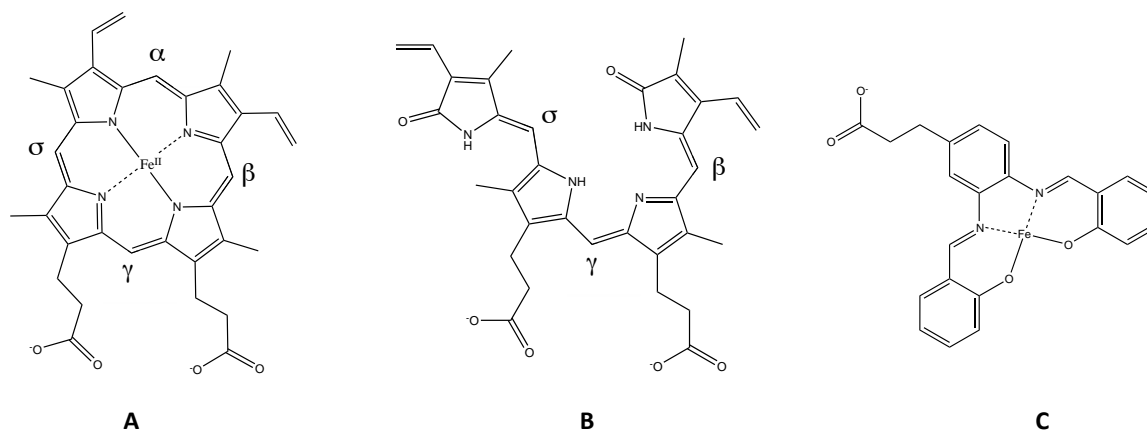


Figure 4.2 - Schematic representation of the heme B (A), the biliverdin (B) and the Fe(Schiff base) (C). The four methylene bridges attaching the pyrrole rings on the heme group are labeled as α, β, γ and σ .

To obtain the artificial metalloenzyme, Ueno and coworkers inserted an iron based salophen (Fe(Schiff base), Figure 4.2) inside the heme oxygenase of *Corynebacterium diphtheriae* (*cdHO*).¹⁰⁸ This kind of organometallic catalysts have long shown similar activities than porphyrin systems and represent interesting candidates in building artificial oxidative enzymes. The main reason behind this is the high structural resemblance between them. The iron salophen represent almost $\frac{3}{4}$ parts of the porphyrin ring: four atoms coordinating the metal center (two N and two O in this case) linked by various methylene bridges. However, they are much more flexible because their central macro-ring is not closed. In this particular case, the salophen used to generate the Fe(Schiff base) \subset *cdHO* hybrid had one propionate group as substituent of the central benzene (Figure 4.2). This group is present in one of the most common substrates of the HO, the heme b, and has two different functions: (i) maintaining the electron flow with the reductase partner and (ii) guiding the binding in the HO active site. In fact, it was observed that the lack of the propionate in the synthetic salophen resulted in the loss of all activity.¹⁰⁸

Despite the high similarities between the porphyrinic catalysts and the salophens (in terms of both structure and reactivity), the artificial Fe(Schiff base) \subset *cdHO* hybrid was only able to activate an oxygen molecule to form the corresponding superoxide and no substrate oxidation could be performed. This artificial metalloenzyme has therefore a rather limited scope for oxidative chemistry with superoxide anions being extremely reactive free radicals.

One of the decisive factors in choosing this system as the benchmark of our integrative protocol were the puzzling structural characteristics observed in the X-ray structure.¹⁰⁸ Both salophens and porphyrinic cofactors usually adopt a planar configuration, regarding the number of coordinations with the metal center. In fact, the heme b group bound to the HO adopts a pentacoordinated square pyramidal configuration (**A**, Figure 4.3), and even after the binding of the oxygen molecule its central macroring does not abandon its general planarity. However, the synthetic cofactor in the Fe(Schiff base) \subset cdHO structure presents a distorted octahedral conformation with two different chelations with residues His20 and Glu24 (**B**, Figure 4.3). This unusual structure is possible thanks to the open macrocycle present in the salophen catalyst, which provides the necessary flexibility to adopt this configuration and allows the Glu24 to enter the first coordination sphere of the metal. Additionally, the cdHO selected as the protein scaffold presents a remarkably high flexibility. Upon the binding of the heme group, both α helix A and B suffer a hydrophobic collapse, closing the cavity around the prosthetic group in a collective movement that involves almost half of the protein.

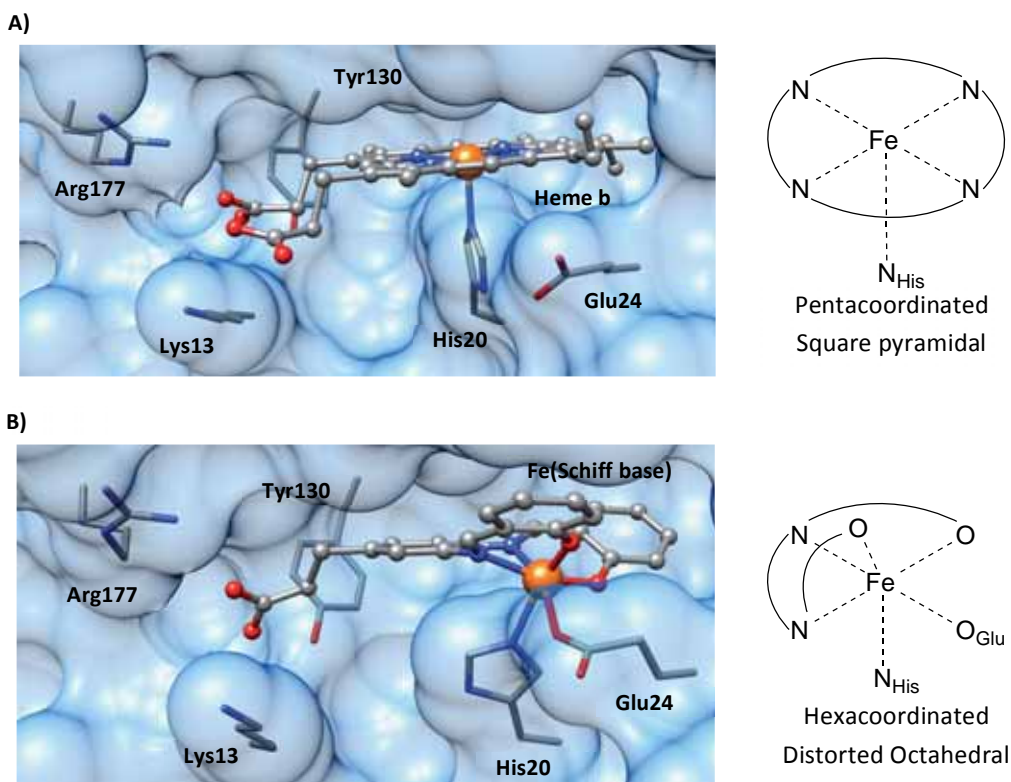


Figure 4.3 - Cartoon representations of the bound forms of the heme b (panel **A** left part, PDB code 1IW1)¹¹⁷ and the Fe(Schiff base) (panel **B** left part, PDB code 1WZD)¹⁰⁸ in the cdHO host. At the right of each panel is the schematic representation of the first coordination sphere of the metal center.

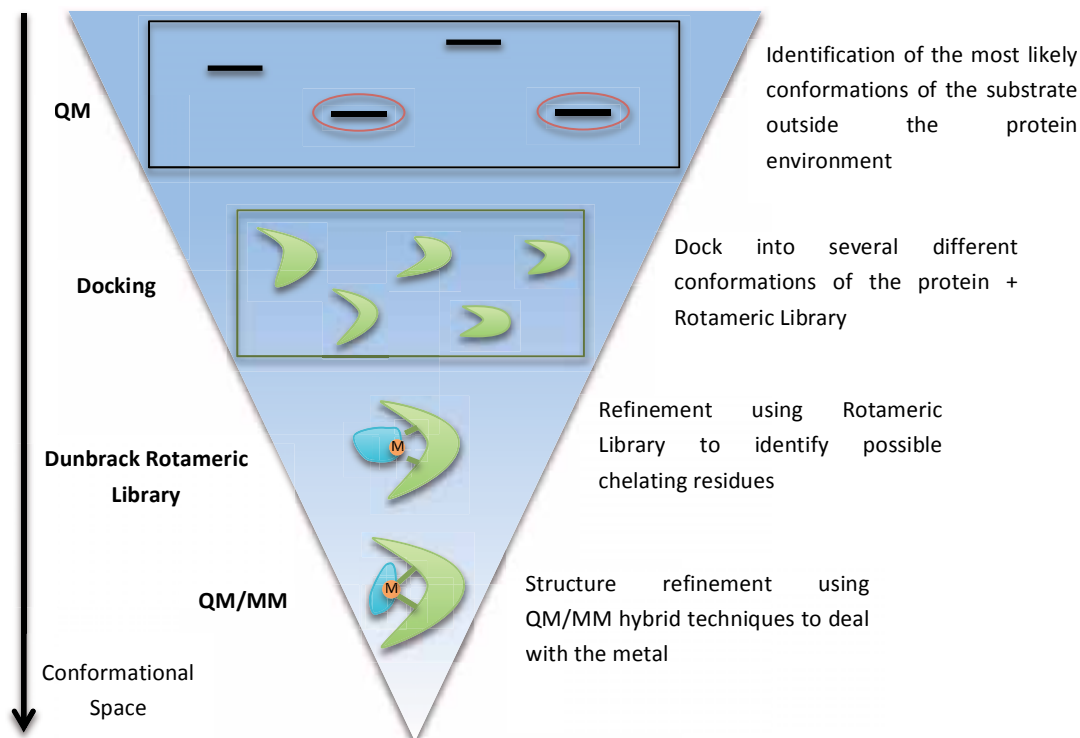
Thanks to the structural data obtained from the crystallographic structure a smaller inorganic cofactor with better affinity for the *cd*HO receptor was built. However, no efficient enzymatic activity was obtained with this novel generation and further artificial enzymes based on this heme oxygenase technology were dropped.¹¹⁹

The active role of the metal in the binding of the Fe(Schiff base), the abnormal configuration adopted by the synthetic cofactor and the high flexibility of the *cd*HO host protein are of the most common challenges encountered by molecular modelling approaches in the study of complex bioorganometallic systems. Our integrative approach combines different QM and MM methodologies to reproduce all these features in a Fe(Schiff base) \subset *cd*HO model structure and can become a great asset in both the design and optimization of these species.

II.III - Presentation of the protocol

From a modelling perspective, the system developed by Ueno and coworkers is of the most complex to reproduce. It is therefore particularly indicated to test of our working hypothesis on a multilevel strategy. The particular protocol developed in this case consists on a four-step scheme combining several molecular modelling techniques (Scheme 4.1).

Similarly to any docking process, the first step of the protocol consists in the characterization of the geometry of the isolated cofactor based on QM calculations. At this point, all the electronic properties of the metal (oxidation and spin states) are considered. In the second step, the cofactor is docked inside the binding site of the protein using the geometry (or the geometries) previously identified. The lowest energy docking solutions are subjected to an exhaustive analysis using an *in house* algorithm to find residues able to reach the metal in the binding site. Applying the Dunbrack Rotameric Library¹²⁰ the selected residues (if any) are replaced for those that can coordinate the metal center. On the final step the resulting structures are refined using a hybrid QM/MM methodology to account for possible changes in the first coordination sphere of the metal due to the protein environment. The QM/MM energies obtained for each structure are then compared to discriminate which one is the most probable model to represent the cofactor \subset protein system. Details on the different steps of the protocol are provided in the very next paragraphs.



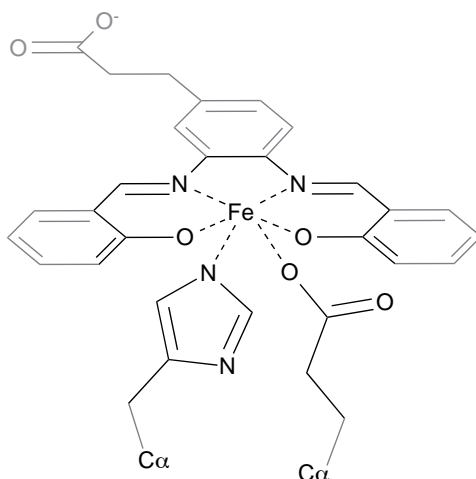
Scheme 4.1 - Schematic representation of the integrative protocol created by combining several molecular modelling techniques. It is designed so that the approaches that are more computationally intensive have to deal with the smaller conformational space.

The structures of the isolated cofactors were prepared using a DFT approach. The functional B3LYP^{121,122} was used as implemented in the program package Gaussian03.¹²³ The basis set LANL2DZ¹²⁴ and its associated pseudo-potential was used to accurately treat the Fe atom, while the rest of the atoms was treated using the 6-31g** basis set.^{125,126}

Protein-ligand dockings were performed using the program GOLD4.1¹²⁷ and the ChemScore¹²⁸ scoring function. This program was selected as it is one of the few commercially available docking software containing parameters for the interaction of organic ligands with metal containing proteins.¹²⁹ For this study, we used those parameters to develop a new atom type to represent the metal center present in the ligand. This new atom type would behave like an hbond donor to mimic the propensity of the metal atoms to interact with Lewis bases. It would be able to recreate regular octahedral geometries and the force of the interactions would be computed in the scoring function as regular hbond contacts. All the structures have been prepared as dictated by the GOLD user manual using the UCSF Chimera¹⁰⁷ visualization program.

An *in house* algorithm was developed for the determination of the residues that are able to coordinate the metal in the docking generated structures. This approach shares some common grounds with some recently published algorithms aimed at predicting the binding of metal ions.¹³⁰ Based on a statistical analysis carried on more than 400 iron containing proteins, a 7 Å distance between the iron center and the C α of the target residue was used as a selection criteria. Only polar residues with a rotameric position able to reach the metal (3.5 Å between the coordinating atom and the metal) without presenting important close contacts are retained as potential targets.

QM/MM calculations have been performed using the ONIOM¹³¹ scheme as implemented in Gaussian03¹²³. The central part of the salophen and the lateral chains of His20 and Glu24 were included in the high layer (Scheme 4.2), which was treated using the same QM methodology as in the isolated cofactor. The rest of the system was included in the low layer (Scheme 4.2) and was treated using the AMBER¹³² force field. The charges of the system were calculated using the Antechamber¹³³ software as implemented in the UCSF Chimera.¹⁰⁷ For a better treatment of the flexibility of the system, the entire cofactor and the α helix containing the chelating residues were allowed to relax during the optimization. In certain model structures some clashes present in the vicinity of the metal could prevent the hypothetical chelating residues to reach the metal. To prevent falling into a local minima as a result of this bad contacts a small distance restraint was added between the hypothetical chelating atoms and the metal center, which was relaxed in a second round of QM/MM minimizations.



Scheme 4.2 - Depiction of the QM/MM partition of the salophen and the chelating residues. The highlighted atoms are included in the high layer while the rest of the system is included in the low one.

II.IV - Analysis of the conformational space of the organometallic cofactor

The first step in our integrative approach is to generate the structure of the isolated Fe(Schiff base). As we are dealing with a metal-containing ligand it is important to take into account each possible electronic state of the metal at this point as they could have a dramatic impact in the configuration of the entire cofactor. For this purpose calculations on the Fe(Schiff base) were performed in both Fe(II) and Fe(III) oxidation states taking into account both the low spin (LS) and high spin (HS) possibilities in each state.

The four different structures generated presented no relevant differences between them. Even though the high spin forms had a little doming of the overall cofactor, all of them maintained the typical planarity of the salophens (Figure 4.4). In fact, the calculated RMSD between them was no higher than 0.3 Å, suggesting that the different electronic states of the metal had essentially no effect on the geometry. Subsequently, the Fe(II) low spin model was selected as the representative model structure for the Fe(Schiff base) cofactor in the docking calculations.

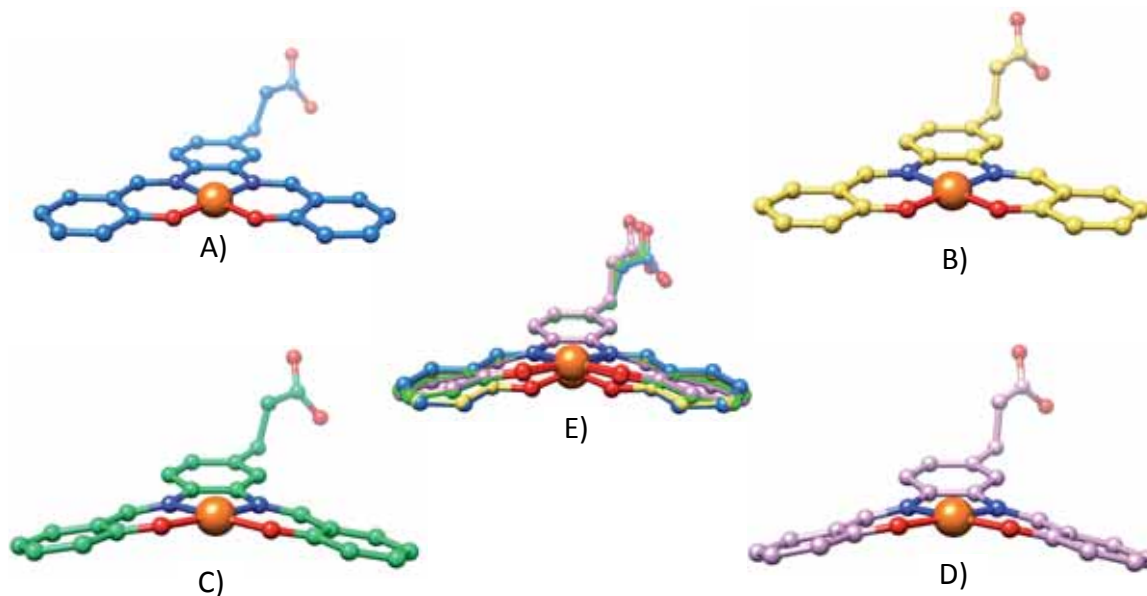


Figure 4.4 - Model structures for the different electronic states of the iron center in the Fe(Schiff base) cofactor: Fe(II) low spin (A), Fe(III) LS (B), Fe(II) HS (C) and Fe(III) HS (D). Panel E shows the superposition of the four of them.

To validate our DFT approach the optimized structures were compared with all the iron based salophen complexes available in the Cambridge Structural Database^a (CSD) (Figure

^a CSD codes of the selected iron salophens for the comparison: BEKZIA, RIVCEF, EXOJUW, SERPUA, FEWVUY, SUDJUW, GURKUZ, SUDJUW10, JUJDUN, JUJDUN01, LALTOH, MIGDAI.

4.5).¹³⁴ The results highlighted a good agreement between the optimized geometries and the crystallographic data, with RMSD no higher than 0.4 Å. Additionally, none of the crystallographic structures adopted the distorted octahedral configuration observed in the X-ray structure of the Fe(Schiff base) \subset cdHO hybrid. This is a reflection of the dramatic impact that the protein environment could have in the overall configuration of the cofactor through influencing the electronic properties of the metal.

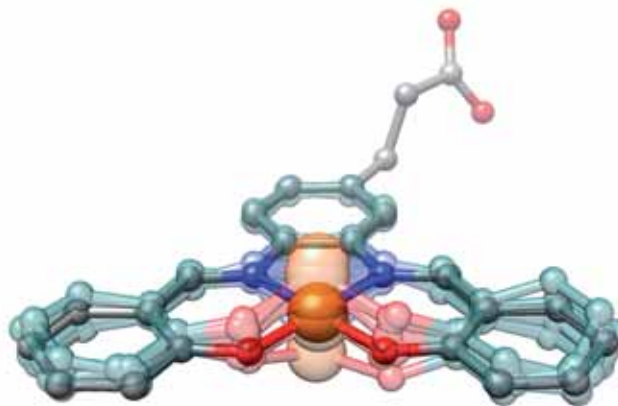


Figure 4.5 - Superposition of all the available iron based salophens on the CSD (ghost color) with the Fe(III) high spin model of the Fe(Schiff base) (solid color). Only the salophen center of the different CSD structures is displayed for clarity reasons.

II.V - Generation of the Fe(Schiff base) \subset cdHO model structures

The Fe(II) low spin model generated using QM approaches was docked into several different HO structures available in the PDB, including: (i) the free-heme (PDB code 1NI6),¹³⁵ (ii) the heme bound (PDB code 1IW1),¹¹⁷ and (iii) the heme-bound with a large inhibitor coordinating the iron (PDB code 3CZY)¹³⁶. All these structures were selected for a better representation of the flexibility of the protein, as they compose a large spectra of possible ligand-state conformations of the HO. However, some considerations were taken into account prior to the docking calculations. Amongst all the selected HO conformations only the 1IW1 belonged to *Corynebacterium diphtheriae*, while the other two were human. The host selected by Ueno and coworkers also belonged to this microorganism, so to give consistency to the study homology models were created for the human HO to obtain their corresponding cdHO model.

The docking simulation provided with a vast set of different orientations with no apparent correlation between them. To shed some light in this chaos, all the binding modes were superimposed. The resulting RMSD presented substantial variations, ranging from 0.42Å

to 6.7Å, and unveiled the existence of 4 major clusters (hereafter referred as A, B, C and D, Figure 4.6).

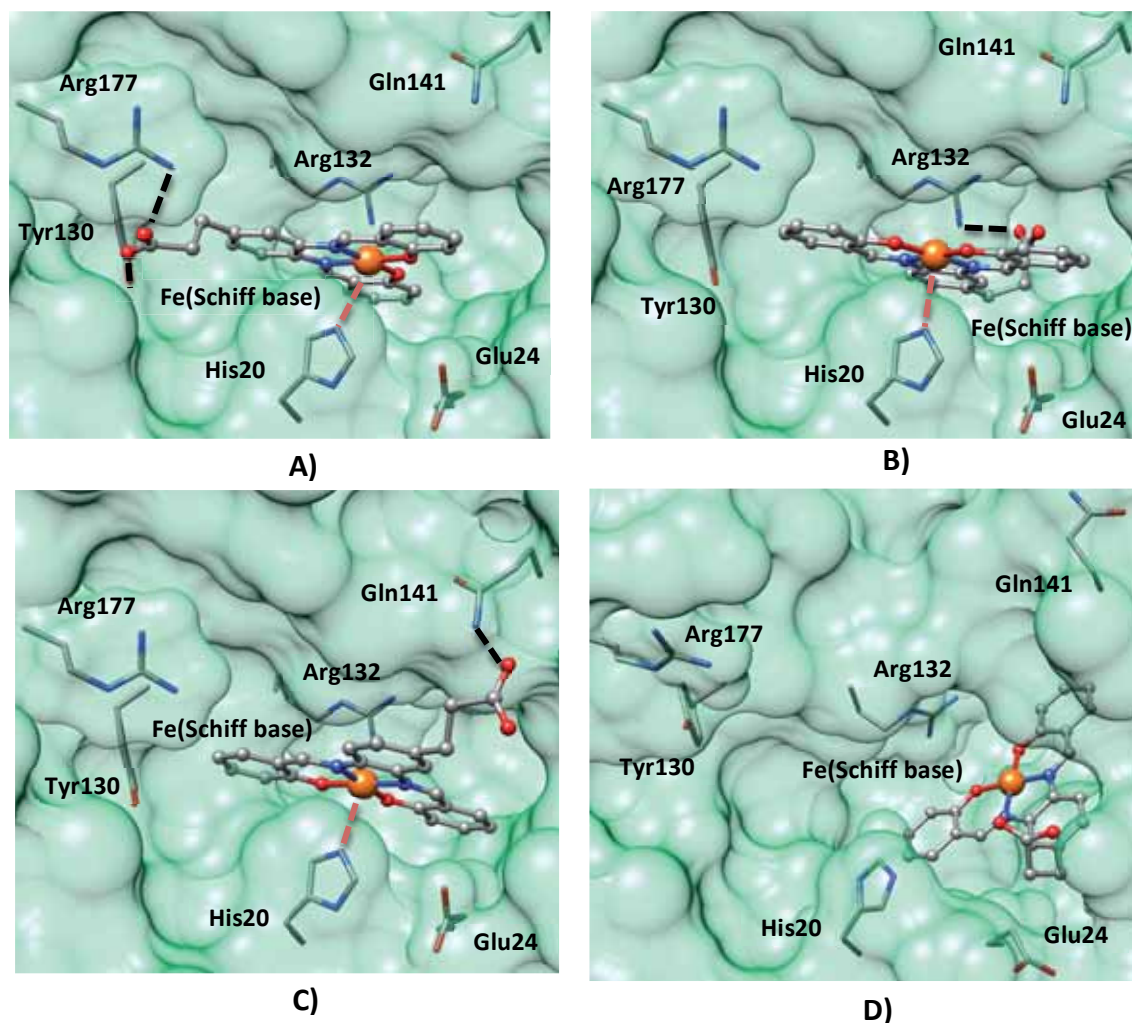


Figure 4.6 - Depiction of a representative model for each one of the four different clusters identified amongst all the different binding modes obtained in the docking simulation. Orientations A, B, C and D are represented in panels **A**), **B**), **C**) and **D**) respectively. Main interactions between the Fe(Schiff base) cofactor and the protein scaffold are highlighted with dashed lines (black for hbond, red for hypothetical chelations of the metal).

Amongst all four, orientations A, B, and C shared some common features, while D was quite different from the rest. Indeed, these three clusters had the central aromatic ring of the Fe(Schiff base) located on the same region of the receptor between α helix A and B, occupying the heme binding site (Figure 4.6). Additionally, the docking results suggested a hypothetical chelation of the iron center by His20 in all three of them. Despite those similarities, the three orientations could be discerned by the position of their propionate

tail. Orientation A had it interacting with the polar patch composed by Arg177 and Tyr130, orientation B had it inserted deep buried inside the binding cavity and interacting with Arg132 and C had it mainly solvent exposed with some hbond contacts with shallow residues like Gln141 (Figure 4.6). Orientation D had a significantly different conformation, with no analogy with any of the other clusters. In this case, the whole synthetic cofactor was deep inserted inside the cavity, in a region where iron chelating inhibitors are usually found.¹³⁶ In this particular orientation, the propionate is not making any interaction with the protein scaffold, neither is there a residue able to reach metal (Figure 4.6).

To unveil the impact of the flexibility of the protein in the binding of the Fe(Schiff base) we classified the solutions obtained in each *cdHO* host according to one of the four main clusters (orientations A, B, C and D) identified in the previously RMSD analysis (Figure 4.7). The data obtained clearly indicated that the protein conformation had a tremendous impact in the binding orientations adopted by the synthetic cofactor. In fact, orientation D was restricted only to those *cdHO* configurations presenting a wider binding cavity (heme free and heme inhibitor, Figure 4.7), while the other orientations were present in almost every receptor structure but more centered on the apo form (heme bound, Figure 4.7).

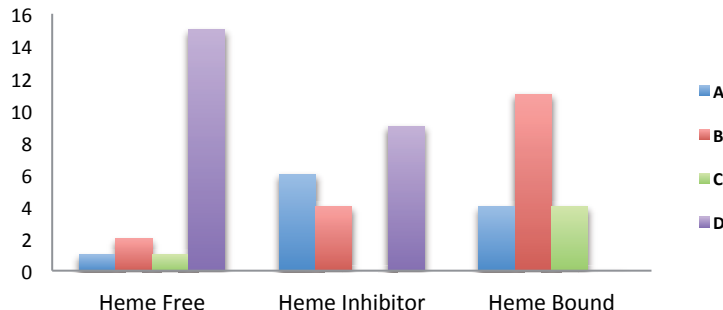


Figure 4.7 - Distribution of the docking solutions obtained on each different *cdHO* receptor. The height of the bars corresponds to the number of solutions belonging to each cluster (binding modes A, B, C and D).

All the lowest energy solutions were close to 30 Score units, with differences no higher than 8 units despite the *cdHO* configuration or the cluster orientations (Table 4.1). Even though these binding energies indicated a good complementarity between all the different protein receptors and the Fe(Schiff base), we could not make any further assumption using an energetic criterion for various reasons. First, we are treating the metal with an approximate pseudo-metal atom type that has not been optimized. Second, the changes on the first coordination sphere as a result of a possible chelation of the

metal could not be reproduced by docking means. Those two factors could have a dramatic effect on the final scores and cannot be neglected.

Orientation	Heme Free	Heme Inhibitor	Heme Bound
A	28.2	31.5	32.7
B	-	30.8	31.4
C	-	24.9	32.7
D	30.5	31.0	-

Table 4.1 - Score values for the lowest energy solution for each one of the four different conformations of the cofactor obtained in the three different *cdHO* conformations using the ChemScore scoring function.

To carefully treat the metal chelation, we need QM/MM techniques. However they are computationally intensive and cannot explore large conformation spaces. Performing such optimizations taking into account the four main orientations identified and all the *cdHO* receptors would be a task too costly for the scope of this work. Our goal is to design an integrative approach able to predict the binding of inorganic cofactors inside biological scaffolds with an efficient “accuracy/resources consuming” ratio. Subsequently, some structural and experimental information was taken into account to narrow the number of Fe(Schiff base) \subset *cdHO* models obtained in the docking calculations that should be refined through QM/MM means.

The first thing to take into account is the chelation of the metal. This interaction should be by far stronger than any other one the Fe(Schiff base) could make with the protein host, thus it can be considered the driving force of the binding. Amongst the four orientations observed, orientation D was the only one in which no possible chelation of the metal center could be identified using our pseudo-metal type atom in the docking calculations. A visual analysis of these docking solutions highlighted that there was no residue able to do this role in the vicinity of the metal. Subsequently, all binding modes belonging to this particular orientation were discarded at this point. Regarding the other orientations the only important structural difference is the position of the propionate tail. Different experiments have pointed out the important role of this functional group on the electron flow in the Fe(Schiff base) \subset *cdHO* catalytic mechanism.¹⁰⁸ From a structural point of view, orientation C is not likely to account for experimental behavior. Its propionate tail is mainly solvent exposed, with some hbond interactions with shallow residues of the protein that are not expected to be very stable due to the presence of the solvent. On the contrary, in the other two clusters the propionate tail was deep buried inside the cavity,

making several hydrophobic and hbond interactions with the host. Consequently, orientation C was also dropped.

Even if two orientations have been discarded, we still have four different *cdHO* conformations to consider in the generation of the Fe(Schiff base) \subset *cdHO* model structures. However, in this case we can also take advantage of the structural information available on the HO to discard some of them.

It has been reported that upon binding of the heme group the HO closes its binding site around the prosthetic group.¹¹⁸ Due to the high structural resemblance between the Fe(Schiff base) and the heme group we expected something similar would happen in the binding of the former. Only those models of the hybrid generated using the 1IW1 (heme bound) structures presented such foreseen conformation in the binding site as they corresponded to the apo form of the *cdHO*. The other two conformations 1NI6 (heme free) and 3CZY (heme bound with a large inhibitor) had wider cavities that do not fit with this expected behavior of the receptor. Additionally, the orientations chosen to reproduce the model of the biometallic hybrid were clear minorities in those two scaffolds (Figure 4.7). Therefore, the 1IW1 was more likely to better reproduce the actual Fe(Schiff base) \subset *cdHO* artificial metalloenzyme and the other two structures were subsequently discarded.

Thanks to the structural and experimental data, we could limit all the possible Fe(Schiff base) \subset *cdHO* models to only two. They were generated with the lowest energy binding modes of orientations A and B in the 1W11 *cdHO* host. Both models were then analyzed with an in house developed algorithm to identify other possible chelating residues that could not be identified using our pseudo-metal atom type in the docking simulations. These way we could narrow even more the conformational space prior to the QM/MM refinement of the Fe(Schiff base) \subset *cdHO* models.

II.VI - Analyzing the metal environment

The qualitative nature of the calculated binding energies and the non-optimized parameters to model the metal center in the docking cannot accurately predict all the possible interactions between this atom and the protein environment. We therefore have to consider that some residues able to coordinate the metal may have been overlooked during the docking simulation. QM methodologies cannot explore large conformational spaces, so identifying those residues before performing the QM/MM simulations is vital. Otherwise the calculations could fall into a local minimum not corresponding with the real

Fe(Schiff base) \subset cdHO structure. To avoid this situation, we developed an exploration algorithm able to identify possible chelating amino acids in the vicinity of the metal and provide with a more realistic conformation of it prior to the QM/MM refinement.

The search algorithm was based on a statistical analysis we performed on several PDB structures with residues coordinating an iron atom. It was observed that all residues have its C α within a 9Å distance to the metal center. Our algorithm written in Python under the UCSF Chimera environment analyzes which residues of the binding site able to chelate a metal have its C α within this distance. The ones identified are then replaced with a rotamer displaying a more suitable configuration for the chelation of the metal. If there is no such rotameric position, the residue is no longer considered.

Apart from His20, which was already suggested as a possible chelating residue in the docking calculations, the algorithm highlighted six other targets: Glu21, Glu24, Arg177, Asp136 and Ser138 (Figure 4.8). The analysis of the Dunbrack Rotameric Library¹²⁰ for each one of them displayed that only Glu24 and His20 had a suitable rotameric position able to reach the iron within an acceptable distance to create a coordination bond. All the other ones were either unable to reach the metal or had too many bad contacts with the surrounding atoms.

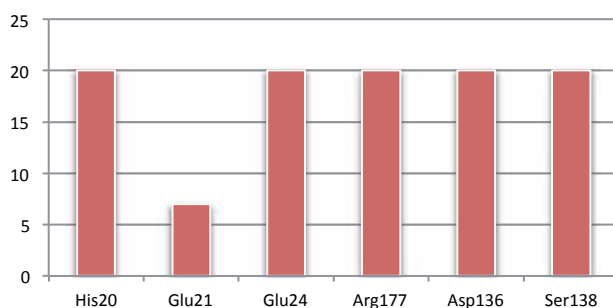


Figure 4.8 - Number of possible residues able to chelate the metal found by the *in house* algorithm when analyzing all the docking solutions of the Fe(Schiff base) \subset cdHO model.

All the possible combinations of rotameric positions for His20 and Glu24 were analyzed to unveil which ones were more likely to chelate the metal. This analysis highlighted two different ones that could satisfy this purpose without major rearrangements of the protein environment (Figure 4.9). The first led to a double chelation of the iron by both His20 and Glu24 (HE1), while in the second only the His20 is able to reach the metal center and the Glu24 is located outside the first coordination sphere (HE2).

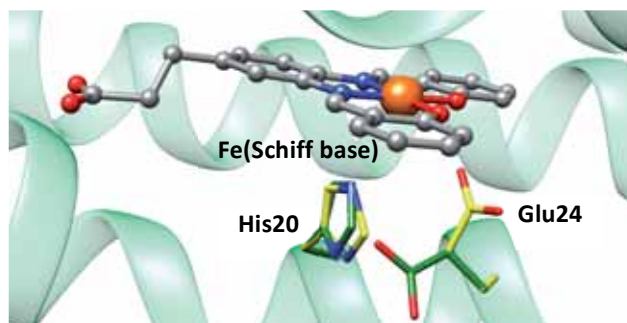


Figure 4.9 - Cartoon of the Fe(Schiff base) \subset cdHO model structure generated using orientation A highlighting the rotameric positions of Glu24 and His20 able to chelate the iron. Combination HE1 is depicted in green and HE2 in yellow.

The configurations of residues His20 and Glu24 in the two models of the Fe(Schiff base) \subset cdHO hybrid were replaced for the two different combinations (HE1 and HE2) of rotamers able to chelate the metal. This resulted in four new models, each one of them containing one of the main docking orientations (A and B) and one of the two rotameric combinations (HE1 and HE2). Those featuring orientation A were referred as model A1 and A2, depending on the rotameric combination (HE1 and HE2 respectively) and the same goes for those displaying orientation B (B1 and B2 for combinations HE1 and HE2 respectively).

The four models were minimized using a QM/MM approach. With this methodology we could take into account all the possible effects that the protein environment could have on the metal center and how this will, in turn, impact on the binding itself. The information provided will allow us to finally discern which of the models generated for the Fe(Schiff base) \subset cdHO hybrid is most likely to correspond to the real system.

II.VII - QM/MM refinement

To have an accurate representation of the metal center in the four Fe(Schiff base) \subset cdHO models generated (A1, A2, B1 and B2) we have to consider all its possible electronic states. Subsequently, all the calculations were performed with the metal in its Fe(II) and Fe(III) oxidation states and with their corresponding low and high spin configurations.

Calculations in both A1 and A2 models indicated that there was no impact in the overall structure of the Fe(Schiff base) cofactor as a result of the different electronic states of the metal considered. However, the divergent configurations adopted by the synthetic cofactors in both models were an indicator of the significant impact that the protein could have in the first coordination sphere of the metal. In the model A1 all calculations led to a

distorted octahedral configuration resulting from the double chelation of His20 and Glu24, abandoning the typical planarity of the salophen catalysts (**A**, Figure 4.10). On the contrary, all minimizations of model A2 resulted in a planar pentacoordinated square pyramidal geometry (**B**, Figure 4.10). In this case, only His20 was coordinated to the metal, while Glu24 was driven out of the first coordination sphere.

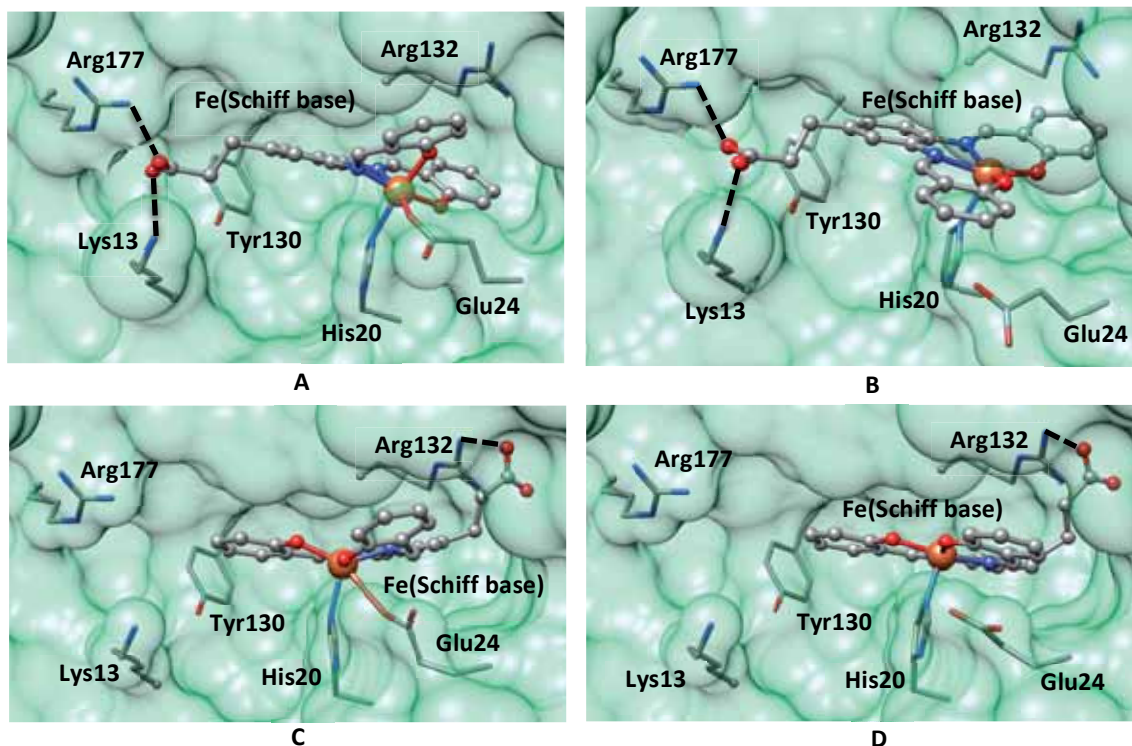


Figure 4.10 - Cartoon representation of the minimized models of the $\text{Fe}(\text{Schiff base})\text{-}cd\text{HO}$ using a QM/MM scheme. Panels **A** and **B** correspond to the hexa- (A1) and pentacoordinated (A2) forms of model A respectively. Panels **C** and **D** correspond to the hexa- (B1) and pentacoordinated (B2) forms of model B respectively. All these structures are in the Fe(II) HS electronic state.

QM/MM calculations on the B2 model also resulted in no major differences between the different electronic configurations of the metal taken into account. All simulations resulted in a pentacoordinated square pyramidal conformation where only His20 was coordinated to the metal center (**D**, Figure 4.10). Interestingly, in this case the central aromatic ring of the homogeneous catalyst suffered a considerable deformation caused by the protein environment. Calculations on the B1 model were the only ones where the different electronic states had a significant impact on the geometry of the cofactor. All minimizations led to a distorted hexacoordinated octahedral configuration with a double chelation by His20 and Glu24 (**D**, Figure 4.10), except for the Fe(III) and Fe(II) in low spin. The former adopted a pentacoordinated square pyramidal conformation where only His20

was inside the first coordination sphere, while no stable minimum could be identified for the later. Interestingly, all stable structures of the B1 model presented the same distortion of the central aromatic ring of the cofactor observed in the B2 models.

All the obtained QM/MM energies were compared to identify the most stable geometries taking the model A1 as the reference value for each possible electronic state of the metal (Table 4.2). The B models were substantially higher in energy than the A ones, with differences higher than 140 kcal mol⁻¹ in all cases. This difference in energy could be a result of the high deformation observed in the central aromatic part of the Fe(Schiff base). Regarding the A models, the energies for the penta- (A1) and hexacoordinated (A2) configurations were very similar with energy differences no higher than 7 kcal mol⁻¹. The only exception was the Fe(III) low spin electronic state, in which this difference was of 33.1 kcal mol⁻¹. Interestingly, the A1 model was more stable than the A2 in almost all the electronic states of the iron considered, except for the Fe(II) low spin in which this tendency is inverted.

Electronic configuration	Model	ΔE_{QM}	ΔE_{MM}	$\Delta E_{QM/MM}$
FeII LS	A1	0	0	0
	A2	-25.7	16.9	-8.8
	B1	-	-	-
	B2	-5.9	147.3	141.3
FeII HS	A1	0	0	0
	A2	-6.0	8.6	2.6
	B1	2.7	151.8	154.5
	B2	13.8	174.2	188.0
FeIII LS	A1	0	0	0
	A2	22.2	10.9	33.1
	B1	20.4	150.3	170.7
	B2	40.1	192.6	232.7
FeIII HS	A1	0	0	0
	A2	10.9	-10.4	0.5
	B1	16.9	124.7	141.6
	B2	-4.4	176.5	172.1

Table 4.2 - QM/MM energies for all the minimized models in kcal mol⁻¹. All energies are compared with the model A1 with the same electronic features

Two different conclusions can be extracted from the analysis of the QM/MM energies. First, we can discard all those models containing the B orientation of the Fe(Schiff base) because they were much higher in energy than the ones containing the A counterpart. Second, the low difference in energy between the penta- and the hexacoordinated forms of these models suggests that the two species could coexist in the same experimental conditions. We therefore hypothesized that both A1 and A2 models could correspond to different intermediates of the catalytic cycle of the Fe(Schiff base)_{cd}HO biometallic hybrid.

II.VIII -Comparison with the experimental system

Our integrative protocol predicted the existence of two different forms for the Fe(Schiff base)_{cd}HO hybrid depending on the electronic configuration of the synthetic cofactor. One was a pentacoordinated configuration presenting a chelation of His20, while the other was an hexacoordinated one with a double chelation of His20 and Glu24. The former is likely to correspond to the experimental structure obtained by Ueno and coworkers.¹⁰⁸ In both cases the Fe(Schiff base) cofactor presented an unusual distorted octahedral hexacoordination with the same ligands on the first coordination sphere of the metal. In fact, the superposition of the two inorganic moieties resulted in an RMSD of 0.6 Å, highlighting the high resemblance between them. However, there was no experimental evidence suggesting the existence of a pentacoordinated form of the Fe(Schiff base) in the protein. At this point, the following question arises: Is this model structure an artifact of the method, or by contrary, it has a relevant role on the catalytic mechanism?

To answer whether the pentacoordinated form of the Fe(Schiff base) exists or not, we should focus on fitting the models obtained in the catalytic mechanism of the Fe(Schiff base)_{cd}HO artificial enzyme. This hybrid is able to activate an oxygen molecule from O_2 to O_2^- . Two different events are needed for this process to take place: (i) the iron (originally in Fe(III) oxidation state at the beginning of the cycle) has to be oxidized to Fe(II) and (ii) it has to be able to fix the oxygen. Looking at the experimental hexacoordinated structure, it seems rather impossible for it to be the active form of the enzyme. The first coordination sphere of the iron is already filled and there is no space in the binding pocket to allocate the oxygen above the iron center (6.1 Å³ available space versus 66 Å³ room needed for the O_2) (**A**, Figure 4.11). So there must be another form of the biometallic complex that could fulfill this role, for which we proposed our pentacoordinated Fe(II) model. In this case, the oxygen could bind the iron because: (i) it is already in its Fe(II) oxidation state, (ii) the first coordination sphere have one vacancy and (iii) there was enough space to allocate the oxygen above the iron (70 Å³) (**B**, Figure 4.11).

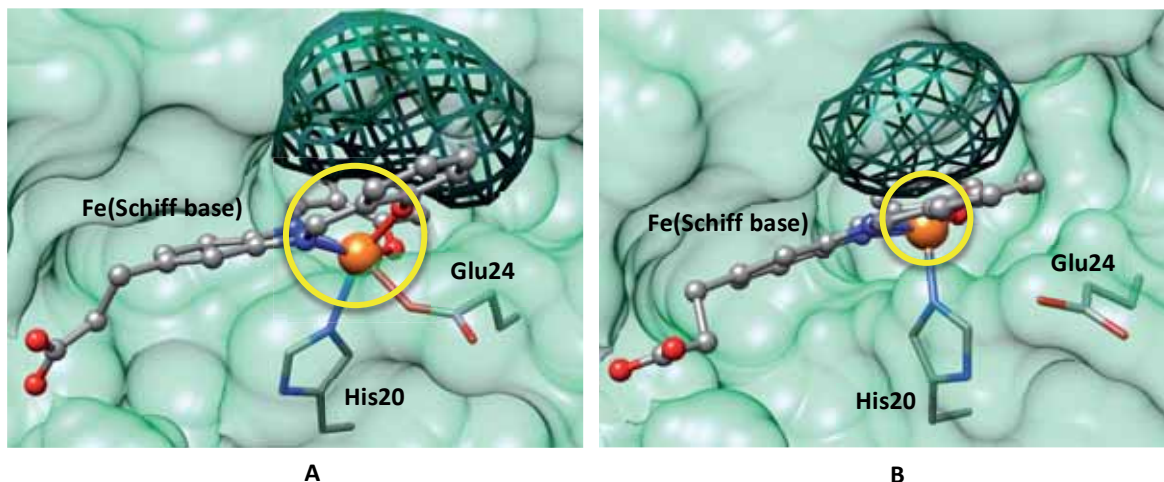


Figure 4.11 - Cartoon depiction of the space available for the fixation of the oxygen in both the hexa- (A) and the pentacoordinated (B) forms of model A. As highlighted by the yellow circle, the O_2 cannot reach the iron in the hexa form, while it is completely accessible in the penta.

In view of the results obtained, it seems plausible to hypothesize that the pentacoordinated Fe(II) models (regarding of the spin state) should represent the reactive form of the $Fe(\text{Schiff base})\text{-}cdHO$ structure. In this situation the catalytic cycle should go as follows: (i) the synthetic catalyst in its Fe(II) pentacoordinated form binds the oxygen molecule, (ii) transfers an electron to it and goes from Fe(II) to Fe(III), (iii) the activated oxygen leaves and the Glu24 enters the first coordination sphere of the metal, thus completing the catalytic cycle. This conversion between the different oxidation states of the iron seems possible due to the small energetic gap between them (Table 4.2). To clearly validate this hypothetical cycle, we need to find the corresponding transition states and obtain the corresponding potential energy surface. However, a study of this kind is far from the scope of this work and is therefore not discussed here.

II.IX - Conclusions

In this section we have demonstrated that combining several molecular modelling techniques could lead to a highly predictive integrative approach. Protein-ligand dockings can offer a good representation of the binding modes of an inorganic cofactor, which can be later refined by a rational analysis of the binding site with QM/MM methodologies. Using this integrative protocol we have obtained two different models for the $Fe(\text{Schiff base})\text{-}cdHO$ structure, each one containing either an pentacoordinated or an hexacoordinated form of the $Fe(\text{Schiff base})$. The formed one was found to be in very good agreement with the challenging crystallographic structure obtained by Ueno et. al.¹⁰⁸ However, the later model was an intriguing structure as no experimental information has

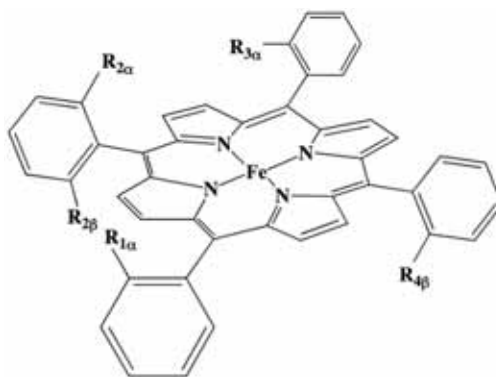
been reported about its existence. Taking into account the experimental information of the system and structural data, we proposed that this model should correspond to the starting specie of the catalytic mechanism.

Although substantial improvements are still required, our integrative approach provided with two model structures for the Fe(Schiff base) \subset cdHO hybrid that could be correlated with the experimental system. We therefore validated our initial hypothesis that the limitations of each individual molecular modelling technique could be overcome by combining them in a multilevel approach. The study performed on this section was published on the *Faraday Discussions* journal.⁷²

III - Binding of porphyrin-like catalysts to monoclonal antibodies: Abzymes

We have successfully benchmarked our protocol reproducing the unusual characteristics of the Fe(Schiff base) \subset cdHO crystallographic structure. With this experience in hand, we decided to focus towards other artificial metalloenzymes, the first of which were the antibodies embedded porphyrin developed by Jean-Pierre Mahy and coworkers.^{137,138} In these systems, monoclonal antibodies were generated to bind porphyrinic cofactors, resulting in a hybrid presenting an oxidative activity. These kind of catalytic antibodies are known as abzymes and in this case their main purpose was to mimic natural peroxidases and other hemoenzymes.

The group of Jean-Pierre Mahy has a large trajectory creating porphyrinic-based homogeneous catalysts as models of natural peroxidases.^{109,111,112,137,139-142} In this case, the synthetic cofactor used to generate the abzyme is an iron-porphyrin with four carboxyphenyl substituents (Fe(ToCPP), Scheme 4.3). The metal center and the porphyrinic ring conferred an oxidative activity while the polar substituents were introduced as potential hbond donors to enhance the binding to the antibody.



Scheme 4.3 - General scheme of the porphyrinic cofactors used to obtain the different abzymes. Depending on the substituents of positions R (-COOH or -H), we have the $\alpha,\alpha,\alpha,\beta$ -Fe(ToCPP), the α,β -Fe(DoCPP), the α,α -Fe(DoCPP) or the Fe(MoCPP).

The monoclonal antibodies against the Fe(ToCPP) cofactor were obtained using the hybridoma methodology. The first step in this technique is to inject the mice with the homogeneous catalyst to stimulate an immune response against it. The lymphocytes (plasma cells producing the target antibody) originated are characterized and the ones producing the antibody with the best affinity for the cofactor are then isolated and fused with cancer plasma cells. This results in a new type of cells, the hybridomas, which are

able to indefinitely yield the desired antibodies. This hybridoma approach led to two different monoclonal antibodies: the 13G10 and the 14H7.^{137,138}

The insertion of the Fe(ToCPP) in both antibodies enhanced the oxidation of 2,2'-azinobis (3-ethylbenzothiazoline-6-sulfonic acid) (ABTS), being 5 to 8 fold more effective than with the isolated cofactor.¹⁰⁹ The protein scaffold provided a great thermostability and protected the cofactor from self-oxidation, allowing more than one thousand turnovers and the use of several hydroperoxides as cosubstrates, including H₂O₂.^{109,110} However, the yields were still far from those obtained in natural hemoenzymes. One of the possible reasons for this low performance could be the lack of a distal residue coordinating the metal, one of the most common features in this kind of enzymes. This was in agreement with the similar binding affinities observed for the Fe(ToCPP) cofactor and its free-iron counterpart (1.4·10⁸ M and 2.6·10⁸ M respectively in 13G10, 1.4·10⁸ M and 2.15·10⁸ M respectively in 14H7).¹³⁸ To overcome this problem, imidazole was added to the reaction with the purpose of emulating a distal histidine chelating the metal. Although the catalytic activity was enhanced at first, imidazole concentrations higher than 50nM ended up inhibiting the artificial metalloenzymes (Table 4.3).¹¹⁰

Cofactor	Without Imidazol			+50 Mm Imidazol		
	K _{cat} (min ⁻¹)	K _M (mM)	K _{cat} /K _M (min ⁻¹ /mM)	K _{cat} (min ⁻¹)	K _M (mM)	K _{cat} /K _M (min ⁻¹ /mM)
Fe(ToCPP)	68 ± 7	37 ± 4	1.8 ± 0.3·10 ³	71 ± 7	8.5 ± 1	8.3 ± 1.5·10 ³
Fe(ToCPP)⊂13G10	109 ± 10	29 ± 3	3.8 ± 0.7·10 ³	32 ± 3	19 ± 2	1.7 ± 0.4·10 ³
α,α-Fe(DoCPP)⊂13G10	32 ± 3	34 ± 3	0.9 ± 0.2·10 ³	152 ± 10	10 ± 1	15.2 ± 2.5·10 ³
α,β-Fe(DoCPP)⊂13G10	16 ± 2	18 ± 2	0.9 ± 0.2·10 ³	96 ± 9	7 ± 1	13.7 ± 2.8·10 ³

Table 4.3 - Kinetics parameters of the isolated Fe(ToCPP) cofactor and three abzymes containing different porphyrinic cofactors

Another cause for the low yields obtained could be the high steric hindrance provided by the volumetric carboxyphenyl substituents. To validate this hypothesis, Mahy and coworkers decided to try other less substituted Fe(ToCPP) catalysts: the α,α- and α,β-Fe(DoCPP) and the Fe(MoCPP) (Scheme 4.3). The three of them were able to effectively bind both the 13G10 and the 14H7 antibody receptors but presented different binding affinities; while in the di-substituted cofactors they were similar as those observed for the Fe(ToCPP) catalyst, in the case of the Fe(MoCPP) there was a 50-fold affinity decrease.¹³⁸ Additionally, as in the Fe(ToCPP)⊂antibody complexes, no chelation of the metal was induced as suggested by the similar affinities observed for the three alternative cofactors and their respective free-iron forms.¹³⁸ The catalytic behavior of these species was enhanced upon incorporation of imidazole and no inhibition was observed at higher

concentrations (Table 4.3).¹¹⁰ Unfortunately, these alternative abzymes still did not reach the high catalytic profiles of natural hemoenzymes.

In view of the low catalytic activity of the abzymes generated, Mahy and coworkers attempted to optimize them using a rational design approach. To gather the necessary structural data for this process, they obtained the X-ray structure of the Fab region^b of both the Fe(ToCPP)C13G10 and the Fe(ToCPP)C14H7 abzymes (Figure 4.12). However, despite many efforts, none of the obtained structures displayed the position of the synthetic cofactor.

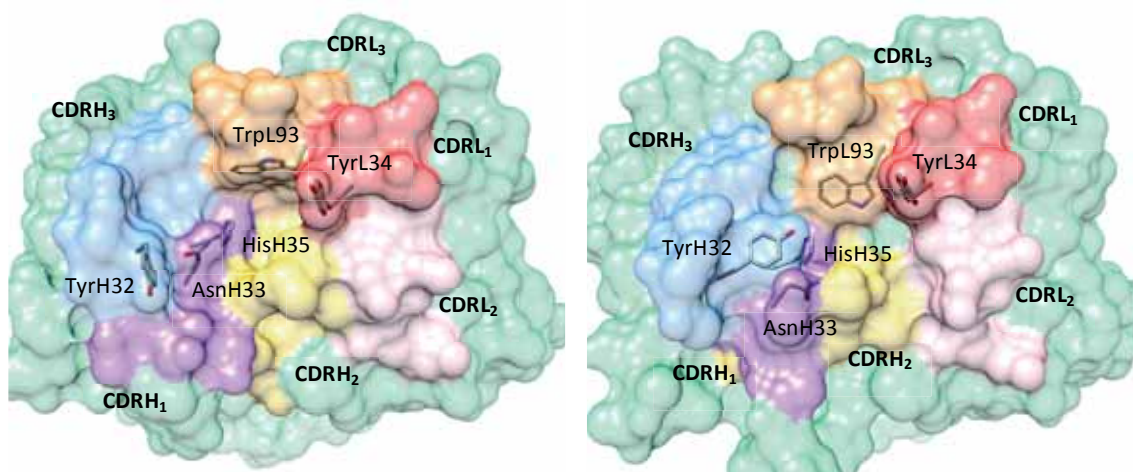


Figure 4.12 - Representation of the binding site of both 13G10 (PDB code 4AMK, **left**) and 14H7 (PDB code 4AT6, **right**). Main CDR regions of both the light and the heavy chain are depicted in different colors. The main residues influencing the binding of the cofactor are displayed.

Both structures presented three different complementarity determining regions (CDR) in both the light (CDRL) and the heavy (CDRH) chains. This regions located at the VH/VL interface is responsible to bind the target hapten, thus it was postulated that the Fe(ToCPP) binding site should be in this area. Even though the CDRs of both antibodies were composed of almost the same residues, there were subtle differences in the configuration they adopted in each scaffold.

^b An antibody is typically composed of two different parts: one variable (change its sequence to match the antigen) and one conserved (determine the overall structure and the identity of the antibody). Both of them are formed by a heavy chain (VH) and a light chain (VL). The fragment antigen binding (Fab) area is located at the interface of these two regions and is the one responsible for the recognition and binding of the hapten.

Knowing the localization of the cofactor is something vital to proceed with the rational optimization of the hybrid. This information allows us to know in which residues of the protein or chemical groups of the cofactor we should focus in the process. To obtain a model structure of the different abzymes designed displaying the position of the synthetic cofactor we applied our *savoir faire* previously gained on the salen \subset cdHO system. The cofactor \subset antibody model structures generated will provide the essential molecular information needed for their rational optimization.

III.I - Computational protocol

Protein-ligand dockings were performed with each different inorganic cofactor into both the 13G10 and the 14H7 antibodies to obtain the model structures of the corresponding abzyme. The structures of the porphyrinic catalysts were obtained using a DFT approach with the density functional B3LYP^{121,122} as implemented in Gaussian09.¹²³ The basis set LANL2DZ¹²⁴ and its associated pseudopotential were used to treat the iron atom while the rest was treated with the 6-31g** basis set.^{125,126} Calculations were carried out in the Fe(III) high spin as it is the usual electronic state of the iron in this kind of systems.

Docking calculations were performed using the program GOLD5.1¹²⁷ and the ChemScore scoring function¹²⁸. For a better representation of the iron in all the docked homogeneous catalysts it was treated using a pseudo-metal atom type able to simulate the capacity of the metal to interact with polar residues. To prevent the docking algorithm of altering the α/β configuration of the carboxyphenyl substituents of the cofactor the bonds linking them to the porphyrinic macro ring were not allowed to rotate. A preliminary set of docking calculations showed that position TyrL34 may have a crucial impact in the binding process, as it could block the entrance of the binding site. Therefore, this residue was allowed flexibility in all the docking simulations using the Dunbrack Rotameric Library.¹²⁰ All structures were prepared as dictated by the GOLD user manual using the molecular visualization UCSF Chimera.¹⁰⁷

III.II - Refinement of the crystallographic structures by protein-ligand dockings

On a first set of docking calculations, the Fe(ToCPP) cofactor was docked into the postulated binding site of both the 13G10 and the 14H7 antibodies. Even though the experimental data suggested no coordination between the iron and the protein scaffold, we still needed a good representation of the metal center to obtain accurate predictions. We therefore took into account some considerations, including: (i) treating the iron with a

pseudo-metal atom type^c and (ii) using a search algorithm^c to identify possible residues able to chelate the metal that may have been overlooked during the docking simulation.

All the lowest energy binding modes generated for the Fe(ToCPP) cofactor were localized at the VL/VH interface of both antibodies (Figure 4.13). However, there were several differences on the orientations obtained depending on the host. In the 13G10 antibody all the lowest energy binding modes had almost the same orientation, with two of the substituents in α,β configuration located at the interface of the two antibody chains and the other α,α solvent exposed. Interestingly, all the solutions presented one carboxyphenil buried deep within the cavity making several hbond interactions with a polar patch constituted by AsnH33, HisH35 and the NH group of GlyH96 and GlyH100a (Figure 4.13). On the other hand, in the 14H7 the cofactor is located in the same area but the lowest energy solutions displayed a wide variety of different orientations with no common pattern. The binding site was narrower than in the 13G10 and no substituent could penetrate deep inside the cavity. As a consequence, the whole cofactor remained mainly exposed to the solvent in all cases, with some hbond interactions between the carboxyphenyls and the surrounding residues (Figure 4.13). Surprisingly, some orientations in the 14H7 antibody presented a possible chelation of the iron by TyrL34, something that has not been observed experimentally.

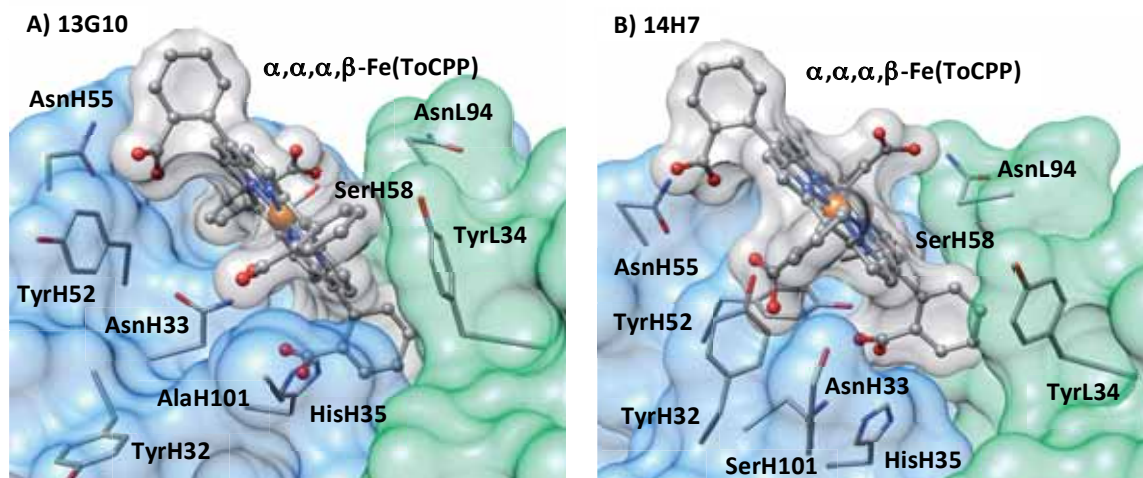


Figure 4.13 - Cartoon representation of the lowest energy solution obtained from the docking of the Fe(ToCPP) cofactor into the binding sites of 13G10 (A) and 14H7 (B) antibodies. Main residues interacting with the inorganic cofactor are displayed in stick representation. The blue surface corresponds to the heavy chain of the antibody, the green to the light chain and the grey one to the docked cofactor.

^c Both implementations belong to our integrative approach and are thoroughly explained in section II.I of this chapter.

The docking binding energies displayed a better affinity between the Fe(ToCPP) cofactor and the 13G10 antibody by almost 10 Score units (Table 4.4). In fact, its binding site had a better complementarity with the inorganic moiety than the 14H7 counterpart, as highlighted by the 100 units difference in the $S_{\text{lipo}}^{\text{d}}$ term. One of the main contributors to that difference is the deep insertion of one of the carboxyphenyl groups in the binding site of the 13G10 structure.

Antibody	Ligand	Score	ΔG (kJ mol ⁻¹)	S_{hbond}	S_{lipo}	S_{clash}
13G10	Fe(ToCPP)	41.5	-49.1	3.4	276.6	3.0
	α,α -Fe(DoCPP)	42.0	-45.3	3.2	248.9	0.9
	α,β -Fe(DoCPP)	40.5	-43.8	2.7	251.1	0.9
	Fe(MoCPP)	42.1	-43.8	2.6	252.5	0.3
14H7	Fe(ToCPP)	31.5	-35.6	3.3	164.6	0.2

Table 4.4 - Analysis of the energetic terms for the lowest energy solutions obtained in the docking of the different porphyrinic cofactors into the two different 13G10 and 14H7 antibody structures.

Using experimental information, Mahy and coworkers proposed a model for both Fe(ToCPP)⊂13G10 and Fe(ToCPP)⊂14H7 hybrids in which two thirds of the porphyrinic cofactor were inserted in the antibody binding site with two carboxyphenyl substituents in α,β position more specifically bound to the protein.¹³⁸ Additionally, they also observed that no chelation of the metal was induced in any of the two antibodies.¹³⁷ All these observations were in agreement with the docking solutions obtained for the Fe(ToCPP) cofactor in the 13G10 antibody. All the lowest energy binding modes displayed almost half of the synthetic catalyst inserted in the antibody with the two carboxyphenyl groups interacting with the protein in α,β configuration. Moreover, none of the solutions presented a chelation of the metal and the search algorithm highlighted that no residue was able to reach the metal. On the contrary, the docking results on the 14H7 antibody did not account for the experimental behavior of the system. In this case, the cofactor remained mainly solvent exposed with no common features regarding the configuration of the substituents and a possible chelation of TyrL34 was observed. Additionally, while the experimental data indicated similar binding affinity of the cofactor for both antibodies,¹³⁷ the docking energies displayed a 10 Score units difference between them being the insertion of the cofactor in the 13G10 much more favorable than in the 14H7.

^d The S_{lipo} term of the docking energetic break down accounts for the hydrophobic interactions calculated during the simulation.

The low correlation between the docking results in the 14H7 antibody and the experimental data suggests that this X-ray structure may belong to a close conformation of the antibody in which the cofactor cannot enter the binding site. As no more structural data on this antibody was available, the model for this hybrid was discarded at this point. On the contrary, the docking solutions on the 13G10 antibody were highly in agreement with the experimental behavior, thus the lowest energy orientation of the cofactor was used to build the corresponding Fe(ToCPP)⊂13G10 model structure.

III.III - Structural basis for the optimization of porphyrinic cofactors

It was postulated that the high steric hindrance provided by the four carboxyphenyl substituents of the Fe(ToCPP) cofactor could negatively affect the reaction. Three other less substituted homogeneous catalysts were therefore synthesized and inserted into both 13G10 and 14H7 antibodies. From them, two were disubstituted (α,α - and α,β -Fe(DoCPP)) and one monosubstituted (Fe(MoCPP)). In agreement with the hypothesis, the hybrids obtained displayed an improvement on the yield and were not inhibited by increasing concentrations of imidazole.^{109,111,137}

To rationalize the observed behavior we aimed to obtain the corresponding model structures. However, the previous results with the Fe(ToCPP) cofactor in the 14H7 antibody highlighted that this structure was not appropriate for such feat as its closed conformation prevented the cofactor to enter the binding site. For this reason, only the 13G10 antibody was considered to generate the corresponding Fe(ToCPP)⊂13G10 model structures.

The docking results obtained for the three less substituted porphyrinic cofactors in the 13G10 antibody were very similar to those of the Fe(ToCPP) in that same scaffold. All the lowest energy modes were located on the VL/VH interface, in the same binding site identified for the original catalyst (Figure 4.14). Additionally, they were very similar to those of this tetrasubstituted cofactor, with one of the carboxyphenyls buried deep within the cavity and interacting with the HisH35-AsnH33 polar patch. The other two substituents in the case of the α,α - and α,β -Fe(DoCPP) catalysts were mainly solvent exposed with some interactions with the polar neighboring residues. Consistently with the experimental observations, both the docking and the search algorithm indicated that no residue in the vicinity of the metal in the three catalysts could chelate it.

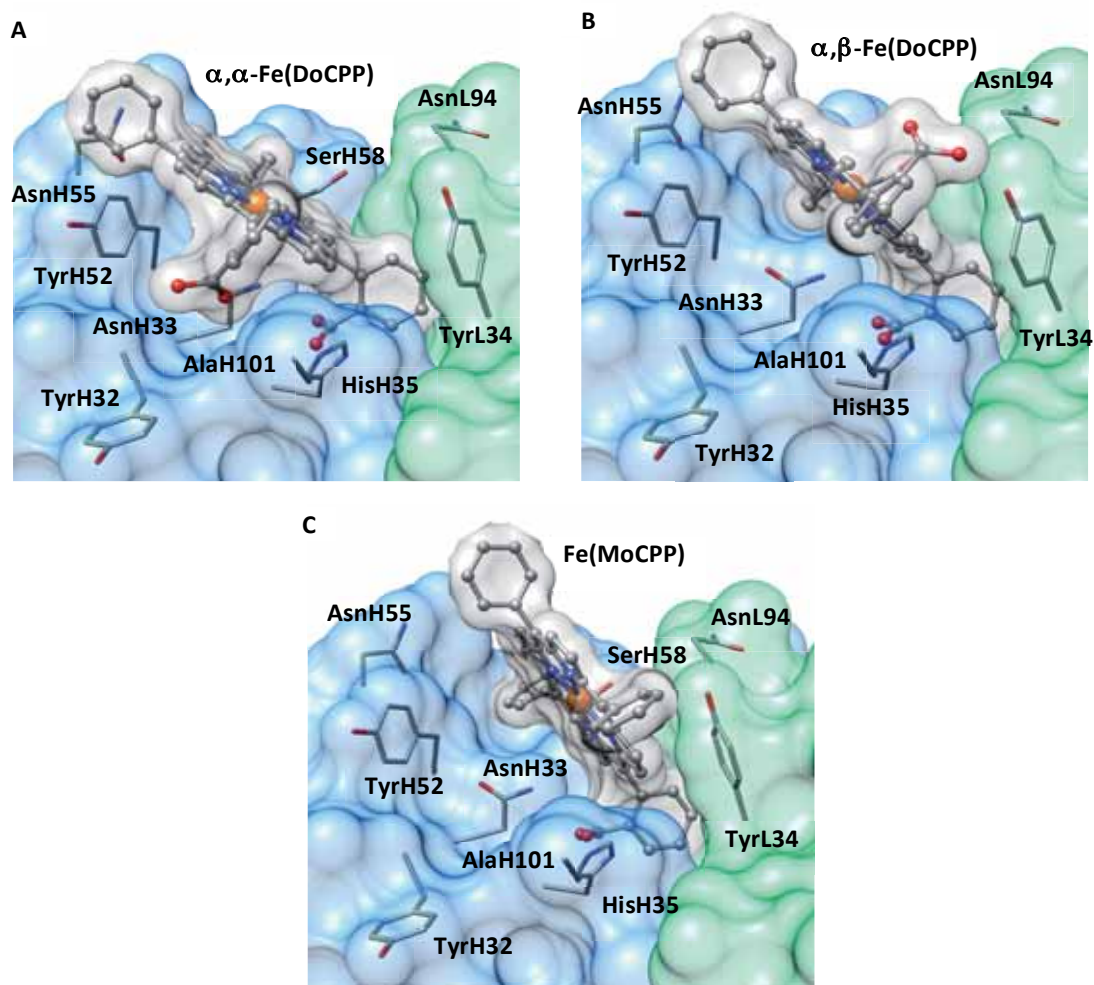


Figure 4.14 - Cartoon representation of the lowest energy solution of the three less substituted porphyrinic cofactors on the 13G10 antibody. α,α -Fe(DoCPP) = **A**, α,β -Fe(DoCPP) = **B**, Fe(MoCPP) = **C**. Main residues involved in the binding are displayed in stick. The heavy and the light chains of the antibody are depicted in blue and green respectively.

The predicted binding energies for the three alternative cofactors were also very similar to those of the Fe(ToCPP)–13G10 system, indicating a good affinity for the 13G10 antibody (Table 4.4). These results were in agreement with the experimental behavior observed for the two disubstituted cofactors since the two of them had similar binding affinities to the Fe(ToCPP).¹¹¹ However, this was not the case for the Fe(MoCPP). This catalyst was observed to have a 50-fold decrease in its binding affinity,¹¹¹ but the predicted energies were similar to the other porphyrinic cofactors.

The docking results for the α,α - and the α,β -Fe(DoCPP) had a good correlation with the experimental data gathered on their corresponding abzymes using the 13G10 antibody as scaffold. Their binding energies were similar to those of the Fe(ToCPP) and the binding

poses were in agreement with the model proposed by Mahy and coworkers with two-thirds of the porphyrinic ring inserted in the antibody host.¹³⁸ However, while in this model they proposed that the antibody should specifically recognize an α,β configuration at the binding pocket, the α,β -Fe(DoCPP) cofactor had only one substituent at the binding site and the other solvent exposed. In fact, the α,α -Fe(DoCPP) cofactor also presented this positioning of the carboxyphenyl groups. An analysis of the Fe(ToCPP) \subset 13G10 model highlighted that, while one of the carboxyphenyls placed on the binding site stabilized the binding by making several interactions with the AsnH33-HisH35 polar patch, the other presented some bad contacts with the surrounding residues. Therefore, the reason why none of the two disubstituted cofactors presented both carboxyphenyl substituents on the binding site was probably to avoid these clashes. This was reflected in the S_{clash} term on the breakdown of the predicted binding energies, which was significantly higher in the Fe(ToCPP) case (Table 4.4).

Regarding the Fe(MoCPP) catalyst, the docking results did not reproduce the 50-fold decrease in the binding affinity of this cofactor in comparison with the other ones. This may be due to some molecular events not taken into account during this simulation, like the induced fit that most likely takes part after the binding of the cofactor. To fully account for them we should perform other molecular modelling techniques, like molecular dynamics. However, this is something far from the scope of this work and the Fe(MoCPP) \subset 13G10 model structures were not further refined.

III.IV - Rationalization of the experimental catalytic profiles

The docking simulations provided with the model structures of different porphyrinic cofactors inserted into the 13G10 antibody. Those belonging to the Fe(ToCPP) \subset 13G10, the α,α - and the α,β -Fe(DoCPP) \subset 13G10 abzymes were in good agreement with the experimental data. However, these models could neither explain the inhibition observed with increasing concentrations of imidazole on the Fe(ToCPP) \subset 13G10 hybrid nor the increasing rate in those same conditions for the abzymes with the two disubstituted cofactors. To explain this behavior protein-ligand dockings were performed with the imidazole to identify where would it stand in the corresponding model structures.

The results from the docking calculations suggested a good affinity between the imidazole molecule and the three abzyme models, with binding energies close to 20 Score units in all cases (Table 4.5). Interestingly, even though the binding modes of the Fe(ToCPP) and the two disubstituted cofactors were very similar, the imidazole preferentially bound to

different faces of the iron depending on the model. The subtle differences in the orientations of the synthetic cofactors resulted in distinct cavities at each side of it, which presented a different degree of complementarity with the imidazole. On the Fe(ToCPP)⊂13G10 model it was located between the porphyrinic cofactor and the light chain of the antibody with one stabilizing hbond with the backbone of AlaL93. On both Fe(DoCPP)⊂13G10 hybrids it bound to the opposite face of the iron, being stabilized by only hydrophobic interactions in the α,β -Fe(DoCPP) case and with a contribution of an additional hbond with one of the carboxylate substituents in the α,α -Fe(DoCPP) (Figure 4.15).

Ligand bound	Score	ΔG (kJ mol ⁻¹)	S_{hbond}	S_{metal}	S_{lipo}	S_{clash}
Fe(ToCPP)	22.6	-23.1	1.0	0.9	76.4	0.5
α,α -Fe(DoCPP)	21.5	-22.9	0.6	0.9	89.0	1.3
α,β -Fe(DoCPP)	19.9	-19.9	0.0	1.0	77.2	0.1

Table 4.5 - Breakdown of the most relevant terms of the scoring for the binding of an imidazole molecule in the different 13G10 abzyme models using the ChemScore scoring function.

The docking of the imidazole indicates that the reactive face of the iron changes depending on the abzyme. This could have a tremendous effect on the catalytic performance as the protein environment is utterly different from one case to the other: one is highly solvent exposed (α,α - and α,β -Fe(DoCPP)⊂13G10) and the other is sheltered by the protein environment (Fe(ToCPP)⊂13G10). Two different events are needed for the reaction mechanism: (i) the formation of the iron(IV)-oxo complex and (ii) the binding of the ABTS substrate. Considering the size of the two possible catalytic pockets, it is more feasible for both the ABTS to reach the iron(IV)-oxo specie if is formed on the solvent exposed face of the iron. As stated by the docking of the imidazole, this side is only accessible in both disubstituted cofactors, as in the Fe(ToCPP) it is occupied by the imidazole (Figure 4.15). These results could shed some light on the different behavior observed upon increasing the concentration of imidazole: an inhibition in the case of the Fe(ToCPP)⊂13G10 and an enhancement in both abzyme α,α - and α,β -Fe(DoCPP)⊂13G10.

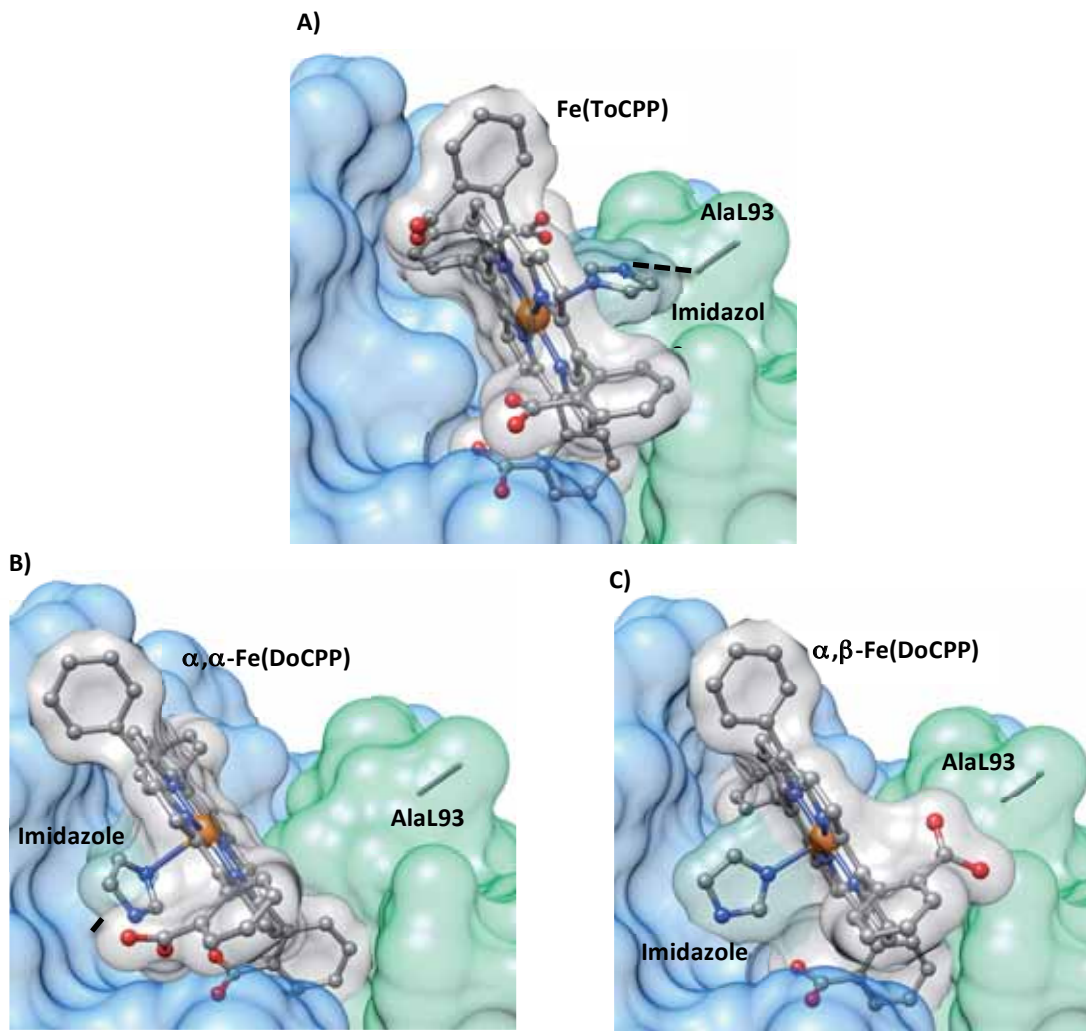


Figure 4.15 - Binding of an imidazole molecule in the model structures of Fe(ToCPP) (A), α,α -Fe(DoCPP) (B) and α,β -Fe(DoCPP) (C). The light chain of the antibody is depicted in green surface, the heavy one in blue. The hbond interactions stabilizing the imidazole binding are represented in black dashed lines.

To form the iron(IV)-oxo complex it is necessary the aid of a residue of the protein in the cleavage of the O-O bond of the H_2O_2 molecule, which in natural peroxidases is normally a histidine or an arginine. The model structures of the three different abzymes displayed two different elements that could assume this role: (i) two polar residues in the vicinity of the iron (AsnH33 and TyrH52 in the Fe(ToCPP) an TyrL34 in the α,α -Fe(DoCPP)) or (ii) one of the carboxyphenyl substituents of the cofactor (in the Fe(ToCPP) and in the α,β -Fe(DoCPP)). The later has also been proposed for an aldoxime dehydration where the acid catalysis is aided by a carboxylate substituent of the porphyrinic catalyst.¹⁴³

III.V - Conclusions

Protein-ligand dockings have been applied to model the structure of four different porphyrinic cofactors bound to two different antibody receptors: 13G10 and 14H7. In the 14H7 case, the narrower binding site prevented an efficient binding of the synthetic cofactors and suggested that the crystallographic data belonged to a closed conformation not suitable to produce a model of the different abzymes. On the contrary, in the 13G10 case all the tested porphyrinic cofactors presented a remarkably good binding affinity where the deep anchoring of one of the carboxylate substituents into the AsnH33/HisH35 polar patch guided the binding process. These results were in agreement with the high affinity observed experimentally for the Fe(ToCPP) and the two Fe(DoCPP) cofactors.

Regarding the catalytic mechanism, the Fe(ToCPP)⊂13G10 and the two Fe(DoCPP)⊂13G10 models showed that the imidazole binds in opposite faces of the iron depending on the model. For the Fe(ToCPP)⊂13G10, the imidazole attaches to the solvent exposed face of the iron, while in both Fe(DoCPP)⊂13G10 structures it prefers the sheltered one. As the solvent exposed side of the iron was postulated to be the reactive face, the location of the imidazole in this area on the Fe(ToCPP) abzyme could explain the inhibition observed at high concentrations. Moreover, our models highlighted different possibilities regarding the cleavage of the O-O bond to form the oxo-iron(IV) specie; it could be aided by either TyrH52, AsnH33, TyrL34 or one of the carboxyphenyl substituents.

The structural data obtained from the docking models can be used to design new optimization strategies. For example, new cofactors could be designed taking into account the high degree of stabilization provided by the Asn33/His35 polar patch or some residues of the binding site could be mutated to induce a chelation of the metal. The study performed on this section is available online in the *Plos One* journal.¹⁴⁴

IV - Decoding biotin-Streptavidin embedded systems

One of the most fruitful strategies in the design of artificial metalloenzymes is the insertion of the homogeneous catalyst inside a biological scaffold using the supramolecular approach.^{43–46,108,145} The foundations of this methodology were laid by Whitesides and coworkers on the late 70's when they created one of the very first artificial metalloenzymes using the Strept(avidin)-biotin technology.⁴⁰ They brilliantly exploit the incredible affinity existing between the Strept(avidin) protein(s) and their natural ligand, the biotin ($K_M \approx 10^{-14}$), which nonetheless is one of the strongest non-covalent interactions known in the natural world.¹⁴⁶ By linking the homogeneous catalyst to this ligand, they ensured the localization of the inorganic moiety inside the binding site of the enzyme – or better said, to the remaining free space of the protein. Nowadays, this system is being extensively used by the group of Thomas R. Ward, which is constantly pushing its limits and have created some of the most efficient artificial metalloenzymes reported so far.^{41,44,45,64}

The Streptavidin (Sav) is a homotetrameric complex that works as a dimer of functional dimers. The binding of two monomers to form a dimer (Sav_{1/2}) results in the formation of two different binding sites that can allocate up to one biotin molecule each one. Each dimer is bound to another dimer to form the tetrameric complex, but in this case there is no relation between the binding sites of the two dimeric units (Figure 4.16). Despite the deep relation existing between the different monomeric units forming the Sav_{1/2}, the binding of the biotin molecules has shown to be non-cooperative.¹⁴⁵

The latest system on which we intended to decode the interactions between an organometallic cofactor and a protein host is precisely one designed by Ward and coworkers using the biotin-streptavidin Trojan horse strategy. In this hybrid a biotinylated Noyori's like catalyst ([Cp*Ir(Biot-*p*-L)Cl]) is inserted inside a Sav scaffold (Figure 4.16), obtaining a hybrid able to reduce cyclic imines.⁴⁵ For its further optimization, an intensive random mutagenesis analysis of the binding site was performed and the results highlighted the dramatic impact that mutations on this area could have in the reaction profile.⁴⁵ In fact, amongst the many different mutants characterized, two of them presented some of the best enantioselectivities reported so far for this kind of artificial metalloenzymes: (i) mutant S112A leading to an (*R*)-salsolidine in 96% *ee* and (ii) mutant S112K leading to an (*S*)-salsolidine in 78% *ee*.^{45,88}

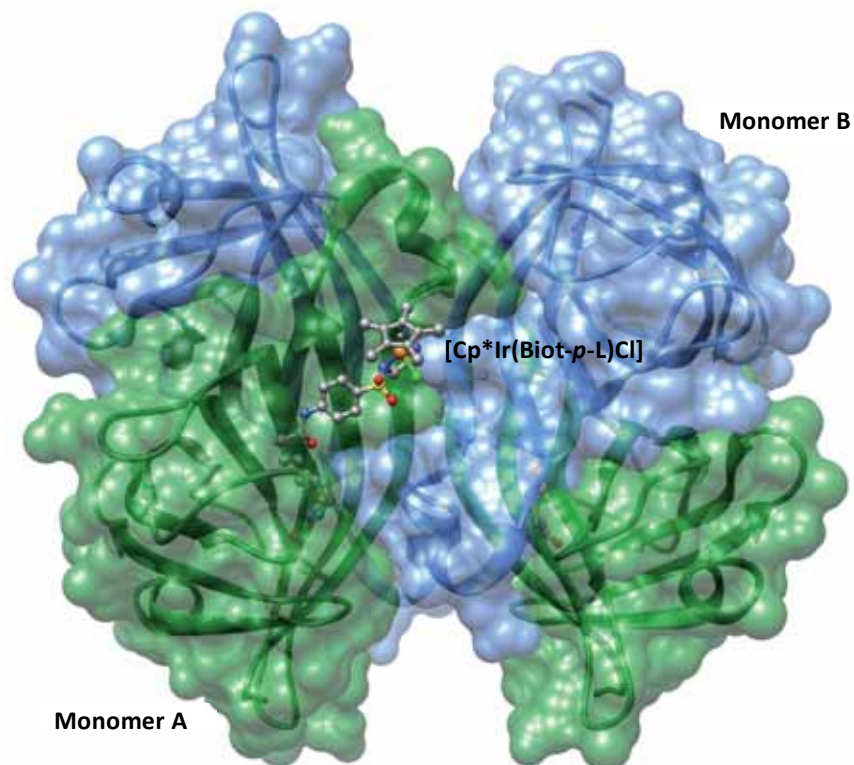


Figure 4.16 - Tetrameric structure of the S112A mutant Streptavidin as determined by X-ray crystallography (PDB code 3PK2). The protein is a dimer of dimers with two close-lying biotin binding sites in each monomer A (green) and B (blue). Only one S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactor is displayed for clarity. Reprinted with permission from (see reference ⁸⁸). Copyright 2014 American Chemical Society.

Ward and coworkers performed a kinetic and an enantioselective study to characterize both the S112A and the S112K Sav mutants bearing the biotinilated catalyst. To rationalize the behavior observed, they attempted to obtain the corresponding crystallographic structure. However, despite many efforts, only the one belonging to the S112A mutant could be obtained and the resulting structure presented several deficiencies.⁴⁵ First, the reported structure corresponded only to one monomeric unit. Without the other monomer of the Sav1/2, this structure leaves too many unanswered questions like: (i) can a second cofactor bind the other monomer?, (ii) how will the already loaded biotinilated catalyst affect the binding of another catalyst on the close-lying biotin binding site? and (iii) where will the substrate stand for the catalysis in each catalytic unit? Additionally, the position of the chloride and the Cp* ligands from the homogeneous catalyst present in the X-ray structure could not be determined, thus the absolute configuration of the metal remained unknown. This was a critical issue as the position of the chloride was essential to determine where the substrate should stand for the catalysis.

In this section, we applied a protein-ligand docking approach to obtain a model structure of both the S112A and the S112K artificial hydrogenases. Together with the experimental behaviors of the two entities, these models are aimed to deal with those questions that the crystallographic structure is unable to answer. Altogether, the information provided will be of vital importance to identify the role of the protein environment in the catalysis in order to further optimize both systems and/or design new ones using the Streptavidin Trojan horse technology.

IV.I - Experimental behavior of the artificial hydrogenases

To characterize both $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112A$ and $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112K$ artificial hydrogenases, Ward and coworkers obtained their Michaelis-Menten kinetics and enantioselectivity profiles at different ratios of the homogeneous catalyst per Sav tetramers (Ir/Sav) (Table 4.6). They started from Ir/Sav = 1, where there should be one homogeneous catalyst inserted in each Sav tetrameric structure, up to Ir/Sav = 4, where all the biotin binding sites should be loaded. Surprisingly, the results obtained were very different depending on the Sav mutant, highlighting the dramatic impact that mutations on position 112 could have.

Entry	Sav mutant	eq [Ir] ^a	ee ^c	$k_{\text{cat}}/[\text{Sav}][\text{min}^{-1}]^{\text{d}}$	$k_{\text{cat}}/[\text{Ir}][\text{min}^{-1}]^{\text{d}}$	$K_{\text{M}}[\text{mM}]^{\text{d}}$
1	No Sav	– ^b	0	4±0.21	8±0.4	120±18
2	S112A	1.0	93	3.5±0.4	14.1±1.7	65±16
3		2.0	92	5.7±0.3	11.4±0.7	74±17
4		3.0	89	5.1±0.3	6.8±0.5	80±19
5		4.0	45	4.3±1.1	4.3±1.1	370±175
6		1.0	–70	0.7±0.2	2.7±0.8	82±47
7	S112K	2.0	–74	1.3±0.0	2.6±0.1	95±9
8		3.0	–76	2.5±0.1	3.3±0.1	119±14
9		4.0	–78	3.3±0.2	3.3±0.2	151±21

Table 4.6 - Data gathered from the kinetic saturation experiments on the isolated cofactor, the $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112A$ and the $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112K$ artificial hydrogenases. ^a $[\text{Cp}^*\text{Ir}(\text{biot-}p\text{-L})\text{Cl}]$ equivalents versus free biotin binding sites; ^bcorresponds to a $[[\text{Cp}^*\text{Ir}(\text{biot-}p\text{-L})\text{Cl}]] = 50 \text{ mM}$, no Sav present; ^cpositive values correspond to (*R*)-salsolidine, negative values correspond to (*S*)-salsolidine; ^derrors represent standard ones derived from triplicate measurements.

On the S112A hydrogenase, the max rate is observed at Ir/Sav=2 (Table 4.6, entry 3), from this point adding more equivalents has a negative effect on the reaction. In fact, at Ir/Sav=4 the rate goes down to its original value of Ir/Sav=1 (Table 4.6, entry 5). Regarding the enantioselectivity, the max ee is already reached at Ir/Sav=1, with a 95% towards the *R*

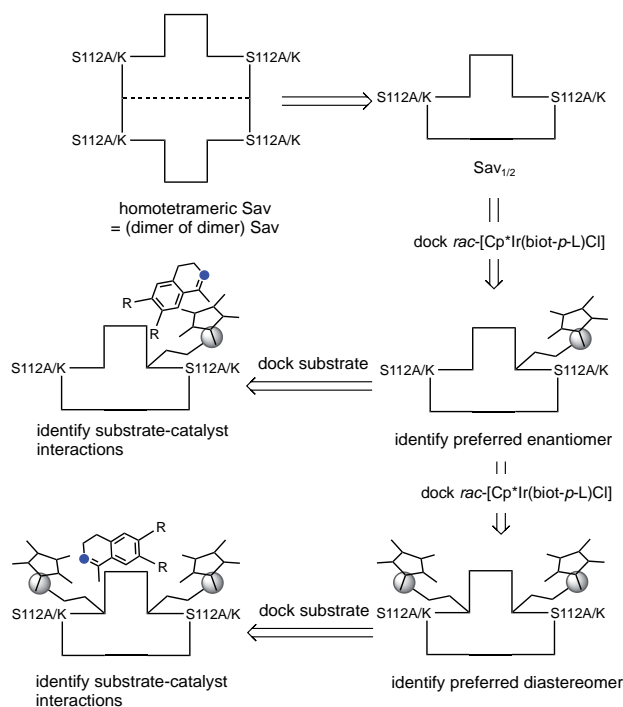
product (Table 4.6, entry 2). Adding more equivalents did not have any relevant effect on it until all the biotin-binding sites are loaded ($Ir/Sav=4$), in which the *ee* sharply decreases up to 45% (Table 4.6, entry 5).

Regarding the S112K ATHase the kinetic results were utterly different. In this case the k_{cat} , the K_M and the *ee* profile remained mainly constant, with no effect as a result of the increasing concentrations of the iridium catalyst (Table 4.6, entry 7-9).

IV.II - Molecular Modelling of the structure of the pre-catalyst [Cp*Ir(Biot-p-L)Cl]₂-S112A and [Cp*Ir(Biot-p-L)Cl]₂-S112K

A two-step protein-ligand docking protocol (Scheme 4.4) was applied to rationalize the enantioselectivity and kinetic profiles of the two different artificial hydrogenases (Scheme 4.4). First, both R_{Ir} - and S_{Ir} -metal cofactors were docked into the Sav dimer ($Sav_{1/2}$) to determine if the specific absolute configuration of the metal could have an impact on the binding affinity of the cofactor to the receptor. The lowest energy solutions for both the R_{Ir} and S_{Ir} metal moieties were used to generate the corresponding S112A or S112K hydrogenases, unless the binding energies displayed a clear preference for one of the two metal forms. In this case only the $Sav_{1/2}$ model structure containing this cofactor was generated. A second biotinilated cofactor in either R_{Ir} or S_{Ir} configuration was then docked into those models to determine if the presence of the first cofactor could influence the binding of a second one. The lowest energy solutions for the R_{Ir} and S_{Ir} metal moieties were used to generate the corresponding S112A or S112K fully loaded $Sav_{1/2}$ model structure. Again, if the docking highlighted a preferential binding for one of the two metal configurations, only this was selected to generate the model.

To identify where the substrate would stand for the catalysis it was docked into every generated model structure. We docked one substrate per each cofactor present on the structure to localize how it would approach each different catalytic unit (Scheme 4.4).



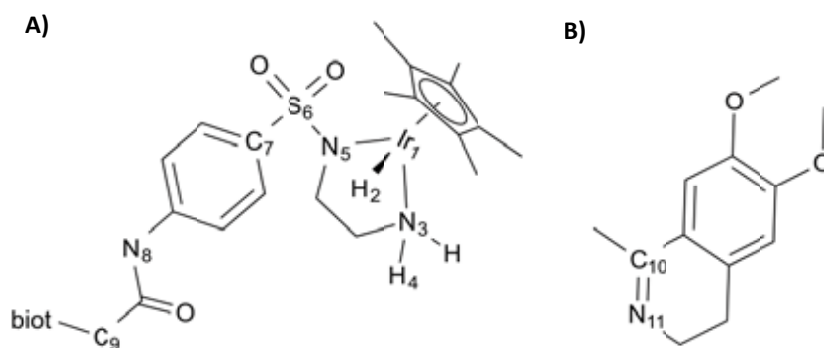
Scheme 4.4 - Docking strategy to identify (i) the absolute configuration at the metal center, (ii) the number of cofactors per Sav_{1/2} and (iii) the localization of the prochiral substrate.

IV.III - Computational Protocol

A first set of QM minimizations was performed to optimize the inorganic moieties considering both the *R* and *S* absolute configurations at the metal. Only the atoms of its first coordination sphere up to the N_{amide} atom (N₈) of the sulfonamide group were considered, while the remaining part was replaced with a methyl group (Scheme 4.5). Calculations were performed within the Kohn-Sham approach to Density Functional Theory (DFT) using the PBE functional^{147,148} as implemented in Gaussian09.¹²³ The basis set Def2-TZVPP¹⁴⁹ and its associated pseudo-potential were used for the iridium and the 6-31G*¹²⁵ basis set for the rest of the atoms. All optimizations were carried out in water as solvent using a polarizable continuum model (CPCM).^{150–153}

Protein-ligand dockings were performed with the minimized catalyst into the dimeric X-ray structure of the [Cp*Ir(Biot-*p*-L)Cl]cS112A hybrid (PDB code 3PK2).⁴⁵ Due to the particularly high affinity between the biotin and the Sav protein (KD≈10⁻¹⁴ M), we postulated that the binding of this ligand should be tightly enough as to not present any kind of flexibility. Therefore, this region of the biotinilated catalyst (up to the N₈ atom, Scheme 4.5) was maintained in the receptor structure and used to perform a covalent docking. The terminal methyl group of the minimized catalyst was substituted for an -NH₂

to regenerate the N₈ atom and both this nitrogen and the N₈ of the biotinic anchoring present in the receptor were selected as the anchoring atoms needed for this covalent approach. Regarding the S112K mutant, no crystallographic data was available for the [Cp*Ir(Biot-*p*-L)Cl]⊂S112K hybrid, thus the S112A structure was *in silico* mutated replacing the Ala₁₁₂ on both subunits for the most probable Lys rotamer using the Dunbrack Rotameric library.¹²⁰ In this case, the biotin part of the catalysts was also maintained in the receptor to perform the same covalent scheme as in the S112A mutant. On the S112A case no flexibility was introduced during the docking as the receptor was the X-ray structure of that same mutant. On the S112K case, due to the lack of structural information, the surrounding residues Lys121_{A,B}, Leu124_{A,B}, and Lys112_{A,B} of both monomers and Leu110 of the corresponding monomer where the cofactor is being docked were allowed flexibility using the Dunbrack Rotameric library.¹²⁰



Scheme 4.5 - Schematic representations of the [Cp*Ir(Biot-*p*-L)Cl] cofactor (A) and the dihydroisoquinoline substrate (B). The biotin anchoring region of the catalyst up to the N₈ atom have been substituted for a methyl group in the QM minimizations.

The iridium center in the homogeneous catalyst has its first coordination sphere completed, thus it is impossible for it to directly participate in the binding process. Subsequently, no additional parameters to model this metal are required to reproduce a possible chelation by any residue of the protein.

In a published study by Petr and coworkers, they reported that the ATH of imines should undergo a two-step mechanism, in which the imine is first protonated in the medium and then the hydride is transferred from the iridium.¹⁵⁴ We performed a similar QM study on the ATH of imines and also arrived at this conclusion.^e Consequently, in the docking analysis of where the substrate would stand for the catalysis we used its protonated form. The chloride atom of the [Cp*Ir(Biot-*p*-L)Cl] cofactor was changed for the corresponding hydride to give consistency with the reaction mechanism. The distance Ir-H⁻ was set to 1.6

^e For more information about this study please refer to section III in Chapter 5

Å as calculated using QM approaches. A small harmonic restraint of 5.0 kJ mol⁻¹ with a distance between 2.5 Å and 3.5 Å was added between the hydride atom and the reactive C_{amine} (C₁₀, Scheme 4.5) of the substrate to ensure a binding orientation consistent with the hydride transfer.

All the dockings carried out in this study (i.e. the ones involving the minimized cofactor and the dihydroisoquinoline substrate) were performed using the program GOLD5.1¹²⁷ and the ChemScore scoring function.^{128,155}

IV.IV - Identification of the preferred metal configuration for the S112A and S112K mutants

Both the *R*_{Ir}- and the *S*_{Ir}-[Cp*Ir(Biot-*p*-L)Cl] catalysts were docked into the two Sav_{1/2} mutant receptors (S112A and S112K) to measure the impact of the metal configuration in the binding process. The structural data gathered from the docking models should also give some information about the differences observed experimentally between both [Cp*Ir(Biot-*p*-L)H]⊂S112A and [Cp*Ir(Biot-*p*-L)H]⊂S112K hybrids.

IV.IV.I - Rationalization of the binding process in the S112A mutant

The docking of the *S*_{Ir}-[Cp*Ir(Biot-*p*-L)Cl] catalyst into the S112A Sav_{1/2} led to two similar low-energy orientations (*S*_{Ir1} and *S*_{Ir2}). The two of them had their Cp* ligand solvent exposed and their chloride pointing at the interface of the two monomers. The main difference observed between the two structures was a little twist of the sulfonamide group, resulting in a turn of the piano-stool towards the center of the cavity in the *S*_{Ir2} orientation. The predicted binding energies were similar for both binding modes with 44.4 and 45.2 Score units for the *S*_{Ir1} and the *S*_{Ir2} respectively (Table 4.7). In both cases there was a good agreement with the crystallographic structure as highlighted by the low RMSD obtained in the superposition; 1.4 Å for the *S*_{Ir1} and 1.0 Å for the *S*_{Ir2} (Figure 4.17).

Host Protein	Cofactor	Score	ΔG (kJ mol ⁻¹)	<i>S</i> _{hbond}	<i>S</i> _{lipo}	<i>S</i> _{clash}
S112A	<i>S</i> _{Ir1}	44.4	-55.0	5.3	314.6	0.4
	<i>S</i> _{Ir2}	45.2	-54.8	5.3	311.7	1.2
	<i>R</i> _{Ir}	43.2	-50.4	4.0	314.0	0.5
S112K	<i>S</i> _{Ir}	51.5	-60.0	6.0	336.2	0.7
	<i>R</i> _{Ir}	58.6	-66.8	7.1	365.1	0.3

Table 4.7 - Breakdown of the docking binding energies for the [Cp*Ir(Biot-*p*-L)Cl] inserted into the two different Sav_{1/2} mutants S112A and S112K.

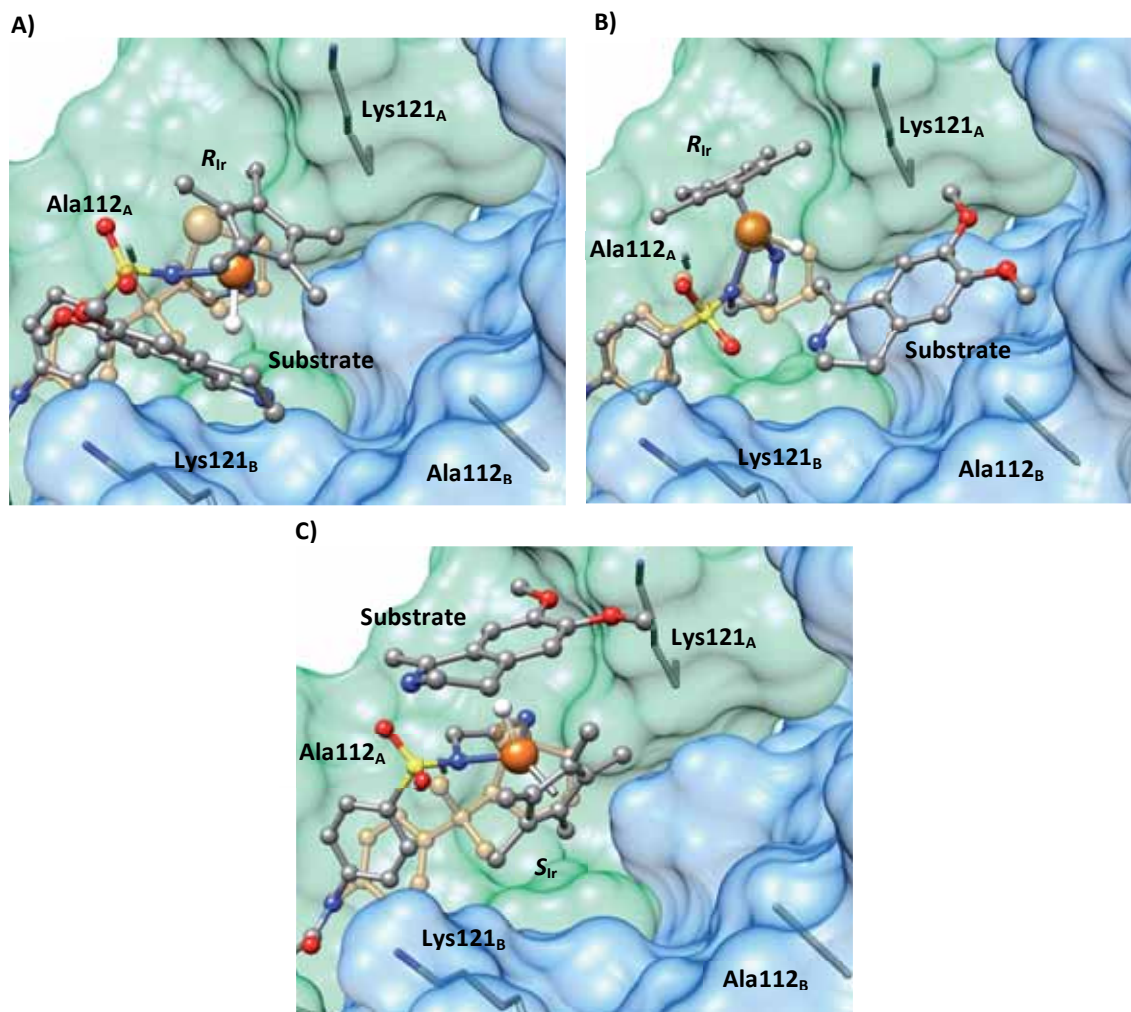


Figure 4.17 - Close-up view of the lowest energy solutions of $[Cp^*Ir(Biot-p-L)Cl] \subset S112A$ Sav with the dihydroisoquinole substrate. The chloride was substituted by a hydride to be consistent with the substrate docking. Panel **A**: ($R_{Ir,1}$); Panel **B**: ($R_{Ir,2}$); Panel **C**: (S_{Ir}). The most relevant aminoacids in the binding process are depicted in stick. For comparison, the X-ray determined position of the $[Cp^*Ir(Biot-p-L)Cl] \subset S112A$ hybrid is represented in ghost orange.

Only one low energy orientation was obtained in the docking of the R_{Ir} metal cofactor. This binding mode was completely opposite compared with the two obtained for the S_{Ir} metal: the Cp^* was looking at the interface of the two monomers and the chloride was solvent exposed. The predicted energy for this orientation was of 43.2 Score units, two units lower than the S_{Ir} counterparts. Moreover, the RMSD obtained from the superposition with the catalyst embedded S112A X-ray structure resulted in 2.2 Å, far higher than those obtained with the S_{Ir} orientations (Figure 4.17).

Even though the docking and the RMSD results with the S112A mutant suggested a slightly better preference for the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalyst, the differences observed between the R_{Ir} and S_{Ir} cofactors were still not conclusive enough. To definitely settle if there was an influence of the absolute metal configuration on the binding of the inorganic moiety, the salsolidine substrate was docked into the model structures containing the $S_{Ir,1}$, the $S_{Ir,2}$ or the R_{Ir} catalysts. Even if the binding energies were similar, those models in which the binding of the substrate was not consistent with the high *ee* observed experimentally (i.e. the binding is strongly influenced by the protein environment) should be discarded.

In the experimental conditions, the chloride of the [Cp*Ir(Biot-*p*-L)Cl] catalyst is replaced for a hydride, which is later transferred to the substrate to produce either the *R* or the *S* amine product. To give consistency with this mechanism we also changed the chloride for a hydride in the three different S112A ATHase models to obtain the corresponding $R_{Ir,1}$ -, $R_{Ir,2}$ - and S_{Ir} -[Cp*Ir(Biot-*p*-L)H]⊂S112A^f models.

In both the $R_{Ir,1}$ - and the $R_{Ir,2}$ -[Cp*Ir(Biot-*p*-L)H]⊂S112A structures, the predicted binding energies for the cyclic imine were about 20 Score units, indicating a good complementarity in both cases (Table 4.8). Interestingly, the binding in both hybrids displayed the substrate deep inserted inside the adjacent empty biotin-binding site (Figure 4.17). Regarding the S_{Ir} -[Cp*Ir(Biot-*p*-L)H]⊂S112A model the lowest binding energy was of 12.4 Score units, far lower than in the R_{Ir} models (Table 4.8). Indeed, the lowest energy binding mode of the substrate was heavy solvent exposed lying at the surface of the catalyst. The only stabilizing interactions were an hbond between the -OMe group and Lys121_A and a few hydrophobic interactions with the loop between Ala112_A and Ser122_A (Figure 4.17). In all the dockings on the three different catalyst-embedded S112A Sav_{1/2} models the constraint between the reactive carbon and the hydride was mainly respected, thus indicating that in these particular cases the hydride transfer did not need a major change of the conformation of the system (Figure 4.17).

^f It is important to notice that, upon substitution of the chloride for the hydride, the order of the ligands is altered and thus following the Cahn-Ingold-Prelog rules the absolute configuration of the metal changes in consequence. Therefore, the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] corresponds to the R_{Ir} -[Cp*Ir(Biot-*p*-L)H] and viceversa.

Host Protein	Score	ΔG (kJ mol ⁻¹)	S_{hbond}	S_{lipo}	S_{clash}	S_{con}
$R_{\text{Ir},1}$ -[Cp*Ir(Biot- <i>p</i> -L)H]⊂S112A	17.0	-18.3	0.9	109.2	0.5	0.3
$R_{\text{Ir},2}$ -[Cp*Ir(Biot- <i>p</i> -L)H]⊂S112A	20.0	-20.3	1.0	123.2	0.2	0.0
S_{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H]⊂S112A	12.4	-13.8	0.9	71.5	1.0	0.3
R_{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H]⊂S112K	5.7	-20.8	0.9	131.6	8.7	6.3
S_{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H]⊂S112K	14.5	-16.1	1.6	71.2	1.2	0.0

Table 4.8 - Breakdown of the docking binding energies for the dihydroisoquinoline substrate on the different structure models generated for the Sav_{1/2} dimer with only one cofactor loaded.

The high enantioselectivity observed for the mono-loaded S112A hydrogenase suggests that the protein environment have a high impact on the catalytic process (Table 4.6, entry 2). From our calculations, this behavior is consistent only in those Sav_{1/2} models presenting one S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalyst. In the two different docking orientations obtained for this cofactor the dihydroisoquinoline substrate is deep inserted at the interface of the two monomers, making strong hydrophobic interactions with the surrounding residues. On the other hand, the model with the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] moiety presented the substrate highly solvent exposed with a minor role of the protein scaffold in the binding event.

The following statements can be concluded from our docking analysis of the S112A ATHase: (i) there was a slightly preference for the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalyst, (ii) this cofactor displayed the best overlap with the [Cp*Ir(Biot-*p*-L)Cl]⊂S112A X-ray structure and (iii) only in this case the binding of the dihydroisoquinoline substrate was in agreement with the high *ee* observed experimentally. Altogether, these results strongly suggested that the S112A empty dimer had a preferential binding for the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] inorganic moiety.

The docking results of the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalyst inside the S112A Sav scaffold highlighted two different low energy orientations: $S_{\text{Ir},1}$ and $S_{\text{Ir},2}$. Given the low energetic differences between them (less than one Score unit, Table 4.7) and the better overlap of the $S_{\text{Ir},2}$ orientation with the X-ray structure, we selected this later binding mode to generate the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]⊂S112A model for the following docking analysis.

IV.IV.II - Rationalization of the binding process in the S112K mutant

Protein-ligand dockings of both the R_{Ir} - and the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactors in the S112K Sav_{1/2} led to lowest energy solutions with divergent binding modes. The R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalyst displayed its chloride widely exposed to the solvent and its Cp* located at the interface of the two monomers while in the S_{Ir} form the position of these

ligands was inverted: the Cp* was solvent exposed and the chloride pointing at the interface between the two monomers (Figure 4.18). The binding energies were also fairly different, with 51.5 Score units for the S_{Ir} cofactor and 58.6 for the R_{Ir} one (Table 4.7). This 7 Score units difference was the result of a better complementarity between the scaffold and the R_{Ir} moiety (S_{lipo} term in Table 4.7) and a wider hbond network (S_{hbond} term in Table 4.7).

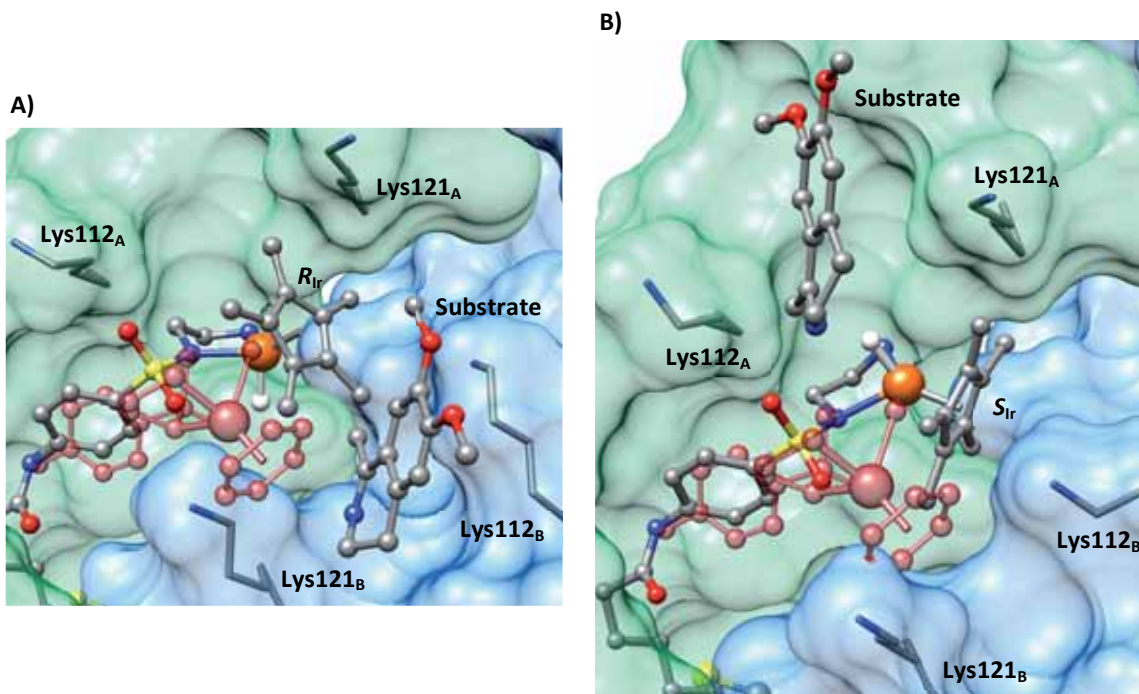


Figure 4.18 - Close-up view of the lowest energy solutions of R_{Ir} - and S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]⊂S112K Sav with the dihydroisoquinoline substrate. The chloride was substituted by a hydride to be consistent with the docking of the substrate. Panel **A**: (R_{Ir}); Panel **B**: (S_{Ir}). The most relevant aminoacids in the binding process are depicted in stick. For comparison, the X-ray determined position of the (R_{Ru})-[C₆H₆Ru(Biot-*p*-L)Cl]⊂S112K hybrid is represented in ghost red.

The difference in energy between the two metal configurations is significant enough to suggest that the S112K Sav_{1/2} has a preferential binding for the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactor. However, from a catalytic point of view, the presence of the substrate and its complementarity with the protein could still influence on the relative orientation of the cofactor. To study such effect, we docked the isoquinoline substrate in the S_{Ir} - and R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]⊂S112K models. Once again, to be consistent with the experimental conditions, the chloride was replaced by a hydride, thus generating the corresponding R_{Ir} - and S_{Ir} -[Cp*Ir(Biot-*p*-L)H]⊂S112K activated forms of the biocomplex.

The docking results for the salsolidine substrate in the two activated forms of the S112K ATHase were completely different. In the S_{Ir^-} -[Cp*Ir(Biot-*p*-L)H]⊂S112K case, the substrate was located in a cleft between Lys121_A and Lys112_A and could satisfy the imposed distance restraint between the C_{imine} and the hydride (Figure 4.18). A part from the hydrophobic interactions with surrounding residues of the cavity, the binding was further stabilized by some hbonds, one between the N_{imine} and the sulfone group of the catalysts and another between the -OMe group of the substrate and Asn118_A (Figure 4.18). All these interactions were translated in a good complementarity between the substrate and the cofactor-protein complex, with a predicted binding energy of 14,5 Score units (Table 4.8). On the other hand the docking of the substrate in the R_{Ir^-} -[Cp*Ir(Biot-*p*-L)H]⊂S112K model suggested that this specie should be inactive. In this case, the hydride was facing the interface of the two monomers and there was no space available for the substrate to bind (Figure 4.18). As a result, the predicted binding energies in this case were fairly lower than in the S_{Ir^-} activated form of the S112K ATHase, being 5.7 Score units (Table 4.8). In fact, there were numerous bad contacts between the substrate and the surrounding atoms (S_{clash} term, Table 4.8). Moreover, the docking could not satisfy the imposed restraint and none of the obtained orientations was consistent with the catalysis (ΔE_{con} term, Table 4.8).

The docking results for the binding of the substrate suggested that only the binding of an R_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl] catalyst should lead to a catalytically competent Sav_{1/2}. The hydride is too sheltered in the S_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl] cofactor for the substrate to reach it. Therefore, the fairly high *ee* and activity rate observed experimentally with one Ir equivalent bound to the S112K Sav (Table 4.6, entry 6) could only be explained if the first catalyst to bind the Sav_{1/2} has an *R* absolute configuration at the iridium.

Two main conclusions could be extracted from the docking analysis of the [Cp*Ir(Biot-*p*-L)Cl] catalysts into the S112K Sav_{1/2}. First, we predicted that the R_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl] homogeneous catalyst should have a better complementarity with the S112K scaffold than the S_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl] counterpart (up to 7 Score units, Table 4.7). Secondly, only the docking of the substrate in the activated form of the R_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl]⊂S112K model was consistent with the experimental observations. Given these results, we can conclude that the S112K Sav mutant should preferentially bind the R_{Ir^-} -[Cp*Ir(Biot-*p*-L)Cl] form of the catalyst.

IV.V - Structural consequences of increasing the Ir/Sav ratio in [Cp*Ir(Biot-*p*-L)Cl]₂ S112A and S112K

Up to this point, two different model structures have been generated: (i) the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂S112A and (ii) the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂S112K. Although these models have shed some light on the molecular behavior of the S112A and S112K ATHases respectively, they could not account for the observations upon increasing the Ir/Sav ratio up to four equivalents (Table 4.6). For this purpose the models of the fully-loaded S112A and S112K Sav proteins are required. It is interesting to notice that we do not need to recreate the whole tetrameric structure. The Sav complex is a dimer of functional dimers that work independently from each other, thus generating a model of one of the dimers with the two biotin-binding sites occupied by the homogeneous catalyst should be enough.

To generate the models of the fully-loaded Sav_{1/2} both the S_{Ir} - and the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalysts were considered to measure the impact of the metal absolute configuration in the binding process. All the dockings were performed in the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂S112A and R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂S112K ATHase models.

IV.V.I - Structural insights on the S112A fully loaded dimer

The docking results in the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂S112A model displayed that both the R_{Ir} - and the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalysts adopted similar orientations, but their respective Cp* and chloride ligands positions were inverted. The S_{Ir} cofactor had its chloride facing the adjacent catalyst, while its Cp* was solvent exposed. Contrary, in the R_{Ir} moiety the Cp* was facing the other cofactor and the chloride was solvent exposed (Figure 4.19). Consistent with these similar binding modes, the predicted binding energies were also close, being 43.1 and 43.9 Score units for the S_{Ir} and R_{Ir} metal forms respectively (Table 4.9).

The small binding energy difference predicted between the R_{Ir} - and the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactors (less than 1 kJ mol⁻¹, Table 4.9) suggested that the metal configuration had no impact on the binding of the second homogeneous catalyst to the S112A Sav mutant. To further validate this assumption we docked the salsolidine substrate in the activated forms of the (S_{Ir}, S_{Ir})- and the (S_{Ir}, R_{Ir})-[Cp*Ir(Biot-*p*-L)Cl]₂S112A models ((R_{Ir}, R_{Ir})- and (R_{Ir}, S_{Ir})-[Cp*Ir(Biot-*p*-L)H]₂S112A, respectively) to analyze whether it could reach or not the hydride in the two available catalytic units. For this purpose two different salsolidine

substrates were docked, each one with a restraint between the C_{imine} and one of the two hydrides available.

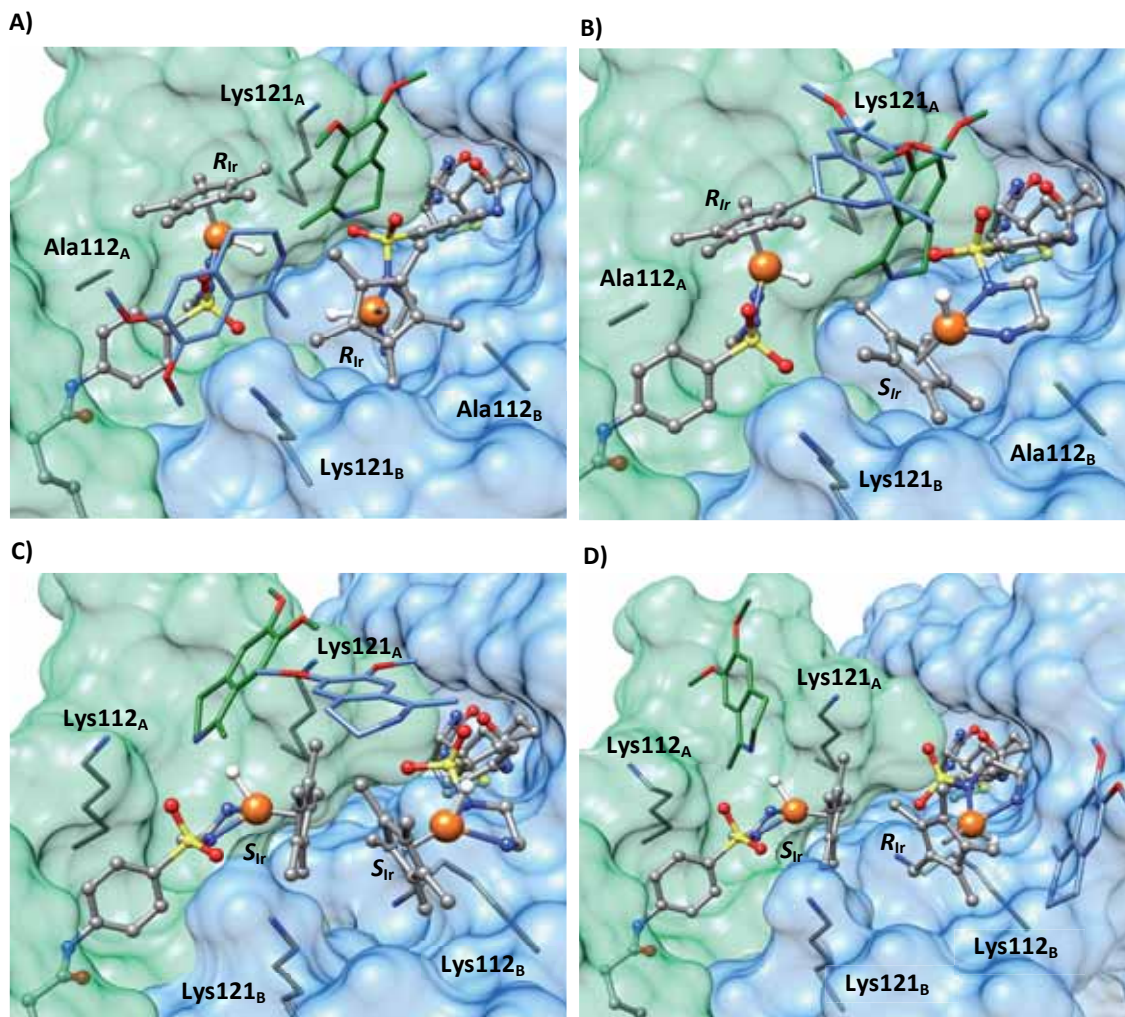


Figure 4.19 - Representation of the models of a fully-loaded dimer of the Sav tetramer. Monomer A is depicted in green, monomer B in blue. The chlorides have been substituted to the corresponding hydrides to give consistency with the docking of the substrate. **Top:** $(R_{\text{Ir}}, R_{\text{Ir}})$ - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{H}]_2<S112A$ (A) $(R_{\text{Ir}}, S_{\text{Ir}})$ - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{H}]_2<S112A$ (B). **Bottom:** $(S_{\text{Ir}}, S_{\text{Ir}})$ - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{H}]_2<S112K$ (C) $(S_{\text{Ir}}, R_{\text{Ir}})$ - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{H}]_2<S112K$ (D). Green substrates were docked with a restraint with the cofactor in monomer A, blue substrates with the one in monomer B.

Host Protein	Cofactor	Score	ΔG (kJ mol ⁻¹)	S_{nbond}	S_{metal}	S_{lipo}	S_{clash}
S_{Ir} - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]<S112A$	S_{Ir}	43.1	-54.2	5.4	0.0	306.7	1.5
	R_{Ir}	43.9	-51.8	4.2	0.0	319.9	0.3
R_{Ir} - $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]<S112K$	S_{Ir}	42.5	-52.2	4.3	0.0	319.1	0.5
	R_{Ir}	44.4	-52.7	3.7	0.0	341.6	0.4

Table 4.9 - Breakdown of the docking binding energies for the docking of the second cofactor in already loaded Sav_{1/2}.

The binding of the substrate in the fully loaded S112A Sav_{1/2} indicated that it could only reach the hydride of the S_{Ir}-[Cp*Ir(Biot-*p*-L)H] cofactor in the R_{Ir},S_{Ir}-[Cp*Ir(Biot-*p*-L)H]₂⊂S112A model. In this case the predicted binding energies were 14.0 Score units, suggesting a good complementarity between the biometallic hybrid and the substrate (Table 4.10). The S_{Ir}-[Cp*Ir(Biot-*p*-L)H] catalyst had the hydride exposed to the solvent, so the substrate was able to bind at the surface of the catalyst in a consistent way with the hydride transfer. Although in this binding mode the substrate was highly exposed to the solvent, it was stabilized by some hbond contacts between the -OMe group and Lys121_A or Lys121_B depending on the docking solution (Figure 4.19). In this same Sav_{1/2} unit the substrate was unable to reach the hydride of the R_{Ir}-[Cp*Ir(Biot-*p*-L)H] cofactor because the Cp* ligand of the adjacent S_{Ir} catalyst is too close and is blocking the path. Even though the predicted binding affinities were also fairly high (up to 14.3 Score units, Table 4.10), the S_{clash} term was far too high (up to 13 units, Table 4.10), indicating that there were too many close contacts with the surrounding atoms.

Host Protein	Catalytic moiety	Score	ΔG (kJ mol ⁻¹)	S _{hbond}	S _{lipo}	S _{clash}	S _{cons}
R _{Ir} ,R _{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H] ₂ ⊂S112A	First R _{Ir}	3.4	-22.5	0.0	170.8	18.0	0.1
	Second R _{Ir}	4.4	-17.7	0.0	130.1	11.5	1.9
R _{Ir} ,S _{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H] ₂ ⊂S112A	R _{Ir}	14.3	-27.2	0.0	211.3	13.0	0.0
	S _{Ir}	14.0	-18.0	1.4	-93.3	3.1	0.8
S _{Ir} ,S _{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H] ₂ ⊂S112K	First S _{Ir}	15.0	-15.2	1.0	80.7	0.2	0.0
	Second S _{Ir}	9.3	-11.4	0.0	76.3	1.6	0.5
S _{Ir} ,R _{Ir} -[Cp*Ir(Biot- <i>p</i> -L)H] ₂ ⊂S112K	S _{Ir}	14.5	-16.0	1.5	71.1	1.0	0.0
	R _{Ir}	12.9	-15.8	0.9	86.5	1.7	1.1

Table 4.10 - Breakdown of the docking binding energies for the dihydroisoquinoline substrate onto the fully loaded Sav_{1/2}.

The docking of the substrate on the R_{Ir},R_{Ir}-[Cp*Ir(Biot-*p*-L)H]₂⊂S112A model suggested that this combination of cofactors should lead to an inactive dimeric unit. The two hydrides are facing each other at a close distance (4.7 Å), leaving no space between them for the substrate to bind (Figure 4.19). This was consistent with the low scores obtained (around 4 Score units, Table 4.10) and the high S_{clash} term in the docking of the two different salsolidine substrates (more than 10 units in both cases, Table 4.10).

The docking analysis of the binding of the substrate provided valuable molecular information about the S_{Ir},S_{Ir}- and S_{Ir},R_{Ir}-[Cp*Ir(Biot-*p*-L)Cl]₂⊂S112A artificial hydrogenases. Summarizing, the dimers with a combination of two S_{Ir}-[Cp*Ir(Biot-*p*-L)Cl] catalysts should

not present any catalytic activity while in the S_{Ir}, R_{Ir} combination only the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] moiety should be active.

The similar predicted binding energies for the S_{Ir} - and R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalysts suggested that, upon the binding of the second cofactor, we could have a mixture of both S_{Ir}, S_{Ir} and S_{Ir}, R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂⊂S112A species. Additionally, the docking analysis of the substrate clearly indicated that only the S_{Ir}, R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂⊂S112A Sav_{1/2} should retain any catalytic activity, while the S_{Ir}, S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂⊂S112A dimers should be inactive. Altogether, these results can be correlated with the decrease in both the rate and the *ee* observed upon the addition of the fourth iridium equivalent to the S112A Sav tetramer. First, the decay in the rate of the reaction could be explained by the apparition of the S_{Ir}, S_{Ir} - inactive dimeric units (Table 4.6, entry 5). Second, the decrease on the enantioselectivity could be due to the solvent exposed binding of the substrate in the active catalyst (R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]) of the S_{Ir}, R_{Ir} - dimers (Table 4.6, entry 5). In comparison, the role of the protein in this event is much lower than when we only had one S_{Ir} - catalyst bound in the dimer.

IV.V.II - Structural insights on the S112K fully loaded dimer

Protein-ligand dockings with both the R_{Ir} and the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalysts on the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]⊂S112K model displayed similar orientations. The two of them presented their Cp* ligand next to Cp* of the adjacent cofactor, but they differed in the position of the chloride. In the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactor it was next to Lys112_B and facing the interface of the two monomeric units, while in the R_{Ir} form it was solvent exposed (Figure 4.19). The predicted binding energies were also similar, being 42.5 and 44.5 Score units for the S_{Ir} - and the R_{Ir} -metal configurations respectively (Table 4.9).

Even though the difference in energy between the S_{Ir} - and the R_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] catalysts indicated a slightly preference for the R_{Ir} one (2 Score units, Table 4.9) it was not high enough as to ascertain that there was indeed a measurable impact of the metal configuration in the binding of the iridium catalyst. To further analyze how the metal configuration influences the resulting S112K Sav_{1/2} unit we docked two different dihydroisoquinoline substrates into the activated forms of both the R_{Ir}, R_{Ir} - and the R_{Ir}, S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl]₂⊂S112K models (the S_{Ir}, S_{Ir} - and the R_{Ir}, S_{Ir} -[Cp*Ir(Biot-*p*-L)H]₂⊂S112K forms respectively). A little restraint was added between the C_{imine} of each substrate and one of the two available hydrides to ensure a binding of the substrate consistent with the hydride transfer.

The binding of the substrate revealed a dramatic impact of the metal configuration in the S112K ATHases. In fact, only the $S_{Ir}, S_{Ir}-[Cp^*Ir(Biot-p-L)H]_2$ combination of catalysts led to a fully-active $Sav_{1/2}$. In this case the substrate could reach the two available hydrides without any steric hindrances and without a major rearrangement of the protein environment (Figure 4.19). Additionally, the binding of the substrate in both monomeric units was stabilized by some hydrogen bonds with the surrounding lysine residues. Accordingly, the predicted binding energies were close to 14 Score units for both catalytic units (Table 4.10). On the other hand, dockings on the $S_{Ir}, R_{Ir}-[Cp^*Ir(Biot-p-L)H]_2 \subset S112K$ model revealed that the substrate could not reach the hydride of the $R_{Ir}-[Cp^*Ir(Biot-p-L)H]$ catalyst in a catalytically consistent orientation. Even though the predicted binding energies were fairly high (12.9 Score units, Table 4.10), this hydride was too sheltered by the protein environment and the C_{imine} could not satisfy the imposed restraint (Figure 4.19 and S_{cons} term in Table 4.10).

The docking analysis suggested a slightly preference of the $R_{Ir}-[Cp^*Ir(Biot-p-L)Cl] \subset S112K$ loaded ATHase for the $R_{Ir}-[Cp^*Ir(Biot-p-L)Cl]$ cofactor. Additionally, only in this R_{Ir}, R_{Ir} -combination of iridium moieties the resulting $Sav_{1/2}$ maintained the catalytic activity in both monomers. The substrate could reach the two hydrides of the $Sav_{1/2}$, thus the presence of the two cofactors did not have a negative impact on the catalysis. Interestingly, this is precisely what is observed experimentally; adding more Ir equivalents resulted in an increase of the catalytic rate while the *ee* did not experience any noticeable change (Table 4.6, entry 6-9). We therefore concluded that the second cofactor to bind the $R_{Ir}-[Cp^*Ir(Biot-p-L)Cl] \subset S112K$ hybrid should have also an *R* configuration at the iridium center.

IV.VI - Rationalizing the kinetic and enantioselective profiles of S112A and S112K ATHase.

Our docking analysis have provided with several model structures of both the S112A and S112K artificial hydrogenases with either one or two homogeneous catalysts bound. Even though these models corresponded to the Sav dimeric units, they could offer valuable molecular information of the behavior of the full tetrameric complex. This inference was possible because the Sav tetramer is composed by a dimer of functional dimers that work independently from each other. Taking into account both the experimental information and the model structures we proposed an explanation of the behavior observed for both S112A and S112K hydrogenases (Table 4.11).

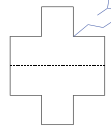
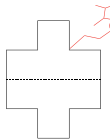
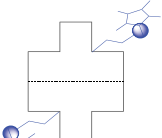
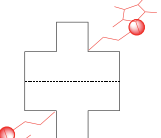
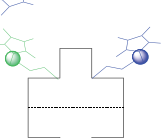
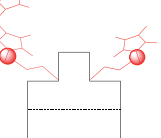
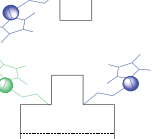
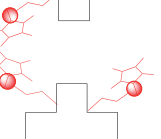
S112A			S112K		
Equivalents of cofactor	<i>ee</i>	Rate	Equivalents of cofactor	<i>ee</i>	Rate
	Max.	1/2		Max.	1/4
	Max.	Max.		Max.	1/2
	Max.	Decreases		Max.	3/4
	Decreases	Decreases		Max.	Max.

Table 4.11 - Rationalization of the experimental behavior with the docking results in function of the Ir/Sav ratio. Blue, red and green cofactors symbolize (S_{Ir}) -[Cp*Ir(Biot-*p*-L)Cl], (R_{Ir}) -[Cp*Ir(Biot-*p*-L)Cl] and (rac) -[Cp*Ir(Biot-*p*-L)Cl] respectively.

IV.VI.I - [Cp*Ir(Biot-*p*-L)Cl]_cS112A

i) The kinetic data indicated that, upon the binding of the first cofactor, the system reaches nearly half the maximum rate and highest enantioselectivity (Table 4.6, entry 2). The docking results indicated that in this case the Sav_{1/2} should preferentially bind the S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] cofactor. This binding was consistent with the experimental behavior; with only one cofactor bound per Sav tetramer it could only account for half the max rate of the hybrid, as only one of the two dimers can be functional. Additionally, the binding of the substrate in this case was identified to be in the adjacent biotin-binding site of the monomer where the cofactor was. The deep insertion of the substrate indicated a high impact of the protein environment, which is in agreement with the high enantioselectivity obtained in this case.

ii) The theoretical results suggested that the second cofactor to bind should be another S_{Ir} -[Cp*Ir(Biot-*p*-L)Cl] homogeneous catalyst in an empty Sav_{1/2}. This way the possible clashes that could appear from the binding of the cofactor in the same dimeric unit are

minimized. Now we have two identical functional dimers that could account for the double on the rate observed experimentally, while maintaining the enantioselectivity (Table 4.6, entry 3).

iii) The binding of a third cofactor must occur in a Sav_{1/2} already containing one S_{Ir}-[Cp*Ir(Biot-p-L)Cl] catalyst. The docking study suggested that the second cofactor to bind the dimer could have the iridium center in either *R* or *S* absolute configuration. Additionally, only the binding of a R_{Ir} cofactor should lead to an active dimeric unit where the substrate should bind in a solvent exposed region of the dimer. The presence of some inactive Sav_{1/2} as a result of the binding of two S_{Ir} homogeneous catalysts should explain the decrease on the rate observed experimentally when adding the third iridium equivalent (Table 4.6, entry 4). Additionally, the scarce role that protein environment has in the binding of the substrate should cause an erosion in the enantioselectivity. However, at ratio Ir/Sav=3 the enantioselectivity did not suffer any important decrease. We hypothesized that, as there was still one Sav_{1/2} with only one S_{Ir} cofactor bound, it could outstrip the performance of the R_{Ir} metal moiety of the fully loaded dimer, thus compensating the loss of enantioselectivity from the S_{Ir},R_{Ir} dimeric units (Table 4.6, entry 4).

iv) The binding of the fourth cofactor led to a fully loaded Sav tetramer. The situation previously described in case iii) for the Sav_{1/2} loaded with two cofactors is now applied in this case. Only the dimers with a R_{Ir}-[Cp*Ir(Biot-p-L)Cl] cofactor may retain some activity with an eroded enantioselectivity as a result of the highly solvent exposed binding of the substrate. These observations could explain the dramatic decrease observed in both the rate and the enantioselectivity (Table 4.6, entry 5)

IV.VI.II - [Cp*Ir(Biot-p-L)Cl]cS112K

The experimental behavior of the [Cp*Ir(Biot-p-L)Cl]cS112K hydrogenase indicate that increasing concentrations of iridium equivalents result in an enhancement of the rate, while the enantioselectivity is always maintained at its maximum value (Table 4.6, entries 6-9). The docking study highlights that this is only possible if all the biotin-binding sites are loaded with the R_{Ir}-[Cp*Ir(Biot-p-L)Cl] catalyst. In this situation, the catalysis takes place in independent regions of the Sav_{1/2}, thus creating a new functional monomer each time we increase the Ir/Sav ratio.

IV.VII - Conclusions

Protein-ligand docking results suggested an opposite enantioselectivity depending on the Sav mutant. This is an important event that determines the enantioselectivity of the system; the metal absolute configuration governs the region of the protein where the substrate should stand for the catalysis, which in turn can by-and-large dictate the enantiomeric product of the reaction. This situation can be described as an “*induced lock-and-key*” fit where all the elements involved are entwined and have an important effect on the final product.

It is pleasing to see that, despite the qualitative nature of the *in silico* modelling approaches, the results obtained can be correlated with the experimental data to provide with an accurate model of the catalytic systems. The several model structures generated of both artificial hydrogenases helped to identify the role of each part of the system in the catalytic process. This information can now be taken into account for a further optimization of the system through a rational design, either by making chemical changes on the inorganic moiety and/or by creating new mutants of the protein.

The results discussed in this section of Chapter 4 have been submitted for publication to the *Chemical Science* journal.¹⁵⁶



**Catalytic Mechanisms of
Artificial Metalloenzymes**

I - General introduction

In the previous chapter we developed and applied several computational methodologies to predict the molecular features for the binding of inorganic moieties to protein hosts. This step is fundamental for the development of artificial metalloenzymes, but is just the tip of the iceberg in the design and optimization of these hybrids. Indeed, there are several other molecular events that need to be taken into account for a successful rational design/optimization of the biometallic hybrid. Especially important are those related to the catalytic process (i.e the binding of the substrate and the stabilization of the transition state) as they dictate the rate and the origin of the region-, chemio- and enantioselectivity. These properties are some of the holy grails in synthetic chemistry and one of the conditions *sine qua non* for the application of artificial enzymes in the industry. However, characterizing those processes is a challenging task due to the vast conformational space available as a result of its size.

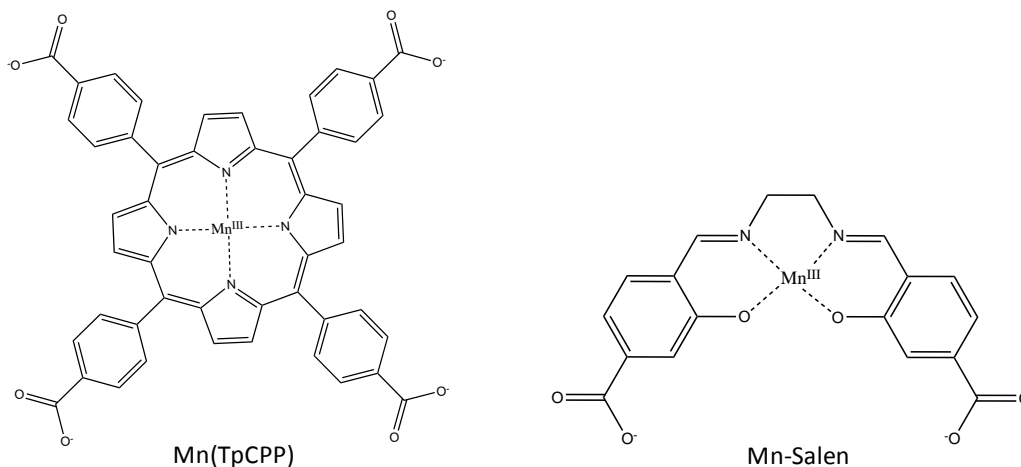
In this part of the thesis we use computational tools to study the catalytic mechanisms of different artificial metalloenzymes. We mainly aim at identifying as accurately as possible the location of the substrate in a catalytic consistent manner and eventually identify the most likely transition states. Considering the high level of complexity, here we studied only systems for which high *ee* have been characterized and their structural information is available. Indeed, at this stage our work is highly conditioned by the information provided from the *wet* side. If this information is scarce or not accurate enough (specially at the structural level) it makes no sense in getting into high level calculations since the error at the beginning of the work could be substantial. Only for well-defined systems such level of accuracy could be reached, minimizing the risks of obtaining artefactual/erroneous results in the workflow and becoming. Furthermore, this kind of systems could also become the perfect test case for our theoretical studies. In this chapter we try to address the enantioselective problem from both the protein-ligand docking and QM/MM perspectives. Special mention needs the latter case, in which we needed to develop a novel integrative strategy that mixes up several MM and QM approaches. This way we could explore large conformational spaces while maintaining an accurate electronic representation, something vital to quantitatively unveil the origin of the enantioselectivity. This part of the work is expected to provide with fundamental basis for future computer aided design/optimization of artificial enzymes and biocatalysts.

II - Enantioselective study through binding modes analysis

The group of Jean-Pierre Mahy is specialized in generating artificial metalloenzymes by inserting porphyrinic-like catalysts into biological scaffolds. Following this line of research, they recently reported two different porphyrinic homogeneous catalysts able to perform the epoxidation of several different styrenes.⁴⁷ However, only one of them could be efficiently inserted into a xylanase scaffold to obtain an artificial metalloenzyme. This hybrid maintained the oxidative activity but presented substantially divergent catalytic behavior depending on the nature of the styrene substrate. Since no crystallographic structure could be produced on any of the attempted biometallic complexes, molecular modelling techniques were involved to catch the molecular grounds of their architecture and rationalize the experimental observations.

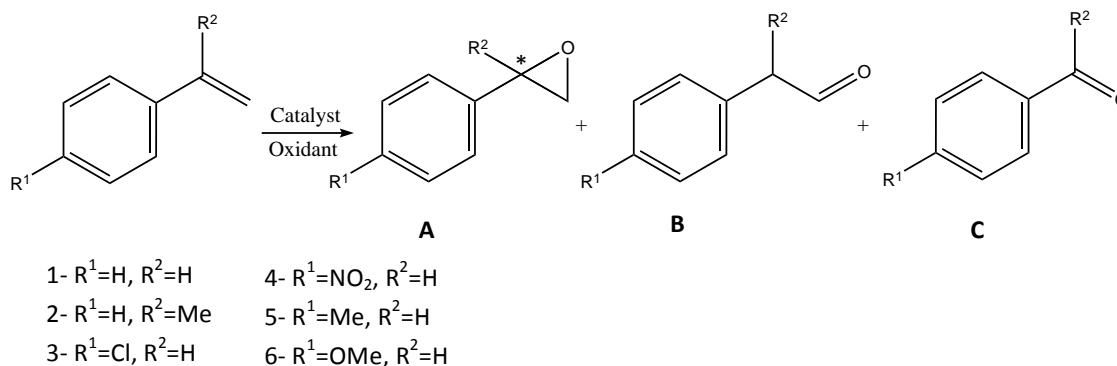
Some natural metalloenzymes can perform the enantioselective epoxidation of styrenes by H₂O₂ (*ee* up to 99%), however their yield is still not high enough for its use on industrial processes (49% at max).^{157,158} These enzymes are generally heme binding proteins that requires the metal cofactor for the functionalization of the oxygen and, in some cases, an external coenzyme (i.e. NADP(H)⁺) to provide with the necessary electrons. Given the biometallic nature of these complexes, it is natural to think of artificial metalloenzymes as a way to obtain highly efficient mimic hybrids. The first attempts along this road have led to conversion yield and enantioselective levels still far from those of natural metalloenzymes.^{159–161} Mahy and coworkers have recently reported two systems obtained through the insertion of a N,N'-ethylenebis(2-hydroxybenzylimine)-5,5'-dicarboxylic acid Mn(III) (Mn-salen) and a Mn(III)-meso-tetrakis(p-carboxyphenyl)porphyrin (Mn(TpCPP) (Scheme 5.1) homogeneous catalysts inside a xylanase 10A from *Streptomyces lividans* (Xln10A).

The enzyme Xln10A can be a great candidate to act as the host protein in the design of artificial metalloenzymes as: (i) it is available at low cost and in large quantities and (ii) it has numerous positively charged residues in the active site to aid in the binding of anionic cofactors¹³⁹ (like the ones used in this work) and offers a significantly large binding pocket to accommodate homogeneous catalysts with a wide range of different sizes. Indeed, the potential of Xln10A has already been tested with the insertion of a Fe(TpCPP) homogeneous catalyst, producing an efficient oxidative hybrid.^{139,140}



Scheme 5.1 - Representation of the two cofactors inserted into the XIn10A scaffold to obtain the artificial metalloenzyme

The experiments pointed out that XIn10A had a better affinity for the Mn(TpCPP) cofactor, with a K_D 20 times better than the Mn-salen counterpart ($1.5\mu\text{M}$ vs $31\mu\text{M}$, respectively). Because of this difference in stability, the catalytic assays were only performed with the Mn(TpCPP)XIn10A hybrid. These assays revealed that the epoxidation of styrenes by this hybrid resulted in three oxidized products^{162–164}: (i) two different enantiomers of the phenyloxirane (**A**, Scheme 5.2), (ii) the phenylacetaldehyde (**B**, Scheme 5.2) and (iii) the benzaldehyde (**C**, Scheme 5.2). This chemical variety represents an excellent illustration of the chemoselectivity (the three different products) and enantioselectivity (the two different enantiomers for product **A**) in the epoxidation of alkenes that artificial metalloenzymes are able to produce. In their experimental study, Mahy and coworkers tested this system with a regular styrene and a set of different substituted ones: β -methylstyrene, *p*-nitrostyrene, *p*-chlorostyrene, *p*-methylstyrene, *p*-methoxystyrene (entries 1-6, Scheme 5.2).



Scheme 5.2 - Representation of the different styrenes used to test the Mn(TpCPP)XIn10A artificial metalloenzyme. The * marks the asymmetric carbon obtained in the reaction.

The catalytic results of the various styrenes oxidized by Mn(TpCPP), either isolated or inserted inside Xln10A, are depicted in Table 5.1.⁴⁷ In general, the protein scaffold enhanced both the total yield and the turn over number (TON) of the reaction from 0.79 to 1.05 min⁻¹. The chemoselectivity of the reaction was also slightly improved, increasing it towards the epoxide product (between 1% and 8%). Of the most important results, a relative low but clearly recognizable enantioselective profile was observed. As of the most promising results, for the *p*-methylstyrene the *ee* increased to 25% towards the *S* product and for the *p*-methoxystyrene it hit up to 80% *R*. In order to carry out further optimizations of the system, information on how the enzyme works at a molecular level is fundamental. Since an initial model of the enzyme-cofactor able to provide this information is missing, a two-steps docking procedure was performed to generate one. The first one deals with the binding of the cofactor to the host and the second with the binding of the substrate to the most plausible models obtained for the cofactor<Xln10A system.

Substrate	Catalyst				
	Mn(TpCPP)		Mn(TpCPP)<Xln10A		
	Total yield (%)	Epoxide yield (%)	Total yield (%)	Epoxide yield (%)	<i>ee</i> (%)
Styrene	18	11.5	26	17	8.5 (<i>S</i>)
β -methylstyrene	16.4	10	22	14	9 (<i>S</i>)
<i>p</i> -nitrostyrene	1.8	2	3	3	3 (<i>S</i>)
<i>p</i> -chlorostyrene	3.5	3.5	7	7	5 (<i>S</i>)
<i>p</i> -methylstyrene	18	15	39.5	23	25 (<i>S</i>)
<i>p</i> -methoxystyrene	39	11.5	49.5	16	80 (<i>R</i>)

Table 5.1 - Results for the oxidation of different styrenes by Mn(TpCPP) homogeneous catalyst alone and by the Mn(TpCPP)<Xln10A hybrid.

II.1 - Computational methods

Two different sets of docking calculations were used to study the mechanistic insights of the artificial epoxidase. In the first one we applied the protocol presented in Chapter IV to obtain a model of the different cofactor<Xln10A hybrids that could explain the major binding affinity observed for the Mn(TpCPP) catalyst. In the second set, the lowest-energy model obtained for the artificial metalloenzyme is used to dock the different styrene substrates to identify the source of the different *ee* profiles identified.

Prior to the docking simulations, we minimized both Mn(TpCPP) and Mn-salen cofactors, as well as the different styrenes. These molecules were prepared using DFT methods with the functional B3LYP^{121,122} as implemented in Gaussian09.¹²³ The LANL2DZ¹²⁴ basis set and its associated pseudopotential was used for the treatment of the Mn center, while the other atoms were treated with the 6-31g** basis set.^{125,126}

To obtain the model structures of the two bioinorganic complexes we applied the same docking procedure described in section II of Chapter IV. First, the minimized cofactors were docked into the different structures available for the Xln10A on the PDB (1V0K¹⁶⁵, 1OD8¹⁶⁶ and 1E0W)¹⁶⁷ for a better treatment of the flexibility of the protein. Additionally, for an accurate representation of the metal we used a pseudo-metal atom type able to mimic the capacity of metals to interact with polar residues.⁷² Finally, an *in-house* developed algorithm to search for some chelating atoms overlooked during the docking simulation was applied. The dockings of the cofactor, as well as those carried out later with the substrate, were performed using the program GOLD5.0¹²⁷ and the ChemScore¹²⁸ scoring function.

It was postulated that the reaction should proceed through a highly reactive Mn(V)=O intermediate, which should donate the oxygen to form the corresponding epoxide.¹⁶⁸ Therefore, for a more realistic docking of the styrene substrates, we model this complex in the theoretical structure of the Mn(TpCPP)⊂Xln10A hybrid. The oxygen atom was bonded to the Mn with a bond length of 1.746 Å as suggested by a previous study on similar Fe(IV)-oxo species.¹⁶⁹ In this specific case, we can take advantage of one of the main assets of using GOLD; it has some well-parameterized atom types to represent metal ions when they are located in the receptor. Therefore, instead of using our pseudo-metal atom type, we used the default metal-atom types from GOLD, and more specifically the ones reported in a previous study of the Cytochrome P450.¹⁷⁰ The binding cavity was limited to a 7 Å sphere radius around the metal atom and surrounding residues His85, Ser86 and Gln87 were allowed flexibility using the Dunbrack Rotameric library¹²⁰. Special mention needs residue Arg139. A preliminary docking analysis of the substrate into the Mn(TpCPP)⊂Xln10A model structure highlighted that this residue could either block or allow the entrance of the substrate in the artificial binding site. We therefore substituted this residue for a rotameric position that allowed the entrance of the substrate. All docked structures were prepared for the docking as specified in the GOLD user manual using the USCF Chimera¹⁰⁷ interface.

Ideally, we should perform QM/MM calculations on the system and identify the transition states leading to both enantiomeric products to quantify the *ee* of the system. However, the limited trust we can have on our starting model and the high resource consuming-ratio of this approach discouraged us from doing this attempt.

II.II - Insertion of the cofactors inside the xylanase scaffold

Both the Mn-salen and the Mn(TpCPP) homogeneous catalysts were docked into the different xylanase scaffolds (1VOK¹⁶⁵, 1OD8¹⁶⁶ and 1EOW¹⁶⁷) to obtain the models of the corresponding biometallic hybrids. The 1VOK and the 1OD8 structures corresponded to the *holo* form of the xylanase; both of them presented different forms of a xylobio substrate in their respective binding sites, resulting in different rotameric positions of the neighboring residues. This is a reflection of the high promiscuity of this protein, which can accept several different substrates thanks to its particularly flexibility of its active site. The 1EOW structure corresponded to the *apo* form and presented a slightly wider binding site in comparison with the two *holo* conformations.

The highly similar docking results obtained for the binding of the two inorganic moieties in the three alternative Xln10A structures suggested that these conformational changes had no impact on the binding process. We therefore performed all the docking analysis of both Mn-salen and Mn(TpCPP) using only the 1VOK Xln10A structure.

The protein-ligand dockings indicated a good complementarity between both the Mn-salen and the Mn(TpCPP) catalysts with the 1VOK protein host. However, the almost 6 Score units difference in the predicted binding energies suggested a far better complementarity with the Mn(TpCPP) cofactor (30.9 vs 25.0 Score units respectively, Table 5.2). The Mn-salen was predicted to be far more solvent exposed with very few interactions with the protein scaffold generally concentrating on the carboxyphenyl substituents and residues Asn45 and Arg139 (**A**, Figure 5.1). On the other hand, the lowest energy modes of the Mn(TpCPP) displayed the inorganic catalyst deep inserted inside the cavity, with two of the carboxyphenyl substituents making strong hbond contacts with different residues of the cavity (Lys48, Asn173, Gln94 and Trp266, **B** in Figure 5.1).

Ligand	Score	ΔG (kJ mol ⁻¹)	S_{Hbond}	S_{lipo}	S_{clash}
Mn(TpCPP)	30.9	-31.6	3.4	124.8	0.4
Mn-salen	25.0	-25.5	2.0	114.7	0.1

Table 5.2 - Breakdown of the most relevant ChemScore terms of the docking of the two homogeneous catalysts into the Xln10A 1V0K structure. Only the most representative terms are displayed.

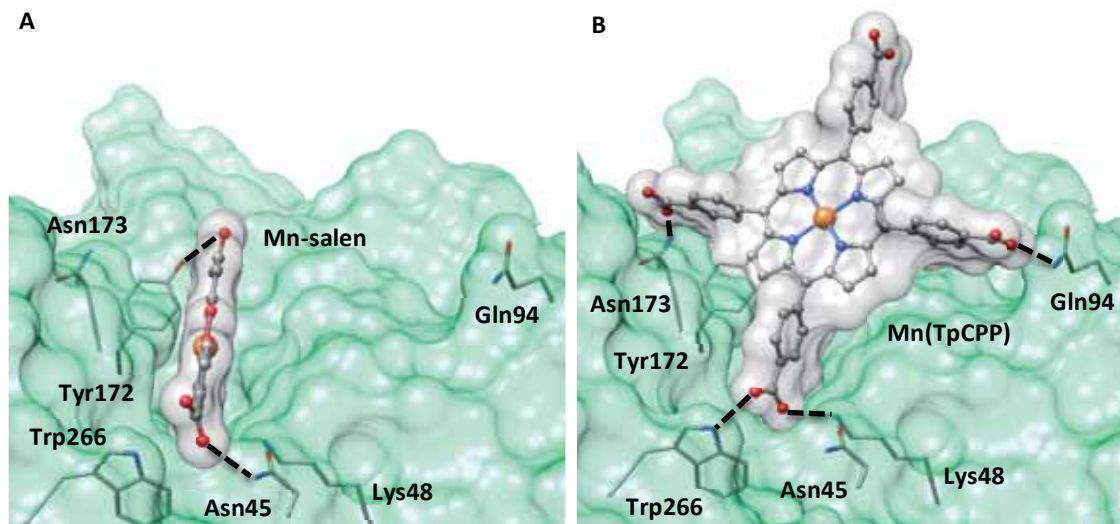


Figure 5.1 - Cartoon representation of the lowest energy binding modes obtained for Mn(TpCPP) (A) and Mn-salen (B) in the Xln10A 1V0K structure. Hbond contacts are represented in black dashed lines

Nor the docking algorithm neither our *in house* search protocol managed to find any residue able to chelate the metal in none of the two tested inorganic moieties. This observation was consistent with the absence of absorption spectra observed experimentally¹³⁹ and suggested that no major changes in the coordination sphere of the metal occur during binding. From the technical point of view, this means that subsequent QM/MM calculations were not necessary to refine the model. However, this also means that the first coordination sphere of the Mn was incomplete. Based on experimental knowledge on Mn-porphyrin systems, it was hypothesized that some water molecules from the media should complete it. We do not expect any relevant conformational change as a result of coordinating this water to the Mn, thus no QM/MM calculations were performed. The water was placed as a distal ligand at 1.746 Å, following an octahedral pattern configuration of the metal.

The docking results indicated the binding of the Mn(TpCPP) catalyst into the Xln10A scaffold to be 5 Score units more stable than the Mn-salen counterpart. Although these docking binding energies can not provide a quantitative answer, qualitatively they could account for the 20-fold increase in the K_D of the Mn(TpCPP) cofactor. The lowest energy

modes displayed a better complementarity between this cofactor and the xylanase host than with the Mn-salen, which in-turn resulted in a deeper insertion and a wider hbond network with the polar residues of the cavity.

No significant differences were observed in the docking results obtained for the binding of the two homogeneous catalysts into the three different conformations of the Xln10A host. Consequently, only the 1V0K protein was selected to generate the Mn(TpCPP)⊂Xln10A model structure.

II.III - Analysis of the substrate binding modes in the modeled artificial metalloenzyme

The Mn(TpCPP)⊂Xln10A artificial metalloenzyme was selected to perform the catalysis of six different styrenes (Scheme 5.2). A part for the standard styrene and the β -methylstyrene, all the others were *para*-substituted. The substituents include polar (chloro, nitro and O-methoxyde) and non-polar (methyl) groups. To rationalize the different experimental results obtained in function of the styrene source the six different substrates were docked into the Mn(TpCPP)⊂Xln10A model structure. For a better representation of the catalytic pocket, an oxygen atom and a water molecule were bonded to the Mn center to recreate the Mn(V)-oxo complex and complete the first coordination sphere of this metal, respectively.

A preliminary set of docking calculations highlighted a relevant role of Arg139 in the binding of the substrate. In fact, depending on the position of this residue, it could either allow or prevent the entrance of the different styrenes to the artificial active site. We therefore changed the position of this Arg139 for a rotamer allowing the binding of the substrates and performed the dockings on this new model of the Mn(TpCPP)⊂Xln10A complex.

The docking of the different styrenes showed good binding affinity with the artificial binding site of the Mn(TpCPP)⊂Xln10A complex, with scores ranging from 25.3 to 21.9 Score units. The only exception was the *p*-nitrostyrene, in which the high volume of the polar substituent prevented it from entering the binding site (Table 5.3). In all the binding orientations, the double bond targeted for the epoxidation was laying near the metal-oxo complex (about 3Å) (Figure 5.2); a distance in agreement with the possible oxidation mechanism as observed in the studies of closely related cytochromes P450.¹⁷¹

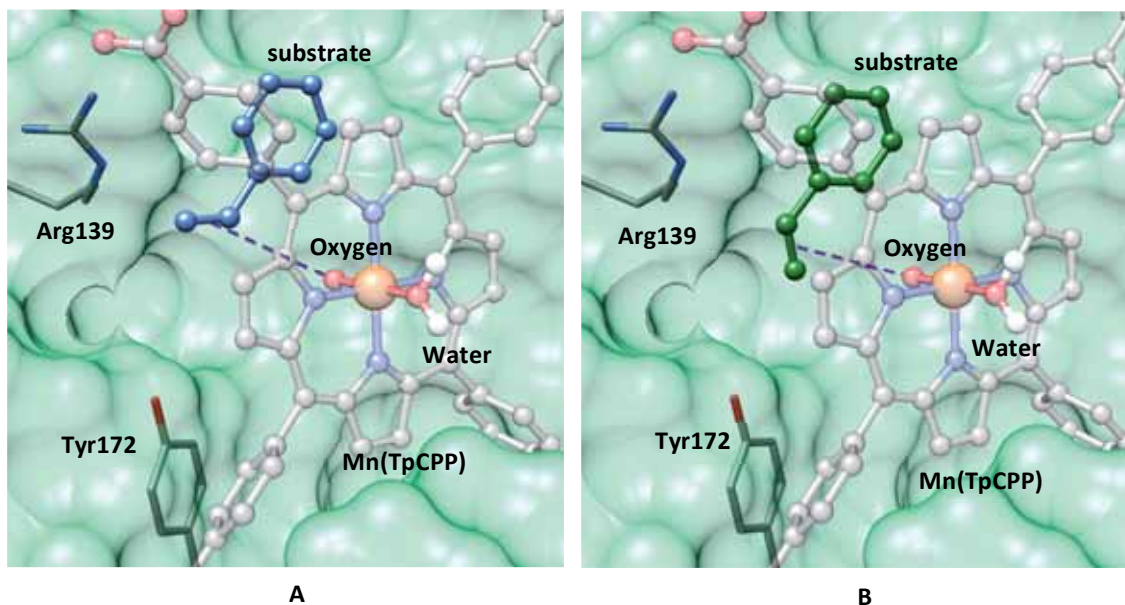


Figure 5.2 - Docking of the styrene substrate on the modeled Mn(TpCPP) \subset XIn10A structure. The proR and the proS substrates are depicted in blue (panel A) and green (panel B) respectively. In both cases the purple dashed line represents the attack of the oxygen atom on the C=C bond.

Substrate	Number of solutions		Lowest binding energy (average)	
	R	S	R	S
Styrene	3	17	22.6 (22.5)	22.96 (22.6)
β -methylstyrene	6	14	25.0 (24.7)	25.3 (25.0)
<i>p</i> -nitrostyrene	0	0	0.0 (0.0)	0.0 (0.0)
<i>p</i> -chlorostyrene	3	17	24.2 (23.9)	24.6 (24.2)
<i>p</i> -methylstyrene	6	13	24.6 (23.8)	24.7 (24.2)
<i>p</i> -methoxystyrene	8	12	22.5 (22.1)	21.8 (21.5)

Table 5.3 - Docking results for the different styrenes tested on the Mn(TpCPP) \subset XIn10A hybrid. The lowest binding energy (and the corresponding averages) are given in absolute non-dimensional values of Scores as given by the ChemScore scheme.

Because of the relatively high variability in the substrate orientations, a cluster analysis on the docking poses was performed. Two major ones were identified for every styrene, which corresponded to the proR and proS forms of the substrate (Figure 5.2). The difference in the predicted binding energies between the two clusters was no higher than 1 Score unit in any case. These results are at the limit of accuracy of the scoring functions and cannot quantitatively describe the enantiomeric excess of the system. However, some qualitative interpretations could still be stated. Except for the *p*-methoxystyrene, the predicted energies suggested a slightly preference for the formation of the S product as

the corresponding cluster was generally more stable and more populated. In the case of the *p*-methoxystyrene this tendency was inverted, thus indicating a slight preference for the formation of the *R* enantiomer. In this case, the main difference between the two clusters was an additional hbond contact between the -Omet group and Tyr172, only present in pro*R* solutions (Figure 5.3). This suggested that the *ee* observed in that particular styrene could be due to the additional stabilization of the pro*R* orientation by Tyr172. Altogether, these results were in agreement with the *ee* obtained experimentally; high *ee* towards the *R* product in the *p*-methoxystyrene and a modest *ee* towards the *S* product in all the others. To quantitatively obtain the *ee* of the system a far more intensive computational method would be required to study the whole catalytic mechanism (including intensive QM/MM calculations), which is far from the scope of this section.

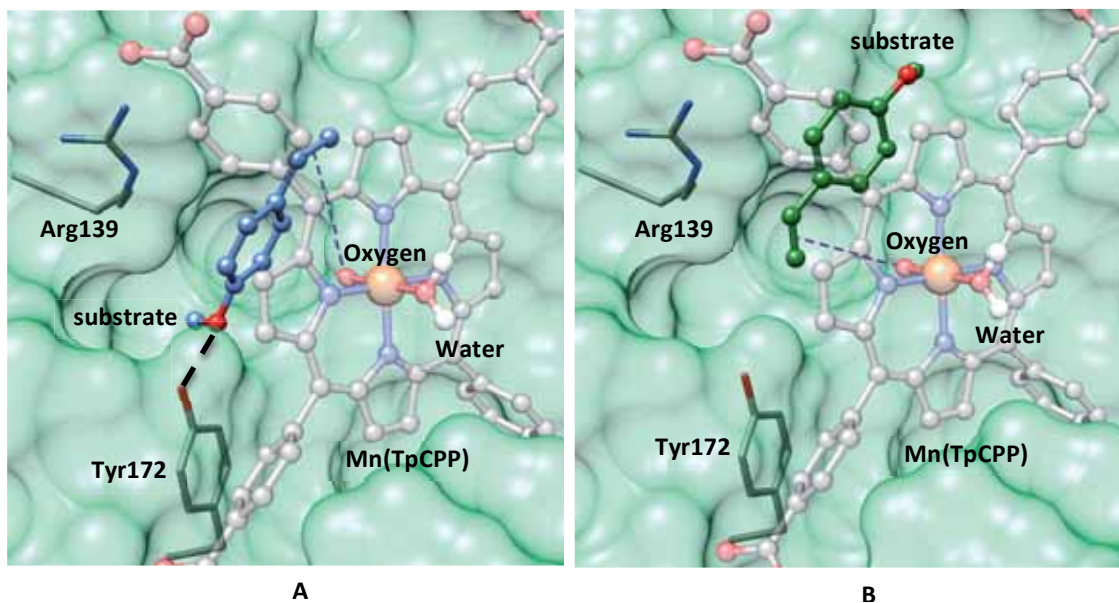


Figure 5.3 - Docking of the substrate of the *p*-methoxystyrene (substrate) on the modeled Mn(TpCPP)Xln10A structure. The pro*R* and the pro*S* substrates are depicted in blue (panel A) and green (panel B) respectively. In both cases the purple dashed line represents the attack of the oxygen atom on the C=C bond.

II.IV - Conclusions

In this part of the Ph. D. dissertation we tried to rationalize the preferential binding of the Mn(TpCPP) cofactor into the Xln10A scaffold and the *ee* observed for the resulting biometallic hybrid. Since many structural information was missing, docking calculations were needed both for the prediction of the binding of the cofactor and for the subsequent binding of the substrate. We believed that extensive QM/MM calculations would

represent unnecessary efforts because of the limited trust on the initial models and tried to rationalize the *ee* from the same docking study. Even though docking calculations alone could not fully account for the *ee* profile, they could still offer some valuable information. Our work highlighted the molecular events dictating the better stability of the Mn(TpCPP)@Xln10A hybrid and suggested that Tyr172 may be one of the main participants involved in the enantioselectivity observed for the *p*-methoxystyrene. Additionally, it also pointed out the importance of Arg139, which could either allow or block the entrance of the substrate towards the binding site.

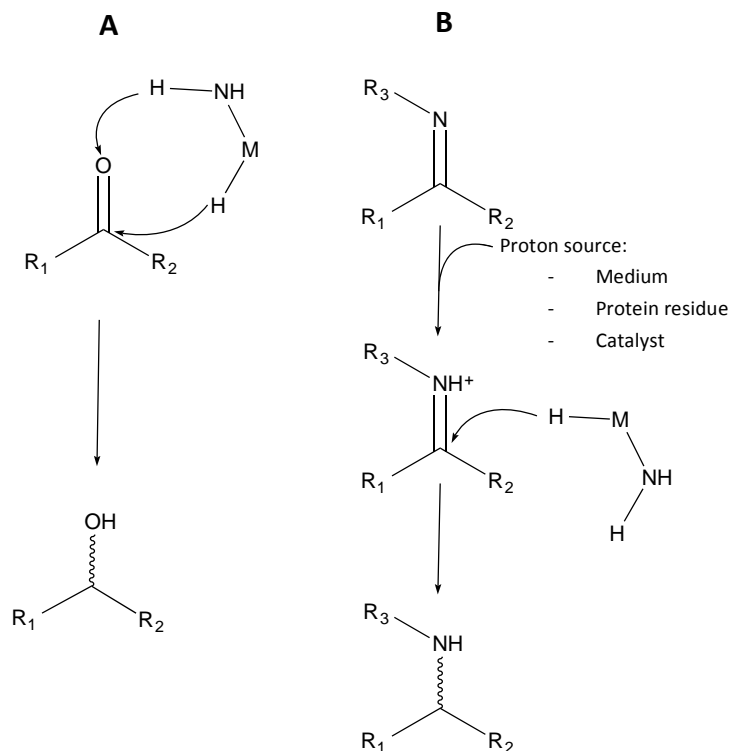
III - Effects of the protein environment in the catalytic mechanism

In the previous section we have demonstrated that, despite their qualitative nature, protein-ligand dockings can provide valuable information in the design/optimization of artificial metalloenzymes. However, in many cases only the characterization of realistic transition state structures allows to clearly evaluate the role of the protein environment in the catalysis. Unveiling the insights of these structures can open the way for the rational optimization of the *ee* of (artificial metallo)enzymes. However, to the best of our knowledge, no study has been reported achieving this objective solely using computational approaches and only a few ones manage to successfully predict the enantioselectivity of a benchmarked system.^{85,88} Main reason is that a 100% *ee* can be achieved with differences in ΔG as small as 2.2 kcal mol⁻¹ between the two possible enantiomeric pathways, thus astoundingly accurate (and by definition intensive) QM calculations are needed to provide with an accurate quantitative prediction of the enantioselectivity.¹⁷²

To test to which point computational chemistry tools can provide with accurate predictions in the study of the *ee* of artificial metalloenzymes, particularly of those of the kind studied in this Ph. D. thesis, we performed a study aimed at the *in silico* identification of their most realistic transition states. For this purpose, we needed a system to use as benchmark, which ideally should have an X-ray structure available and a well-characterized and outstanding enantioselective profile. From all the artificial bioentities reported so far, the artificial ATHases obtained by the group of Thomas R. Ward are of the most efficient in terms of both rate and *ee*.⁴³⁻⁴⁵ In particular, the one obtained from the insertion of a [Cp*Ir(biot-*p*-L)Cl] catalyst into the S112A mutant of streptavidin (Sav) is of the most competent ones, with a 96% *ee* towards the *R* product for the ATH of a salsolidine cyclic imine.⁴⁵ This system was already studied in section IV of Chapter IV to determine the enantiodiscrimination for a specific form of the inorganic catalyst by the protein scaffold. Our knowledge of this particular hybrid and its experimental features made this system an ideal candidate in which to perform our benchmark.

Although the [Cp*Ir(Biot-*p*-L)Cl]⊂S112A ATHase represented an ideal candidate to test the predictivity of *ee* by computational means, some considerations must be taken into account before starting our computational study. First, we needed to fully understand the catalytic mechanism of the reaction: ATH of cyclic imines by “Noyori-type” d⁶-piano-stool complexes, which was in itself a rather complex problem. While the ATH of ketones by this

kind of homogeneous catalysts is well-established,¹⁷³ their role in the reduction of imines is still a matter of active debate (Scheme 5.3).^{154,174–177} Up to date, the most accepted mechanism is the so-called “*ionic*” mechanism, in which the reaction should go through the creation of an iminium intermediate prior to the transfer of the hydride to the reactive C_{imine} , both processes being step-wise.^{173,174,176–181} However, to the best of our knowledge, there is still a lack of a unified view on this matter.



Scheme 5.3 - The mechanism of ATH performed by Noyori’s type catalyst for ketones (A) and the “*ionic*” mechanism for imines (B). The first one is a concerted mechanism in which both the hydride and the proton are delivered in the same step from the catalyst to the ketone. The second is a step-wise mechanism in which the N_{imine} is first protonated by a proton donor and then the hydride is delivered to the reactive C_{imine} from the metal.

An additional problem was the vast conformational space we needed to explore in order to find the most probable transition states in the protein. Indeed, there were several questions on the catalytic mechanism that the $[Cp^*Ir(\text{Biot-}p\text{-L})Cl] \subset S112A$ X-ray structure could not answer, like: (i) where would the substrate stand for the catalysis?, (ii) which would be the preferred proR and proS orientations of the salsolidine? and (iii) how would the binding of the substrate in each of those orientations affect the loaded catalyst?

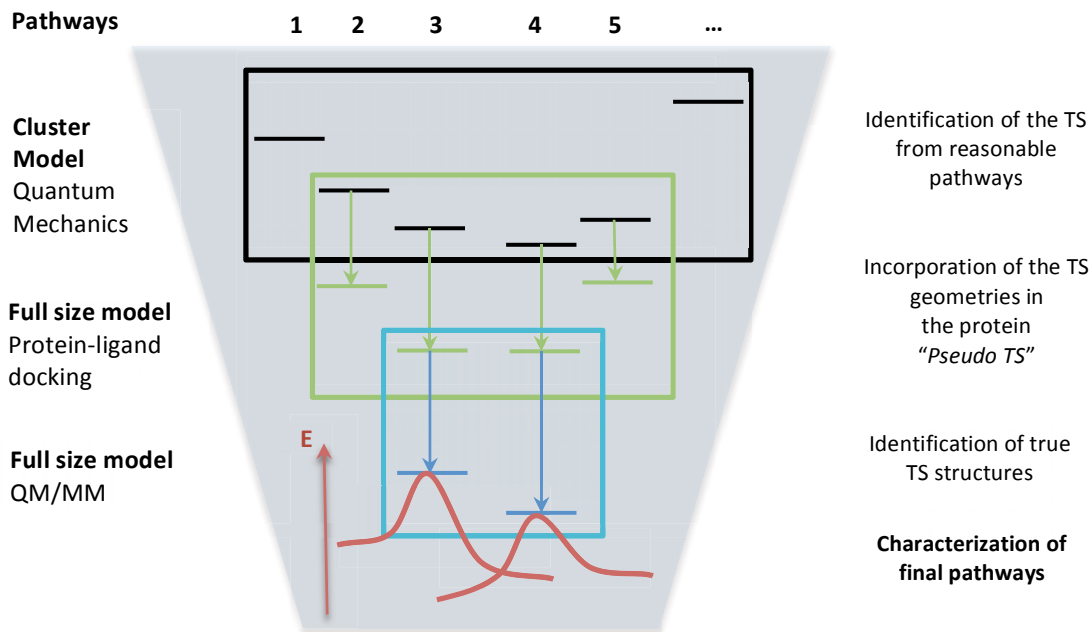
To accurately predict the *ee* profile of the $[Cp^*Ir(\text{Biot-}p\text{-L})Cl] \subset S112A$ hybrid we need to identify: (i) which would be the most probable ATH mechanism for imines, (ii) how the

substrate would approach the embedded catalyst and (iii) the most-likely transition states for each enantiomeric pathway. It is important to notice that all those molecular problems cannot be tackled using a brute force approach. This would imply to perform MD simulations taking into account each possible orientation of the substrate and a posterior QM/MM refinement of the most representative frames to identify all the possible transition states. We would need a humongous amount of time and resources and eventually we could end up with artefactual results due to the limited trust we could have in the MD models. Instead, we designed an integrative protocol aimed at decreasing the conformational space to explore in each consecutive step by mixing QM, protein-ligand dockings and QM/MM approaches. Using a small cluster of the system real transition states taking into account different pathways can be identified and docked in the protein. This way we would obtain reliable models of the biometallic hybrid displaying the proR and proS orientations of the substrate in which to perform the QM/MM calculations to identify realistic transition states. If this protocol can offer a good prediction of the *ee* of the [Cp*Ir(Biot-*p*-L)Cl]C₅H₁₁ artificial entity it could become a precious asset in the *in silico* design/optimization of complex bioinorganic systems.

III.1 - Presentation of the protocol

Our integrative protocol is divided in three different steps. They are designed to decrease the conformational space along the workflow prior to the final identification of the most likely transition states in the protein (Scheme 5.4).

The first step is directed to answer all those chemical questions emerging from the computational study of the reaction mechanism (i.e. identification of possible stabilizing interactions, characterization of the most probable pathways, etc.). Even if the protein environment could have a dramatic impact in the stabilization of the transition state, we do not expect it to have such influence in all those intrinsic catalytic features. Based on this hypothesis, we performed this analysis in a small model of the real protein system. This reduction in size is translated in faster and less intensive calculations, which allowed us to explore larger conformational spaces. From all the possible reaction mechanism/pathways identified, the lowest ones were selected for the next stage of our protocol. Those higher than this difference are not expected to be possible mechanisms/pathways even with the stabilization factor provided by the protein environment.



Scheme 5.4 - Schematic representation of the designed integrative protocol to characterize the transition state in the protein. The transition state structures of the unbound cofactor are docked into the protein receptor. Then QM/MM approaches are used to identify the minima and the real transition state inside the protein.

In the second step the transition state structures identified in the previous phase are docked into the protein binding site, obtaining what we called the "*pseudo-transition state*" structures (Scheme 5.4). Those models are not real transition states (we need QM approaches to find them), but they give us a clue on the complementarity between the realistic transition states of the isolated system and the protein binding site. Consequently, all those orientations presenting a low complementarity with the protein should be discarded.

There might be some relevant interactions between the protein/cofactor/substrate triad that docking algorithms cannot take into account (i.e. CH- π or cation- π interactions). This situation needs to be dealt with care. Those interactions can play a vital role in the catalytic mechanism and might be overlooked by the following QM/MM scheme as it cannot explore large conformational spaces. To overcome this problem, these interactions can be identified by a visual inspection of the pseudo-transition state structure and the implicated residues replaced for a more suitable rotamer.

In the final step the generated *pseudo-transition states* are used as the starting point to find the real transition states using a QM/MM scheme (Scheme 5.4). Each reaction

pathway is fully characterized (from reactants to products) to identify the lowest ones in energy. The structural information gathered from these models will aid in the rational design/optimization of these hybrids.

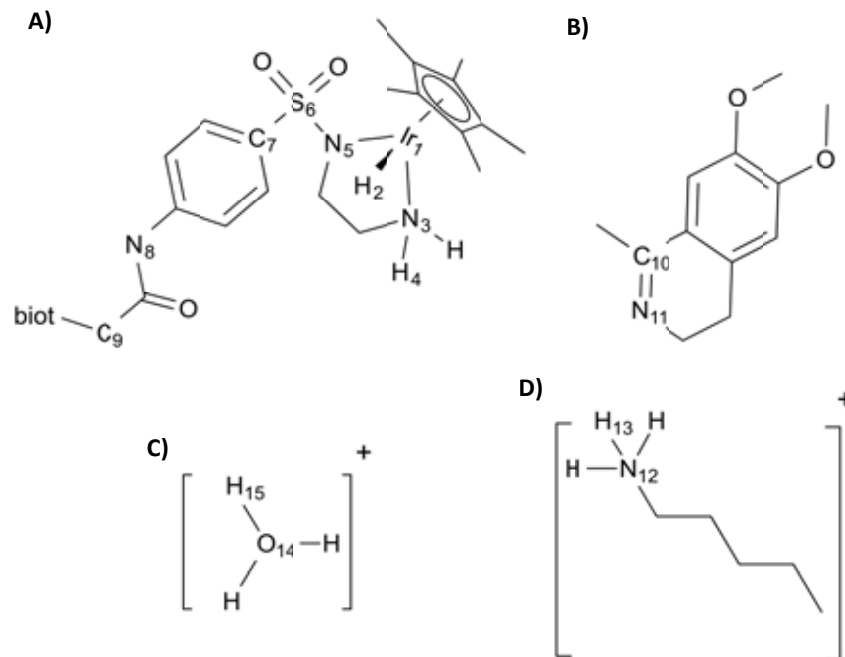
III.II - Computational Methods

Calculations on the small cluster model were performed within the DFT approach using the PBE^{147,148} functional as implemented in Gaussian09¹²³. The basis set Def2-TZVPP¹⁴⁹ and its associated pseudo-potential were used for the treatment of the iridium atom. The other atoms were treated with the 6-31g*¹²⁵ basis set. All optimizations were performed with water as implicit solvent using a polarizable continuum model (CPCM)¹⁵⁰⁻¹⁵³ as implemented in Gaussian09. All transition states and minima were validated performing a frequency analysis at the same level of theory.

To generate the *pseudo-transition state* structures a covalent docking approach was performed as implemented in the program GOLD5.1¹²⁷ using the ChemScore^{128,155} scoring function and the N₈ as the anchoring atom^a (Scheme 5.5). The surrounding residues, Leu110_A and Ser88_A as well as Lys121 and Leu124 from both monomers were allowed flexibility using the Dunbrack Rotameric library.¹²⁰ All structures were prepared as specified in the GOLD user manual using the UCSF Chimera interface.¹⁰⁷

QM/MM calculations were performed within the two-layer ONIOM¹³¹ approach as implemented in Gaussian09 using the electronic embedding¹⁸² scheme. The salsolidine substrate and the biotinilated catalyst up to C₉ (this included) were considered within the QM region (Scheme 5.5) and treated using the same functional and basis sets used in the cluster calculations. For a better representation of the interactions between the substrate and the protein environment, the residues directly interacting with it were also included in the QM partition. The rest of the system was included in the MM region and was treated with the AMBER force field.¹⁸³ Residues Leu124_A, Leu124_B, Leu110_A, Ala112_A, Lys121_A, Lys121_B and Ser88_A as well as the whole biotinilated cofactor were allowed flexibility during the calculations. All minima and transition state structures were validated by performing a frequency analysis at the same level of theory.

^a For more information about the covalent docking approach with the Sav-biotin technology, please refer to Materials and Methods from Section IV in Chapter 4.



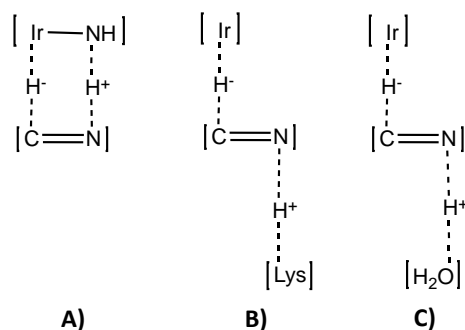
Scheme 5.5 - Schematic representation of **A**) R_{ir} -[Cp*Ir(Biot-*p*-L)H] catalyst, **B**) 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline substrate **C**) hydronium ion and **D**) protonated lysine side chain. **C**) and **D**) can both donate the proton to the imine nitrogen, defining different ATH mechanisms (H_{ir}^-/H_{med}^+ or H_{ir}^-/H_{Lys}^+ respectively).

The enantiomeric excess of the reaction was calculated from the difference in Gibbs energy between the energy barriers of the most likely *R* and *S* enantiomeric pathways at 298K applying the following equation:¹⁷²

$$\% ee = \frac{1 - e^{-\Delta G_{R/S}^\ddagger/RT}}{1 + e^{-\Delta G_{R/S}^\ddagger/RT}} \cdot 100 \quad \text{Eq 5.1}$$

III.III - Unveiling the ATH mechanism for cyclic imines in a small cluster model

The identification of the most likely transition states was performed on different cluster models of the [Cp*Ir(Biot-*p*-L)Cl]⊂S112A system. All of them included the salsolidine substrate and a simplified version of the catalyst, in which the biotinic anchor part up to the N_{amide} link (N_8 in Scheme 5.5) was replaced by a methyl group. Those models differ from each other in the proton source, which defined the different mechanisms studied: (i) the N-ligand of the piano-stool moiety of the catalyst (H_4 in Scheme 5.6, H_{ir}^-/H_{ir}^+ mechanism), (ii) the only polar residue able to donate a proton in the protein environment, the Lys121 (H_{13} in Scheme 5.6, H_{ir}^-/H_{Lys}^+ mechanism) or (iii) a hydronium from the media (H_{15} in Scheme 5.6, H_{ir}^-/H_{med}^+ mechanism).



Scheme 5.6 - Representation of the different mechanisms taken into account for the mechanistic study of the ATH of imines in the cluster model. **A)** H^-_{Ir}/H^+_{Ir} , **B)** H^-_{Ir}/H^+_{Lys} , **C)** H^-_{Ir}/H^+_{med} .

In our previous docking study on the $[Cp^*Ir(Biot-p-L)Cl] \subset S112A$ ATHase,^b we concluded: (i) the highest activity/enantioselectivity is only obtained when there is one cofactor loaded in every dimer and (ii) in this situation there is a preference to bind the S_{Ir} - $[Cp^*Ir(biot-p-L)Cl]$ form of the catalyst. Subsequently, all the calculations were performed only taking into account the activated form of this pseudo-enantiomer of the catalyst (R_{Ir} - $[Cp^*Ir(biot-p-L)H]$).^c

In stark contrast with the ATH of ketones, the mechanism for the ATH of imines is still under a strong controversy. For this reason, we first need to unveil whether in our system the transfer of both the hydride and the proton is concerted (like in the ATH of ketones) or step-wise, and in this case whether the proton is transferred prior to the hydride or the other way around. For this purpose, we calculated the 3D potential energy surface to identify the lowest energy pathway preferred for each of the studied mechanisms (Scheme 5.6). However, as this task is highly resource-consuming, some simplifications were taken into account. First, we further simplified the cluster models: the biotinilated part of the catalyst up to the C_7 was replaced by a methyl group (Scheme 5.5). Second, due to the resemblance between the H^-_{Ir}/H^+_{Lys} and the H^-_{Ir}/H^+_{Med} mechanisms (external proton source), only the former was taken into account. Both the proR and proS substrate orientations were considered.

The computed 3D energy surfaces clearly indicated that the H^-_{Ir}/H^+_{Ir} and the H^-_{Ir}/H^+_{Lys} mechanisms followed utterly different pathways (Figure 5.4). In the H^-_{Ir}/H^+_{Ir} case (**top**, Figure 5.4) the lowest pathway started from the cyclic imine (**1**) and went through a

^b Please refer to section IV in Chapter 4.

^c When replacing the chloride with a hydride the chirality of the metal changes from S_{Ir} to R_{Ir} as dictated by the Cahn–Ingold–Prelog priority rules.

concerted transition state (**2**) in which both the proton and the hydride are transferred in the same step to arrive to the final amine product (**3**). On the other hand, the lowest energy pathway in the H_{Ir}^-/H_{Lys}^+ mechanism followed a step-wise fashion (**bottom**, Figure 5.4). Starting from the imine substrate (**1**) it went first through a little energy barrier (**2**) in which the proton is transferred to the N_{imine} , obtaining the protonated imine (**3**). Afterwards it continues through a much higher transition state (**4**) corresponding to the delivery of the hydride to the C_{imine} to obtain the final amine (**5**). This preference for the concerted mechanism in the H_{Ir}^-/H_{Ir}^+ case and for the step-wise in the H_{Ir}^-/H_{Lys}^+ one is observed in all the calculations performed regardless of which enantiomeric amine is produced.

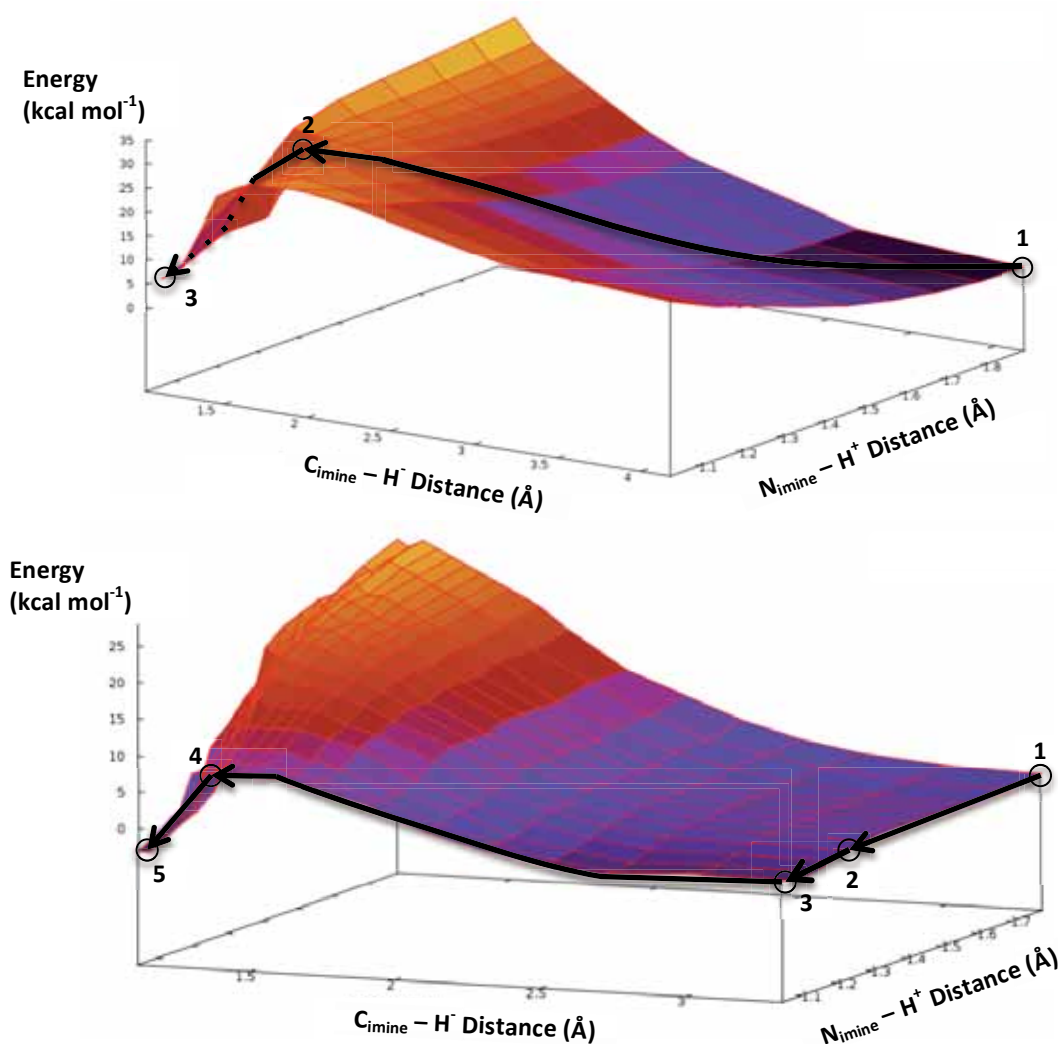


Figure 5.4 - 3D potential energy surfaces obtained for the H_{Ir}^-/H_{Ir}^+ (**top**) and the H_{Ir}^-/H_{Lys}^+ (**bottom**) mechanisms resulting in the *R* amine product. The black lines mark the lowest energy pathways in both cases. The $C_{imine}-H^-$ and the $N_{imine}-H^+$ distances were taken as the reactions coordinates to draw the surface.

Interestingly, the 3D potential energy surface indicated that the nature of the transfer of both the proton and the hydride was subject to which was the proton source. If we have an external proton source (e.g. a lysine) the transfer follows a step-wise fashion while if the proton comes from the inorganic moiety (e.g. the piano-stool) a concerted transfer is preferred. Thus the H^-_{ir}/H^+_{ir} mechanism was considered to be concerted (like in the ATH of ketones)¹⁷³ whereas in the H^-_{ir}/H^+_{Lys} and in the H^-_{ir}/H^+_{med} mechanisms it was considered step-wise.^d This last result was in agreement with a computational study on the ATH of imines reported by Petr and coworkers.¹⁵⁴

Regarding the activation energies of both the H^-_{ir}/H^+_{ir} and the H^-_{ir}/H^+_{Lys} mechanisms, in the later case they were far lower than in the former. In fact, both the *R* and the *S* enantiomeric pathways in the H^-_{ir}/H^+_{Lys} presented energetic barriers close to 10 kcal mol⁻¹ in their rate-limiting step (the transfer of the hydride), while those same barriers were close to 25 kcal mol⁻¹ in the H^-_{ir}/H^+_{ir} . It is highly unlikely that the protein environment could overcome this 15 kcal mol⁻¹ difference, thus the H^-_{ir}/H^+_{ir} mechanism was discarded at this point.

At this point two different mechanisms were under revision to explain the ATH of imines in our system: the H^-_{ir}/H^+_{med} and the H^-_{ir}/H^+_{Lys} . Taking into account the results of the 3D potential energy surface, the whole pathway of the reaction was recalculated in both cases taking into account the attack of the hydride in both the *re* and *si* faces of the substrate and considering the transfer to be step-wise. As the benzene moiety of the catalyst could have an impact in the catalysis, the original cluster models were used in the simulations in which only the biotinilated part up to the N₈ (this included) was replaced by a methyl group (Scheme 5.5).

In the two mechanisms studied the protonation of the N_{imine} to form the iminium was barrierless (H^-_{ir}/H^+_{med}) or almost barrierless (H^-_{ir}/H^+_{Lys}) with a ΔG barrier of around 0.5 kcal mol⁻¹ in the later case (Table 5.4). Interestingly, in the H^-_{ir}/H^+_{med} mechanism once the iminium was formed the substrate presented two different orientations in which the transfer of the hydride could take place. In the first one the iminium is making a hbond contact with the sulfone moiety of the catalyst (TSO models, Figure 5.5), while in the other this interaction is between the iminium and the -NH₂ of the piano stool moiety (TSN models, Figure 5.5).

^d This is the so-called “*ionic*” mechanism, the most accepted hypothesis for the ATH of imines.

Mechanisms	Chirality	Energy (kcal mol ⁻¹)					TS2 Freq.	Distances (Å)	
		Reactant	TS1	Intermediate	TS2	Product		C – H ⁻	Ir – H ⁻
H ⁻ _{Ir} /H ⁺ _{Lys}	<i>R</i>	0.0 (0.0)	0.5 (-4.4)	-2.0 (-4.5)	10.2 (10.3)	-9.1 (-7.3)	-398.7	1.62	1.75
	<i>S</i>	0.0 (0.0)	0.6 (-2.2)	-1.3 (-3.8)	12.5 (12.4)	-3.9 (-4.5)	-343.4	1.62	1.73
H ⁻ _{Ir} /H ⁺ _{med}	<i>R</i> _{TSO}	0.0 (0.0)	-	-	12.0 (13.5)	-1.45 (0.4)	-325.5	1.55	1.74
	<i>R</i> _{TSN}	0.0 (0.0)	-	-	11.3 (12.4)	-13.0 (-8.8)	-325.0	1.69	1.74
	<i>S</i> _{TSO}	0.0 (0.0)	-	-	13.9 (18.0)	-0.8 (3.6)	-368.2	1.55	1.75
	<i>S</i> _{TSN}	0.0 (0.0)	-	-	8.9 (8.7)	-7.0 (-6.9)	-353.0	1.62	1.74

Table 5.4 - Energies obtained for the different minima points and transition states on the small clusters models. Gibbs energies are given in brackets. The distances and the frequencies are associated to the hydride transfer transition state.

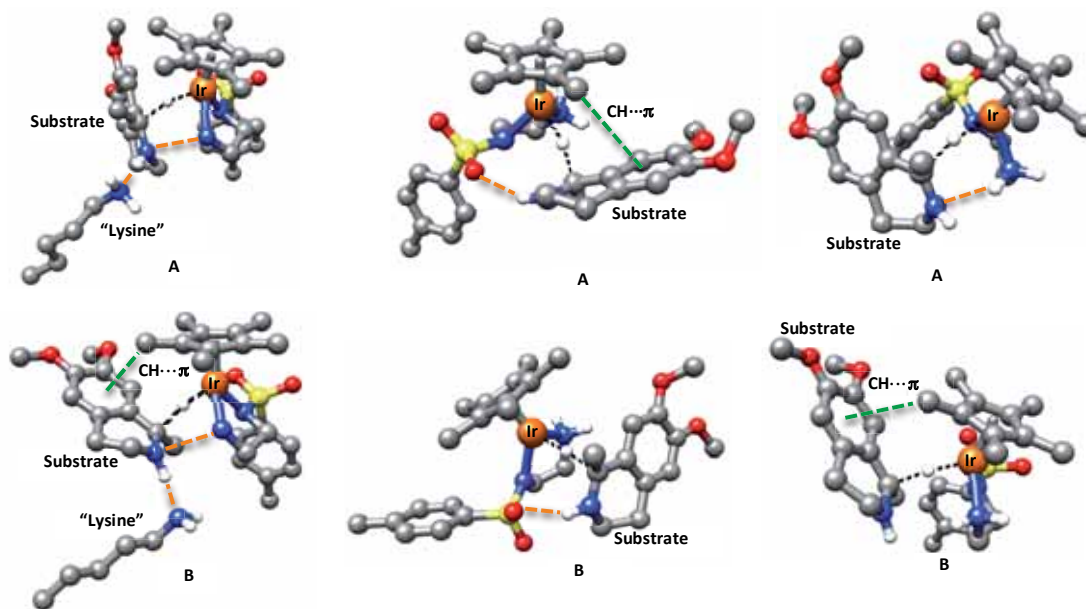


Figure 5.5 - Transition state structures corresponding to the different mechanisms of the ATH of the cyclic imine 1-methyl-6,7-dimethoxy-3,4-dihydroisoquinoline by the iridium catalyst (*R*)-[Cp*Ir(Biot-*p*-L)H]. **Left:** H⁻_{Ir}/H⁺_{Lys} mechanism. **Middle:** TSO H⁻_{Ir}/H⁺_{med} mechanism. **Right:** TSN H⁻_{Ir}/H⁺_{med} mechanism. All the transition state at the top of the image (**A** cases) lead to the *R* amine while the ones at the bottom (**B** cases) correspond to the *S* product. The black dotted lines represent the hydride transfer, the orange ones the hydrogen bonds and the green ones the CH-π interactions.

In the H⁻_{Ir}/H⁺_{med} mechanism the protonation step was rendered totally barrierless. To determine the impact of the metal in this stage of the catalysis, we performed the same protonation step without the homogeneous catalyst. This simulation led to the exact same results, thus we hypothesized that the cyclic imine could be protonated in the media, even before entering the binding site of the enzyme. Consequently, the resulting

water molecule from the proton transfer was not considered in the cluster model of the H^-_{ir}/H^+_{med} mechanism.

The almost inexistent energy barrier from the protonation step in the H^-_{ir}/H^+_{med} and the H^-_{ir}/H^+_{Lys} mechanisms clearly indicated that the posterior attack of the hydride was indeed the limiting step of the ATH in those cases.¹⁷⁴ We therefore characterized the transition state of the hydride transfer for the H^-_{ir}/H^+_{Lys} and the H^-_{ir}/H^+_{med} mechanisms, considering the two possible modes of interactions between the catalysts and the substrate in the latter case. The energy barriers obtained were compared with those of the H^-_{ir}/H^+_{ir} mechanisms to shed light on which should be the preferred mechanisms for the ATH of imines in this case.

Both attacks on the *re* and *si* faces of the substrate in the H^-_{ir}/H^+_{ir} mechanism presented the highest energy ($>25 \text{ kcal mol}^{-1}$) from all the studied mechanisms (Table 5.4). All the other step-wise mechanisms, whether having the *proS* or *proR* orientation of the substrate, were much more favorable with energies ranging from 9 to 14 kcal mol^{-1} for the hydride transfer (Table 5.4). Interestingly, those energies were similar to those reported by Petr and coworkers on their work on ATH of imines.¹⁷⁶

Several different interactions between the catalyst moiety and the substrate were contributing to the stabilization of the transition state structures, including some CH- π and several hbond contacts between the N_{imine} and: (i) the pseudo-lysine (H^-_{ir}/H^+_{Lys}), the $-NH_2$ moiety of the pianostool (H^-_{ir}/H^+_{ir} , H^-_{ir}/H^+_{Lys} and some TSN models) or the sulfone group (TSO models). Interestingly, there was a 2 kcal mol^{-1} overstabilization in those enantiomeric pathways presenting a CH- π interaction that belonged to the same ATH mechanism (Figure 5.5 and Table 5.4).

The QM study on the cluster models suggested that the step-wise mechanisms H^-_{ir}/H^+_{Lys} and H^-_{ir}/H^+_{med} were much more favorable ($>10 \text{ kcal mol}^{-1}$) than the concerted H^-_{ir}/H^+_{ir} one. This indeed proves that the ATH of imines must undergo a different mechanism to that reported for the ATH of ketones.¹⁷³ The lowest transition state identified was that belonging to the hydride attack on the *re* face of the substrate in the TSN orientation ($\approx 9 \text{ kcal mol}^{-1}$). These results differed from the ones obtained by Petr and coworkers, who determined that the polar interaction between the protonated substrate and the sulfone moiety of the catalyst was one of the key elements stabilizing the transition state.¹⁷⁶ Those discrepancies could be due to some disparities between the two computational approaches: (i) the different DFT functional (PBE used in this work, B3LYP used by Petr et

al.),¹⁷⁶ (ii) the solvent conditions (here we optimized using always implicit solvent conditions, while Petr optimized in gas phase and performed single points calculations in implicit solvent) and (iii) the nature of the catalyst (here an $\{\eta^5\text{-Cp}^*\text{Ir}\}$ based catalyst with an ethylenediamine ligand, while Petr focused on a $\{\eta^6\text{-}p\text{-cymeneRu}\}$ system with an 1,2-diphenylethylenediamine ligand).

We do not expect that the protein environment could overcome the high difference in energy between the concerted and step-wise mechanisms studied. Therefore we discarded the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Ir}}$ mechanisms for subsequent steps of our integrative protocol.

III.IV - Modelling the pseudo-transition state structures in the protein

The characterized transition states of the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{med}}$ and the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Lys}}$ mechanisms were docked into the $\text{Sav}_{1/2}$ to generate the pseudo-transition state structures using a covalent docking approach. Both transition states leading to the *R* and *S* enantiomeric products were considered for each mechanism.

In the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Lys}}$ mechanism an additional restraint was added between the N_{imine} of the substrate and the $\text{N}\zeta$ of the Lys121 to obtain the pseudo-transition state structures compatible with the proton transfer (N_{11} and N_{12} in the cluster model respectively, Scheme 5.5). It is important to notice that there were two different Lys121 on the binding site that could fulfill the proton-donor role, one for each Sav monomer. Consequently, we performed each docking of the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Lys}}$ mechanism in duplicate applying the restraint with either Lys121_A (monomer A, where the covalent docking was actually performed) or Lys121_B (monomer B).

A total of eight different docking simulations were performed: four for the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{med}}$ mechanism (*R* and *S* product pathways, TSN and TSO orientations) and four for the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Lys}}$ mechanism (*R* and *S* product pathways, Lys121_A and Lys121_B as proton sources). All the predicted binding affinities in both mechanisms were close to 37 Score units with a maximum difference of 9.4 Score units (S_{TSO} and S_{TSN} , Table 5.5). These results suggested a good complementarity of all the pseudo transition state like structures with the protein environment. Interestingly, those of the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{Lys}}$ mechanism were 3 ChemScore units better in average than those of the $\text{H}^-_{\text{Ir}}/\text{H}^+_{\text{med}}$. Nonetheless, the energetic breakdown of the docking terms pointed out a higher number of clashes in all the transition states of the latter case (S_{clash} , Table 5.5).

Model	Score	ΔG (kJ mol ⁻¹)	S_{hbond}	S_{lipo}	S_{clash}
$R_{K121.B}$	37.1	-42.5	1.9	334.4	0.8
$R_{K121.A}$	38.3	-45.1	2.3	346.0	0.8
$S_{K121.B}$	42.1	-49.2	3.1	356.2	0.8
$S_{K121.A}$	40.1	-48.6	2.8	361.2	2.5
R_{TSO}	34.9	-45.4	2.3	348.4	6.7
R_{TSN}	33.9	-45.4	1.7	363.4	6.5
S_{TSO}	42.9	-49.0	2.2	380.5	2.6
S_{TSN}	33.5	-44.3	1.8	352.4	4.6

Table 5.5 - Breakdown of the energetic terms obtained in the docking simulation to obtain the pseudo-transition state structures.

In the $H_{\text{Ir}}^-/H_{\text{med}}^+$ mechanism the most stable solutions for the R_{TSN} , R_{TSO} and S_{TSN} models presented similar binding energies (Table 5.5) and were consistent with the published structure of the $R_{\text{Ir}}-[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\subset\text{S112A}$:⁴⁵ in both cases the pianostool and the sulfonamide group of the catalyst were located inside the hydrophobic vestibule present at the interface of the two monomers. Additionally, those binding were further stabilized by an hbond contact between the $-\text{NH}_2$ group of the pianostool and the carbonyl of the backbone of Lys121_A. In those three cases, the substrate is deep inserted in the binding pocket, located in the cavity between the cofactor and the backbone of Lys121_B (Figure 5.6). On the contrary, the S_{TSO} model presented a rather different binding mode. In this case the Cp^* was deep buried in-between the two Sav monomers on the same hydrophobic pocket where the pianostool was located in the other three orientations and this moiety of the catalyst was now solvent exposed (Figure 5.6). These changes in the orientation of the homogeneous catalysts also altered the localization of the substrate, which was now located in the cavity between the catalysts and Lys121_A. The S_{TSO} binding mode seemed to have a better complementarity with the protein environment as highlighted by the improved binding energy with respect of the other $H_{\text{Ir}}^-/H_{\text{med}}^+$ models (Table 5.5).

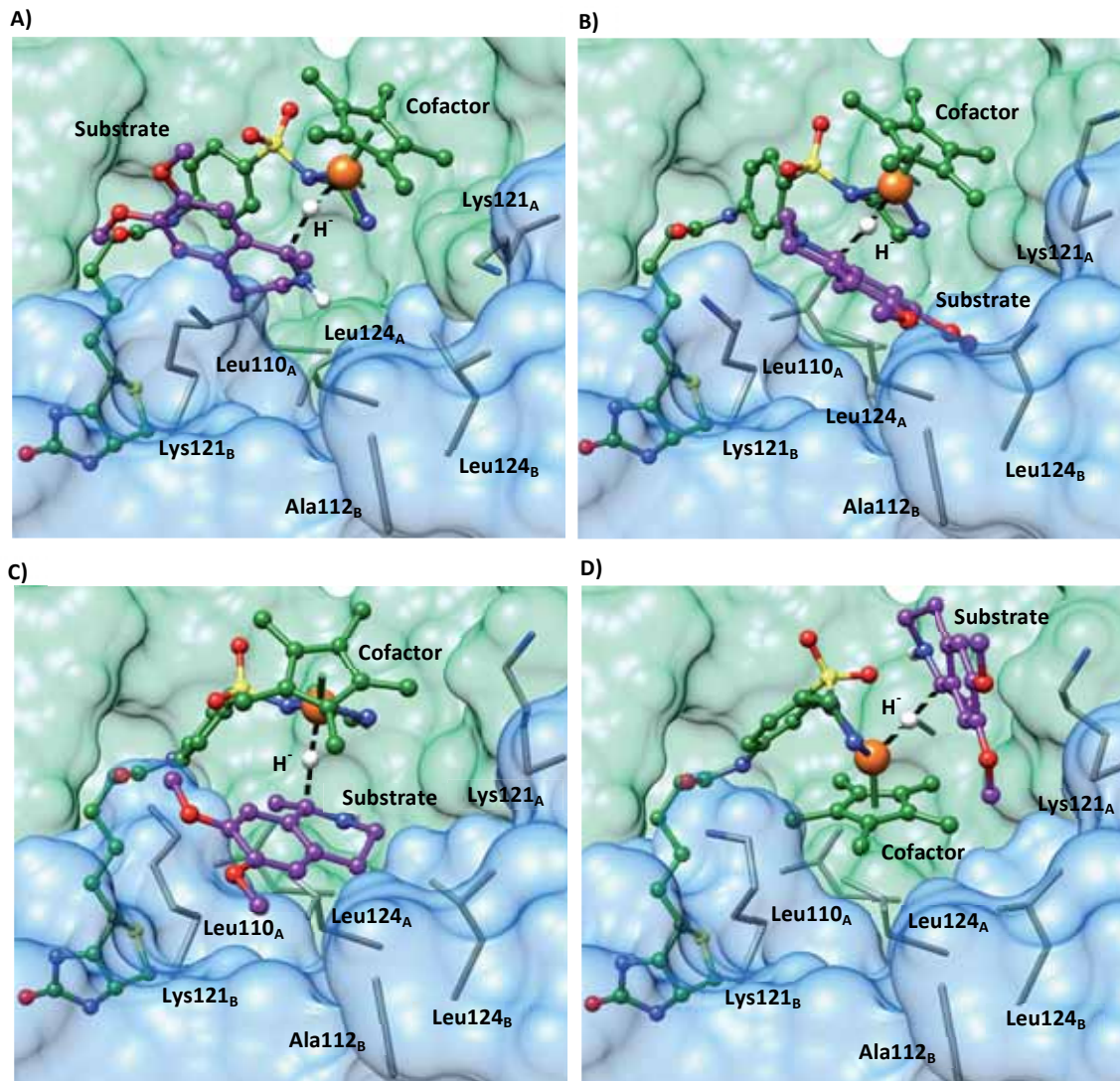


Figure 5.6 - Lowest energy orientations obtained for the docking of the different models of the H^-_{ir}/H^+_{med} mechanism: R_{TSN} (A), R_{TSO} (B), S_{TSN} (C) and S_{TSO} (D). The different monomers of the $Sav_{1/2}$ are depicted in green (monomer A) and blue (monomer B). The iridium moiety is colored in green, the isoquinoline substrate in purple. The black dashed lines represent the halfway transfer of the hydride to the C_{imine} in the pseudo-TS structure.

In the H^-_{ir}/H^+_{Lys} mechanism the binding orientations were substantially different depending on which lysine donated the proton. In the case of $Lys121_A$ the Ir moiety presented similar binding modes in both R and S pathways ($R_{K121.A}$ and $S_{K121.A}$ respectively), with the Cp^* located at the interface between the two Sav monomers and the pianostool exposed to the solvent. This orientation is inverted when the proton donor is the opposite $Lys121_B$: both pro R and pro S solutions ($R_{K121.B}$ and $S_{K121.B}$ respectively) presented the Cp^* ligand located at the solvent exposed region of the active site and the pianostool deep inserted at the interface of the two monomers (Figure 5.7). Whichever was the lysine

giving the proton, in all cases the substrate was highly solvent exposed and only in some cases there was an interaction with the protein scaffold through a hydrogen bond between the donor lysine and the N_{imine} (Figure 5.7). Indeed, even if all the docking solutions could satisfy the distance restraint imposed the orientations of the substrate were not always consistent with the transfer of the proton from the lysine to the N_{imine} . Consequently, in the $H^-_{\text{Ir}}/H^+_{\text{Lys}}$ mechanism a major rearrangement of the protein environment/cofactor/substrate triad was needed for this event to happen.

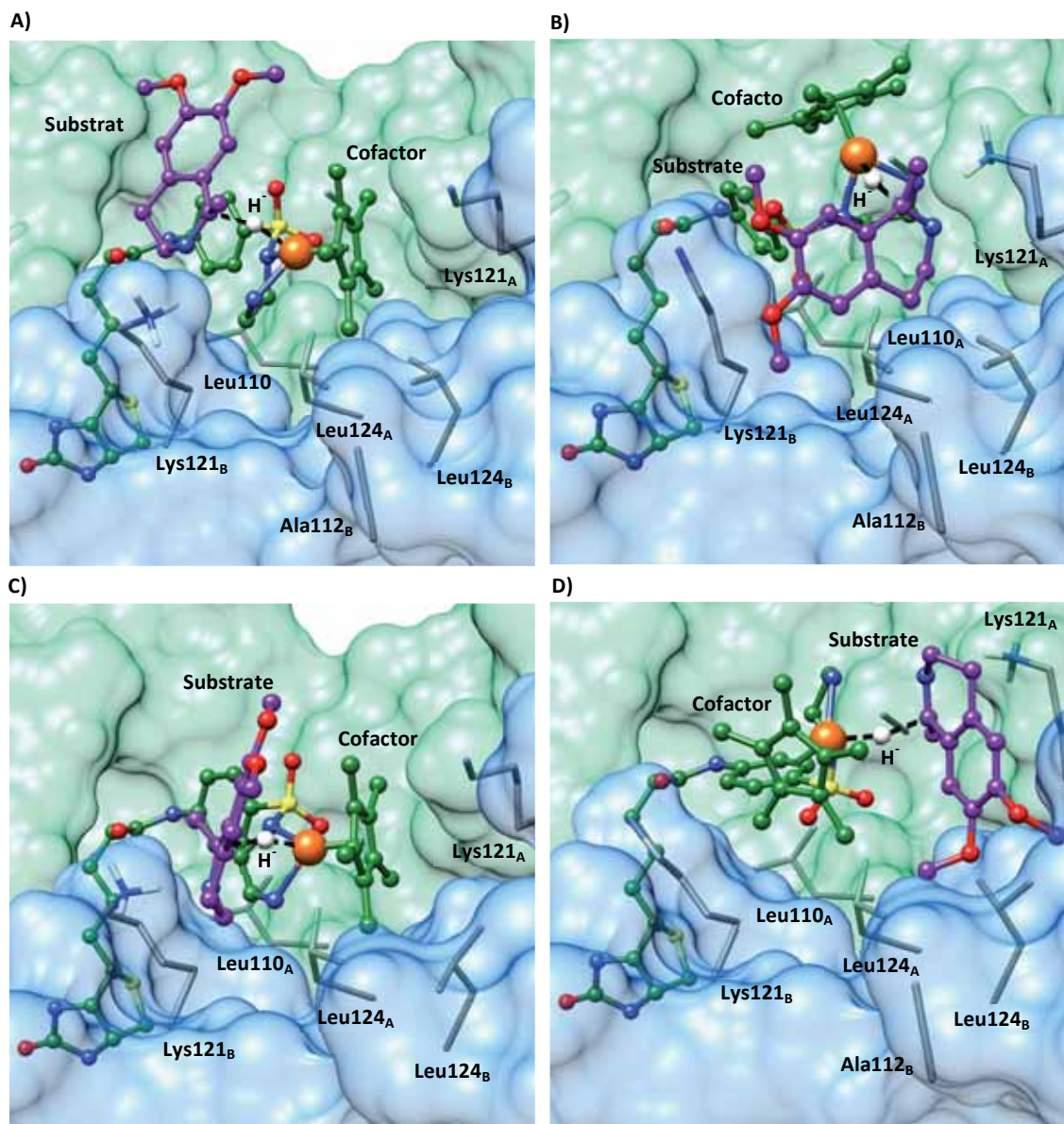


Figure 5.7 - Lowest energy orientations obtained for the docking of the different models of the $H^-_{\text{Ir}}/H^+_{\text{Lys}}$ mechanism: $R_{K121.B}$ (A), $R_{K121.B}$ (B), $S_{K121.B}$ (C) and $S_{K121.A}$ (D). The different monomers of the $Sav_{1/2}$ are depicted in green (monomer A) and blue (monomer B). The iridium moiety is colored in green, the isoquinoline substrate in purple. The black dashed lines represent the halfway transfer of the hydride to the C_{imine} in the pseudo-TS structure.

Despite the substantial differences in binding energies predicted in some cases (i.e. 9.4 Score difference between the S_{TSO} and S_{TSN} models) we could not discard any mechanism or model on energy basis for two different reasons. First, in this docking procedure we neither parametrized the metal center nor any of the corresponding ligands. The first coordination sphere of the metal was fixed and could not be optimized during the docking calculations. This was something specially important in the case of the Cp* ligand as it could have a significant influence in the binding process because of its high volume. Secondly, we are dealing with transition state structures presenting *quasi* broken/formed bonds ("bonds" linking the Ir₁-H₂-C₁₀ atoms).

To quantitatively discern between each possible mechanism/model we needed QM/MM approaches. Therefore, all the lowest energy solutions obtained in the docking were used as the starting points in a QM/MM analysis to find the corresponding real transition states on the protein. It is important to notice that this approach is extremely sensitive; clashes present in the starting structure could cause the simulation to fail. To relax the clashes that might be present in the docking structures, a few minimization steps were performed using the MMTK¹⁰⁶ minimizer as implemented in the USCF Chimera with the AMBER¹³² force field. To prevent this minimization to take us away from the transition state structure the first coordination sphere of the metal center was kept fix.

III.V - Finding the real transition state structures

A total of eight different structures - four for the H^-_{Ir}/H^+_{med} mechanism and four for the H^-_{Ir}/H^+_{Lys} one - were evaluated using QM/MM techniques and their transition states were characterized. Once located, the nearest minima of the potential energy surface were also identified leading to the corresponding reactants and products.

The QM/MM partition was different depending on the mechanism/model studied. In all of them the high layer included the substrate, the iridium moiety of the catalyst (up to the N₉ atom) and a neighboring lysine up to their C α (this excluded). This residue was the one varying from one case to the other; for the H^-_{Ir}/H^+_{med} TSO model it was Lys121_A, for the rest of the H^-_{Ir}/H^+_{med} models it was Lys121_B and for the H^-_{Ir}/H^+_{Lys} mechanisms it was the lysine donating the proton. These lysine residues were included because they were directly interacting with the substrate, thus improving the representation of the system and the results of the simulation.

The QM/MM calculations on all the models from the H^-_{Ir}/H^+_{Lys} mechanism followed the same pattern as in the cluster calculations: (i) protonation of the substrate by a close-lying

lysine (Lys121_A or Lys121_B depending on the model) and (ii) transfer of the hydride from the iridium center to the C_{imine}. The formation of the iminium was always barrierless with no transition structure between the two states, thus again in this case the limiting step of the reaction was the hydride transfer. However, the energetic barrier for all the models was substantially higher compared to those of the cluster models, ranging from 17.6 (*S*_{Lys121.A}) to 27.5 (*S*_{Lys121.B}) kcal mol⁻¹ (Table 5.6). This energy difference was always higher than 10 kcal mol⁻¹, thus suggesting that this mechanism was dramatically impaired by the protein environment.

Model	Energy (kcal mol ⁻¹)			TS Frequencies	Distances	
	Reactant	TS	Product		C-H ⁻	Ir-H ⁻
<i>R</i> _{TSO}	0.0 (0.0)	8.8 (9.1)	-17.8 (-13.7)	-333.6	1.52	1.77
<i>R</i> _{TSN}	0.0 (0.0)	0.7 (1.9)	-26.3 (-23.5)	-90.1	1.92	1.69
<i>S</i> _{TSO}	0.0 (0.0)	1.9 (3.2)	-2.9 (-1.1)	-147.0	1.65	1.73
<i>S</i> _{TSN}	0.0 (0.0)	13.3 (11.2)	-13.4 (-5.5)	-333.2	1.54	1.77
<i>R</i> _{K121.A}	0.0 (0.0)	21.7 (22.3)	2.5 (5.0)	-96.2	1.32	1.91
<i>S</i> _{K121.A}	0.0 (0.0)	17.6 (19.7)	-	-189.6	1.37	1.87
<i>R</i> _{K121.B}	0.0 (0.0)	23.3 (26.7)	10.1 (15.1)	-360.6	1.47	1.78
<i>S</i> _{K121.B}	0.0 (0.0)	27.5 (27.0)	13.3 (13.8)	-432.4	1.58	1.74

Table 5.6 - Energies and relevant structural information from the QM/MM calculations. Gibbs energies are given in brackets.

Regarding the H⁻_{Ir}/H⁺_{med} mechanism, the lowest energy barriers were observed for the *R*_{TSN} and the *S*_{TSO} models with $\Delta E^\ddagger=0.7$ and 1.9 kcal mol⁻¹ respectively (Table 5.6). The other models were found to be significantly higher, with an 8.8 kcal mol⁻¹ energy barrier for the *R*_{TSO} model and 13.3 kcal mol⁻¹ for the *S*_{TSN} (Table 5.6). While the *R*_{TSO} and *S*_{TSN} models presented similar energies than those obtained in their respective cluster versions with differences no higher than 3 kcal mol⁻¹, the *R*_{TSN} and *S*_{TSO} did not follow this tendency. The protein environment lowered their energy barriers more than 10 kcal mol⁻¹ if compared with their respective cluster calculations (Table 5.4 and Table 5.6).

The QM/MM calculations highlighted substantial differences between the mechanisms. Concerning the H⁻_{Ir}/H⁺_{Lys} the limiting step of the reaction (the hydride transfer) presented ΔE^\ddagger ranging from 17 to 27 kcal mol⁻¹ (Table 5.6). On the contrary, the energy barriers identified for the different models of the H⁻_{Ir}/H⁺_{med} mechanism ranged from 0.7 to 13.3 kcal mol⁻¹ (Table 5.6). From all the cases studied, the lowest ones leading to the *R* and *S*

enantiomeric products were the R_{TSN} and S_{TSO} respectively, with energy differences higher than 8 kcal mol^{-1} with all the other possible pathways. Consequently, these two pathways were taken as the most realistic pathways of the ATH in the protein and all the others were discarded.

The great stabilization provided by the protein environment in both the R_{TSN} and the S_{TSO} models could not be reduced to a single factor but rather as a result of the many different weak interactions between the protein/cofactor/substrate triad (Figure 5.8). In both cases the substrate was deep inserted inside the binding pocket and making several hydrophobic interactions with the protein environment. Especially important was the role of both Lys121 in the stabilization. In the S_{TSO} model the Lys121_A was making a CH- π and a cation- π interaction with the substrate, while in the R_{TSN} was Lys121_B the one making the CH- π and an hbond with the -OMe group of the substrate. In fact, mutating this lysine was observed to dramatically decrease the *ee* of the $R_{Ir}[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112A$ ATHase as highlighted by the 52% *ee* towards the *R*-salsolidine in the $S112A\text{-K121T}$ variant, a decrease of almost 45%.⁴⁵ Additionally, the hbond contact between Lys121_B and the -OMe group of the substrate on the R_{TSN} model could be a key element in the enantioselective mechanism of the $[\text{Cp}^*\text{Ir}(\text{Biot-}p\text{-L})\text{Cl}]\text{C}S112A$ hybrid. Nonetheless, the catalysis of the same cyclic imine lacking this functional group resulted in a dramatic drop of the *ee* from 93% to 50%.¹⁸⁴

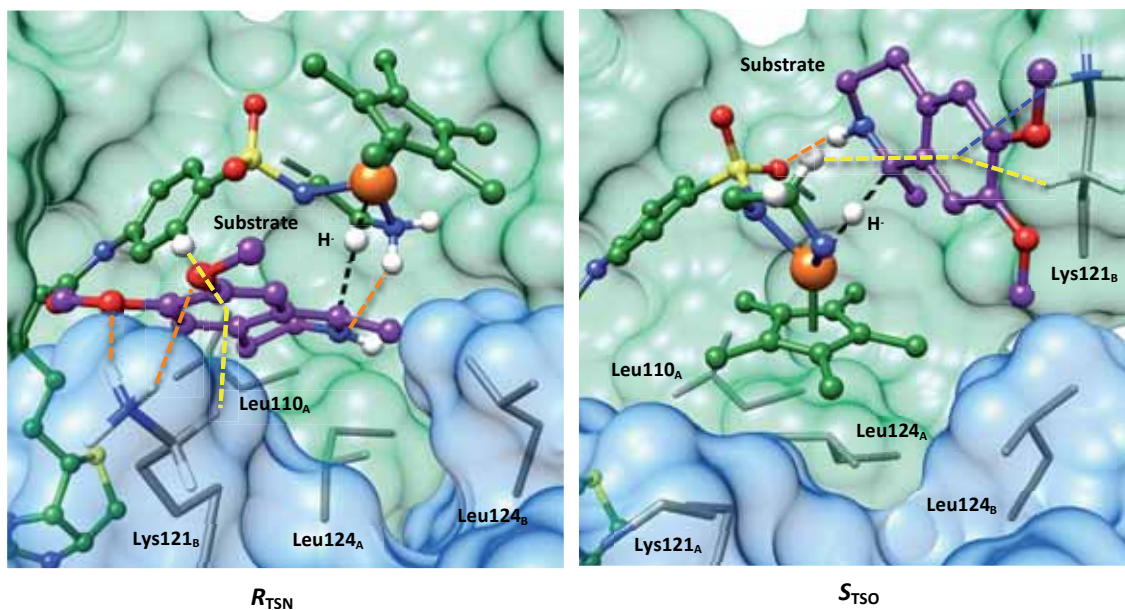


Figure 5.8 - Lowest energy transition state structures leading to the *R* (A, R_{TSN}) and *S* (B, S_{TSO}) enantiomeric products obtained in the protein using the QM/MM approach. Main interactions between the substrate, the catalyst and the protein environment are highlighted: CH- π (yellow dashed lines), Hbond (orange dashed

lines) and cation- π (blue dashed lines). Black dashed lines represent the halfway transfer of the hydride from the Iridium to the C_{imine} .

From the energetic point of view, the most likely reaction mechanism of the ATH of imines by the artificial hydrogenase $R_{\text{Ir}}\text{-[Cp}^*\text{Ir(Biot-}p\text{-L)H]}\subset\text{S112A}$ should be the $H_{\text{Ir}}^-/H_{\text{med}}^+$, and more specifically the R_{TSN} and the S_{T50} models leading to the *R* and *S* enantiomeric products respectively. The computed difference in Gibbs energy between the two pathways was $1.2 \text{ kcal mol}^{-1}$ in favor of the *R* product (Table 5.6), which was predicted to lead to an 80% enantioselectivity in favor of the *R*-salsolidine using the Eq 5.1 reported by Maseras and coworkers.¹⁷² Although one should remain cautious in quantitative interpretations of *ee* from potential energy calculations, this value is in line with the experimental *ee* of 96% obtained at 278K for the *R*-salsolidine.⁴⁵ Additionally, the QM/MM simulation highlighted the vital role of the two Lys121 in the stabilization of the transition state of these two models. These results were in agreement with the decrease on the *ee* observed experimentally when mutating this position or deleting the -OMe group of the substrate.

III.VI - Conclusions

The integrated computational approach presented in this part of the thesis successfully provided with the key molecular events responsible for the enantioselectivity observed for the $R_{\text{Ir}}\text{-[Cp}^*\text{Ir(Biot-}p\text{-L)Cl]}\subset\text{S112A}$ artificial hydrogenase. It sheds light onto the homogeneous and enzymatic mechanism in which Noyori's like catalyst reduce imines. The protocol unveiled how the protein environment helped in stabilizing the transition state by simultaneously making hbond contacts and CH- π interactions between Lys121 and the substrate. The gathered structural data will help in future rational design of the hybrid.

The good agreement between the experimental observations and the theoretical model confirms the efficacy of our integrative protocol in the study of complex biological/biometallic systems. The combination of several different molecular modelling techniques allowed us identify the most likely transition states while exploring large conformation spaces. For this reason, our approach can be of particular interest in the study of those bioinorganic entities where the characterization of the transition state is troublesome or there is a large conformational space available.

All images, tables and schemes from this section have been reprinted/modified with permission from (see reference ⁸⁸). Copyright 2014 American Chemical Society.



**Towards a Novel
Integrative
Computational Platform**

I - Introduction

The study of the kind of artificial enzymes reported in this Ph. D. thesis is of the most difficult tasks a molecular modeller could encounter. The events we need to model involve a broad spectrum of molecular processes including recognition, dynamics and catalysis which, at the current state of molecular modelling, can only be dealt separately with different kind of methodologies (i.e. MM and QM approaches). One of the natural solutions to study such complex systems is to combine these methodologies together in a unique workflow. The application of these integrative or multi level strategies is spreading over the entire chemical biology community but, despite the promising results reported so far^{72,84,88,185} their range of applications is still rather unexplored and their availability is fairly limited.

One of the key elements for an integrative approach to be efficient is to enable the free flow of data between the different modelling softwares it encompasses without losing track of all the molecular data and allowing a competitive graphical section. Nowadays, there are commercially available suite packages following this principle for the study of complex biological problems.¹⁰³⁻¹⁰⁵ However, they present some limitations that may restrict their usability. First, they use a “black box” approach in which the user has neither control over the algorithm nor over the data flow. In fact, the user cannot think out-of-the-box of what has been dictated in the program, thus making it particularly difficult to study non-standard problems. Additionally, many of them use their own scripting language, which implies that the user has to learn it even if it is not spread amongst the computational community. Finally, they are particularly expensive and not every group can afford them.

I.I - An integrative platform made for everyone

The computational chemistry community is vivid and rapidly growing. Many groups around the world offer freely (or for a relatively low amount of money) efficient softwares designed for the study from simple to complex biological and chemical problems. Moreover, most of them share an open-code concept in beneficial of the end user, thus fostering the creation of numerous communities and promoting the exchange of ideas and information. However, a solid wall is encountered when trying to combine them in an integrative approach. They have not been designed to be used this way and these attempts may result in the loss of accuracy or data due to the different way each program interpret/express the molecular information. An integrative platform based on those same free-of-charge and open-code standards would overcome those problems and

enable the design of multilevel strategies combining several different molecular modelling approaches for the study of complex systems.

In this chapter we present our on-going efforts to create an open-code and freely available integrative platform, which cornerstone will be centered around two different modelling packages: the UCSF Chimera¹⁰⁷ and the Molecular Modelling Toolkit (MMTK)¹⁰⁶. The former provides an intuitive and friendly user interface, as well as later-generation graphical environment, while the later contains the core algorithms to perform the molecular modelling calculations (Molecular Dynamics (MD), Normal Modes Analysis (NMA) and many other approaches). Both programs are free-of-charge, open-code and based on the same programming language: python (Figure 6.1). Our platform will also inherit all those basic features with the aim of becoming a widespread tool in the modellers community.

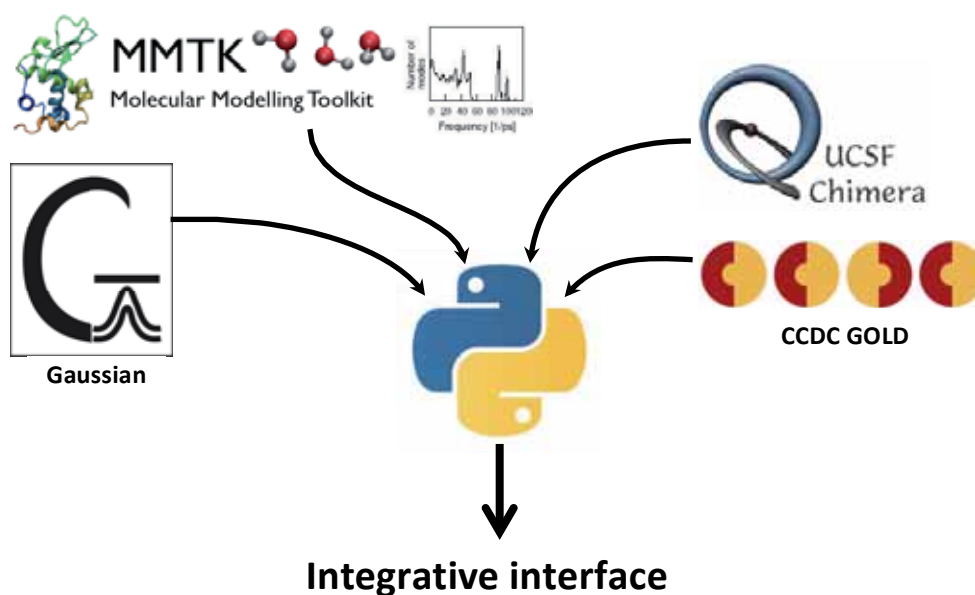


Figure 6.1 - We used python to merge the MMTK and the UCSF Chimera in what will be the cornerstone of our integrative platform, but we also designed some interfaces to analyze the results coming from the Gaussian suite (including QM and QM/MM calculations) and from the CCDC GOLD package (Docking). Altogether, this is particularly beneficial as: (i) it allows the free flow of information between the two programs, (ii) it is easy and fast to learn with a relatively low learning curve and (iii) it is a widespread programming language with many other applicabilities.

At the time of publication of this Ph. D. thesis dissertation, our integrative platform allowed to perform MD (MD) simulations and NMA and contained a graphical interface to prepare/analyze QM/MM calculations (using the ONIOM¹³¹ scheme of Gaussian¹²³) and to analyze docking results (from GOLD)¹²⁷ and QM simulations (from Gaussian).¹²³ This is just

the beginning of the road towards a complete integrative platform for molecular modelling. Several more implementations will be made in the future, including an algorithm able to provide with enhancing mutations/chemical modifications for the design/optimization of biological complexes.

II - Molecular Dynamics simulation interface

In a MD simulation the trajectory of all the atoms of a system at a given temperature are calculated in order to study how it evolves through time. As this approach is based on molecular mechanics it is not restricted to small size systems and big ones (e.g. proteins) can be studied. For this reason, this methodology is a standard on the modellers community used to explore conformational spaces of chemical and biochemical systems. MMTK provides with an efficient and relatively easy driver for MD simulations and we therefore embarked in set up such integration in our Chimera platform. One of the particular interests of the MMTK implementation is the possibility to run highly scalable calculations on the same desktop without requiring direct interaction with clusters and to be applicable to small systems as well as large ones. This represents a unique feature of the MMTK suite and can have a huge potential on the study of molecular problems.

In the MD interface we designed all the parameters needed are classified in different tabs (Figure 6.2). The first two (*Set Up* (1) and *Solvation* (2)) contain all the options needed to prepare the system for the simulation. On the following three (*MD Options* (3), *Constraints* (4) and *Miscellanea* (5)) the user can define all those parameters affecting the actual MD (e.g. number of steps, temperature, pressure, etc.). The final one (*Running* (6)) is where the user can define how to launch the simulation (i.e. number of processors to use, background or foreground, etc.). All the different tabs are thoroughly explained in the following pages.



Figure 6.2 - The different tabs available on the MD general interface

Set Up tab

First thing to do is to *Set Up* the system for the MD simulation (Figure 6.3). In this tab the user can add whatever is missing in the system prior to begin the simulation (e.g. hydrogens, charges, etc.) by calling the *Dock Prep* (2) routine of Chimera. This interface will check that everything is fine with the chimera model (e.g. no missing/incomplete

residues) and fix all the wrong aspects it encounters. The selected options in this interface can be stored for later use in another system (1).

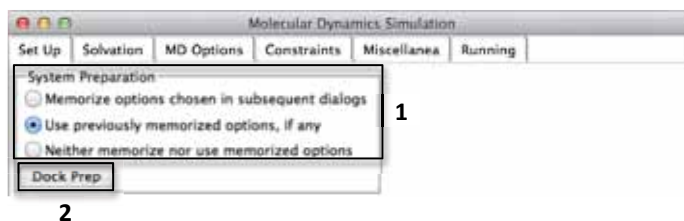


Figure 6.3 - Options available on the *Set Up* tab of the MD interface

Solvation tab

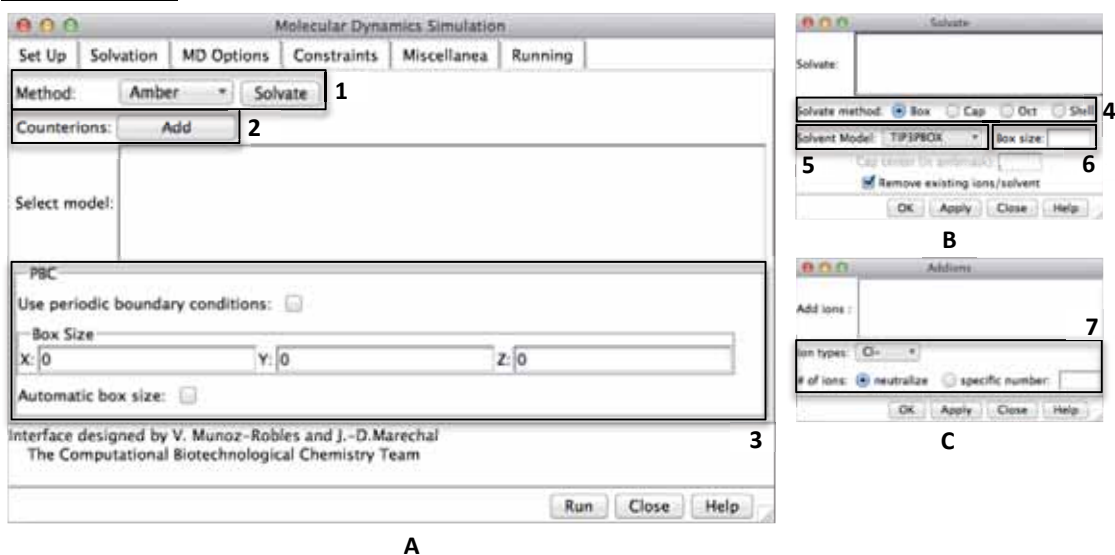


Figure 6.4 - Interfaces dedicated to the solvation of the system. Panel A: Solvation tab on the MD interface, Panel B: Amber solvate routine interface of Chimera, Panel C: Addions interface of Chimera.

To *Solvate* the system (Figure 6.4) our interface makes use of the Amber method (1) as implemented in Chimera (Panel B). The user has to select the solvation method (4) (with the corresponding parameters (5)) and the solvent model (6). The counterions needed to neutralize the system can also be added at this point by invoking the *Addions* interface of Chimera (2 and Panel C). Additionally, the user can also define whether to use or not Periodic Boundary Conditions (PBC) (3) along the MD simulation, but it can only be used in box-shaped systems in the current version of the interface. Additional implementations to enable the use of Stochastic Boundary Conditions (SBC) are currently under development.

MD Options tab

In our interface, the MD simulation is divided in three different stages: (i) *Minimization*, (ii) *Equilibration* and (iii) *Production* (Figure 6.5). To obtain consistent results, the user needs to previously have some knowledge on MD simulations to set up what he/she considered the most convenient parameters in each stage. For clarity, all the details of the different interfaces are not reported in this part of the thesis and the reader can refer to the manual pages of the interface in the UCSF Chimera web page. Novel functionalities are frequently added to merge chimera with MMTK.

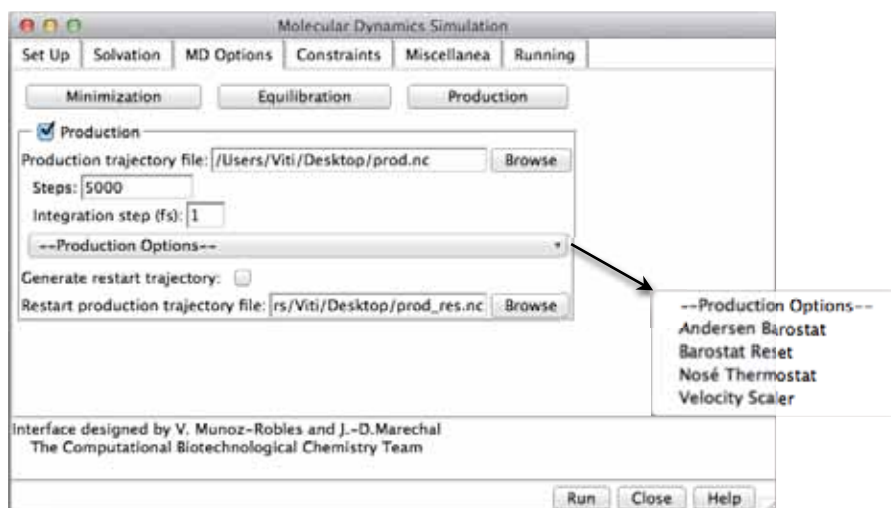


Figure 6.5 - Different options available for the production stage of the MD simulation on the Chimera interface.

Constraints tab

Depending on the purpose of the study the MD can be applied to only one part of the system, thus reducing the conformational space available. For those cases, our interface offers the possibility to add a motion restraint on a selected group of atoms, which will be kept frozen all along the simulation. This option can be found on the *Constraints* tab (Figure 6.6).

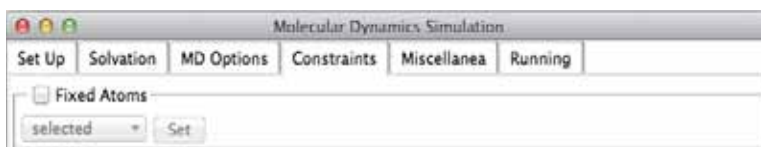


Figure 6.6 - At the current stage of development in the MD interface the constraint tab only allows to freeze a selection of atoms during the MD simulation.

Miscellanea tab

Certain parameters of the force field can be modified and new ones can be added on the *Miscellanea* tab (Figure 6.7). At the moment, the user can include the *Translational* and *Rotational Remover* (1) features in the simulation, which will prevent to obtain any frame of the trajectory presenting those kinds of movements and dramatically simplify the output. Additionally, the standard *Electrostatics* and *Lennar-Jones* potentials of the force field can be modified in several different ways (2). One must be very cautious with those changes as they can tremendously affect the results of the simulation. For more information about this matter, please refer to the MMTK User's Guide.¹⁸⁶

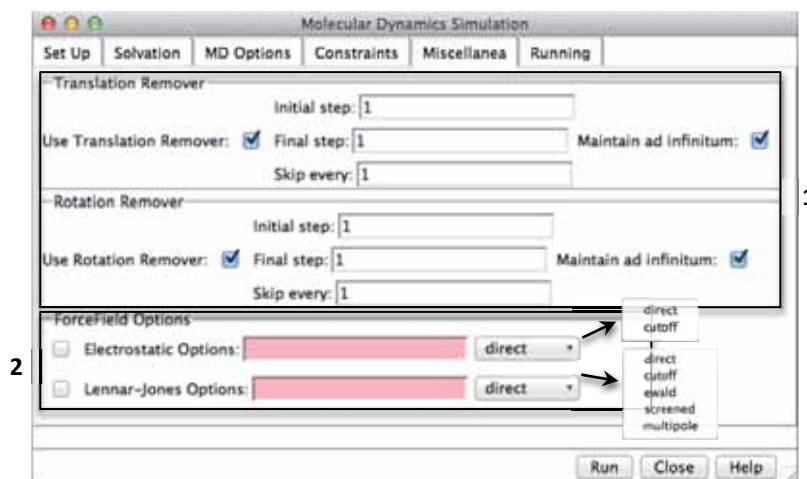


Figure 6.7 - The Miscellanea tab of the MD interface showing additional force field parameters than could be added/modified

Running tab

The *Running* tab is where the different options to run the MD simulation can be defined, including: (i) writing an MMTK input to perform the simulation using the python interpreter instead of Chimera (1), (ii) running the MD on Chimera and visualize the trajectories as they are being computed (4) or (iii) running the simulation normally using chimera (Figure 6.8). The number of processors that will be used for the simulation can also be defined at this point (2). To simplify the resulting trajectory, chimera can ignore some of the frames and load only one in every pre-defined interval (3).

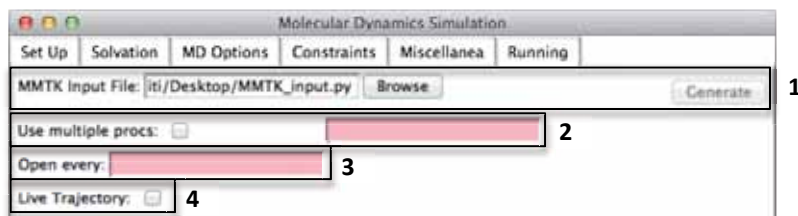


Figure 6.8 - Print screen of the running tab.

II.I - MD visualization

Once the simulation is finished, the resulting trajectory is loaded into the *MD Movie Dialog* interface of Chimera, which allows the user to move through the different frames (Figure 6.9). Additionally, this interface contains several different tools to analyze the trajectory (i.e. clusterization tool, RMSD maps, etc.), including the plotting of several different parameters (i.e. potential energy, temperature, etc.).

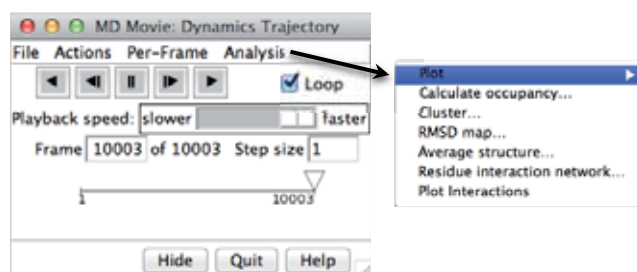


Figure 6.9 - MD Movie Dialog interface of Chimera adapted to load the trajectories of our MD implementation

III - Normal Modes Analysis interface

Calculating the normal modes of a protein allows us to harmonically move along the potential energy well where the system is located to study large collective movements (e.g. opening/closing the active site through a molecular hinge). The low amount of time/resources needed to run this kind of simulations is one of the main assets of the NMA. In fact, representing those kinds of motions by other means (e.g. MD simulations) would need to run large resource-consuming simulations. NMA has since-long become a widespread methodology in the modellers community with some very successful results reported up to date.^{96,187-190}

From all the available molecular modelling softwares, MMTK is one of the best options to perform a NMA. It has a highly efficient code allowing different kinds of NMA calculations, including: (i) calculations using a standard force field (Amber) and (ii) a simplified Elastic

Network scheme. This last option deserves a special mention; it is only available in the MMTK package and allows much faster NMA using a reduced form of the molecular system and a simplified force field. But even with all these simplifications, it has proven to be particularly useful and accurate in the motion study of big systems.^{190,191}

One of the advantages of the implementation of the NMA into the UCSF Chimera is that the user can have a visual description of all the calculated normal modes once the routine is finished. Additionally, several other analysis tools (some already present in Chimera, some implemented by us) are available on the interface to analyze the results, including: structural analysis tools (i.e. angles, distances, dihedrals), visual depiction of the modes displacement, conformation analysis of different configurations of the same protein using NMA, etc.

The interface can be accessed directly from the menu and is available since version 1.8 of UCSF Chimera. The main interface (Figure 6.10) is structured in six different parts, representing each of the different aspects of the NMA calculation the user will be able to personalize.

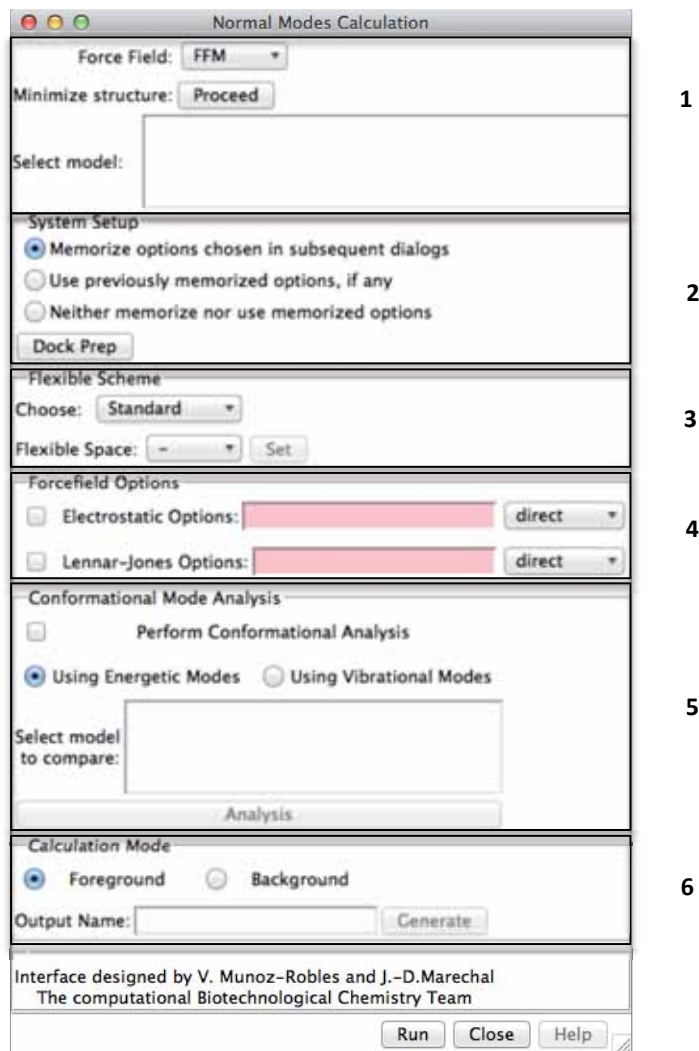


Figure 6.10 - General NMA interface implemented in the Chimera suite. Each of the different parts of the interface have been highlighted and numbered.

First, the user can select on which of the Chimera models the NMA will be performed and the kind of NMA simulation (1). In its actual version, the available NMA schemes are^a: (i)

^a The simplified NMA using the C α of the protein (using an optimized force field for this purpose) is not yet available at the time of publication of this thesis.

Full Atoms^b (*FFM*, standard AMBER force field) or (ii) Elastic Network (*ENM*, simplified force field from the MMTK). Most of the properties needed in the set up of the system (i.e. add missing hydrogens, add charges, eliminate ions, etc.) can be specified by the user (2) and will be added automatically by the *DockPrep* routine of Chimera. By default the NMA will take into account the whole chimera model, but the user can also choose to compute the normal modes only on a specified part of the system (3). This is a straightforward process where the atom selection can be performed directly by picking atoms on the screen or using specialized functions of chimera like command lines and python scripting. Because of the modulability of MMTK, several internal functions of the force field can be defined like how the Electrostatic and the Lennard-Jones parameters are computed in the force field (4). It is important to be extremely cautious when modifying those features of the force field, as these changes could dramatically impact the resulting normal modes and lead to artefactual results.^c The user can compute the normal modes in two different ways (6): (i) in the foreground or (ii) in the background. The former will hang Chimera until the calculation is finished and display the results on the screen while the later allows the user to continue working with the same Chimera session and give a notification once the NMA is finished to load the normal modes.

One additional feature added on the interface is a conformational analysis between two forms of the same protein using a NMA (5). To do so the two conformations must be loaded in separate Chimera models (one of them being the reference model) and have the same atom numeration. The user can select which kind of normal modes (energetic or vibrational)^d will be used to perform the analysis. The algorithm will calculate the modes on the reference model and provide a plot indicating the similarities between the second conformation and each of the computed modes (Figure 6.11).

^b Please notice that for full atom calculations it is mandatory to perform an intensive minimization of the system; something the user can do directly from the interface by invoking the Chimera minimizer

^c For specific information about each options, the user need to refer to the MMTK user's guide for which direct access to the webpage will be set up in a near future.¹⁸⁶

^d Energetic Modes find the eigenvalues and eigenvectors of the plain Hessian (second derivative matrix of the potential energy), whereas Vibrational Modes calculates the Hessian in mass-weighted coordinates. The mass-weighted normal modes are then re-weighted to produce standard cartesian coordinates, which can be used for visualization or various calculations. But the re-weighted modes are not long orthogonal, which is why you can't use them to do decompositions of distance vectors.

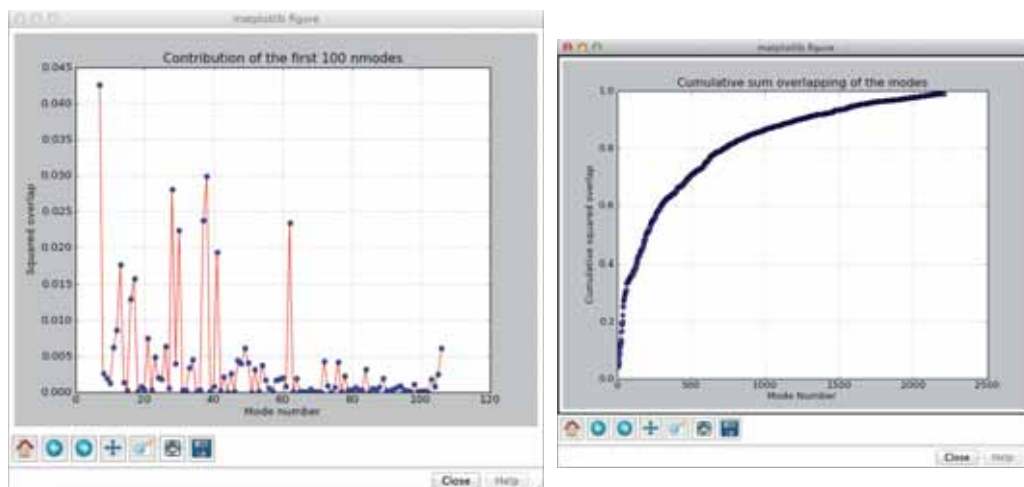


Figure 6.11 - Designed interface to plot the results from the structural NMA analysis between two different conformations of the same protein. **Left:** Contribution of the first computed 100 modes. **Right:** Cumulative sum overlapping of the modes. By clicking on the different modes on the left panel the corresponding mode will load on the screen.

III.I - NMA visualization

Once the calculation is finished, the resulting normal modes will be transferred into the visualization interface. This interface is composed of two parts, the MD Movie and the Mode List (Figure 6.12).

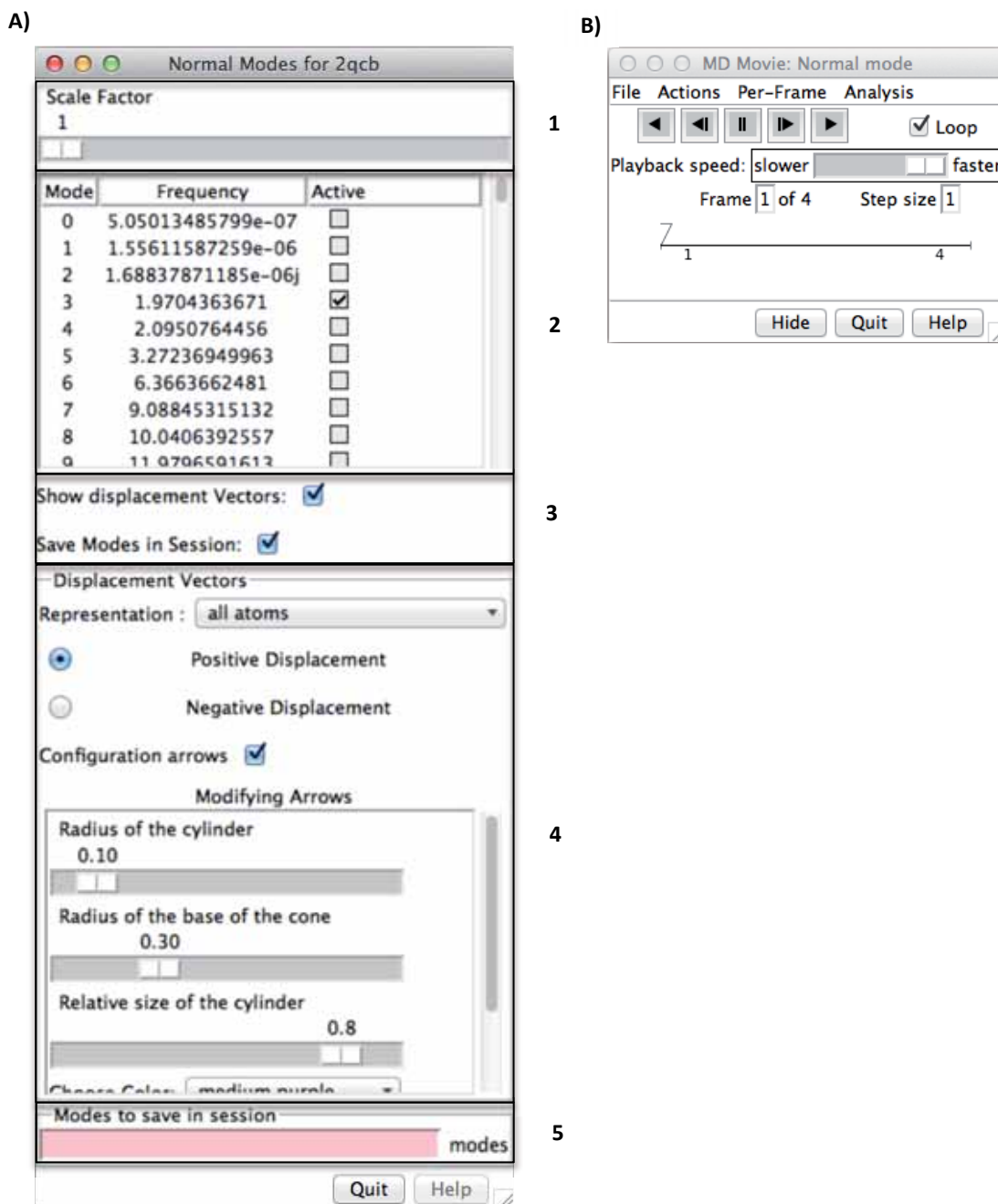


Figure 6.12 - Designed interfaces for the visualization of the normal modes. Panel A represents the interface containing the modes list with all the tools to visually analyze the corresponding displacement. Panel B is the interface used to move through the different frames associated to each mode.

The Mode List (Figure 6.12, panel **A**) encompasses a table with all the calculated normal modes, sorted by their frequency value. The selected mode on the table (**2**) will be visualized on the Chimera main screen. It is to notice that only one mode may be selected at each time. The scale of the movement along the selected mode can be modified using the *Scale Factor* bar (**1**). Additionally, in this interface the user can display the arrows showing the displacement vectors of the mode (**3** and Figure 6.13), with several options available to modify its settings (i.e. color, size, etc.) (**4**). Moreover, as the number of modes can be dramatically high depending on the size of the system (and in many occasions only those of lowest energy might have any relevance), the user can define how many modes to save on the session (**5**), making the storing file significantly smaller and more easy to handle.

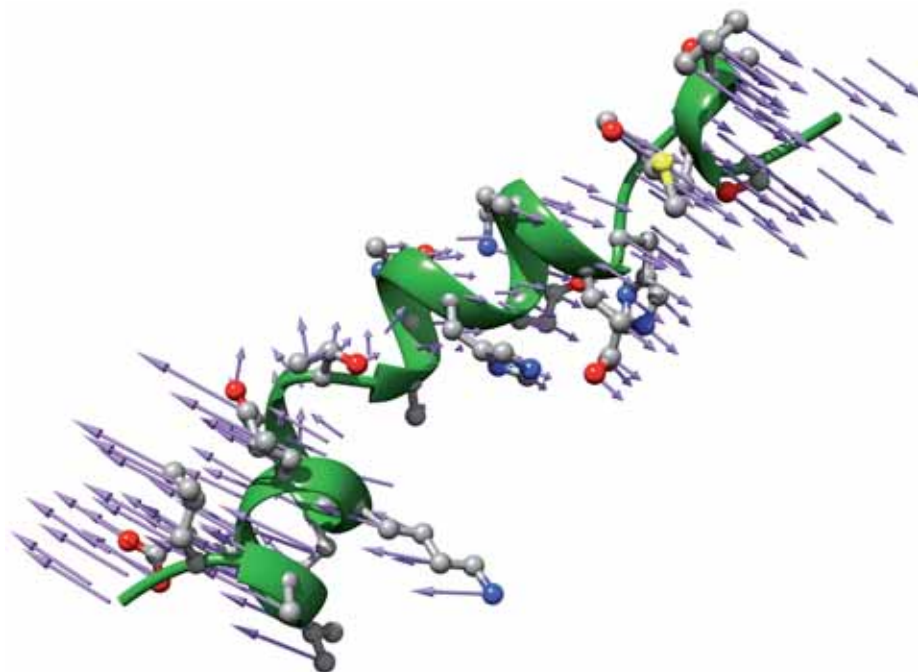


Figure 6.13 - Arrow representation of the computed normal modes for a peptide.

The MD Movie interface (Figure 6.12, panel **B**) is used to move the structure along the selected normal mode on the Mode List interface. Each mode is a displacement vector, which can either be summed or subtracted to the coordinates of the system, thus creating two different movements (a positive and a negative one, respectively). In this interface the user can choose several playing options to move along those frames. Additionally, other standard analysis options from the default MD Movie of Chimera are available on the top menus of this interface.

IV - Novel interfaces for Gaussian and GOLD

IV.I - Gold visualization interface

In our different studies of artificial metalloenzymes we selected the CCDC GOLD¹²⁷ package to perform the docking calculations as it is the one on which the group has major expertise and also because it is of the few which incorporates parameters to describe metal atoms.^e However, GOLD has not been optimized for being part of the computational chemistry pipeline neither for high quality graphics.

To optimize the analysis of the dockings we designed an interface in Chimera able to read and load on the screen all the results and their related information (Figure 6.14). Each entry of the table belongs to one docking solution and displays the break down of the energetic terms of the Scoring function used. By clicking on the entries of the table the corresponding docking solution and the protein receptor will load on the screen. This allows a rapid visualization of both the energetic terms and the resulting structure.

S	Ranked	Dock_Solution	Score	DG	S(hbond)	S(metal)	S(lipo)	H(rot)	S(int_hb)	DE(clash)	DE(int)	S(protein)	time	Rotamers
V	1	9	31.35	-21.64	0.0	0.0	163.42	1.16	0.61	-12.53	2.43	0.0	39.062	LYS109:1,LEU112:1,LYS
V	2	12	31.29	-22.05	0.0	0.0	166.95	1.16	0.61	-12.43	2.57	0.0	41.291	LYS109:1,LEU112:1,LYS
V	3	20	31.16	-22.38	0.0	0.0	169.79	1.16	0.61	-11.62	2.45	0.0	39.05	LYS109:1,LEU112:1,LYS
V	4	5	31.11	-21.31	0.0	0.0	160.64	1.16	0.61	-12.55	2.41	0.0	39.474	LYS109:1,LEU112:1,LYS
V	5	14	31.11	-23.21	0.0	0.0	176.86	1.16	0.61	-11.06	2.57	0.0	41.535	LYS109:1,LEU112:1,LYS
V	6	4	31.03	-21.41	0.0	0.0	161.52	1.16	0.61	-12.53	2.41	0.0	39.39	LYS109:1,LEU112:1,LYS
V	7	7	30.65	-21.0	0.0	0.0	157.95	1.16	0.61	-12.74	2.4	0.0	39.174	LYS109:1,LEU112:1,LYS
V	8	19	30.64	-22.39	0.0	0.0	169.82	1.16	0.61	-11.05	2.38	0.0	39.77	LYS109:1,LEU112:1,LYS

Chimera Model #0.1

Change Compound State

Viable Deleted Purged

Hide Quit Help

Figure 6.14 - Interface for the visualization of the docking results. All the docking solutions are ranked according to their Scoring. Each entry of the list displays all the energetic terms of the Scoring function.

Unfortunately, at the moment, this interface cannot be released because few elements for the parsing come from a protected code. A modification of the interface is actually under development.

^e The metal-atom types are used to describe the metal in a metal-containing ligand when it is bound to a protein (e.g. the iron in the bound porphyrin of cytochrome P450).

IV.II - Gaussian visualization and ONIOM inputs set up interfaces

We used the Gaussian package¹²³ (both Gaussian03 and Gaussian 09) to perform all our QM calculations (and ONIOM¹³¹ in the QM/MM cases). For our integrative platform to gain access to all the molecular information coming from the Gaussian outputs we created an interface to read those files in Chimera. It loads the QM results in the screen using the MD Movie Dialog to store the different points (for optimizations) or frequencies (for frequency analysis) (Figure 6.15). All the related information of the structure (e.g. forces, energies, etc.) can be found in the Analysis menu of the MD Movie (Figure 6.16).

Loading the Gaussian results on Chimera is particularly useful when dealing with QM/MM ONIOM results. In contrast with other visualization programs, Chimera can display those results using the simplified ribbon scheme for the protein (Figure 6.17). This greatly simplifies the visual analysis of the resulting QM/MM structures.

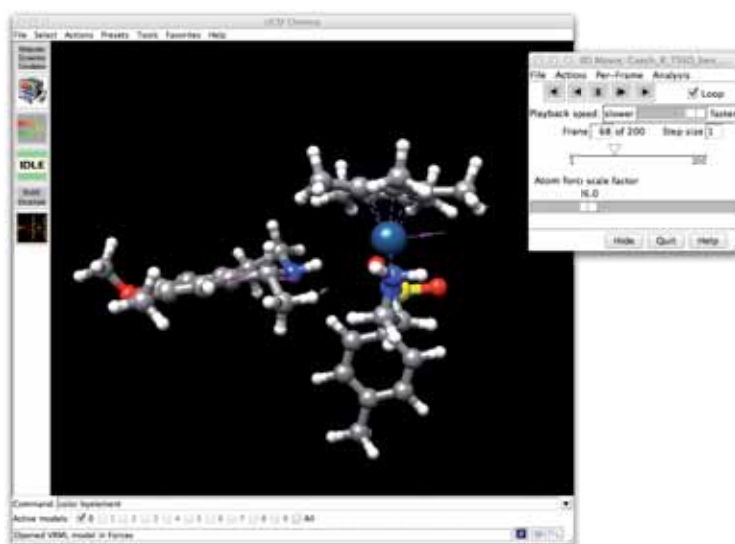


Figure 6.15 - Chimera loading the Gaussian results for an optimization. The MD Movie allows moving through the different points of the calculation. The arrows represent the atom forces and their direction.

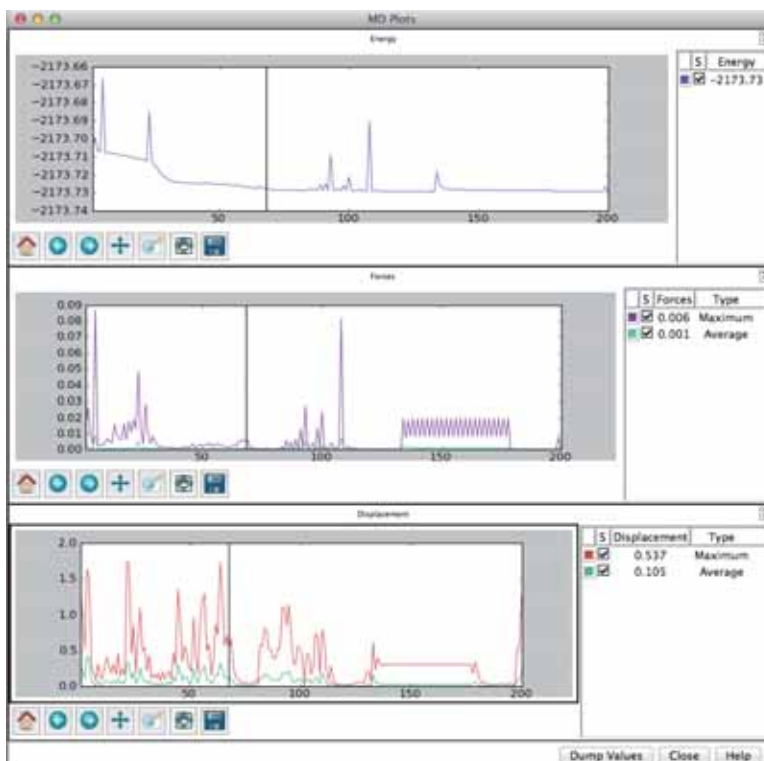


Figure 6.16 - Plotting of the different energetic terms of the Gaussian calculations. By clicking on each point of the plots, Chimera will instantly go to that particular frame and load the corresponding structure on the screen.

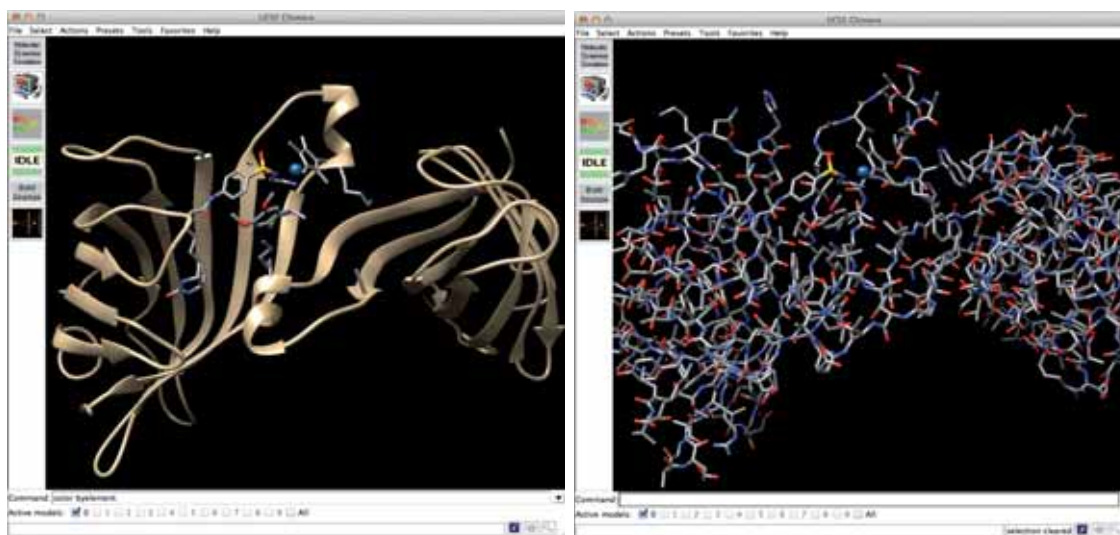


Figure 6.17 - QM/MM results displayed in Chimera with (Left) and without (Right) ribbon simplification of the protein.

Additionally, we added an interface to prepare QM/MM inputs for the ONIOM Scheme in Gaussian (Figure 6.18). This way we can prepare the system in Chimera and directly start the ONIOM calculation. By selecting the atoms on the screen (or using simple python

commands) the different layers of the ONIOM calculation can be defined, as well as specifying which of them will be flexible or will be kept fixed during the simulation. All the other information needed (e.g. input name, charges, multiplicity, etc.) can be manually introduced by the user.^f

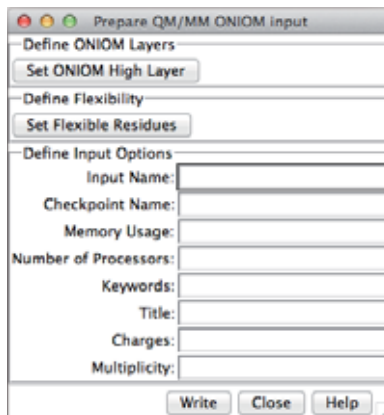


Figure 6.18 - Interface to set up the ONIOM inputs using Chimera.

V - Perspectives

In this part of the thesis we have set up the cornerstone of our integrative platform using the UCSF Chimera and the MMTK. At the current stage of development we can perform MD simulations and NMA. Additionally, we also added some implementations for an enhanced analysis of QM and QM/MM calculations (Gaussian) as well as docking simulations (GOLD). But all these implementations are just the tip of the iceberg of what our platform can really do. Now the box is open to add several new tools in the future for the study of complex biological and chemical systems.

^f At the current of development, the resulting input must be loaded with the GaussView program to correctly add the *Z-matrix* of the system.



General Conclusions

I - Conclusions

With the raise of chemical and structural biology, the design of artificial metalloenzymes is becoming a viable alternative to classical homogeneous catalysts and natural enzymes. Several examples have now showed that they can perform complex chemical reactions in a selective way and perfectly maintain the environmental friendliness of natural biocatalysts. Particularly fruitful are the artificial metalloenzymes based on the insertion of an homogeneous catalyst inside a biological scaffold. In this thesis we have intensively applied several different computational approaches in the study of those hybrids in order to aid in both their design/optimization processes.

Our early studies focused on characterizing the binding of the inorganic cofactor inside the biological scaffold.

- 1) We designed an integrative protocol involving different molecular modelling techniques, including docking and QM/MM approaches, and used an artificial oxidase reported by Ueno and coworkers to benchmark it. We successfully reproduced all the particularities of the X-ray structure in an *in-silico* model, thus validating our multilevel protocol and confirming our initial hypothesis: the combination of different modelling approaches could overcome the limitations found in their standalone application.
- 2) A computational study was performed on different abzymes designed by the group of Jean-Pierre Mahy. These species were composed by carboxyphenyl substituted porphyrinic-like catalysts inserted into two different antibodies: (i) 13G10 and (ii) 14H7. Although the X-ray structure of the latter antibody was not suitable for the docking study, we successfully obtained the models of the hybrids obtained with the former one. The molecular data gathered from them unveiled the vital role played by the AsnH33/HisH35 polar patch in the binding process of the Fe(ToCPP) and the two Fe(DoCPP) cofactors. Moreover, a docking study with the imidazole in those models also shed some light on the catalytic mechanism. First, the imidazole showed to preferentially bind the active face of the iron in the Fe(ToCPP) \subset 13G10 hybrid, while in the two Fe(DoCPP) \subset 13G10 it prevailed the opposite side. This could explain why large concentrations of imidazole inhibited the Fe(ToCPP) \subset 13G10 abzyme, as they could compete with the ABTS substrate. Additional, a structural inspection of our models identified probable residues/functional groups that could aid in the cleavage of the water peroxide,

including TyrH52, AsnH33, TyrL34 or one of the carboxyphenyl substituents. All this structural information could be used in the design of novel mutations/chemical modifications towards an improved generation of oxidative abzymes.

- 3) Finally, we performed a docking study in two different ATHases designed by Thomas R. Ward and coworkers. The aim was to characterize the enantioselectivity in the binding of a biotinylated Noyori's like catalysts in two different mutants of the Streptavidin complex: the S112A and the S112K. The results described a situation we defined as *induced lock-and-key fit*: the protein environment favors the binding of an enantiopure configuration at the metal center which, in turn, determines to which prochiral face of the substrate the hydride will be delivered. Additional interactions between the protein environment and the substrate helped to fine-tune the final enantioselectivity.

After studying the binding of the organometallic compounds to the proteinic receptor, we decided to study the catalytic mechanism of different artificial metalloenzymes through computational means.

- 4) The first systems were two artificial metalloenzymes obtained by Jean-Pierre Mahy and coworkers. Both of them used a Xln10A as scaffold but differ in the inorganic cofactor. The first was a porphyrinic-like catalyst while the second one was a salophen-like. We used a docking approach to decode the interactions between the inorganic cofactors and the xylanase receptor. The results suggested that the Mn(TpCPP) cofactor had a better stability because: (i) it was deeper inserted in the binding pocket and (ii) it could create a wider hbond network with the different carboxyphenyl substituents. To unveil the origin of the *ee* of the Mn(TpCPP)-Xln10A we should ideally resort to QM/MM approaches. However, their elevated resource-consuming ratio and the limited trust of our model discouraged us from applying them. Instead, we performed a docking study with the different styrenes tested experimentally with the Mn(TpCPP) based metalloenzyme and see if they could shed some light on this matter. The results systematically displayed a little preference (up to 1 Score Unit) for the *S* product except for the *p*-nitrostyrene (did not enter the binding site) and the *p*-methoxystyrene. In this latter case the results were inverted and the *R* enantiomer prevailed by 1 Score unit. The origin of this shift was an additional hbond between the -Ome group of the substrate and residue Tyr172. Additionally, our model structure also suggested that Arg139 could have a relevant role in the catalysis as it

could either allow or deny the entry of the substrates in the active site of the enzyme. It is pleasing to see that despite the qualitative nature of the docking results they could be largely corroborated by the experimental observations, thus giving us confidence in our models and allowing us to use the gathered structural data for future rational optimizations.

- 5) Our final study was performed on the most effective ATHase designed by the group of Thomas R. Ward. This metalloenzyme was obtained by incorporating a biotinilated Noyori's like catalyst inside the S112A mutant of the Streptavidin complex. The resulting enzyme afforded an astoundingly 96% *ee* for the *R* amine in the ATH of a cyclic imine. First focused on understanding the catalytic cycle in a small cluster model of the whole biometallic complex. We concluded that the ATH of imines should undergo the so-called "*ionic mechanism*" and selected two different mechanism to scale the study up to the real system size. They differ in the proton source of the reaction: (i) in the H^-_{Ir}/H^+_{Lys} it was the nearby Lys121 and (ii) in the H^-_{Ir}/H^+_{med} it was a hydronium from the media. In this later case two different pathways were considered depending on the orientation of the substrate: (i) the TSN and (ii) the TSO. We docked the transition state of all these pathways and the resulting structures were then passed onto a QM/MM scheme to obtain their corresponding reaction profiles. The results allowed us to discard the H^-_{Ir}/H^+_{Lys} mechanism and highlighted the R_{TSN} and S_{TSO} pathways from the H^-_{Ir}/H^+_{med} mechanism as the most probable ones for both the *R* and *S* enantiomeric products respectively. The energetic difference between them accounted for up to 80% *ee*, pretty close to the actual 96% value observed. The protein environment played a vital role in reducing the energetic barrier of the reaction, lowering it by almost 10 kcal mol⁻¹ in some cases. In this aspect, the neighboring Lys121_A and Lys121_B had a vital role in the stabilization of the S_{TSO} and R_{TSN} pathways respectively. These residues participated in most of the stabilization interactions in the protein/catalyst/substrate triad, including hbonds, CH- π and cation- π . However, these interactions alone could not account for the dramatic degree of stabilization. In fact, this event cannot be traced down to a single cause but rather to a many-interaction effect provided by the protein environment. The structural information obtained from the characterizations of the different transition states could now be used for the design of novel mutations allowing further optimization of the system.

In this first part of the Ph. D. dissertation we extensively demonstrated the effectiveness of molecular modelling approaches in the structural study of the artificial metalloenzymes. Depending on the nature of the particular molecular problem, in some cases we can deal with it using standard molecular modelling approaches, while in others we need to combine them in an integrative protocol. Despite the situation we are dealing with, these theoretical methodologies could successfully provide novel mutations of the biological scaffold or chemical modifications of the homogeneous catalyst and guide towards an efficient rational design/optimization of bioinorganic complexes like artificial metalloenzymes.

During this Ph. D. thesis we needed to combine several modelling softwares to tackle most of the molecular problems studied due to their high complexity. We realized that this task was far from easy due to the differences between each program to store/analyze the molecular data. An integrative platform encompassing all those methodologies is therefore needed to overcome these problems. Unfortunately, the few ones commercially available have a rather restricted usability as they follow a “black-box” concept and are usually out of the league for most groups as they are particularly expensive. All these reasons encouraged us to design an integrative platform that should be “open-code” and free-of-charge, available for every member of the modellers community.

- 6) The cornerstones of our integrative platform were the UCSF Chimera and the MMTK. The later provided a user-friendly interface with a high-standard graphic visualization while the former contained all the algorithms needed to perform the molecular simulation. At the current stage of development our platform contained all the necessary elements to perform MD simulations and NMA, as well as several tools to analyze the results. Additionally, two different interfaces were also developed to provide a highly efficient way to analyze the results from two of the most used modelling softwares: (i) the docking suite GOLD and (ii) the QM package Gaussian. Specially interesting is the Gaussian interface in the case of QM/MM calculations as it allowed to simplify the system using the ribbon representation scheme and to efficiently prepare the corresponding inputs.

Several more implementations to our integrative platform are on its way, including a chemogenetic algorithm designed to analyze the protein/cofactor/substrate triad an offer novel mutations and chemical modifications guiding the rational design of the system. Once the whole interface is available in the stable version of UCSF

Chimera, we expect the community to have an active role developing their own tools and expanding its range of applicability.

B

Bibliography

- (1) Berzelius, J. J. *Annls. Chim. Phys.* **1836**, *61*, 146.
- (2) Mathews, C. K.; Holde, K. E. van; Ahern, K. G. *Biochemistry*; 1999.
- (3) Fischer, E. J. *Am. Chem. Soc.* **1894**, *27*, 2985–2993.
- (4) D. E., K. *Proc. Natl. Acad. Sci. U. S. A.* **1958**, *44*, 98–104.
- (5) Monod, J.; Wyman, J.; Changeaux, J.-P. *J. Mol. Biol.* **1965**, *12*, 88–118.
- (6) Tummino, P. J.; Copeland, R. A. *Biochemistry* **2008**, *47*, 5481–5492.
- (7) Zhou, H. X. *Biophys. J.* **2010**, *98*, 15–17.
- (8) Vogt, A. D.; Cera, E. Di *Biochemistry* **2012**, *51*, 5894–5902.
- (9) Changeaux, J.-P.; Edelstein, S. *Biol. Reports* **2011**, *3*, 19.
- (10) Straathof, A. J. J.; Adlercreutz, P. *Applied Biocatalysis*; 2nd Editio.; Harwood academic publishers: Amsteldijk, 2000; p. 443.
- (11) Schmid, A.; Dordick, J. S.; Hauer, B.; Kiener, A.; Wubbolts, M.; Witholt, B. *Nature* **2001**, *409*, 258–268.
- (12) Otten, L. G.; Hollmann, F.; Arends, I. W. C. E. *Trends Biotechnol.* **2010**, *28*, 46–54.
- (13) Reetz, M. T. *Angew. Chem. Int. Ed.* **2013**, *52*, 2658–2666.
- (14) Horsman, G. P.; Liu, A. M. F.; Henke, E.; Bornscheuer, U. T.; Kazlauskas, R. J. *Chem. Eur. J.* **2003**, *9*, 1933–1939.
- (15) Morley, K. L.; Kazlauskas, R. J. *Trends Biotechnol.* **2005**, *23*, 231–237.
- (16) Chica, R.; Doucet, N.; Pelletier, J. N. *Curr. Opin. Biotechnol.* **2005**, *16*, 378–384.
- (17) Reetz, M. T.; Wang, L.-W.; Bocola, M. *Angew. Chem. Int. Ed.* **2006**, *118*, 1258–1263.
- (18) Dill, K. A.; MacCallum, J. L. *Science* **2012**, *338*, 1042–1046.
- (19) Hanes, M. S.; Handel, T. M.; Chowdry, A. B. In *Protein Engineering and design*; CRC press: New York, 2010; pp. 313–325.
- (20) Maierov, V. N.; Crippen, G. M. *Proteins Struct. Funct. Genet.* **1994**, *20*, 167–173.
- (21) Pollack, S. J.; Jacobs, J. W.; Schultz, P. G. *Science* **1986**, *234*, 1570–1573.
- (22) Ringenberg, M. R.; Ward, T. R. *Chem. Commun.* **2011**, *47*, 8470–8476.
- (23) Malakauskas, S. M.; Mayo, S. L. *Nat. Struct. Biol.* **1998**, *5*, 470–475.
- (24) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R. E.; Grütter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D. *Nature* **2013**, doi:10.103.
- (25) Oelschlaeger, P.; Mayo, S. L.; Pleiss, J. *Protein Sci.* **2005**, *14*, 765–774.
- (26) Bolon, D. N.; Mayo, S. L. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14274–14279.
- (27) Richter, F.; Blomberg, R.; Khare, S. D.; Kiss, G.; Kuzin, A. P.; Smith, A. J. T.; Gallaher, J.; Pianowski, Z.; Helgeson, R. C.; Grjasnow, A.; Xiao, R.; Seetharaman, J.; Su, M.; Vorobiev, S.; Lew, S.; Forouhar, F.; Kornhaber, G. J.; Hunt, J. F.; Montelione, G. T.; Tong, L.; Houk, K. N.; Hilvert, D.; Baker, D. *J. Am. Chem. Soc.* **2012**, *134*, 16197–16206.
- (28) Khersonsky, O.; Röthlisberger, D.; Wollacott, A. M.; Murphy, P.; Dym, O.; Albeck, S.; Kiss, G.; Houk, K. N.; Baker, D.; Tawfik, D. S. *J. Mol. Biol.* **2011**, *407*, 391–412.
- (29) 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.; Barry, L. S.; Tawfik, D. S.; Baker, D. *Nat. Chem. Biol.* **2012**, *8*, 294–300.

- (30) Chevalier, B. S.; Kortemme, T.; Chadsey, M. S.; Baker, D.; Monnat, R. J.; Stoddard, B. L. *Mol. Cell* **2002**, *10*, 895–905.
- (31) Bury, C. R. *J. Am. Chem. Soc.* **1921**, *43*, 1602–1609.
- (32) Elschenbroich, C. *Organometallics*; 3rd editio.; Wiley Co.: Munich, 2006; p. 817.
- (33) Pace, N. R. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 805–808.
- (34) Zalups, R. K.; Koropatnick, J. *Molecular Biology and Toxicology of Metals*; Taylor & Francis Inc.: London, 2000; p. 596.
- (35) Butler, C. F.; Peet, C.; Mason, A. E.; Voice, M. W.; Leys, D.; Munro, A. W. *J. Biol. Chem.* **2013**, *288*, 25387–25399.
- (36) Ward, T. R. *Bio-inspired Catalysts*; Ward, T. R., Ed.; First Edit.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; Vol. 25, p. 115.
- (37) Koder, R. L.; Anderson, J. L. R.; Solomon, L. A.; Reddy, K. S.; Moser, C. C.; Dutton, P. L. *Nature* **2009**, *458*, 305–309.
- (38) Fleishman, S. J.; Whitehead, T. a; Ekiert, D. C.; Dreyfus, C.; Corn, J. E.; Strauch, E.-M.; Wilson, I. a; Baker, D. *Science* **2011**, *332*, 816–21.
- (39) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *109*, 3790–3795.
- (40) Wilson, M. E.; Whitesides, G. M. *J. Am. Chem. Soc.* **1978**, *100*, 306–307.
- (41) 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. *Angew. Chem. Int. Ed.* **2011**, *123*, 11055–11058.
- (42) Mayer, C.; Gillingham, D. G.; Ward, T. R.; Hilvert, D. *Chem. Commun.* **2011**, *47*, 12068–12070.
- (43) Creus, M.; Pordea, A.; Rossel, T.; Sardo, A.; Letondor, C.; Ivanova, A.; Letrong, I.; Stenkamp, R. E.; Ward, T. R. *Angew. Chem. Int. Ed.* **2008**, *47*, 1400–1404.
- (44) Hyster, T. K.; Knörr, L.; Ward, T. R.; Rovis, T. *Science* **2012**, *338*, 500–503.
- (45) Dürrenberger, M.; Heinisch, T.; Wilson, Y. M.; Rossel, T.; Nogueira, E.; Knörr, L.; Mutschler, A.; Kersten, K.; Zimbron, M. J.; Pierron, J.; Schirmer, T.; Ward, T. R. *Angew. Chem. Int. Ed.* **2011**, *50*, 3026–3029.
- (46) Bos, J.; García-Herraiz, A.; Roelfes, G. *Chem. Sci.* **2013**, *4*, 3578–3582.
- (47) Allard, M.; Dupont, C.; Muñoz Robles, V.; Doucet, N.; Lledós, A.; Maréchal, J.-D.; Urvoas, A.; Mahy, J.-P.; Ricoux, R. *Chembiochem* **2012**, *13*, 240–251.
- (48) Esmieu, C.; Cherrier, M. V; Amara, P.; Girgenti, E.; Marchi-Delapierre, C.; Odon, F.; Iannello, M.; Jorge-Robin, A.; Cavazza, C.; Ménage, S. *Angew. Chem. Int. Ed.* **2013**, *52*, 3922–3925.
- (49) J. Boersma, A.; Coquière, D.; Geerdink, D.; Rosati, F.; L. Feringa, B.; Roelfes, G. *Nat. Chem.* **2010**, *2*, 991–995.
- (50) Roelfes, G.; Feringa, B. L. *Angew. Chem. Int. Ed.* **2005**, *44*, 3230–3232.
- (51) J. Boersma, A.; L. Feringa, B.; Roelfes, G. *Angew. Chem. Int. Ed.* **2009**, *121*, 3396–3398.
- (52) Rosati, F.; Roelfes, G. *ChemCatChem* **2010**, *2*, 916–927.
- (53) Steinreiber, J.; Ward, T. R. *Coord. Chem. Rev.* **2008**, *252*, 751–766.
- (54) Yamamura, K.; Kaiser, E. *J. Chem. Soc., Chem. Commun.* **1976**, 830–831.

- (55) Mahammed, A.; Gross, Z. *J. Am. Chem. Soc.* **2005**, *127*, 2883–2887.
- (56) Ohashi, M.; Koshiyama, T.; Ueno, T. *Angew. Chem. Int. Ed.* **2003**, *115*, 1035–1038.
- (57) Ueno, T.; Koshiyama, T.; Ohashi, M.; Kondo, K.; Kono, M.; Suzuki, A.; Yamane, T.; Watanabe, Y. *J. Am. Chem. Soc.* **2005**, *127*, 6556–6562.
- (58) Ueno, T.; Koshiyama, T.; Abe, S.; Yokoi, N.; Ohashi, M.; Nakajima, H.; Watanabe, Y. *J. Organomet. Chem.* **2007**, *692*, 142–147.
- (59) Reetz, M. T.; Jiao, N. *Angew. Chem. Int. Ed.* **2006**, *45*, 2416–2419.
- (60) Carey, J. R.; Ma, S. K.; Pfister, T. D.; Garner, D. K.; Kim, H. K.; Abramite, J. a; Wang, Z.; Guo, Z.; Lu, Y. *J. Am. Chem. Soc.* **2004**, *126*, 10812–10813.
- (61) Haquette, P.; Salmain, M.; Svedlung, K.; Martel, A.; Rudolf, B.; Zakrzewski, J.; Cordier, S.; Roisnel, T.; Fosse, C.; Jaouen, G. *Chembiochem* **2007**, *8*, 224–231.
- (62) Kruithof, C. A.; Casado, M. A.; Guillena, G.; Egmond, M. R.; van der Kerk-van Hoof, A.; Heck, A. J. R.; Klein Gebbink, R. J. M.; van Koten, G. *Chem. Eur. J.* **2005**, *11*, 6869–6877.
- (63) Reetz, M. T.; Peyralans, J. J.-P.; Maichele, A.; Fu, Y.; Maywald, M. *Chem. Commun.* **2006**, 4318–4320.
- (64) Pordea, A.; Creus, M.; Panek, J.; Duboc, C.; Mathis, D.; Novic, M.; Ward, T. R. *J. Am. Chem. Soc.* **2008**, *130*, 8085–8.
- (65) Deuss, P. J.; den Heeten, R.; Laan, W.; Kamer, P. C. J. *Chem. Eur. J.* **2011**, *17*, 4680–4698.
- (66) Roelfes, G.; J. Boersma, A.; L. Feringa, B. *Chem. Commun.* **2006**, 635–637.
- (67) Coquière, D.; L. Feringa, B.; Roelfes, G. *Angew. Chem. Int. Ed.* **2007**, *46*, 9308–9311.
- (68) W. Dijk, E.; J. Boersma, A.; L. Feringa, B.; Roelfes, G. *Org. Biomol. Chem.* **2010**, *8*, 3868–3873.
- (69) Boersma, A. J.; Klijjn, J. E.; Feringa, B. L.; Roelfes, G. *J. Am. Chem. Soc.* **2008**, *130*, 11783–11790.
- (70) Oltra, N. S.; Roelfes, G. *Chem. Commun.* **2008**, 6039–6041.
- (71) Fournier, P.; Fiammengo, R.; Jäschke, A. *Angew. Chem. Int. Ed.* **2009**, *48*, 4426–4429.
- (72) Robles, V. M.; Ortega-Carrasco, E.; Fuentes, E. G.; Lledós, A.; Maréchal, J.-D. *Faraday Discuss.* **2011**, *148*, 137–159.
- (73) Pleiss, J. *Curr. Opin. Biotechnol.* **2011**, *22*, 611–617.
- (74) Liu, D.; Trodler, P.; Eiben, S.; Koschorreck, K.; Mueller, M.; Pleiss, J.; Maurer, S. C.; Branneby, C.; Schmid, R. D.; Hauer, B. *ChemBioChem* **2010**, *11*, 789–795.
- (75) Morra, G.; Colombo, G. *Proteins* **2008**, *72*, 660–672.
- (76) Gatti-Lafranconi, P.; Natalello, A.; Rehm, S.; Doglia, S. M.; Pleiss, J.; Lotti, M. *J. Mol. Biol.* **2010**, *395*, 155–166.
- (77) Henke, E.; Bornscheuer, U. T.; Schmid, R. D.; Pleiss, J. *Chembiochem* **2003**, *4*, 485–493.
- (78) Stjerschantz, E.; Vugt-Lussenburg, B. M. van; Bonifacio, A.; Beer, S. B. de; Zwan, G. van der; Gooijer, C.; Commandeur, J. N.; Vermeulen, N. P.; Oosterbring, C. *Proteins* **2008**, *71*, 336–352.
- (79) Guieysse, D.; Cortes, J.; Puech-Guenot, S.; Barbe, S.; Lafaquiere, V.; Monsan, P.; Simeon, T.; Andre, I.; Remaund-Simeon, M. *Chembiochem* **2008**, *9*, 1308–1317.

- (80) Pavlova, M.; Klvana, M.; Prokop, Z.; Chaloupkova, R.; Banas, P.; Otyepka, M.; Wade, R. C.; Tsuda, M.; Nagata, Y.; Damborsky, J. *Nat. Chem. Biol.* **2009**, *5*, 727–733.
- (81) Bocola, M.; Otte, N.; Jaeger, K. E.; Reetz, M. T.; Thiel, W. *Chembiochem* **2004**, *5*, 214–223.
- (82) Benkovic, S. J.; Hammes, G. G.; Hammes-Schiffer, S. *Biochemistry* **2008**, *47*, 3317–3321.
- (83) Kamp, M. W. van der; McGeagh, J. D.; Mulholland, A. J. *Angew. Chem. Int. Ed.* **2011**, *50*, 10349–10351.
- (84) Polyak, I.; Reetz, M. T.; Thiel, W. *J. Am. Chem. Soc.* **2012**, *134*, 2732–2741.
- (85) Lind, M. E. S.; Himo, F. *Angew. Chem. Int. Ed.* **2013**, *125*, 4661–4665.
- (86) Frushicheva, M. P.; Warshel, A. *ChemBioChem* **2012**, *13*, 215–223.
- (87) Althoff, E. a; Wang, L.; Jiang, L.; Giger, L.; Lassila, J. K.; Wang, Z.; Smith, M.; Hari, S.; Kast, P.; Herschlag, D.; Hilvert, D.; Baker, D. *Protein Sci.* **2012**, *21*, 717–726.
- (88) Muñoz Robles, V.; Vidossich, P.; Lledós, A.; Ward, T. R.; Maréchal, J.-D. *ACS Catal.* **2014**, d.o.i: 10.1021/cs400921n.
- (89) Hohenber, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, 864–871.
- (90) Kohn, W. *Lu Jeu, Sham* **1965**, *140*, A1133–A1138.
- (91) Höltje, H.-D.; Sippl, W.; Rognan, D.; Folkers, G. *Molecular Modeling: basic principles and applications*; 3rd editio.; Wiley Co.: Weinheim, 2008.
- (92) Borrelli W., K.; Vitalis, A.; Alcantara, R.; Guallar, V. *J. Chem. Theory Comput.* **2005**, *1*, 1304–1311.
- (93) Frauenfelder, H.; Parak, F.; Young, R. D. *Ann. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 451–479.
- (94) Elber, R.; Karplus, M. *Science* **1987**, *235*, 318–321.
- (95) Kitao, A.; Hayward, S.; Go, N. *Proteins* **1998**, *33*, 496–517.
- (96) Ma, J. *Curr. Protein Pept. Sci.* **2004**, *5*, 119–123.
- (97) Ma, J. *Structure* **2005**, *13*, 373–380.
- (98) Trakhanov, S.; K. Vyas, N.; Luecke, H.; M. Kristensen, D.; Ma, J.; A. Quioco, F. *Biochemistry* **2005**, *44*, 6597–6608.
- (99) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
- (100) Gonzalez-Lopez de Turiso, F.; Sun, D.; Rew, Y.; Bartberger, M. D.; Beck, H. P.; Canon, J.; Chen, A.; Chow, D.; Correll, T. L.; Huang, X.; Julian, L. D.; Kayser, F.; C., L. M.; Long, A. M.; McMinn, D.; Oliner, J. D.; Osgood, T.; Powers, J. P.; Saiki, A. Y.; Schneider, S.; Shaffer, P.; Xiao, S. H.; Yakowec, P.; Yan, X.; Ye, Q.; Yu, D.; Zhao, X.; Zhou, J.; Medina, J. C.; Olson, S. H. *J. Med. Chem.* **2013**, *56*, 4053–4070.
- (101) Yuriev, E.; Ramsland, P. A. *J. Mol. Recognit.* **2013**, *26*, 215–239.
- (102) Byun, K. S.; Morokuma, K. *J. Mol. Struct. THEOCHEM* **1999**, *461-462*, 1–21.
- (103) Molecular Operating Environment (MOE) **2013**.
- (104) Schrödinger **2013**.
- (105) Discovery Studio Modeling Environment **2013**.
- (106) Hinsen, K. *J. Comput. Chem.* **2000**, *21*, 79–85.

- (107) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. *J. Comput. Chem.* **2004**, *25*, 1605–1612.
- (108) Ueno, T.; Yokoi, N.; Unno, M.; Matsui, T.; Tokita, Y.; Yamada, M.; Ikeda-Saito, M.; Nakajima, H.; Watanabe, Y. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9416–9421.
- (109) De Lauzon, S.; Desfosses, B.; Mansuy, D.; Mahy, J.-P. *FEBS Lett.* **1999**, *443*, 229–243.
- (110) De Lauzon, S.; Mansuy, D.; Mahy, J.-P. *Fed. Eur. Biochem. Soc. J.* **2002**, *269*, 470–480.
- (111) Mahy, J.-P.; Desfosses, B.; de Lauzon, S.; Quilez, R. *Appl. Biochem. Biotech.* **1998**, *75*, 103–112.
- (112) Ricoux, R.; Sauriat-Dorizon, H.; Girgenti, R.; Blanchard, D.; Mahy, J.-P. *J. Immunol. Methods* **2002**, *269*, 39–57.
- (113) Lu, Y.; Yeung, N.; Sieracki, N.; Marshall, N. M. *Nature* **2009**, *460*, 855–862.
- (114) Nanda, V.; Koder, R. L. *Nat. Chem.* **2010**, *2*, 15–24.
- (115) Maglio, O.; Natri, F.; Rosales, R. T. M. de; Faiella, M.; Pavone, V.; DeGrado, W. F.; Lombardi, A. *Comptes Rendus Chim.* **2007**, *10*, 703–720.
- (116) Summa, C. M.; Rosenblatt, M. M.; Hong, J.-K.; Lear, J. D.; DeGrado, W. F. *J. Mol. Biol.* **2002**, *321*, 923–938.
- (117) Hirotsu, S.; Chu, G. C.; Unni, M.; Lee, D. S.; Yoshida, T.; Park, S. Y.; Shiro, Y.; Ikeda-Saito, M. *J. Biol. Chem.* **2004**, *279*, 11937–11947.
- (118) Maines, M. D. *FASEB* **1988**, *2*, 2557–2568.
- (119) Yokoi, N.; Ueno, T.; Unno, M.; Matsui, T.; Ikeda-Saito, M.; Watanabe, Y. *Chem. Commun.* **2008**, 229–231.
- (120) Chase, F.; Avenue, B. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440.
- (121) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (122) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B Condens. Matter* **1988**, *37*, 785–789.
- (123) 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, Revision A.1* **2009**.
- (124) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.
- (125) Hehre, W. J. *J. Chem. Phys.* **1972**, *56*, 2257–2261.
- (126) Hariharan, P. C.; Pople, J. A. *Theor. Chim. Acta* **1973**, *28*, 213–222.
- (127) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609–623.

- (128) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V; Mee, R. P. *J. Comput. Aided. Mol. Des.* **1997**, *11*, 425–445.
- (129) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. D. *Proteins Struct. Funct. Bioinforma.* **2005**, *58*, 836–844.
- (130) Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M. *Proteins Struct. Funct. Bioinforma.* **2008**, *70*, 208–217.
- (131) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. THEOCHEM* **1999**, *461-462*, 1–21.
- (132) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. a. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (133) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.
- (134) Allen, F. H. *Acta Crystallographica Sect. B Struct. Sci.* **2002**, *58*, 380–388.
- (135) Lad, L.; Schuller, D. J.; Shimizu, H.; Friedman, J.; Li, H.; Ortiz de Montellanos, P. R.; Poulos, T. L. *J. Biol. Chem.* **2003**, *278*, 7834–7843.
- (136) Rahman, M. N.; Vlahakis, J. Z.; Szarek, W. A.; Nakatsu, K.; Jia, Z. *J. Med. Chem.* **2008**, *51*, 5943–5952.
- (137) Quilez, R.; De Lauzon, S.; Desfosses, B.; Mansuy, D.; Mahy, J. P. *FEBS Lett.* **1996**, *395*, 73–76.
- (138) De Lauzon, S.; Quilez, R.; Lion, L.; Desfosses, B.; Lee, I.; Sari, M. A.; Benkovic, S. J.; Mansuy, D.; Mahy, J. P. *Fed. Eur. Biochem. Soc. J.* **1998**, *257*, 121–130.
- (139) Ricoux, R.; Dubuc, R.; Dupont, C.; Marechal, J.-D.; Martin, A.; Sellier, M.; Mahy, J.-P. *Bioconjug. Chem.* **2008**, *19*, 899–910.
- (140) Ricoux, R.; Allard, M.; Dubuc, R.; Dupont, C.; Maréchal, J.-D.; Mahy, J. P. *Org. Biomol. Chem.* **2009**, *7*, 3208–3211.
- (141) Mahy, J.-P.; Raffy, Q.; Allard, M.; Ricoux, R. *Biochimie* **2009**, *91*, 1321–1323.
- (142) Ricoux, R.; Raffy, Q.; Mahy, J.-P. *Comptes Rendus Chim.* **2007**, *10*, 684–702.
- (143) Hart-Davis, J.; Battioni, J. P.; Boucher, J. L.; Mansuy, D. *J. Am. Chem. Soc.* **1998**, *120*, 12524–12530.
- (144) Muñoz Robles, V.; Maréchal, J.-D.; Bahloul, A.; Sari, M.-A.; Mahy, J.-P.; Golinelli-Pimpaneau, B. *PLoS One* **2012**, *7*, e51128.
- (145) Loosli, A.; Rusbandi, U. E.; Gradinaru, J.; Bernauer, K.; Schlaepfer, C. W.; Meyer, M.; Mazurek, S.; Novic, M.; Ward, T. R. *Inorg. Chem.* **2006**, *45*, 660–668.
- (146) Green, N. M. *Adv. Protein Chem.* **1975**, *29*, 85–133.
- (147) Perdew, J.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- (148) Perdew, J. P.; Ernzerhof, M.; Burke, K. *J. Chem. Phys.* **1996**, *105*, 9982–9985.
- (149) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (150) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.
- (151) Andzelm, J.; Kölmel, C.; Klamt, A. *J. Chem. Phys.* **1995**, *103*, 9312–9320.
- (152) Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

- (153) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V.; Chimica, D.; Li, F.; Angelo, C. M. S. *J. Comput. Chem.* **2003**, *24*, 669–681.
- (154) Šot, P.; Kuzma, M.; Václavík, J.; Pecháček, J.; Přeč, J.; Januščák, J.; Kačer, P. *Organometallics* **2012**, *31*, 6496–6499.
- (155) Baxter, C. a; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. *Proteins* **1998**, *33*, 367–382.
- (156) Muñoz Robles, V.; Dürrenberger, M.; Lledós, A.; Ward, T. R.; Maréchal, J.-D. *Chem. Sci.* **2013**, -, Submitted.
- (157) Van Hellemond, E. W.; Janssen, D. B.; Fraaije, M. W. *Appl. Environ. Microbiol.* **2007**, *73*, 5832–5839.
- (158) Tuynman, A.; Spelberg, J. L.; Kooter, I. M.; Schoemaker, H. E.; Wever, R. *J. Biol. Chem.* **2000**, *117*, 6412–6413.
- (159) Reetz, M. T.; Rentzsch, M.; Pletsch, A.; Maywald, M. *Chimia (Aarau)*. **2002**, *56*, 721–723.
- (160) Okrasa, K.; Kazlauskas, R. J. *Chem. Eur. J.* **2006**, *12*, 1587–1596.
- (161) Fernández-Gacio, A.; Codina, A.; Fastrez, J.; Riant, O.; Soumillion, P. *ChemBioChem* **2006**, *7*, 1013–1016.
- (162) Meunier, B. *Chem. Rev.* **1992**, *92*, 1411–1456.
- (163) Battioni, P.; Renaud, J. P.; Bartoli, J. F.; Reina-Artiles, M.; Fort, M.; Mansuy, D. *J. Am. Chem. Soc.* **1988**, *110*, 8462–8470.
- (164) Ostovic, D.; Bruce, T. C. *Acc. Chem. Res.* **1992**, *25*, 314–320.
- (165) Gloster, T. M.; Williams, S. J.; Roberts, S.; Tarlinga, C.; Wicki, J.; Withers, S. G.; Davies, G. J. *Chem. Commun.* **2004**, 1794–1795.
- (166) Gloster, T. M.; Williams, S. J.; Roberts, S.; Tarlinga, C.; Wicki, J.; Withers, S. G.; Davies, G. J. *Chem. Commun.* **2003**, 944–945.
- (167) Charnock, S. J.; Derewenda, U.; Derewenda, Z. S.; Dauter, Z.; Dupont, C.; Morosoli, R.; Kluepfel, D.; Davies, G. J. *J. Biol. Chem.* **2000**, *275*, 23020–23026.
- (168) Groves, J. T.; Lee, J.; Marla, S. S. *J. Am. Chem. Soc.* **1997**, *119*, 6269–6273.
- (169) Maréchal, J.-D. Estudio del mecanismo cooperativo de la hemoglobina por métodos de mecánica cuántica y mecánica cuántica/mecánica molecular, Universitat Autònoma de Barcelona (Tesis Doctoral), 2003, p. 188.
- (170) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. D. *Proteins Struct. Funct. Bioinforma.* **2005**, *58*, 836–844.
- (171) Yu, J.; Paine, J. I.; Maréchal, J.-D.; Kemp, C. A.; Ward, C. J.; Brown, S.; Sutcliffe, M. J.; Roberts, G. C. K.; Rankin, E. M.; Wolf, C. R. *Drug. Met. Disp.* **1996**, *34*, 1386–1392.
- (172) Balcells, D.; Maseras, F. *New J. Chem.* **2007**, *31*, 333–343.
- (173) Yamakawa, M.; Yamada, I.; Noyori, R. *Angew. Chem. Int. Ed.* **2001**, *40*, 2818–2821.
- (174) Martins, J. E. D.; Clarkson, G. J.; Wills, M. *Org. Lett.* **2009**, *11*, 847–850.
- (175) Soni, R.; Cheung, F. K.; Clarkson, G. C.; Martins, J. E. D.; Graham, M. a; Wills, M. *Org. Biomol. Chem.* **2011**, *9*, 3290–3294.
- (176) Jiří Václavík, Marek Kuzma, J. P. and P. K. *Organometallics* **2011**, *30*, 4822–4829.

- (177) Åberg, J. B.; Samec, J. S. M.; Bäckvall, J.-E. *Chem. Commun.* **2006**, 70, 2771–2773.
- (178) Fabrello, A.; Bachelier, A.; Urrutigoity, M.; Kalck, P. *Coord. Chem. Rev.* **2010**, 254, 273–287.
- (179) Hopmann, K. H.; Bayer, A. *Organometallics* **2011**, 30, 2483–2497.
- (180) Pablo, Ó.; Guijarro, D.; Kovács, G.; Lledós, A.; Ujaque, G.; Yus, M. *Chem. Eur. J.* **2012**, 18, 1969–1983.
- (181) Zhu, Y.; Fan, Y.; Burgess, K. *J. Am. Chem. Soc.* **2010**, 132, 6249–6253.
- (182) Bakowies, D.; Thiel, W. *J. Phys. Chem.* **1996**, 3654, 10580–10594.
- (183) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *Amber 12* **2012**.
- (184) Köhler, V.; Wilson, Y. M.; Dürrenberger, M.; Ghislieri, D.; Churakova, E.; Quinto, T.; Knörr, L.; Häussinger, D.; Hollmann, F.; Turner, N. J.; Ward, T. R. *Nat. Chem.* **2013**, 5, 93–99.
- (185) Polyak, I.; Reetz, M. T.; Thiel, W. *J. Phys. Chem. B* **2013**, 117, 4993–5001.
- (186) Hinsen, K. MMTK User's Guide <http://dirac.cnrs-orleans.fr/MMTK/using-mmtk/MMTK-Manual.pdf>.
- (187) Floquet, N.; Maréchal, J.-D.; Badet-Denisot, M.-A.; Robert, C. H.; Dauchez, M.; Perahia, D. *FEBS Lett.* **2006**, 580, 5130–5136.
- (188) Maréchal, J.-D.; Perahia, D. *Eur. Biophys. J.* **2008**, 37, 1157–1165.
- (189) Hinsen, K. *Proteins Struct. Funct. Genet.* **1998**, 33, 417–429.
- (190) Hinsen, K. *Bioinformatics* **2008**, 24, 521–528.
- (191) Yang, L.; Song, G.; L. Jernigan, R. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, 106, 12347–12352.
- (192) Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, 24, 760.
- (193) Vreven, T.; Morokuma, K. *Annu. Rep. Comput. Chem.* **2006**, 2, 35–51.
- (194) Szabó, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Publications: Mineola, 1996; p. 466.



Annex A: Computational Methods

I - Quantum Mechanics

Until the 20th century, physics described natural phenomena using classical mechanics, which main ideas were centered on the Newton laws of dynamic¹ and Galilei transformation for relativity.² As science and technology advanced, classical physics failed to describe new experimental phenomena, such as the wave-particle duality of light or the black body radiation.³ To deal with them, two alternative mechanics were developed during the 20th century: *Quantum Mechanics* and *Relativistic Mechanics*. The former is a series of postulates that describes what happens in a short-scale systems, typically 100 nm or less, while the later is used to describe large-scale high-massed systems travelling at velocities similar to the speed of light. Atoms and molecules are clearly inside the quantum realm, so quantum mechanics is the reference in physics to describe how they behave. Some relativistic effect should also be taken into account to describe the electrons due to its high velocity (similar to the speed of light). It is to notice that quantum and classical mechanics are still entwined. Objects belonging to the macroscopic world can also be described using quantum mechanics if they are treated as a large collection of particles. In fact, classical mechanics could be considered as an approximation of quantum mechanics for large objects. This correlation between the two mechanics is known as the *correspondence principle*.

I.I - Schrödinger equation resolution: An approximate approach

A wave function is a mathematical function that describes the quantum state of a system and how it behaves. Usually having complex numbers, it is both time and space dependent and it correlates to a single particle. The Schrödinger equation describes how this function evolves over time (Eq A.1):

$$i\hbar \frac{\partial}{\partial t} \Psi(x, t) = H\Psi(x, t) \quad \text{Eq A.1}$$

In this last equation the H represents the *Hamiltonian*, an operator that describes how the quantum state of the system evolves over time and provides with its total energy. In

¹ Three Newton laws describing the inertia, relationship between the acceleration of a body and forces and the action-reaction force pairs.

² Galilei transformation for velocities of different inertial systems is $v_1 = v_2 + u$, where the different v_i are the velocities of a particle in different reference systems and u is the velocity of movement of the system 2 taking as a movement reference the system 1.

³ Ron, J. M. S. *Espacio-Tiempo y átomos. Relatividad y mecánica cuántica*; Edicion AKAL: Los Berrocales del Jarama, 1992.

correlation with classical mechanics, the Hamiltonian of quantum mechanics can also be considered as a sum of the total potential (V) and kinetic (T) energy of the system (Eq A.2).

$$H = T + V \quad \text{Eq A.2}$$

Eq A.1 is often regarded as the time-dependent Schrödinger equation as time is one of the variables. From a computational chemist point of view, a given reaction can be studied as the sum of the stationary states corresponding to the minima (reactants, intermediates, product) and maxima (transition states) points of the Potential Energy Surface (see section IV of Chapter 2). For this reason, the resolution of a time-independent version of the Schrödinger equation (Eq A.3) is more interesting as it is a simple way to obtain the wave function of the stationary states.⁴

$$H\Psi_x = E\Psi_x \quad \text{Eq A.3}$$

The resolution of Eq A.3 for a chemical system leads to a set of very complex reactions that are only feasible for very small systems. For large systems, which tend to be the interesting ones, several approximations are needed to obtain simple forms of the time-independent Schrödinger equation. Although this can compromise the final result, it is the only way for computational chemical studies to study chemical systems. Knowing how and when to apply the different existing approximations is one of the crucial points in such analysis.

I.I.I - Born-Oppenheimer approximation

One of the most successfully applied approximations in Theoretical Chemistry is the Born-Oppenheimer approximation.⁵ In general terms, this approximation allows the separation of the wave function in two different terms: (i) a term that describes the electron wave function for a fixed nuclei position and (ii) another term that describes the nuclei wave function, where the first term (i) plays the role as the potential energy. This approximation can be securely used because of the high mass ratio between the electrons and the nuclei.⁶ For this reason, the electrons are much faster than the nuclei and we can assume that the electron population adapts itself instantaneously to the nuclear distribution. Therefore, for a given chemical system the electron density will only depend on the

⁴ Levine, I. N. *Quantum Chemistry*; Pearson Prentice Hall: New Jersey, 2009.

⁵ Szabó, A.; Ostlun, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover Publications: Mineola, 1996.

⁶ For the lightest nuclei, a proton (H^+), this ratio is already $\frac{m_{H^+}}{m_{e^-}} = 1836$.

nuclear position and not on their movement. Applying this approximation on Eq A.2, a much simple form of the time-independent Schrödinger equation can be obtained (Eq A.4):

$$H = T + V = \left[\sum_i \frac{P_i^2}{2} + V_{NN} + V_{Ne} + V_{ee} \right] \quad \text{Eq A.4}$$

In the last equation, P_i is the momentum operator for the electron i and V_{xx} represent the potential energy due to the Coulomb interaction for every particle pair (NN for nuclei-nuclei repulsion, Ne for nuclei-electron attraction and ee for electron-electron repulsion).

I.1.II - Hartree-Fock Self Consistent Field (HF-SCF)

The Hartree-Fock approximation is one of the cornerstones of all the theoretical chemistry based on resolving the wave function of a given system. Instead of trying to resolve the many-electrons wave function, it considers that the N -electron wave function can be considered as the antisymmetrized⁷ product of N one-electron wave functions $\chi_i(\vec{x}_i)$, where \vec{x}_i are the coordinates of the electron i in polar notation. This product is often related to as the *Slater Determinant*, Φ_{SD} (Eq A.5):

$$\Psi_0 \approx \Phi_{SD} = \frac{1}{\sqrt{N!}} \det\{\chi_1(\vec{x}_1) \chi_2(\vec{x}_2) \dots \chi_N(\vec{x}_N)\} \quad \text{Eq A.5}$$

Replacing the N -electron wave function by a single Slater determinant will lead to a high-quantitative inaccuracy as the electron-electron interactions are neglected. To optimize this approximation the *Variational Principle* is used to find the best Slater determinant (the one that yields the lowest energy, Eq A.6). The only variable in the Slater determinant are the spin orbitals (χ_i), which in the Hartree-Fock approach will vary under the premise that they must remain orthonormal so the energy obtained from the Slater determinant is minimal.

$$E_{HF} = \min_{\Phi_{SD} \rightarrow N} E[\Phi_{SD}] \quad \text{Eq A.6}$$

⁷ The wave function describing the electrons behavior must be keep antisymmetric to meet with the Pauli exclusion principle.

Optimizing the set of \mathcal{X}_i functions present on the Φ_{SD} will lead to the minimum energy result using the variational approach (Eq A.6). The resulting equations for that minimization are the Hartree-Fock equations (Eq A.7).

$$\hat{f}\mathcal{X}_i = \varepsilon_i\mathcal{X}_i, \quad i = 1, 2, \dots, N \quad \text{Eq A.7}$$

The *Lagrangian multiplier* (ε_i) is needed to fulfill the orthonormality of the spin orbitals through the minimization. The N equations have the appearance of eigenvalue equations, where the ε_i are the eigenvalues of the Fock operator (\hat{f}). These ε_i are the physical interpretation of orbital energies. The \hat{f} is an effective one-electron operator defined as (Eq A.8):

$$\hat{f} = -\frac{1}{2}\nabla_i^2 - \sum_A^M \frac{Z_A}{r_{iA}} + V_{HF}(i) \quad \text{Eq A.8}$$

The first two terms of this equation account for the kinetic and the potential energy resulting from the nucleus attraction. $V_{HF}(i)$ is the Hartree-Fock potential and represents the average repulsive potential experienced by the i 'th electron due to the remaining $N-1$ electrons. This way, the complicated two-electron repulsion operator in the Hamiltonian is replaced by the simple one-electron operator $V_{HF}(i)$, where the electron-electron repulsion is taken into account only in an average way.

The Fock operator is dependent of the spin orbitals through the Hartree Fock potential (V_{HF} in Eq A.8), which is the very solution of the eigenvalues that need to be solved. For this reason, Eq A.7 is a pseudo-eigenvalue problem that can be solved iteratively. Starting with a "guessed" set of orbitals, the Hartree Fock equations are solved. Then, the resulting new set of orbitals is used for another iteration, solving again the Hartree Fock equation. This cycle goes on until the output and the input differ by less than a predefined threshold. This method is called *Self-Consistent Field* (SCF).

In 1951, Roothaan proposed a new method to rapidly optimize the \mathcal{X}_i on the SCF. This approach considered the \mathcal{X}_i functions as a linear combination of a set of known and invariable functions. This group of functions is known as *Basis Function* (Eq A.9).

$$\mathcal{X}_i = \sum_j c_{ij}f_j \quad \text{Eq A.9}$$

Finding the exact solution using this expansion is not practical because an infinite number of terms is needed. However, we can restrict the size of the basis set to find an approximate resolution. One should take into account that the bigger this set of functions is, the more accurate the result will be but the computing time will dramatically increase. Finding a suitable basis set with a good compromise between accuracy and calculation time is one of the cornerstone of any quantum mechanical simulation.

II - Force Field

II.I - Bonding terms

The *bond-stretching* term (Figure 2.2) is defined as follows (Eq A.10):

$$U_{AB} = \frac{1}{2} k_{AB} (R_{AB} - R_{e,AB})^2 \quad \text{Eq A.10}$$

and describes the potential energy between two different bonded atoms A-B using an harmonic oscillator. k_{AB} is the force constant, R_{AB} is the instantaneous bond length and $R_{e,AB}$ is the length bond at the equilibrium. The sum of all the bonds presents on the molecule will provide the bond-stretching term of the ff.

The *bond-bending vibrations* term (Figure 2.2) also uses an harmonic oscillator to calculate the potential energy for the connected atoms A-B-C (Eq A.11).

$$U_{ABC} = \frac{1}{2} k_{ABC} (\theta_{ABC} - \theta_{e,ABC})^2 \quad \text{Eq A.11}$$

Similarly to Eq A.10, k_{ABC} is the force constant and the subscript “e” refers to the equilibrium position of the angle A-B-C.

Two different dihedral angles can be formed between four atoms ABCD (Figure 2.2). If full rotation is allowed in all bonds, then we have a *proper dihedral*. If the rotation of some bonds is restricted, then we are dealing with an *improper dihedral*. The last is used to maintain the chirality and the planarity of some molecules, like in a benzene molecule. A typical way to calculate the potential energy of a dihedral is given by (Eq A.12):

$$U = \frac{U_0}{2} (1 - \cos(n(\chi - \chi_e))) \quad \text{Eq A.12}$$

Here, χ accounts for the ABCD angle and n is the periodicity term that will define if the dihedral is proper or improper.

Finally, we have the out-of-plane angle potential or inversion term. Imagine Figure A.1 to be ammonia. The inversion of the molecule will depend on the angle ψ , which can be negative or positive depending on the side of the plane atom D is occupying. The potential energy for this inversion can be calculated either in terms of height (h) or angle (ψ) and considering or not the periodicity of the inversion (Eq A.13).

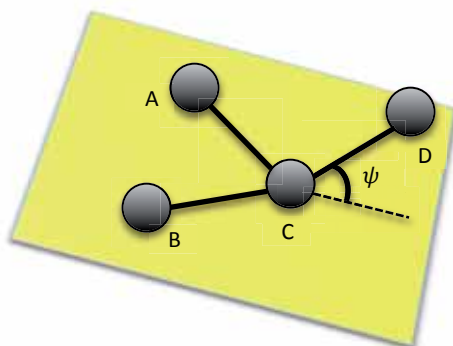


Figure A.1 – Schematic representation of an out-of-plane inversion.

$$U = \frac{k_1}{2\sin^2 \psi_e} (\cos \psi - \cos \psi_e)^2$$

$$U = k_2 h^2$$

$$U = k_3 (1 + k_4 \cos(n\psi))$$

$$U = k_5 (1 + \cos(n\psi - k_6))$$

Eq A.13

In the previous equations, the n terms accounts for the periodicity, the ψ_e for the angle at equilibrium and the different constants k have to be fixed against experimental values.

II.II - Non-Bonding terms

As discussed earlier, the non-bonded terms are the sum for the van der Waals and the electrostatic interactions. The former is calculated using a *Lennard-Jones potential* and the later using the *Coulomb electrostatic law*.

The Lennard-Jones potential is calculated between all non-bonded pair of atoms using the following equation (Eq A.14):

$$U_{L-J} = \frac{C_{12}}{R^{12}} - \frac{C_6}{R^6} \quad \text{Eq A.14}$$

Where the C term is the distance at which the minimum of energy is found between the two atoms and R is the distance between them.

The electrostatic term is calculated using the Coulomb electrostatic law. This term is calculated for all non-bonded pair of atoms in the system and the charges are assigned to each atom according to the rules of each particular ff. Eq A.15 shows the typical Coulomb electrostatic potential between two atoms A and B, where Q_A and Q_B are their respective charges.

$$U_{AB} = \frac{1}{4\pi\epsilon_0} \frac{Q_A Q_B}{R_{AB}} \quad \text{Eq A.15}$$

Putting together all the previously discussed terms will define the general equation in which most modern ffs are based nowadays (Eq A.16):

$$U = \sum_{stretch} \frac{1}{2} k_{AB} (R_{AB} - R_{e,AB})^2 + \sum_{bend} \frac{1}{2} k_{ABC} (\theta_{ABC} - \theta_{e,ABC})^2 + \sum_{dihedrals} \frac{U_0}{2} (1 - \cos(n(\chi - \chi_e))) \\ + \sum_{inversion} \frac{k}{2 \sin^2 \psi_e} (\cos \psi - \cos \psi_e)^2 + \sum_{non-bonded} \left(\frac{C_{12,AB}}{R_{AB}^{12}} - \frac{C_{6,AB}}{R_{AB}^6} \right) + \frac{1}{4\pi\epsilon_0} \sum_{charges} \frac{Q_A Q_B}{R_{AB}}$$

Eq A.16

III - Molecular Mechanics

III.I - Docking Limitations

Most scoring functions are essentially focused in accurately represent the enthalpic effects (e.g. the energy) of the binding, neglecting the entropic effects. Both enthalpic and entropic effects could be the driven force of the binding, so neglecting any of those terms could lead to incorrect results.⁸

Many efforts are being focused in the treatment of the receptor flexibility. Inherent flexibility of the receptor, induced fit and other conformational changes that could occur during the binding are usually not considered during the simulation. Many attempts to solve those problems have been reported, including minimization, side-chain rotameric libraries, and minimized Monte Carlo via normal-mode analysis. Those methodologies

⁸ Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discov.* **2004**, 3, 935–949.

showed that best results are obtained with a proper opening of the binding site via Monte Carlo methods and a side-chain optimization afterwards around the flexible ligand. Despite those improvements, several studies still highlight the need of methodologies for a better treatment of receptor flexibility, special with full receptor flexibility (e.g. backbone atoms) and collective movements.⁹

Water molecules can have an important role in the binding of ligands by creating a hydrogen bond network. Placing waters on the interface of biomolecular complexes has proven to increase the binding affinities of ligands, however this could lead to an increase of false positive bindings. Properly accounting for the specific waters involved in the binding, as well as the solvation effect, is still a challenging issue in docking simulations. This problem is particularly difficult when dealing with polar or charged systems, as waters molecules have a crucial effect on the binding event.⁹

III.II - MD algorithms

A typical MD algorithm should be able to compute all the following features needed during the MD simulation to describe the evolution of the system: (i) the initial velocities, (ii) the forces acting on each atom and (iii) their acceleration.

When starting a classical MD simulation the initial structure contains the coordinates of every atom but it does not have any information regarding their velocities, which cannot be determined experimentally. First thing the MD algorithm needs to do is to provide for the initial velocities of each atom. They are randomly assigned to each atom using the Maxwell distribution of velocities (Eq A.17) at a given temperature (T):

$$P(v)dv = \left(\frac{m}{2\pi k_B T}\right)^{1/2} e^{\left[\frac{-mv^2}{2k_B T}\right]} dv \quad \text{Eq A.17}$$

With every atom having its initial velocities associated, the forces acting on each one of them can be calculated. This is the most computational intensive part of the whole MD simulation, as we need to know the potential applied to each molecule. The most usual way to compute the forces is to use a force field to account for all the potentials acting on each atom and selecting an appropriate cut-off to simplify the long-range interactions. Then, the forces acting on each atom of the system can be calculated as the derivative of the potential energy respect the coordinates of the atom (Eq A.18):

⁹ Yuriev, E.; Ramsland, P. A. *J. Mol. Recognit.* **2013**, *26*, 215–239.

$$\vec{F}_i = -\frac{dU}{d\vec{r}_i} \quad \text{Eq A.18}$$

Once the force is calculated, the acceleration can be calculated applying Newton second law (Eq A.19):

$$\vec{a}_i = \frac{\vec{F}_i}{m_i} \quad \text{Eq A.19}$$

Now the acceleration, the forces and the velocities of each atom are calculated at a given time t , so the next step is to determine the displacement of each atom at time $t+\Delta t$. To do so, *Taylor expansion* is used to solve the Newton equation of movement (Eq A.20):

$$r_{(t+\Delta t)} = r_t + \frac{dr}{dt} \Delta t + \frac{1}{2} \frac{d^2r}{d^2t} \Delta t^2 + \dots \quad \text{Eq A.20}$$

On this expansion both the velocity ($v_i = \frac{dr_i}{dt}$) and the acceleration ($a_i = \frac{d^2r_i}{d^2t}$) terms are found. Eq A.20 can be rewritten in a discrete form truncating the Taylor expansion at the second derivative, where r_n indicates the position at step n and time t and r_{n+1} indicates the position at step $n+1$ and time $t+\Delta t$ (Eq A.21):

$$r_{n+1} = r_n + v_n \Delta t + \frac{1}{2} a_n \Delta t^2 \quad \text{Eq A.21}$$

Doing the same procedure to obtain the positions at the previous step (r_{n-1}) using the Taylor expansion (Eq A.20) and adding it to Eq A.21 the so-called Verlet algorithm¹⁰ is obtained (Eq A.22):

$$r_{n+1} = 2r_n - r_{n-1} + a_n \Delta t^2 \quad \text{Eq A.22}$$

Although many MD simulations procedures are based on this algorithm, it is not exempted of problems. For example, the velocities are not directly represented on the equation. To solve that and other numerical issues, other algorithms can also be used (Leap-Frog¹¹ and

¹⁰ Verlet, L. *Phys. Rev.* **1967**, *158*, 98–103.

¹¹ Hockney, R. W.; Eastwood, J. W. *Computer Simulation Using Particles*; Mac Graw Hill: New York, 1981.

Velocity Verlet¹²) during a MD simulation. For more information about the advantages and disadvantages of every method please refer to a general MD manual.^{13,14}

III.III - How to run a MD simulation

A typical MD simulation is composed of the following phases: (i) creation of the system, (ii) initialization, (iii) equilibration and (iv) production.

The starting point to build the system is to get the coordinates of the atoms. The usual way to do that is to get a crystallographic structure extracted from the *Protein Data Bank* (PDB). Obtaining these structures is not something trivial, and in many cases some parts of the protein (specially the most flexible ones) could be missing.¹⁵ After checking that the structure is okay, hydrogen atoms can be added using any of the available algorithms existing for that purpose (HBuild in the CHARMM suite,¹⁶ etc...). The system also will need to be solvated and charges neutralized (AmberTools,¹⁷ etc...). At this point it is recommended to minimize the system to avoid clashes that could have appeared in the previous processes.

Now that the system is prepared we can start the real MD with the initialization phase. First step is to assign the initial velocities to each atom, something needed to calculate their respective forces and accelerations. If the MD is a re-start of a previous one, these velocities can be extracted from the preceding MD data. However, if it is a new MD simulation, we have to randomly assign these velocities. To avoid overestimation they are assigned at a low temperature, otherwise some collisions may appear and lead to a destabilization of the system. Through a series of small MD simulations, the temperature

¹² Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637.

¹³ Frenkel, D.; Smit, B. *Understanding Molecular Simulation: from Algorithms to Applications*; Academic Press: New York, 2002.

¹⁴ Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press: New York, 1989.

¹⁵ Explaining how to deal with those problems is out of the scope of that thesis. For more information please refer to a general MD manual (Haile, J. M. *Molecular Dynamics Simulation: Elementary Methods*; Wiley Co.: Clemson, 1997.)

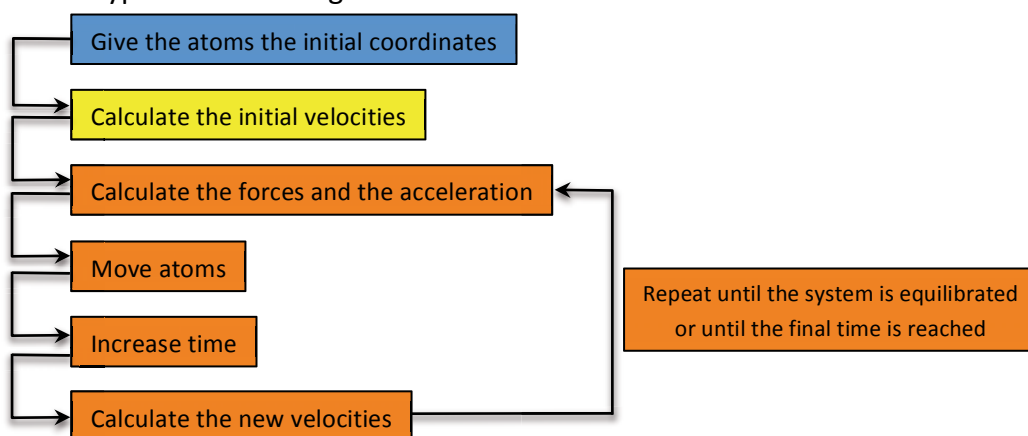
¹⁶ Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

¹⁷ Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *Amber 12* **2012**.

is gradually raised while the velocities are scaled to match it in a process called *heating*. Once the desired temperature is reached, the equilibration of the system begins.

The equilibration phase is a large MD simulation in which there is an exchange between the potential and the kinetics energy, thus maintaining the total energy of the system. This phase should last until both of them start oscillating between their medium values, something needed to obtain stable MD simulations without erratic fluctuations.

Finally, the production phase begins and the actual evolution of the system will be obtained. A thermostat^{18,19} and a barostat²⁰ can be added to account for the temperature and the pressure. While the system evolves, its configuration (velocities and coordinates of each atom) is stored within a given sampling interval, which should be carefully considered depending on the time-scale of the phenomena under study. If it is too high we can lose some vital frames of the trajectory, while if it is too low we would get humongous large files, thus increasing the difficulty of the posterior analysis. A simplified scheme of a typical MD running is showed in Scheme A.1.



Scheme A.1 - Simplification of a MD simulation. The creation of the system is highlighted in blue and the initialization in yellow. Depending on the objective, the steps highlighted in orange could either represent the equilibration or the production phase.

IV - QM/MM hybrid approaches: ONIOM

The ONIOM (Our own N-layered Integrated molecular Orbital and molecular Mechanics) approach implemented in the Gaussian program suite has been the elected QM/MM

¹⁸ Nosé, A. *J. Chem. Phys.* **1984**, *81*, 511–519.

¹⁹ Berendsen, H. J. C.; Postma, J. P. M.; Gunsteren, W. F. van; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

²⁰ Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

methodology in this work. This method was designed by Morokuma and coworkers²¹ and allows the treatment of large systems by defining two or three regions or layers that are treated at a different level of theory.²² Several calibration studies^{21,23,24 102,192,193} have demonstrated that the results obtained using this approach are similar to those obtained by treating the system with a high accuracy method alone.

IV.I - Defining the layers

The two layers in the ONIOM approach are the *High Layer* (HL) and the *Low Layer* (LL) (Figure A.2):

- The HL corresponds to the one treated with the highest accuracy level, normally a QM approach. This layer is the smallest one and corresponds to the area where the catalytic process takes place.
- The LL corresponds to the whole system treated with the lowest accuracy level, normally a MM approach. The modelling of this region affects the environment of the HL.

²¹ Byun, K. S.; Morokuma, K. *J. Mol. Struct. THEOCHEM* **1999**, 461-462, 1–21.

²² Only the two-layers approach will be explained, as it is the one used in this thesis. For more information please refer to general Gaussian manual (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, Revision A.1 2009*).

²³ Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **2003**, 24, 760.

²⁴ Vreven, T.; Morokuma, K. *Annu. Rep. Comput. Chem.* **2006**, 2, 35–51.

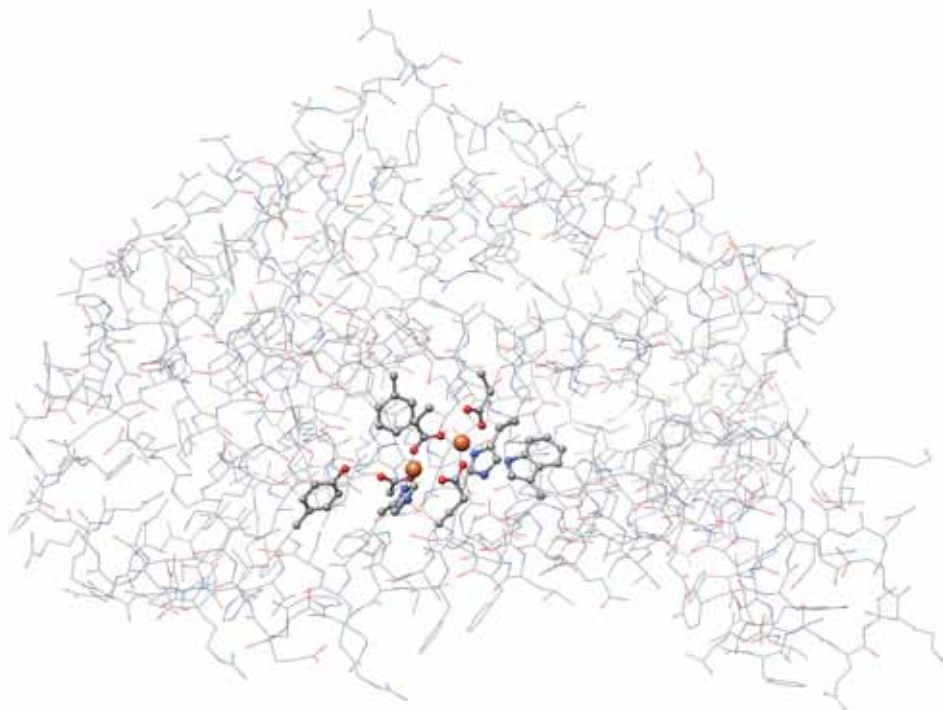


Figure A.2 - Structure of a Ribonucleotide Reductase (PDB code: 1XIK) prepared for an ONIOM calculation. Only the atoms surrounding the iron atoms are treated in the *HL* (ball and stick representation), the rest of the protein is in the *LL* (wire representation)

Some considerations must be taken into account in the ONIOM approach when assigning which part of the system will be included in each layer:

- 1- Bond breaking or formation should not take place in any region treated with a MM approach alone.
- 2- Atoms linked by double or triple bonds should be placed in the same ONIOM layer, including resonant structures like aromatic rings.
- 3- In a catalytic process, breaking and forming new bonds in the *HL* could affect some angular and dihedral MM parameters of the atoms in the boundary region. It is recommended that the MM parameters remain unaltered in the *LL* region (except from those included in the *HL* region) in both reactants and products. To be safe, the *HL* should include at least three atoms distance between the place where the catalysis takes place and the boundary.

IV.II - The link atoms

When partitioning a system to treat it with a QM/MM approach, two different things can happen. If there is no covalent bond between the two regions created, then no special treatment is needed for the boundary between the two layers. However, this is just a

special case and in many occasions covalent bonds are cut as a result of the partition. When this happens, dangling bonds appear at the interface of the two layers, creating an unrealistic chemical model in the *HL*. To avoid this problem, all the unpaired atoms as a result of the partitioning are paired with a hydrogen atom, also called *link atom*. This atom is placed along the vector of the previously existing bond and the length of this new bond is a function of the nature of the original bond. No considerations have to be made in the *LL* as it engulfs the whole system.

IV.III - Calculating the energy

The total energy is calculated from the energies of the different parts of the system (*HL* and *LL*):

$$E_{ONIOM} = E_{MM}^{LL} + E_{QM}^{HL} - E_{MM}^{HL} \quad \text{Eq A.23}$$

The *HL* region is also treated with a MM approach (the *LL* encloses the whole system), so the calculated energy of this region with that approach (E_{MM}^{HL}) has to be subtracted from the total in order to avoid an overestimation of the energy in that region.



Annex B: Integrative Interface Examples

I - Programs used on the Integrative Interface

I.I - UCSF Chimera



UCSF Chimera is a highly extensible visualization program developed by the Resource for Biocomputing, Visualization and Informatics (RBVI) department at the University of California, San Francisco (UCSF). Chimera can be downloaded free of charge for academic, government, non-profit, and personal use from its web page (<http://www.cgl.ucsf.edu/chimera/>) and its development is funded by the National Institute of Health (NIH) of the USA government. It has a highly intuitive and friendly-user interface for the visualization of molecular structures and provides several different structural analysis tools, including hbond analysis, geometrical features, clusterization of MD trajectories, RMSD maps, etc.

I.II - Molecular Modelling Toolkit



The Molecular Modelling Toolkit (MMTK) is an open source program library for molecular simulations. It provides a highly extensive ready-to-use library of different standard molecular modelling techniques and provides an easy way to import and modify these algorithms. This allows the usage of the MMTK to deal with both standard and non-standard molecular problems. Many advantages of the MMTK in respect to other similar molecular modelling approaches arise from its scripting language: Python. In addition to perform molecular simulations by writing simple python scripts importing the appropriate python MMTK molecules, the user can use at the same time several custom python modules to tune the simulation for its own necessities. Additionally, the user can write or modify existing modules to solve problems for which no standard solution exists (i.e. adding a particular force field term). This last feature is one of the main assets of the MMTK as the major part of the available molecular modelling suites tends to use a black box¹ approach rather to allow the user to easily modify the internal codes.

¹ A software uses a black box approach when it does not allow the user to modify or analyze what is happening inside the core algorithms. The user can only provide with the initial input and the program will generate the corresponding output.

I.III - Python



Python has consecrated as one of the most used programming languages in the science field for several reasons. Primarily, it is considered a high level language², which makes it easier to learn by scientists that are not specialist programmers. Python code is easy to read and write than other widespread programming languages, which facilitates the writing and diffusion of the developed algorithms. The syntax is easy-to-read and it contains several different built-in functions to interact with data structures and objects. Additionally, it offers an interactive command line (IDLE) which makes code development much more dynamic. The researcher can directly start playing with the code rather than to spend time thinking about how to write it. Python integrates really well with other programming languages (i.e. C/C++), which allows us to write some parts of the algorithm in other programming languages more efficient for that specific purpose.

² A programming language is considered high level if it can encode many different actions in just a few lines

II - MD example

In this example we will use the MD interface to describe the behavior of a glucagon peptide (hereafter referred as gluc, PDB code 1GCN)³ in a water box using Periodic Boundary Conditions (PBC). This peptide hormone is secreted by the α -cells of the pancreas when sugar levels in blood are too low, raising them and restoring the body homeostasis. Its small size (29 peptides) and its stable folding (α helix) turn it into a good example to show how the MD interface works. The steps described here should be transferable to any protein system, including protein-ligand complexes.

II.I - Preparing the PDB file

First we will download the PDB file (code 1GCN) containing the gluc X-ray structure. We can either go to “**File/Open...**” or use the “**open**” command in the command line. As usually in the raw PDB files we need to add the missing hydrogens and assign the charges before running the simulation (Figure B.1). To do so first open the MD interface (“**Tools/MD Ensemble Analysis/Molecular Dynamics Simulation**”) and invoke the **DockPrep** routine of Chimera from the **Set Up** tab. In the DockPrep interface select the **Add hydrogens** and **Add charges** options.

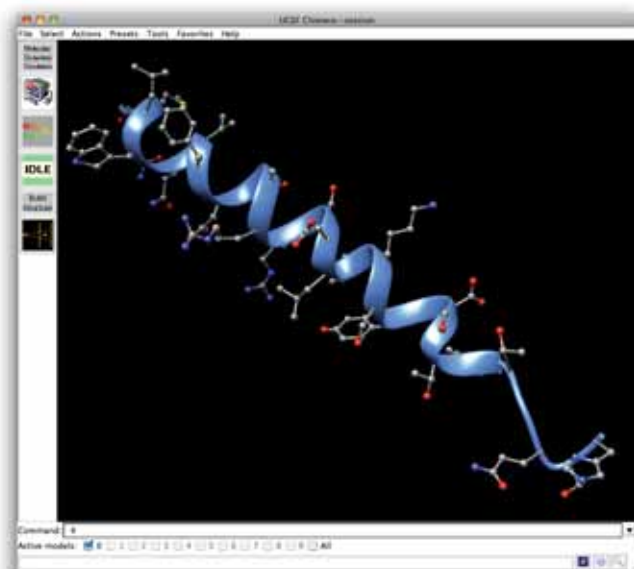


Figure B.1 - Gluc PDB file loaded in Chimera. The raw PDB file does not contain neither the hydrogens nor the associated charges to each atom so we have to add them prior to the MD simulation.

³ Sasaki, K.; Dockerill, S.; A. Adamiak, D. *Nature*. **1975**, 257, 751-757.

Next we will put our system in a water box with PBC. Go to the **Solvation** tab and click on the **Solvate** button. For the purpose of this example we selected a box shape with a size of 6Å filled with TIP3PBOX water molecules. We do not need to add the counterions in this case because the total charge of the system is already zero. Check the “**Use periodic boundary conditions**” option and let the program to automatically adjust the size of the system (“**Automatic box size**”). Figure B.2 displays how the system should look like at this point.

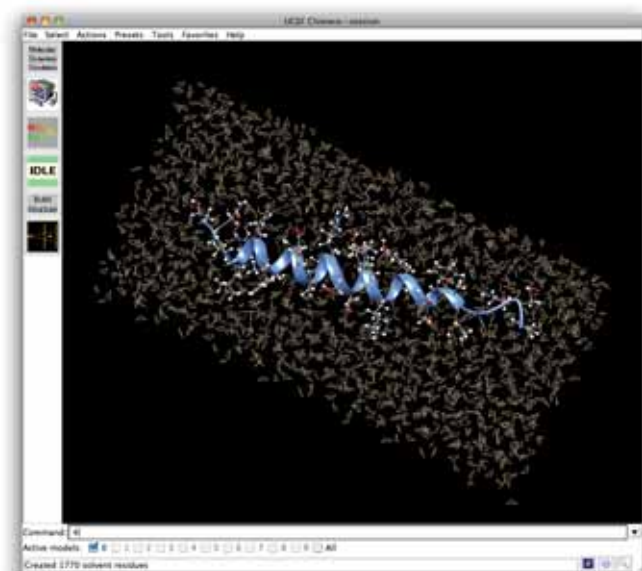


Figure B.2 - Gluc peptide solvated with TIP3PBOX waters

II.II - Setting up the MD simulation

We will use the following scheme for the MD simulation: (i) minimize the system, (ii) heat it up until 298K, (iii) run a short equilibration and finally, (iii) run the production stage. All these steps can be defined in the **MD Options** tab.

To minimize the system click the **Minimization** button. For the purpose of this example we will use 1000 steepest descend and 10 conjugate gradient steps with the default step size values in both cases.

To define the heating process and the equilibration click on the **Equilibration** button. We will heat the system up from 0K to 298K using a gradient of 0.5K/ps and maintaining the heater on during all the equilibration (“**Maintain ad infinitum**” option). In the example we will run 10000 steps with the default integration step value. At this point we need to

define two trajectory files. The first one will store the actual trajectory created during the equilibration (“**Equilibration trajectory file:**”). The second one will be used to re-assign the initial velocities to the system once the production stage begins (“**Restart equilibration trajectory file:**”).

Finally, to set up the production phase click on the **Production** button. In this example we will run 25000 steps using the default integration step value. Additionally, we will include all the available options (Andersen Barostat, Barostat Reset, Velocity Scaler and Noose Thermostat) with the default values in each option (298K and 1.0132bars). You can define where to store the final trajectory in the “**Production trajectory file:**” option.

II.III - Running the MD

In this example we will neither add any restraint to the force field (**Constraint** tab) nor modify the default electrostatic and Lennard-Jones potential values (**Miscellanea** tab). The translational and rotational removers are added by default in the simulation (**Miscellanea** tab).

To run the simulation we will go to the **Running** tab and define multiple processors to speed up the process (in our case we used 8 cores). Additionally, we will open the whole trajectory every 10 frames to simplify the posterior analysis. Last thing to do is to click on the **Run** button and wait for the results.

II.IV - Analyzing the results

Once the simulation is finished, the MD movie dialog will appear on the screen with the corresponding trajectory loaded. Now we can visualize the movement of the gluc peptide alongside the 35000 steps of the simulation (Figure B.3). To plot the temperature and both the kinetic and the potential energies go to “**Analysis/Plot**” and select the physical variable you want to represent (Figure B.4).

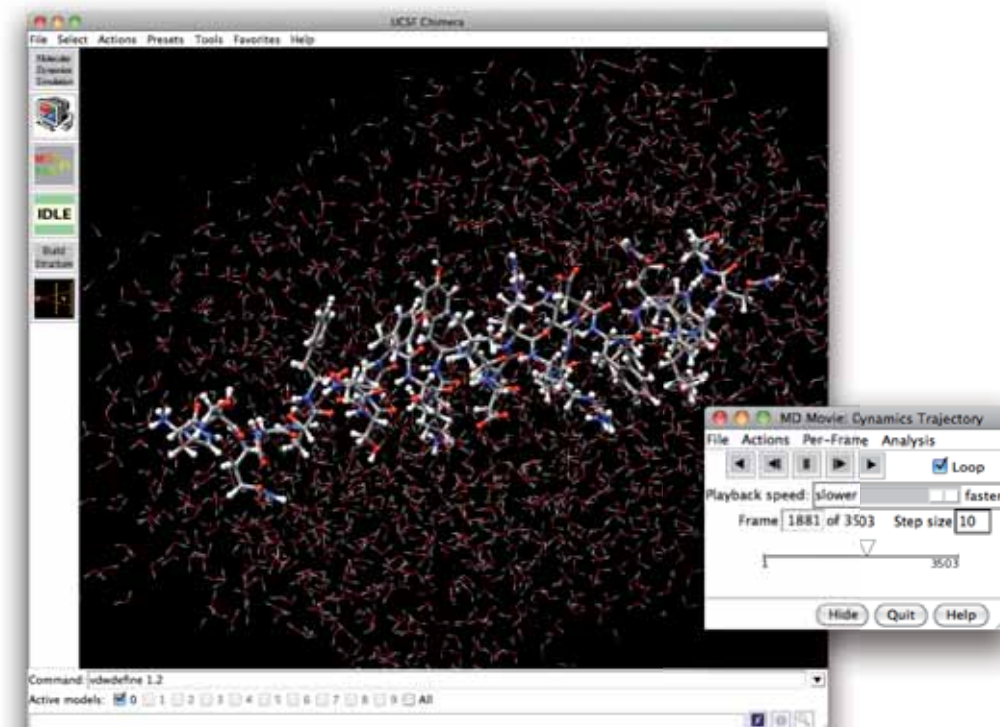


Figure B.3 - MD trajectory resulting from the simulation of the gluc peptide

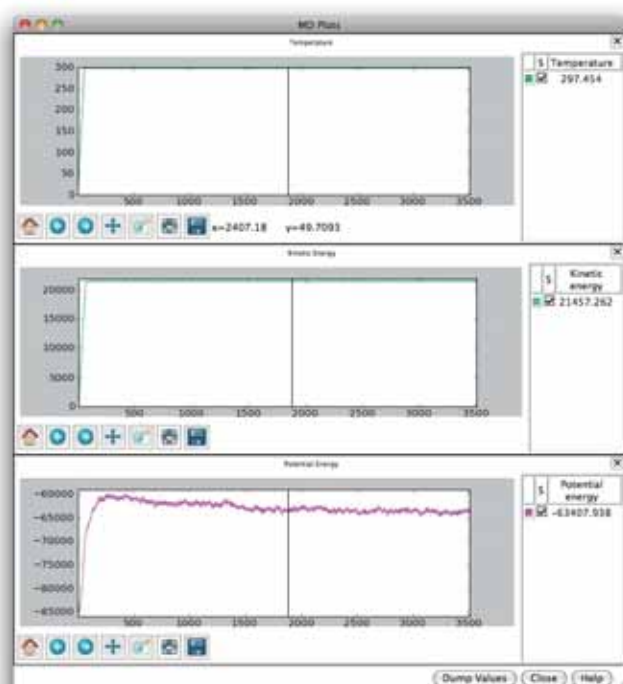


Figure B.4 - Plot of temperature (**top**), kinetic (**middle**) and potential (**bottom**) energies during the gluc MD simulation

III - NMA example

In this example we will calculate the normal modes of the hinge domain of a mouse condensin, a protein involved in the repairing mechanism of single-strand DNA damage. This region is highly flexible and allows large collective movements of the protein, thus we can study them through NMA.

To perform a standard NMA simulation first we need to minimize the system. Once loaded the PDB file (code 3L51)⁴ in Chimera, invoke the NMA interface (“**Tools/Structure Analysis/Calculate Normal Modes**”) and click on the “**Minimizer...**” button. For the purpose of this example we will run 2000 steepest descend steps and 100 conjugate gradient steps using the default step size values in both cases. Once the minimization is over select the “**FFM**” option on the “**Calculation procedure:**” menu. We will use the standard parameters for the simulation (e.g. Lennard-Jones and electrostatic potentials), so we can directly click the “**Run**” button to launch it. Due to the large size of the system, the calculation may take a while.

Once the simulation is over, the visualization interface will appear with all the calculated modes. The first six ones (from mode 0 to 5) correspond to translational/rotational movements, so they do not offer any relevant information. However, starting from mode 6, they describe the large collective movements we are seeking. In fact, mode number 6 (Figure B.5) displays a large opening/closing movement of the large cavity between the two regions of the protein. This kind of movement is typical in the hinge domains.

⁴J. Griese, J.; Witte, G.; Hopfner, K.-P. *Nucleic Acids Res.* **2010**, *38*, 3454–3465.

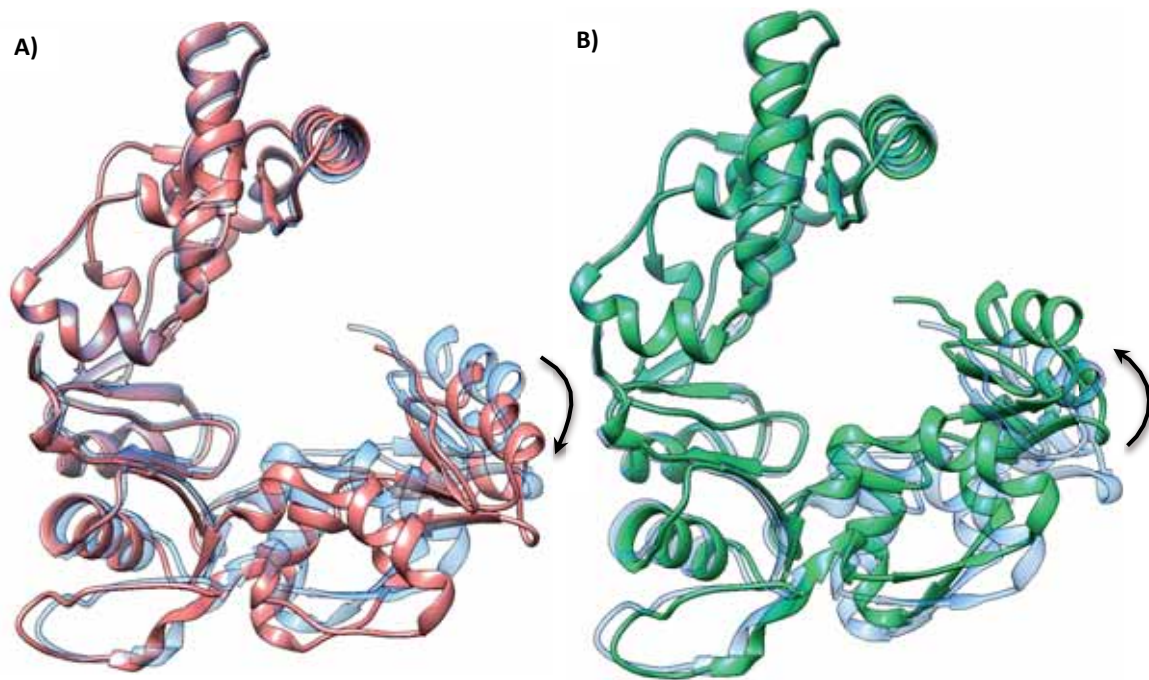


Figure B.5 - Calculated mode number 6 of the hinge domain of the mouse condensin (PDB code 3L51). The blue colored protein is the minimized form of the PDB structure while the red (A) and green (B) ones correspond to the positive and negative displacement along the normal mode, respectively.



Annex C: Articles

What can molecular modelling bring to the design of artificial inorganic cofactors?[†]

Victor Muñoz Robles, Elisabeth Ortega-Carrasco,
Eric González Fuentes, Agustí Lledós and Jean-Didier Maréchal*

Received 23rd March 2010, Accepted 28th April 2010

DOI: 10.1039/c004578k

In recent years, the development of synthetic metalloenzymes based on the insertion of inorganic catalysts into biological macromolecules has become a vivid field of investigation. The success of the design of these composites is highly dependent on an atomic understanding of the recognition process between inorganic and biological entities. Despite facing several challenging complexities, molecular modelling techniques could be particularly useful in providing such knowledge. This study aims to discuss how the prediction of the structural and energetic properties of the host–cofactor interactions can be performed by computational means. To do so, we designed a protocol that combines several methodologies like protein–ligand dockings and QM/MM techniques. The overall approach considers fundamental bioinorganic questions like the participation of the amino acids of the receptor to the first coordination sphere of the metal, the impact of the receptor/cofactor flexibility on the structure of the complex, the cost of inserting the inorganic catalyst in place of the natural ligand/substrate into the host and how experimental knowledge can improve or invalidate a theoretical model. As a real case system, we studied an artificial metalloenzyme obtained by the insertion of a Fe(Schiff base) moiety into the heme oxygenase of *Corynebacterium diphtheriae*. The experimental structure of this species shows a distorted cofactor leading to an unusual octahedral configuration of the iron with two proximal residues chelating the metal and no external ligand. This geometry is far from the conformation adopted by similar cofactors in other hosts and shows that a fine tuning exists between the coordination environment of the metal, the deformability of its organic ligand and the conformational adaptability of the receptor. In a field where very little structural information is yet available, this work should help in building an initial molecular modelling framework for the discovery, design and optimization of inorganic cofactors. Moreover, the approach used in this study also lays the groundwork for the development of computational methods adequate for studying several metal mediated biological processes like the generation of realistic three dimensional models of metalloproteins bound to their natural cofactor or the folding of metal containing peptides.

Departament de Química, Universitat Autònoma de Barcelona, Edifici C.n., 08193 Bellaterra, Barcelona, Spain. E-mail: jeandidier.marechal@uab.es; Fax: +34 93 581 2920; Tel: +34 93 581 4936

[†] Electronic supplementary information (ESI) available: Further structural analysis on experimental and computed structures are provided as figures and data tables. See DOI: 10.1039/c004578k

CHEMBIOCHEM

DOI: 10.1002/cbic.201100659

Incorporation of Manganese Complexes into Xylanase: New Artificial Metalloenzymes for Enantioselective Epoxidation

Mathieu Allard,^[a] Claude Dupont,^[b] Victor Muñoz Robles,^[c] Nicolas Doucet,^[b] Agustí Lledós,^[c] Jean-Didier Maréchal,^[c] Agathe Urvoas,^[d] Jean-Pierre Mahy,^{*[a]} and Rémy Ricoux^[a]

Here we report the best artificial metalloenzyme to date for the selective oxidation of aromatic alkenes; it was obtained by noncovalent insertion of Mn^{II}-meso-tetrakis(*p*-carboxyphenyl)porphyrin [Mn(TpCPP), 1-Mn] into a host protein, xylanase 10A from *Streptomyces lividans* (Xln10A). Two metallic complexes—*N,N'*-ethylene bis(2-hydroxybenzylimine)-5,5'-dicarboxylic acid Mn^{II} [(Mn-salen), 2-Mn] and 1-Mn—were associated with Xln10A, and the two hybrid biocatalysts were characterised by UV-visible spectroscopy, circular dichroism and molecular

modelling. Only the artificial metalloenzyme based on 1-Mn and Xln10A was studied for its catalytic properties in the oxidation of various substituted styrene derivatives by KHSO₅; after optimisation, the 1-Mn-Xln10A artificial metalloenzyme was able to catalyse the oxidation of *para*-methoxystyrene by KHSO₅ with a 16% yield and the best enantioselectivity (80% in favour of the *R* isomer) ever reported for an artificial metalloenzyme.

Introduction

Epoxidation of olefins is a widely used process in the chemical industry. Epoxides are valuable intermediates for laboratory syntheses and chemical manufacturing because they can easily be transformed into a large variety of compounds through regioselective ring-opening reactions.^[1] The two main methods for obtaining epoxides are homogeneous catalysis, using classical metalloporphyrin catalysts,^[2] and biocatalysis;^[3] the two approaches are in many aspects complementary. Enzymes work under mild conditions and provide high regioselectivities. Synthetic catalysts are more widely applicable and accept a wider range of substrates, but have some disadvantages, one of the most crucial being their destruction during the course of the reaction.

With regard to biocatalysts, for example, many naturally occurring peroxidases, such as myeloperoxidase (MPO), and *Coprinus cinereus* peroxidase (CIP), have been reported to catalyse the enantioselective epoxidation of styrene by H₂O₂ with moderate levels of conversion but good enantiomeric excesses (16% conversion after 16 h and up to 80% ee).^[4] Furthermore, a novel styrene monooxygenase was found to be able to catalyse the epoxidation of styrene with good yield and very high stereoselectivity (49% yield and up to 99% ee).^[5] Of the several strategies explored to induce stereoselectivity into chemical reactions, the construction of artificial metalloenzymes has appeared to be one of the most promising. Indeed, such hybrid biocatalysts combine the efficiency and wide scope of reactions of synthetic catalysts with enzymes' high selectivity and ability to function under mild conditions. Different strategies for creating artificial metalloenzymes by incorporation of metal complexes into a protein-binding site—including covalent,^[6–8] dative^[9,10] or noncovalent^[11–13] anchoring strategies—have been investigated. By these approaches it has been possible to

obtain different artificial metalloenzymes capable of catalysing hydrogenation,^[14,15] sulfoxidation,^[16,17] dihydroxylation,^[18] Diels–Alder,^[19,20] transamination^[21] and epoxidation^[22–24] reactions.

For the last of these, only a few results have been obtained with artificial metalloenzymes. Two different strategies have been investigated. The first was the incorporation of a manganese-salen complex through a covalent linkage into papain, which led to hybrid catalysts that were able to perform the enantioselective epoxidation of styrene but the ee values turned out to be less than 10%.^[22] The second, developed by the teams of Kazlauskas^[23] and Soumillion,^[24] was the substitution of the zinc ion of carbonic anhydrase by a manganese ion, known to promote oxidation reactions. Both groups obtained artificial metalloenzymes that catalysed the enantioselective epoxidation of styrene, but with poor levels of conversion

[a] M. Allard, Prof. J.-P. Mahy, Dr. R. Ricoux
Institut de Chimie Moléculaire et des Matériaux d'Orsay
UMR 8182 CNRS, Laboratoire de Chimie Bioorganique et Bioinorganique
Bâtiment 420, Université Paris XI, 91405 Orsay Cedex (France)
E-mail: jean-pierre.mahy@u-psud.fr

[b] Prof. C. Dupont, Prof. N. Doucet
INRS-Institut Armand-Frappier, Université du Québec
531 Boulevard des Prairies, Laval, Québec, H7V 1B7 (Canada)

[c] V. Muñoz Robles, Prof. A. Lledós, Dr. J.-D. Maréchal
Departament de Química, Universitat Autònoma de Barcelona
Edifici C.n., 08193 Cerdanyola del Vallès, Barcelona (Spain)

[d] Dr. A. Urvoas
Institut de Biochimie et de Biophysique Moléculaire et Cellulaire
UMR 8619 CNRS, Laboratoire de Modélisation et d'Ingénierie des Protéines
Bâtiment 430, Université Paris XI, 91405 Orsay Cedex (France)

Crystal Structure of Two Anti-Porphyrin Antibodies with Peroxidase Activity

Victor Muñoz Robles^{1*}, Jean-Didier Maréchal¹, Amel Bahloul^{2,3a}, Marie-Agnès Sari³, Jean-Pierre Mahy⁴, Béatrice Golinelli-Pimpaneau^{2*,†}

1 Département de Chimie, Université Autonoma de Barcelona, Edifici C.n., 08193 Cerdanyola del Vallès, Barcelona, Spain, **2** Laboratoire d'Enzymologie et Biochimie structurales, CNRS, Centre de Recherche de Gif, Gif-sur-Yvette, France, **3** Laboratoire de Chimie et Biochimie Pharmacologiques et Toxicologiques, CNRS, Université Paris Descartes, Paris, France, **4** Institut de Chimie Moléculaire et des Matériaux d'Orsay, CNRS, Laboratoire de Chimie Biorganique et Bioinorganique, CNRS, Bât 420, Université Paris 11, Orsay, France

Abstract

We report the crystal structures at 2.05 and 2.45 Å resolution of two antibodies, 13G10 and 14H7, directed against an iron(III)- α -carboxyphenylporphyrin, which display some peroxidase activity. Although these two antibodies differ by only one amino acid in their variable λ -light chain and display 86% sequence identity in their variable heavy chain, their complementary determining regions (CDR) CDRH1 and CDRH3 adopt very different conformations. The presence of Met or Leu residues at positions preceding residue H101 in CDRH3 in 13G10 and 14H7, respectively, yields to shallow combining sites pockets with different shapes that are mainly hydrophobic. The hapten and other carboxyphenyl-derivatized iron(III)-porphyrins have been modeled in the active sites of both antibodies using protein ligand docking with the program GOLD. The hapten is maintained in the antibody pockets of 13G10 and 14H7 by a strong network of hydrogen bonds with two or three carboxylates of the carboxyphenyl substituents of the porphyrin, respectively, as well as numerous stacking and van der Waals interactions with the very hydrophobic CDRH3. However, no amino acid residue was found to chelate the iron. Modeling also allows us to rationalize the recognition of alternative porphyrinic cofactors by the 13G10 and 14H7 antibodies and the effect of imidazole binding on the peroxidase activity of the 13G10/porphyrin complexes.

Citation: Muñoz Robles V, Maréchal J-D, Bahloul A, Sari M-A, Mahy J-P, et al. (2012) Crystal Structure of Two Anti-Porphyrin Antibodies with Peroxidase Activity. PLoS ONE 7(12): e51128. doi:10.1371/journal.pone.0051128

Editor: Claudine Mayer, Institut Pasteur, France

Received: September 6, 2012; **Accepted:** October 30, 2012; **Published:** December 11, 2012

Copyright: © 2012 Muñoz Robles et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Financial support was provided to B.G.-P. and J.-P.M. by the Centre National de la Recherche Scientifique (Programme Physique et Chimie du Vivant) and to J.-D.M. and V.M. by the Spanish "Ministerio de Ciencia e Innovación" through projects CTQ2008-06866-C02-01 and consolidat-inferio 2010 and by the Generalitat de Catalunya through project 2009SGR68. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: beatrice.golinelli@college-de-france.fr

†a Current address: Département de Neurosciences, Institut Pasteur, Paris, France

†b Current address: Laboratoire de Chimie des Processus Biologiques, Collège de France, CNRS, Paris, France

† These authors contributed equally to this work.

Introduction

Hemoproteins contain iron-protoporphyrin IX or heme as the prosthetic group, whose divalent iron atom can reversibly bind molecules such as molecular oxygen, leading to a wide range of biological functions [1]. Chemical or biotechnological models of hemoproteins have thus long been developed in order to create selective catalysts for industrial and fine chemistry and to predict the oxidative metabolism of new drugs [2,3,4,5]. Examples include the *de novo* design of heme proteins, including that of membrane-soluble proteins [6,7]. Peroxidases appear to be the easiest hemoproteins to be mimicked. Indeed, their active site consists of the iron(III)-porphyrin moiety encapsulated in the apoprotein. On one side, the heme iron is bound to an axial histidine residue (proximal ligand) and on the other side to the peroxide substrate to lead to an iron-oxo complex. The radical cation on the iron (IV)-oxo porphyrin ring can be delocalized onto proximal protein side chains [8]. The reducing cosubstrate does not bind to a well-defined site on the inside of the protein, as peroxidases restrict access of substrates to the heme-oxo complex, so that the electron

transfer occurs to the *meso* edge of the heme [9]. Heterolytic cleavage of the O-O bond is assisted by general acid base catalysis through the concerted action of the distal histidine and arginine residues [10]. A major problem in homogeneous metalloporphyrin systems mimicking hemoproteins is that the catalyst is often destroyed by oxidation during the course of the reaction and it is difficult to combine reactivity and selectivity in these models. The use of a protein such as sylanase A [11] or an antibody mimicking the protein matrix of heme enzymes not only prevents aggregation and intermolecular self-oxidation of the catalyst, but can also influence the selectivity of the reaction [12]. As the antibody has the role of a host molecule that enhances the function of porphyrin, the porphyrin itself can be used as the hapten to induce the antibodies.

In order to generate antibodies with peroxidase activity, mice have been immunized against iron(III)- α , α , α , β -*meso*-tetra*ortho*-carboxyphenylporphyrin (Fe(ToCPP)) (Figure 1) [13,14]. Two antibodies, 13G10 and 14H7, were found to bind the porphyrin haptens with nanomolar affinities and enhance its peroxidase

J Biol Inorg Chem
DOI 10.1007/s00775-013-1088-z

ORIGINAL PAPER

Monitoring lactoferrin iron levels by fluorescence resonance energy transfer: a combined chemical and computational study

Fernando Carmona · Víctor Muñoz-Robles ·
Rafael Cuesta · Natividad Gálvez · Mercè Capdevila ·
Jean-Didier Maréchal · José M. Domínguez-Vera

Received: 30 August 2013 / Accepted: 27 December 2013
© SBIC 2014

Abstract Three forms of lactoferrin (Lf) that differed in their levels of iron loading (Lf, LfFe, and LfFe₂) were simultaneously labeled with the fluorophores AF350 and AF430. All three resulting fluorescent lactoferrins exhibited fluorescence resonance energy transfer (FRET), but they all presented different FRET patterns. Whereas only partial FRET was observed for Lf and LfFe, practically complete FRET was seen for the holo form (LfFe₂). For each form of metal-loaded lactoferrin, the AF350–AF430 distance varied depending on the protein conformation, which in turn depended on the level of iron loading. Thus, the FRET patterns of these lactoferrins were found to correlate with their iron loading levels. In order to gain greater insight into the number of fluorophores and the different FRET patterns observed (i.e., their iron levels), a computational analysis was performed. The results highlighted a number of lysines that have the greatest influence on the FRET profile. Moreover, despite the lack of an X-ray structure for any LfFe

species, our study also showed that this species presents modified subdomain organization of the N-lobe, which narrows its iron-binding site. Complete domain rearrangement occurs during the LfFe to LfFe₂ transition. Finally, as an example of the possible applications of the results of this study, we made use of the FRET fingerprints of these fluorescent lactoferrins to monitor the interaction of lactoferrin with a healthy bacterium, namely *Bifidobacterium breve*. This latter study demonstrated that lactoferrin supplies iron to this bacterium, and suggested that this process occurs with no protein internalization.

Keywords Lactoferrin · Iron metabolism · Protein–ligand docking · FRET · Structural analysis

Introduction

Lactoferrin is a glycoprotein (80 kDa) of the transferrin family with a high affinity for iron(III) [1, 2]. Lactoferrin possesses various biological functions, including antibacterial, antiviral, and antiparasitic activities [3]. High lactoferrin levels are found in colostrum and milk, and this protein is also present in most mucosal secretions, including uterine fluid, vaginal, and nasal secretions, as well as in tears [4, 5].

The extraordinary affinity of lactoferrin for iron undoubtedly determines part of its functionality. Indeed, lactoferrin is considered to form part of the innate immune system due to its effects on pathogen growth. Iron is essential for life and is a key nutrient for pathogenic microorganisms, which require this metal to survive and replicate. Life can to some extent be considered a battle for iron, so hosts must deprive undesirable guests of iron in order to combat the infections they cause. As a result, iron

Electronic supplementary material The online version of this article (doi:10.1007/s00775-013-1088-z) contains supplementary material, which is available to authorized users.

F. Carmona · N. Gálvez · J. M. Domínguez-Vera (✉)
Departamento de Química Inorgánica, Facultad de Ciencias,
Instituto de Biotecnología, Universidad de Granada,
18071 Granada, Spain
e-mail: josema@ugr.es

V. Muñoz-Robles · M. Capdevila · J.-D. Maréchal (✉)
Departament de Química, Facultat de Ciències,
Universitat Autònoma de Barcelona,
Cerdanyola del Vallès, 08193 Barcelona, Spain
e-mail: jeandidier.marechal@uab.cat

R. Cuesta
Departamento de Química, Escuela de Linares,
Universidad de Jaén, 23700 Linares, Spain

Published online: 18 January 2014

 Springer

Computational Insights on an Artificial Imine Reductase, Based on the Biotin–Streptavidin Technology

Victor Muñoz Robles,[†] Pietro Vidossich,[‡] Agustí Lledós,[‡] Thomas R. Ward,[‡] and Jean-Didier Maréchal^{*†‡}

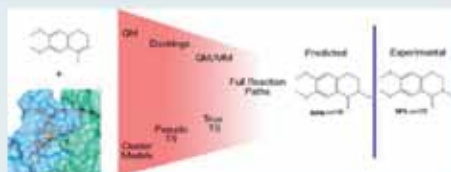
[†]Departament de Química, Universitat Autònoma de Barcelona, Edifici C.n., 08193 Cerdanyola del Vallés, Barcelona, Spain.

[‡]Department of Chemistry, University of Basel, Spitalstrasse 51, CH-4056 Basel, Switzerland.

Supporting Information

ABSTRACT: We present a computational study that combines protein–ligand docking, quantum mechanical, and quantum mechanical/molecular calculations to scrutinize the mechanistic behavior of the first artificial enzyme able to enantioselectively reduce cyclic imines. We applied a novel strategy that allows the characterization of transition-state structures in the protein host and their associated reaction paths. Of the most striking results of our investigation is the identification of major conformational differences between the transition-state geometries of the lowest energy paths leading to (*R*)- and (*S*)-reduction products. The molecular features of (*R*)- and (*S*)-transition states highlight distinctive patterns of hydrophobic and polar complementarities between the substrate and the binding site. These differences lead to an activation energy gap that stands in very good agreement with the experimentally determined enantioselectivity. This study sheds light on the mechanism by which transfer hydrogenases operate and illustrates how the change of environment (from homogeneous solution conditions to the asymmetric protein frame) affect the reactivity of the organometallic co-factor. It provides novel insights on the complexity in integrating unnatural organometallic compounds into biological scaffolds. The modeling strategy that we pursued, based on the generation of “pseudo-transition state” structures, is computationally efficient and suitable for the discovery and optimization of artificial enzymes. Alternatively, this approach can be applied on systems for which a large conformational sampling is needed to identify relevant transition states.

KEYWORDS: artificial enzymes, asymmetric catalysis, imine reduction, computational chemistry, QM/MM, protein–ligand dockings



INTRODUCTION

Artificial enzymes, which result from the incorporation of a catalytically competent metal co-factor within a protein environment, are attracting attention as alternatives to more-traditional homogeneous and enzymatic catalysts.^{1–5}

Several complementary strategies to create non-natural enzymes have been pursued, with varying degrees of success.^{6–9} As generating catalytic function *ex nihilo* remains challenging, a promising approach consists of the incorporation of a catalytically competent moiety within a biomolecular environment (protein or oligonucleotide).¹⁰ These hybrids, in essence, follow the conceptual framework of natural metalloenzymes: the co-factor, including its first coordination sphere, by and large dictates the catalytic activity, while the protein, which provides the second coordination sphere environment, controls substrate selectivity.¹¹ Within this framework, the choice of macromolecular scaffold dictates the anchoring strategy (i.e., dative,¹² covalent,¹³ or supramolecular,^{8,14}) whereby the artificial co-factor determines which transformation can be catalyzed.^{5–9} Since the selected macromolecular host was not optimized by nature to accommodate the artificial co-factor or to catalyze the transformation considered, genetic optimization offers nearly limitless opportunities to improve the catalytic performance of artificial metalloenzymes.

Inspired by a visionary report by Whitesides in 1978,¹⁵ the Ward group and others have been pursuing the biotin–(strept)avidin technology to create artificial metalloenzymes.^{7,15–19} Following this supramolecular anchoring strategy, the introduction of a biotinylated organometallic moiety within streptavidin (STREP) affords artificial metalloenzymes for a variety of transformations, including hydrogenation, allylic alkylation, metathesis, C–H activation, alcohol oxidation, sulfoxidation, dihydroxylation, and transfer hydrogenation.^{7,10,20–23} These designs take advantage of the high affinity of STREP for its natural ligand biotin ($K_M \approx 10^{-13}$ M). The structure of STREP is best described as a homotetrameric eight-stranded β -barrel with two close-lying biotin-binding sites. It is generally accepted that biotin-binding events are noncooperative.²⁴ In the context of asymmetric transfer hydrogenation, we have relied on the introduction of a biotin anchor on a Noyori-type aminosulfonamide-bearing d⁶-piano-stool moiety. In the presence of STREP, the biotinylated piano-stool is quantitatively incorporated in the bowl-shaped biotin-binding vestibule STREP (Figure 1). Following promising results obtained for the enantioselective reduction of

Received: October 14, 2013

Revised: December 11, 2013

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Structural-, Kinetic- and Docking Studies of Artificial Imine Reductases Based on the Biotin-Streptavidin Technology: An Induced Lock-and-Key Hypothesis

Victor Muñoz Robles,^{a,†} Marc Dürrenberger,^{b,†} Tillmann Heinisch,^c Agustí Lledós^a Tilman Schirmer,^c Thomas R. Ward^{b,*} and Jean-Didier Maréchal^{a,*}

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

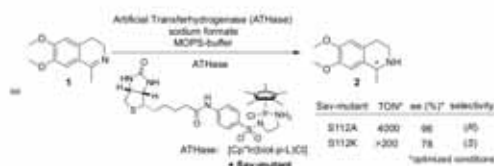
An artificial imine reductase results upon incorporation of a biotinylated Cp*Ir moiety (Cp*: C₅Me₅) within homo-tetrameric streptavidin (Sav). Mutation of S112 reveals a marked effect of the Ir/streptavidin ratio on both saturation kinetics as well as enantioselectivity for the production of salsolidine. For [Cp*Ir(Biot-*p*-L)Cl] ⊂ S112A Sav, both the reaction rate and the selectivity (up to 96% ee (*R*)-salsolidine, k_{cat} 14 min⁻¹ – 4 min⁻¹ vs. [Ir], K_M 65 mM – 370 mM) decrease upon fully saturating all biotin binding sites (the ee varying between 96% ee and 45% ee (*R*)). In contrast, for [Cp*Ir(Biot-*p*-L)Cl] ⊂ S112K Sav, both the rate and the selectivity remain nearly constant upon varying the Ir/Streptavidin ratio (up to 78% ee (*S*)-salsolidine, k_{cat} 2.6 min⁻¹, K_M 95 mM). X-ray analysis complemented with docking studies highlight a marked preference of the S112A and S112K Sav mutants for the (*S*₀) and (*R*₀) enantiomeric forms of the cofactor respectively. Combining both docking and saturation kinetic studies lead to the formulation of an enantioselection mechanism relying on an “induced lock-and-key” hypothesis: the host protein dictates the configuration of the biotinylated Ir-cofactor which, in turn, by-and-large determines the enantioselectivity of the imine reductase.

Introduction

Artificial metalloenzymes result from the incorporation of a catalytically competent organometallic moiety within a macromolecule.^{1–13} Thus far, three anchoring strategies have been pursued to ensure localization of the abiotic cofactor within a well defined second coordination sphere environment:¹⁴ covalent, dative or supramolecular. One of the most attractive features of such systems results from combining both chemical- and genetic-optimization strategies. In this context, the biotin-streptavidin technology has provided a propitious playground for the creation and optimization of artificial metalloenzymes.^{1,15–18} Tethering a biotin-anchor to a catalyst precursor ensures that, in the presence of streptavidin (Sav hereafter), the metal moiety is quantitatively incorporated within the host protein. Importantly, the dimer of dimer nature of the Sav homo-tetramer provides two ideally sized biotin-binding vestibules, each capable of accommodating (up to) two biotinylated catalysts as well as the corresponding substrates.¹⁹ However, as the biotin-binding vestibule is fairly shallow, upon incorporation the biotinylated catalyst tends to be poorly localized, as reflected by the low metal occupancy in the corresponding X-ray structures.^{19–22} This ill-defined cofactor localization, combined with the vast genetic optimization potential, offer an opportunity but also a challenge for rational structure-based design relying on *in silico* modeling.

We recently reported on an artificial imine reductase (Asymmetric Transfer Hydrogenase, ATHase hereafter) resulting

from incorporation of a biotinylated Cp*Ir-moiety within streptavidin ([Cp*Ir(Biot-*p*-L)Cl] ⊂ Sav hereafter) (Cp*: C₅Me₅, Scheme 1). We showed that, upon substituting Sav Ser112 by either an alanine or a lysine (i.e. S112A or S112K) both enantiomers of salsolidine **2** could be produced in 96 % ee (*R*-) and 78 % ee (*S*)-configuration respectively, Scheme 1.^{21,22} This single point mutation thus leads to a difference in transition state free energy $\Delta\Delta G^\ddagger$ 3.5 kcal·mol⁻¹ for both enantiomers at room temperature.²⁴ Herein, we present our efforts to rationalize the effect of point mutations on the structure and catalytic performance of both ATHases.



Scheme 1 - Artificial imine reductase (ATHase) for the production of both enantiomers of salsolidine **2**.