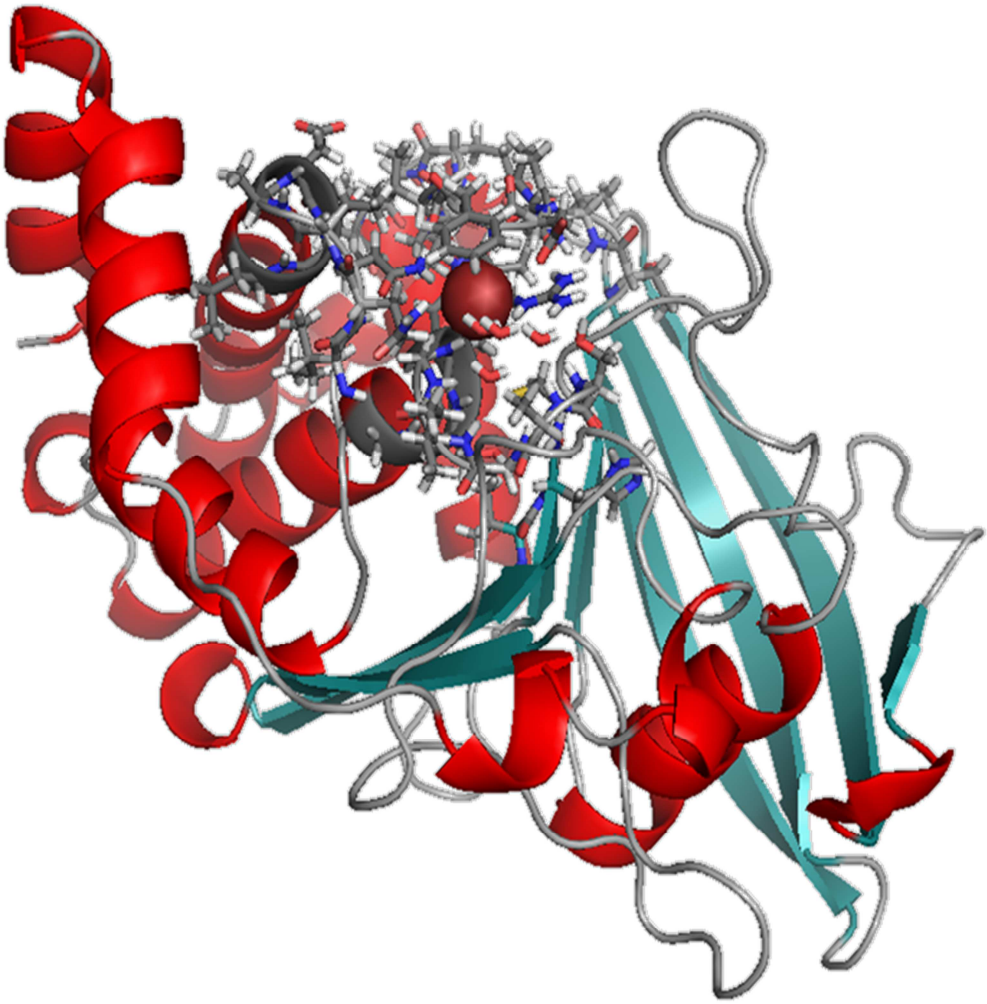
Study of the movement of the WPD flexible loop of human protein tyrosine phosphatase PTP1B and the factors that influence it



by

Chemist Aline Katz Wisel

A dissertation submitted in partial fulfillment of the requirements for Doctor's Degree in Chemistry

September, 2011

THESIS

Submitted to fulfill the requirements for achievement of the degree of

Doctor of the University of Strasbourg
and
Doctor of the Universidad de la República

Discipline: Life Sciences

Specific Field: Molecular and cellular Biology

Study of the movement of the WPD flexible loop of human protein tyrosine phosphatase PTP1B and the factors that influence it

Aline KATZ WISEL

Public PhD defense October 25th, 2011

Dr. Alberto D. PODJARNY	THESIS DIRECTOR
Dr. Oscar N. VENTURA	THESIS DIRECTOR
Dr. Sergio PANTANO	RAPPORTEUR
Dr. Juan BUSSI LASA	RAPPORTEUR
Dr. Eric WESTHOF	EXAMINER

Institute de Génétique et de Biologie Moléculaire et Cellulaire
Computational Chemistry and Biology Group

THESIS

Submitted to fulfill the requirements for achievement of the degree of

Doctor of the University of Strasbourg
and
Doctor of the Universidad de la República

Discipline: Life Sciences

Specific Field: Molecular and cellular Biology

Study of the movement of the WPD flexible loop of human protein tyrosine phosphatase PTP1B and the factors that influence it

Aline KATZ WISEL

Public PhD defense October 25th, 2011

Dr. Alberto D. PODJARNY	THESIS DIRECTOR
Dr. Oscar N. VENTURA	THESIS DIRECTOR
Dr. Sergio PANTANO	RAPPORTEUR
Dr. Juan BUSSI LASA	RAPPORTEUR
Dr. Eric WESTHOF	EXAMINER

Institute de Génétique et de Biologie Moléculaire et Cellulaire
Computational Chemistry and Biology Group

To my parents, for being who they are,

and helping me become who I am.

For raising me in a house full of love,

and always being proud of me,

no matter what.

To Seba, for standing strong by my side

during this roller-coaster, giving me the

love and support I needed,

never letting me fall.

La vida

Hubo una época en que todo
era más fácil.
Tu mamá te decía qué ropa
te ponías;
te peinaba;
te cuidaba.
Y cuando tenías hambre,
sólo llorabas.
Ibas a ser abogado,
o, tal vez, ingeniero.
Pero un día, sin que te dieras
cuenta, creciste.
Y aprendiste a decir que no.
No te conformaste.
Empezaste a tomar tus
propias decisiones.
Y sentiste que querías
cometer tus propios errores.
Entonces tomaste el camino
más difícil.
El que no estaba hecho.
Te dedicaste a lo que
realmente querías.
Te animaste a ser distinto.
Escuchaste a esa voz que te
salía de adentro.
Y, por primera vez, sentiste
que podías.
Era tu lucha.
Tu convicción.
Y, sin dudarlo, arriesgaste
todo lo que tenías.
Porque, en el fondo, sabías
que había algo mucho peor
que fracasar…

NO HABERLO INTENTADO
(Autor anónimo)

Life

There was a time when everything
was easier.
Your mom told you what
to wear;
she combed your hair;
she looked after you.
And whenever you were hungry,
all you had to do was cry.
You were going to become a lawyer,
or, maybe, an engineer.
But one day, without even noticing,
you grew up.
And you learned to say no.
You didn't resign yourself.
You started making your
own decisions.
And you felt like you wanted
to make your own mistakes.
Then, you took the
hardest path.
The one that wasn't already laid out.
You committed yourself to what you
really wanted.
You dared to be different.
You listened to that voice
inside of yourself.
And, for the first time,
you felt you could.
It was your struggle.
Your conviction.
And, without hesitation, you
risked everything you had.
Because deep down, you knew
there was something far worse
than failing…

NOT EVEN TRYING
(Anonymous author)

ACKNOWLEDGMENTS

I always imagined that when this time came, when I finally had to sit down and write all I've been doing for the past five years, it would be difficult… And, for once, I was right! The worst of all was the sight of that blank page, waiting for me to fill it up with ideas, experience and conclusions. Luckily, a lot of people came to the rescue: whether it was with ideas, or sharing their own experiences, or simply by showing me they were there, by my side, listening to me… just like they have been for the past five years.

So being finally here, giving this thesis the 'finishing touches' and, with that, closing an important chapter in my life, all the people who helped me along the way, making the ride so much more enjoyable, come to mind. And, since they were part of the experience, they should also be part of the result.

I want to begin this manuscript by remembering all of them, since all that comes after this, in one way or another, was possible because of them.

To all of you, just… thank you!!

- To Oscar… I will never forget the day he came up to our office (the old one), sat down and asked how would I like to go to France about a PhD… Who could say no to that?? And so, the story begun… So, thank you Oscar, not only for the opportunity, but also for fighting against the force fields and the programs with me, for your patience when the frustration brought out the worst in me. Most of all, thank you for your confidence in me.
- To Alberto, for taking the bet of working with me, even when he knew I had no background whatsoever in macromolecular crystallography. Thank you for encouraging me to read about it, for answering all my questions (as basic as I now know they were), for listening to my ideas and actually taking them into account. Thank you for helping me make the best of my Strasbourg experience, for the days off to go and see the world. Above all, thank you for taking a chance on me.

- To Patricia, Pati… hard to summarize all the things I have to thank her for! For always taking the time for me (even with the million other things she has to do), either for answering my questions, giving me her input, sitting in front of the computer with me, staring at a file, until the answer kind of pops out of the screen, or for reading my manuscripts. But most of all, thanks for your friendship, for those long talks, for those Ismael Serrano's recitals we shared, for listening to me when I was frustrated and saw everything black, and point out the silver lining. I probably wouldn't have gotten this far without you, so… thank you.
- To Alex, of course for the most obvious things: growing the crystals, sharing her knowledge on PTP1B with me, showing me the ropes around the IGBMC, explaining me everything in my first (and only) trip to a synchrotron. But also for the other stuff, those small things that really do make the difference: her tips on the city, going with me to Auchan and helping me to set up my very first home away from home, her invitations to lunch. Simply, thank you for helping me feel a little less lost in Strasbourg.
- To André, for the data collection, for teaching me how to mount the crystals in the synchrotron, for sharing his expertise and his intuition (which he called his sense of smell) about crystals with me, and for always having a smile for me, making me feel welcomed.
- To Marc, a great co-worker, always willing to help. Thank you for sharing your experience on NWChem and other programs with me, as well as the evening talks and radio shows.
- To Fiorentina, for all those classes we prepared together, for sharing the office and bearing with my radio. For our common struggle with force fields and programs.
- To Eduardo, thanks to whom I now have two monitors working. For his patience, for always being willing to lend a hand, for always taking the time to help me out, and, of course, for introducing me to the best desserts.
- To Gastón and Camila, the "newcomers", with whom is a joy to share the office, for their contagious enthusiasm.

- To Maitia, Mai… for those Sunday nights listening to "La Celeste" (I know that was in the pre-PhD era, but… still wanted to thank you for them!), and, especially, for those chat conversations where you helped me focus and get the writing going. Your tips worked like a charm!!
- To Denise, for sharing the Strasbourg experience with me. For those trips to IKEA, the visits to the museums, the nights out in town, the birthdays celebrated at your studio, the dinners together at mine, for teaching me songs as good as "Pasame la botella", and for bearing with my crazy ideas. Thank you for receiving me with open arms every time I went back to Strasbourg, and for keeping this friendship alive through the distance. I am really grateful to the teacher who paired us to work together, since we were both from Argentina…
- To Vero, my friend since… forever. It is hard to try and write here all the things I am thankful for. Thanks for always being there, for listening to my problems and always trying to come up with a solution… and understanding me when I get frustrated. Thanks for those study afternoons (some of them when you didn't even had to study!), for the trips to Buenos Aires, for all the concerts we shared, for always being close, even when I was 10000 km away. Thank you for being the great friend you are.
- To Ri, also my friend since forever. Thank you for always being around, for having my back whenever I needed it. Thanks for all those night and afternoons out, for all the talks over coffees, for always understanding what I am going through, and for being my friend no matter what.
- To Dri, sis, my sister from Brasil. For all those phone calls that always came at the exact time I needed her, as if she telepathically knew I needed her support. For being a part of my life and my family for over twelve years, and for always being there for me. For being the friend and sis she is.
- To Alejandra, for her constant friendship since that special year spent abroad, and, in particular, for those Thursday evenings dancing rikudim… They were critical in releasing some of the stress!
- To Pablo, for helping me out, listening to me and always supporting me. For offering me new opportunities and helping me grow as a scientist.
- To Pao and Emi, for those Friday nights out, were we blew out steam. For those chat sessions, for listening to me, and having so much faith in me.

- To Mirel and Virginia, for being there, and offering help in all kinds of ways, from a trip to the movie to clear my head, to a long-distance reiki session to help with my muscles' contractures.
- To Giu and Inés, for having my back all these years, and for their unbreakable faith in me, and in the fact that I was actually going to be able to finish this, even in those moments when I doubted it.
- To my parents, for… everything! For rising me in a house full of love, where I was encouraged to pursue whatever made me happy, giving me the tools to confront whatever came up. For always supporting me, always being on my side and by my side. For being the great role model they are, and for always being so proud of me, and showing it. For really trying to understand what is exactly what I do, as far away from their field as it is. I love you both, and I am proud of being your daughter.
- To Gabriel, Gabo, my brother, who walked by my side during this whole ordeal. We went from being brother and sister to being friends, and I could not be happier about being in each other's life. Thank you for your support, it was a constant source of strength.
- Last, but definitively not least, I want to thank Sebastian, Seba… For all the things we have lived together and for all the things to come. Thank you for being there, for your understanding, for giving up the desk so I had a place to write. For trying to make this time as easy as possible for me, so I could focus on writing. Thank you for understanding what this meant for me, and supporting me all the way through it. More than anything, thank you for your patience, your support, for looking after me and for your love. I love you with all my heart.

The last couple of months have been a great ordeal, definitively like a roller-coaster: I went from feeling down when I could not write what I wanted exactly as I wanted, to feel up in the clouds when I finally finished each chapter. So, I would like to thank all of the people who bore with me during this time… and I promise I'll go back to my usual self as soon as I get some sleep!!

Of course, I also want to thank all the institutions that made this thesis possible:

- Facultad de Química, for providing me the place and the tools to carry out this work.
- Universtié Louis Pasteur, for giving me the opportunity of doing this work in a joint supervision scheme.
- Institute de Génétique et de Biologie Moléculaire et Cellulaire, for receiving me with open arms and letting me do the experimental part of the work over there.
- PEDECIBA Química (Program for the Development of Basic Sciences – Chemistry) for the postgraduate studies scholarship and for subsidizing some of the time spent in Strasbourg, which was vital for this work.
- ANII (Research and Innovation National Agency), for the magister and PhD scholarships, and for the Fondo Clemente Estable (Clemente Estable's fund), which help me dedicate myself completely to this thesis work.
- CNRS (Centre National de la Recherche Scientifique) for their support.
- Institut National de la Santé et de la Recherche Médicale for their support.
- H.U.S (Hôpital Universitaire de Strasbourg).

Index

General index

Schemes index

Figures index

Tables index

Abbreviations and Glossary

ACE	N-terminus blocking group in the AMBER force field.
AMBER	Assisted Model Building with Energy Refinement.
BSSE	Basis Set Superposition Error.
cDNA	Complementary deoxyribonucleic acid.
CHARMM	Chemistry at HARvard Macromolecular Mechanics.
CM-Sephadex column	Weak cation exchanger column, with carboxymethyl as main functional group and sodium as counterion. It has a pH working range of 6-10.
Completeness	Number of crystallographic reflections measured in a data set, expressed as a percentage of the total number of reflections present at a specific resolution.
EDTA	Ethylenediaminetetraacetic acid.
Escherichia coli BL21(DE3)	Chemically competent E. coli cells suitable for transformation and protein expression.
Free R-factor (R_{free})	Measures how well the current atomic model predicts a subset of the measured reflection intensities, randomly chosen, not employed during refinement (the "test set"). It is computed in the same way as the R-factor, but using only the "test set".
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid.
I/σ(I)	Intensity divided by the standard deviation of a reflection, averaged for a group of reflections. Reports signal over noise.
IPTG	Isopropyl β-D-1-thiogalactopyranoside.

Lattice	A lattice in the vector space $\mathbf{V^n}$ is the set of all integral linear combinations $t = u_1a_1 + u_2a_2 + \cdots + u_ka_k$ of a system $(a_1, a_2, \ldots, a_k)$ of linearly independent vectors in $\mathbf{V^n}$.
LB medium	Lysogeny broth medium.
MM	Molecular Mechanics.
MD	Molecular Dynamics.
NaBr	Sodium bromide.
$NaBrCH_2CH_2SO_3$	Sodium bromoethanesulphonate.
NaI	Sodium iodide.
NME	C-terminus blocking group in the AMBER force field.
ONIOM	Our own N-layered Integrated molecular Orbital and molecular Mechanics.
PEG	Polyethylene glycol.
QM	Quantum Mechanics.
R-factor	Measure of agreement between the amplitudes of the structure factors calculated from a crystallographic model and those from original X-ray diffraction data.

$$R = \frac{\sum|F_{obs} - F_{calc}|}{\sum|F_{obs}|}$$

R_{symm}	A measure of agreement among the independent measurements of symmetry-related reflections in a crystallographic data set. Symmetry related reflections should have identical intensities. If they do not, it suggest some type of measurement error. $R_{symm}$is calculated as follows, with I and $\bar{I}$ representing intensities of two symmetry-related reflections: $$R_{symm} = \frac{\sum_A (I_A - \bar{I}_A)}{\sum_A \bar{I}_A}$$
Redundancy	Average number of independent measurements of each reflection in a crystallographic data set.
Resolution	Ability to distinguish between neighboring features in an electron density map. By convention, it is defined as the minimum plane spacing given by Bragg's law (which provides the condition for a plane wave to be diffracted by a family of lattice planes) for a particular set of X-ray diffraction intensities.
SDS-polyacrylamide	Sodium dodecyl sulfate polyacrylamide
Structure factor (F_{hkl})	Mathematical function describing the amplitude and phase of a wave diffracted from crystal lattice planes characterized by Miller indices h, k, l.
Tris-HCl	Tris(hydroxymethyl)aminomethane-hydrochloride

Resumen

El presente trabajo de tesis se enmarca dentro del estudio de la estructura de diferentes complejos entre una proteína de interés farmacológico y ligandos, y el análisis de las consecuencias que dichos modelos estructurales tienen sobre los mecanismos tanto de inhibición como catalítico.

En particular, este trabajo apunta hacia el refinamiento a resolución mediana de la enzima humana PTP1B y, más específicamente, al estudio estructural por cristalografía y modelización del movimiento del lazo flexible WPD de esta enzima frente a la presencia de diferentes iones halógenos.

La proteína de interés fue co-cristalizada en diferentes concentraciones de buffer acetato y de iones halógeno (Br^- y I^-). Estos cristales fueron medidos en fuentes de radiación sincrotrónica, y las estructuras tridimensionales de los complejos fueron determinadas. Estos datos experimentales permitieron comprobar que en los complejos formados por la proteína y el ion bromuro el lazo WPD fue encontrado tanto en la conformación cerrada como la abierta, mientras que en los complejos con el ion ioduro sólo se obtuvieron estructuras en las que el mencionado lazo se encontraba en conformación cerrada. Esto estaría indicando una relación entre la identidad del ion presente en el complejo y la fuerza de las interacciones establecidas entre éste y los residuos proteicos circundantes.

Los estudios cristalográficos también permitieron determinar una correlación entre la concentración de haluro presente en la mezcla en la que el cristal fue crecido y la conformación del lazo en estudio en dicho cristal. A mayor concentración del haluro, mayor es la ocupación tanto del ion como de la conformación cerrada del lazo. Con respecto a esto, se observó que cuando la solución presentaba concentraciones elevadas tanto de acetato como del haluro, el efecto del primero prevalecía, y la ocupación del haluro disminuía.

En busca de una correlación entre el tipo de ion presente en el complejo y la ocupación relativa de cada uno de los iones presentes, se llevó a cabo una serie de estudios teóricos, empleando diferentes técnicas y niveles de teoría.

Los movimientos colectivos de la enzima PTP1B en complejo con los dos haluros considerados fueron estudiados a través de una simulación de Dinámica Molecular, empleando el módulo SANDER del programa AMBER10.

La información obtenida respecto al movimiento del mencionado lazo en ambos complejos fue consistente con las observaciones experimentales obtenidas en este mismo trabajo: mientras que el complejo proteína-bromuro presenta un cambio conformacional a lo largo de la simulación, esto no se observó para el complejo proteína-ioduro. Esto apoya la teoría previamente planteada de la existencia de interacciones diferenciadas entre los dos haluros y los residuos circundantes.

Para lograr una mejor comprensión de las interacciones actuantes entre ambos iones y los residuos proteicos de la región catalítica, se están llevando a cabo optimizaciones de geometría para cuatro de los complejos (PTP1B-bromuro con el lazo en conformación cerrada, el mismo complejo con el lazo en la conformación abierta, PTP1B-ioduro en conformación abierta y el mismo complejo en conformación cerrada). Estos cálculos se están llevando a cabo empleando el modelo multicapa ONIOM del programa Gaussian09, el cual nos permite estudiar las partes más relevantes de la estructura a un nivel de funcionales de la densidad, empleando el funcional M06, y empleando la base DGDZV para representar los orbitales de Kohn-Sham de ambos iones, y la base 6-31G* para representar los residuos proteicos del sitio catalítico. El resto de la proteína se está estudiando a nivel de Mecánica Molecular, empleando el campo de fuerza AMBER para representar las interacciones existentes. Una etapa crítica para poder emplear métodos híbridos es la correcta determinación de la región a estudiar al nivel más elevado de teoría (DFT). En este trabajo se presenta la metodología empleada para lograrlo, y se deja constancia de los residuos que se consideran dentro de la región más relevante de la estructura.

El presente trabajo muestra que ciertas interacciones entre los residuos del sitio catalítico y el ligando presente en el complejo serían los responsables de la conformación adoptada por el lazo flexible WPD de la proteína PTP1B. Una vez se terminen los estudios híbridos y se logre determinar cuáles son los principales residuos catalíticos, se tendrá un punto de partida confiable para el diseño costo-efectivo de posibles inhibidores específicos para esta enzima.

Résumé

Le présent travail de thèse fait partie de l'étude de la structure de différents complexes entre une protéine d'intérêt pharmacologique et des ligands et l'analyse des conséquences que ces modèles structurels ont sur les mécanismes aussi bien d'inhibition que de catalyse.

En particulier, ce travail présente l'affinement à une résolution moyenne, de l'enzyme humaine PTP1B, et plus précisément, l'étude structurale par cristallographie et modélisation, du mouvement de la boucle flexible WPD de cette enzyme en présence de différents ions halogènes.

La protéine d'intérêt a été co-cristallisée avec différentes concentrations de solution tampon acétate et d'halogènes (Br^- et I^-). Ces cristaux ont été mesurés dans des sources de rayonnement synchrotron et les structures tridimensionnelles des complexes ont étés déterminées. Ces données expérimentales ont confirmé que dans les complexes formés par la protéine et l'ion bromure, la boucle WPD était aussi bien dans la conformation fermée qu'ouverte, tandis que dans les complexes avec l'ion iodure uniquement la conformation fermée a été observée. Cela indiquerait une relation entre la nature de l'ion présent dans le complexe et la force des interactions entre celui-ci et les résidus environnants de la protéine.

Les études cristallographiques ont également permis de déterminer une corrélation entre la concentration d'halogénure présent dans la solution dans laquelle le cristal a été obtenu et la conformation de la boucle. Plus la concentration de l'halogénure est élevée, plus l'occupation de l'ion est importante et par conséquent, la conformation de la boucle est fermée. Il a été observé, que lorsque la solution tampon contient à la fois des concentrations élevées d'acétate et d'halogénure, l'occupation de l'acétate augmente et l'occupation de l'halogénure diminue.

Pour vérifier la corrélation entre le type d'ions présent dans le complexe et l'occupation de chacun de ces ions, une série d'études théoriques a été menée, en utilisant différentes techniques de calculs.

Les mouvements d'ensemble de l'enzyme PTP1B en complexe avec les deux halogénures ont été étudiés par simulation de dynamique moléculaire, en utilisant le module SANDER du programme AMBER10. L'information obtenue par rapport au mouvement de cette boucle dans les deux complexes est en accord avec les observations expérimentales obtenues dans ce travail: le complexe protéine-bromure présente lors de la simulation un changement de conformation, qui n'a pas été observé pour le complexe protéine-iodure. Ceci confirme la théorie précédemment exposée concernant l'existence d'interactions différentes entre les deux halogénures et les résidus environnants de la protéine.

Pour parvenir à une meilleure compréhension des interactions qui existent entre les ions et les résidus de la protéine dans la région catalytique, une optimisation de la géométrie est en cours, pour quatre des complexes (PTP1B-bromure avec le boucle en conformation fermée, le même complex avec le boucle en conformation ouvert, PTP1B-iodure en conformation ouvert et le même complex avec le boucle en conformation fermée). Ces calculs sont effectués en utilisant le modèle multi-couche ONIOM du programme Gaussian09, qui permet d'étudier les parties les plus déterminants de la structure au niveau DFT (DensityFunctionalTheory), à l'aide de la fonctionnelle M06, et en utilisant la base DGDZV pour représenter les orbitales Kohn-Sham des ions et la base 6-31G* pour représenter les résidus du site catalytique de la protéine. Le reste de la protéine est étudiée au niveau de la Mécanique Moléculaire, en utilisant le champ de force AMBER pour représenter les interactions existantes. Une étape critique pour pouvoir utiliser les méthodes hybrides est la détermination précise de la région à étudier au niveau DFT. Ce travail présente la méthodologie utilisée pour atteindre cet objectif et l'identification des résidus les plus importants de cette région.

Le présent travail montre que certaines interactions entre les résidus du site catalytique et le ligand présent dans le complexe seraient responsables de la conformation adoptée par la boucle flexible WPD de la protéine PTP1B. Lorsque les études hybrides seront achevées et les principaux résidus catalytiques identifiés, nous aurons un point de départ pour la conception d'inhibiteurs adaptés à cette enzyme.

CHAPTER 1 INTRODUCTION

The current global situation regarding health and disease turns drug development into a continuously increasing need. The lack of effective therapies for several diseases, such as the AIDS virus, tuberculosis, malaria and other parasitic sicknesses, continue to take millions of lives a year around the world, turning this into a pressing matter [1].

Therapeutic drugs are substances that exert some kind of physiological or biochemical effect on the human body. They may be a single compound or a mixture of them, interacting with specific targets within the physiological environment, usually proteins but sometimes DNA or RNA. Drugs work either by stimulating or blocking the activity of their targets, thus having an incidence on the biological processes in which the latter participate [2].

Novel pharmaceutical discovery and development is a complex, expensive (it may cost as much as $880 million to discover a drug), and inefficient process, primarily due to the lack of models that accurately present the appropriate condition or that reflect the appropriate response. It is a high-risk-high-reward business, that requires a long-term vision, considerable technical and strategic experience, and multifaceted expertise [2 ,3].

For several thousand years, man has used herbs and potions as medicines, but it is since the mid-nineteenth century that serious efforts were made to isolate and purify the active principles of these remedies. Since then, the process of drug development has revolved mostly around a screening approach: a large variety of biologically active compounds have been obtained and their structures determined. These natural products became the lead compounds of a major synthetic effort, where chemists made literally thousands of analogues in an attempt to improve on what Nature had provided, aiming to obtain molecules with an increased activity, reduced side-effects, and an easier and more efficient administration to the patient. The vast majority of this work was carried out with no real design or reason, turning the process in a large scale adventure in serendipity.

This approach required a great amount of resources, both in manpower as well as material supplies, turning it into a non-profitable process [4].

The shortcomings of traditional drug discovery along with the development of sophisticated technologies and information generation platforms led to search for more deterministic approaches in the drug discovery process. These are known as 'rational drug design' (sometimes denominated 'mechanistic' or 'structure-based'), and they allow the rapid development of medicines with improved selectivity and safety profiles [5,6]. In its simplest formulation, the principle of structural complementarity is exploited to yield target-specific antagonists [1].

The advent of computers and their increasing computational power has had a positive impact on several scientific disciplines involved in the drug-discovery field, allowing for faster and cost-effective drug design. Among these disciplines, we can find macromolecular crystallography and Quantum Mechanics, which have shown to play critical roles in the designing of new pharmacologically important molecules.

Knowledge of accurate atomic structures of small molecules, obtained by X-ray diffraction experiments, has assisted the medicinal chemists in their endeavors to modify many of these molecules for the combat of disease. In addition to that, the three-dimensional structures of proteins provide insight into their physiological functions, facilitating the design of more efficient drugs and therapeutic agents. Actually, among all the technologies in drug discovery, macromolecular crystallography is one of the most powerful. Even though crystallography has been successfully used in the *de novo* design of drugs, its most important use has been, and will continue to be, in lead optimization: optimization of the affinity and specificity of compounds to their drug target [3,7].

Meanwhile, over the past three decades, Quantum Chemistry (the application of Quantum Mechanics to solve chemistry problems) has become an essential tool in the study of atoms and molecules and, increasingly, in modeling complex systems such as those arising in biology. The continued improvements in the theory and implementation, as well as the reduction in the cost/performance of computing, ensure that dramatic progress will continue to be made in the years to come. This discipline aims to understand the reasons and conditions (the why's and how's) in which different atoms and molecules interact with each other [8,9].

This thesis work is enclosed within the structural study of different complexes between enzymes of pharmacological interest and their corresponding inhibitors, and the analysis of the consequences that these structural models have both on the inhibition and catalytic mechanisms.

In particular, this work is focused on understanding the human enzyme protein tyrosine phosphatase 1B (PTP1B), a protein phosphatase that plays an important role in diseases such as type 2 diabetes mellitus (T2DM), obesity, and even certain types of cancer. An exhaustive understanding of this protein, as well as the structure, movement and flexibility of its WPD loop will entail the possibility of designing specific and potent inhibitors. This will mean an important advance not only in the therapeutic possibilities for the treatment of T2DM, but also a deeper understanding of the enzymatic mechanism of the PTPs in general, given the fact that some key structural features critical for catalysis are highly conserved within the PTP family [10,11]. Moreover, this will open the doors to the development, in the future, of specific therapeutic agents for the diverse affections derived from the inefficient or unsuitable operation of the enzymatic network carried out by the PTKs and the PTPs.

Even though the most effective approach for the high affinity inhibitor synthesis focuses on the enzyme's active site, the important conservation of this sequence among different PTPs generates the need to search for new strategies to pursue a greater specificity. In order to obtain it, it is vital to as reach a deep knowledge, both of the enzymatic structure, and the catalytic (and inhibitory) mechanism. In this aspect, it should be emphasized that the study of the PTP1B structures has suggested that the WPD loop contains variable residues, which could contribute to the substrate's specificity [10]. This, along with its inherent flexibility, which plays an important role in the enzyme's catalytic efficiency, induces to consider this loop when designing the synthesis of efficient inhibitors.

Since the active site of PTP1B binds negative ions, we have chosen halogen ions as probes to study the response of the active site to ligand binding. In particular, we focus on the closing of the WPD loop, since the understanding of the mechanism of this motion can lead to the design of new inhibitors.

Both structural and computational studies will be used in this thesis work, combining their potential, in the hope of achieving a better understanding of the interactions that govern

the WPD loop conformation and, through it, lay the groundwork for achieving an even better drug-development technique in the future.

REFERENCES

[1] Sun, E.and Cohen, F. E., *Gene 137*, **1993,** 127-132.

[2] *Drug development*; http://genome.wellcome.ac.uk/doc_WTD020915.html; September 19th.,2011

[3] Abdel-Meguid, S. S. Macromolecular crystallography in drug design, in *Macromolecular crystallography conventional and high-throughput methods*; Sanderson, M. R., Skelly, J. V., Eds.; Oxford University Press.

[4] Patrick, G. L. *An introduction to medicinal chemistry*; (Oxford University Press, 1995).

[5] Bureeva, S.,Andia-Pravdivy, J.and Kaplun, A., *Drug Discovery Today 10*, *22*, **2005,** 1535-1542.

[6] Grossman, M. R. B. J.-B. S. C.and Rouvray, D. H., *Math. Modell. 8*, **1987,** 571-582.

[7] Azarani, A.,Segelke, B. W.,Toppani, D.and Lekin, T., *JALA 11*, **2006,** 7-15.

[8] Friesner, R. A., *Proc. Natl. Acad. Sci.USA 102*, *19*, **2005,** 6648-6653.

[9] Levine, I. N. *Quantum chemistry, 5th. Ed.*; (Prentice Hall, 1999).

[10] Zhang, Z.-Y., *Curr. Opin. Chem. Biol. 5*, **2001,** 416-423.

[11] Burke, T. R.and Zhang, Z.-Y., *Biopolymers (Peptide Science) 47*, **1998,** 225-241.

CHAPTER 2 BACKGROUND

Throughout this chapter, the most relevant information available in literature regarding this thesis work will be presented and discussed, thus establishing the background for this project. There are three main aspects that need to be considered in order to gain the necessary insight into the problem at hand: a biological, an experimental and a theoretical one. Even though each of this facets is going to be dealt with in a separate chapter, they all have to be kept in mind while searching for answers for our problem, since, as it usually occurs in any scientific field, it is unrealistic to consider them as completely independent.

2.1- Biological background

2.1.1- Protein tyrosine phosphorylation

Protein tyrosine phosphorylation is a post-translational modification mechanism, conserved throughout the evolution, crucial for the regulation of a nearly all aspects of cell life, such as cell growth, induction of mitosis, differentiation, metabolism, cell cycle control, motility, cell adhesion, cell-cell communication, cell migration, transcriptional activation, neural transmission, ion channel activity, the immune response and apoptosis/survival signalling [1-4]. As an indication of the wide range of processes that are affected by tyrosine phosphorylation, 30% of intracellular proteins are subject to these phenomena [5]. It is a dynamic, reversible process, regulated by the balanced opposing activity of protein tyrosine kinases (PTKs), which catalyse the tyrosine phosphorylation, and protein tyrosine phosphatases (PTPs), responsible for catalysing the specific hydrolysis of the phosphate moiety of phosphotyrosine residues [6,7]. The importance of PTKs and PTPs in regulating cellular activities is highlighted by their strong presence in the genome: higher eukaryotes encode up to 2000 and 1000 genes

for protein kinase and phosphatase genes, respectively [5,8]. Hundreds of PTKs, PTPs and their associated substrates are integrated within elaborate signal transduction networks, whose defective or inappropriate operation can degenerate in such diverse and generalized diseases such as diabetes, cancers and immune dysfunctions [2]. Moreover, 51 of the 90 human protein kinases known, are implicated in cancer, whether it is by mutation, over- or under-expression [4]. The role of PTKs in promoting the signalling responses has been well documented, but the complexity in the structure, function and regulation of protein tyrosine phosphatases, as well as possible mechanisms by which both types of enzymes cooperate with each other to control cellular phosphotyrosine levels has only become object of study in the last decades [9]. The fundamental role that PTPs play in the regulation of diverse cellular mechanisms makes them potential targets for the development of new drugs.

2.1.2- Protein tyrosine phosphatase 1B and type 2 diabetes mellitus

One of the more thoroughly studied examples where is possible to clearly appreciate the fundamental role PTPs play in the organism is the case of the type 2 diabetes mellitus (T2DM). This type of non-insulin-dependent diabetes is present in 80-90% of the population affected with diabetes, and the excess of hepatic glucose production, as well as the liver, skeletal muscle and adipose tissue resistance to the insulin are the prime factors that contribute to it. Type 2 diabetes and obesity are often linked in human, and can lead to the metabolic syndrome, which dramatically increases the risk of cardiovascular incidences and decreases lifespan [10]. T2DM is associated to a deficit in protein tyrosine phosphorylation in the insulin signal transduction cascade that derives from the binding of insulin to the receptors located in the sensible tissues, deriving in a reduction of its metabolic effect and hyperglycaemia. This decrease in the phosphorylation does not seem to derive from an inherent problem in the kinase receptor of insulin (IRK), but from an increase in the activity of the protein tyrosine phosphatase [11]. Several PTPs have been implicated in modulating insulin signal transduction, but the protein tyrosine phosphatase 1B (PTP1B) seems to be a key regulator of the insulin-receptor activity, acting both at the receptor level and on

downstream signalling components [12]. In **Fig. 1** a schematic representation of PTP1B's incidence in both insulin's and leptin's pathways is shown.

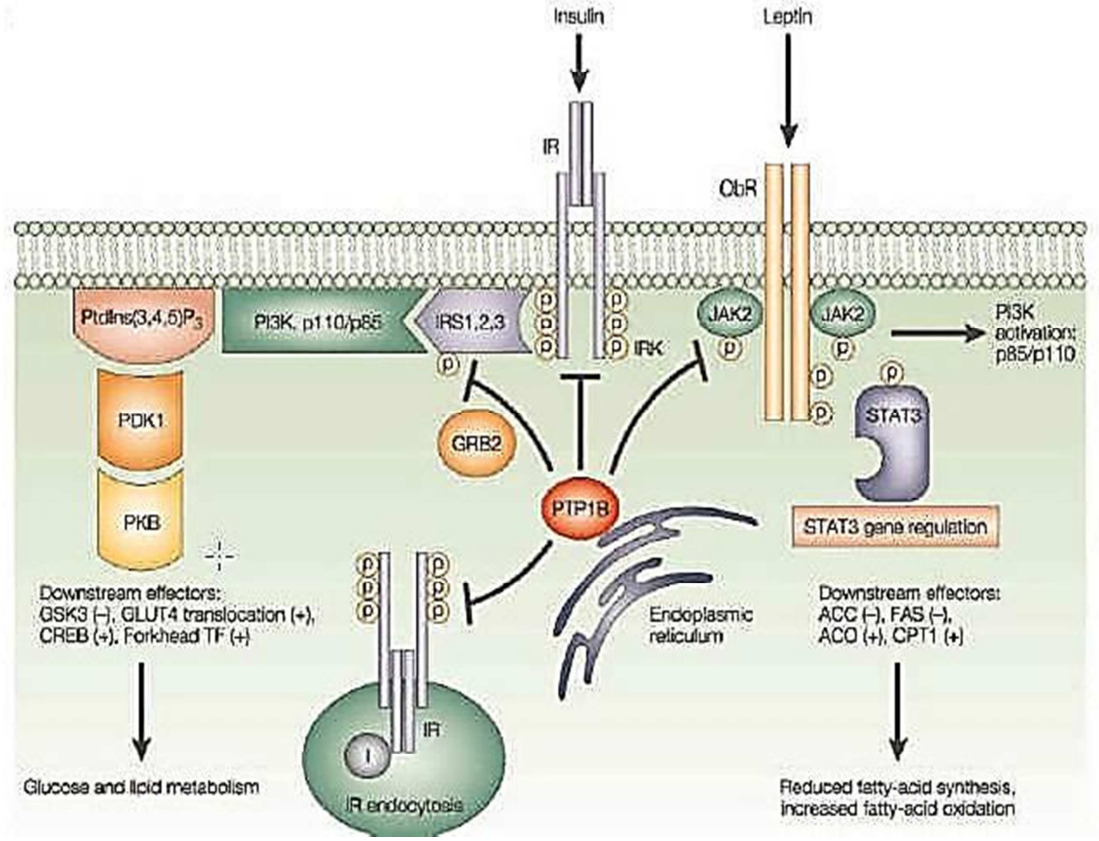


Fig. 1 - PTP1B's incidence in leptin's and insulin's signalling pathways [12]

The specific inhibition of enzyme PTP1B could be beneficial from a therapeutic point of view in the treatment of T2DM, insulin resistance and obesity [13]. This hypothesis is sustained by several experiments, which proved that mice lacking this enzyme show resistance to diabetes and do not develop diet-induced obesity [14,15]. Given that the estimate for the year 2015 is that 70% of the population of the western hemisphere will be overweight, 40% of which would be obese, turns this enzyme into a primary study target, since the social impact and costs associated with this health risk will be substantial, generating a strong demand for highly selective PTP1B inhibitors [10].

Aside from the central role this enzyme plays in both insulin- and leptin-stimulated signal transduction pathways [3], PTP1B is also involved in a series of clinically relevant metabolic routes, and it has been found overexpressed or overregulated in human breasts, colon and ovary cancers [10,16].

Despite all of this, obtaining clinically useful inhibitors has proven to be extremely difficult, due to the limited selectivity and low bioavailability of the proposed molecules [10].

2.1.3-Structure of the enzyme

PTPs constitute a large family of enzymes, present in all eukaryotic organisms (around 112 PTPs have been found to be codified in the human genome) [7], which displays great complexity and structural diversity. Genetic and biochemical studies have demonstrated that these enzymes exert both positive and negative effects on signalling pathways and play crucial physiological roles in a variety of mammalian tissues and cells [17]. Although they are structurally very diverse, it is possible to classify them into three major subfamilies: dual-specific, low molecular weight and tyrosine-specific phosphatases. The latter can be further divided into two groups, based on cellular localization: the receptor-like PTPs, which contain a trans-membrane domain, an extracellular receptor-like domain and two intracellular domains (generally only one of these domains accounts for the majority of the catalytic activity); and intracellular PTPs, which consist of a single catalytic domain with flanking regions, often containing novel protein-protein interaction domains that direct the enzyme towards specific locations within the cell [7,18].

Even though PTPs are very diverse in size (can be proteins of over 400 amino acids), sequence and structural organization, the catalytic domains are usually contained in a region that usually does not surpass the 250 residues, and represents the only structural elements that have sequence identity among all PTPs, from bacteria as in mammals [2]. That domain is characterized by a sequence of 11 amino acids (Ile/Val)-His-Cys-X_5-Arg (Ser/Thr)-Gly (where the "X" represents varying amino acids), denominated "signature motif" (also known as the "P-loop", since it constitutes the phosphate binding loop), containing the arginine and cysteine residues essential for displaying catalytic activity. The structural diversity within the PTP family lies in the different sequences presented by the NH_2- or COOH- termini of the catalytic domain [9].

PTP1B was the first PTP enzyme to be isolated in homogenous form from human placental tissue, and it is a cytoplasmic, nonreceptor PTP, bound to the endoplasmic reticulum (ER) by the presence of a hydrophobic sequence at its C-terminus. In its native form, it consists on 435 amino-acid residues, with an N-terminal catalytic phosphatase region (residues 1 to 300, among which the catalytically significant residues can be found), which shares an average of 40% of sequence identity with other members of the PTP family. The following 80 to 100 residues constitute the regulatory section and, finally, the 35 carboxy-terminal amino acidic residues (residues 400 to 435) are rich in Pro, and constitute the membrane localization region, responsible for binding the enzyme to the cytoplasmic face of the ER. However, for biochemical purposes, shorter versions of the protein (298 or 321 residues) are usually employed. The 321-residue version is composed of a single domain, organized in eight α helices and 12 β strands [5,12,19-21]. **Fig. 2** presents a cartoon representation of these secondary structure characteristics.

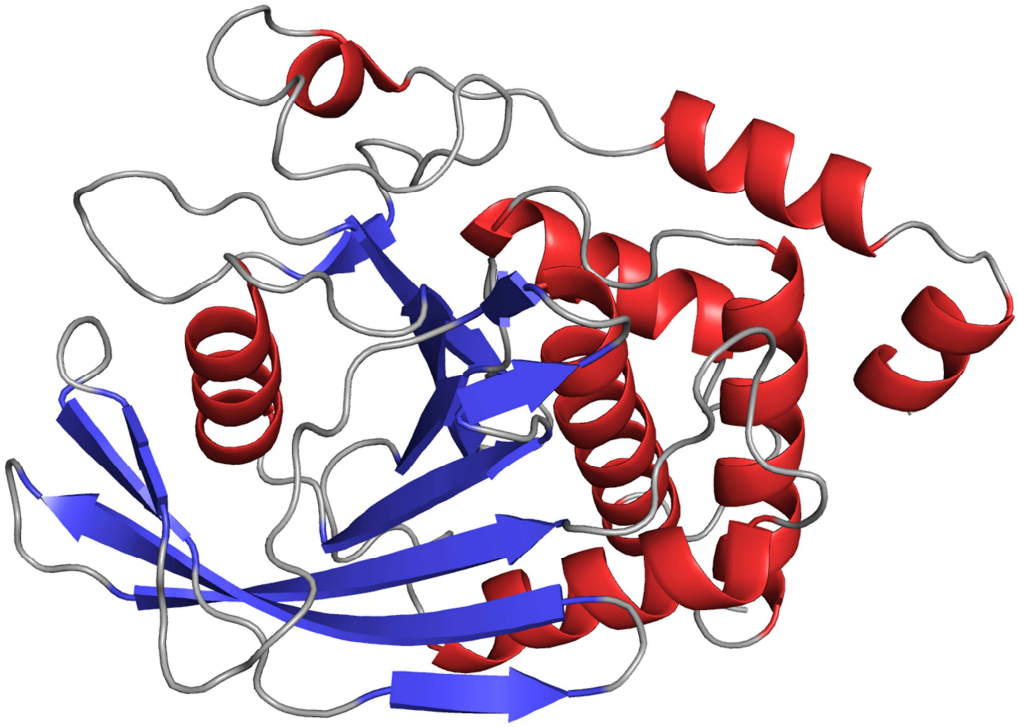


Fig. 2 – Cartoon representation of the 321-residue version of PTP1B.

The *active site* of the enzyme is located within a crevice on the molecular surface. It contains the common structural motif of PTPs, and the base of such cleft is defined by "signature motif", which includes the 214-221 residues (His-Cys-Ser-Ala-Gly-Ile-Gly-Arg), a loop of 8 amino-acid residues that forms a rigid, cradle-like structure that coordinates to the aryl-phosphate moiety of the substrate. This loop also contains the active-site nucleophile, Cys215. Four other loops bearing invariant residues form the sides of the catalytic cleft and contribute to catalysis and substrate recognition (Asp181, Phe182, Tyr46, Val49, Lys120 and Gln262). The depth of the catalytic cleft (which measures 9 Å from the Cys215 at the base to its entrance) provides substrate selectivity, as Ser and Thr (which are usually phosphorylated in human metabolism) are not long enough to reach the nucleophilic site of PTP1B. This enzyme undergoes important conformational changes during substrate-binding, associated to its catalytic activity. The *WPD loop* (thus denominated for including the invariant tripeptide Trp-Pro-Asp- residues 179-181 in PTP1B), which includes the residues 175-184, moves up to 12 Å to close down on the phenyl ring of the substrate, maximizing the hydrophobic interactions. The Asp181 residue moves into a position in which it can act as a general acid to protonate the tyrosyl leaving group, and the Arg221 residue also re-orientates into a position to optimize salt-bridge interactions with the phosphate group of the substrate. In this new conformation, the catalytic cysteine is in a position appropriate to undergo a nucleophilic attack on the substrate phosphorous atom. Another fundamental characteristic of enzyme PTP1B is the *secondary aryl-phosphate-binding site*, adjacent to the catalytic site. This site is catalytically inactive and provides much more weaker binding interactions than the primary site, due, fundamentally, to its more open exposure to the solvent [5,9,12]. In **Fig. 3** a close-up of the active site of PTP1B can be seen.

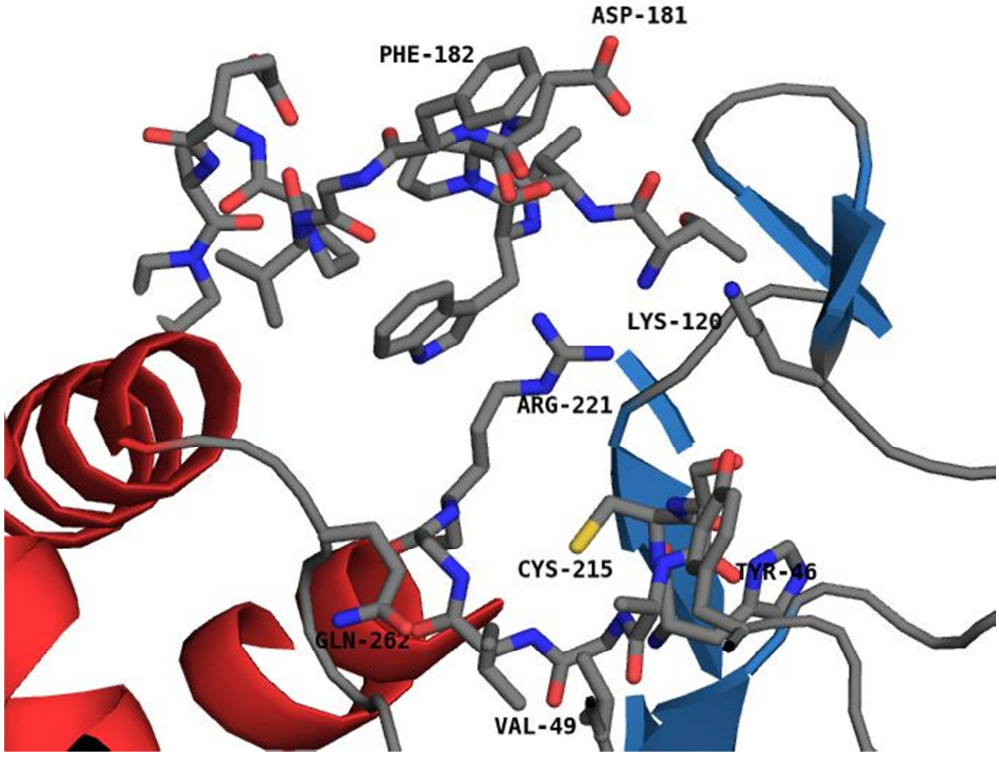
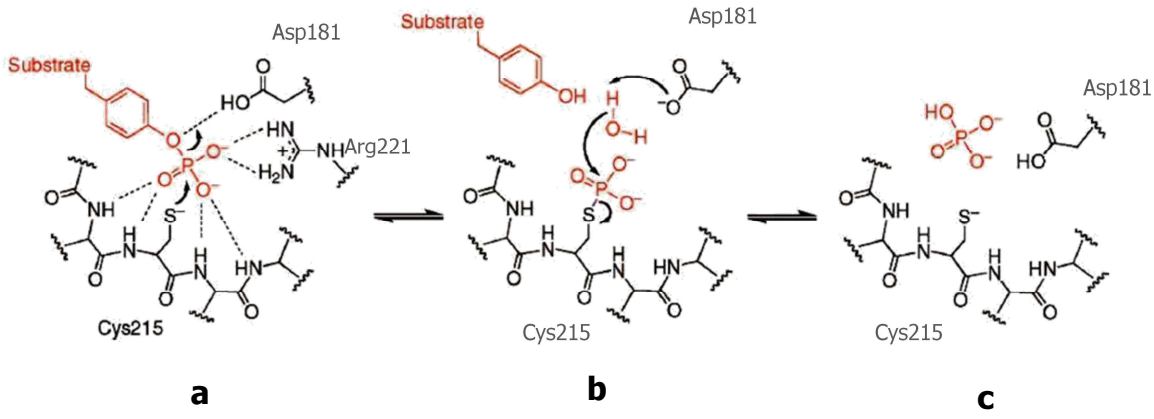


Fig. 3 – PTP1B's active site. The most important residues for catalysis are shown.

2.1.4- Mechanism of catalysis

All PTPs share a common catalytic mechanism to effect catalysis, which consists, basically, of two stages. In a first stage, the enzyme is bound to the di-anion of a phosphate-containing substrate, to form the catalytically competent enzyme-substrate complex. The highly conserved Asp residue on the flexible WPD loop is protonated at physiological pH, whereas the nucleophilic Cys is deprotonated under these conditions. The binding of the substrate to the enzyme induces an important conformational change in the latter, inducing the WPD loop to close on the phenyl ring of the phosphotyrosine, thus positioning the aspartic acid side chain to act as a general acid. In the resulting enzyme-substrate intermediate complex, the three non-bridging oxygen atoms of the phosphoryl group are coordinated by hydrogen bonds to the guanidinium group of the arginine, and by the backbone amide N-H groups of the active site loop (**Scheme 1a**). The nucleophilic cysteine is located underneath the phosphoryl group and at the base of the active site cleft, and it binds covalently the phosphate moiety.

The aspartic residue then acts as a general acid, breaking the bond between the phenol and the phosphate, and protonating the phenolic oxygen of the product (**Scheme 1b**). In the second step of the reaction, the phosphoenzyme intermediate undergoes hydrolysis, breaking the bond between the enzyme and the phosphate group (**Scheme 1c**). In this stage, the aspartic acid has, once again, a key role, acting this time as a general base by proton extraction from a water molecule, thus activating it. After hydrolysis, the WPD loop reopens, releasing the product. The dissociation of an inorganic phosphate from the enzyme completes the catalytic cycle [7,18,22] .



Scheme 1- Catalytic mechanism of PTP1B, figure by Seiner *et al [23]*

The above described mechanism suggests that the WPD loop has an inherent flexibility, and that it plays a fundamental role in the enzymatic catalysis. Usually, the loop is found in an 'open' conformation when no ligand is bound in the active site, and in a 'closed' conformation in substrate or inhibitor complexes. However, cases have been reported in which an open WPD loop was detected even after an inhibitor is bound to the enzyme [24-26], as well as structures where this loop acquires a closed conformation even in the absence of phosphotyrosine [19,27]. Recent studies have demonstrated that, even though the general structure of the protein is relatively rigid, it has localized regions of relatively high flexibility; among these regions is included the WPD loop. These works have demonstrated that, in addition to the expected open conformation, the catalytically important WPD loop of PTP1B can also attain the closed conformation without bound substrates or inhibitors. Nevertheless, it is important to take into account that in all the published cases there is a presence of chloride ion in the mix, due to the crystallization conditions for PTP1B [28]. In addition to the indirect

evidence provided in by several crystal structures, molecular-dynamics simulations confirm the flexibility within this loop in the apo state of the enzyme PTP1B. The movement of this loop would be associated to the movement of other close-by residues, such as the Lys120 loop and the Arg221 residue [27].

Even though the most effective approach for the high affinity inhibitor synthesis focuses on the enzyme's active site, the important conservation of this sequence among different PTPs generates the need to search for new strategies to pursue a greater specificity. In order to obtain it, it is vital to gain a deep knowledge, both of the enzymatic structure, and the catalytic (and inhibitory) mechanism. Along this lines, it should be emphasized that the study of the PTP1B structures has suggested that the WPD loop contains variable residues, which could contribute to the substrate's specificity [7]. This, coupled with its inherent flexibility, which plays an important role in the enzyme's catalytic efficiency, makes this loop an important factor to be considered when designing new and more efficient inhibitors.

2.1.5- Ions in biological macromolecules

The presence of ions in natural systems has proven to be of critical importance in biology, chemistry and life in general. Salting out, denaturation, regulation of the homeostasis and the electric potentials of cells are some of the significant aspects influenced by the presence of different ions in biological environments. Furthermore, monovalent ions stabilize proteins, lipids and nucleic acids through both specific and non-specific interactions [29,30]. Several physiological processes are affected by the presence of water molecules and ions, such as biomolecular stability, folding, dynamics, catalytic activity and ion transport through membranes [30,31]

In particular, halogen ions play key roles in biological systems: for starters, they contribute to the stabilization of DNA, by the presence of chlorine ions in the first hydration shell [32,33]. They are also present in important natural systems, such as the thyroid hormonal system (thyroid hormones are iodinated molecules in which halogen bonds play an important role in their recognition [34]). Another example of the incidence of halogen ions in biological systems is the effect both iodide and chloride have on the structure and stability of the human plasma protein transthyretin (TTR)

[35]. Moreover, halogenation is implicated in inflammatory responses, as in the case of chronic respiratory disease in children, which is correlated with the levels of chlorotyrosines in their system, and allergen-induced asthma, associated with bromotyrosines [36,37].

2.2- Experimental background

The three-dimensional arrangement of a biologically active molecule is deeply related to its physiological function. Having a deep knowledge of a protein's structural form, can lead not only to a better understanding of what role it plays in the organism, but also to the design of a better inhibitor or enhancement factor [38,39]. That is why an important goal of structural biology is obtaining a good model for the structure of proteins, as fast an accurate as possible [40], and several methods have been developed throughout the years, in order to achieve that goal. Among them, macromolecular crystallography (often referred to as protein crystallography) is one of the most widely used. This technique implies an X-ray beam striking on single crystals, and analyzing the resulting diffraction pattern. This method has progressed considerably in the last decades, since the elucidation of the first structures (myoglobin and haemoglobin, in the 1960's), when solving a protein structure with less than a hundred atoms was extremely difficult and took years, until nowadays, when the same process (which consists of the exact same steps: crystallization, data collection, phasing, model building and refinement, validation and presentation of the results) takes a few days [41,42]. A clear indication of the evolution of this field is the fact that the 23 structures included in the Protein Data Bank (PDB) in January 1976 [43] have turned into 75105 by August 2011 [44].

In a thumbnail, macromolecular crystallography requires a series of steps: firstly, it is necessary to grow high-quality crystals of the pure protein. Once the crystal is available, an X-ray beam is aimed at it, and the directions and intensities of the diffraction pattern are measured. With this information, and using computer software, the effects of an objective lens are simulated, thus producing an image of the crystal's contents. Finally, the crystallographer must interpret the obtained image, that is, build a molecular model consistent with the image previously generated [45].

Despite the important progress made in this field, solving the structure of a new macromolecule continues to be a long and demanding job, with the crystal growth acting as the 'bottleneck' of the global process, since technical and theoretical advances in this particular procedure have proceeded very slowly, and there is still a huge empirical component on it .

2.2.1- Protein crystals

In order to use X-ray diffraction to determine the three-dimensional structure of a given protein, it must be crystallized. X-ray scattering from a single molecule would be too weak to be detected above the noise level (which includes air and water scattering). Additionally, not any crystal is suitable to be used for such purpose, but it has to be reasonably large, and of high quality. A crystal, in general, is a three-dimensional array of building blocks (in the case or protein crystals, these building blocks would be the macromolecules themselves). Proteins, as many molecular substances, need particular circumstances to form crystals. When this occurs, individual molecules of the substance adopt one or a few identical orientations, resulting in an orderly three-dimensional arrangement of molecules, which are held in place by weak non-covalent protein-protein interactions (van der Waals and hydrogen bonds), and numerous water-mediated hydrogen bonds [45-48].

At first, protein crystals were obtained from bulk solutions (known as 'the batch mode'), but since then many different techniques to crystalize them have been developed, but some general considerations stayed the same. For example, biocrystallization (just like any other crystallization) is a process which depends on many parameters, but it still has to go through the three basic steps: nucleation, growth and cessation of growth. The differences between protein and small molecule crystals arise, in the first place, in the number of parameters involved in each case (much larger in the first), and, secondly, in the fact that the higher complexity shown by proteins implies that their optimal stability in aqueous phase is restricted to a narrow range of both temperature an pH. There is a third difference, that makes crystallization of proteins a fine art, and that is the higher sensitivity of macromolecules to external conditions, due to their conformational flexibility and chemical versatility [42,49].

In a protein crystal, the solid phase (the actual crystal array) coexists with a high solvent content (30 – 75%, or even more), and this strongly influences the behavior of the crystal. This has a series of implications in the characteristics of the obtained crystal, such as the fact that crystal proteins are much less ordered than classical crystalline arrays; it also provides a higher motility to the surface groups in the crystal, due to their contact with the solvent. While these are clear disadvantages when considering its use for X-ray diffraction, the same high percentage of solvent can result beneficial, since it makes the environment of the macromolecule in the crystal really similar to the one in which it was obtained (it is necessary to take into account the influence of the solvent in the protein conformation), and also it allows for the solvent to be used in the preparation of derivatives that can result helpful in order to solve the structure [41].

Protein crystallization remains the mayor obstacle in the determination of a protein's structure, and that's why some important principles must be taken into account, in order to assure the growth of good, high-quality crystals, useful for X-ray diffraction.

2.2.1.1- Principles of protein crystallization

As it has been already stated, the process of crystallization is very complex, involving a lot of factors, and, up to this day, very poorly understood. Luckily, a lot of experience on the crystallization of water-soluble proteins has been acquired over the years, consequently providing any crystallographer beginning with a considerable amount of pure protein a good chance of obtaining X-ray quality crystals. Protein crystals are almost always grown in solution, involving a phase transition in which the macromolecules eventually come out of the solution to form crystals. Starting from a supersaturated solution of the protein, the solubility of these molecules is gradually lowered until a thermodynamically stable partition of the macromolecules between the solid phase and the liquid is achieved. This lowering of the solubility is done very gradually; otherwise, if it is done too rapidly, the molecules will precipitate from solution, forming useless dust or amorphous gel in the bottom of the container. Following the formation of the nuclei, the concentration of the protein gradually decreases, driving the system into a metastable state where growth occurs without the

formation of further nuclei. It is important to control the process of nuclei formation, since an excess of nuclei (resulting from a very high supersaturation) would lead to the accumulation of structural defects, due to the lack of space for an adequate growth of crystals. High supersaturation also translates in a rapid incorporation of impurities, leading to a crystal not apt for X-ray diffraction [41,50,51].

The solubility of proteins in water depends on a variety of properties, such as temperature, pH and the presence of other solution components, as well as the amino acid composition. The solution becomes supersaturated when it is brought above its solubility limit, and that can be achieved by the addition of precipitating agents and/or modification of some of the internal parameters of the solution. Given the liable character of proteins, extreme conditions of precipitation, pH and temperature should be avoided. Moreover, the crystal growth takes place in conditions close to the physiological environment at which the protein operates, the more information regarding its biological and physiological properties can be obtained [41,52].

There are four big mechanisms to control the precipitation of proteins in a supersaturated solution:

◊ proteins are zwitterionic (different parts of the molecule may be either positively or negatively charged), which means that the solubility of these molecules is highly dependent on the solution's pH, with a minimum at the isoelectric point;

◊ variation of the salt concentration in the solution, since ionic strength has an opposite effect on the protein's (or any other biomolecule's) solubility, i.e. solubility decreases exponentially with increasing ionic strength (the **salting out** phenomenon), but it also presents a minimum at very low ionic strength (the **salting in effect**). In practical crystallization, precipitation can be achieved either by increasing the salt concentration (sodium chloride, ammonium sulfate, sodium sulfate, potassium chloride, ammonium chloride, magnesium sulfate, calcium chloride, ammonium, lithium chloride, citrate and acetate are some of the most commonly used salts);

◊ the addition of organic solvents (such as ethanol, isopropanol, acetone, dioxane, MPD) can also be used to control the protein's precipitation, due to the double effect they have by subtracting water molecules and decreasing the dielectric constant of the medium;

◊ the addition of polymers, in particular PEG, which is available in different molecular weights (ranging from about 200 to 20000 Da), presents a similar effect on the proteins solubility to the organic solvents and salts: it restructures the solvent, thus promoting the phase separation.

Of course, these are not the only factors that can affect the crystallization process: protein concentration, temperature, presence of cations (they can stabilize a given conformation of the protein) and the purity of the sample, since the presence of contaminants can prevent the formation of X-ray suitable crystals [41,51].

The most commonly used methods for non-membrane protein crystallization are vapor diffusion and batch crystallization. Methods based on vapor diffusion have produced the largest number of proteins through the years than all the other methods combined. Two variants can be found of the vapor diffusion method: the **hanging drop** and the **sitting drop** methods. In both cases, the method consists of an aqueous drop containing the purified protein, buffer and the crystallization agents, mixed in a smaller concentration than that required for the formation of crystals. This mixture is placed in the vicinity of a larger reservoir containing similar buffers and precipitants, but in higher concentration. As the equilibrium between both of the systems progress, the water from the drop slowly vaporizes and transfers to the concentrated solution of precipitant (salt or PEG) in the reservoir, due to the different osmotic pressures. This diffusion increases the precipitant concentration in the drop to an optimal level for crystallization. Since the system is in equilibrium, these optimum conditions are maintained until the crystallization is complete. The best crystallization conditions are usually identified by a 'trial and error' previous step, in which variable ratios of solutions of protein, precipitating agents and additives are pipetted together, conducting a self-screening process. Basically, the difference between the hanging drop and the sitting drop methods differ in the position of the protein solution with respect to the reservoir: in the first one, the drop is placed on a glass cover slide, which covers the reservoir, while in the last one the drop is sitting on a pedestal, near to the reservoir. It is important to mention that in both cases a closed system is required, so it must be sealed off from the outside [50]. **Fig. 4** and **5** show a representation of the hanging drop and sitting drop, respectively.

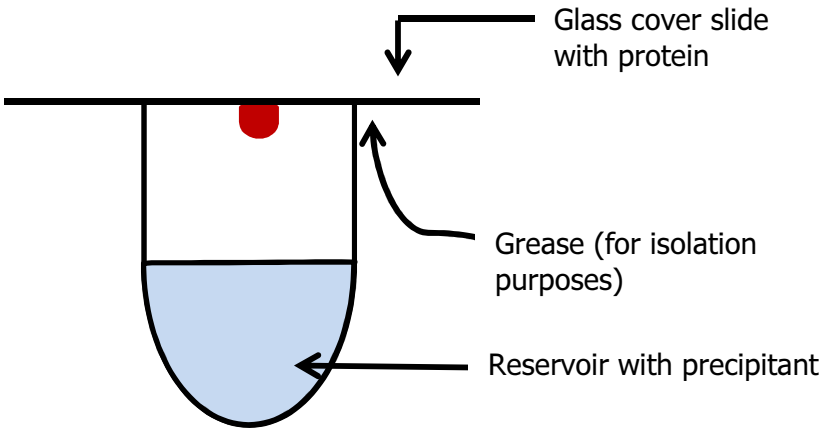


Fig. 4 – Diagram of hanging drop method

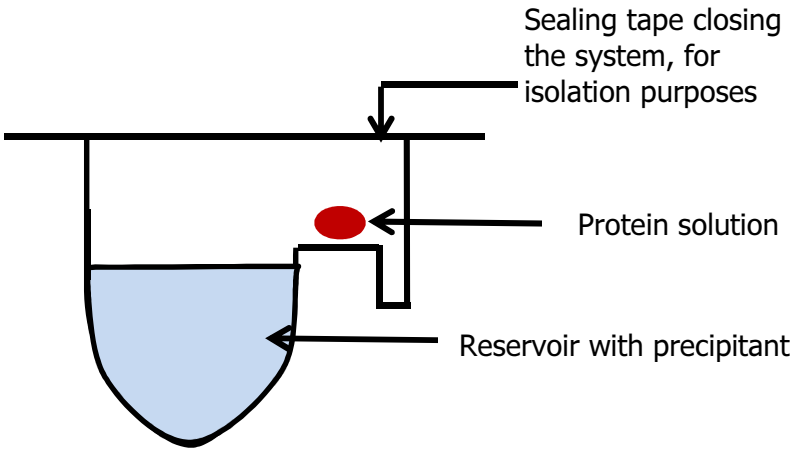


Fig. 5 – Diagram of sitting drop method

The other technique frequently used is the batch method, in which the protein and the crystallizing agents are mixed at the beginning of the experiment in their final concentrations. It is the simplest and oldest crystallization method. The basic difference with the vapour diffusion methods is that these consist of dynamic systems, in which conditions are varying during the experiment, while in the batch technique the conditions are constant within the normal time of a crystallization experiment.

2.2.1.2- Crystallization of protein-ligand complexes

X-ray analysis progressively became an important technique in the analysis of protein-ligand complexes. The fact that macromolecular crystallography can provide a direct view of the whole complex is a great advantage, since no other biophysical technique can give such a straightforward information. Once the X-ray diffraction data is analyzed, and the Fourier transform is applied to it, the electron-density map of the complete complex is obtained. There are different ways to calculate these maps, and one type of maps, commonly referred to as 'difference maps' focuses on the difference between the structure of interest (for example, a protein-ligand complex) and the considered model (for example, the structure of the same protein, alone). In such a way, not only can the structure of the ligand be determined, but also its binding site, changes in the solvent structure and any conformational changes in the protein that results from the ligand binding [53].

There are two methods employed by crystallographers to obtain the crystals required for studying protein-ligand interactions: co-crystallization of the protein and the ligand, and soaking a preformed protein crystal in a mother-liquor containing the ligand. From these two approaches, **co-crystallization** is the most widely employed, and it implies that the complex between the protein and the ligand is obtained previous to the crystallization experiments. Depending on the solubility of the ligand of interest, there are two different protocols that can be followed. The first of them, known as the 'fast co-crystallization protocol' requires a ligand with a solubility in the protein's concentration range, either in the protein buffer or the crystallization solution. When the ligand has a very poor solubility, it is more appropriate to follow a 'slow co-crystallization protocol'. A second means to obtain the protein-ligand complexes is employing the **soaking approach**. This approach is preferred when well diffracting crystals of the unbound protein are available, or if the crystallographer is interested in comparing the structures of the pure protein with the one in the complex. [45,53].

An important fact to keep in mind is the fact that, in a similar way to crystallization, binding events are highly dependent on the assay conditions. Since the experiment conditions for X-ray diffraction are optimized with crystal growth in mind, they may not be optimal for ligand binding. In some cases, the binding can be even stronger in the

experiment's conditions, but it could also go the other way. The same goes for both pH and salt conditions, which play critical roles in the two events. Another thing to bear in mind is that the binding of any ligand stabilizes the protein, thus facilitating the crystallization. For all of the above is that it is worth the while to check the binding of the ligand with that particular protein under the optimized crystallization conditions, to know how the latter affects it [53].

2.2.2- The solution of the phase problem: Molecular Replacement

Once crystal is exposed to an intense beam of X-rays, it scatters the original emission into several discrete and determined waves, generating a pattern of spots or *reflections*, which can be seen in an appropriate detector. The greater the intensity of the X-ray reaching a given position, the more intense the reflection produced, generating a darker spot in the detector. The diffracted waves are affected by a number of factors (temperature, absorption, etc.), but, from a structural point of view, the most important element that influences it is the geometry of the atomic array within the crystal. Each reflection can be assigned three coordinates or *indices* in the imaginary three-dimensional space of the diffraction pattern, called the *reciprocal space*. The position of a given reflection in this reciprocal space is identified by three integer numbers (h, k and l, known as 'the Miller indices'). The diffracted wave in a given direction (h,k,l) is defined by its amplitude ($|\boldsymbol{F_{hkl}}|$, also known as the 'structure factor'), and a phase ($\boldsymbol{\theta_{hkl}}$). The relationship between the electron-density in a given point of space, $\rho(xyz)$, and the diffracted wave in the (h,k,l) direction is given by the Fourier transform depicted in **Eq.1**:

$$\rho(xyz) = V^{-1} \sum \sum \sum |F_{hkl}| e^{-2\pi i(hx+ky+lz-\theta_{hkl})}$$

(Eq. 1)

(V is the volume of the unit cell).

In order to reconstruct the electron-density of the protein, the phase and the amplitude of the diffraction must be retrieved. In practice, what can be measured in the X-ray diffraction patter are the hkl position, and the intensity ($\mathbf{I_{hkl}}$) of each reflection. The

latter is proportional to the square of $\mathbf{F_{hkl}}$, a factor that involves two terms: one of them is the atom factor, and the other one is dependent on the geometric distances [45,53,54].

Given the direct relation between the structure factor and the diffracted intensities, the amplitudes of the different diffracted waves can be derived directly from the experiment, simply by applying the Fourier transform previously presented. The phase cannot be directly obtained from the experiment's measures, and that is the central problem of crystallography, commonly known as '*the phase problem* '. There are three major techniques used to solve this problem, but it is important to keep in mind that any of them only produce estimates for the phases, that have to be improved afterwards. The classical technique for solving this problem is the **isomorphous replacement method**, which was used to solve the phase problem for the first protein structures. It consists in the preparation of a minimum of two compounds similar in form (isomorphous, hence the name) to the original material, but that contain one or many heavy atoms. The amplitudes of all diffraction spectra are compared, and the phases calculated by vector subtraction. In order to have a useful heavy-atom derivative, it is important that the addition of heavy atom does not disturb either the crystal packing or the protein's conformation. Another important aspect for this method to result useful is that there have to be measurable changes in at least a modest number of reflection intensities, since these changes are what make the phase estimation possible [41,45,55,56].

A second method to solve the phase problem, and that is also based in the use of derivatives (or, in case there are any, of the specific atoms naturally present in the biomolecule), is **anomalous scattering**. Electrons bound in atomic orbitals have specific resonant frequencies, which correspond to the allowed transitions. When the wavelength of the source X-ray beam matches up these frequencies, the atom, instead of scattering the incident beam, absorbs it. This absorption of the incident beam breaks down Friedel's law (which states that the intensities of the hkl reflections are equal to the $\bar{h}\bar{k}\bar{l}$ ones), and the inequality of these related reflections is called an *anomalous scattering*. These resonant frequencies in light atoms, such as carbon, oxygen or nitrogen, are far below the energies used for the diffraction experiments, rendering their anomalous scattering negligible, but that is not the case for heavier atoms, where

the absorption range is close to the energies employed in these experiments. The difference in the hkl and $\bar{h}\bar{k}\bar{l}$ intensities can be used to establish the phase of the scattered waves. The methodology and underlying principles in this method do not differ much from the ones in isomorphous replacement, with two slight differences: in the case of anomalous scattering, the heavy atoms may be an integral part of the macromolecule as well as incorporated by diffusion or chemical means into the crystals. The advantage of this phasing method lies in that all the data needed for phase determination is collected from a single crystal, as long as varying the X-ray beam wavelength is available. Optimal use of anomalous scattering requires the possibility to change the wavelength of the incident beam to have photons of specific energy according to the absorption edges. It can be done with multiple wavelengths (MAD method) or with a single wavelength plus density modifications (SAD method), and has become the method of choice for solving new structures [45,46,57,58].

The observed fact that proteins with homologous sequences present similar folding of their polypeptide chain provides a third and easier way of overcoming the phase problem. Since phases of atomic structure factors (and, consequently, of molecular structure factors) depend on the location of the atoms in the unit cell, the availability of a known structure (referred to as *phasing model)* allows the placement of this model in the unit cell of the new protein, and use it as a basis for the calculation of the initial phases. This method is known as **molecular replacement**, and it can be also applied if two structures (one known and the other unknown) are expected to have similar three-dimensional structures for whatever reason. The structure of the model must be superimposed on the new protein's structure (find in the new unit cell the position and orientation of the model that completely overlaps the target protein), and then proceed to calculate the phases for the newly oriented model. When the model is placed correctly, its phases are comparable to the ones of the target protein. Placement of the molecule in the target unit cell involves two aspects: rotation (in order to obtain the correct orientation of the molecule) and translation (to place the model in the correct position). Since the number of possible orientations and positions is colossal, it is practically impossible to conduct a full search, covering every possible combination of position and location. In order to simplify it, the procedure is separated into two steps: rotation, aiming to find the best orientation, and translation, which provides with the

best position possible, each one of which has its own function and methodology [41,45,59].

The rapid progress of the structural genomics field, aided in no small extent to the augmented computer power available, has increased significantly the number of structures deposited in the different structural databases, thus increasing the possibilities of solving structures employing the molecular replacement methods. In 2006, half of the structures deposited in the PDB had been solved using this methodology, a proportion that by 2009 had grown to around two-thirds. Given the high deposit rate for new structures, and the constantly evolution of the field, it is very likely that these numbers will continue to grow over the next few years [42,60].

2.2.3- The interpretation of the electron-density

Albeit the ultimate objective of an X-ray diffraction experiment is to obtain the atomic coordinates of the studied biomolecule, when the X-ray beam is directed to a crystal, its actual diffractors are the clouds of electrons in the macromolecule. Thus, when **Eq. 1** is solved, what is obtained are not the x, y, z coordinates for the atoms, but the distribution of the electronic cloud in space, that is, the *electron density* of the macromolecule. Obviously, since the electrons are distributed around the atom nucleus and the different bonds formed, the shape of this density is a reflection of the protein's structure. Because the protein is crystallized, that is, the molecules are placed in an ordered array, the electron density in a crystal can be described mathematically by a periodic function, more specifically, a Fourier sum. In this way, once a phase angle is estimated for the protein structure factors, the calculation of the electron density is quite straightforward. It is important to obtain the best map possible, since both the quality and the accuracy of the final structural depend heavily in the interpretation of that map [41,45,61].

The initial interpretation of the map is usually not very easy, since usually there are no good phases available, and so, the usual strategy implies an iterative process, where a first electron map is calculated, at low resolution (5-6 Å). Such a map, even though does not provide a lot of information at molecular level, does permit to differentiate the contours of the molecule, distinguishing solvent regions from protein. Even more, some

elements of secondary structure can sometimes be observed at this level. Once the localization of the protein in the unit cell is established, a medium-resolution map is calculated (3.5-2.5 Å), and the trace of the polypeptide chain is established. Some prior knowledge about the primary structure and protein chemistry can be really helpful at this stage, making it easier or sometimes even possible the chain trace. Even though a lot of mistakes can be made at this stage, it is important to start the construction of an atomic model that provides a reasonable 'fit' for the map, allowing the calculation of higher resolution phases (around 2 Å), and thus correcting those inaccuracies in the model. In this way, the first complete model of the three dimensional structure for the macromolecule is obtained [41,62].

2.2.4- Refinement

Given the limited resolution and imperfect phase information available, building and refining a macromolecular structure is not an exact science. It is a rather subjective process, since the choices made by the crystallographer (such as what program to use, whether or not to include alternative conformations, which peaks of the map to interpret) deeply influence the resulting model. Once the first complete macromolecular model is obtained, it is necessary to improve it, through cycles of map calculation and model building, attempting to improve, in each step, the agreement of the model with the experimental data available. This process, known as **refinement**, constitutes, essentially, a function minimization problem, since the idea is to minimize the difference between the information calculated from the current model with the information obtained from the X-ray experiment. The great number of atoms included in a protein turns this into a large problem, since the parameters to optimize are numerous. The fact that the resolution of the diffraction data obtained in most cases is not sufficient to obtain atomic-level information, compels to use energy or stereochemical restraints in order to refine single atoms, thus increasing the number of observations available [45,63-65].

The definition of this procedure at a basic level is the optimization of a function of a set of observations by changing the parameters of a model. The *observations* are all the information regarding the crystal obtained prior to the refinement stage, such as unit-

cell parameters, structure factor amplitudes, standardized stereochemistry and experimentally determined phase information, as well as the primary structure of the macromolecule, and the mean electron density of the mother liquor [63].

A key feature in model refinement is the parameterization of the atomic model. In the PDB format (the general format for molecular models), the molecule is viewed as a collection of atoms, each defined by three parameters: spatial coordinates (position of the atoms); atomic displacement parameters, which portray the movement of the atoms from their average position (the **B factor**, which can be defined as *isotropic*, when vibrations take place equally in all directions, or *anisotropoic*, when the vibration is described within an ellipsoid centered at the atomic coordinate); and an occupancy factor, which measures the fraction of molecules in the crystal in which a given atom actually occupies the position specified [66]. Given this description of the model there are three positional parameters, one or six parameters to describe the B factor (whether it is isotropic or anisotropic, correspondingly), and one for the occupancy. In any case, this implies a huge number of factors to optimize, thus the difficulty of the task. In order to simplify the problem, restraints may be used to provide additional information (imposing a penalty when the model parameters deviate from the ideal stereochemistry, or when there are deviations from observed structure-factor amplitudes), or constraints can be applied to reduce the number of parameters. It is important to keep in mind that even though the number of parameters increase with the size of the molecule, so does the unit cell, and, with it, the number of reflections at a given resolution. This means that the ratio of observations to parameters depends essentially only on resolution. Another fact to consider is that, unless the available diffraction data is extremely high, refinement of a model containing anisotropic B factors results in physically unreasonable structures, and this is why refinement is usually performed considering isotropic B factors. Basically, there are four methods to reduce the number of parameters to optimized: in the rigid-body parameterization the new parameters are obtained simply by rotation and translating a previously known coordinate system (in a very similar way to what was done in Molecular Replacement), and the orientation and location parameters are the ones subjected to optimization. This method is used when the model consists of a molecule whose structure is already known, but with unknown location and position in the crystal. In the NCS-constrained

parameterization there is a single set of atomic coordinates for each type of molecule, and an orientation and location for each copy, and the latter are optimized separately from the atomic positions. This methodology is useful when the asymmetric unit of the crystal contains multiple copies of the same type of molecule, and the diffraction data is not enough as to define the differences between the copies. In the torsion-angle parameterization the atomic coordinates are replaced by torsion angles, thus decreasing the number of parameters to optimize. This method proves to be useful when the diffraction data presents a low resolution, below 3 Å. Finally, the Transition-Libration-Screw (TLS) Bfactor parameterization improves the fit between the model and experimental data, in particular because it allows the description of anisotropic motion with fewer parameters. In this case, the motion of a group of atoms is described by three matrices, one for a translational vibration, another one for libration of the group about a fixed point, and a third one for a translation and a libration occurring in a concertated manner. There is an explicit assumption here that the group of atoms move as a rigid unit [60,63,64].

Having chosen the parameters to consider for the refinement, the function to use has to be determined. The empirical energy function has been used since the early 1970s, and it is quite intuitive: it considers that the best model of a macromolecule is the one with the lowest energy. The problems with this function lie in the fact that up until now it has been impossible to formulate an empirical energy function accurate enough as to reproduce experimental results and that there is no statistical theory sustaining this function. Another option is the least squares residual function, which is the simplest statistical method employed in macromolecular refinement. This approach has also been applied since the 1970s, and it is still widely used. This function is depicted in **Eq 2**. :

$$\mathrm{f(p)} = \sum_{\mathrm{i}}^{\text{all data}} \frac{\left[\mathrm{Q_0(i)\text{-}Q_c(i,p)}\right]^2}{\sigma_0(\mathrm{i})^2}$$

(Eq. 2)

In this equation, $Q_0(i)$ and $\sigma_0(i)$ are the value and standard deviation for observation number i, and $Q_c(i,\boldsymbol{p})$ is the model's prediction for observation i using the set of parameters $\boldsymbol{p}$. The smaller the difference between the actual observation and the

model's prediction, the better the model, and the parameters of the model are varied in order to minimize the difference. The problem with this function is that is based on two assumptions (that the errors in the observation follow a normal distribution with known variances, and that with perfect observations and the best parameters, the predictions obtained with the model would fit the observations perfectly) which have recently been proven incorrect in many refinement problems, for example if the model is incomplete. It is necessary that the refinement function accounts for the unknown contribution of the unmodeled part, and that is something that least squares cannot do [63,67]. Finally, there is maximum likelihood, which is a generalization of least squares, and formalizes the idea that the quality of a model must be judged by the accuracy of its consistency with observations, which means that, would the model be correct, the probability of making an observation with that value is reasonably high. The basic maximum-likelihood residual is shown in **Eq. 3** :

$$f(\mathbf{p}) = \sum_{i}^{\text{all data}} \frac{[Q_0(i) - \langle Q_c(i,\mathbf{p})\rangle]^2}{[\sigma_0(i)^2 + \sigma_c(i,\boldsymbol{p})^2]}$$

(Eq. 3)

The function, as well as the symbols, in this equation is pretty much the same as in **Eq. 2**, but with a few differences: for instance, the quantity subtracted from $Q_0(i)$ is not the actual model's prediction but its expectation value, calculated using the all the plausible models similar to $\boldsymbol{p}$. The new parameter $\sigma_c(i,\boldsymbol{p})$ is the width of the distribution of values for $Q_c(i,\boldsymbol{p})$ over the probable values of $\boldsymbol{p}$. When applying maximum likelihood, given a model, its error and the measurement errors, the probability of making a measurement must be calculated. Compared to least square refinement, this system can obtain over two times the improvement on phase error, leading to clearer electron-density maps, with less bias [63,68].

Having determined the molecular model to consider, and the function to be employed, the process of improving the model and the corresponding map implies a series cycles of computed, reciprocal space refinement followed by map fitting, or real-space refinement, cycles. Usually, the map fitting is a manual procedure, carried out with the

aid of visualization programs (such as COOT [69], among others), but some software packages have been programmed to allow an automated map fitting, alternating with the reciprocal space refinement. As the refinement proceeds, the difference map becomes emptier, allowing placing the crystallization waters. Finally, the only signals in that map are those of problem areas, thus pointing to the errors in the model. When the process is considered to be almost finished (the model includes protein molecules, as well as crystallization water and any other molecules that may come from the mother liquor), a new type of map is computed, the $2F_o - F_c$ (with F_o being the observed structure factors, while F_c are those calculated from the model). If the refinement has been successful, this map should look almost as a space-filling model of the protein. Then it's time to evaluate the resulting model, and one of the most widely employed statistics are the **R-** and **R_{free}-factors**. The value of the R-factor is a measure of the agreement between the amplitudes of the structure factors calculated from the crystallographic model, and those from the original X-ray diffraction data. The problem with this factor is that it's not very sensitive to errors, particularly in the case of low resolution structures with low observations-to-parameters ratios, overall good structures with a few localized errors, or structures with distorted geometries. The R_{free}-value is based on the statistical method of cross-validation, and it measures the degree to which the atomic model predicts a subset of the observed diffraction data (usually, 5-10% of the dataset) that was not used during the refinement process. This value presents a significant reduction in coupling to the target function used during the minimization that the R-factor, and it tends to be 2 to 8% higher than the regular R-factor. As a general rule, structures with a resolution of 2.5 Å or better with R-factors below 25% (or the corresponding value for R-free factor) can be considered to be correct [45,63,65,70].

2.3- Theoretical background

Since the 1970's, there has been a growing application of the results of theoretical chemistry incorporated into efficient computer programs, applied to chemical and biological problems that would be otherwise intractable or inaccessible. This practice has been called the *computational molecular sciences,* and has rapidly become one of the backbones of modern academic research. This discipline is grounded in the techniques of computer-based modeling and analysis, and it has connections to theoretical chemistry, biology and materials sciences, as well as chemical biology, systems biology and biophysics, playing an important role in studies such as protein folding and characterization of biochemical pathways, among others. The key aspect of this area of study is that it allows the application of theoretical studies to large sets of molecules, thus being an important part of, for example, the drug design process [71-73].

The theoretical chemistry aspects involved in this discipline are wide-ranging, and, in this section, the basis for all of them are covered, in order to have a better understanding on the studies carried out in this thesis work.

Chemistry can be defined as the science that deals with the construction, transformation and properties of molecules, which are "composed" of atoms. The behavior of such small particles is not accurately described by classical mechanics; it is necessary to resort to a new set of laws, known as *quantum mechanics*, in order to understand their interactions. The application of this set of laws to the solution of chemical problems is what is known as quantum chemistry [74,75].

The information on the electronic structure of molecules, and therefore, all the information on how they are constructed, how they interact with each other, their spatial arrangement and their properties are all dependent on their electronic structure, which, in turn, lies on one of the basic principles of quantum mechanics: the **Schrödinger equation**. Any problem in the electronic structure of matter is covered by the time-dependent form of that equation but, for most cases, the problems include atoms and molecules without time-dependent interaction, so the simplified time-independent equation can be used. For an isolated M-nuclei and N-electron molecular

system, the time-independent Schrödinger equation for a given state i has the following form:

$$\widehat{H}\Psi_i(r_{1,}r_2,\dots,r_N,R_1,R_2,\dots,R_M) = E_i\Psi_i(r_1,r_2,\dots,r_N,R_1,R_2,\dots,R_M)$$

(Eq. 4)

E_i and Ψ_i are the total energy and the wave function of that state, respectively, and $\widehat{H}$ is the *Hamiltonian operator* for the studied system, a differential operator that represents the total energy of the system.

$$\widehat{H} = -\frac{\hbar^2}{2}\sum_{\alpha}\frac{1}{m_\alpha}\nabla_\alpha^2 - \frac{\hbar^2}{2m_e}\sum_i\nabla_i^2 + \sum_\alpha\sum_{\beta>\alpha}\frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} - \sum_\alpha\sum_i\frac{Z_\alpha}{r_{\alpha i}} + \sum_i\sum_{j>i}\frac{1}{r_{ij}}$$

(Eq. 5)

In **Eq. 5**, α and β represent the nuclei present in the system, and *i* and *j* represent the electrons. Since **Eq. 5** represents the total energy of the system, the different potential and kinetic energies are included: the first and second term represent the electron and nuclei kinetic energy, correspondingly (with m_e representing the electron mass, while m_α represents the mass of nucleus α); the third term on the equation represents the potential energy due to the repulsion between nuclei (with $R_{\alpha\beta}$ the distance between nuclei α and β, with atomic number Z_α and Z_β respectively; the fourth term corresponds to the potential energy of the electrostatic attraction between nucleus α and electron *i,* and the last term corresponds to the electrostatic repulsion between two electrons *i* and *j.* $\hbar$ corresponds to Planck's constant h divided by 2π, and ∇ is the *Laplace operator*, also denominated *Nabla squared* [74,75].

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\frac{\partial^2}{\partial x^2}$$

(Eq. 6)

The wave function Ψ is also known as state function, and it contains all the information there is about that particular state of the system (it depends not only on how the system is composed, but also on where each particular component is located [75]).

The Schrödinger equation has an exact solution when the system consists of a monoelectronic atom (such as the hydrogen atom). However, when considering 'real' chemical problems, which involve polielectronic and polinuclear molecules, the resulting differential equation has not been analytically solved, up to this date. This is equally true for electronic and nuclear-motion problems. This is why it has proven essential to develop and implement mathematical methods that provide approximate solutions to Schrödinger's equation. In this context, two methods are commonly applied: the *variational method*, and the *perturbation theory*. The former ones, in particular the linear variational method, are the most widely used approximation techniques in quantum chemistry, minimizing the energy by optimizing some parameters within the wave function [75,76].

There are four big approaches to the calculation of the molecular properties of a given system:

◊ ***ab initio* calculations**, also called of **first principles**, employ the correct Hamiltonian for each system, with no experimental data included, other than the fundamental physical constants. The name of this method implies a calculation "from the beginning", but this does not imply that the results obtained can be considered as "100% accurate", since, in order to be able to perform them, it is necessary to include some approximations. The biggest drawback of this method is the huge amount of calculation needed to solve the Schrödinger equations, requiring a large amount of time and resources (such as computational time) in order to acquire the needed information;

◊ **semiempirical methods** use a simpler Hamiltonian operator than the one actually corresponding to the molecular system at study, and they include certain constants or *parameters* that are derived either from experimental data, or previous *ab initio* calculations. Since there is no need to calculate everything from the very beginning, these are less time-consuming calculations, and they also require fewer resources. On the downside, the results obtained are too dependent on the parameters included, so a poorly chosen parameter set will derive in non-reliable results;

◊ **density functional methods** have a completely different approach on the subject: they do not attempt to determine the molecular wave function, but they calculate the molecular electron density, **ρ**, and derive the molecular electronic energy from it;

◊ **Molecular Mechanics (MM)** is not, in fact, a mechanoquantic approach; it does not consider neither a Hamiltonian nor a molecular wave function. Instead, this methodology visualizes the whole system as a collection of atoms attached to each other by bonds, and expresses the molecular energy in terms of different parameters (such as strength constants for bond stretching and bending, among others) [75].

The characteristic of the system studied in this thesis work, and the different aspects that have to be taken into account to solve the proposed problem, three of the previously presented methods were used to approach it: *ab initio* calculations, density functional methods, and MM. The basic fundamentals for the three of them are, then, introduced in the following pages.

2.3.1- The Born Oppenheimer and the fix nuclei approximations

The molecular Hamiltonian for a given molecule, as described in **Eq. 5**, is extremely complex, making it almost impossible to work with it. Because of it, some accurate approximations need to be considered in order to solve Schrödinger's equation. These approximations are based on the fact that nuclei are a lot heavier than electrons ($m_\alpha \gg m_e$), leading to two big consequences: on the one hand, the movement of nuclei and electrons can be separated (known as the *Born-Oppenheimer approximation,* and is essential in quantum chemistry). The second consequence is that, since electrons are so much smaller than nuclei, the difference in the velocity at which these two particles move is such that it can be considered that the nuclei are fixed while the electrons are moving about them. Classically speaking, it could be said that during a cycle of the electron's motion, the change in nuclear configuration is practically null. Having this into account, the nuclear kinetic terms can be omitted from the Hamiltonian, consequently obtaining the equation for the electronic movement:

$$(\widehat{H}_{el} + V_{NN})\Psi_{el} = U\Psi_{el}$$

(Eq. 7)

$\widehat{H}_{el}$ represents the *purely electronic Hamiltonian*, and has the form presented in **Eq. 8**:

$$\widehat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e^2}{r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e^2}{r_{ij}}$$

(Eq. 8)

V_{NN} represents the nuclear repulsion, and is given by the formula expressed in **Eq. 9**:

$$V_{NN} = \sum_\alpha \sum_{\beta>\alpha} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}}$$

(Eq. 9)

The term U in **Eq. 7** represents the electronic energy including internuclear repulsion. The internuclear distances $R_{\alpha\beta}$ are not variable; they are fixed at a constant value. The electronic wave function and energy depend parametrically on the nuclear configuration. In **Eq. 7**, the variables are the electronic coordinates. V_{NN} does not depend on the position of the electrons, and it is constant for a given nuclear configuration. The omission of a constant term on the Hamiltonian does not affect the wave functions, and the only effect this has is the diminishing of the energy in the same amount. That can be stated as follows:

$$U = E_{el} + V_{NN}$$

(Eq. 10)

Having all of this into consideration, the internuclear repulsion in Schrödinger's electronic equation can be omitted, thus finding the electronic energy (E_{el}) for a given nuclear configuration as:

$$\widehat{H}_{el}\Psi_{el} = E_{el}\Psi_{el}$$

(Eq. 11)

The next step is to solve the electronic Schrödinger's equation, and determine the energy of the system on a given state [74,75]. In order to do that, two different approaches can be used: either determining the wave function of the system and then finding the correct value for the electronic energy (employing the *Hartree-Fock method*), or considering the energy as a function of the electronic density of the molecule (*Density Functional Theory*).

2.3.2- The Hartree-Fock method

Once the limitations have been stated (the Born-Oppenheimer and the fix nuclei approximations), it is now evident that the potential hypersurface and, therefore, the movement of the nuclei are mainly determined by the solution of the electronic Schrödinger equation for a fixed nuclear configuration. As it has been already stated, this equation can be solved exactly for any one-electron system, but not for more complex ones. In the general case, it is necessary to rely on approximate methods to solve the Schrödinger equation for a multi-electron system.

2.3.2.1- The variational principle

To generate approximate solutions to the Schrödinger equation the *variational principle* is employed. This principle states that any approximate wave function has an energy above or equal to the exact energy. In other words, given an approximate function ø depending on the system's coordinates and spin variables, normalized and well behaved, the inequality presented in **Eq. 12** applies:

$$E_\phi = \frac{\int \phi^* \hat{H} \phi}{\int \phi^* \phi} \geq E_\Psi = \frac{\int \Psi_0^* \hat{H} \Psi_0}{\int \Psi_0^* \Psi_0}$$

(Eq. 12)

The variational principle, then, allows us to establish an upper limit to the energy of the system's fundamental state (Ψ_0). The approximate functions ϕ are known as *variational test function*, the first half of **Eq. 12**, is denominated the *variational integral*, and the equality is only met when $\phi = \Psi_0$. The whole idea behind the variational principle is to employ different test functions, in order to minimize the variational integral. The lowest the value of E_ϕ, the more the test function considered resembles the actual wave function of the system [74,75].

It is important to keep in mind that, in ordered to be considered well behaved, the system's wave function must satisfy certain conditions (and, of course, so do the considered test functions): it must be continuous, monovaluated and quadratically integrable [75].

2.3.2.2- Slater determinants

One of the oldest methods, which is also the basis of other, more refined theories, is the *Hartree-Fock self-consistent field* method (**HF-SCF**). This method gives an approximate solution to the electronic Schrödinger equation using the electronic Hamiltonian ($\widehat{H}_{el}$) as obtained from the Born-Oppenheimer approximation [77].

The HF method is a non-relativistic approach, in which the single electrons are described by single particle functions $\phi_n(x)$, consisting of a product of a spatial orbital $\psi(r)$, depending on the position of the electron, and a spin orbital $\alpha(\omega)$ or $\beta(\omega)$, depending only on the spin coordinate:

$$\phi_n(x) = \phi_n(r,\omega) = \psi(r).\begin{cases}\alpha(\omega)\\ \beta(\omega)\end{cases}$$

(Eq. 13)

The total electronic wave function must be antisymmetric (change sign) with respect to exchange of any two electron coordinates (since electrons are fermions, having a spin of $^1/_2$). The antisymmetry of the wave function can be achieved by building it from *Slater determinants* (SD), where the columns are single-electron wave functions

(*orbitals*), while the rows represent the electron coordinates. If the system at study is a molecule, the one-electron functions are *molecular orbitals* (MO), given as the product of a spatial orbital and a spin function (α or β), also known as *spin-orbitals*, which may be taken as orthonormal. For the general case of N electrons and N spin-orbitals, the corresponding Slater determinant is given in **Eq. 14**:

$$\Phi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1.) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{vmatrix}$$

(Eq. 14)

One last approximation is made, by taking the trial ∅ function to consist of a single Slater orbital. This is quite a drastic approximation, and has, as will be seen later, an important effect, since it implies that electron correlation is neglected, or, equivalently, that the electron-electron repulsion is only included as an average effect [74,77].

2.3.2.3- The energy of a Slater determinant

The HF method seeks the single particle functions $\phi_n(x)$ which minimize the variational integral (the first half of **Eq. 12**). To perform this minimization, it is convenient to split the electronic Hamiltonian $\widehat{H}_{el}$ presented in **Eq. 8** into a sum of single electron operators $\hat{\mathrm{h}}(r_i)$, which affect only the ith electron and the non-separable interaction potential $\widehat{V}_{el-el}$ between all electrons:

$$\widehat{H}_{el} = \sum_i \hat{\mathrm{h}}(r_i) + \frac{1}{2}\sum_{i \neq j} \frac{e^2}{|r_i - r_j|}$$

(Eq. 15)

where

$$\hat{\mathrm{h}}(r_i) = \frac{p_i^2}{2m_e} + \hat{V}_{el-el}$$

(Eq. 16)

Each one of the single particle functions considered are assumed to be normalized and orthonormal.

Once the variational optimization has been performed, the *Hartree-Fock (HF) equations* for the single particle spin orbitals are obtained:

$$\hat{F}\phi_n(x) = \epsilon_n \phi_n(x)$$

(Eq. 17)

with the Fock operator $\hat{F}$, and ϵ_i representing the orbital energy. The Fock operator consists of the single particle operator $\hat{\mathrm{h}}$ and the so-called *Coulomb* and *exchange operators* $\hat{J}$ and $\hat{K}$, respectively:

$$\hat{F}\phi_i(x) = \hat{\mathrm{h}}\phi_i(x) + e^2 \sum_j^{N_{el}} \int \frac{\phi_j^*(x')\phi_j(x')}{|r-r'|} dr' \phi_i(x) - e^2 \sum_j^{N_{el}} \int \frac{\phi_j^*(x')\phi_j(x')}{|r-r'|} \phi_i(x') dr'$$

$$= \hat{J}\phi_i(x) \qquad = \hat{K}\phi_i(x)$$

$$= \epsilon_i \phi_i(x)$$

(Eq. 18)

This form can be interpreted as a single particle operator consisting of the basic operator $\hat{\mathrm{h}}$ for one electron with an additional effective Hartree-Fock potential term

$$V^{HF} = e^2 \sum_{i}^{N_{el}} (\hat{J}_i(x) - \hat{K}_i(x))$$

(Eq. 19)

which can be written in the simple form

$$\underbrace{(\hat{\mathrm{h}} + V^{HF}(x))}_{\hat{F}_i}\phi_i(x) = \epsilon_i \phi_i(x)$$

(Eq. 20)

The Coulomb part of this potential describes the interaction of one electron with the charge of all other electrons located in all the other single particle orbitals, hence it is similar to the Coulomb interaction in classical mechanics. The exchange part of the potential is a purely quantum mechanical effect (it has no classical counterpart), and it is caused by the requirement for the wave function to be anti-symmetric for the exchange of two given electrons [74,75,77].

The energy of the HF molecular orbital is then calculated as:

$$E_{HF} = \sum_{i} h_{ii} + \frac{1}{2}\sum_{i,j}(J_{ij} - K_{ij})$$

(Eq. 21)

with $h_{ii} = \int \phi_i^* \hat{\mathrm{h}} \phi_i$, and the Coulomb and correlation energies derived from the corresponding operators as

$$J_{ij} = \int \frac{\phi_j^*(x')\phi_j(x')\phi_i^*(x)\phi_i(x)}{|r - r'|} dr' dr$$

(Eq. 22)

and

$$K_{ij} = \int \frac{\phi_j^*(x')\phi_i(x')\phi_i^*(x)\phi_j(x)}{|r - r'|} dr'dr$$

(Eq. 23)

2.3.2.4- The Self-Consistent Field method

It is important to remember that the real Hamiltonian and wave function for the system involve the coordinates for the n electrons, while the Fock operator is a one-electron operator (it only includes the coordinates for one electron). Since the latter depends on its own eigenfunctions, which are not known, an iterative process must be employed to solve the Hartree-Fock equations. It is necessary to start with an initial guess for the $\phi_i(x)$, use them to calculate an approximate $V^{HF}(x)$ and, from this, the Fock operator $\hat{F}$, recalculate the $\phi_i(x)$ with this operator, and repeat the whole procedure until the single particle functions converge to a stable solution [75,77].

2.3.3- The basis set approximation

For small highly symmetric systems, the Hartree-Fock equations may be solved by mapping the orbitals on a set of grid points, and these are referred to as *numerical Hartree-Fock* methods. Nonetheless, all calculations use a basis set expansion to further simplify the problem, thus reducing the numerical effort carried out. This approximation was proposed by Roothan in 1951, and implies the expression of the unknown molecular orbitals (MO) in terms of a set of known functions. This is not an approximation if the basis set is complete, but this is not possible in actual calculations (due to the high computational costs), so finite basis sets are employed. This means that only the components of the MO along those coordinate axes corresponding to the selected basis functions can be represented. Both the size of the basis set and the type of functions consider influence the quality of the representation obtained. There are two basic guidelines for choosing the basis functions: one is that they should have a

behavior that agrees with the physics of the problem, to ensure that the convergence as more basis functions are added is reasonably rapid. For molecular systems, this means that the functions should go toward zero as the distance between the nucleus and the electron becomes larger. The second guideline is of a practical nature: the chosen functions should make it easy to calculate all the required integrals [74,75,77].

2.3.3.1- The Roothan-Hall equations

Each molecular orbital ϕ is expanded in terms of the basis functions φ, conventionally called atomic orbitals (AO) (although they are not solutions to the atomic HF problem), thus giving this approach its name: *linear combination of atomic orbitals (LCAO)*.

$$\phi_i = \sum_{\alpha}^{M} c_{i\alpha}\varphi_\alpha$$

(Eq. 24)

As has been stated above, the basis functions φ_α should form a complete series, which implies an infinite number of functions. This is not possible in practice, and a finite number M of basis functions is considered. Taking Roothan's expansion into account, the Hartree-Fock equations can be written as

$$\sum_{\alpha}^{M_{basis}} c_{i\alpha}\hat{F}\varphi_\alpha = \epsilon_i \sum_{\alpha}^{M_{basis}} c_{i\alpha}\varphi_\alpha$$

(Eq. 25)

Multiplying from the left by a specific basis function and integrating yields the *Roothaan-Hall equations*. These are the Hartree-Fock equations in the atomic orbital basis, and all the M_{basis} equations may be collected in a matrix notation:

$$FC = SC\varepsilon$$

(Eq. 26)

where

$$F_{\alpha\beta} = \int \varphi_{\alpha}^{*} \hat{F} \varphi_{\beta}$$

(Eq. 27)

and

$$S_{\alpha\beta} = \int \varphi_{\alpha}^{*} \varphi_{\beta}$$

(Eq. 28)

The S matrix contains the overlap elements between basis functions, and the F matrix contains the Fock matrix elements. Each $F_{\alpha\beta}$ element contains two parts from the Fock operator, integrals involving the one-electron operators, and a sum over occupied MOs of coefficients multiplied with two-electron integrals involving the electron-electron repulsion [74].

The Roothaan-Hall equation is a determination of the eigenvalues of the Fock matrix. To determine the unknown MO coefficients $c_{i\alpha}$ the Fock matrix must be diagonalized, but this matrix is only known if the MO coefficients are known. So, the procedure starts off by some guess of the coefficients, forms the F matrix and diagonalizes it. The new set of coefficients is then used to calculate a new Fock matrix, and so on. The procedure continues until the set of coefficients used for constructing the Fock matrix is equal to those resulting from the diagonalization (within a certain threshold). This set of coefficients determines a self-consistent field solution [74].

The Fock matrix and, therefore, the total energy, only depends on the occupied MOs. Solving the Roothaan-Hall equation produces a total of M_{basis} MOs, with N occupied and $M_{basis} - N$ unoccupied, or *virtual* MOs. The virtual orbitals are orthogonal to all the occupied orbitals, but have no physical interpretation [74].

2.3.3.2- Orbital types

Choosing the correct basis set is essential for the success of the calculation, and there are a series of factors to consider when deciding which to employ. There are two types of AO commonly used in electronic structure calculations: Slater type orbitals (STO) and Gaussian type orbitals (GTO). Slater type orbitals have the functional form:

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{n-1}e^{-\zeta r}$$

(Eq. 29)

Where N is a normalization constant and $Y_{l,m}$ are spherical harmonic function. The exponential dependence on the distance between the nucleus and electron mirrors the exact orbitals for the hydrogen atom, but these type of functions do not have any radial nodes; nodes in the radial part are introduced by making linear combinations of STOs. The exponential dependence ensures a fairly rapid convergence with increasing number of functions, making this type of functions fundamentally better suited for electronic structure calculations. The problem lies in the fact that calculation of three- and four-center two-electron integrals cannot be performed analytically. Gaussian type orbitals can be written in either polar or Cartesian coordinates:

$$\chi_{\zeta,n,l,m}(r,\theta,\varphi) = NY_{l,m}(\theta,\varphi)r^{2n-2-l}e^{-\zeta r^2}$$

(Eq. 30)

$$\chi_{\zeta,n,l,m}(x,y,z) = Nx^{l_x}y^{l_y}z^{l_y}e^{-\zeta r^2}$$

The sum of l_x, l_y and l_z determines the type of orbital. The r^2 dependence of the GTOs makes them inferior to the STO in two respects: they do not show the correct behavior close to the nucleus, since GTOs show a flat slope at that point (in contrast to STOs, that present a discontinuous derivative at that point). Therefore, is a poor representation of an atomic orbital for short distances between the nucleus and the

electron. The second problem arising from the quadratic dependence is that GTOs fall too rapidly far from the nucleus, making for a poor representation of the "tail" of the wave function. Even when both AO can be chosen to form a complete basis set, three times as many GTOs than STOs are required for reaching a given level of accuracy. However, even when this implies the evaluation of a higher number of integrals, the calculation of GTOs is a lot easier than for STO. This is because the product of two GTO centered in two different points is equal to a simple Gaussian function centered in a third point, thus simplifying all the three- and four-centered two-electron integrals to simpler two-centered integral. Therefore, GTOs are usually preferred, and are used almost universally as basis sets [74,75]. Once the decision on the type of AO (STO/GTO) and the location (usually, the nucleus) is made, the most important factor is the number of functions to be used. It is important to keep in mind that, as the number of functions increases, the accuracy of the MOs improves, so, in the limit of a complete basis set the results are identical to those obtained by a numerical HF method, and this is known as the *Hartree-Fock limit*. This is not the exact solution to the Schrödinger equation, but the best single-determinant wave function that can be obtained [74]. However, the more functions are considered, the greater the increase in the computational resources needed to solve the problem at study, so achieving a compromise is important, using a cost-effective basis set: it should be big enough to provide the information needed to answer the questions posed, but small enough as to consume the least possible computational resources. The smallest number of functions possible is a *minimum basis set:* only enough functions are employed to contain all the electrons of the neutral atom.

2.3.3.3- Improving the basis sets

The first improvement of the basis sets is a doubling of all the basis functions, producing a *Double Zeta basis set (DZ).* The term zeta stems from the ζ (zeta) orbital exponents. The chemical bonding occurs between valence orbitals, and a variation of the DZ basis sets that takes this fact into consideration is the *split valence basis (SV),* which only doubles the number of valence orbitals, and is usually called valence double zeta (VDZ). The next step up in basis set size is a *Triple Zeta basis set (TZ)*, which

contains three times as many functions as the minimum basis set. The split valence can be also used in this case, thus only tripling the number of functions for the valence layer, producing a triple split valence basis set. Up until now, only the number of s- and p-functions has been taken into consideration.

In most cases, higher angular momentum functions are also important, and these are called *polarization functions.* The polarization functions are added to the chosen sp-basis; adding a single set of polarization functions to the DZ basis forms a *Double Zeta plus Polarization (DZP)* type basis. In a similar way, multiple sets of polarization functions with different exponents may be added: for example, if two sets of polarization functions are added to a TZ sp-basis, a *Triple Zeta plus Double Polarization (TZ2P)* type basis is obtained. For larger basis sets with many polarization functions the explicit composition in terms of number and types of functions is usually given. Basis sets are also frequently augmented with the so-called *diffuse functions*, basis functions with very small exponents, which decay slowly with distance from the nucleus. They are usually of *s* and *p* type; however, sometimes diffuse polarization functions are also used. Diffuse functions are necessary for correct description of anions and weak bonds, such as hydrogen bonds, and are frequently used for calculations of properties [74,75,78].

Combining a full set of basis functions, known as the *primitive* GTOs (PGTOs) into a smaller set of functions by forming fixed linear combinations is known as basis set contraction, and the resulting functions are called *contracted* GTOs (CGTOs). This approximation reduces the computational effort of representing the energetically important but chemically unimportant core electrons. The formerly introduced acronyms DZP, TZ2P, and so on, refer to the number of contracted basis functions. Contraction is especially useful for orbitals describing the inner (core) electrons, since they require a relatively large number of functions to represent the wave function cusp near the nucleus, and they are largely independent of the environment. Contracting a basis set will always increase the energy, since it is a restriction on the number of the variational parameters, and makes the basis set less flexible, but it will also reduce significantly the computational cost. The degree of contraction is the number of PGTOs entering the CGTO, typically varying between one and ten [74].

2.3.3.4- Commonly used basis sets

There are many different contracted basis sets available in the literature, and below there is a very short description of often used basis sets:

◊ Pople style basis sets: *STO-nG basis sets* are Slater type orbitals consisting of *n* GTOs. This is a minimum type basis set. It has been found that using more than three PGTOs for representing the STO gives little improvement, and the STO-3G is a widely used minimum basis.

k-nlmG basis sets are of the split valence type, with the *k* indicating how many PGTOs are used for representing the core orbitals. The nlm indicate both how many functions the valence are split into, and how many PGTOs are used for their representation. Two values (*nl*) indicate a double split valence, while three values (*nlm*) indicate a triple split valence. The values before the G indicate the s- and p-functions in the basis, the polarization functions are placed after the G. These types of basis have the restriction that the same exponent is used for both the s- and p-functions in the valence, thus increasing the computational efficiency but diminishing the flexibility of the basis. *6-31G* is a very widely example of these basis sets, and it is a split valence basis, where the core orbitals are a contraction of six PGTOs, the inner part of the valence orbitals is a contraction of three PGTOs and the outer part of the valence is represented by one PGTO [74].

Polarization and/or diffuse functions can be added to any of these basis sets: diffuse functions are usually *s*- and *p*-functions, thus they are placed before the G. They are denoted by + or ++ (the first + indicates the inclusion of diffuse functions only on heavy atoms, while the second one indicates that a diffuse function is also included for the hydrogen atoms in the system). Polarization functions are indicated after the G, with a separate designation for heavy atoms and hydrogen. Usually, the polarized functions are designed with their name (*d* or *f*), but if only one set of polarization functions is used, an alternative notation in terms of * can be used. For example, 6-31G* represents the 6-31G(d) basis set, while the 6-31G** represents the 6-31G(d,p) basis set [74].

◊ Dunning-Huzinaga basis sets: Huzinaga has determined uncontracted energy-optimized basis sets up to (10s6p) for first row elements; later, this was extended to (14s9p) and up to (18s13p) by van Duijneveldt and Partridge, respectively. Duning has used the Huzinaga primitive GTOs to derive various contraction schemes, and these are known as the *Dunning-Huzinaga (DH)* type basis set. These type of basis sets do not have the restriction of equal exponents for *s*- and *p*-functions, making them more flexible but also more computationally expensive [74].

2.3.3.5- Basis Set Superposition Error (BSSE)

Fixing the position of the basis functions at the nuclei allows for a compact basis set, otherwise, sets of basis functions positioned at many points in the geometrical space would be needed. When comparing energies at different geometries, the nuclear fixed basis set introduces an error. The quality of a given basis set is not the same at all geometries, due to the fact that the electron density around one nucleus may be described by functions centered at another nucleus, thus compensating for the basis set incompleteness on that nucleus. This effect is known as the *Basis Set Superposition Error (BSSE)*, and is arises when two chemical fragments, A and B, approach to form the AB supermolecule. The description of fragment A within the complex can be improved by the functions of fragment B and vice versa, while such extension is not possible in the calculation of the isolated fragments. Subsequently, in the process $A + B \rightarrow AB$, the total energy decreases by two factors: the stabilization of the system due to the fragments interaction, and the improvement in the individual atomic description. This latter effect is the actual BSSE, and is an artifact, which causes an unphysical overestimation of the interaction energy. This implies that it is not accurate enough to calculate the interaction energy simply as

$$E_{interaction}^{AB} = E_{AB}^{Opt} - E_{A}^{Opt} - E_{B}^{Opt}$$

(Eq. 31)

where E_{AB}, E_A and E_B represent the energies of the molecule and the fragments, respectively, and the superscript Opt denotes that their geometries have been

optimized. In order to obtain the accurate interaction energy, a positive correction is needed to correct the interaction energy in the supermolecule:

$$E_{interaction}^{AB} = E_{AB}^{Opt} - E_{A}^{Opt} - E_{B}^{Opt} + \delta_{AB}^{BSSE}$$

(Eq. 32)

The most widely used method to correct for BSSE effects (that is, calculating the δ_{AB}^{BSSE} value) is the *counterpoise (*CP^n, n =number of fragments*)* correction of Boys and Bernardi [79], an *a posteriori* correction method where the energy calculations for the individual monomers are performed using the whole supermolecular basis sets instead of only the appropriate fragment set. That means that the energy of A is calculated in the presence of both the normal basis set for that fragment and with the basis set functions of fragment B located at the corresponding nuclear positions, but without the B nuclei present, and vice versa, using so-called 'ghost orbitals'.

$$\delta_{AB}^{BSSE}(CP^n) = \sum_{i=1}^{n}(E_i^F - E_i^{F*})$$

(Eq. 33)

where the superscript F denotes that the fragments are frozen in their AB geometries, and the asterisk (*) denotes the presence of the ghost orbitals in the calculation [74,80-82].

2.3.4- Møller-Plesset perturbation theory

As it has been stated before, the HF method generates solutions to the Schrödinger equation where the real electron-electron interaction is replaced by an average interaction. If the considered basis set is sufficiently large, the HF wave function is able to account for approximately 99% of the total energy, but, as small as it seems, the remaining ~1% is often very important in describing chemical phenomena. This energy error is known as the *total correlation energy*, and it is due to the instant interaction

between electrons (since electrons tend to repel each other, they tend to place themselves as further apart as possible, thus the movement of a given electron is not completely independent from the movement of the others). This energy is defined as the difference between the true energy and the Hartree-Fock energy in a complete basis (the already defined Hartree-Fock limit):

$$E_{corr} = \varepsilon_{exact} - E_{HF}^{\infty}$$

(Eq. 34)

Physically, the correlation energy corresponds to the motion of the electrons being correlated. In practice, we usually don't know the exact energy ε_{exact}, but the exact energy for a given one-electron basis set can be computed. This allows the computation of the *basis set correlation energy*, which is what is usually considered as the correlation energy.

$$E_{corr}^{basis} \equiv E_{exact}^{basis} - E_{HF}^{basis}$$

(Eq. 35)

Several methods have been developed throughout the years in order to try to recover the correlation energy, thus improving the results obtained with HF. One of the possible approaches is based in the *perturbation methods*, based in the idea that the problem at hand only differs slightly from a problem that has already been solved (either exactly or approximately). The solution to the present problem, then, should in some sense be close to the solution already known, and this is described mathematically by defining a Hamiltonian operator that consists of two parts, a reference (H_0) and a perturbation (H'). Thus, the Hamiltonian for the actual problem can be described as in **Eq. 36**:

$$\widehat{H} = \widehat{H}_0 + \lambda \widehat{H}'$$

(Eq. 36)

The premise of perturbation methods is that H' is small compared to H_0. The solution to Schrödinger's equation when considering the perturbed Hamiltonian is expressed as a Taylor series in λ, the perturbation strength, as

$$\varepsilon_i = E_i^{(0)} + \lambda E_i^{(1)} + \lambda^2 E_i^{(2)} + \cdots$$

(Eq. 37)

$$\Psi_i = \Psi_i^{(0)} + \lambda \Psi_i^{(1)} + \lambda^2 \Psi_i^{(2)} + \cdots$$

(Eq. 38)

The perturbation methods developed in order to deal with systems that consist of several interacting particles is known as the *many-body perturbation theory (MBPT)*. In 1934, Møller and Plesset postulated what is called the *Møller-Plesset perturbation theory (MP),* a perturbation treatment for atoms and molecules in which the unperturbed Hamiltonian is the sum of the one particle Fock operators:

$$\hat{H}_0 \equiv \sum_{i=1}^{N} \hat{F}(i)$$

(Eq. 39)

The difference between the real r_{ij}^{-1} repulsion and the Fock operator (the average Hartree-Fock potential) becomes the perturbation ('fluctuation potential') $\hat{H}'$.

$$\hat{H}' = \sum_{i<j} r_{ij}^{-1} - v^{HF} = \sum_{i<j} r_{ij}^{-1} - \sum_i v^{HF}(i)$$

(Eq. 40)

The wave functions for the unperturbed Hamiltonian $\widehat{H}_0$ are the zero order wave functions (unperturbed), so that the Hartree-Fock function for the fundamental state ϕ_0 is one of them. That is:

$$\Psi_0^{(0)} = \phi_0$$

(Eq. 41)

The first order MP correction to the ground state energy, $E_0^{(1)}$ is given in **Eq. 42**:

$$E_0^{(1)} = \int \Psi_0^{(0)*} \widehat{H}' \Psi_0^{(0)} = \int \phi_0^* \widehat{H}' \phi_0$$

(Eq. 42)

The subindex 0 indicates the ground (or fundamental) state.

$$E_0^{(0)} + E_0^{(1)} = \int \Psi_0^{(0)*} \widehat{H}_0 \Psi_0^{(0)} + \int \phi_0{}^* \widehat{H}' \phi_0 = \int \phi_0{}^* (\widehat{H}_0 + \widehat{H}') \phi_0 = \int \phi_0^* \widehat{H}' \phi_0$$

(Eq. 43)

But $\int \phi_0^* \widehat{H}' \phi_0$ is the variational integral for the ϕ_0 Hartree-Fock wave function and, therefore, it's equal to the Hartree-Fock energy E_{HF}. Therefore,

$$E_0^{HF} = E_0^{(0)} + E_0^{(1)}$$

(Eq. 44)

Electron correlation energy, thus, starts at the second order correction with this choice of zeroth order Hamiltonian. The expression for the second-order energy involves

matrix elements of the perturbation operator between the HF reference and all possible excited states.

$$E_0^{(2)} = \sum_{I \neq 0} \frac{\left|\phi_0^* \hat{H}' \phi_I\right|^2}{E_0^{(0)} - E_I^{(0)}}$$

(Eq. 45)

Since the perturbation is a two-electron operator, all matrix involving triple, quadruple, etc., excitations are zero. The second-order correction, then, only involves a sum over doubly excited determinants, which can be generated by promoting two electrons from occupied orbitals i and j to virtual orbitals a and b. The summation must be restricted in order to count each excited state only once. The matrix elements between the HF and a doubly excited state are given by two electron integrals over MOs. The difference in total energy between two Slater determinants becomes a difference in MO energies, and the explicit formula for the second order Møller-Plesset correction becomes:

$$E_0^{(2)} = \sum_{i<j}^{occ} \sum_{a<b}^{vir} \frac{\left(\left(\int \phi_i^*(1)\phi_j^*(2)\phi_a(1)\phi_b(2)\right) - \left(\int \phi_i^*(1)\phi_j^*(2)\phi_b(1)\phi_a(1)\right)\right)}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b}$$

(Eq. 46)

The second-order energy correction is negative: electron correlation stabilizes the energy. Considering the molecular energy as $E_0^{(0)} + E_0^{(1)} + E_0^{(2)} = E_0^{HF} + E_0^{(2)}$, the calculation is denominated MBPT(2) or, more commonly, MP2, where the 2 is indicating the inclusion of the second-order energy correction. The formulas for the $E_0^{(3)}$, $E_0^{(4)}$ and so on, have been deduced in a similar manner, but usually the perturbation theory is only taken up to the second order correction, since MP2 typically accounts for 80-90% of the correlation energy. The third-order energy correction (MP3), even when it is not so computationally expensive, does not provide an important improvement on the energy correction; for the fourth-order correction, there is an improvement in the energy, but the computational cost is very big, thus explaining the prevalence of MP2 as the most economical method for including electron correlation. Further corrections,

such as MP5 or MP6, even though they have been deduced, have such complex formulas, that the actual calculation can only be performed in very small systems. In order to carry out an MP calculation (whatever the order of the correction considered), the first step is to select a basis set, and perform an SCF calculation, in order to obtain ϕ_0, E_0^{HF} and the virtual orbitals; only then is possible to evaluate the energy corresponding to the MP corrections [74,75,83,84].

2.3.5- Density Functional Theory

All *ab initio methods* start with a Hartree-Fock approximation that results in spin orbitals, and then electron correlation is taken into account. Even when the results of such calculations are reliable, the major disadvantage is that they are computationally expensive, and cannot be readily applied to large molecules of interest. The Density Functional Theory (DFT) offers a completely different approach to the calculation of molecular potentials, providing results comparable to configuration interaction (CI, another correlation energy correction method) and MP2 computational results, but more cost-effective, making it possible to employ DFT computations on molecules with over 100 heavy atoms [74,80].

The complexity of the N-electron wave function Ψ and the associated Schrödinger equation considered by the HF-SCF method has prompted the search for simpler functions, dependent on less parameters and useful in the determination of the system's energy as well as other molecular properties. This search has led to a long history of such theories, which, until 1964, only had the status of models. The work of Thomas and Fermi in the 1920s set the ground basis for these models, who were the first to realize that statistical considerations can be used to approximate the distribution of electrons in an atom. Thomas, in 1927, stated that "Electrons are distributed uniformly in the six-dimensional phase space for the motion of an electron at the rate of two for each h^3 of volume" and that there is an effective potential field that "is itself determined by the nuclear charge and this distribution of the electrons". Unfortunately, the method derived by Thomas and Fermi just described founder when one comes to molecules, but no molecular binding was predicted. This, plus the fact that the accuracy for atoms is not as good as with other methods, caused the method to be viewed as

oversimplified of no real importance for quantitative predictions. It was only in 1964, with the publication of Hohenberg and Kohn's paper that this situation changed, turning the Thomas and Fermi model into the basis of an actual theory [85].

2.3.5.1- The Hohenberg-Kohn theorems

In 1964, Pierre Hoghenberg and Walter Kohn stated: "The external potential $v(r)$ is determined, within a trivial additive constant, by the electron density $\rho(r)$" [86]. Since $\rho\ (r)$ determines the number of electrons, it follows that that function also determines the ground-state wave function Ψ and all other electronic properties of the system. It is important to notice that $v(r)$ is not restricted to Coulomb potentials. Then, the electronic energy of the molecule's ground state, E_0, can be identified as a *functional* (a functional $F[f]$ is a rule that associates a number to every function f) of $\rho(r)$, and, as such, can be written as $E_0 = E_0[\rho_0]$, where the subscript 0 denotes the ground-state. This idea comes from considering the electronic wave function for the ground-state, Ψ_0, in a molecule with n- electrons, is an eigenfunction of the purely electronic Hamiltonian (**Eq. 8**), which, in atomic units, is:

$$\widehat{H}_{el} = -\frac{1}{2}\sum_{i=1}^{n}\nabla_i^2 + \sum_{i=1}^{n} v(r_i) + \sum_{j}\sum_{i>j}\frac{1}{r_{ij}}$$

(Eq. 47)

$$v(r_i) = -\sum_{\alpha}\frac{Z_\alpha}{r_{i\alpha}}$$

(Eq. 48)

$v(r_i)$ represents the potential energy arising from the interaction between the ith electron and the different nuclei, and is dependent on the x_i, y_i, z_i coordinates for electron i, and the nuclear coordinates. Since the electronic Schrödinger equation is solved considering the Born-Oppenheimer and fix nuclei approximations, the nuclear coordinates are not variable in this equation. Thus, $v(r_i)$ depends only on x_i, y_i, z_i. In

DFT, $v(r_i)$ is considered an **external potential** acting on electron i, since it's a potential generated outside of the electron system [75,85].

Once the external potential $v(r_i)$ and the number of electrons are specified, the electronic wave functions can be determined, as well as the allowed energies for the molecule, as solutions to the electronic Schrödinger equation. Hohenberg and Kohn proved that in systems where the ground-state is not degenerate, the electron density $\rho_0(r)$ determines the external potential (except for an additive arbitrary constant) and it also determines the number of electrons. Consequently, the ground-state wave function and its energy (and, in this case, all the wave functions and energies for the excited states) are determined by the ground-state's electron density. This can be written as $E_0 = E_v[\rho_0]$, where the subscript v is indicating the dependence of E_0 on the external potential $v(r_i)$, which is different for different molecules [75].

The purely electronic Hamiltonian is the sum of the electronic kinetic energy, electron-nucleus attraction and electron-electron repulsion terms:

$$E_0 = E_v[\rho_0] = \bar{T}[\rho_0] + \bar{V}_{Ne}[\rho_0] + \bar{V}_{ee}[\rho_0]$$

(Eq. 49)

where the upper bar represent the mean values. $\bar{V}_{Ne}$ can be determined, considering **Eq. 48**, but the functionals $\bar{T}[\rho_0]$ and $\bar{V}_{ee}[\rho_0]$ remain unknown.

$$E_0 = E_v\lfloor\rho_0\rfloor = \int \rho_0(r)v(r).\,dr + F[\rho_0]$$

$$F[\rho_0] \equiv \bar{T}[\rho_0] + \bar{V}_{ee}[\rho_0] \qquad \textbf{(Eq. 50)}$$

Since $F[\rho_0]$ remains unknown, the previous equation does not provide a practical way to determine E_0 from ρ_0. To solve this problem, the second theorem of Hohenberg and Kohn has to be considered, which provides the energy variational principle. This theorem states: "For a trial density $\tilde{\rho}(r)$ such that $\tilde{\rho}(r) \geq 0$ and $\int \tilde{\rho}(\mathrm{r}).\,\mathrm{dr} = N$,

$$E_0 \leq E_v[\tilde{\rho}(r)]$$

(Eq. 51)

where $E_v[\tilde{\rho}(r)]$ is the energy functional of **Eq. 50** [86]. It is evident that this is analogous to the variational principle for wave functions, and, as it happened in that case, the only occasion when the equality is achieved, is when the real electron density for the ground state is considered. Hohenberg and Kohn only proved both of their theorems for non-degenerated ground states, but later on, Levy proved them for degenerate ground states as well [75,85].

2.3.5.2- The Kohn-Sham method

Even though Hohenberg and Kohn established a relationship between the electron density and the molecular properties of a given system, their theorems did not establish a practical way to obtain the energy or the any other property from ρ_0. The foundation for the use of DFT methods in computational chemistry is in the introduction of orbitals, as suggested by Kohn and Sham in 1965 [87]. On principle, this method could provide exact results, but, since the equations included in it include an unknown functional that has to be approximated, the Kohn-Sham (KS) formulation of DFT only provides approximate results. In this method, Kohn and Sham considered a fictional reference system (often called non interacting system), with n electrons that do not interact, all of them subjected to the same potential energy function $v_s(r_i)$ (the s subscript indicates that is acting on the fictional system). $v_s(r_i)$ is such that it makes the ground-state electron density of the reference system ($\rho_s(r)$) equal to the exact ground-state electron density for the problem molecule, $\rho_0(r)$. Since the electron density function of the ground-state determines the external potential, as proven by Hohenberg and Kohn, once $\rho_s(r_i)$ is defined, $v_s(r_i)$ for the reference system is unequivocally determined for the reference system. Since the electrons in the reference system do not interact, the Hamiltonian for this system is:

$$\hat{H}_s = \sum_{i=1}^{n} \left[-\frac{1}{2}\nabla_i^2 + v_s(r_i) \right] \equiv \sum_{i=1}^{n} \hat{h}_i^{KS}$$

(Eq. 52)

$\hat{h}_i^{KS}$ is the Hamiltonian for a Kohn-Sham electron. Since the reference system consists of particles that do no interact with each other, the ground state wave function $\Psi_{s,0}$ for it is a Slater determinant of the Kohn-Sham's spin orbitals of lower energy (as it was the case when considering the HF-SCF method), u_i^{KS}, where the spatial part $\theta_i^{KS}(r_i)$ is an eigenfunction of the one-electron operator $\hat{k}_i^{KS}$. In not so many words:

$$\Psi_{s,0} = |u_1 u_2 \dots u_n|, \qquad u_i = \theta_i^{KS}(r_i)\sigma_i$$

(Eq. 53)

$$\hat{h}_i^{KS}\theta_i^{KS} = \varepsilon_i^{KS}\theta_i^{KS}$$

(Eq. 54)

with σ_i being the spin function (α ó β) and ε_i^{KS} are the Kohn-Sham orbital energies.

2.3.5.3- The exchange-correlation potentials

Now, some definitions need to be introduced:

$$\Delta\bar{T}[\rho] \equiv \bar{T}[\rho] - \bar{T}_s[\rho]$$

(Eq. 55)

The subscript 0 is omitted from this equation on, in order to simplify the notation.$\Delta\bar{T}$ is the difference in the average electronic kinetic energy of the ground state between the molecule and the reference system.

$$\Delta\bar{V}_{ee}[\rho] = \bar{V}_{ee} - \frac{1}{2}\iint \frac{\rho(1)\rho(2)}{r_{12}} dr_1 dr_2$$

(Eq. 56)

where $r_{12}$is the distance between two points with coordinates x_1, y_1, z_1 and x_2, y_2, z_2 respectively. $\frac{1}{2}\iint \frac{\rho(1)\rho(2)}{r_{12}} dr_1 dr_2$ is the classical expression, in atomic units, of the interelectronic electrostatic repulsion when the electrons are scattered on a continuous charge distribution with electron density ρ. Substituting in **Eq. 50**:

$$E_v[\rho] = \int \rho(r)v(r).dr + \bar{T}_s[\rho] + \frac{1}{2}\iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta\bar{T}[\rho] + \Delta\bar{V}_{ee}[\rho]$$

The functionals $\Delta\bar{T}$ and $\Delta\bar{V}_{ee}$ are not known, thus the *exchange-correlation energy functional*, $E_{xc}[\rho]$, is defined as

$$E_{xc}[\rho] \equiv \Delta\bar{T}[\rho] + \Delta\bar{V}_{ee}[\rho]$$

(Eq. 57)

then

$$E_0 = E_v[\rho] = \int \rho(r)v(r).dr + \bar{T}_s[\rho] + \frac{1}{2}\iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + E_{xc}[\rho]$$

(Eq. 58)

Since the first three terms in **Eq. 58** can be easily evaluated once ρ is known, and are the main contributions to the energy in the ground-state, the key to carry out a precision KS DFT calculation of molecular properties is to have a good approximation for E_{xc}.

The electron density of an n-particle system with a Slater determinant of spin-orbitals $u_i^{KS} = \theta_i^{KS}(r_i)\sigma_i$ as wave function is obtained as $\sum_{i=1}^{n}\left|\theta_i^{KS}\right|^2$, hence

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

(Eq. 59)

Now, with all the definitions considered, and taking into account that $\bar{T}_s[\rho]$ can be determined considering the Kohn-Sham orbitals, E_0 can be calculated from the electron density if the KS orbitals are obtained and the E_{xc} functional is known. In order to find the Kohn-Sham orbitals, the Hohenberg and Kohn variational theorem needs to be applied: the energy for the ground-state can be determined varying ρ (and, according to **Eq. 59**, this is the same than vary the KS θ_i^{KS} orbitals) in such a way that $E_v[\rho]$ is minimized.

The *exchange-correlation potential*, v_{xc}, is the derivative of the exchange-correlation energy functional:

$$v_{xc}(r) \equiv \frac{\delta E_{xc}[\rho(r)]}{\delta\rho(r)}$$

(Eq. 60)

It is important to notice that both E_{xc} and v_{xc} are not known, so they must be approximated in order to solve **Eq. 54**.

The Kohn-Sham orbitals do not have a physical meaning (they were obtained for the reference fictitious system) other than allowing the determination of ρ; in fact, there is no molecular wave function in DFT. However, in practice, since the occupied KS orbitals are similar to MOs calculated employing the HF-SCF method, they can be used in qualitative MO discussions of reactivity and molecular properties [75].

The exchange-correlation energy E_{xc} includes the kinetic correlation energy, the exchange energy (associated to the anti-symmetry requirement), the Coulomb's correlation energy (associated to the interelectronic repulsions) and a self-interaction

correction (SIC). This last correction is derived from the fact that the classical expression for the electrostatic repulsion wrongly allows a given electron to interact with its own charge contributions to ρ, and that is not actually possible [75].

2.3.5.4- The exchange-correlation functionals

The difference between various DFT methods is the choice of functional form for the exchange-correlation energy. It can be proven that the exchange-correlation potential is a unique functional, valid for all systems, but an explicit functional form for it has not yet been found. It is customary to separate E_{xc} into two parts: a pure exchange E_x and a correlation part E_c, but it must be noted that only the combined exchange-correlation hole has a physical meaning, calling for a combined E_{xc}. E_x is defined by the same formula employed for the exchange energy in Hartree-Fock theory, only that in the current case the Kohn-Sham orbitals are the ones considered. Earlier works tended to work on the two components separately, and then combining them, while the newest trend tend to construct the two parts in a combined fashion [74,75].
Each of the exchange and correlation energies is often written in term of the energy per particle (energy density), ε_x and ε_c.

$$E_{xc}[\rho] = E_x[\rho] + E_c[\rho] = \int \rho(r)\varepsilon_x[\rho(r)]dr + \int \rho(r)\varepsilon_c[\rho(dr)]dr$$

(Eq. 61)

The correlation between electrons of parallel spin is different from the one between electrons of opposite spin. The exchange energy is given 'by definition' as a sum of contributions from the α and β spin densities, as exchange energy only involves electrons of the same spin. The kinetic energy, the nuclear-electron attraction and Coulomb terms are trivially separable in terms of electron spin

$$E_x[\rho] = E_x^{\alpha}[\rho_\alpha] + E_x^{\beta}[\rho_\beta]$$

(Eq. 62)

$$E_c[\rho] = E_c^{\alpha\alpha}[\rho_\alpha] + E_c^{\beta\beta}[\rho_\beta] + E_c^{\alpha\beta}[\rho_\alpha, \rho_\beta]$$

The total density is the sum of the α and β contributions, $\rho = \rho_\alpha + \rho_\beta$. Functionals for the exchange and correlation energies may be formulated in terms of separate spin densities, but, instead, they are often given as functions of the *spin polarization,* ς (normalized difference between ρ_α and ρ_β), and the radius of the effective volume containing one electron, r_s:

$$\varsigma = \frac{\rho_\alpha - \rho_\beta}{\rho_\alpha + \rho_\beta}$$

(Eq. 63)

$$\frac{4}{3}\pi r_s^3 = \rho^{-1}$$

The different exchange-correlation functionals considered in the various DFT methods are, to a large extent, empirical. In the next lines, a brief introduction to the different DFT methods is introduced.

2.3.5.5- The Local Density Approximation (LDA) and the Local Spin Density Approximation (LSDA)

The *Local Density Approximation (LDA)* is the simplest approach to a DFT study. It assumes that the density locally can be treated as a uniform electron gas, thus consider the density a slowly varying function. The correlation-exchange energy is determined as an integral of some function of the total electron density:

$$E_{xc}^{LDA}[\rho] = \int \rho(r)\varepsilon_{xc}(\rho)dr$$

(Eq. 64)

$\varepsilon_{ci}(\rho)$ is the exchange energy plus the electron correlation in the uniform electron gas with electron density ρ.

$$v_{xc}^{LDA} = \frac{\delta E_{xc}^{LDA}}{\delta\rho} = \varepsilon_{xc}(\rho(r)) + \rho(r)\frac{\delta\varepsilon_{xc}(\rho)}{\delta\rho}$$

(Eq. 65)

ε_{xc} can be expressed as

$$\varepsilon_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho)$$

(Eq. 66)

with

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}(\rho(r))^{1/3}$$

(Eq. 67)

The correlation part has been determined as a very complex function, $\varepsilon_c^{VWN}(\rho)$, by Vosko, Wilk and Nusair [88]. Then,

$$v_{xc}^{LDA} = v_x^{LDA} + v_c^{LDA} = -\left[(3/\pi)\rho(r)\right]^{1/3} + v_c^{VWN}$$

(Eq. 68)

$$E_i^{LDA} \equiv \int \rho \varepsilon_i dr = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int [\rho(r)]^{4/3} dr$$

(Eq. 69)

The LDA approach is accurate when the problem consists of closed shell molecules. When open-shell molecules or molecular geometries closed to dissociation are considered, it is necessary to consider the *Local Spin Density Approximation (LSDA),* which shields better results in such cases. While LDA considers that electrons with opposite spin, paired between them, 'occupy' the same KS orbital, LSDA allows these electrons to have different spatial KS orbitals, $\theta_{i\alpha}^{KS}$ and $\theta_{i\beta}^{KS}$. The Hohenberg, Kohn and Sham theorems do not require the use of different orbitals for electrons with different spins, and, if the exact $E_{xc}[\rho]$ was known, this would not be necessary. With the approximated functionals employed in DFT KS, it is beneficial to allow the possibility of different orbitals for electrons with different spins, improving the calculated properties for open-shell molecules and species with geometries close to dissociation [74,75].

2.3.5.6- Gradient corrected methods

Both LDA and LSDA are based in the uniform gas model, which is appropriate for a system in which ρ varies slowly with the position. In order to improve these approximations, a non-uniform electron gas must be considered, correcting the LSDA to include the variation of the electron density with the position. In order to achieve this, the first derivative of the density is included as a variable:

$$E_{xc}^{GGA}\left[\rho^{\alpha}, \rho^{\beta}\right] = \int f(\rho^{\alpha}(r), \rho^{\beta}(r), \nabla\rho^{\alpha}(r), \nabla\rho^{\beta}(r)) dr$$

(Eq. 70)

f is a function dependent on spin densities and their gradients, and *GGA* stands for *Generalized Gradient Approximation*. The term *gradient-corrected functional* is also employed to describe this type of approximations. In these methods, the exchange-

correlation functional is usually divided into an exchange and a correlation parts, modeled separately.

$$E_{xc}^{GGA} = E_{x}^{GGA} + E_{c}^{GGA}$$

(Eq. 71)

Some GGA exchange functionals commonly used are the Perdew-Wang functional from 1986, which does not contain any empirical parameters, and that is usually found as PW86 or PWx86; the Becke functional from 1988, usually noted as B88, Bx88 or B, and the Perdew and Wang functional from 1991, PWx91. PWx86 and B88 provide similar results when predicting molecular properties. As for the correlation functionals, some of the more usually employed are the Lee-Young-Parr functional (LYP), the Perdew 1986 functional (P86 or Pc86), the parameter-free Perdew-Wang 1991 (PW91 or PWc91), and the Becke correlation functional (Bc95 or B96) [74,75]. The exchange-correlation functional by Perdew, Burke and Ernzenhof (PBE) does not have empirical parameters [89].

2.3.5.7- Higher order gradient or meta*-GGA methods*

The logical extension to the GGA methods is to allow the exchange and correlation functionals to depend on higher order derivatives of the electron density, with the Laplacian being the second-order term. Alternatively, the functional can be taken to depend on the orbital kinetic energy density. Both options (the Laplacian of the density and the orbital kinetic energy) carry essentially the same information, since they are related via the orbitals and the effective potential of the system. Inclusion of either of them as a variable leads to the so-called *meta-GGA functionals*, and any functional that, in general, uses orbital information may be included in this category too. Calculation of the orbital kinetic energy density is numerically more stable than calculation of the Laplacian of the density. One of the earliest attempts to include kinetic energy functionals was by Becke and Russel (BR), followed by a similar correction considered by Becke (B95), which is one of the few functionals that does not have the self-interaction problem. The VSXC (Voorhis-Scuseria-eXchange-Correlation) functional includes the kinetic energy density and contains 21 parameters, which are fitted to

experimental data. Other examples of this type of functionals are the Tao-Perdew-Staroverov-Scuseria (TPSS) exchange-correlation functional, which is non-empirical, and the Perdew-Kurth-Zupan-Blaha (PKZB) functional. Any exchange functional can be combined with any correlation functional: for example, BLYP/6-31G* indicates a DF calculation with the exchange functional Becke 1988 and the Lee-Young-Parr (BLYP) correlation functional, and where the Kohn-Sham orbitals are expanded in the 6-31G* basis [74,75].

2.3.5.8- Hybrid or hyper*-GGA methods*

Finally, the hybrid or hyper-GGA methods are a class of approximations to the exchange-correlation energy functional that incorporate a portion of exact exchange from Hartree-Fock theory with exchange and correlation from other sources (such as LSDA, for example). The exact exchange energy functional is expressed in terms of the Kohn-Sham orbitals rather than the density, so it also receives the name of *implicit density functional*. One of the most commonly used versions is B3LYP (Becke, 3-parameter functional, Lee-Yang-Parr); other options are B3PW1 and B1B96 [90], and the M06 suite of functionals [91], among many others . The inclusion of exact HF exchange is often found to improve the calculated results, although the optimum fraction to include depends on the specific properties of interest. At least part of the improvement may arise from reducing the self-interaction error, since HF is completely self-interaction free [74,75].

2.3.5.9- DFT pros and cons

DFT has become very popular, since it exhibits most of the advantages of the HF-SCF method, but including electron correlation. This method can, in principle, provide the exact (non-relativistic) results for the electronic structure of a molecule or solid. However, it is not the panacea, since, as in everything, it presents some flaws. For example, the theory was developed for ground-state systems, and, even when some versions have been derived to be applied in excited states, it is still not possible to carry out accurate practical calculations for them. Another fact is that, since the functionals

employed are actually approximations, DFT KS is not variational, and it is possible to obtain a value for the energy that is actually lower than the real ground-state energy. Even when DFT KS allows to achieve good results for most molecular properties with the functionals available nowadays, it cannot match the accuracy obtained with the application of electron correlation methods of high order; on the other hand, these methods are only applicable in small molecules, while DFT can manage quite large molecules [74,75,92].

As always, the decision on the approach to use, as well as the specific method within that approach, will finally depend on the problem at hand (including the molecule itself and what needs to be determined for that molecule) and, of course, the resources available.

2.3.6- Molecular Mechanics

All the methods previously described methods are based on Quantum Mechanics (QM), and even though they provide very accurate molecular information, including energy and other molecular properties, they are typically limited to small systems (even in the case of DFT, that allows to study bigger systems than the HF-SCF method, the limit is around the 100 heavy atoms), due to the computational cost that they imply. Systems of biochemical interest usually involve macromolecules that contain 1000-5000 or more atoms, plus a condensed phase environment that sometimes needs to be represented explicitly since it can affect the conformation or activity of the macromolecule. This can lead to biochemical systems of 20000 atoms or even more, which can hardly be analyzed with electronic structure methods. In addition to their size, the dynamical nature and the environmental mobility of biochemical systems require that a large number of conformations, generated via various methods, be subjected to energy calculations. Is in such cases that *Molecular Mechanics (MM)* is the method of choice [93].

Molecular Mechanics methods, also known as *force field (FF)* methods, do not deal with a Hamiltonian, a wave function or an electron density. Instead of that, these methods are a strict empirical, or deductive, technique for describing the potential energy of a molecule. They consider the electronic energy to be a parametric function of the

nuclear coordinates, and the parameters are fit to experimental or higher level computational data. The 'building blocks' in force field methods are atoms, and molecules are described by a 'ball and spring' model, with the atoms presenting different sizes and 'softness', and bonds having different lengths and 'stiffness'. The electrons are not considered as individual particles; hence bonding information needs to be provided explicitly, rather than being the result of solving the electronic Schrödinger equation. MM employs classical mechanics in its calculations, making them computationally faster [74,75,80,93,94].

MM works due to the validity of several assumptions, the first of which is the Born-Oppenheimer approximation, without which it would be impossible to contemplate writing the energy as a function of the nuclear coordinates at all. MM is based upon a rather simple model of interactions within a system, with contribution from different processes, such as the stretching of bonds, the opening and closing of angles and the rotations about single bonds. Even when simple functions are used to describe these contributions the force field can perform quite acceptably [95]

The foundation of force field methods is the observation that molecules tend to be composed of units that are structurally similar in different molecules: all C-H bond lengths, for example, are roughly constant in all molecules, and the same happens with the C-H stretch vibrations, indicating comparable force constants. If the C-H bonds are further divided into groups, considering the environment around that bond (for example, if the hydrogen atom is attached to a single-, double- or triple-bonded carbon), the variation within each one of these groups is even smaller. This considerations hold for other features as well (for example, all C=O bonds have approximately the same length, all double-bonded carbons are essentially planar, and so on), including the energetic features. The idea of molecules being composed of atoms, structurally similar in different molecules, is implemented in these methods as atom *types*. The atom type depends on the atomic number and the type of chemical bonding it is involved in. Transferability is a key attribute of a force field, for it enables a set of parameters developed and tested on a relatively small number of cases to be applied to a much wider range of problems. Moreover, parameters developed from data on small molecules can be used to study much larger molecules such as polymers [74,96].

Each force field is then an empirical fit of a mathematical expression in function of molecular coordinates on the potential energy surface, describing entire classes of molecules as an extrapolation from the experimental data of a representative set of molecules, and, in many cases, trading accuracy for generality [97].

2.3.6.1- The force field energy

Molecular Mechanics methods use empirical energy functions, which are called force fields (hence the alternative name for the method). Many of the molecular modeling force fields in use today for molecular systems can be interpreted in terms of a simple six-component picture of intra- and inter-molecular forces within the system. They may have different functional forms, but always include energetic penalties associated with the deviation of bonds and angles away from their reference values, functions describing how the energy changes as bonds are rotated and terms describing interactions between non-bonded parts of the system. Every functional form includes certain constants, or parameters, which allow approximating the mathematical expressions to empirical reference values [95]. The energy of a given configuration is calculated as follows:

$$E = \overbrace{\sum bond\ stretching}^{\text{pairs of atoms}} + \overbrace{\sum angle\ bending}^{\text{triplets of atoms}} + \overbrace{\sum dihedral}^{\text{quartets of atoms}} + \overbrace{\sum van\ der\ Waals}^{\text{pairs of atoms}} + \underbrace{\sum Coulombic}_{\text{pairs of atoms}} + \underbrace{\sum crossterms}_{\text{coupling between the first three terms}}$$

(Eq. 72)

The sum of all the potential energy terms for a particular atom gives a mathematical representation for how that atom would move under the influence of the motions or displacements of all the other atoms in the system: the potential energy is a representation of the forces experienced by that atom. The contribution that kinetic

energy (temperature) makes to a molecular energy can be estimated either by carrying out a statistical thermodynamic analysis (which will be discussed later on this chapter), or a Monte Carlo study [74,94]. In the next few lines, a more detailed explanation on the equations used to describe the interactions in a FF.

Bond stretch: is the energy function for stretching a bond between two atom types A and B. Different function forms have been used throughout time, with the simplest approach being the harmonic form the simplest possible, and sufficient for most systems.

$$V_r = \frac{1}{2} k_{AB} (r - r_{AB})^2$$

(Eq. 73)

Where k_{AB} is the force constant for the stretch between A and B, and r_{AB} is the natural bond distance between those two atoms. Both k_{AB} and r_{AB} are parameters of the force field. Another very used approach is the Morse function [98] :

$$V_r = D_{AB}\left[e^{-\alpha(r-r_{AB}} - 1\right]^2$$

$$\alpha = \left[\frac{k_{AB}}{2D_{AB}}\right]^{1/2}$$

(Eq. 74)

D_{AB} is the bond energy for the bonds between centers A and B, r_{AB} is the unstrained or natural bond distance, and k_{AB} is the force constant. The Morse function represents dissociation much better than the harmonic approach, so is the function of choice when bonds are being strained (for example, in bond braking); away from this situation, it is more cost-effective to use the harmonic approach. Series approximations to the Morse function are used in some modern force fields.

It is important to notice that for each bond type there are at least two parameters to be determined. Higher order expansions, and the Morse potential, have one additional parameter that needs to be determined [74,94].

Angle bend: is the energy required for bending an angle formed by three atoms A-B-C, where there is a bond between A and B, and between B and C. The simple harmonic expansion is adequate for most applications:

$$V_\theta = \frac{1}{2} k_{ABC}(\theta - \theta_{ABC})^2$$

(Eq. 75)

The bending force constant (k_{ABC}) and the strain-free bond angle (θ_{ABC}) are the parameters associated to this function. There may be some cases in which higher accuracy is required, and, in these cases, usually a third-order term is usually enough [74,94].

Torsion: it describes part of the energy change associated with rotation around a B-C bond in a four-atom sequence A-B-C-D, where A-B, B-C and C-D are bonded. The torsional angle is defined as the angle formed by the A-B and C-D bonds. These potentials are used to mimic the preference for staggered conformations about $sp^3 - sp^3$ bonds, and the preference for eclipsed conformations about $sp^2 - sp^3$ bonds. To encompass the periodicity, this potential is written as a Fourier series:

$$V_\phi = K_{ABCD} \sum_{n=0}^{m} C_n \cos n\phi$$

(Eq. 76)

where ϕ is the torsional angle, and $K_{ABCD}$is the force constant. The coefficients C_n are determined by the rotational barrier V_ϕ, the periodicity of the potential, and the natural angle, ϕ_{ABCD}. For molecules that are composed of atoms having a maximum valence of four (essentially all organic molecules) the first three terms of the Fourier series are sufficient for qualitatively reproducing their rotational profiles. Force fields that are

aimed at large systems often limit the Fourier series to only one term, depending on the bond type [74,94].

van der Waals: this term describes the repulsion or attraction between atoms that are not directly bonded. Together with the electrostatic term they describe the non-bonded energy. This term is very positive at small distances (due to the overlap of the electron clouds of the two atoms, both negative, hence, they repel each other), has a minimum that is slightly negative at a distance corresponding to the two atoms just 'touching' each other, and approaches zero as the distance becomes large. A general functional form that fits this condition is

$$V_{vdW}(R^{AB}) = V_{repulsion}(R^{AB}) - \frac{C^{AB}}{(R^{AB})^6}$$

(Eq. 77)

It is not possible to derive theoretically the functional form of the repulsive interactions, it is only required that it approaches zero as R (distance between the two atoms considered) goes to infinity, and it should approach zero faster than the R^{-6} term, as the energy should go towards zero from below. A popular function that obeys these general requirements is the Lennard-Jones (LJ) potential, where the repulsive part is given by a R^{-12} dependence (C_1 and C_2 are suitable constants):

$$V_{LJ}(R^{AB}) = \frac{C_1}{(R^{AB})^{12}} - \frac{C_2}{(R^{AB})^6}$$

(Eq. 78)

Another way of writing **Eq. 78** is

$$V_{LJ}(R) = \varepsilon\left[\left(\frac{R_0}{R}\right)^{12} - 2\left(\frac{R_0}{R}\right)^{6}\right]$$

(Eq. 79)

where R_0 represents the minimum energy distance and ε the depth of the minimum. There are no theoretical arguments for choosing the exponent in the repulsive part to be 12, this is a purely computational choice, and there is actual evidence that an exponent of 9 or 10 gives better results. Different force fields can present different expressions for this term of the potential energy, depending on the origin of the FF, and its primary target systems. The Lennard-Jones potential considers two parameters, while some other expressions for the van der Waals potential energy include up to three parameters, and these usually allow a better representation of the non-bonded interaction [74,94].

Coulomb: this term of the potential energy represents the internal (re)distribution of the electrons, creating positive and negative parts of the molecules, thus giving place to electrostatic interactions. These are the other part of the non-bonded interactions. Since it represents the interaction between point charges, the functional form is the Coulomb potential, with q_1 and q_B representing the point charges, R^{AB} is the non-bonded distance between the interacting atoms, and ε is a dielectric constant:

$$V_{el}(R^{AB}) = C\frac{q_A q_B}{\varepsilon R^{AB}}$$

(Eq. 80)

This interaction can also be represented as a dipole description, such as:

$$V_{el}(R^{AB}) = \frac{\mu_a \mu_b}{\varepsilon(R^{AB})}(\cos\chi - 3\cos\alpha_A \cos\alpha_B)$$

(Eq. 81)

The angles χ, α_A and α_B are described in **Fig. 6**:

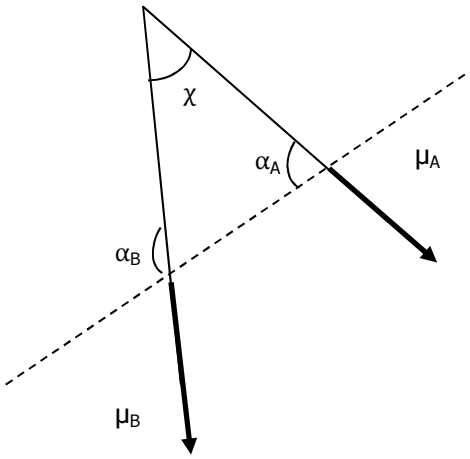


Fig. 6 – Geometric representation of dipole-dipole interaction

When properly parameterized, there is not a big difference between the results obtained with both models [74,94].

Cross terms: The first five terms in the general energy expression are common to all force fields. The last term, V_{cross} covers couplings between these fundamental, or diagonal, terms. It may also include inversions, or specific hydrogen-bond potentials, and there is no general form for them, they are different for each force field [74,94].

To define a force field not only the functional form must be specified, but also the parameters: two force fields may use an identical functional form yet have very different parameters. Moreover, force fields with different functional forms but different parameters, and force fields with different functional forms may give results of comparable accuracy. A FF should be considered as a single entity: it is not possible to take the parameters from one of them and mix them with the parameters of another one. Force fields are primarily design to reproduce structural properties, but they can also be used to predict other properties. A given force field is generally designed to predict certain properties and is parameterized accordingly [95] .

2.3.6.2- Existing force fields

Over the years, there have been a large number of force fields used for simulation of proteins, each of them with its own particularities [99]. In the next few lines, the more commonly used molecular mechanics force fields are summarized:

CHARMM: Chemistry at HARvard Macromolecular Mechanics is both the name of a force field and a program incorporating that force field, originally developed in the 1980s. It was originally devised for proteins and nucleic acids, but it has now been applied to a range of biomolecules, molecular dynamics, solvation, crystal packing, vibrational analysis and QM/MM studies. CHARMM uses five valence terms, one of which is an electrostatic term [99,100].

AMBER: Assisted Model Building with Energy Refinement is also the name of a force field and a Molecular Mechanics program. This FF was parameterized specifically for proteins and nucleic acids, and it uses only five bonding and non-bonding terms along with a sophisticated electrostatic treatment, but no cross terms are included. It provides very good results for proteins and nucleic acids, but can be somewhat erratic for other systems [95,100].

GROMOS: as in the previous two cases, GRonigen MOlecular Simulation is the name of both a force field and the program incorporating that force field. This FF is popular for predicting the dynamical motion of molecules and bulk liquids. It is also used for modeling biomolecules, and it uses five valence terms, one of which is an electrostatic term.

UFF: the Universal Force Field is the most promising full periodic table FF available at this time. UFF is most widely used for systems containing inorganic elements. It was designed to use four valence term, but not an electrostatic one [100].

MM1, MM2, MM3, MM4: these are general-purpose organic force fields. There have been many variants of the original methods, particularly MM2. MM1 is seldom used since newer versions show measurable improvements. The MM3 method is probably one of the most accurate ways of modeling hydrocarbons. MM4 is the most recent version, and even though there are not enough data as to allow a broad generalization of the results, the ones published are encouraging. These are some of the most widely used FF for organic molecules, given the accuracy of their representation for these kind

of molecules. MMX and MM+ are variations on MM2. These force fields use five to six valence terms, one of which is an electrostatic term and one to nine cross terms [100].
MMFF: the Merck Molecular Force Field is one of the most recently published in the literature. It is a general-purpose method, particularly popular for organic molecules. MMFF94 was originally intended for Molecular Dynamic simulations, but has also been much use for geometry optimization. It uses five valence terms, one of which is an electrostatic one, and one cross term [100].
The above described FFs are not the only ones in existence, but they do represent the most widely used ones, depending on the characteristics of the system at study. In the case of this thesis work, since the object of study is a biomolecular system, both CHARMM and AMBER represent a good choice of FF to represent it, with CHARMM being our first choice, given the fact that hybrid methods (including both QM and MM methodologies) are necessary to correctly represent the interactions taking place in the catalytic site.

2.3.7- Combined Quantum Mechanical/Molecular Mechanical approaches: the ONIOM method

One of the main themes in computational chemistry is to find a balance between the accuracy of the results and the computational cost. On the one hand, FF methods are inherently unable to describe the details of bond breaking/forming, or electron transfer reactions, since there is an extensive rearrangement of electrons. On the other hand, QM is unable to model very large compounds in a cost-effective fashion, since the high accuracy model chemistries scale unfavorably with the size of the problem, resulting in a practical limit on how large a system can be studied, placing a lot of interesting chemical and/or biological systems out of reach of traditional approaches. If the system of interest is too large to be treated entirely by electronic structure methods, there are two approximate methods that can be used. In some cases, the system can be 'pruned' to a size that can be treated by replacing 'unimportant' (or less critical regarding the reaction at study) parts of the molecule with smaller groups. This approximation, however, does not apply for studying enzymes or solvation, where it is not possible to substitute any fraction of the system without seriously affecting the accuracy of the

model. In such cases, the so-called *hybrid methods* offer a solution, combining either QM and MM methods, or QM methods with different level of theory in one calculation. Over the years, a variety of hybrid methods have been presented, which are conceptually quite similar, but differ in a number of details. Most methods only combine a Quantum Mechanical method with a Molecular Mechanics method, where a very large compound is model using MM, and the most crucial section of the molecule is modeled with QM, and these are generally referred to as *QM/MM*. Only several hybrid methods can also combine QM with QM, or more than two computational methods. Other distinctions involve the description of the interaction between the regions, or how the regions are connected when there is a covalent interaction between them. In particular, the *ONIOM* method (the name stands for Our own N-layered Integrated molecular Orbital and molecular Mechanics), developed by Morokuma and co-workers, can in principle use any computational method, combining QM with QM as well as QM with MM, and do so for any number of layers (usually two or three). This method uses link atoms to saturate the dangling bonds resulting from cutting covalent bonds between the QM and the MM regions. Even though the original ONIOM method only employed mechanical embedding for the QM/MM interface, more recent extensions have also included electron embedding [101]. This method incorporates the partial charges of the MM region into the quantum mechanical Hamiltonian, thus providing a better description of the electrostatic interaction between the QM and MM regions (as it is treated at the QM level) and allows the QM wave function to be polarized [74,100,102]. The ONIOM method employs an extrapolation scheme based on assumed additivity. For a two-layer scheme, the small (*model*) system is calculated at both levels of theory, while the large (*real*) system is calculated at the low level of theory. The link atoms are considered in the model system. The result for the real system at the high theoretical level is estimated by adding the change between the high and low levels of theory for the model system to the low level results for the real system [74], as illustrated in **Fig. 7**:

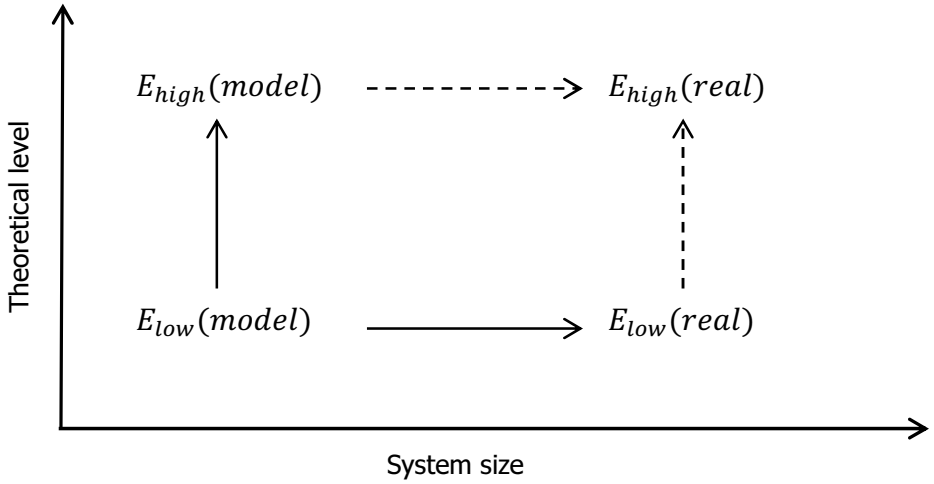


Fig. 7 – Illustration of the ONIOM extrapolation method

In such a two-layer ONIOM calculation, the total energy of the system is obtained from three independent calculations:

$$E^{ONIOM(QM:MM)} = E^{QM}_{model} + E^{MM}_{real} - E^{MM}_{model} = E^{high}_{model} + E^{low}_{real} - E^{low}_{model}$$

(Eq. 82)

The *real* system in this case contains all the atoms, and is calculated only at the MM level, while the *model* system contains the part of the system that is treated at the QM level, along with the link atoms. To evaluate the ONIOM energy, both QM and MM calculations need to be carried out for the model system. Because the positions of the link atoms are defined in terms of the atoms in the real system, the potential energy surface (PES), and therefore geometry optimization, is well defined [103]

Although ONIOM in its original formulation follows the mechanical embedding scheme, the formalism of ONIOM (QM:MM) can be extended to include electronic embedding [104,105]. Because the model system needs to be identical for both the QM and MM calculation, the environmental charges are included in both, not changing the real system calculation:

$$E^{ONIOM(QM:MM)-EE} = E^{v,model,QM} + E^{real,MM} - E^{v,model,MM}$$

(Eq. 83)

To avoid overpolarization, the charges close to the QM region are scaled. Because these charges will be scaled in both the $E^{v,model,QM}$ and $E^{v,model,MM}$ terms, the balance will not change. The charge interactions that are overcounted or undercounted at the QM level in the $E^{v,model,QM}$ will be balanced at the MM level in the $E^{v,model,MM}$ term [102].

An important and critical feature of the ONIOM method is the treatment of the link atoms. In a two-layer ONIOM scheme, represented in **Fig. 8**, the atoms present both in the model and the real systems are called set 1 atoms, and their coordinates are denoted by R_1. The set 2 atoms are the artificially introduced link atoms. They only occur in the model system, and their coordinates are described by R_2. In the real system they are replaced by the atoms described by R_3. Atoms that belong to the outer layer and are not substituted by link atoms are called set 4 atoms with the coordinates R_4. The geometry of the real system is then described by R_1, R_3 and R_4, and they are the independent coordinates for the ONIOM energy:

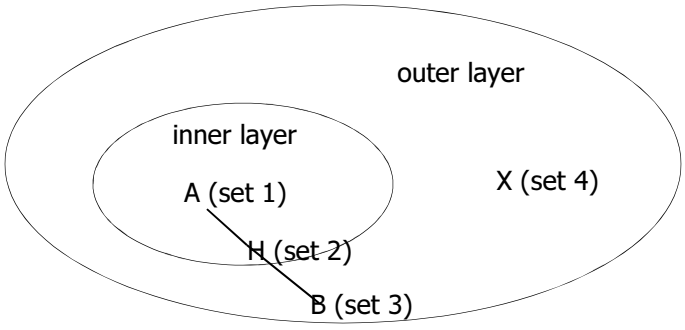


Fig. 8 – Definition of atom sets within the ONIOM scheme

$$E_{ONIOM} = E_{ONIOM}(R_1, R_3, R_4)$$

(Eq. 84)

In order to generate the model system, described by R_1 and the link atoms R_2, the latter is defined as a function of R_1 and R_3:

$$R_2 = f(R_1, R_3)$$

(Eq. 85)

The explicit functional form of the R2 dependency can be chosen arbitrarily. However, since link atoms are introduced to mimic the corresponding covalent bonds of the real system, they should follow the movement of the atoms they replace. Therefore, a coupling scheme should be adopted: if atom A belongs to set 1 and atom B belongs to set 3, the set 2 link atom (symbolized by H in **Fig. 8**) is placed onto the bond axis A-B. In terms of internal coordinates, the same bond angles and dihedral angles are chosen for both set 2 and set 3 atoms, making the link atoms always aligned along the bond vectors of the real system. For the exact position r_2 of a single H atom along an A-B bond (r_3-r_1), a fixed scale factor (or distance parameter) g is introduced.

$$r_2 = r_1 + g(r_3 - r_1)$$

(Eq. 86)

If the A-B bond distance $|r_3 - r_1|$ changes during a geometry optimization, the A-H bond distance $|r_2 - r_1|$ also changes. The value of the parameter g depends on the nature of the cut bond, the atoms A and B, and the link atom, and the levels of theory used for the two layers connected by the cut bonds [106].

If no covalent bonds are broken in the real system to form a model system (as it happens when the real system contains a solute molecule and solvent molecules, and the model system containing only the solute molecule is formed by simply removing the solvent molecules), there is no need to introduce a link atom [106].

The calculation of chemical shifts studied with hybrid methods has received much attention in the literature, ranging from large systems studied with QM/MM to small organic systems, studied with QM/QM methods. Finally, QM/MM methods have been used in numerous studies on excited state surfaces, including the dynamical

investigation of surface crossings. The ONIOM approach has shown to be successful in reproducing benchmark calculations and experimental results [107-109].

2.3.8- Potential Energy Surfaces (PES)

Different structures of the same molecule have distinct physicochemical properties, being its energy one of the most explicit posteriori ones. The energy of a molecule is dependent on parameters such as its environment and conformation. Considering the Born-Oppenheimer and the fix nuclei approximations, the result of the Schrödinger equation will be exclusive for a fixed geometry of nuclei, and describes the absolute energy of the molecular structure. Basically, for many-atom molecules, the number of internal coordinates that define the structure of the molecule is $(3N-6)$, where N is the number of the participating atoms. The number 6 is subtracted from the total number of nuclear coordinates due to the fact that there are three translational and three rotational degrees of freedom that leave the energy unaltered.
As the variables bond length $(r_1, r_2, \dots, r_i)$, bond angles $(\alpha_1, \alpha_2, \dots, \alpha_{i-1})$ and torsional angles $(\theta_1, \theta_2, \dots, \theta_{i-2})$ in a molecule are changed, the energy of such molecule,

$$E = E(r_1, r_2, \dots, r_i, \alpha_1, \alpha_2, \dots, \alpha_{i-1}, \theta_1, \theta_2, \dots, \theta_{i-2})$$

(Eq. 87)

will be different. When the conformational space of structurally flexible molecules is evaluated, the *potential energy surface (PES)* is obtained. The resulting surface is actually a *hyperface (PEHS)*, since it depends on several variables, but it is usually referred to as PES. Every point on the surface represents one allocation of the nuclei, thus, one conformation. If the first derivative of an energy surface associated to a particular conformation along one of the dimensions is zero, this indicates a critical point. Furthermore, the second derivative by a variable can also be determined:

$$H_{ij} = \frac{\delta^2 E}{\delta x_i \delta y_i}$$

(Eq. 88)

where x_i and y_i are any of the internal coordinates of the molecule. This will represent one element of the Hess-matrix. If the matrix does not have any negative eigenvalues, the critical point is a minimum (zero-order critical point). If every eigenvalue in the matrix is negative, the critical point is a maximum. If there are both negative and positive eigenvalues, the critical point is a *saddle point*. Saddle points (and, in particular, first-order critical points) are important in determining the pathways for a conversion from a minimum to another one. Primary objectives in conformational studies, then, are the localization of minima and saddle points. When working with big molecules, several minima can be found: the point that presents the lower energy in the whole surface is called the *global minima*, while those points on the surface that present a lower energy value than the surrounding PES points but higher than the global minima are called *local minima* [75,110].

In **Fig 9** the representation of a general PES is presented:

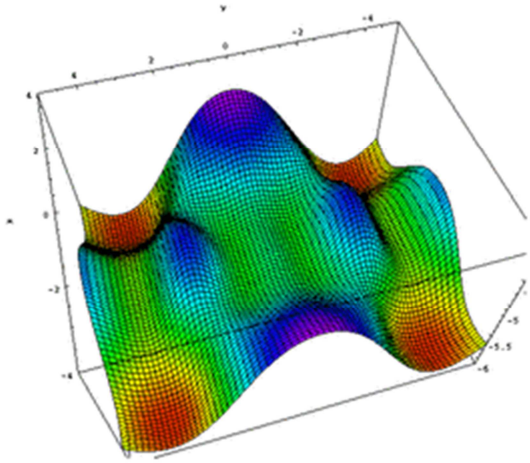


Fig. 9 – General form of a potential energy surface (PES) [111]

There are several mathematical procedures (algorithms) that allow the localization of local minima in a function dependent on several variables. These procedures will find a local minimum in a given PES in the vicinity of the initially considered geometry. The

procedure of finding such minimum is called geometry optimization or energy minimization. Given a molecule that presents several conformations, the procedure of finding the minimum must be repeated for each conformation, in order to find the global minimum. If the molecule is too big, there may be too many conformations, making the geometry optimization of all of them practically impossible. In such cases, alternative methods for sampling the PES should be considered [75].

2.3.9- Geometry optimization techniques

Geometry optimization methods start with an initial geometry and then change it until they find a lower-energy shape. This usually results in finding a local minimum of the energy, which corresponds to the conformer closest to the starting geometry. There are many different algorithms for finding the set of coordinates corresponding to the minimum energy, commonly known as *optimization algorithms* [100].
Since optimization problems in computational chemistry tend to have many variables, essentially all commonly used methods assume that, at least, the first derivative of the function with respect to all variables, the gradient **g**, can be calculated analytically. Some methods also assume that the second derivative matrix, the Hessian **H**, can be calculated. There are three classes of commonly used optimization methods for finding minima, each having their advantages and disadvantages [74].
In the next few lines, the different algorithms employed in this thesis work are briefly presented.

2.3.9.1- Steepest Descent

The gradient vector **g** points in the direction where the function increases most: the function value can always be lowered by stepping in the opposite direction. In the *Steepest Descent (SD)* method, a series of function evaluations are performed in the negative gradient direction, i.e. along a search direction defined as $d = -\mathbf{g}$**.** Once the function starts to increase, an approximate minimum may be determined by interpolation between the calculated points. At this interpolated point, a new gradient is calculated and used for the next line search [74].

If the line minimization is carried out sufficiently accurately, it will always lower the function value, therefore guaranteeing an approach to a minimum. This method has, however, two main problems: the first one lies in the fact that two subsequent line searches are necessarily perpendicular to each other; if there was a gradient component along the previous search direction, the energy could be further lowered in this direction. The steepest descent, therefore, has a tendency for each line search to partly spoil the function lowering obtained by the previous search. The steepest descent path oscillates around the minimum path, which is particularly problematic for surfaces having long narrow valleys. The other problem lies in the fact that as the minimum is approached, the rate of convergence slows down. This algorithm actually never reaches the minimum, it crawls towards it at an ever decreasing speed [74].

This is a very simple algorithm, and requires only storage of a gradient vector. It is one of the few methods that is guaranteed to lower the function value. The main use of this algorithm is to quickly relax a poor starting point, before some of the more advanced algorithms take over, or as a 'backup' algorithm if the more sophisticated methods are unable to lower the function value [74].

2.3.9.2- Conjugate Gradient methods

The *Conjugate Gradient (CG)* method tries to improve on one of the main problems of SD (the partial 'undoing' of the previous step) by performing each line search not along the current gradient but along a line that is constructed such that it is 'conjugate' to the previous search directions. If the surface is purely quadratic, the conjugate direction criterion guarantees that each successive minimization will not generate gradient components along any of the previous directions, and the minimum is effectively reached. The first step is equivalent to a steepest descent step, but subsequent searches are performed along a line formed as a mixture of the current negative gradient and the previous search direction.

$$d_i = -\mathbf{g}_i + \beta_i \mathbf{d}_{i-1}$$

(Eq. 89)

There are several ways of choosing the scaling factor β_i, and is in this choice that the different CG methods, such as *Fletcher-Reeves (FR)* and *Polack-Ribiere (PR)* among others, differ [74].

For non-quadratic surfaces the conjugate property does not hold rigorously, and the CG algorithm must often be restarted (by setting $\beta = 0$) during the optimization process. The conjugate property holds best for near-quadratic surfaces, and the convergence properties of CG methods can be improved by scaling the variables by a suitable pre-conditioner matrix, for example containing approximate inverse second derivatives [74]. Conjugate gradient methods have much better convergence characteristics than the steepest descent, but they require slightly more storage, since two vectors (current gradient and previous search direction) must be stored, but this is rarely a problem [74].

2.3.9.3- The Berny algorithm

The Berny algorithm is the default optimization algorithm included in the Gaussian09 program, and it was developed by Bernhard Schlegel. This algorithm uses the forces acting on the atoms of a given structure together with the second derivative matrix (the Hessian matrix) to predict energetically more favorable structures and thus optimize the molecular structure towards the next local minimum on the potential energy surface. Since an explicit calculation of the second derivative matrix is quite costly, this algorithm constructs an approximate Hessian at the beginning of the optimization procedure through application of a simple valence force field, and then uses the energies and first derivatives calculated along the optimization pathway to update this approximate Hessian matrix. The success of the optimization procedure therefore depends to some degree on how well the approximate Hessian represents the true situation at a given point. For many "normal" systems, the approximate Hessians work quite well, but in a few cases a better matrix has to be used. In such cases, it is often sufficient to calculate the second derivative matrix explicitly once at the beginning of the calculations, and then use the standard updating scheme of the Berny algorithm. In some very rare cases, the Hessian changes considerably between optimization steps, and it must be recomputed after each optimization step [112].

2.3.10- Molecular Dynamics simulations

During the last decades, increasing attention has been focused on the dynamic aspects of protein structure and function. It has long been inferred from a variety of experimental studies that substantial structural fluctuations occur in these molecules, and that these fluctuations are essential to biological activity. One of the most widely used methods to study atomic motion in biomolecules is Molecular Dynamics simulations (MD). These interactions then provide information on the structure, dynamics and thermodynamics of biological molecules and their complexes. In particular, detailed information regarding fluctuations and conformational changes in proteins and nucleic acids was obtained through MD studies [113,114].
Since the late 50s, when the MD method was first introduced by Alder and Wainwright [115,116], this method has evolved into an important and widely used theoretical tool that allows researchers in chemistry, physics and biology to model the detailed microscopic dynamical behavior of many different types of systems, including gases, liquids, solids, surfaces and clusters [117].

2.3.10.1- Solving the classical equations

In a MD simulation, the classical equations of motion governing the microscopic time evolution of a many-body system are solved numerically subject to boundary conditions appropriate for the geometry or symmetry of the system. Molecular Dynamics involves the calculation of the time dependent movement of each atom in a molecule, obtaining consecutive configurations for the system. The result is a trajectory (a series of time-correlated points in phase space), which specifies how the positions and velocities of the particles composing the system vary through time. This is achieved by solving Newton's equations of motion (in most cases, this is done numerically). For this process, the energy surface and the derivative of the energy in terms of the nuclear coordinates are required:

$$F = m.a = -\frac{dE}{dr} = m\frac{d^2r}{dt^2}$$

(Eq. 90)

$$\frac{d^2r_i}{dt^2} = a_i = \frac{F_i}{m_i};\ F_i = \frac{\delta E}{\delta r_i}$$

(Eq. 91)

where m is the mass, a is the acceleration, E represents the potential energy, and r and t represent the coordinates and time, respectively. The potential energy, as well as the coordinates, are determined using one of the empirical FF already described. Once the trajectories are obtained, the mean value of different molecular properties can be determined [95,117-119].

2.3.10.2- Temperature in Molecular Dynamics

Structures for starting geometries are normally sampled as a function of time or geometry during a molecular dynamics run of at the most a few nanoseconds. At a finite temperature, the average kinetic energy is directly related to the temperature, and the molecule(s) explores a part of the surface with energies lower than the typical kinetic energy. Therefore, molecular dynamics is efficient at exploring local conformational space but it is not effective at crossing large energy barriers, and so it is not suited for global searches. Given a high-enough energy, the dynamics will sample the whole surface, but this will also require an impractically long simulation time. Since quite small time steps must be used for integrating Newton's equation, the simulation time is short (pico- or nano-seconds). The maximum time step that can be considered in the simulation depends on the highest frequencies, usually R-H bonds. One of the main issues in dynamics simulations is that the movements leading to appreciable conformational changes usually have lower frequencies (milliseconds, for example in the torsions). Depending on the size of the molecule the computation of such long time intervals is usually prohibitive due to the CPU time and storage space involved [74,118].

To achieve faster and more complete molecular dynamics, searching at high temperatures can be used. However, sampling at temperatures that are too high (over 1000 K) is not constructive, since a large proportion of the resulting minima are high-energy conformers. In order to prevent the simulation from going to areas of the potential energy surface that have already been searched, a penalty can be assigned to sampled points, making the molecular dynamic search more global [118].

2.3.10.3- Molecular Dynamics methods

A dynamics simulation requires a set of initial coordinates and velocities, and an interaction potential (energy function). For a short time step, the interaction may be considered constant, allowing a set of updated positions and velocities to be estimated, at which point the new interaction can be calculated. By taking a (large) number of (small) time steps, the time behavior of the system can be obtained. Since the phase space is huge, and the fundamental time step is short, the simulation will only explore the region close to the starting point, and several different simulations with different starting conditions are required for estimating the stability of the results. Given the number of atoms usually involved the systems studied with Molecular Dynamics, several algorithms to solve Newton's equation numerically, such as the *Verlet* , the *velocity Verlet* and the *leap-frog* algorithms. All integration algorithms assume the positions, velocities and accelerations can be approximated by a Taylor series expansion.
Since the MD programs used in this thesis work (NAMD [120] and AMBER [121]) use the velocity verlet and the leap-frog algorithm respectively, a brief summary of both of this methods is included in the following lines.

2.3.10.3a- The velocity Verlet algorithm

The *velocity Verlet* algorithm computes the particle velocity and position at a time $t + \Delta t$ as follows:

$$x(t + \Delta t) = x(t) + v(t)\Delta t + \frac{1}{2}\frac{f(t)}{m}\Delta t^2$$

(Eq. 92)

$$v(t + \Delta t) = v(t) + \frac{f(t) + f(t + \Delta t)}{2m}\Delta t$$

(Eq. 93)

where Δt is a small time increment, m is the particle mass and $f(t)$ is the total force acting on a particle at time t. Given the initial conditions $x(0)$ and $v(0)$, $v(t)$ and $x(t)$ can be computed simply by applying **Eq. 90** and **Eq. 91** successively n times, with $n = {}^{t}/_{\Delta t}$ [122,123].

2.3.10.3b- The leap-frog algorithm

This algorithm is similar to the velocity Verlet algorithm, except it calculates the positions and velocities at interleaved time points, in such a way that they 'leapfrog' over each other. For example, the position is known at integer time steps and the velocity is known at integer plus half time steps. The equations can be written as:

$$x_{i+1} = x_i + v_i\Delta t + \frac{f(t)}{m}\frac{\Delta t^2}{2}$$

(Eq. 94)

$$v_{i+1} = v_i + \frac{f_i + f_{i+1}}{2m}\Delta t$$

(Eq. 95)

As in the previous case, Δt is a small time increment, m is the particle mass and $f(t)$ is the total force acting on a particle at time t. Given the initial conditions, the successive application of Eq. 92 and Eq. 93 allows to determine the position and velocity at any given time t [74,124].

2.4- Current state of the field

Given the important role that PTP1B plays not only in T2DM and obesity, but also in certain kind of cancers, it comes as no surprise that a large number of studies on this enzyme are being carried out at this moment. The large number of publications available to this day (by August 11th, 2011, 1759 papers concerning this enzyme can be found in the Science Direct database, 404 of which were published between 2010 and 2011 [125]) is a good testimony of that. Most of these works are focused on the evaluation of possible novel inhibitors for PTP1B, which, in most cases, are derivates from molecules with proven biological activity, either on PTP1B itself or other similar proteins [126-132]. Some Molecular Dynamics simulations have been carried out throughout time, both for the complexed and uncomplexed protein [20,133].

This thesis work is still novel, due, mostly, to the simplicity of the complexes considered. The fact that the considered "ligands" are halide ions allows the analysis to focus exclusively in the conformational changes the protein suffers upon complex formation, without including any misleading secondary interactions that can arise when a more complex is considered.

REFERENCES

[1] Puius, Y. A.,Zhao, Y.*et al.*, *Proc. Natl. Acad. Sci. USA 94*, **1997,** 13420-13425.

[2] Burke, T. R.and Zhang, Z.-Y., *Biopolymers (Peptide Science) 47*, **1998,** 225-241.

[3] Liu, G.-X.,Tan, J.-Z.*et al.*, *Acta Pharmacol. Sin. 27*, *1*, **2006,** 100-110.

[4] Hunter, T., *Curr. Opin. Cell Biol. 21*, **2009,** 140-146.

[5] Barford, D.,K.Das, A.and Egloff, M.-P., *Annu. Rev. Biophys. Biomol. Struct. 27*, **1998,** 133-164.

[6] Juszczak, L. J.,Zhang, Z.-y.,Wu, L.,Gottfried, D. S.and Eads, D. D., *Biochemistry 36*, **1997,** 2227-2236.

[7] Zhang, Z.-Y., *Curr. Opin. Chem. Biol. 5*, **2001,** 416-423.

[8] Hunter, T., *Cell 80*, *2*, **1995,** 225-263.

[9] Barford, D.,Flint, A. J.and Tonks, N. K., *Science 263*, **1994,** 1397-1404.

[10] Vintonyak, V. V.,Antonchick, A. P.,Rauh, D.and Waldmann, H., *Curr. Opin. Chem. Biol. 13*, **2009,** 272-283.

[11] Wrobel, J.,Sredy, J.*et al.*, *J. Med. Chem 42*, **1999,** 3199-3202.

[12] Johnson, T. O.,Ermolieff, J.and Jirousek, M. R., *Nat. Rev. Drug Discovery 1*, **2002,** 696-709.

[13] Sarmiento, M.,Wu, L.*et al.*, *J. Med. Chem 43*, **2000,** 146-155.

[14] Elchebly, M.,Payette, P.*et al.*, *Science 283*, *1544-1548*, **1999,**

[15] Klaman, L. D.,Boss, O.*et al.*, *Mol. Cell. Biol. 20*, *15*, **2000,** 5479-5489.

[16] Shen, K.,Keng, Y.-F.*et al.*, *J. Biol. Chem. 276*, *50*, **2001,** 47311-47319.

[17] Aoyama, H.,Silva, T. M. A.,Miranda, M. A.and Ferreira, C. V., *Quim. Nova 26*, *6*, **2003,** 896-900.

[18] Li, L.and Dixon, J. E., *Semm. Immun. 12*, **2000,** 75-84.

[19] Li, S.,Depetris, R. S.,Barford, D.,Chernoff, J.and Hubbard, S. R., *Structure 13*, **2005,** 1643-1651.

[20] Kamerlin, S. C. L.,Rucker, R.and Boresch, S., *Biochem. Biophys. Res. Commun. 356*, **2007,** 1011-1016.

[21] Yip, S.-C.,Saha, S.and Chernoff, J., *Trends Biochem. Sci. 35*, *8*, **2010,** 442-449.

[22] Pannifer, A. D. B.,Flint, A. J.,K.Tonks, N.and Barford, D., *J. Biol. Chem. 273*, *17*, **1998,** 10454-10462.

[23] Seiner, D. R.,LaButti, J. N.and Gates, K. S., *Chem. Res. Toxicol. 20*, *9*, **2007,** 1315-1320.

[24] Liu, G.,Szczepankiewicz, B. G.*et al.*, *J. Med. Chem 46*, **2003,** 2093-2103.

[25] Bleasdale, J. E.,Ogg, D.*et al.*, *Biochemistry 40*, *19*, **2001,** 5642-5654.

[26] Groves, M. R.,Yao, Z.-J.,Roller, P. P.,Burke, T. R.and Barford, D., *Biochemistry 37*, *51*, **1998,** 17773-17783.

[27] Pedersen, A. K.,Peters, G. H.,Møller, K. B.,Iversen, L. F.and Kastrup, J. S., *Acta Cryst. D 60*, **2004,** 1527-1534.

[28] Barford, D.,Keller, J. C.,Flint, A. J.and Tonks, N. K., *J. Mol. Biol 239*, **1994,** 726-730.

[29] Heyda, J.,Hrobárik, T.and Jungwirth, P., *J. Phys. Chem. A 113*, **2009,** 1969-1995.

[30] Joung, I. S.and Cheatham, T. E., *J. Phys. Chem. B 112*, *30*, **2008,** 9020-9041.

[31] Laage, D.and Hynes, J. T., *Proc. Natl. Acad. Sci. USA 104*, *27*, **2007,** 11167-11172.

[32] Auffinger, P.,Bielecki, L.and Westhof, E., *Structure 12*, *3*, **2004,** 379-388.

[33] Feig, M.and Pettitt, B. M., *Biophys. J. 77*, *4*, **1999,** 1769-1871.

[34] Lu, Y.,Shi, T.*et al.*, *J. Med. Chem. 52*, *9*, **2009,** 2854-2862.

[35] Hömberg, A.,Hultdin, U. W.,Olofsson, A.and Sauer-Eriksson, A. E., *Biochemistry 44*, *26*, **2005,** 9290-9299.

[36] Auffinger, P.,Hays, F. A.,Weshof, E.and Ho, P. S., *Proc. Natl. Acad. Sci. USA 101*, *48*, **2004,** 16789-16794.

[37] Auffinger, P.,Bielecki, L.and Weshof, E., *Structure 12*, **2004,** 379-388.

[38] Azarani, A.,Segelke, B. W.,Toppani, D.and Lekin, T., *JALA 11*, **2006,** 7-15.

[39] Deschamps, J. R., *Life Sci. 86*, **2010,** 585-589.

[40] Pusey, M. L.,Liu, Z.-J.*et al.*, *Prog. Biophys. Mol. Biol. 88*, **2005,** 359-386.

[41] Zanotti, G. Protein crystallography, in *Fundamentals of crystallography, 2nd. Ed.*; Giacovazzo, C., Ed.; Oxford University Press, 2002.

[42] Dauter, Z., *Acta Cryst. D 62*, **2006,** 1-11.

[43] Berman, H. M., *Acta Cryst. A 62*, **2008,** 88-89.

[44] *RCSB PDB Protein Data Bank*; http://www.pdb.org/pdb/home/home.do; August 15th.,2011

[45] Rhodes, G. *Crystallography made crystal clear, 3rd Ed.*; (Academic Press - Elsevier, 2006).

[46] McPherson, A. *Introduction to macromolecular crystallography, 2nd. Ed.*; (Wiley-Blackwell, 2009).

[47] *Protein crystallography course*; http://www-structmed.cimr.cam.ac.uk/Course/Overview/Overview.html; August 16th.,2011

[48] Mikol, V.,Hirsch, E.and Giegé, R., *J. Mol. Biol 123*, **1990,** 187-195.

[49] *Crystallization of nucleic acids and proteins, 2nd. Ed.*; Ducruix, A.and Giegé, R., Eds.; (Oxford University Press, 1999).

[50] Bolanos-Garcia, V. M.and Chayen, N. E., *Prog. Biophys. Mol. Biol. 101*, **2009,** 3-12.

[51] Chernov, A. A., *J. Struct. Biol. 142*, *1*, **2003,** 3-21.

[52] Smatanová, I. K., *Mat. Struct. 9*, *1*, **2002,** 14-15.

[53] Rondeau, J.-M.,Klebe, G.and Podjarny, A. Ligand binding: the crystallographic approach, in *Biophysical approaches determining ligand binding to biomolecular targets - Detection, measurement and modelling*; Podjarny, A., Dejaegere, A. P., Kieffer, B., Eds.; RSC Biomolecular Sciences, 2011.

[54] Smith, J. F., *J. Phase. Equilib. Difuss. 25*, *6*, **2004,** 497-506.

[55] Green, D. W.,Ingram, V. M.and Perutz, M. F., *Proc. R. Soc. London Ser. A 225*, **1954,** 287-307.

[56] Rossman, M. G., *Acta Cryst. 13*, **1960,** 221-226.

[57] *Chemical crystallography lab*; http://xrayweb.chem.ou.edu/notes/crystallography.html; August 17th.,2011

[58] *Friedel's law*; http://reference.iucr.org/dictionary/Friedel's_law; August 17th.,2001

[59] Drenth, J. *Principles of protein x-ray crystallography, 3rd Ed*; (Springer, 2007).

[60] Adams, P. D.,Afonine, P. V.*et al.*, *Curr. Opin. Struct. Biol. 19*, **2009,** 566-572.

[61] Greer, J., *J. Mol. Biol 100*, **1976,** 427-458.

[62] Richards, F. M., *J. Mol. Biol 37*, **1968,** 225-230.

[63] Tronrud, D. E., *Acta Cryst. D 60*, **2004,** 2156-2168.

[64] Messerschmidt, A. *X-ray crystallography of biomacromolecules - A practical guide*; (Wiley-VCH Verlag GmbH & Co.: Weinheim, 2007).

[65] Kleywegt, G. J., *Acta Cryst. D 56*, **2000,** 249-265.

[66] Bernstein, F. C.,Koetzle, T. F.*et al.*, *Eur. J. Biochem. 80*, **1977,** 319-324.

[67] Read, R. J., *Acta Cryst. A 46*, **1990,** 900-912.

[68] Pannu, N. S.and Read, R. J., *Acta Cryst. A 52*, **1996,** 659-668.

[69] Emsley, P.,Lohkamp, B.,Scott, W. G.and Cowtan, K., *Acta Cryst. D 66*, **2010,** 486-501.

[70] *R-factor calculations and their significance*; http://bass.bio.uci.edu/~hudel/mbb254/lecture2/lecture2_1.html; August 20th.,2011

[71] Hann, M.and Green, R., *Curr. Opin. Chem. Biol. 3*, **1999,** 379-383.

[72] Agrafiotis, D. K.,Bandyopadhyay, D.,Wegner, J. K.and Vlijmen, H. v., *J. Chem. Inf. Model. 47*, **2007,** 1279-1293.

[73] Willet, P., *Wiley Interdiscip. Rev. Comput. Mol. Sci. 1*, *1*, **2011,** 46-56.

[74] Jensen, F. *Introduction to computational chemistry, 2nd. Ed.*; (Wiley, 2007).

[75] Levine, I. N. *Quantum chemistry, 5th. Ed.*; (Prentice Hall, 1999).

[76] Simmons, J.and Nichols, J.in *Quantum mechanics in chemistry, 1st. Ed.*; Oxford University Press, 1997.

[77] Naundorf, H., *Short introduction to quantum chemistry methods,* 2005.

[78] *Simplified introduction to ab initio basis sets. Terms and notations.*; http://www.ccl.net/cca/documents/basis-sets/basis.html; September 4th.,2011

[79] Boys, S. F.and Bernard, F., *Mol. Phys. 19*, *4*, **1970,** 553-566.

[80] Mueller, M. *Fundamentals of Quantum Chemistry. Molecular spectroscopy and modern electronic structure computations*; (Kluwer Academic Publishers, 2001).

[81] Walczak, K.,Friedrich, J.and Dolg, M., *Chem. Phys. 365*, **2009,** 38-43.

[82] Galano, A.and Alvarez-Idaboy, J. R., *J. Comput. Chem. 27*, *11*, **2006,** 1203-1210.

[83] *Introduction to electron correlation*; http://vergil.chemistry.gatech.edu/courses/chem6485/pdf/intro-e-correlation.pdf; August 5th.,2011

[84] *Møller-Plesset perturbation theory*; http://www.cem.msu.edu/~cem883/topics_pdf_2008/Moller-Plesset.pdf; August 5th.,2011

[85] Parr, R. G.and Yang, W.; Oxford Science Publications, 1989.

[86] Hohenberg, P.and Kohn, W., *Phys. Rev. 136*, **1964,** B864-B871.

[87] Kohn, W.and Sham, L. J., *Phys. Rev. 140*, *4A*, **1965,** A1133-A1138.

[88] Vosko, S. H.,Wilk, L.and Nusair, M., *Can. J. Phys. 58*, *8*, **1980,** 1200-1211.

[89] Perdew, J. P.,Burke, K.and Ernzerhof, M., *Phys. Rev. Lett. 77*, **1996,** 3865-3865.

[90] Becke, A. D., *J. Chem. Phys 107*, **1997,** 8554-8560.

[91] Zhao, Y.and Truhlar, D. G., *Theor. Chem. Acc 120*, **2008,** 214-241.

[92] Bartlett, R. J.,Lotrich, V. F.and Schweigert, I. V., *J. Chem. Phys. 123*, *6*, **2005,** 062205.

[93] MacKerell, A. D. Atomistic models in force fields, in *Computational biochemistry and biophysics*; Becker, O. M., MacKerell, A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker, Inc.: New York, 2001.

[94] Rappé, A. K.and Casewit, C. J.; University Science Books, 1997.

[95] Leach, A. R. *Molecular modelling. Principles and applications. 2nd Ed.*; (Pearson Education Limited, 2001).

[96] Leach, A. R. *Molecular Modelling - principles and applications, 2nd. Ed.*; (Prentice Hall: Harlow, 1996).

[97] Sarzyńska, J.,Kulińska, K.and Kuliński, T., *Comp. Meth. Sci. Tech. 9*, *1-2*, **2003,** 127-135.

[98] Morse, P. M., *Phys. Rev. 34*, **1929,** 57-64.

[99] Ponder, J. W.and Case, D. A., *Adv. Protein. Chem. 66*, **2003,** 27-86.

[100] Young, D. C. *Computational chemistry: a practical guide for applying techniques to real-world problems*; (John Wiley and Sons, Inc, 2001).

[101] Prabhakar, R.,Musaev, D. G.,Khavrutskii, I. V.and Morokuma, K., *J. Phys. Chem. B 108*, **2004,** 12643-12645.

[102] Vreven, T.and Morokuma, K. Hybrid mehtods: ONIOM (QM:MM) and QM/MM, in *Annual reports in computational chemistry, Volume 2*; Elsevier B.V., 2006.

[103] Vreven, T.,Morokuma, K.,Farkas, Ö.,Schlegel, H. B.and Frisch, M. J., *J. Comput. Chem. 24*, *6*, **2003,** 760-769.

[104] Vreven, T.,Byun, K. S.*et al.*, *J. Chem. Theory Comput. 2*, *3*, **2006,** 815-826.

[105] Vreven, T.and Morokuma, K., *Theor. Chem. Acc. 109*, **2003,** 125-132.

[106] Dapprich, S.,Komáromi, I.,Byun, K. S.,Morokuma, K.and Frisch, M. J., *J. Mol. Struct. (THEOCHEM) 461-462*, **1999,** 1-21.

[107] Svensson, M.,Humbel, S.*et al.*, *J. Phys. Chem. 100*, **1996,** 19357-19363.

[108] Cui, Q.and Karplus, M., *J. Phys. Chem. B 104*, **2000,** 3721-3743.

[109] Karadakov, P. B.and Morokuma, K., *Phys. Lett. 317*, *6*, **2000,** 589-596.

[110] Láng, A.,Füzéry, A. K.,Beke, T.,Hudáky, P.and Perczel, A., *J. Mol. Struct. (THEOCHEM) 675*, **2004,** 163-175.

[111] *Research area: surfaces and interfaces*; http://cst-www.nrl.navy.mil/ResearchAreas/SurfacesAndInterfaces/; August 11th.,2011

[112] *Electronic structure calculations in Gaussian*; http://www.chem.cornell.edu/dbc6/documents/Gaussian_optimization.pdf; September 20th.,2011

[113] *Theory of molecular dynamics simulations*; http://www.ch.embnet.org/MD_tutorial/pages/MD.Part1.html; September 11th.,2011

[114] McCammon, J. A.and Harvey, S. C. *Dynamics of proteins and nucleic acids*; (Cambridge University Press, 1987).

[115] Alder, B. J.and Wainwright, T. E., *J. Chem. Phys. 27*, **1957,** 1208-1209.

[116] Alder, B. J.and Wainwright, T. E., *J. Chem. Phys. 31*, *459-466*, **1959,**

[117] Tuckerman, M. E.and Martyna, G. J., *J. Phys. Chem. B 104*, **2000,** 159-178.

[118] Comba, P.and Hambley, T. W. *Molecular modeling of inorganic compounds, 2nd. Ed.*; (Wiley-VCH, 2001).

[119] Phillips, J. C.,Braun, R.*et al.*, *J. Comput. Chem. 26*, *16*, **2005,** 1781-1802.

[120] Phillips, J. C.,Braun, R.*et al.*, *J. Comput. Chem. 26*, **2005,** 1781-1802.

[121] Case, D. A.,Cheatham, T. E.*et al.*, *J. Comput. Chem. 26*, *16*, **2005,** 1668-1668.

[122] *Velocity verlet algorithm* http://xbeams.chem.yale.edu/~batista/vaa/node60.html; September 20th.,2011

[123] Swope, W. C.,Andersen, H. C.,Berens, P. H.and Wilson, K. R., *J. Chem. Phys. 76*, *1*, **1982,** 637-651.

[124] Skeel, R. D. Integration schemes for Molecular Dynamics and related applications, in *The graduate student's guide to numerical analysis (SSCM)*; Ainsworth, M., Levesley, J., Marletta, M., Eds.; Springer-Verlag, 1999; p 119-176.

[125] *Science Direct*; http://www.sciencedirect.com/; August 11th.,2011

[126] Thareja, S.,Aggarwal, S.,Bhardwaj, T. R.and Kumar, M., *Eur. J. Med. Chem. 45*, **2010,** 2537-2546.

[127] Bhattarai, B. R.,Kafle, B.*et al.*, *Bioorg. Med. Chem. Lett. 20*, **2010,** 6758-6763.

[128] Nguyen, P. H.,Dao, T. T.*et al.*, *Bioorg. Med. Chem. 19*, **2011,** 3378-3383.

[129] Cheng, Y.,Zhou, M.,Tung, C.-H.,Ji, M.and Zhang, F., *Bioorg. Med. Chem. Lett. 20*, **2010,** 3329-3337.

[130] Tong, Y. F.,Zhang, P.*et al.*, *Chin. Chem. Lett. 21*, **2010,** 1415-1418.

[131] Song, Z.,He, X.-P.*et al.*, *Carbohydr. Res. 346*, **2011,** 140-145.

[132] Sun, L.-P.,Shen, Q.*et al.*, *Eur. J. Med. Chem. 46*, **2011,** 3630-3638.

[133] Kamerlin, S. C. L.,Rucker, R.and Boresch, S., *Biochem. Biophys. Res. Commun. 345*, **2006,** 1161-1166.

CHAPTER 3

OBJECTIVES AND STRATEGIES

The present work aims to achieve a deeper understanding of the structure, movement and flexibility of PTP1B's WPD loop, entailing the possibility, in the future, of designing specific high affinity and efficiency PTP1B inhibitors. This would mean an important breakthrough not only in the therapeutic possibilities for the T2DM treatment, but also a deeper understanding of the enzymatic mechanism of the PTPs in general, given the existing similarities in this aspect between the diverse enzymes that constitute this family. Moreover, this could open the doors to the future development of specific therapeutic agents for the diverse affections derived from the inefficient or unsuitable operation of the enzymatic network carried out by the PTKs and PTPs.

Even though the most effective approach to the synthesis of a high affinity inhibitor focuses on the enzyme's active site, the important conservation of the WPD loop sequence among different PTPs generates the need to search for new strategies to obtain greater specificity. The fact that some residues of the WPD loop presents certain variability between PTPs, and the inherent flexibility of this loop, make it an interesting target for designing efficient and selective inhibitors.

3.1- GENERAL OBJECTIVE

This work focuses on the study of loop WPD and the effect that different halide ions have on its molecular conformation and its general behaviour. In order to achieve this, the structures of the tyrosine phosphatase PTP1B in complex with different halide ions (Br^-, I^-) are refined, considering the two possible conformation of the flexible loop. After that, theoretical tools of molecular modelling and molecular dynamics simulations are used, in order to obtain a correlation between the type and occupation of the halide ion in the complex with the relative occupation of the loop of interest.

3.2- SPECIFIC OBJECTIVES

3.2.1- Crystallographic Objectives

- Obtainment and crystallization of the PTP1B-halide ion complexes, in various conditions.
- Measurement of the different complexes in synchrotron radiation sources.
- Obtainment and refinement of molecular models for the PTP1B-halide ion complexes, incorporating two copies of the WPD loop (an open and a closed one).
- Determination of the relative occupations of each copy, as well as the occupation of the halide ions.

3.2.2- Molecular Modeling Objectives

- Obtainment and optimization of the appropriate Molecular Mechanics parameters required to model the halide ions of interest.
- Validation of the chosen Molecular Mechanics method.
- Determination of the region that needs to be studied at a high theory level.
- Geometry optimization of the protein- halide ion complex model.
- Study of the global movement of the PTP1B, and, in particular, the flexible WPD loop, using Molecular Dynamics.

CHAPTER 4

RESULTS AND DISCUSSION-

4.1- EXPERIMENTAL RESULTS

4.1.1- Crystallization of PTP1B-ion complexes and measurement of X-ray diffraction data

A series of different ions were tested, as well as different crystallization conditions, varying both the concentration of the ions and the concentration of the acetate buffer solution (the detailed crystallization procedure is specified in Chapter 6). The different complexes conditions are detailed in **Table 1**.

Data set	ION	[Ion] (mM)	[Acetate] (mM)	Space group
1	Br^-	200	20	$P\,3_121$
2	Br^-	25	25	$P\,3_121$
3	Br^-	25	200	$P\,3_121$
4	Br^-	150	200	$P\,3_121$
5	Br^-	100	300	$P\,3_121$
6	Br^-	150	300	$P\,3_121$
7	$Br\text{-}SO_3^-$ (as $Br\text{-}CH_2\text{-}CH_2\text{-}SO3^-$)	150	25	$P\,3_121$
8	I^-	25	50	$P\,3_121$
9	I^-	150	50	$P\,3_121$
10	I^-	25	200	$P\,3_121$
11	I^-	100	200	$P\,3_121$

Table 1- Different conditions tested for the co-crystallization of PTP1B with ions.

It is important to notice that the case of data set 7 is the only one in which the halide was not complexed with the protein as the single atom, but in a molecular form. As it can be seen from the previous table, all crystals obtained presented a trigonal $P3_121$ space group. This is consistent with the previously reported crystallographic data for PTP1B [1].

In **Table 2** the unit cell parameters for each obtained crystal are detailed.

Data set	a	b	c	α	β	γ
1	88.4	88.4	103.8	90.0	90.0	120.0
2	88.4	88.4	104.2	90.0	90.0	120.0
3	88.6	88.6	104.5	90.0	90.0	120.0
4	88.5	88.5	104.4	90.0	90.0	120.0
5	88.5	88.5	104.2	90.0	90.0	120.0
6	88.4	88.4	104.0	90.0	90.0	120.0
7	88.5	88.5	104.5	90.0	90.0	120.0
8	89.1	89.1	104.5	90.0	90.0	120.0
9	88.3	88.3	103.9	90.0	90.0	120.0
10	88.0	88.0	103.6	90.0	90.0	120.0
11	88.1	88.1	104.2	90.0	90.0	120.0

Table 2- Unit cell parameters for the different crystals.

The data collection characteristics in each case are depicted in **Table 3**, allowing us to assess the data collection strategy. In every case, statistics indicate that the data collected meet the quality requirements essential to obtain a reliable macromolecular model.

Data set	Resolution (in Å)	Measurements. (N)	Completeness (%)	R_{symm}	I/σ(I)	Redundancy
1	35 – 1.85	38552	99.5	4.2	35.24	6.23
2	35 – 1.71	47461	97.5	3.9	49.04	5.94
3	35 – 1.75	45631	99.7	5.3	28.67	6.04
4	35 – 1.85	38210	98.5	4.1	32.54	5.89
5	35 – 1.85	37978	98.3	3.8	31.57	6.30
6	35 – 1.71	47684	98.2	3.8	42.89	12.28
7	35 – 1.70	49561	99.5	3.8	38.22	6.25
8	35 – 1.80	41433	97.1	5.8	32.86	6.62
9	35 – 1.69	49662	99.4	3.7	34.12	6.24
10	35 – 2.00	30122	99.7	6.9	23.95	9.75
11	35 – 1.80	40002	95.8	5.8	37.20	7.43

Table 3- Data collection and processing statistics.

4.1.2- Crystal structure determination

The structure in each case was determined by molecular replacement, using the 1AAX pdb deposition in the RCSB PDB [2], and optimized employing a model which already included a double conformation for the WPD loop (the software employed to do so, as well as the conditions considered are described in Chapter 6).

Data set	Resolution range (in Å)	R value (%)	R_{free} value (%)
1	1.89 - 1.85	20	22
2	1.75 – 1.71	22	25
3	1.80 – 1.75	21	24
4	1.90 – 1.85	21	24
5	1.90 – 1.85	22	25
6	1.76 – 1.71	22	26
7	1.74 – 1.70	24	26
8	1.85 – 1.80	22	26
9	1.74 – 1.69	22	25
10	2.05 – 2.00	22	25
11	1.84 – 1.80	23	27

Table 4- Refinement statistics

The R and R_{free} values shown in **Table 4** indicate that the refined structures in each case are correct (R~20% and R_{free} is in the accepted range of 2-8% over R value) [3-5].
The obtained structures for the active site of each data set (with both possible conformations) and the corresponding $2F_o - F_c$ maps are shown in Appendix A.

4.1.3- Refinement of the flexible WPD loop occupation in each crystal

Having obtained an acceptable structure for each complex, the next step was to refine the atomic positions for the atoms of residues 177 to 189 of the protein, in order to determine which conformation of the WPD loop was present in each case.
The Debye-Weller factor (also known as 'B factor' or atomic temperature factor) is usually regarded as a measure of how much an atom oscillates or vibrates around a position. When considering atoms with a low degree of freedom (such as those that are far away from the N- and C-terminus of the protein), this value should be relatively low. As a general rule, a B-factor lower than 30 $Å^2$ (indicating a displacement of approximately 0.62 Å) for a given atom indicates confidence in its position, while a B-factor over 60 Å (a

displacement of as much as 0.87 Å) indicates a less trustable position [6,7] . Of course, these are not absolute values, and the mean values for B in a given protein should be taken into account. In a refinement with fixed occupancy values there is an inverse correlation between the B factor and the real occupancy of a given atom. Therefore, in this case a high B-factor can be interpreted as indicating a low occupancy [8].

Data set	B values (in Å^2)				
	Residues 180 to 183		**Ion**	**Base line estimation**	
	Open conformation	Closed conformation		Residues 1 to 20	Residues 120 to 139
1	38.8	*12.1*	**34.8**	36.7	23.0
2	29.0	*17.4*	**62.1**	36.6	22.2
3	23.4	22.1	159.6*	38.2	21.8
4	21.9	27.0	204.0*	44.0	23.4
5	22.4	33.1	101.2*	47.8	24.1
6	30.4	*17.0*	**35.4**	39.2	24.4
7	*26.2*	32.2	137.0*	26.2	46.5
8	40.5	*19.7*	**27.6**	45.0	30.9
9	31.8	*10.1*	**14.7**	32.4	21.2
10	41.4	*22.8*	**37.6**	50.0	33.1
11	33.6	*11.1*	**13.8**	32.4	21.0

Table 5- Summary of B factor values for the WPD loop (residues 180 to 183) for the different crystal complexes

* These high B-values indicate that the ion has a very low occupancy

In **Table 5** the different B values for some residues of interest are shown: the first two rows represent the temperature values for residues 180 to 183 (the most flexible residues of the WPD loop) for the two conformations considered; in the third place the motility of the ion in each complex is represented. The cases where the B-factors indicate a clear preference for the closed WPD loop conformation are written in italics. The cases where the B-factor indicate significant occupancy for the ions are written in bold characters.

The last two rows are intended as an estimation of a base line in each complex: the first 20 residues are in the N-terminal segment of the chain, and for that they present the highest B values for the polypeptide chain, giving a comparing value as to what should be considered 'high' B values. Residues 120 to 139 are buried into the three-dimensional structure of the protein, which constraints their movement, and that's why they are considered as the base-line for 'low' B values.

4.1.3.1- Analysis of case 1

There is a clear preference for the WPD loop for the ligand-bound conformation: not only the B values for the open conformation are in the same range that those shown for the more disordered regions of the protein, but the values for the closed conformation are notoriously lower, being even quite below the base line for the ordered residues. As for the ion's occupancy, it presents a B value of around 35 $Å^2$. Even when this would represent a low value for a protein's region, it is important to take into account the fact that it is an unbound atom, therefore intrinsically more mobile. That is, in the particular case of a single atom, this can be considered to represent an atom with high occupancy, and present in most of the complex copies present in the crystal.

4.1.3.2- Analysis of case 2

Similarly to case 1, the closed conformation appears to be the preferred structure for the complex. In this case, the difference in occupancy between both possibilities is not as big as it was in the previous set. Even more, the values obtained for the open conformation are below the disordered level calculated for the same structure. At the same time, the ion's occupancy appears to be lower than in the previous case (the temperature factor for the bromide in this case is around 60 $Å^2$). This could be interpreted as a prevalence of the ligand-bound conformation for the flexible loop, but not as big as in data set **1**.

4.1.3.3- Analysis of case 3

When considering data set **3**, the resulting crystal presents both loops conformations in basically the same proportions (the B values for both of them are almost identical). Moreover, the values are really close to those presented by the highly ordered residues. The ion presents a very high B value (almost 160 $Å^2$), indicating a low occupancy.

4.1.3.4- Analysis of case 4

When considering the crystal conditions depicted in data set **4**, the situation is similar to the previous one, but in this case the so-called equilibrium between both conformations is slightly displaced towards the unbounded conformation. Albeit they are both close to the values expected for ordered regions, the difference between both possible conformations is augmented. Sustaining this impression is the fact that the B value for the ion is even higher, reaching 200 $Å^2$, thus indicating a low occupancy.

4.1.3.5- Analysis of case 5

Data set **5** is the first case in which a prevalence of the open conformation can be seen, although it is not a particularly clear one. This consideration is based in the difference between the temperature factors for each possible conformation, which amounts to about 10 $Å^2$ (even when it is a significant difference, it is smaller than the difference between the B factors for the ordered and disordered regions in the same model). In any case, both of them are quite below the base line for disordered regions, indicating that both of them appear to be present in the crystal. The ion shows a low occupancy, but significantly higher than in the prior two cases.

4.1.3.6- Analysis of case 6

When analyzing the model obtained for data set **6**, once again a clear dominance of the ligand-bound structure for the WPD loop can be found. The temperature factor for this

arrangement is even below the one shown for the well-ordered regions of the protein while the unbounded structure shows a B factor somewhere in between the well-organized and the highly mobile regions. Once again, the ion shows a low B value, indication of a high occupancy. This is coherent with what expected, since seem to be more copies of the macromolecule forming a complex with the halide than in a free form.

4.1.3.7- Analysis of case 7

The following crystal (**7**) is particular, since the halide was incorporated in the complex not by itself, but as part of a bigger molecule (bromoethanesulphonate). The diffraction data in this case suggest that both possible spatial arrangements for the flexible loop of the protein are present in similar proportions, since the B values they present are not only close to each other, but closer to the expected value for well-ordered regions, particularly in the case of the unbounded conformation. The ligand shows a high B factor, implying a low rate of complex formation.

4.1.3.8- Analysis of case 8

When considering the complexes with iodide, the first scenario is the crystal corresponding to data set **8**. The ligand bound conformation of the WPD loop appears to be clearly favored, since the closed conformation presents a B value notoriously beneath the 'basis line' for ordered residues, while the corresponding values for the open conformation are in the vicinity of the disordered regions of the protein. As for the iodide, it shows a high occupancy (the B factor presents a value beneath any of those achieved by the bromine ion). All of this would indicate that the PTP1B-ion complex is more present in the crystal than the unbounded protein.

4.1.3.9- Analysis of case 9

When the ion concentration is augmented versus the buffer concentration, as in data set **9**, the ligand-free conformation is clearly favored compared to the alternative one: in the

former, the temperature factor is half the one presented by the highly ordered residues of the protein, while the latter presents a value in the close proximity to the values shown by the disordered regions. In this case, the halide presents a very high occupancy, since the B value for this atom is even below the base line considered for the highly occupied regions. Even though the iodide is bigger than bromide, the size of the catalytic cleft is enough to allow its mobility, hence this low B factor value can be considered as a sign of an important presence of this ion in this specific position in the crystals. This would mean that a large percentage of the protein molecules are complexed with the halide.

4.1.3.10- Analysis of case 10

In the case of the crystal that originated diffraction data set **10**, the WPD conformation corresponding to the ligand-bound structure of PTP1B is preferred over the open conformation, but, once again, the B factor values for the latter are beneath the upper limit established by the more disordered regions of the structure. As for the ion's occupancy, it shows a low B value, but not as low as the previous data set. Overall, the temperature factors found for the optimized structures would be an indication of a crystal with a high percentage of complex, and very low presence of unbound protein.

4.1.3.11- Analysis of case 11

Finally, the last data set (**11**) presents characteristics quite similar to those of data set **9**: the closed conformation shows a higher occupancy than the residues buried in the three-dimensional structure of the protein is obviously preferred over the open one, while the B values for the ligand-free conformation are in the same range than those for the N-terminal residues. Again, the halide shows a B factor value extremely low, indicating a high prevalence of the desired complex over the free protein in this crystal.

Summarizing the previous analysis,it is easy to notice that in the iodide-PTP1B complexes, the WPD loop is always found in the conformation that the protein adopts upon ligand or inhibitor binding, commonly referred to as closed conformation in the consulted bibliography. The temperature factors for the adopted conformation are always

significantly lower than the base line considered, and in most cases they are even lower than what is considered as an acceptable value for amino acidic residues, indicating that these residues are well ordered in that conformation.

4.1.3.12- Correlation of B-factors of the flexible loop and the ions

Another interesting aspect of the collected data is the correlation between the B values of the flexible loop and the corresponding ions. In all cases, the ions show higher motility than the residues considered, but that is actually what is expected, since there are no covalent interactions linking the halide into a specific position, providing it with a higher oscillation capability. Nonetheless, it is interesting to note that a correlation can be found between the conformation adopted by the loop and the variability in the ion's position.

4.1.3.13- Bromide complexes

In the case of the bromide complexes, both modeled conformations can be found. In the complexes where the open conformation is observed, or even in the case where it is not clear whether one conformation is predominant over the other one, the ions show a lower occupancy (even though the signal in the X-ray data was clear enough as to place the atom accurately).On the other hand, the ions interacting with a closed conformation seem to be less mobile. In particular, data set **7** is a very interesting case to analyze, given the particularity of the ligand used to form the complex (compared to the other crystals), which also places three oxygen atoms within the active site, mimicking, in some way, the natural ligand for the protein. In this circumstance, all the oxygen atoms are oriented in such a way that they form several hydrogen bonds with some main-chain NH groups, as well as the side chains of Asp181, Arg221 and Gln262, as expected not only from the catalysis mechanism [9], but also from a previously reported work were a complex between PTP1B and vanadate was reported [10]. These extra interactions should stabilize the resulting complex, so what is expected is to find a ligand-bound conformation for the WPD loop, with an elevated occupancy-factor for the ligand, but that is not the case. The diffraction data collected showed, instead, a preference for the unbounded conformation

of the loop (even though the difference is not as remarkable as in some other cases, and the B factor for the closed conformation is below the base line established for the disordered regions), and, more important, a very low occupancy for the $BrCH_2CH_2SO_3^-$ .This leads to believe there is a low rate of complex conformation, with the crystal being richer in free protein. The difference between the actual findings and what was expectedcould be due to the steric effects derived from the fact of using a bigger molecule than vanadate to mimic the ligand. Moreover, the presence of both the bromine atom and the three oxygen atoms forces a particular three-dimensional arrangement for the ligand, in order to maximize all the possible attractive interactions while diminishing the repelling ones, as well as the repelling steric effects. This lessens the possibilities of achieving the desired complex, hence explaining the differences with the vanadate complex.

Fig. 10 shows a superposition of all the active site crystallographic structures obtained for the PTP1B-bromide complex.

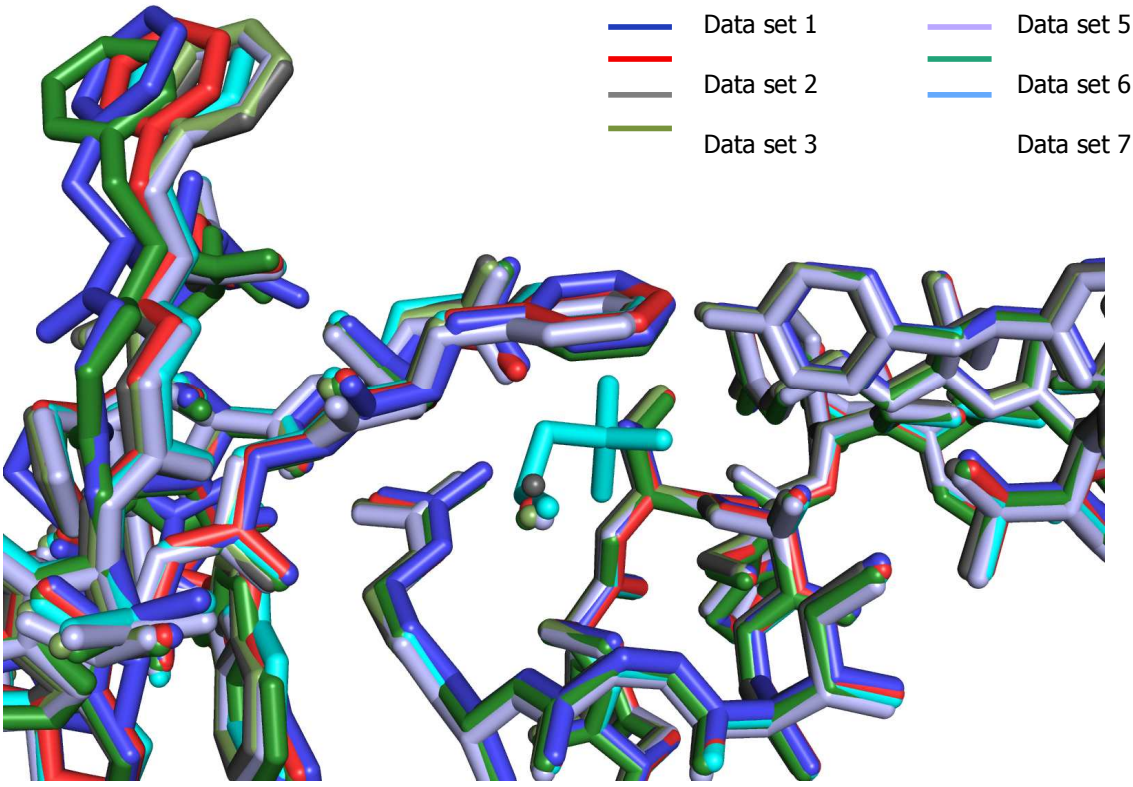


Fig. 10 – Superposition of crystallographic structures obtained for the active site, data sets **1-7**.

It is interesting to notice that the resulting structures are virtually the same for the most part, presenting the bigger differences in the residues involved in the WPD loop. In particular, the differences are mostly noticeable in the open conformation, while the residues in the ligand-bound conformation present a very good superposition (very little differences). The position of the ion (or the bromine atom, in the case of **7**) presents a very low variability between the different structures (it is possible to explain it considering little positional variations that still maximize the stabilizing interactions). This leads to assume that the interactions between the WPD residues and the bromide stabilize the former (and, in particular, the phenyl-ring of residue Phe182, which is placed directly above the ion) in a particular structure, where all the possible attractive interactions are maximized, while reducing the repulsive ones. When the ion is not located in the catalytic cleft, the open conformation is preferred (as observed before), and no extra interactions are stabilizing the three-dimensional arrangement of the loop, therefore permitting a more flexible arrangement of this residues in space.

4.1.3.14- Iodine complexes

The superposition of the obtained active site structures for the PTP1B-iodide complexes is presented in **Fig. 11**.

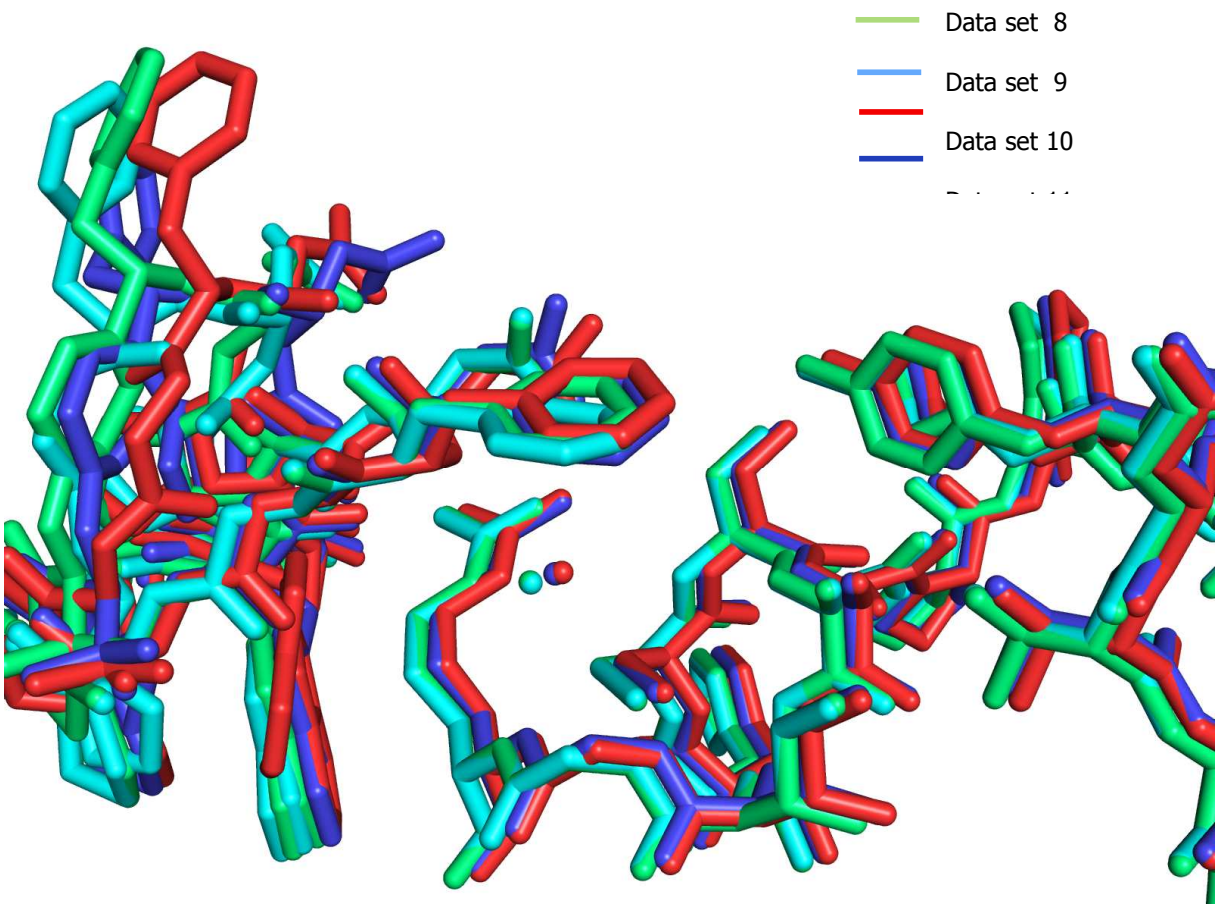


Fig. 11 – Superposition of crystallographic structures obtained for the active site, data sets **8-11**.

The first thing that can be noticed upon analyzing this figure is that, as in the previous case, all closed conformations seem to present the same three-dimensional structure. The superposition of the WPD loop in these complexes is almost perfect, once again hinting the existence of stabilizing interactions that favor this spatial orientation. The superposition in the halide is even better than for the bromide, and there are two possible (and not mutually exclusive) explanations for this: first, since the iodide is a bigger atom than the bromide, a little variation in its coordinates could generate repulsive sterical interactions. In addition to this, this halide presents a lower B-value, thus an increased occupancy regarding that of bromide, thus generating a better diffraction signal, which, in turn, allows a better positioning of the atom. The unbounded conformation presents a higher variability than in the previous scenario, and this can also be explained with the

data collected, since in all cases the open conformation showed a low occupancy, limiting the amount of structural information available.

4.1.3.15- Effects of concentration

Another interesting aspect to take into account is the possible effect that the halide and buffer concentration could have in the formation of the complex. In **Table 6** are summarized the different crystallization conditions, ion:buffer concentration ratio as well as the occupancy of the corresponding halide.

Data set	[Ion] (mM)	[Ac] (mM)	[Ion]: [Ac] ratio	Ion B value (Å^2)
1	200	20	10:1	34.8
2	25	25	1:1	62.1
3	25	200	0.125:1	159.6
4	150	200	0.75:1	204.0
5	100	300	0.33:1	101.2
6	150	300	0.5:1	35.4
7	150	25	6:1	137.0
8	25	50	0.5:1	27.6
9	150	50	3:1	14.7
10	25	200	0.125:1	37.6
11	100	200	0.5:1	13.8

Table 6- Correlation between the ion:buffer concentration ratio and ion's occupancy

In data sets **1** to **6** (in all of which the bromide is present as the unbounded ion), a correlation between the ratio of ion and buffer concentrations can be noticed. In the first two data sets, where the ion shows a high occupancy, the ion concentration is important (either the same or ten times higher than that of the buffer solution). In the other hand, data sets **3**, **4** and **5**, where the presence of the ion is exceeded by the acetate's, the B value for the ion is notoriously increased (thus indicating a lower presence of complex in the crystal). Data set **6** present a particular situation, since even when the ratio between

the acetate and the ions concentration is lower than it was for data set **4**, the bromide presents a higher occupancy than in the latter. This is the only crystal where such a behavior is seen, and it contradicts the general behavior, hinting that there might be an external factor affecting this particular case.

When the bromine atom is present as a part of a larger molecule, as is the case in data set **7**, the presence of the complex seems to be overrode by the free protein (the occupancy of the ligand appears quite small), but that could be due to the facts already discussed.

In the case of the iodide complexes (data sets **8** to **11**), even when the halide shows a high occupancy in every case, it is possible to notice that the complex with the lowest occupancy corresponds to the crystal with the lower iodide concentration.

This could be related to the fact that a higher concentration of ion implies an augmented chance of the halide to actually enter the catalytic cleft, thus interacting with the surrounding residues (fixing its position).The elevated buffer solution concentration could somehow be interacting either with the ion or with the protein preventing these interactions, in some sort of competition. Nonetheless, it is interesting to note that if both the buffer solution and the halide are concentrated, the effect of the buffer solution is overcome. In such cases, the B value for the position of the ion is low (for example, the case of data set **5**, where even when the ratio between concentrations is low, the bromide shows a slightly higher occupancy that in the case of **3**, where not only the ratio is low, but there is also a small concentration of bromide). The same pattern can be seen for the corresponding WPD's residues occupancies (the cases where the ion shows a higher concentration are the ones presenting a high occupancy for the closed conformation of this loop), hence supporting this hypothesis

4.1.3.16- Occupation refinement

Based on these results, the occupancy of the WPD loop's residues as well as the corresponding ion was refined in some of the data sets obtained (once again, the experimental procedure is detailed in Chapter 6). In the case of bromide complexes, three data sets were considered for further refinement: two of them were selected to see the

differences between the open and closed conformation;the third one was selected in order to analyze the influence of the buffer and halide concentrations. The data sets selected were **1**, **2** and **5**, because they showed the best combination of open/closed conformation and ion B values). In the case of the iodide complexes, only one data set is further refined, since no open conformation is present (once again, the selection on what data set to continue to refine is based on the combination of B values for the three regions of interest; in this case, it is the **11**). The results for this further refinement are shown in **Table 7**.

Data set	Occupancy (in %)		
	Residues 180 to 183		Ion
	Open conformation	Closed conformation	
1	20	80	70
2	35	65	25
5	65	35	20
11	25	75	75

Table 7- Refined occupancy for selected data sets

As expected, the crystal that showed an open conformation for the flexible WPD loop (**5**), show a very low occupancy for the ion (in this case, bromide), while the difference between both conformations is not too big. This agrees with the previous theory where the non-covalent interactions between the ion and the surrounding residues could be a factor in stabilizing the presence of the halide in the catalytic cleft.

In the case of **1** and **11**, both crystals where the halide concentration was elevated, the closed conformation is clearly preferred to the alternative one, and both ions show an elevated occupancy, just as expected from the previous analysis.

Finally, for the combination of the ligand-bound conformation but with lower halide concentration (in this case, bromide – **2**), the results are also coherent with the previous analysis, since the difference between the occupancies for both conformations has diminished, as well as the halide's.

The structures of the complexes with the two different halides were compared, considering in each case the crystal with the highest occupancy for the ion (data set **1** for PTP1B-bromide, and data set **11** for the PTP1B-iodide complex), as is shown in **Fig. 12**.

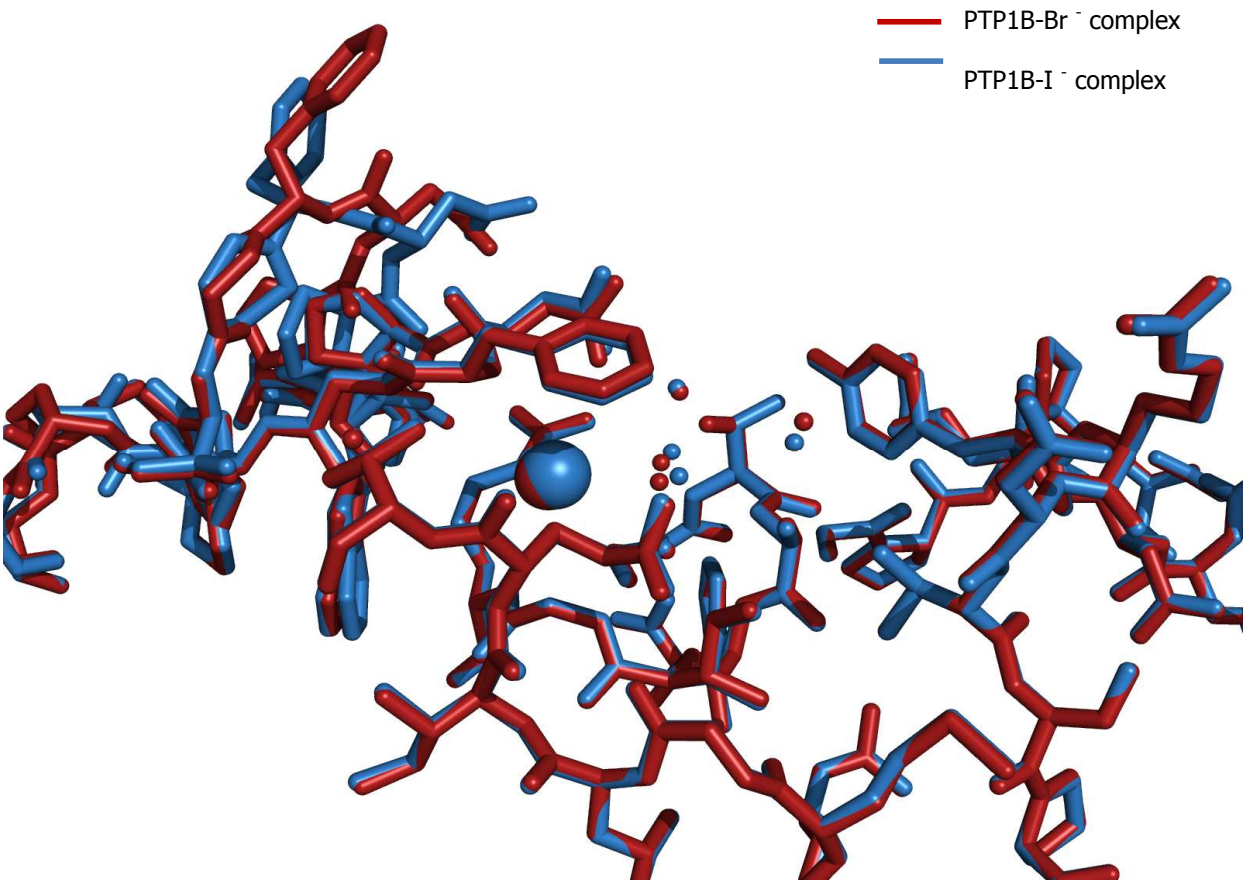


Fig. 12 – Superposition of crystallographic structures obtained for the active site for the bromide and iodide complexes, including water molecules (represented by the smaller isolated dots).

4.1.4- Discussion on crystallographic results

It is interesting to notice that even when the residues in the open conformation differ quite a lot between both complexes, that is not the case for the closed one, nor for the rest of the residues surrounding the halide; moreover, the superposition in these residues is almost complete. The same happens with the halides themselves, which are placed in the same position in both crystals. Even more, the oxygen atoms of the crystallographic

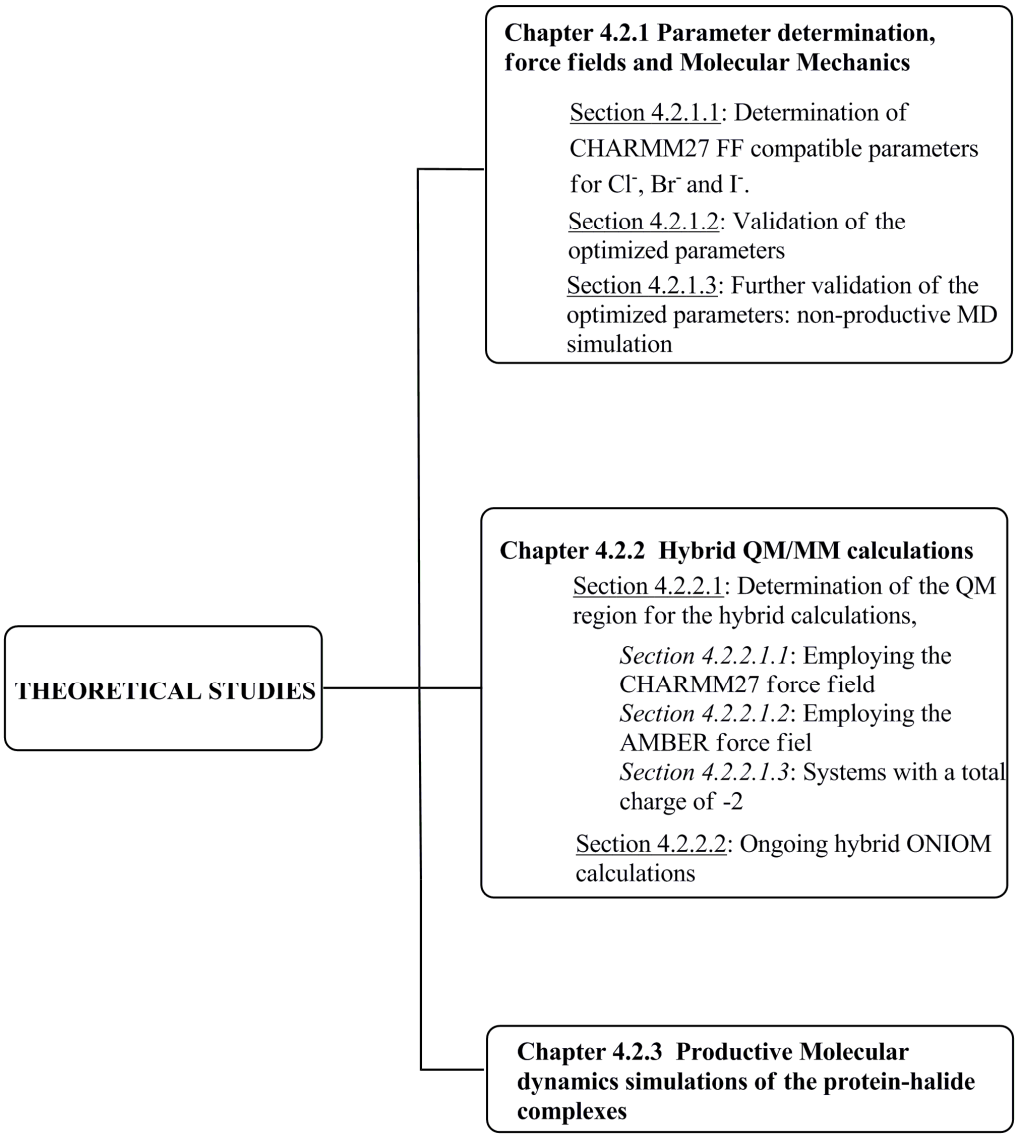
water molecules within the catalytic site are located almost in the same position in both structures.The very slight difference can be seen in some of these atoms is probably due to either the different size of the ions, or the natural error associated to the method, if not to a combination of both factors. The importance of this observation relies in that this would be indicating that the position of both the ions and the water molecules is not casual, but rather a consequence of the interactions taking place in the catalytic cleft.

Based on the crystallographic data and the previous discussion, some interesting observations can be drawn: there appears to be a correlation between the identity of the ion present at the catalytic cleft and the conformation adopted by the flexible WPD loop. This is probably due to the fact that the interactions taking place between the halides and the surrounding amino acidic residues are not the same in both cases. Further studies should be conducted in order to gain a deeper understanding of which they are, and how they influence the adopted conformation.

A second observation that rises from the analysis of the experimental data is the existence of a correlation between the conditions in which the crystals were grown, and the structural conformation of the complexes obtained. The higher the halide's concentration in the preparation mix, the more the ligand-binding conformation is favored.

4.2- THEORETICAL RESULTS

Since the experimental data indicates a correlation between the identity of the ion present in the PTP1B-halide complex and the three-dimensional WPD loop conformation in the resulting complex, the next step is to conduct a series of theoretical studies in order to understand what the plausible causes for this correspondence are. The proper analysis of these complexes implies the use of a wide range of theoretical methods, covering not only different programs, but also different requirements and characteristics. This is why each method as well as the particular features chosen for its implementation is presented along with the results. A summary of the different studies carried out is shown in Scheme 2.



Scheme 2 – Summary of the different theoretical methods employed throughout this thesis work.

4.2.1- Parameter determination, force fields and Molecular Mechanics

The study of a macromolecular system employing Molecular Mechanics (MM) implies the determination of the force field (FF) to use. It is an important step, since the characteristics of the empirical function chosen to describe the system will deeply influence the outcome of any study conducted with it. Several FF have been developed through the last decades, each of them with its own characteristics, advantages and disadvantages [11]. In particular, the CHARMM FF, first described in 1983, is one of the

most successful in the study of biomolecular systems, and it has been used in many Molecular Dynamics (MD) simulations [12]. As described in Chapter 2, force fields are actually empirical mathematical functions, which are approximations to the exact potential energy of the biomolecule at study, and, as such, they must be improved over time [13,14]. This constant enhancement of the force fields results in the existence of several versions for each of them, which can turn out to be both confusing and problematic when trying to carry out a comprehensive study of a biological system, were the use of different versions of the same FF can derive in unreliable results. CHARMM27 [15,16] is the latest version of CHARMM, released in 1999, but it still does not include any description for iodide and bromide ions as independent atoms (the way they are present in the complex studied in this work).This description is mandatory, since without the inclusion of these descriptions in the employed force field, no further MM nor MD analysis can take place with it.

4.2.1.1- Determination of chloride, bromide and iodide ions parameters for use in molecular simulations of TIP3P compatible solvated systems with CHARMM27 force field

Ions interact through non-bonded, 'through-space' interactions, which include electrostatic and van der Waals (VDW) contributions. Since ions have a defined, well localized charge, the electrostatic factor is easily determined, but the VDW interaction remains to be modeled, so that was the first step in the theoretical study of the PTP1B-halides problem. Since chloride ion is already described in the aforementioned force field, and it is critical that such a function is internally consistent, the parameters for the chloride atom were optimized along with the determination of bromide and iodide, to ensure the compatibility of the obtained parameters with the rest of the FF.

One of the possible approaches for the optimization of van der Waals parameters is primarily based on the use of Quantum Mechanical (QM) data [17], and that was the chosen methodology in this case. In order to do so, a water molecule was optimized at an *ab initio*[18,19] level of theory, and the Møller-Plesset perturbation theory, truncated at second order (MP2) used to calculate the correlation energy [20,21]. The full electron *split-valence*[18,19] basis set DGDZVP was employed. The choice of the basis set is

grounded in the fact that the same conditions were used to study the interaction of the water molecule with the different ions, and previous studies have shown that this approach gives better results than pseudo-potentials when employed for the study of molecules containing fifth period atoms (including iodine compounds) [22,23]. Since CHARMM27 parameters were optimized relative to model compounds in condensed phase [16], and in order to guarantee a good correlation between the halide parameters to be determined and the rest of the force field, all the QM calculations were carried out in the presence of aqueous solvent, employing the Polarizable Continuum Model (PCM) [24,25]. The QM interaction energies between the different ions and the optimized water molecule were then determined at the same level of theory, and the inherent basis set superposition error (BSSE) [19] the Counterpoise method was employed [26-28]. The interaction energies were calculated as the total BSSE-corrected energy for the water-ion complex (without optimization of the complex three-dimensional structure), minus the sum of the individual energies of the water molecule and the corresponding ion in each case. The distance between themwas changedfrom 2.0 to 4.0 Å, in 0.1 Å steps (all distances being larger than bond distance, since this procedure aims to determine the non-bounded interactions). Calculations were carried out using the Gaussian09 program [29].

Molecular Mechanics calculations were performed on the same set of systems (using the same structures employed in the previous step) with NAMD [30] and VMD [31], using the CHARMM27 FF and the newly determined trial parameters for the ions. The water molecules were modeled using the TP3M water model [15], an analog of TIP3P water model employed when the SHAKE constraint algorithm [32] is not used. This analog is described to be employed with the MMFF force field [33], but the advantage it presents is that is not a rigid water molecule, (the definition of the TP3M model includes only the two O-H bonds, while the TIP3P water molecule also has the H-H distance parameter defined). This allows the TP3M model to incorporate the polarization effects, thus making it a better model for the analysis of its interaction with ions. The final FF parameters for the ions were obtained using a best fit method between the MM and QM energy curves.

The CHARMM force field represents the VDW interactions through a potential function known as the *Lennard-Jones 12-6function*, with the form presented in **Eq. 96**:

$$V(Lennard-Jones) = \varepsilon_{i,j}\left[\left(\frac{R_{i,j}^{min}}{r_{i,j}}\right)^{12} - \left(\frac{R_{i,j}^{min}}{i,j}\right)^{6}\right]$$

(Eq. 96)

In this mathematical expression, there are only two adjustable parameters: the collision parameter $R_{i,j}^{min}$, which represents the distance at which two given atoms i and $j$present the minimum Lennard-Jones interaction (it is related to the VDW radius of the atoms involved); and the well depth $\varepsilon_{i,j}$, which indicates the magnitude of the favorable London's dispersion interactions between the same two atoms. In order to properly represent the interactions taking place, two components are included: an attractive part, which varies as r^{-6}, and a repulsive term, dependent on r^{-12}. This is a pairwise model, but it was found that a significant proportion of the many-body effects that can be incorporated into the Lennard-Jones model if properly parameterized. It is, then, an 'effective' pairwise potential, not representing the true interaction energy between two isolated particles, but including many-body effects in the energy [17,34].

The form of **Eq. 96** implies that the potential energy of the system depends on each atom's interaction with the others present in the system. This is why, in the particular case of the water – halide complex, the interaction between each ion and all the other atoms (oxygen and hydrogen) must be calculated, and the resulting contributions added, in order to represent the system's energy accurately. According to the FF definitions, $\varepsilon_{i,j}$ and $R_{i,j}^{min}$can be calculated as **Eq. 97** and **Eq.98** show:

$$\varepsilon_{i,j} = \sqrt{\varepsilon_{ion} * \varepsilon_j}$$

(Eq. 97)

$$R_{ion,j}^{min} = \frac{R_{ion}^{min}}{2} + \frac{R_j^{min}}{2}$$

(Eq. 98)

In order to accurately represent the different interactions that take place in the biological environment, two possible conformations ($C_s$andC_{2v}) were considered **(Fig. 13a**-**b**): the

first geometry represents the interaction between the ions and heavy atoms, while, in the second one, the ion primarily interacts with the hydrogen atoms.

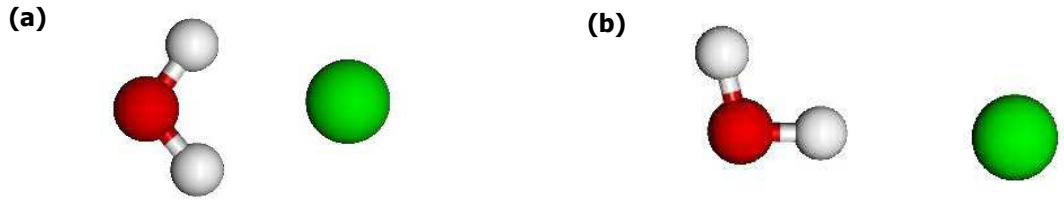


Fig. 13 – Graphical representation of the (a) C_s and (b) C_{2v} molecular geometry employed for the parameter optimization

To begin the iterative process of parameter optimization an initial guess for them is estimated (except for the case of the chloride ion, where the starting parameters were those already described in the considered force field), summarized in **Table 8**,and employed to calculate the systems' energies. In this first step, not only the total energy is obtained, but also the contribution of the electrostatic interactions, which will remain constant through the entire process, since they only depend on the electric charges and the system's geometry. The ion VDW parameters were then obtained by fitting the Lennard-Jones potential function to the MP2 energy profile, employing an iterative least-square method. **Table 8** also shows the optimized parameters for the three halide ions considered.

ION	Initial parameters		Optimized parameters	
	$R_{min}/_2$ (Å)	ε $(kcal/mol)$	$R_{min}/_2$ (Å)	ε $(kcal/mol)$
Cl^-	2.27	-0.0130	2.52	-0.0329
Br^-	2.40	-0.0497	2.80	-0.0312
I^-	2.60	-0.0342	2.86	-0.0990

Table 8– Initial and optimized parameters for chloride, bromide and iodide ions

The comparison between the energy profiles obtained with the *ab initio* calculations and those obtained employing MM, both with the initial and the optimized parameters, for the two geometries, are shown in **Fig. 14a-f**. In each case, the recently determined parameters allow a more accurate representation of both geometries simultaneously, almost reproducing the QM curves. As expected, in all cases (both with the QM and MM calculations), the potential energy presents a rapid decrease, reaching a minimum, and stabilizes afterwards.

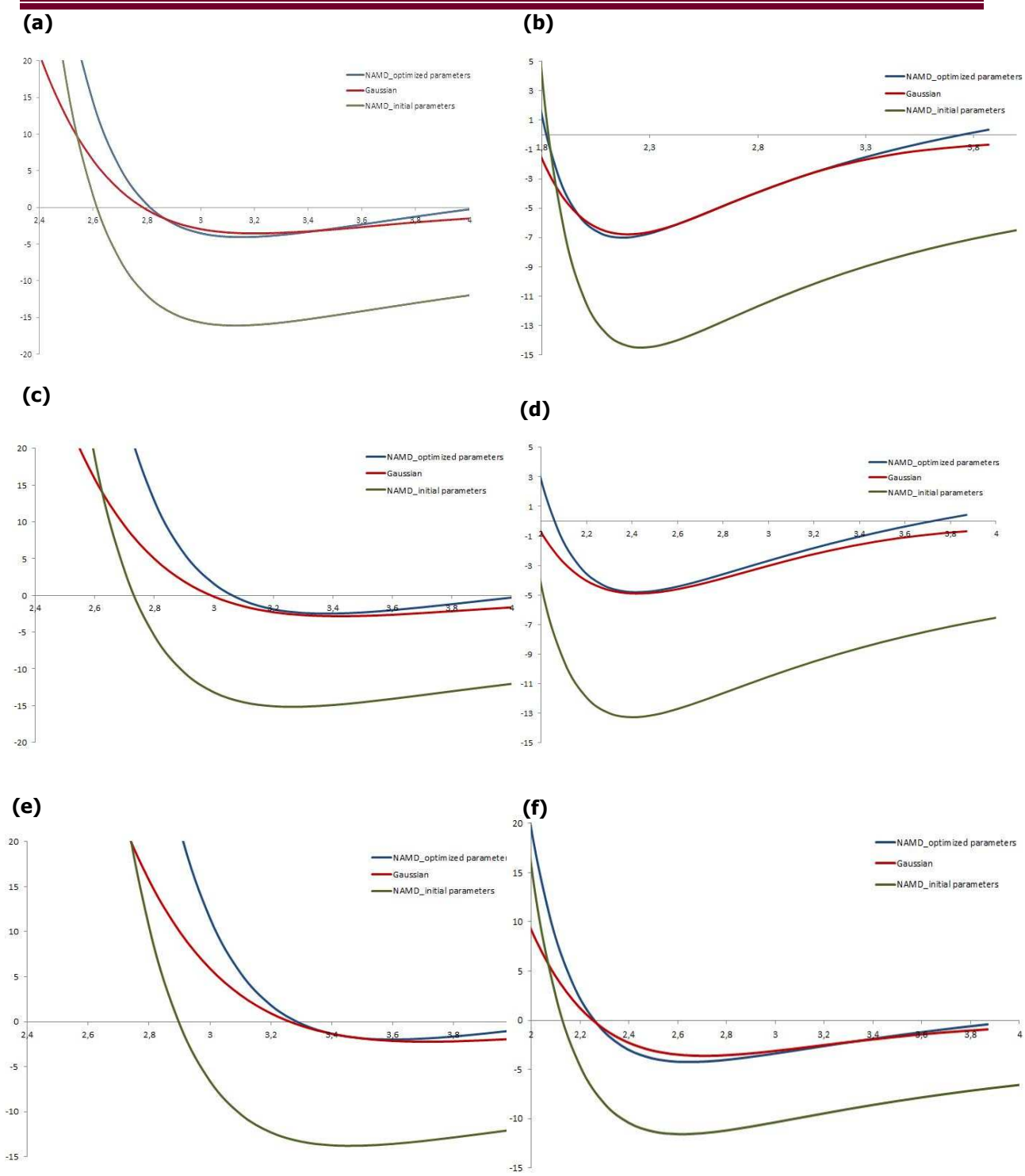


Fig. 14– Energy profiles as a function of the distance between the ion and the oxygen atoms for: chloride ion interacting with (a) oxygen atom and (b) hydrogen atoms; bromide ion interacting with (c) oxygen atom and (d) hydrogen atom, and iodide ion interacting with (e) oxygen atom and (f) hydrogen atom, calculated with Gaussian09 and NAMD. For the MM calculations, the profiles obtained with both the initial guess parameters and the optimized ones are shown.

4.2.1.2- Validation of the optimized parameters

Validation of the newly determined FF parameters for the halide ions was carried out optimizing different models of single ions with one or more water molecules *in vacuo*, employing both QM and MM methodologies, and comparing the resulting structures. In the case of the Quantum Mechanics optimizations, the same level of theory employed for the parameterization step was used (MP2/DGDZVP), while the MM optimization was carried out using the TP3M water model, and the new parameters. The different starting geometries are shown in **Fig. 15**, and were obtained from a similar previous work by Joung and Cheatham, where the parameters corresponding to a series of ions were determined for the AMBER force field [35].

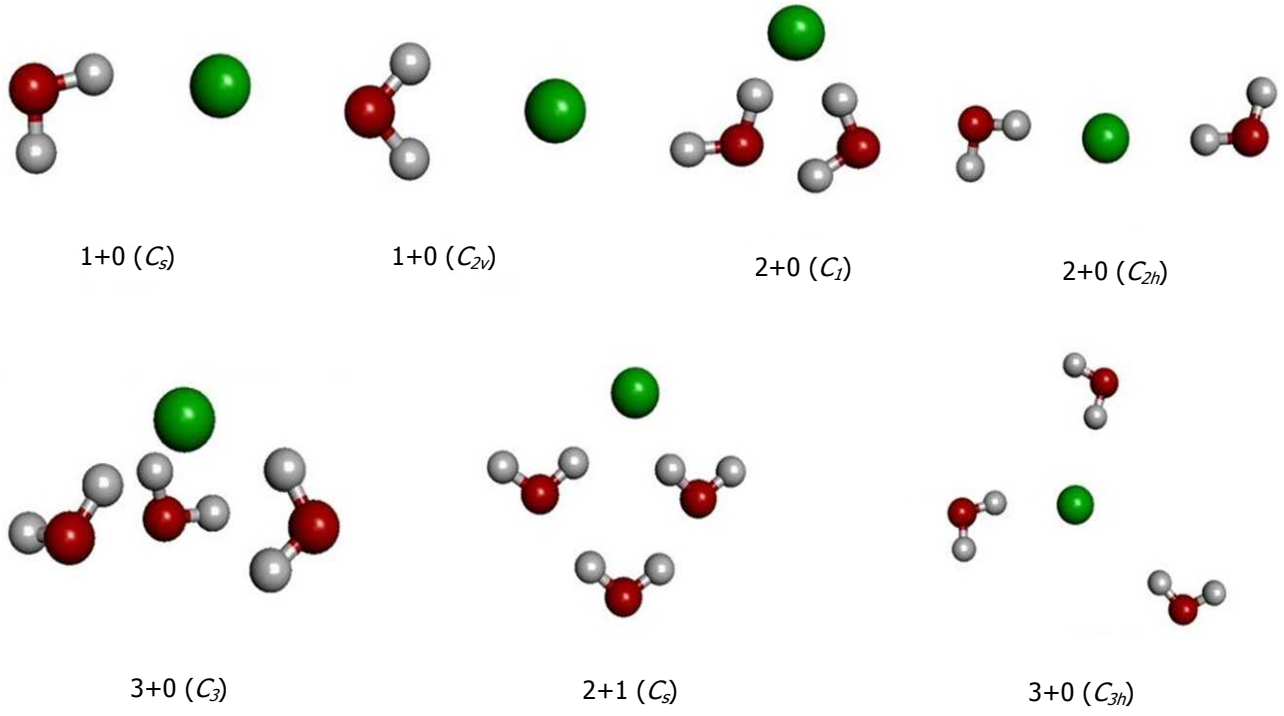


Fig. 15– Graphic visualization of the initial geometries of the halide ions – water molecules systems, thereafter optimized with both QM and MM

The indices indicated in the figure include two numbers, indicating the number of water molecules in the first and second hydration shell, respectively.

4.2.1.2.a- Systems with a single water molecule

In the case of the single water – halide systems, it is interesting to compare not only the structures resulting from both optimization methodologies, but also the halide – oxygen distances in each case. The optimized structures are quite similar in each case, thus proving that the newly determined parameters allow the MM calculations to reproduce satisfactorily the results obtained with QM optimizations. The rmsd values between the structures obtained with either method are shown in **Table 9**.

ION	Structure	rmsd (Å)
Cl^-	C_s	0.35
	C_{2v}	0.35
Br^-	C_s	0.93
	C_{2v}	0.03
I^-	C_s	1.34
	C_{2v}	0.03

Table 9 – rmsd values between QM and MM optimized structures for the single water – halide complexes

It is interesting to note that as the halide being considered becomes bigger, the difference in the structure between both methodologies increases. This is consistent with the halide behavior in general, where it can be seen that the bigger halide, the more its behavior averts from the expected one.

The distances between the water oxygen and the corresponding halides are shown in **Table 10**, as well as these distances deviations from several reference values. In the table are included not only the distances obtained using our parameters and the ones obtained using QM calculations, but also the distances obtained employing different parameter sets. The reference values are averages of one or more *ab initio* calculations, as considered in Joung& Cheatham [35], and all the other distances considered for the comparison are also obtained form that same work.

STRUCTURES		New parameters	QM (MP2 full)	Joung& Cheatham	Jensen & Jorgensen	Reference values
Cl -	C_s	3.33	3.40	3.326	3.38	3.32
	C_{2v}	3.33	3.39	3.30	3.42	3.35
Br -	C_s	3.55	3.70	3.48	3.64	3.61
	C_{2v}	3.55	3.69	3.51	3.65	3.61
I -	C_s	3.09	3.18	3.09	3.26	3.12
	C_{2v}	3.09	3.18	3.15	3.30	3.20
rmsd		**0.02**		**0.03**	**0.03**	

Table 10– Single water oxygen atom-ion distances for the optimized geometries for different ion parameter sets with TP3M water model. The distances are expressed in Å, and the rmsd values are calculated against the reference values.

The previous comparison shows that the deviations obtained with the newly determined parameters are not only within the range of acceptable ratios, but they are also lower than those obtained with accepted parameter sets.

4.2.1.2.b- Systems with multiple water molecules

Fig. 16 shows an example of the comparison of the structures resulting from optimizing the systems with multiple water molecules employing the two previously described methodologies. The rmsd values between both structures are depicted in **Table 11**.

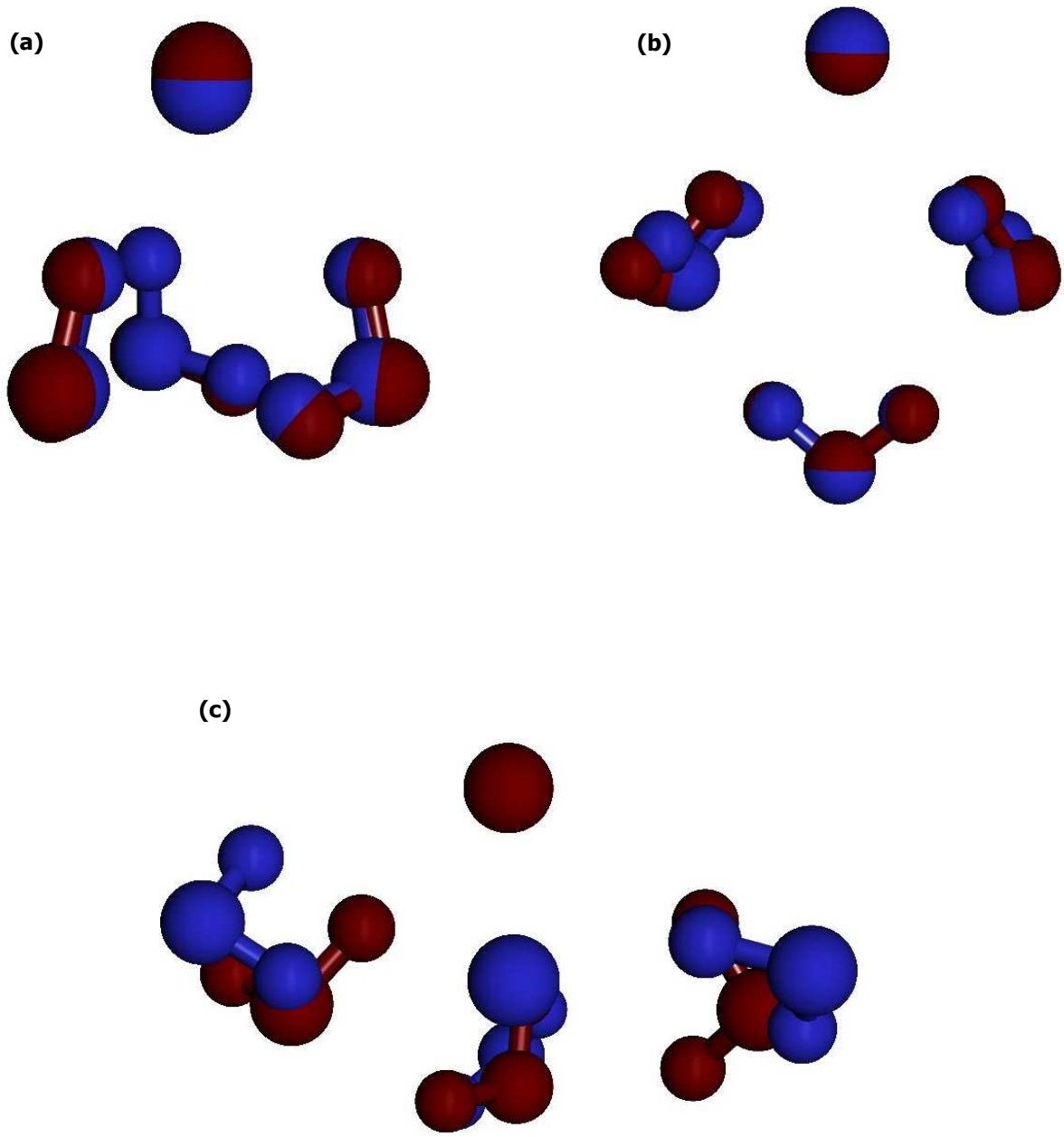


Fig. 16 - Example of the comparison of the optimized structures for halide – three water molecules systems, in the (a) C_3, (b) C_s and (c) C_{3h} configurations. In red is the optimized structure obtained with QM calculations, while the blue structure represents the optimized structure obtained with MM calculations

System	Ion	Structures	rmsd (Å)
Two water molecules	Cl^-	C_1	0.22
		C_{2h}	1.62
	Br^-	C_1	0.22
		C_{2h}	1.81
	I^-	C_1	0.24
		C_{2h}	2.04
Three water molecules	Cl^-	C_3	0.10
		C_s	0.27
		C_{3h}	1.40
	Br^-	C_3	1.86
		C_s	0.30
		C_{3h}	1.71
	I^-	C_3	2.07
		C_s	0.30
		C_{3h}	1.91

Table 11 – rmsd values between the QM and MM optimized structures for the different water-halide systems

The most noticeable thing is that when the water molecules are close to each other (as in the C_1 and $C_s$configurations, for the two and three water molecules systems, respectively) the optimization employing the parameters determined in this work reproduces almost perfectly the structures obtained with QM.In the case were all the water molecules are further apart from each other (C_{2h}, C_3 and C_{3h} configurations) the MM optimization does not succeed in reproducing the *ab initio* structures. This could be indicating that the problem lies in the representation of the long-range water-water interactions rather than in the recently determined halide parameters. Furthermore, in the case of the chloride ion, which presents a smaller van der Waals radius (allowing the water molecules to arrange themselves spatially closer to each other), the optimization of the C_3configuration with both methodologies produces really similar structures, thus supporting the proposed hypothesis.

4.2.1.3- Further validation of the new parameters: MD simulation

Since the determination of this parameters aims to properly represent the interaction between halide ions and biomolecules, their performance was also studied by performing MD simulations for the PTP1B-ion complexes, using the NAMD and VMD software. The quality and analysis of any molecular simulation are strictly related to the quality of the parameters considered in the force field employed, that is why it is necessary to confirm that any MD simulations performed with the newly determined parameters behave correctly (that is, as expected). Initial structures for these simulations were obtained from some of the previously described X-ray experiments, including the water molecules in the catalytic cleft. Since the whole interest of this study is to see how the ions behave in normal MD simulations conditions, and not obtain any properties from the complexes, the PTP1B – Cl^- complex was obtained simply by substituting the bromide ion for a chloride in the former's initial structure. The aqueous environment was represented by adding a TP3M water cube measuring 12 Å per side. Each simulation consisted of a single trajectory of 10 ps of equilibration time; followed by 10 ns of data collection for analysis, with a 1 fstimestep (no constraining algorithm was employed). A modified Nose-Hoover method was employed to ensure a constant pressure simulation, which is a combination of the Nose-Hoover constant pressure method with piston fluctuation control, implemented using Langevin dynamics [36-38].This allows to maintain a physiological temperature of 310 K. All simulations were carried out using periodic boundary conditions, and applying the particle mesh Ewald (PME) [39] method for the long range electrostatic interactions. The nonbonded cutoff distance was 12 Å (with an additional 1.5 Å as allowable distance between atoms for inclusion in the pair list), and smoothing functions were applied to electrostatic and van der Waals forces beginning at 10 Å.

The first two aspects to be analyzed in a molecular dynamic simulation are the temperature and the total energy throughout the whole run, to confirm that the system at study has reached equilibrium. In **Fig. 1a-f** the temperature and total energy for each one of the three systems is displayed for the whole length of the simulation run.

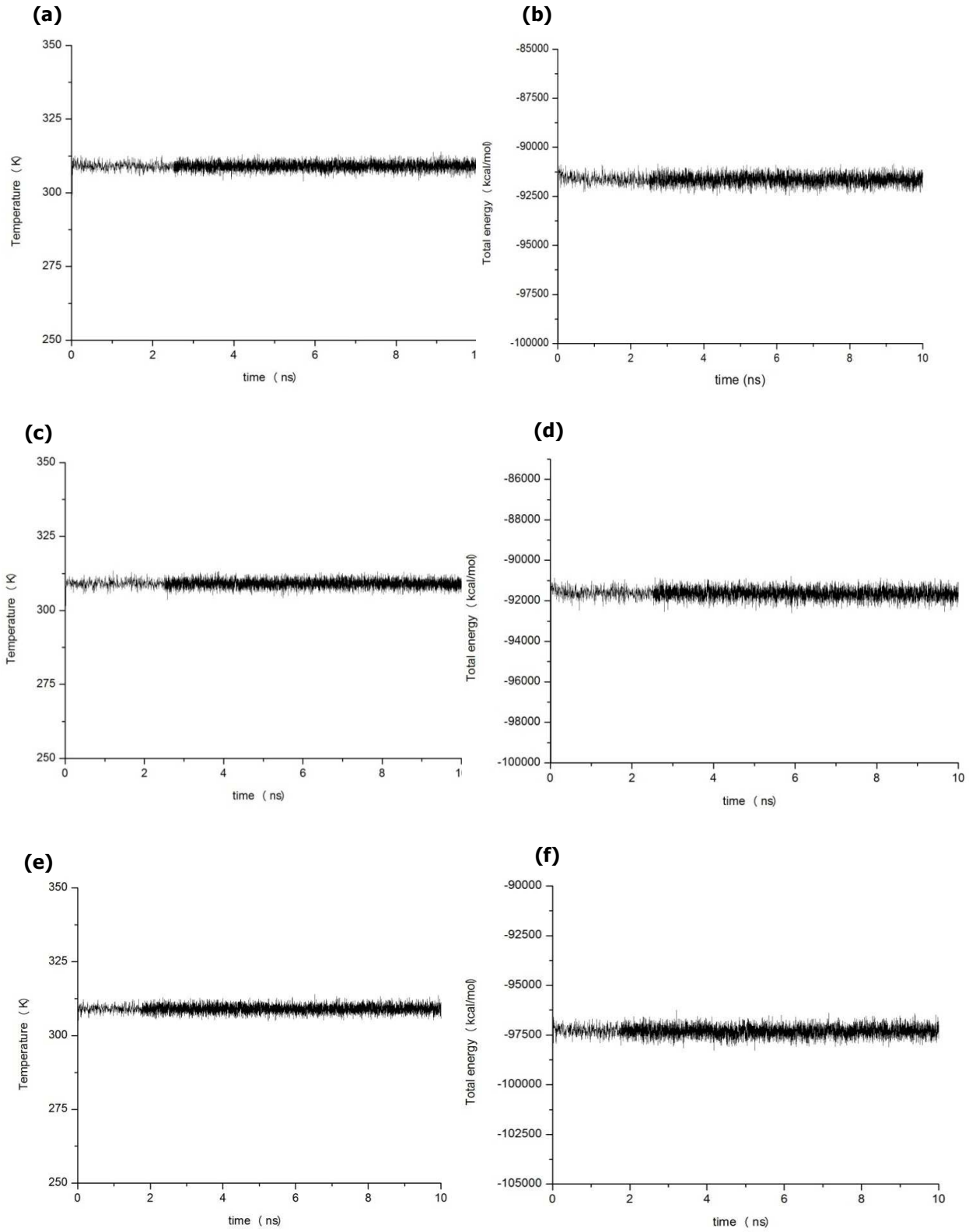


Fig. 17 – Temperature and total energy variation throughout the Molecular Dynamics simulation for chloride, (a) and (b) respectively, bromide, (c) and (d), and iodide, (e) and (f).

As the previous figure shows, all three systems reached equilibrium quite rapidly, and the energies remain stable throughout the rest of the simulation.

The next parameter to analyze is the root mean square deviations (rmsd) values for the different components of each system, which indicates their relative motions at all points of the simulation. The corresponding values for the whole protein (considering the backbone atoms) and for the ion in each case are presented in **Fig. 18a-f** below.

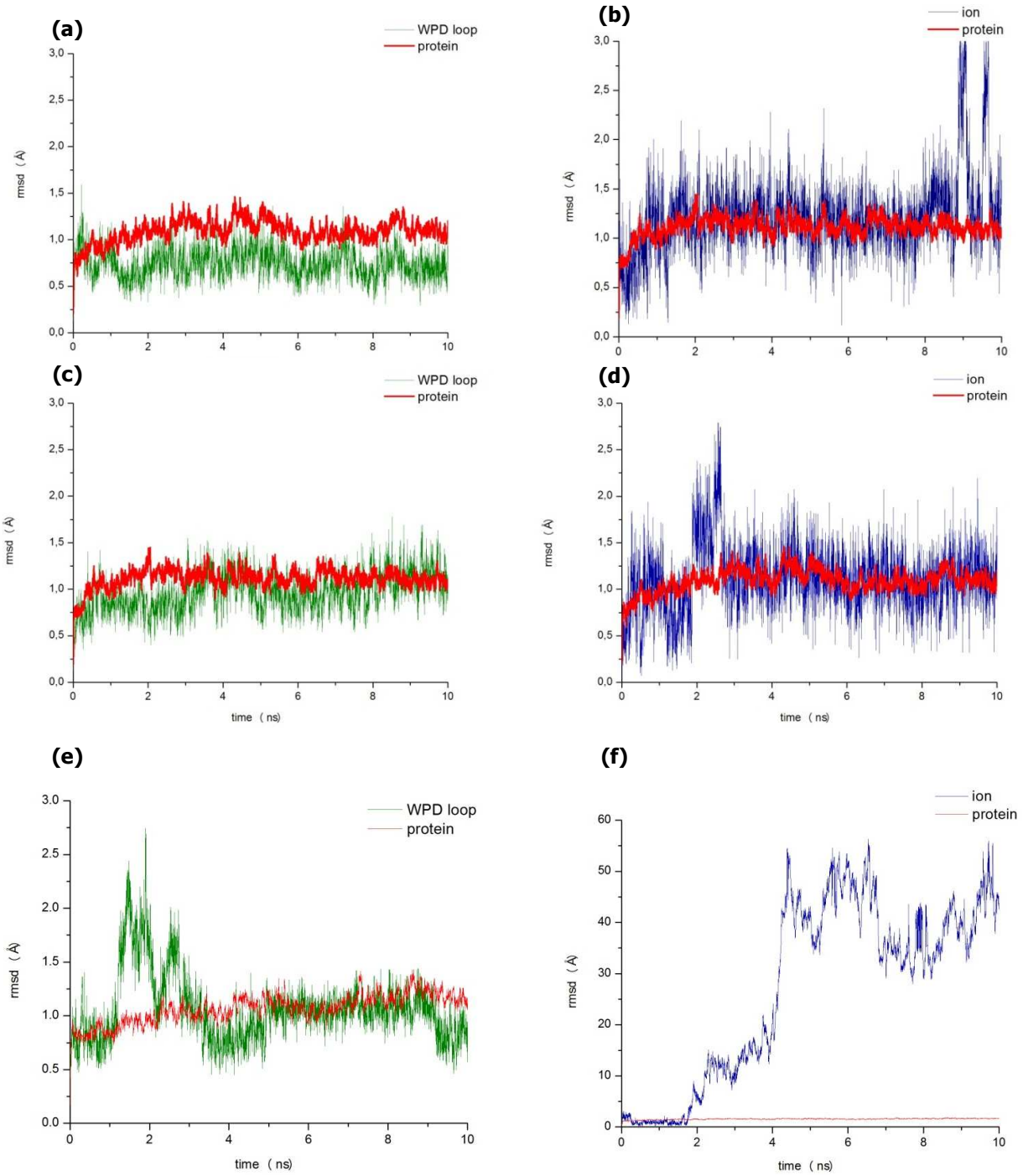


Fig. 18 – rmsd values for the whole protein and residues 177 to 184 of the WPD loop throughout the molecular dynamic simulation for (a)chloride, (c) bromide and (e) iodide; rmsd values for the whole protein and the ion for (b) chloride, (d) bromide, and (f) iodide

In all cases, the rmsd plot for the protein after the initial frames (during which the water molecules adjust around the protein surface) grows quickly up to around 1.1 Å, where it levels off and oscillates around this value for the rest of the simulation. The leveling of the rmsd means that the protein has relaxed from its initial crystal structure (which is affected by the crystal packaging and the fact that some atoms are missing, such as hydrogen atoms) to a more stable one, and hence is an indication that an equilibrium has been reached. Since 1.5 Å is the acceptable value in most protein simulations, the fact that in all three cases the protein stabilized below that limit indicates that these simulations are acceptable. When considering the WPD loop in particular (residues 177 to 184), the rmsd values found supports previous observations on the flexibility of this loop [40,41]. In this case, the mean values are comparable to those obtained for the whole protein, but they present a more pronounced oscillation around that value, indicating a greater mobility of the considered residues during the course of the simulation. In the case of the PTP1B-iodide complex, the mobility of this loop is even higher, inducing to assume a possible conformational change for these residues.

Analyzing the rmsd profile for the different halides,it is interesting to notice that chloride and bromide present a similar behavior, while the iodide presents important difference with them. As it has been stated before, this is coherent with the difference in size between the considered ions. For the first two cases, it can be seen that, even when the mean value is not far away from that of the protein (around 1.2 Å), the variations are more pronounced, indicating a larger mobility for this atoms. This is,actually, what was expected, since the halides in the different complexes do not present any stabilizing covalent interactions, thus allowing them more freedom degrees. Nonetheless, these two ions still oscillate around the initial position, not leaving the catalytic cleft; this is an indication of the strength of the non-bonded interactions that keep each of them in their respective place. In the case of the iodide ion, after the first nanosecond of simulation, the rmsd values show an important increase, indicating that the ion leaves the catalytic cleft. The point in the simulation at which this rmsd increase begins coincides with the beginning of the higher rmsd values for the WPD loop: once the WPD loop leaves the closed conformation, the remaining interactions are not strong enough as to maintain the halide in the catalytic cleft. This could be indicating that the strongest interactions keeping

the iodide in place are those with the flexible loop and, in particular, with the phenylalanine, since previous works have proven that non polar surfaces (as the benzyl side chain of this amino acid) interact more strongly with ions of larger size [42]. Even though the dynamical studies carried out so far are not strong enough to draw any strong conclusions from them regarding the interactions that are actually taking place in the catalytic site, the results obtained in all three cases are not only what expected, but also coherent with previous studies, showing that the different ions present different behavior throughout the molecular dynamic simulations [43,44].

Taking into account all of the above, it can be considered that the obtained force field parameters for chloride, bromide and iodide ions are adequate, and can be further used to model biological systems including those ions using the CHARMM27 FF.

4.2.2-Combined Quantum Mechanical/Molecular Mechanical calculations

The complexity of biological systems makes their molecular modeling a very difficult task. Both the nature of the problem at hand (that is, what particular aspects or interactions taking place within the system need to be studied), together with the size of the biomolecule (or complex) at study are the main considerations to take into account when deciding on a methodology. As seen in Chapter 2 of this thesis work, hybrid methods are a good alternative when there is a large system but with a well-defined region where the most interesting interactions are taking place, while the rest of the biomolecule only present a steric or electrostatic effect. This is the case in the system studied in this thesis work: the interactions taking place in the catalytic cleft between the ion and the surrounding residues is the aim of our studies, while the rest of the protein plays more of a secondary role in the definition of the conformation of the studied loop. Nonetheless, it is not possible to completely disregard the electrostatic or steric effect of the biological environment could be exerting on the adoption of one conformation or the other. Based on all the stated above, and taking into account the available resources, the ONIOM method was chosen to study the PTP1B-halide complexes. A two-layer scheme was considered, where the ion and the surrounding residues in the catalytic cleft are considered the *model* system (as described in Chapter 2), and thus studied at the QM

level of theory, and the rest of the protein and the crystallization waters constitute the *real* system, studied at a MM [45,46].

4.2.2.1- Determination of the QM region

4.2.2.1.1- Working with the CHARMM27 force field

One of the critical steps in the study of a biological system with a hybrid method is defining the constituting layers. This is not a trivial task, since it will determine the outcome of the whole calculation, and it requires a compromise. The need for this compromise lies in the fact that, while it is important to include all atoms directly involved in the studied interactions in the *model* system, it is also important to keep that system small enough to ensure that cost-effective calculations. If the considered *model* system is too small, some important interactions may be lost, thus rendering the results useless, but if it is too big, the calculations can turn out to be practically impossible (either because the resources available are insufficient, or because the time needed to obtain an answer is too long). Another aspect to consider is the fact the frontier between the layers: it is important that the calculations with both theory levels in that region are convergent. In order to determine the best possible model system, then, a series of geometry optimizations were carried out using both levels of theory, considering in each case a small subgroup of atoms. The objective is to find the smallest *model* system that better reproduces the experimental results (where the optimized structures are relatively close to the crystallographic structures, since these are the result of all the interactions present in the system) and, simultaneously, renders similar structures with both methodologies, hence providing a good border region. Since this project aims to determine the interactions taking place between the ion and the residues in the catalytic cleft, in all cases the corresponding halide ion was present, and the protein residues as well as the crystallographic waters included in the subgroup were determined considering their distances with such atom.

A parenthesis needs to be made at this point: as in any optimization procedure, a starting geometry is required, and, in this case, the best choice is to consider the experimental

ones, that is, the ones obtained from the X-ray diffraction experiments. Even though several crystallographic structures are available for each complex, it is important to remember that, as noticed in sections *4.1.3.13* and *4.1.3.14*, the differences between them are basically located in the position of the residues of the WPD in the open conformation and, even then, they are not too big. This makes any of the crystallographic structures eligible as starting geometry. This choice has an impact on the calculations, since the better the starting geometry, the sooner the minima located closer to it on the potential energy surface will be found; that is why the choice was based on which structure provided more reliable structural information. Keeping this in mind, for the protein-bromide complexes, the structure resulting from treating **data set1**, which showed the highest occupancy for the closed conformation of the WPD loop, was chosen as a starting geometry the complex in such a conformation,while **data set 5** was chosen as the starting structure for the complex while in the open conformation. In the case of the iodide complex, no crystal showed a clear open conformation, so this conformation was not modeled. **Dataset 10**was chosen to represent the closed conformation.

The QM geometry optimizations were carried out using the Gaussian09 program. The number of atoms included in these calculations makes density functional theory the best choice for this calculation, since it depends on fewer variables while recovering the electron-electron correlation energy [19]. Based on a previous work by Zaho and Truhlar, the M06 density functional was chosen, since it proved to be very good for evaluating noncovalent interactions, as the ones between the halide ions and the surrounding atoms [47]. In order to have consistent results, the full electron *split valence* basis set DGDZVP was used to describe both ion, for the reasons already described in section *4.2.1.1*; for the protein residues, the 6-31G* basis set was chosen, since it is defined for first and second row elements, thus providing good results without an elevated computational cost [18]. The optimization algorithm is the default algorithm considered in Gaussian09, the Berny algorithm; as for the convergence criteria considered, there are two options: either an average force of $3x10^{-4}$ atomic units (a.u.) on all atoms, or an energetic difference of less than 1 kcal/mol between ten consecutive conformations. This last criteria is added because the size of the model consider can make convergence difficult, thus increasing the computational cost with no significant changes in the structure.

The Molecular Mechanics calculations were carried out using the NAMD software, considering the CHARMM27 FF, and including the newly determined parameters for bromide and iodide. The structure and initial structure files were prepared using the *psfgen* module, included in the NAMD software package. In this case, the minimization algorithm employed is conjugate gradient for a maximum of 10000 steps, or until a gradient of 0.001 kcal/mol is achieved. The water molecules were represented using the TP3M water model, the same one already introduced when the parameterization process was described.

The residues constituting the catalytic cleft in PTP1B are not consecutive within the amino acidic chain, so the residues considered for these optimizations are not necessarily bound to each other. Two problems arise from this fact: on the one hand, it is necessary to saturate the dangling bonds resulting from cutting peptide bonds, and this is achieved by capping the residues in each extreme with hydrogen atoms. The second problem lies in the fact that since not all the protein is included, the residues to be optimized present more freedom degrees than it occurs in the reality (when the atomic movements are restrained by the surrounding residues). In order to avoid this issue, the Cα in all the residues are left fixed, thus mimicking the restraint forces these atoms are subjected to in the protein, while leaving the side chains free to be reoriented according to the interactions they are exposed to.

In both methods the optimizations are carried out in Cartesian coordinates, in order to simplify the posteriori comparison of the resulting structures.

4.2.2.1.1.1- 4Å radius

In a first instance, the residues and water molecules considered for geometry optimization were those that presented at least one atom in a 4Å radius from the ion, in each conformation. Thus, three sets of geometries were defined: two of them represented the protein-halide complex with the loop in the closed conformation, and the other one represents the protein-bromide complex with the loop in the open conformation. Since the phenyl-ring of the Phe182 residue in the WPD loop seemed to play an important role in the interaction between the ions and the surrounding protein (as was noticed in this same

chapter, while analyzing the X-ray diffraction results), this residue was included in all the models, even when in the case of the complex in the open conformation no atom of this residue was within the designed radius).**Table 12** shows which residues and how many water molecules were included in each case:

Complex	Residues	Total charge	Water molecules
PTP1B-Br^-, open conformation	Phe182, Gly220, Arg221,Gln262, Gln266	0	1
PTP1B-Br^-, closed conformation	Asp181, Phe182, Gly220,Arg221, Gln266	-1	3
PTP1B-I^-, closed conformation	Asp181, Phe182, Gly220, Arg221, Gln266	-1	3

Table 12 – Residues and water molecules that present at least one atom within a 4Å radius from the corresponding ion

In **Fig. 19a-c** the superposition of the optimized geometries for the three complexes with both methods are shown.

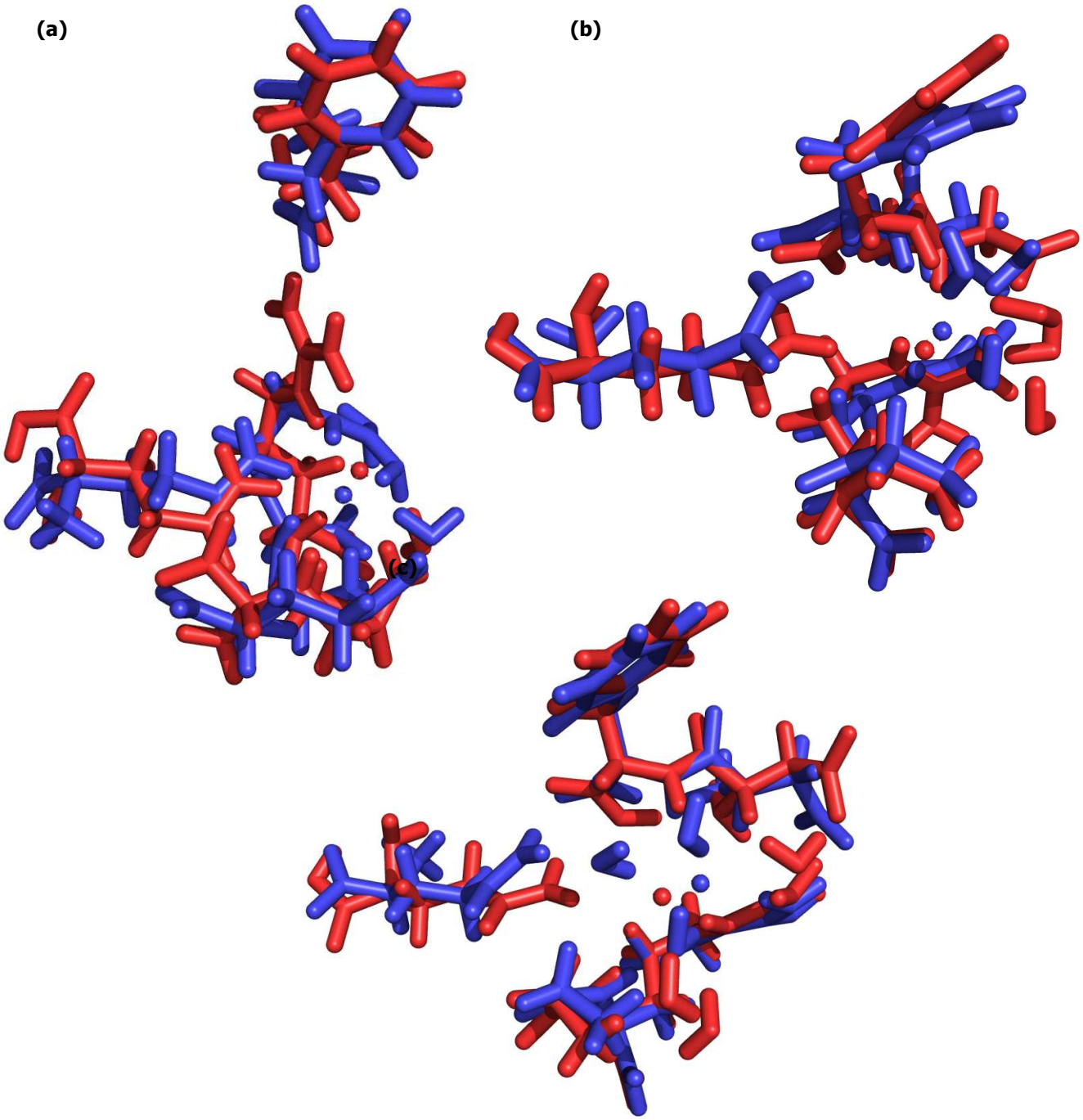


Fig. 19– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a)PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation and (c) PTP1B-I- in closed conformation, considering the residues in a 4Å radius.

In **Table 13** the difference in the obtained structures is resumed, by presenting the root mean square deviation (rmsd) between the QM and MM optimized structures and the initial crystallographic structure for the three complexes.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Phe182	2.50	2.03
	Gly220 - Arg221	2.28	2.02
	Gln262	1.49	1.51
	Gln266	6.69	6.69
	Br^-	0.58	1.02
	water molecules	2.28	2.49
	total (not including water molecules)	2.25	1.85
PTP1B – Br^- closed conformation	Asp181 - Phe182	1.32	1.53
	Gly220 – Arg221	1.40	0.69
	Gln266	0.89	1.61
	Br -	0.41	1.18
	water molecules	0.71	1.88
	total (not including water molecules)	1.28	1.30
PTP1B – I^- closed conformation	Asp181 – Phe182	1.51	1.62
	Gly220 – Arg221	0.60	0.74
	Gln266	0.58	1.39
	I^-	0.63	1.43
	water molecules	2.66	2.28
	total (not including water molecules)	1.05	1.30

Table 13 – rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 4Å radius.

A series of conclusions can be drawn from the previous table, the first of which is the fact that the model of the open conformation presents, as a whole, larger rmsd values than the two that represent the open closed conformation. This is probably due to the fact that in the latter the distances between atoms are smaller, thus allowing the model to better reproduce the interactions taking place in the biological environment. This is clearly the case of the Phe182 residue, which was included in the first model even when it was

further away from the ion, and in both optimizations methods the optimized structure is over 2 Å away than in the experimental one.

The water molecules present the highest rmsd values. In the QM case, this is probably due to the fact that in the crystallographic structure these water molecules are kept in that position by long-range interactions with several surrounding residues, some of which are not represented in these models. In the MM case, this could be also influenced by the inadequate parameters for representing long range interactions with this water noticed during the parameterization process.

As a whole, and as expected, the Quantum Mechanics method reproduces the experimental structure better than Molecular Mechanics. In any case, the rmsd values for the whole system (without including the water molecules) are still not good enough as to consider that a good representation of the experimental data is achieved, and the optimized structures with both methods cannot yet be considered convergent. Hence, the model including this set of parameters is not good enough as to represent the interactions of interest.

4.2.2.1.1.2- 6Å radius

In this second stage, the residues and water molecules considered for geometry optimization were those that presented at least one atom in a 6Å radius from the ion, in each conformation. Once again, three sets of geometries were defined: two of them represented the protein-halide complex with the loop in the closed conformation, and the other one represented the PTP1B-Br^- complex with the loop in the open conformation. In the case of the models representing the closed conformation for the WPD loop include the Phe182 residue within their span, while this is not the case for the model for the protein-bromide complex in the open conformation. In these last cases, that residue is once again included even though it is not within the designed radius. **Table 14** summarizes the residues and the amount of water molecules included in each model:

Complex	Residues	Total charge	Water molecules
PTP1B-Br^-, open conformation	Phe182, Cys215, Gly220, Arg221, Gln262, Thr263, Gln266	-1	4
PTP1B-Br^-, closed conformation	Asp181, Phe182, Gly183, Cys215, Ser216, Ile219, Gly220, Arg221, Gln262, Gln266	-2	4
PTP1B-I^-, closed conformation	Asp181, Phe182, Gly183, Cys215, Ile219, Gly220, Arg221, Gln262, Gln266	-2	4

Table 14– Residues and water molecules that present at least one atom within a 6Å radius from the corresponding ion

These larger models include the catalytically important Cys215, which was not present in the previous ones. This is not a trivial fact, since this residue plays a key role in the catalytic mechanism, turning its inclusion in the model fundamental to understand the interactions taking place in the catalytic site. Since the physiological conditions are considered, the Cys215 is modeled deprotonated, with a negative charge, as reported in several previous works [48-50].

In **Fig. 20a-c** the structures optimized with each level of theory are superposed, in order to compare them:

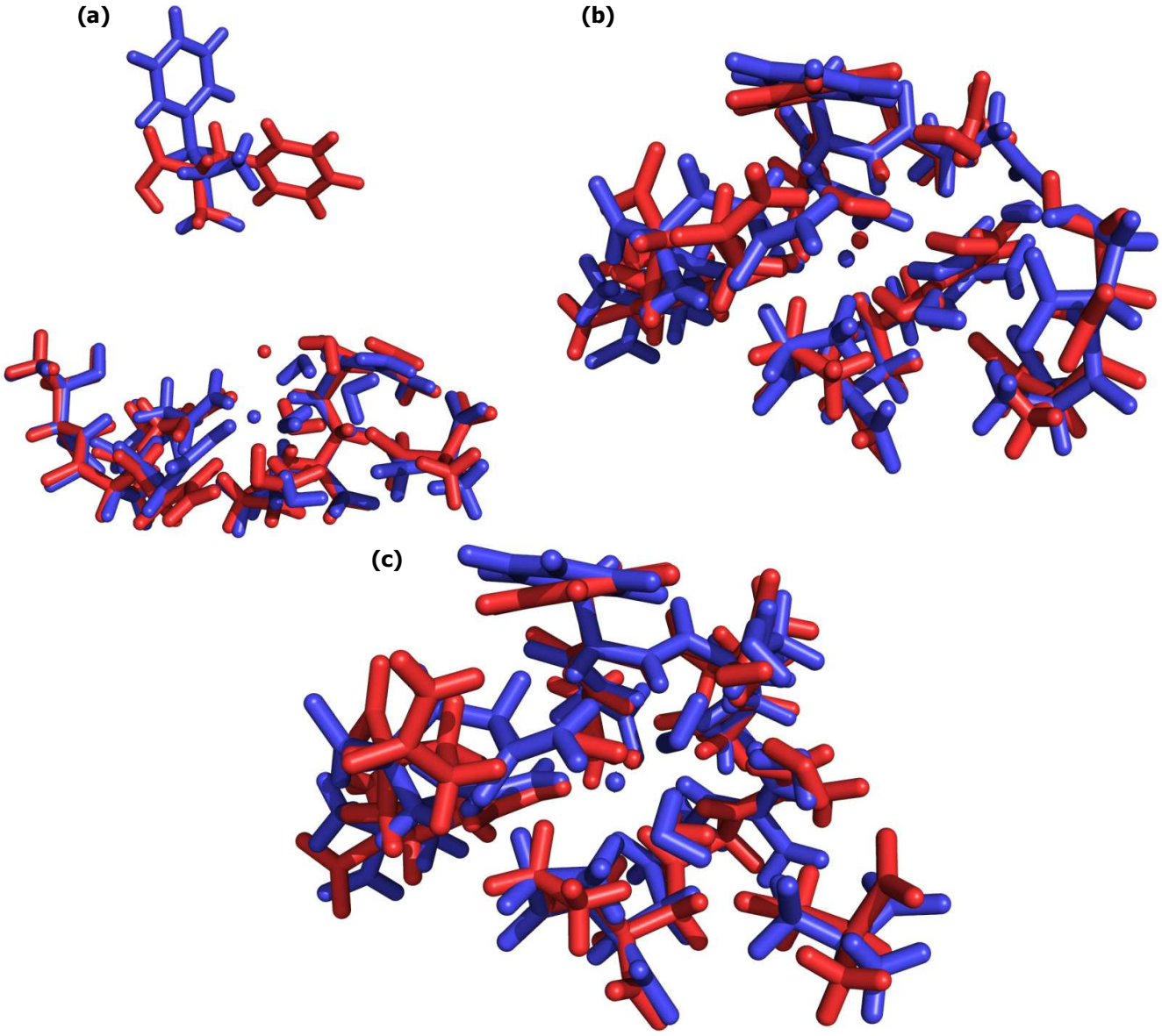


Fig. 20– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a)PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation and (c) PTP1B-I- in closed conformation, considering the residues in a 6Å radius.

In **Table 15**the rmsd values between the structures optimized with both theory levels and the crystallographic experimental data are depicted.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Phe182	4.33	2.01
	Cys215	1.89	1.61
	Gly220 - Arg221	1.23	1.25
	Gln262 – Thr263	1.75	1.52
	Gln266	4.95	1.60
	Br^-	1.50	0.82
	water molecules	5.31	1.92
	total (not including water molecules)	3.00	1.58
PTP1B – Br^- closed conformation	Asp181 - Phe182 – Gly183	0.39	1.60
	Cys215 – Ser216	0.91	1.75
	Ile219 - Gly220 – Arg221	0.53	1.34
	Gln262	0.67	1.57
	Gln266	0.68	1.44
	Br -	0.18	0.79
	water molecules	0.56	1.85
	total (not including water molecules)	0.62	1.54
PTP1B – I^- closed conformation	Pro180 - Asp181 – Phe182 - Gly183	0.64	0.91
	Cys215	0.55	1.61
	Gly220 – Arg221	0.68	1.28
	Gln2623	0.47	1.91
	Gln266	0.30	1.39
	I^-	0.05	0.95
	water molecules	0.67	1.18
	total (not including water molecules)	0.59	1.32

Table 15– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 6Å radius.

When considering the two models that represent the complex with the WPD loop in a closed conformation, it is interesting to notice the improvement on the structure optimized with QM. The rmsd values between these structures and the experimental data could indicate that these models would be sufficient to correctly reproduce the interactions taking place in the crystal packing, more closely resembling the actual biological environment. When considering the MM optimized geometries, the difference regarding

the previous results is not so big, but this is coherent with the fact that this method employs an empirical parameterized force field. However, there does not seem to be a clear improvement in the resemblance between the optimized geometries.

In the case of the model representing the open loop conformation for the PTP1B-Br^- complex, there is no improvement in the optimization at the QM level. This can be easily explained by the fact that the Phe182 residue is still apart from the rest of the model, thus the stabilizing interactions generated by the surrounding residues that determine this residue's orientation in the biological system are not reproduced. Once again, this effect is less pronounced in the MM level, due to the empirical information included in the force field.

Since the model for the open conformation is not good enough to represent the empirical data, and there seems to be no consistency between the methods (thus introducing some error in the frontier region), it is important to consider a model where more residues are included.

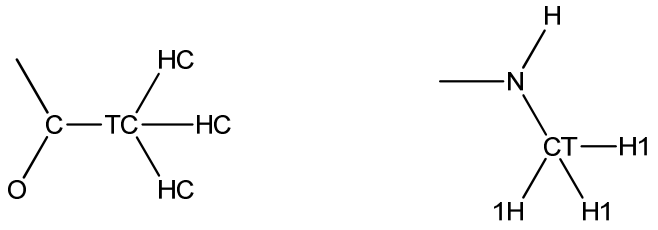
4.2.2.1.2-Changing the force field: the AMBER FF

At this point, an unexpected issue aroused: the original intention was to employ the NorthWest computational Chemistry (NWChem) software package version 6.0 [51] to carry out the hybrid calculations. This was based on the fact that this package was designed to run on high-performance parallel computers, and that it is scalable, treating large problems efficiently. Since the program's documentation indicated that the software package supported the CHARMM force field for classical methods, and no indication contrary to this was found in the section on MM parameters for hybrid QM/MM calculations [52], all the previous work was based on the assumption that the specified FF could be used in the hybrid methods optimization. Since we had no previous experience with this program, even when the final QM region for the hybrid calculation was not yet defined, some experiments were conducted in order to gain a good insight into the requirements to run this program properly. It was at this stage that serious problems were encountered while trying to use the CHARMM force field, since all attempts to prepare the required files ended abruptly with error messages. This problem did not occur when the other supported

force field (the AMBER force field [11,53,54]) was considered, thus indicating that the problem was not in the input files but in the choice of the force field. Trying to solve this problem, a communication with one of the program's developer was initiated, and that is how we came to the knowledge that at this moment, QM/MM calculations with CHARMM force field are not recommended with this software package. Since the other computer program available to us for carrying out hybrid calculations (Gaussian09) does not support the CHARMM force field, it was imperative at this point to change force fields. Since the before mentioned AMBER FF is supported in both programs and, as described in Chapter 2, it is also a force field developed specifically for proteins and nucleic acids, this was the obvious choice.

The parameters for all halide ions compatible with the AMBER force field have already been determined by Joung and Cheatham [35], so it is not necessary to redo the whole determination. Considering the new conditions, the work is resumed from the determination of the QM region for the hybrid calculation.

In the case of the AMBER force field, the cap procedure for the dangling bonds was developed as part of the force field design, involving the use of N-terminal and C-terminal blocking groups (ACE and NME respectively)[53]. These are shown in **Fig. 21**:



ACE NME

Fig. 21– ACE and NME capping residues defined in the AMBER force field

The atom names correspond to the nomenclature used by the AMBER force field.

This implies that all the geometry optimizations have to be redone, since the initial structures have to be the same for both the QM and MM optimization procedures to compare the obtained results. In all cases, the residues considered are the same that were

contemplated in the previous step, but the atoms corresponding to the capping residues are added to the initial structures.

The initial structures and parameter files required for the MM calculation were carried out using the AmberTools 1.5 program, which is associated with the AMBER software [55,56], and considering the ff10 version of the AMBER force field [57]. This latest version of the FF implement the ff99SB protein force field, and the ff99bsc0 force field for DNA/RNA, while using atom and residue names from version 3 of the pdb. The optimizations were carried out employing the SANDER energy minimizer, for a maximum of 100000 cycles or until a root mean square for the Cartesian elements of less than 1.0^{-4} kcal/mole-Å is obtained. Two minimization methods are used: the first 10 cycles are carried out with steepest descent, then switching to conjugate gradient for the rest of the optimization cycles.

The parameters for the QM calculations are the same than in the previous stage.

As in the previous case, the C$\alpha$are held fixed during the geometry optimization process for both methodologies (for the same considerations discussed above).

It is important to keep in mind that, as was mentioned in Chapter 2, a different FF is being used (both in the functional forms and the parameters), even when the initial geometry is the same that in the previous stage, the optimized structures will most likely differ from the ones obtained when the CHARMM force field was considered.

The same starting geometries are considered, but, in this case, the open conformation for the iodide complex was also considered. Even though none of the crystals showed a clear open conformation, in all of them there was enough signal as to model a possible structure for that conformation. In this case, the modeled open conformation from **data set 10** is the one considered as starting structure.

4.2.2.1.2.1- 4Å radius

The region considered is the same as when this comparison was calculated with the CHARMM force field. For the protein-iodide complex, with the loop in the open conformation, the residues considered are **Phe182** (as in the previous case, at this stage this residue is included due to its important role, since its placed further away than the

considered 4 Å), **Gly220**, **Arg221** and **Gln266**. In this system, the total charge of the system is zero.

In **Fig. 22a-d** the optimized structure obtained with the different methodsincluding the ACE and NME capping residues are shown (which are not considered for the rmsd calculation, since they have no biological interest).

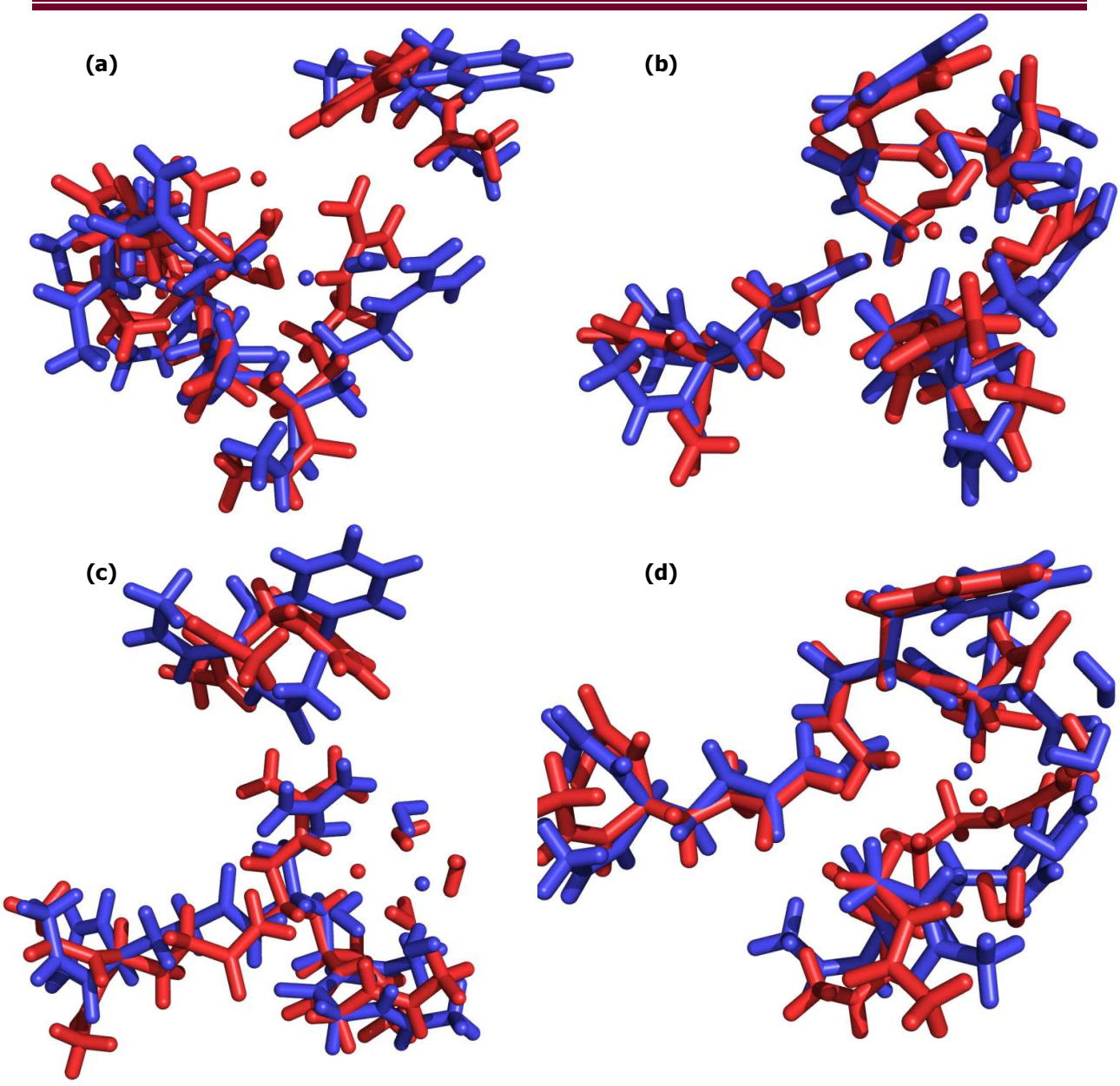


Fig. 22– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a)PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation, (c) PTP1B-I- in open conformation and (d)PTP1B-I- in closed conformation, considering the residues in a 4 Å radius. In the figure, the ACE and NME residues are also shown.

In **Table 16** the rmsd values between the structures optimized with both theory levels and the crystallographic experimental data are detailed. Only the protein residues were considered when calculating the rmsd values (the NME and ACE capping residues were not included, since they provide no useful information).

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Phe182	3.99	3.04
	Gly220 - Arg221	2.95	1.35
	Gln262	1.33	1.03
	Gln266	1.04	1.38
	Br^-	3.06	1.06
	water molecules	2.77	2.97
	total (not including water molecules)	2.74	1.84
PTP1B – Br^- closed conformation	Asp181 - Phe182	0.58	1.14
	Gly220 – Arg221	0.67	0.67
	Gln266	1.12	0.35
	Br -	0.43	0.66
	water molecules	1.48	1.90
	total (not including water molecules)	0.76	0.85
PTP1B – I^- open conformation	Phe182	4.43	3.02
	Gly220 – Arg221	2.92	2.79
	Gln266	0.96	1.94
	I^-	1.26	2.08
	water molecules	3.85	3.42
	total (not including water molecules)	3.12	2.67
PTP1B – I^- closed conformation	Asp181 – Phe182	1.52	2.20
	Gly220 – Arg221	1.05	2.94
	Gln266	1.16	1.35
	I^-	1.07	1.89
	water molecules	2.44	3.78
	total (not including water molecules)	1.28	2.38

Table 16– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 4Å radius.

As expected, the open conformations of both complexes show the higher rmsd values. The explanation probably lies, once again, in the lack of representation of the stabilizing interactions established with surrounding residues (which are not included in the model), since in these cases the Phe182 residues were added even when they are out of the considered radius.

In the case of the closed conformations, both the QM and MM optimized structures are still not close enough to the experimental data as to consider these residues to provide a good representation of the interactions taking place within the catalytic site.

4.2.2.1.2.2-6 Å radius

Since the starting geometries are the same that the ones considered in the previous optimization, the residues included in these models are the same ones discussed in section *4.2.2.1.1.2*. In the particular case of the open loop conformation for the protein-iodide complex, which was not modeled before, the residues considered are:**Trp179**, **Phe182**, **Cys215**, **Ile219**, **Gly220**, **Arg221**, **Gln262** and **Gln266**. The total charge for this system is **-1**. As in the other models, the important cysteine is included in this model, since it is within the considered range.

In **Fig. 23a-d** the superposition of the resulting structures is shown, including the ACE and NME capping residues (which are not considered for the rmsd calculation, since they have no biological interest).
Table **17 shows** the rmsd values calculated for each optimized structured with respect to the empirical (crystallographic) data.

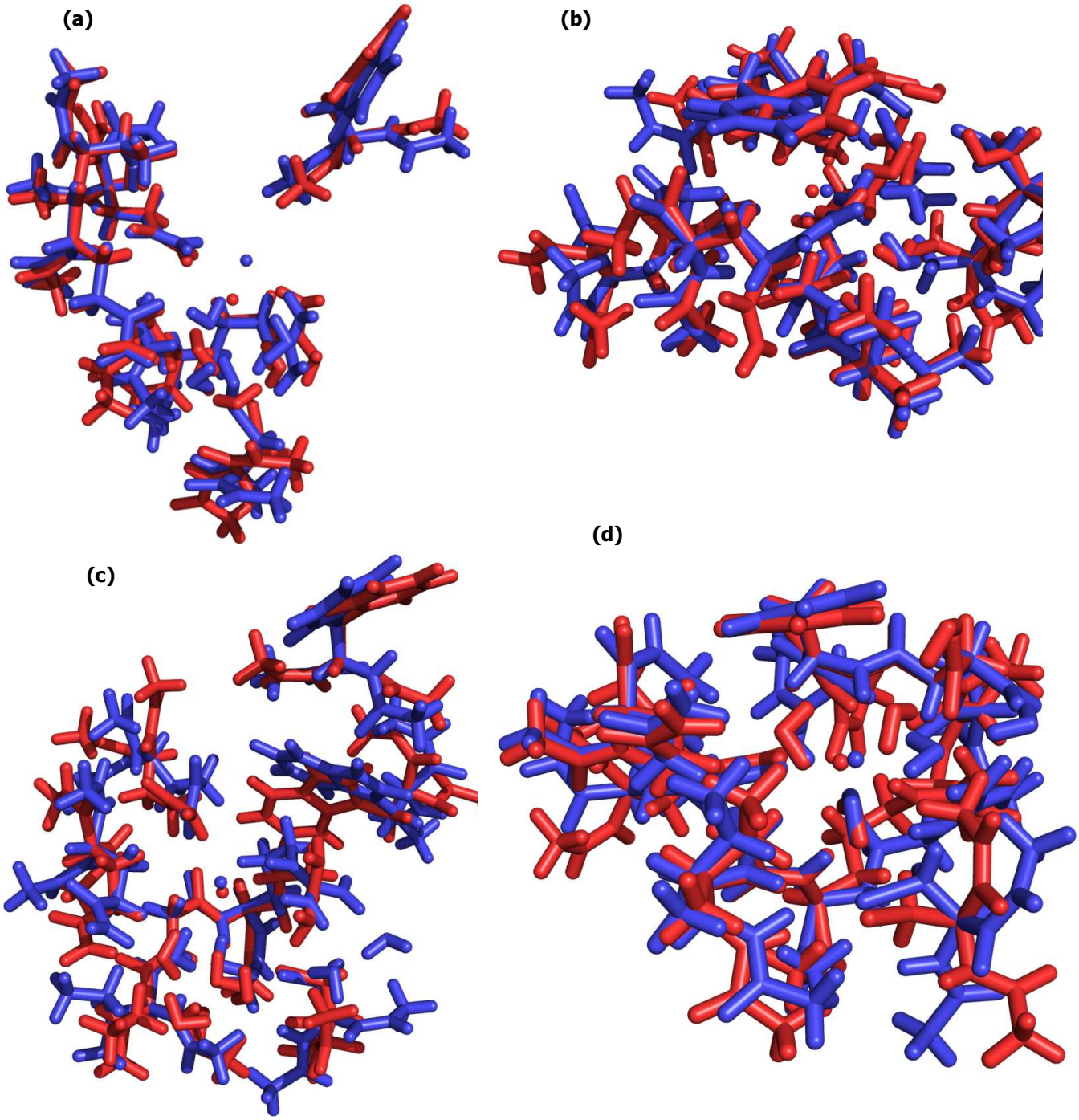


Fig. 23– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a)PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation, (c) PTP1B-I- in open conformation and (d)PTP1B-I- in closed conformation, considering the residues in a 6Å. radius. In the figure, the ACE and NME residues are also shown.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Phe182	2.86	1.99
	Cys215	1.85	1.66
	Gly220 - Arg221	1.38	1.32
	Gln262 – Thr263	0.79	0.84
	Gln266	0.61	0.63
	Br^-	1.51	2.60
	water molecules	2.05	2.34
	total (not including water molecules)	1.60	1.34
PTP1B – Br^- closed conformation	Pro180 - Asp181 - Phe182 – Gly183	0.67	0.90
	Cys215 – Ser216	0.69	1.22
	Ile219 - Gly220 – Arg221	0.73	1.03
	Gln262	2.62	1.39
	Gln266	0.71	0.86
	Br -	0.27	0.97
	water molecules	1.33	1.02
	total (not including water molecules)	1.01	1.04
	Trp179	1.11	2.95
PTP1B – I^- open conformation	Phe182	1.04	2.25
	Cys215	1.06	0.92
	Ile219 – Gly220 – Arg221	1.38	1.28
	Gln262	1.18	1.23
	Gln266	1.31	1.62
	I^-	1.15	1.42
	water molecules	2.87	3.10
	total (not including water molecules)	1.23	1.84
PTP1B – I^- closed conformation	Pro180 - Asp181 – Phe182 – Gly183	1.09	0.99
	Cys215	0.94	1.29
	Ile219 - Gly220 – Arg221	1.15	0.72
	Gln262	1.01	0.90
	Gln266	1.06	0.80
	I^-	0.54	0.90
	water molecules	1.45	1.79
	total (not including water molecules)	1.08	0.90

Table 17– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 6Å radius.

When more residues are included in the model, in three of the four cases an improvement in the reproduction of the experimental data is noticed. The only case this is different, is for the protein-bromide complex in the closed conformation, where the rmsd value when all residues except the water molecules are included is higher in the last model. However, a deeper analysis of the data shows that when the individual residues are considered, an improvement can be seen with respect to the at 4 Å mode, there is only one residue that presents an elevated rmsd value, thus increasing the average value. The problematic residue is Gln262, which was not considered in the first model.

In the other three models, all present lower rmsd values when compared to the previous model, but this improvement in the agreement between the theoretical and experimental data is more pronounced in the case of the models with an open loop conformation. This supports the previously stated hypothesis, in which the lack of stabilizing interactions was the explanation for the high rmsd values.

Overall, the level of agreement between the structure obtained with QM and MM methods is better in this model than in the previous one.

Taking all of the above into consideration, it is decided to carry out a new set of geometry optimizations, considering even larger models.

4.2.2.1.2.3-8 Å radius

In this case, the residues included in each model are those that have at least one atom within an 8 Å radius from the corresponding ion, and the residues considered in the open conformation PTP1B-I $^{-}$ model are: **Leu110**, **Asn111**, **Glu115**, **Lys116**, **Trp179**, **Pro180**, **Asp181**, **Phe182**, **Gly183**, **Cys215**, **Ser216**, **Ala217**, **Gly218**, **Ile219**, **Gly220**, **Arg221**, **Ser222**, **Gly223**, **Thr224**, **Gly259**, **Leu260**, **Ile261**, **Gln262**, **Thr263**, **Ala264**, **Asp265**, **Gln266**, and has a total charge of **-3**. The residues included in each complex are summarized in **Table 18**:

Complex	Residues	Total charge	Water molecules
PTP1B-Br $^-$, open conformation	Glu115, Lys116, Trp179, Pro180, Asp181, Phe182, Gly183, Cys215, Ser216, Ala217, Gly218, Ile219, Gly220, Arg221, Gly223, Thr224, Gly259, Leu260, Ile261, Gln262, Thr263, Ala264, Asp265, Gln266	-3	8
PTP1B-Br $^-$, closed conformation	Leu110, Asn111, Glu115, Lys116, Trp179, Pro180, Asp181, Phe182, Gly183, Val184, Cys215, Ser216, Ala217, Gly218, Ile219, Gly220, Arg221, Ser 222, Gly223, Thr224, Gly259, Leu260, Ile261, Gln262, Thr263, Ala264, Asp265, Gln266	-3	6
PTP1B-I $^-$, open conformation	Leu110, Asn111, Glu115, Lys116, Trp179, Pro180, Asp181, Phe182, Gly183, Cys215, Ser216, Ala217, Gly218, Ile219, Gly220, Arg221, Ser222, Gly223, Thr224, Gly259, Leu260, Ile261, Gln262, Thr263, Ala264, Asp265, Gln266	-3	6
PTP1B-I $^-$, closed conformation	Leu110, Asn111, Glu115, Lys116, Trp179, Pro180, Asp181, Phe182, Gly183, Cys215, Ser216, Ala217, Gly218, Ile219, Gly220, Arg221, Ser222, Gln223, Thr224, Gly259, Leu260, Ile261, Gln262, Thr263, Ala264, Asp265, Gln266	-3	6

Table 18– Residues and water molecules that present at least one atom within a 8 Å radius from the corresponding ion

It is interesting to notice that when this radius around the ion is considered, almost all of the models include the same residues.

An important fact to take into account when considering these new models is that the number of atoms included are over two times the atoms present in the previous ones (in

the 6 Å models there were around 130 atoms, while in the new ones almost 400 atoms are involved). This increase is more pronounced than the one going from the first models to the second ones (the 6 Å models are around 1.5 times than the corresponding 4 Å models). Even though the inclusion of so many atoms presents an advantage from the accuracy point of view, it is important to take into account that these models are largely over the size limit for the electronic structure methods; the inclusion of this amount of atoms could not only make the calculation much longer, but it could even make it impossible.

In fact, only two of the four complexes could be optimized using the QM method, the ones where the closed conformation was represented. For the two open loop models it was not possible to complete the mechanoquantum calculations: in the case of the PTP1B-bromide complex, the optimization kept ending due to an error in the internal coordinate system (several options were tested in order to overcome this problem, but none of them worked as expected. As for the iodide-PTP1B complex in the open conformation, the calculation had just started when analysis of the interaction energies suggested the need to consider the region's charge.

In **Fig. 24a-b** the superposition of the optimized structures for the two closed loop complexes is shown. A visual analysis on the resulting structures already shows a very good superposition between the geometries optimized with both levels of theory (actually, the differences are basically localized on the ACE and NME capping residues, which present no biological interest).

In **Table 19** the corresponding rmsd values are shown, supporting the previously stated observation.

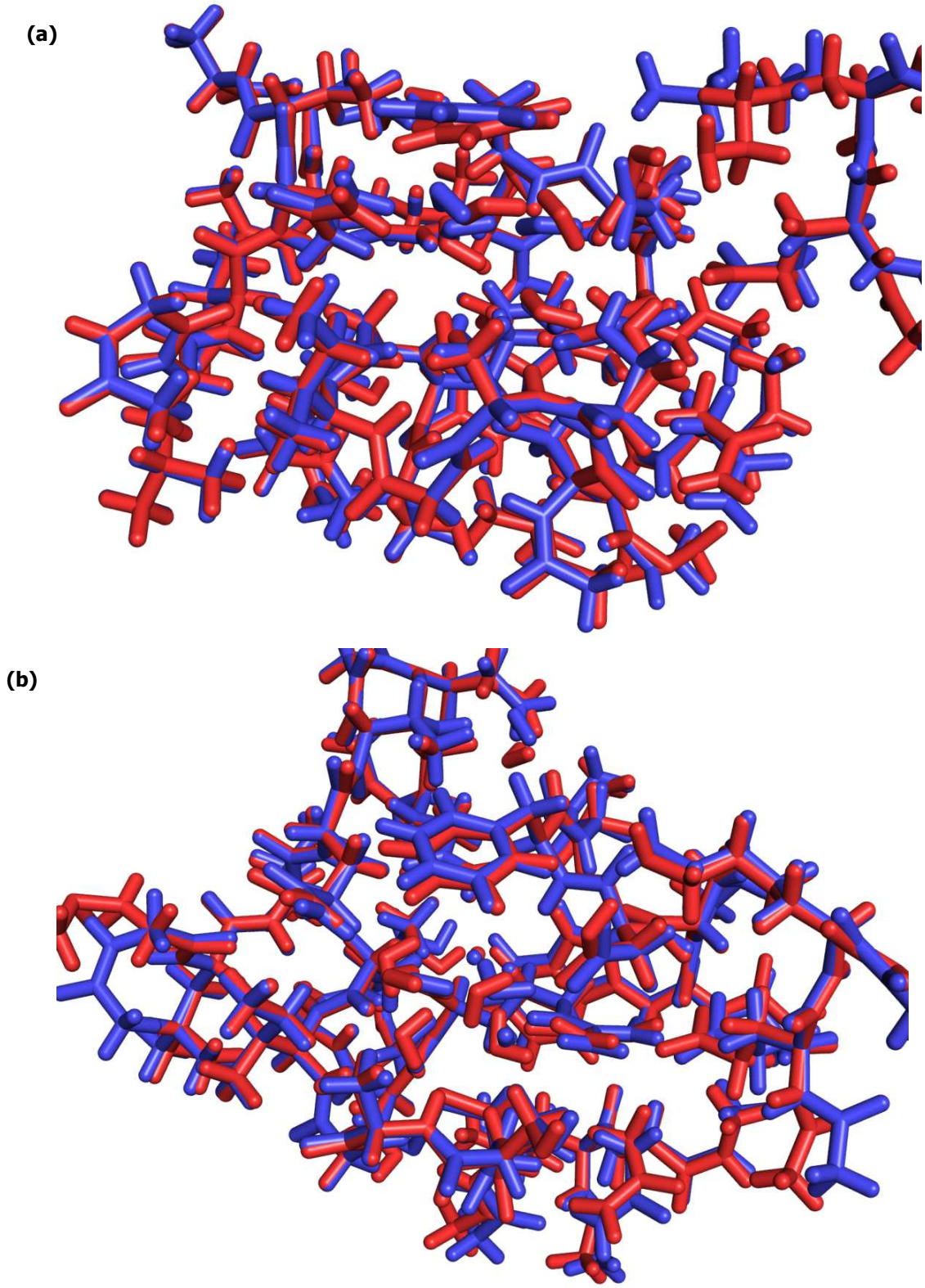


Fig. 24– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a) PTP1B-Br- complex in closed conformation and (b) PTP1B-I- in closed conformation, considering the residues in an8 Å radius. In the figure, the ACE and NME residues are also shown.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- closed conformation	Leu110 – Asn11	0.68	0.66
	Glu115 – Lys116	2.02	0.81
	Trp179 – Pro180 – Asp181 – Phe182 – Gly183 – Val184	0.34	0.36
	Cys215 – Ser216 – Ala217 – Gly218 – Ile219 – Gly220 – Arg221 – Ser222 – Gly223 – Thr224	0.43	0.44
	Gly259 – Leu260 – Ile261 – Gln262 – Thr263 – Ala264 – Asp265 – Gln266	0.49	0.49
	Br -	0.22	0.53
	water molecules	0.94	0.93
	total (not including water molecules)	0.76	0.51
PTP1B – I^- closed conformation	Leu110 – Asn111	0.87	0.48
	Glu115 – Lys116	0.81	0.80
	Trp179 – Pro180 – Asp181 – Phe182 – Gly183	0.38	0.34
	Cys215 – Ser216 – Ala217 – Gly218 – Ile219 – Gly220 – Arg221 – Ser222 – Gly223 – Thr224	0.40	0.40
	Gly259 – Leu260 – Ile261 – Gln262 – Thr263 – Ala264 – Asp265 – Gln266	0.58	0.61
	I^-	0.17	0.30
	water molecules	0.90	1.33
	total (not including water molecules)	0.55	0.52

Table 19– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 8Å radius.

Analyzing the table it becomes evident that not only the difference between the optimized structures and the experimental data diminished (with the only exception of the residues Glu115 and Lys116, in the protein-bromide complex), but also the resulting optimized models are almost completely superimposable (the rmsd values between them and the empirical data are almost identical in most cases).

The results described above would suggest that the best choice as a QM region is to consider the residues that present at least one atom within an 8 Å from the ion; the problem with this model is that, as it has discussed above, the computational cost to

achieve such accuracy is too high. Not only the optimization of the QM models took a long time (around 30 calendar days), but some of the models could not be optimized; hence, the region considered for the high level region in the hybrid calculations should be smaller.

4.2.2.1.2.4-Interaction energies

Along with the rmsd values, the interaction energies in each model were calculated. These energies are calculated as the total QM energy of the model, minus the QM energy corresponding to the isolated ion and the isolated amino acidic system (including the water molecules). It is necessary to carry out a *single point calculation* (calculate the energy of a given structure, without any geometry optimization) for the protein residues and water molecules as one system, and the isolated ion as the other. The energies must be corrected to account for the basis set superposition error, as it was done in the parameterization process. In order to do so, counterpoise calculations were carried out, employing the Gaussian09 program.

Table 20 shows the corrected interaction energy for the models considered so far, and **Fig. 25** presents a graphic of the interaction energy vs. the radius considered when defining the models. Since the systems including every protein residue with at least one atom within an 8 Å radius of the ion have been already discarded as possible QM regions, their interaction energies were not calculated (since they were no longer a possible model, it was not worth the computational cost of calculating these energies).

Radius	System	Interaction energy (kcal/mol)
4 Å	PTP1B-Br^- open conformation	-65.64
	PTP1B-Br^- closed conformation	-34.20
	PTP1B-I^- open conformation	-75.97
	PTP1B-I^- closed conformation	-21.27
6 Å	PTP1B-Br^- open conformation	-11.68
	PTP1B-Br^- closed conformation	57.37
	PTP1B-I^- open conformation	3.61
	PTP1B-I^- closed conformation	57.04

Table 20– BSSE-corrected interaction energies, in kcal/mol, for the 4 Å and 6 Å models.

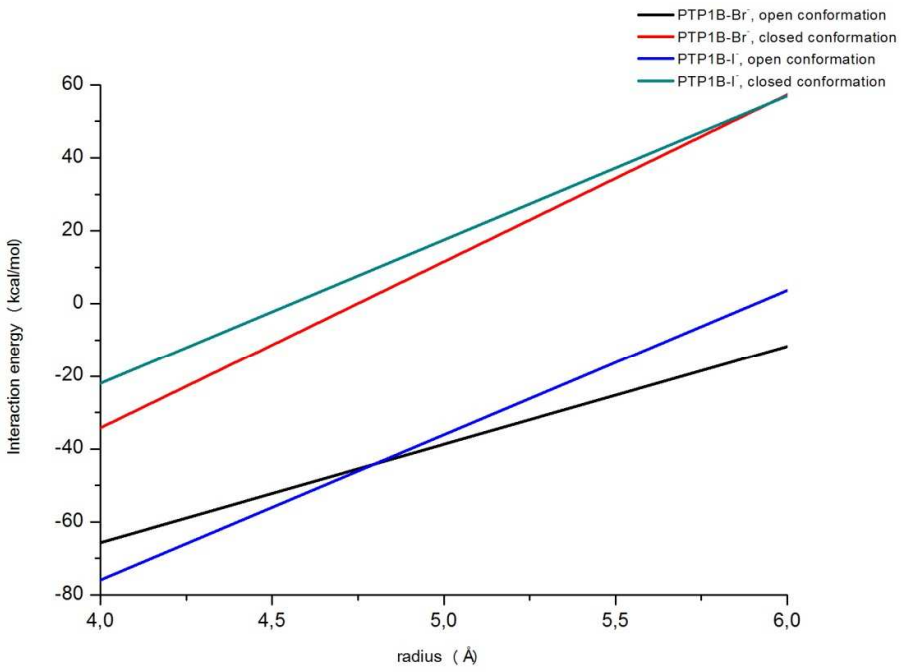


Fig. 25– BSSE-corrected interaction energies, in kcal/mol, vs. radius considered, in Å.

The interaction energies do not present the expected behavior: as the considered radius increases, more residues are included in the models, establishing new interactions with the previously considered peptide residues, water molecules and the ion. These new

interactions should increase the stabilizing interactions with the ion and the rest of the system. On the contrary, the obtained energy profile indicates a stability that diminishes as the number of residues considered increases. When analyzing this unexpected behavior, a possible answer emerged: as the number of residues considered increases, the total charge of the amino acidic residues diminishes, becoming more negative. This means each system presents a bigger negative charge confronted to the negatively charged ion; hence, the increased stabilizing interactions are overcome by the increased coulombic repulsion. This hypothesis is supported by the observation that the lowest energies (the most stable systems) are found for the open loop models at a 4 Å radius, which presented a global charge of zero: the protein residues considered present a positive charge, therefore presenting an attractive interaction with the anion. The iodide model presents a slightly lower energy, but this is probably due to the size of the ion: being the iodide a bigger anion than bromide, it probably establishes a few more stabilizing interactions.

As the models' size increase, so does the negative charge of the included residues, destabilizing the systems (hence, increasing the interaction energy).

In the case of the iodide complex in the open loop conformation at 6 Å, the higher interaction energy when compared to the complex with bromide in the same conformation can easily be explain, once again, by the difference in size between the two ions. Since the iodide is a bigger ion, it probably presents a closer contact with the negatively charged residues, thus increasing the destabilization.

4.2.2.1.3-Taking the charge into account: models with total charge -2

In an attempt to diminish the effect of the repelling negative charges, yet another set of geometry optimizations were carried out. This time, the models where defined in such a way that the total charge (thus, the protein charge, since the ion's is fixed) remains the same in all cases. Trying to keep to the minimum the number of residues that needs to be added, the total charge for the protein is -1 (hence showing a total charge of each model of -2).

4.2.2.1.3.1-4 Å radius, total charge -2

In order to achieve a total charge of -2, it is necessary to add some residues to the previously determined models at 4 Å. Both complex in closed conformation bear a total charge of -1 (the ion's charge, since the residues considered present a neutral charge); the -2 charge is achieved by including the **Cys215** residue in both cases. This approach presents the additional advantage of introducing a catalytically relevant residue in the model. As for the models representing the open conformation of the complex, both of them present a total charge of 0, so two negatively charged amino acidic residues need to be incorporated into the model: once again, the **Cys215** residue is included, but in this case the **Asp181** is added too.

In **Fig. 26a-d** the optimized structures obtained both with QM and MM are superposed, and the rmsd values calculated with respect to the crystallographic structure are shown in **Table 21**.

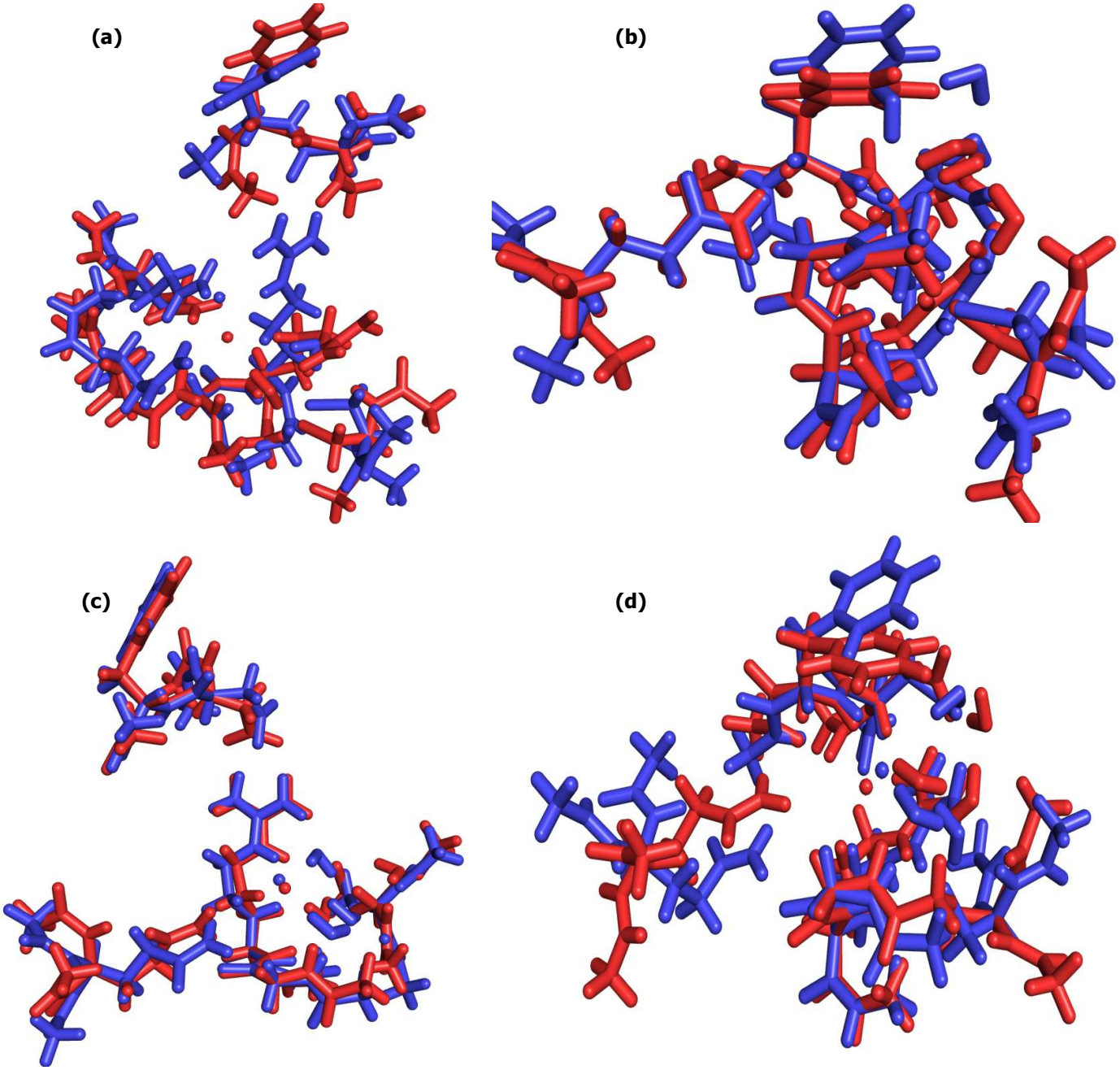


Fig. 26– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a) PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation, (c) PTP1B-I- in open conformation and (d) PTP1B-I- in closed conformation, considering the residues in a 4Å radius. In the figure, the ACE and NME residues are also shown.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Asp181 - Phe182	2.11	0.69
	Cys215	2.77	1.33
	Gly220 - Arg221	1.08	2.32
	Gln262	1.38	1.47
	Gln266	1.24	0.56
	Br^-	0.08	1.91
	water molecules	2.37	2.10
	total (not including water molecules)	1.70	1.50
PTP1B – Br^- closed conformation	Asp181 - Phe182	1.16	1.10
	Cys215	1.26	2.86
	Gly220 – Arg221	0.92	1.14
	Gln266	1.71	1.63
	Br -	0.34	0.93
	water molecules	1.90	3.25
	total (not including water molecules)	1.22	1.51
PTP1B – I^- open conformation	Asp181 – Phe182	1.78	1.79
	Cys215	1.93	0.98
	Gly220 – Arg221	1.57	1.57
	Gln266	1.49	1.49
	I^-	0.83	0.77
	water molecules	2.31	1.67
	total (not including water molecules)	1.66	1.58
PTP1B – I^- closed conformation	Asp181 – Phe182	0.79	2.13
	Cys215	1.15	2.41
	Gln220 – Arg221	0.75	0.88
	Gln266	1.68	2.30
	I^-	0.16	0.84
	water molecules	2.00	1.99
	total (not including water molecules)	1.04	1.87

Table 21– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 4Å radius.

Once again, the higher rmsd values are found for the open loop conformations. This could be indicating that the residues here considered are not enough as to reproduce the

interactions present in the real biological system. In this model, the WPD loop residues are too far away from the other residues as to establish the stabilizing interactions that affect them in the real biological system.

Even though the optimized structures are closer to the empirical data than in the previous modelization studies, they are still not close enough to consider this radius as a good enough representation of the interactions taking place in the biological system. In addition to this, the agreement between the geometries obtained with the different methods employed is still not sufficient as to consider both structures consistent. More residues need to be included in the models considered, attempting to achieve a better representation.

4.2.2.1.3.2-6 Å radius, total charge -2

Both closed conformation models considered in the previous optimization stage presented a total charge of -2, thus there is no need to add any residues.Moreover, since there is no change in the residues considered, there is no need to re-do the calculations. The already presented structures are considered for the comparison. This is not the case for the models representing the open conformation of the complex, which in the previous models presented a total charge of -1: the desired -2 charge is achieved by including the **Asp181** residue.

In **Fig. 27a-d** the optimized structures obtained both with QM and MM are superposed (for comparison's sake, the already shown superposition of the QM and MM structures is reproduced here again), and the rmsd values calculated with respect to the crystallographic structure are shown in **Table 22**.

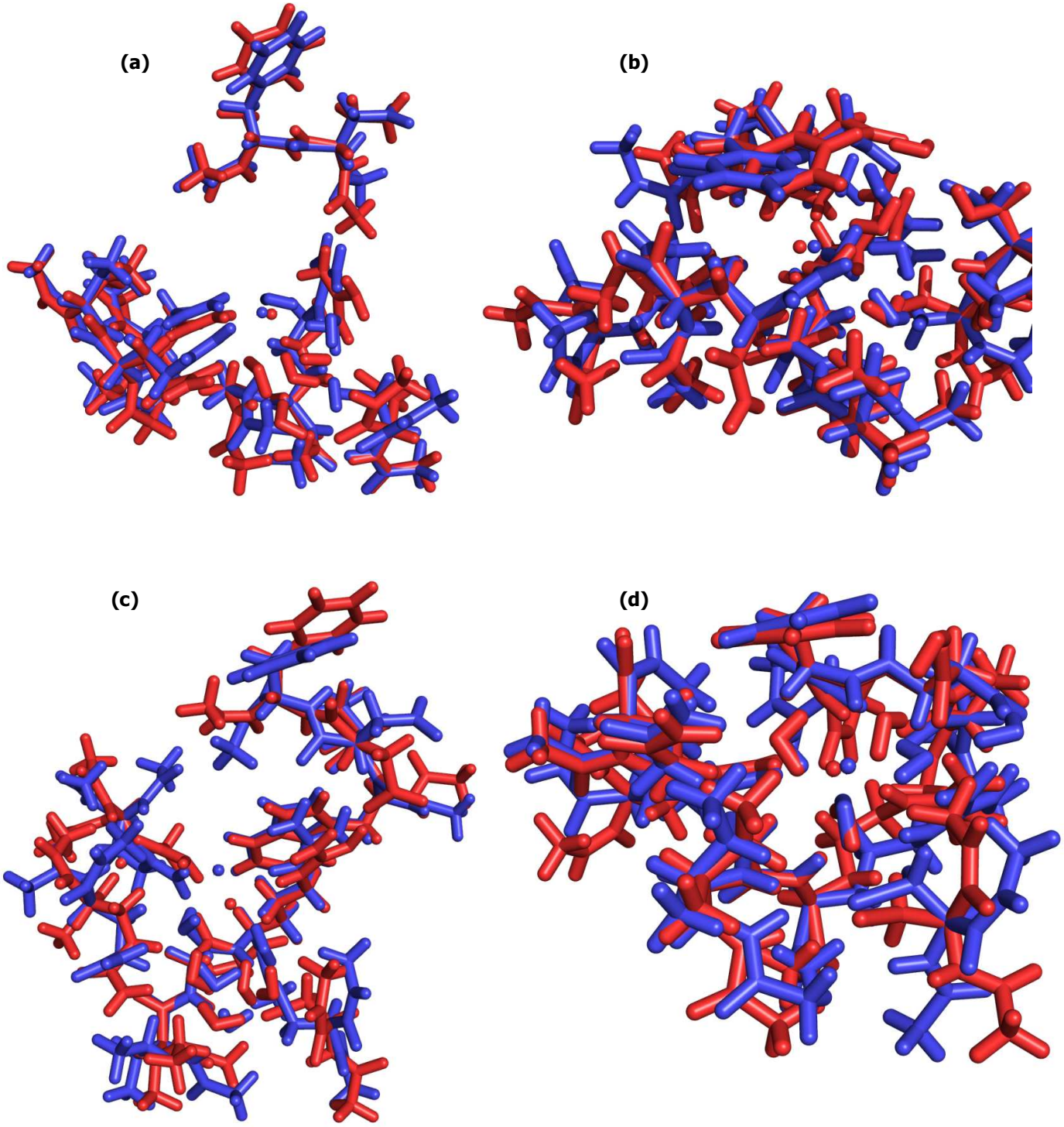


Fig. 27– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a) PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation, (c) PTP1B-I- in open conformation and (d) PTP1B-I- in closed conformation, considering the residues in a 6Å radius. In the figure, the ACE and NME residues are also shown.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br $^-$ open conformation	Asp181 - Phe182	2.03	1.39
	Cys215	1.91	1.36
	Gly220 - Arg221	1.43	2.01
	Gln262 – Thr263	0.57	1.28
	Gln266	0.72	0.81
	Br $^-$	1.46	1.20
	water molecules	1.64	2.56
	total (not including water molecules)	1.64	1.48
PTP1B – Br $^-$ closed conformation	Pro180 - Asp181 - Phe182 – Gly183	0.67	0.90
	Cys215 – Ser216	0.69	1.22
	Ile219 - Gly220 – Arg221	0.73	1.03
	Gln262	2.33	1.39
	Gln266	0.71	0.86
	Br -	0.27	0.97
	water molecules	1.33	1.02
	total (not including water molecules)	1.01	1.04
PTP1B – I $^-$ open conformation	Trp179 – Pro180 - Asp181 – Phe182	1.33	1.19
	Cys215	0.91	1.62
	Ile219 - Gly220 – Arg221	1.67	1.68
	Gln262	1.23	1.20
	Gln266	1.24	1.91
	I $^-$	0.87	2.00
	water molecules	1.90	2.31
	total (not including water molecules)	1.40	1.47
PTP1B – I $^-$ closed conformation	Asp181 – Phe182	1.09	0.99
	Cys215	0.94	1.29
	Gly220 – Arg221	1.15	0.72
	Gln262 – Thr263	1.01	0.90
	Gln266	1.06	0.80
	I $^-$	0.54	0.90
	water molecules	1.45	1.79
	total (not including water molecules)	1.08	0.90

Table 22– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the residues in a 6Å radius.

From the analysis of the previous resultsit is evident that these models show an improvement, both in the reproduction of the crystallographic data as well as in the consistency between the two methodologies. Once again, and as expected, the best reproduction of the experimental data is achieved for the models representing the closed loop conformation (for the reasons discussed above).

As determined from the previous set of geometry optimization calculations, it is not convenient to consider a larger model, since the increase in the computational cost is proportionally bigger than the improvement obtained.

4.2.2.1.3.3-Interaction energies

The interaction energies in the new systems were calculated, in the same way as described in section *4.2.2.1.2.4* .

Table 23 shows the corrected interaction energy for these new models, and **Fig. 28** presents a graphic of the interaction energy vs. the radius considered when defining the models.

Radius	System	Interaction energy (kcal/mol)
4 Å	PTP1B-Br^- open conformation	29.70
	PTP1B-Br^- closed conformation	42.76
	PTP1B-I^- open conformation	37.62
	PTP1B-I^- closed conformation	47.52
6 Å	PTP1B-Br^- open conformation	48.05
	PTP1B-Br^- closed conformation	57.38
	PTP1B-I^- open conformation	56.09
	PTP1B-I^- closed conformation	57.04

Table 23– BSSE-corrected interaction energies, in kcal/mol, for the models with global charge -2.

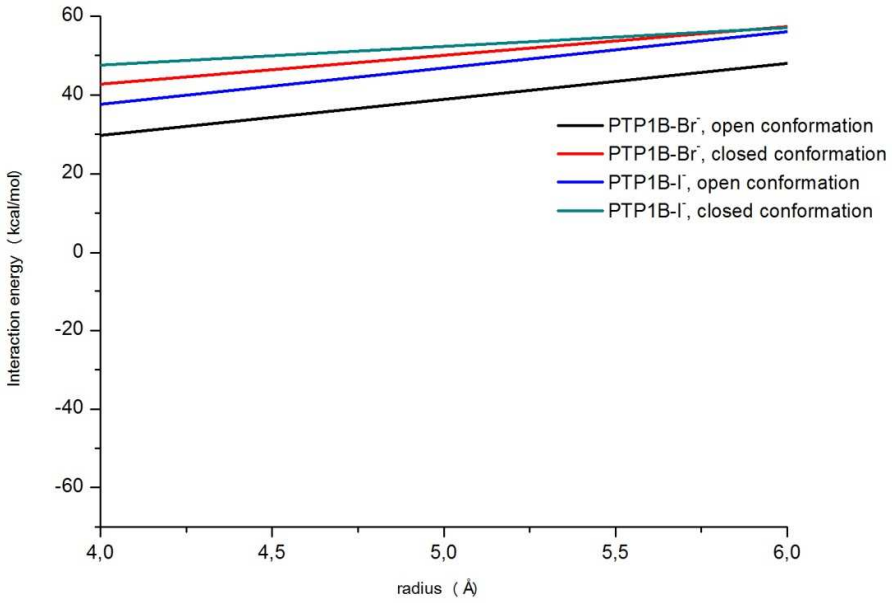


Fig. 28– BSSE-corrected interaction energies, in kcal/mol, vs. radius considered, in Å.

The first observation that arises from the analysis of the previous table is the fact that in this case there are no negative values for the energy. This is due to the fact that in these models there are always repelling coulombic interactions, since both the amino acidic residues and the ion present negative charges. The fact that the lower energies are found for the models representing the open loop conformations is actually expected, since in those systems the negative charges have more space to rearrange themselves, diminishing the repulsive forces.

Another interesting observation that can be made, is the fact that in this case the interacting energies are more similar than when the previous models where considering, thus sustaining the previously presented hypothesis. Moreover, the difference in these energies diminishes when the 6 Å model is considered, showing a converging tendency. This is coherent with the fact that, as a consequence of adding residues to achieve the -2 total charge, the different models presented almost the same residues, which derived in the same interactions. The protein-bromide complex representing the open conformation loop presents the lowest energy, and this can easily be explained by taking into account not only the capacity of rearrangement of the charges (as already discussed), but the fact

that bromide is quite smaller when compared to iodide, thus allowing it a greater movement within the catalytic site, diminishing the repulsive interactions.

4.2.2.1.4-Final model for the QM region

Since the final objective is to compare the ONIOM results for all four systems, the QM region should be the same in all cases. Despite the fact that the residues included to achieve the total charge of -2 has increased the similarity of the models considered, there are still some residues included in one model that are not considered in another one. This is why a final set of geometry optimization is carried out: in this case, any residue including at least one atom within a 6 Å radius from the corresponding ion is included in all models. Hence, all systems include the same residues, thus allowing the results comparison, both in structure and energetic considerations.

The residues included in this final set of optimizations are:**Trp179**, **Pro180**, **Asp181**, **Phe182**, **Gly183**, **Cys215**, **Ser216**, **Ile219**, **Gly220**, **Arg221**, **Gln262**, **Thr263**and **Gln266**. Another consideration is to include the same number of water molecules in the model (since they may play a stabilizing role in the biological structure). In all cases, the total charge is -2.

In such a system, all the catalytically important residues are included, such as Asp181, Phe182, Cys215 (responsible for the nucleophilic on the ligand's phosphate group) and the Arg221 (which plays an important role in the stabilization of the negatively charged cysteine) [48,50].

In Fig. **29a-d** the superposition between the geometries optimized with the different methods is presented, and the comparison between these structures and the empirical data is shown in **Table 24**.

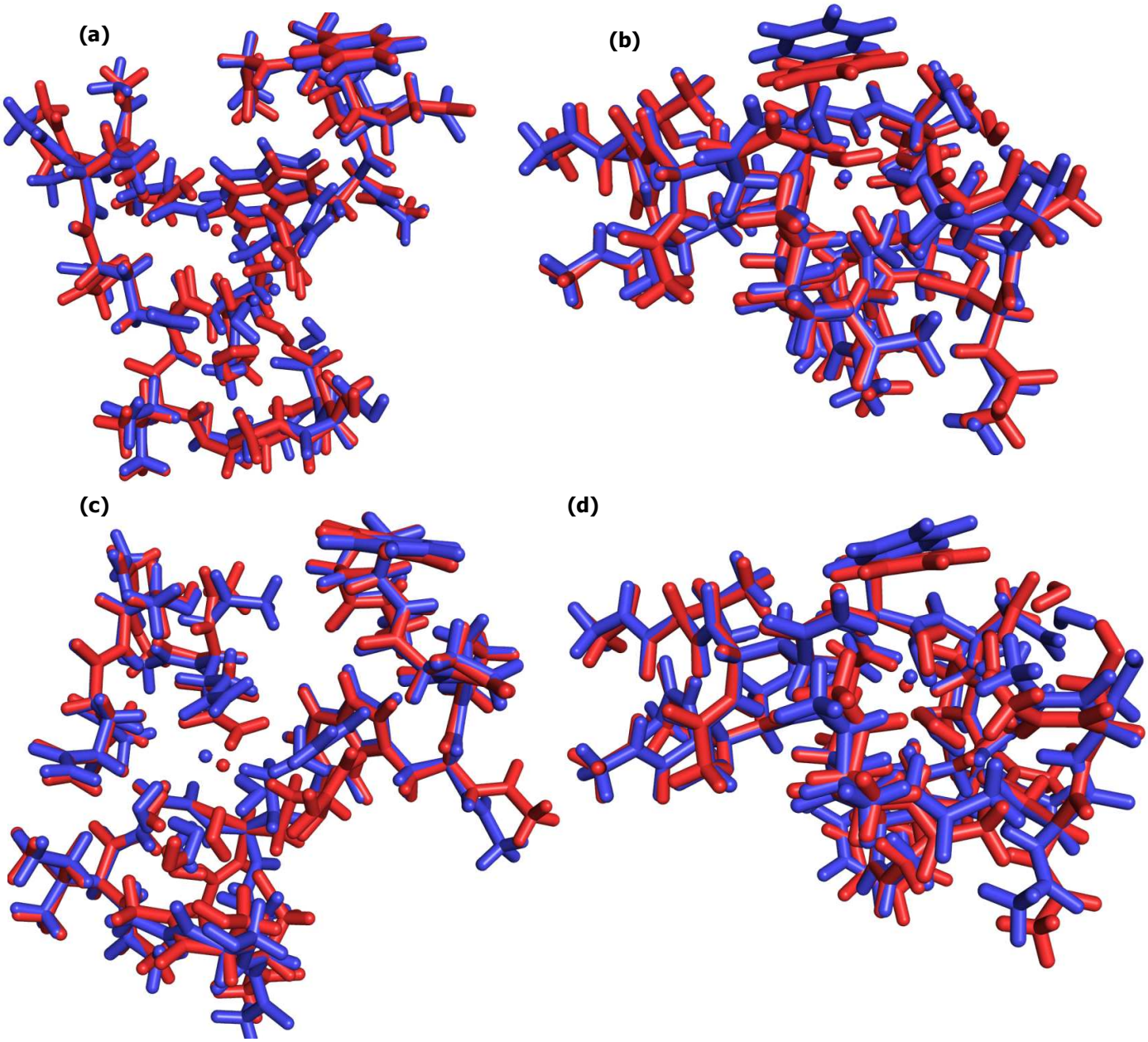


Fig. 29– Superposition of the optimized structures obtained with QM method (in red) and MM method (in blue) for (a) PTP1B-Br- complex in open conformation, (b) PTP1B-Br- complex in closed conformation, (c) PTP1B-I- in open conformation and (d) PTP1B-I- in closed conformation, considering the final models. In the figure, the ACE and NME residues are also shown.

System	Residues	rmsd (Å)	
		QM	MM
PTP1B – Br^- open conformation	Trp19 – Pro180 - Asp181 – Phe182 – Gly183	1.33	1.31
	Cys215 – Ser216	0.76	0.68
	Ile219 - Gly220 - Arg221	1.59	1.43
	Gln262 – Thr263	0.88	0.57
	Gln266	1.00	0.48
	Br^-	2.10	2.85
	water molecules	1.29	1.16
	total (not including water molecules)	1.27	1.16
PTP1B – Br^- closed conformation	Trp19 – Pro180 - Asp181 – Phe182 – Gly183	0.63	0.52
	Cys215 – Ser216	1.36	0.90
	Ile219 - Gly220 - Arg221	0.47	0.60
	Gln262 – Thr263	0.68	0.43
	Gln266	0.18	0.18
	Br^-	0.29	0.45
	water molecules	1.51	1.51
	total (not including water molecules)	0.69	0.56
PTP1B – I^- open conformation	Trp19 – Pro180 - Asp181 – Phe182 – Gly183	1.15	1.21
	Cys215 – Ser216	0.96	1.06
	Ile219 - Gly220 - Arg221	0.97	1.79
	Gln262 – Thr263	0.82	0.75
	Gln266	0.44	0.60
	I^-	0.57	0.92
	water molecules	1.44	1.78
	total (not including water molecules)	0.99	1.28
PTP1B – I^- closed conformation	Trp19 – Pro180 - Asp181 – Phe182 – Gly183	0.48	0.84
	Cys215 – Ser216	0.78	0.98
	Ile219 - Gly220 - Arg221	0.50	0.60
	Gln262 – Thr263	0.59	0.68
	Gln266	0.54	0.55
	I^-	0.09	0.36
	water molecules	1.55	1.69
	total (not including water molecules)	0.55	0.76

Table 24– rmsd values between the optimized geometries at the different theory levels and the corresponding crystallographic structures, considering the final models.

The superimposed structures as well as the rmsd values presented above show that these models are a good choice as QM region in the ONIOM calculation. The inclusion of these residues results in a good reproduction of the empirical data, thus indicating that the important interactions are included. At the same time, a good agreement between the QM and MM optimized geometries is achieved, thus ensuring a consistent calculation in the limit between the two layers studied at different theory levels.

4.2.2.2-Ongoing hybrid ONIOM calculations

Having determined the region to be studied at a Quantum Mechanics level of theory is now possible to move forward with these calculations.

The hybrid ONIOM calculation is implemented in the Gaussian09 program and, since this is a program widely used in molecule modeling, it was our choice for carrying out these calculations. The previously defined QM region (the *model* system in the hybrid calculation, as defined in Chapter 2) was studied at the same level of theory employed for its determination (that is, a DFT level of theory, employing the M06 density functional, and considering the DGDZVP basis set for the anions' Kohn-Sham orbitals, and the 6-31G* basis set for the rest of the system). The rest of the system (the *real* system) was studied at the MM level of theory, using the version of the AMBER force field included in the program, ff98. Since the bromide and iodide parameters are not included in that version of the force field, they had to be included. The Joung and Cheatham parameters [35] were considered, the same that were employed in the determination of the high level region.

As discussed in Chapter 2, the use of electronic embedding provides a better representation of the interactions, but it is recommended to optimize first employing mechanical embedding and then perform a second optimization using electronic embedding, using the optimized structure as a starting point, since this approach is more efficient. Taking this recommendation into account, the ONIOM hybrid calculations are currently being carried out, employing the mechanical embedding scheme. Due to the amount of time required to complete these calculations it is not yet possible to draw any conclusions from them.

4.2.3- Molecular Dynamics simulations of the protein-halide complexes – Production run

In biological systems it is not enough to analyze the crystallographic or optimized geometries to gain insight into the structural basis of the potency and selectivity of any compound for a given protein. To investigate the importance of protein flexibility in any ligand binding event, Molecular Dynamics (MD) simulations need to be carried out. This is particularly important in PTP1B complexes, given the inherent flexibility of the WPD loop and the role it plays in the protein's function. So, in order to achieve further information on the possible reasons explaining the different behavior of PTP1B-bromide and PTP1B-iodide complexes, MD simulations were carried out for both of them.

The initial structures for these simulations were the obtained through the X-ray diffraction experiment, including all the crystallographic water molecules; in both cases, the geometry considered was the one where the WPD loop is present in the closed conformation. All simulations were carried out in an explicit aqueous environment, represented by a TIP3P waterbox, (a water model specially parameterized to be used with the SHAKE algorithm). Since the crystallographic structure has no information regarding the position of the hydrogen atoms, and these were added automatically by the AmberTools program while preparing the initial structure and parameter files (the ff10 AMBER force field was employed), a short minimization (1000 steps) needs to be performed in order to remove any bad contacts that may lead to unstable molecular dynamics. The SANDER module of the AMBER10 program is used in this minimization stage.

The initial structures are the result of the protein's conformation and stabilization interactions at 277 K (temperature at which the crystals were grown).Since the object of these simulations is to analyze the interactions occurring at physiological conditions, it is necessary to heat up the system until the normal physiological temperature, 310 K. This is done slowly, over 40 ps in a total of four stages (considering a coupling constant of 1.0 ps), using periodic boundary conditions at constant volume, with the AMBER10 program. Doing the heating in stages allows the system to equilibrate at each step, thus diminishing the chances of the whole system blowing up.

The final step of the simulation was to run a production simulation at a constant temperature of 310 K, employing the Berendsen temperature coupling [58]. In this approach, the system is coupled to an external heat bath with a fixed temperature T_0, and the velocities are scaled at each step, such that the rate of change of temperature is proportional to the difference in temperature:

$$\frac{dT(t)}{dt} = \frac{1}{\tau}(T_0 - T(t)$$

(Eq. 99)

where τ is the coupling parameter, which determines how tightly the bath and the system are coupled together [59,60]. For the data collection for analysis stage, a coupling parameter of 0.5 ps was chosen, since now that the system was heated up it appears to be more stable, thus allowing the use of a more closely coupled thermostat. The PME method for long range interactions was employed and at this stage, the simulation was carried out at periodic boundary conditions at constant pressure, in order to get a proper density. The data collecting phase of the MD run was carried out considering a 2 fstimestep, since the SHAKE constraining algorithm was employed. The nonbonded cutoff distance was 12 Å, and the maximum distance between atom pairs that were considered during the pairwise summation was 10 Å.

To evaluate the stabilization of the system for the two systems considered, both the temperature and total energy fluctuation during the whole simulation are analyzed. In **Fig 30** the temperature for both system through the whole length of the simulation, while in **Fig.31** the potential energy for the same systems is displayed.

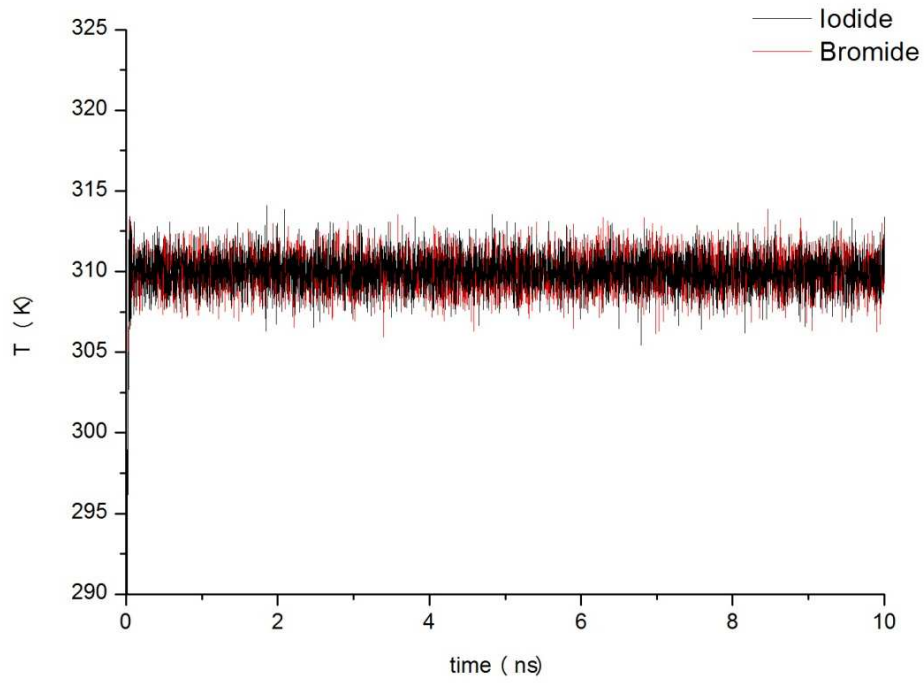


Fig. 30– Temperature fluctuationsfor both protein-halide complexes throughout the simulation.

From the previous figure it is obvious that the temperature oscillated around the desired value of 310 K during the whole simulation.

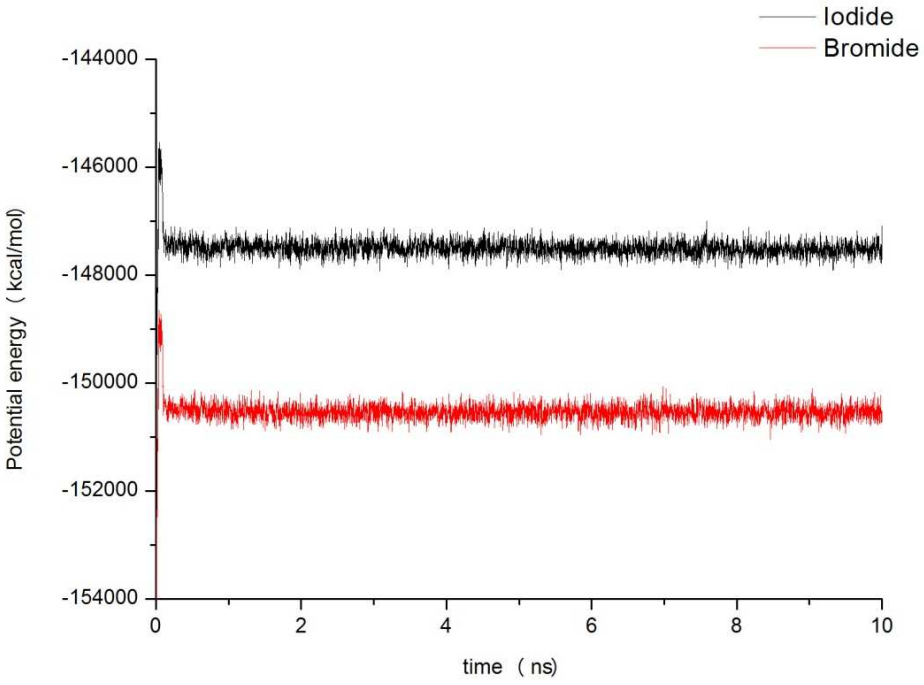


Fig. 31 – Potential energy for both systems during the MD simulation.

As it becomes evident from the previous figure, once the equilibrium is reached, both systems remain stable during the whole simulation runs. Hence, it is possible to obtain reliable information from the analysis of these MD simulations.

In the next figure (**Fig. 32**), the rmsd values for the whole protein in both complexes are shown, compared to the initial structure. This values where calculated considering both the main and side chain non-hydrogen atoms.

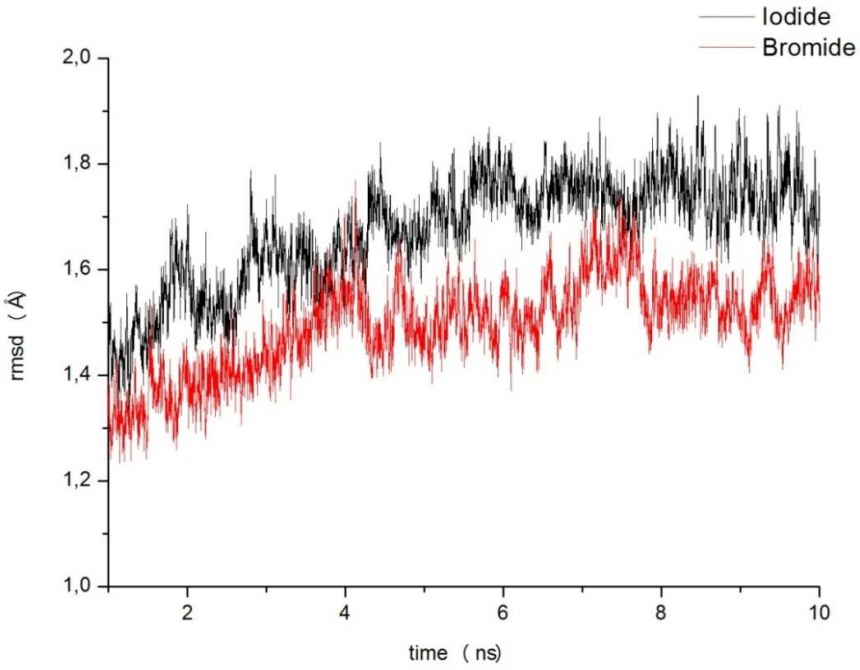


Fig. 32 – rmsd values for the protein atoms with respect to the initial structure.

In both cases the rmsd values present an increasing tendency at the beginning of the simulation, but tend to stabilize after the eight nanosecond of MD run. At this point, a particular fact should be noted: in this same chapter it was stated that an acceptable rmsd value for most protein simulations is 1.5 Å, in that case only the backbone atoms were considered. In the present simulations, due to the characteristics of the studied systems, the side chain atoms were included as well, and these atoms present a higher mobility than the ones in the main chain. In such cases, an rmsd value of 2 Å is still considered acceptable. Since in both cases the rmsd values stabilize at values below that limit (1.5 Å for the protein-bromide complex, and 1.7 Å for the iodide complex), the simulations can

be considered adequate. In any case, since the stabilization is achieved towards the end of the simulation, it is necessary to repeat the molecular dynamics simulation, but for a longer time (at least 20 ns).

It is interesting to notice that, while the iodide complex shows a regular oscillation around the mean value, in the case of the PTP1B-Br^- complex there is a slightly more pronounced variation around seventh nanosecond of simulation, where an increase in the rmsd value is noticed, and it goes back to oscillate around the mean value at the eighth nanosecond. This could be an indication of an important conformational change during this time; once again, a longer simulation is needed in order to actually define whether that is the case or not.

Meanwhile, the other two regions of interest regarding their variation during the simulation are the behavior of the ion, and, of course, an insight into the WPD loop. The rmsd values both the ions and the residues included in the WPD loop are shown in **Fig. 33** and **Fig. 34** respectively.

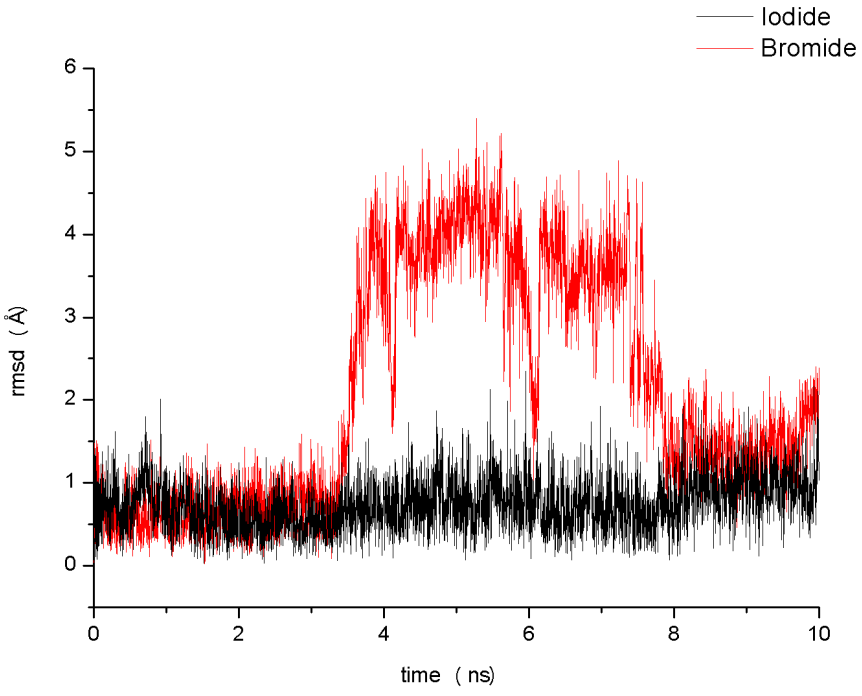


Fig. 33 – rmsd values for the both ions with respect to the initial structure.

During most of the simulation, both ions present quite a stable performance, with a low rmsd value (the mean value is 0.8 Å, indicating that they do not deviate too much from their original position), but their behavior are quite different in the 4 ns to 8 ns lapse. During this time, while the iodide stays stable in its position in the catalytic site, the bromide shows a large deviation, up to 4.5 Å. This is a sign of the existence of strong interactions between the iodide ion and the surrounding residues, which keep the anion in the complexed position. On the case, the bromide complex, even though it is evident that some interactions are acting on the ion, keeping it in the catalytic cleft during most of the simulation time, these interactions would not be as strong as in the iodide's case, thus explaining bromide's larger drift from the initial position. The existence of such forces in this last complex is sustained by the fact that the anion goes back to its original position, instead of leaving the catalytic site completely.

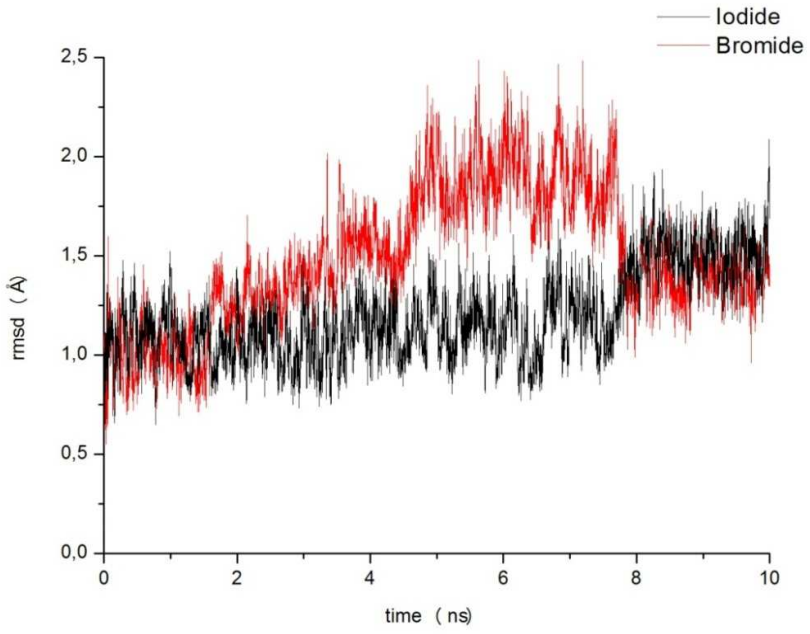


Fig. 34 – rmsd values for the WPD loop (residues 177 to 184) with respect to the initial structure.

As expected, the oscillations of the rmsd values around the mean value (around 1.3 Å for this particular region) are more pronounced in both cases than when the whole protein is

considered. This is a clear indication of the loop's flexibility, and is coherent with the findings of the previous studies already mentioned in this chapter.

When these residues are considered, once again a particular behavior can be observed during the 4 – 8 ns lap. Even when at the beginning of the simulation and towards the end the mean values for the WPD loop in both complexes is very similar (practically the same), this does not happen during that particular period of time. Then, both complexes present different tendencies: while the WPD loop in the protein-iodide complex remains in the same range of rmsd values, the bromide complex shows a clear leap, increasing the rmsd values up to a maximum of 2.5 Å, and a media of 2 Å. This is probably indicating that a big conformational change is taking place during that time. The most interesting part of this observation lies in the fact that the movements of the ion and the WPD loop in the bromide complex seem to be coordinated, since they both occur at the same time. This is consistent with the experimental observations: there is no clear indication of a conformational change in the iodide complex (which, empirically, was only found in the closed conformation), while there are indications of such conformational change for the bromide complex (found in both conformations in the X-ray diffraction experiment).

Having found possible evidence of a conformational change during the dynamics simulations, the next step was to find out if it was the conformational change we are looking for. The rsmd values shown above only indicate a larger movement of both the WPD loop and the ion during that time, but they show no information regarding the characteristic of the movements. That is why both trajectories were compared with the crystallographic structure of each complex in the open and in the closed conformation. The rmsd values for the trajectories of both the protein-bromide and protein iodide complexes, consideringthe available X-ray structures (open and closed conformation) as referenceare shown in **Fig.35** and **Fig. 36**. Only theresidues corresponding to the WPD loopare considered, since they are the only ones that present any evidence of a conformational change: when the entire protein is considered, no such indication is found. This is probably indicating that these residues are involved in an important conformational change, while the rest of the protein maintains the same structure (this is consistent with the evidence found in the mentioned previous works).

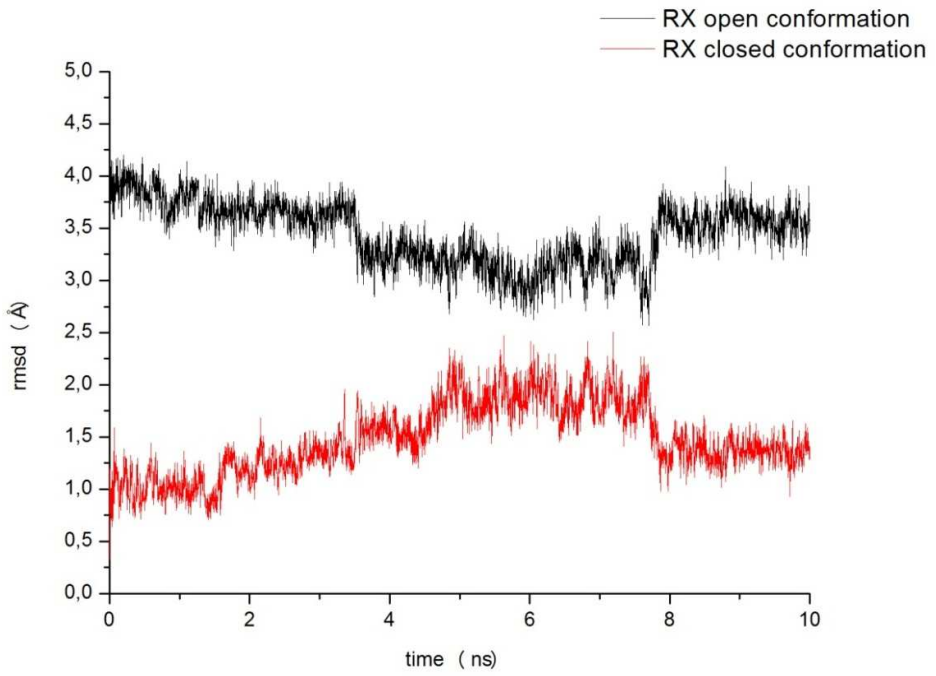


Fig. 35 – rmsd values for the WPD loop (residues 177 to 184) in the bromide complex, with respect to the two available crystallographic structures (complex in both open and closed conformation).

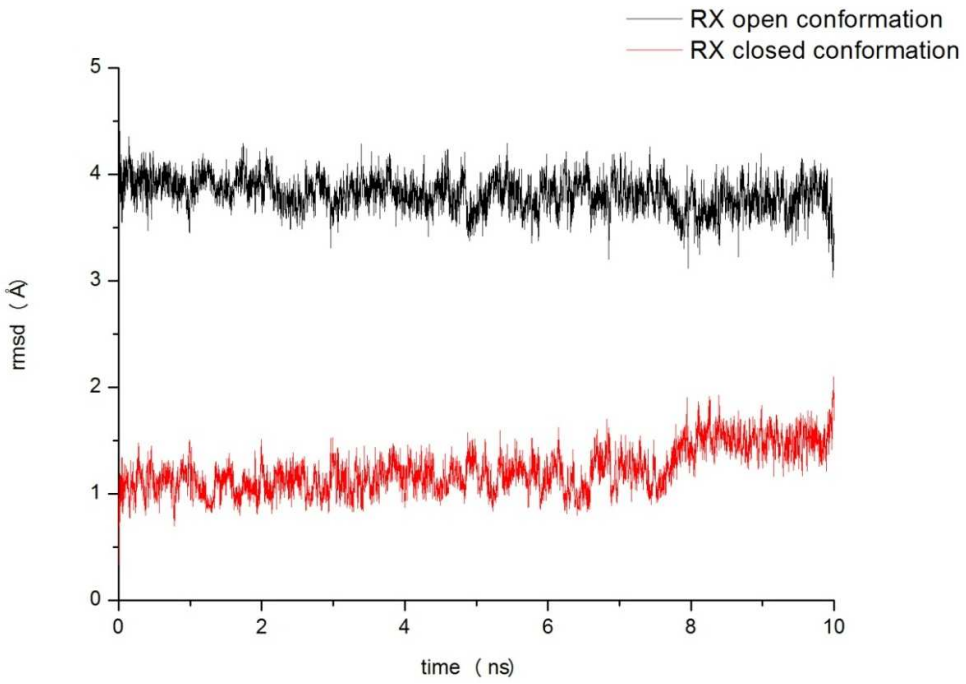


Fig. 36 – rmsd values for the WPD loop (residues 177 to 184) in the iodide complex, with respect to the two available crystallographic structures (complex in both open and closed conformation).

In both cases, the rmsd values are lower when the closed conformation is considered, consistent with the fact that these structures were the ones used as starting geometries. For the bromide simulation run, it is revealing to notice that during the 4 ns to 8 ns period of time (the same where the bigger rmsd for this region were observed) the rmsd values with respect to the closed conformation increase, while decreasing the deviation from the open conformation. This is a clear indication that during that time the observed conformational change is from the closed to the open conformation, as suspected. Moreover, in the case of the iodide complex, where no such conformational change was noticed, the distance between the simulation trajectories to both empirical structures remains stable during the whole run, as expected. A slight increase is noticed by the end of the simulation (after 8 ns), much smaller than the one found for the bromide complex. Since the run finishes only 2 ns after that, there is not enough information as to draw any conclusions from this fact. A longer simulation run is needed to see if this has any conformational consequences.

Since the cysteine residue in the bottom of the catalytic cleft plays a key role in the enzyme's mechanism, binding the substrate, it may also play a part in the stabilization of the halide complexes. That is why the behavior of this residue during the simulation is also analyzed. **Fig. 37** represents the rmsd values for this residue during both complexes' simulations.

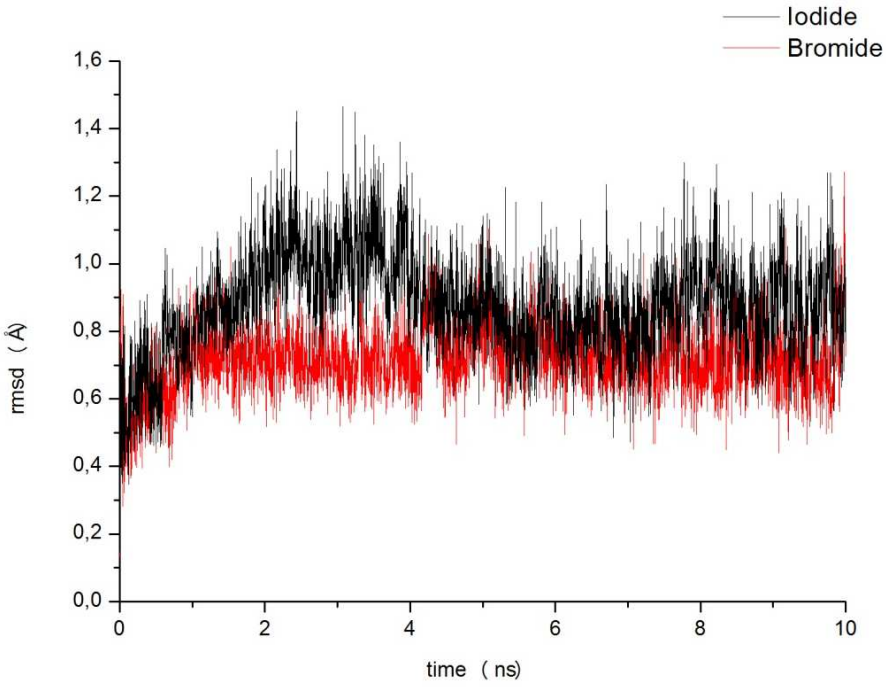


Fig. 37 – rmsd values for the Cys215 residue during the MD simulations in both complexes.

The Cys215 residue does not seem to present a conformational change during the length of both simulation runs. This could be an indication that this residue would not be involved in the forces and interactions that determine the open or closed loop conformation in the PTP1B-halide complexes. Even though there is a slight increase in the rmsd values for this residue in the iodide complex (no difference is found for the bromide complex), this occurs between the 2 ns and 4 ns of the simulation run, before the stabilization of the rmsd values was observed. Hence, it is not reliable to extract any structural conclusions with this data. In addition to that, this motion (which only presents a 0.4 Å amplitude, quite small compared with the changes in the WPD loop and ion positions) is not correlated with any other motion within the studied regions. All of the above considered, this rmsd variation is probably meaningless regarding the interactions of interest.

In order to obtain more reliable information, a larger MD simulation must be carried out; also a targeted Molecular Dynamics (tMD) study [61] should be conducted, allowing the study of the pathway of the desired conformational change. This analysis could allow us to determine the energetic barrier between the two conformational thus providing a better explanation on the observed experimental and theoretical results.

REFERENCES

[1] Barford, D.,Flint, A. J.and Tonks, N. K., *Science263*,**1994,** 1397-1404.

[2] *RCSB PDB Protein Data Bank*; http://www.pdb.org/pdb/home/home.do; August 15th.,2011

[3] *R-factor calculations and their significance*; http://bass.bio.uci.edu/~hudel/mbb254/lecture2/lecture2_1.html; August 20th.,2011

[4] Brünger, A. T.and Rice, L. M. Crystallographic refinement by simulated annealing: methods and applications, in *Methods in Enzymology*; Carter, C. W., Sweet, R. M., Eds.; Academic Press, 1997; Vol. 277; p 243-269.

[5] Brünger, A. T. Free R value: cross-validation in crystallography, in *Methods in Enzymology*; Carter, C. W., Sweet, R. M., Eds.; Academic Press, 1997; Vol. 277; p 366-396.

[6] *Macromolecular crystallography*; http://www.proxychem.com/macromolecular_crystallography.html; August 20th.,2011

[7] *Judging the quality of macromolecular models - a glossary of terms from crystallography, NMR and homology modeling*; http://spdbv.vital-it.ch/TheMolecularLevel/ModQual/; August 20th.,2011

[8] Altman, R. B.,Hughes, C.and Jardetzky, O. "Compositional characteristic of disordered regions in proteins," Stanford University, 1994.

[9] Li, L.and Dixon, J. E., *Semm. Immun.12*,**2000,** 75-84.

[10] Pannifer, A. D. B.,Flint, A. J.,Tonks, N. K.and Barford, D., *J. Biol. Chem.273*, *17*, **1998,** 10454-10462.

[11] Ponder, J. W.and Case, D. A., *Adv. Protein. Chem.66*,**2003,** 27-86.

[12] Zhu, X.,Lopes, P. E. M.and MacKerell, A. D., *Wiley Interdiscip. Rev. Comput. Mol. Sci.00*,**2011,** 1-19.

[13] Zhu, Y.,Su, Y.,Li, X.,Wang, Y.and Chen, G., *Chem. Phys. Lett.455*,**2008,** 354-360.

[14] Brooks, B. R.,III, C. L. B.*et al.*, *J. Comput. Chem.30*, *10*, **2009,** 1545-1614.

[15] MacKerell, A. D., 1999.

[16] MacKerell, A. D.,Bashford, D.*et al.*, *J. Phys. Chem. B102*,**1998,** 3586-3616.

[17] MacKerell, A. D. Atomistic models in force fields, in *Computational biochemistry and biophysics*; Becker, O. M., MacKerell, A. D., Roux, B., Watanabe, M., Eds.; Marcel Dekker, Inc.: New York, 2001.

[18] Levine, I. N. *Quantum chemistry, 5th. Ed.*; (Prentice Hall, 1999).

[19] Jensen, F. *Introduction to computational chemistry, 2nd. Ed.*; (Wiley, 2007).

[20] Møller, C.and Plesset, M. S., *Phys. Rev.46*,**1934,** 618-622.

[21] Head-Gordon, M.and Pople, J. A., *Chem. Phys. Lett.153*, *6*, **1988,** 503-506.

[22] Yurieva, A. G.,Poleshchuk, O. K.and Filimonov, V. D., *J. Struct. Chem.49*, *3*, **2008,** 548-552.

[23] Poleshchuk, O. K.,Yureva, A. G.,Filimonov, V. D.and Frenking, G., *J. Mol. Struct. (THEOCHEM)912*,**2009,** 67-72.

[24] Miertuš, S.and Scrocco, E., *Chem. Phys.55*,**1981,** 117-129.

[25] Tomasi, J.,Mennucci, B.and Cammi, R., *Chem. Rev.105*,**2005,** 2999-3093.

[26] Loushin, S. K.,Liu, S.-y.and Dykstra, C. E., *J. Chem. Phys.84*, *5*, **1986,** 2720-2725.

[27] Galano, A.and Alvarez-Idaboy, J. R., *J. Comput. Chem.27*, *11*, **2006,** 1203-1210.

[28] Walczak, K.,Friedrich, J.and Dolg, M., *Chem. Phys. 365*,**2009,** 38-43.

[29] Frisch, M. J.,Trucks, G. W.*et al.*; Gaussian Inc., Wallingford CT, 2009.

[30] Phillips, J. C.,Braun, R.*et al.*, *J. Comput. Chem.26*,**2005,** 1781-1802.

[31] Humphrey, W.,Dalke, A.and Schulten, K., *J. Mol. Graphics14*,**1996,** 33-38.

[32] Ryckaert, J.-P.,Ciccotti, G.and Berendsen, H. J. C., *J. Comput. Phys.23*,**1977,** 327-341.

[33] Halgren, T. A., *J. Comput. Chem.17*, *5&6*, **1995,** 490-519.

[34] Leach, A. R. *Molecular Modelling - principles and applications, 2nd. Ed.*; (Prentice Hall: Harlow, 1996).

[35] Joung, I. S.and Cheatham, T. E., *J. Phys. Chem. B112*,**2008,** 9020-9041.

[36] Bhandarkar, M.,Brunner, R.*et al.*; Theoretical Biophysics Group
University of Illinois and Beckman Institute: Urbana, IL 61801, 2008.

[37] Martyna, G. J.,Tobias, D. J.and Klein, M. L., *J. Chem. Phys.101*, *5*, **1994,** 4177-4189.

[38] Feller, S. E.,Zhang, Y.,Pastor, R. W.and Brooks, B. R., *J. Chem. Phys.103*, *11*, **1995,** 4613-4621.

[39] Darde, T.,York, D.and Pedersen, L., *J. Chem. Phys.98*, *12*, **1993,** 10089-10092.

[40] Kamerlin, S. C. L.,Rucker, R.and Boresch, S., *Biochem. Biophys. Res. Commun.356*,**2007,** 1011-1016.

[41] Kamerlin, S. C. L.,Rucker, R.and Boresch, S., *Biochem. Biophys. Res. Commun.345*,**2006,** 1161-1166.

[42] Lund, M.and Jungwirth, P., *Phys. Rev. Lett100*, *25*, **2008,** 258105.

[43] Jungwirth, P.and Tobias, D. J., *Chem. Rev.106*,**2006,** 1259-1281.

[44] Dang, L. X., *J. Phys. Chem. B106*,**2002,** 10388-10394.

[45] Young, D. C. *Computational chemistry: a practical guide for applying techniques to real-world problems*; (John Wiley and Sons, Inc, 2001).

[46] Vreven, T.and Morokuma, K. Hybrid mehtods: ONIOM (QM:MM) and QM/MM, in *Annual reports in computational chemistry, Volume 2*; Elsevier B.V., 2006.

[47] Zhao, Y.and Truhlar, D. G., *Theor. Chem. Acc120*,**2008,** 214-241.

[48] Lohse, D. L.,Denu, J. M.,Santoro, N.and Dixon, J. E., *Biochem. 36*,**1997,** 4568-4575.

[49] Hansson, T.,Nordlund, P.and Åqvist, J., *J. Mol. Biol.265*, *2*, **1997,** 118-127.

[50] Peters, G. H.,Frimurer, T. M.,Andersen, J. N.and Olsen, O. H., *Biophys. J.77*,**1999,** 505-515.

[51] Valiev, M.,Bylaska, E. J.*et al.*, *Comput. Phys. Commun.181*,**2010,** 1477-1489.

[52] *NWChem documentation*; http://www.nwchem-sw.org/index.php/NWChem_Documentation; September 14th.,2011

[53] Cornell, W. D.,Cieplak, P.*et al.*, *J. Am. Chem. Soc.117*,**1995,** 5179-5197.

[54] Hornak, V.,Abel, R.*et al.*, *Proteins Struct. Funct. Bioinf.65*, *712-725*, **2006,**

[55] Case, D. A.,Darden, T. A.*et al.*; University of California, 2010.

[56] Case, D. A.,Cheatham, T. E.*et al.*, *J. Comput. Chem.26*, *16*, **2005,** 1668-1668.

[57] 2011.

[58] Berendsen, H. J. C.,Postma, J. P. M.,Gunsteren, W. F. v.,DiNola, A.and Haak, J. R., *J. Chem. Phys.81*,**1984,** 3684-36870.

[59] Rühle, V., 2007.

[60] *Brief introduction to the thermostats*September 16th.,2011

[61] Schlitter, J.,Engels, M.and Krüger, P., *J. Mol. Graph.12*,**1994,** 84-89.

CHAPTER 5

CONCLUSIONS AND PERSPECTIVES-

5.1- CONCLUSIONS

PTP1B-halide ion complexes were obtained in different conditions: both employing different halides as ligands, and different buffer:halide ratios.

The crystals corresponding to the different complexes were obtained, and exposed to synchrotron radiation sources.

Employing different computational tools, the data obtained from the X-ray diffraction experiment were analyzed and molecular models were determined. In all cases two copies of the flexible WPD were modeled, one in the open (unbound) conformation and the other in the ligand-bound (closed) conformation.

The refinement of the different data sets obtained from the synchrotron radiation showed a distinct influence of the ion present in each complex on the conformation adopted by the WPD loop. While some of the crystals where the bromide was present showed a higher occupancy for both the closed conformation copy of the WPD loop and the halide, in others the open conformation was clearly preferred. In the latter, this low occupancy for the closed conformation was coincident with a low occupancy of the bromide itself, thus indicating a correlation between the loop's conformation and the presence or absence of the ion in the catalytic cleft. As for the iodide complexes, in all cases the empirical data indicated the WPD was present in a closed conformation, and a high occupancy for the ion. These data both confirms the correlation previously established between the presence of the ion and the conformation adopted by the loop at study, and indicates a relationship between the identity of the ion present in the complex and the strength of the interactions established between the protein and the halide.

A correlation between the concentration level of the halide and the loop's conformation was found. It was observed that the higher the concentration of the halide, the higher the occupancy of both the ion and the closed conformation for the WPD loop; this was the case in the complexes formed with both halides. In this regard, an interesting observation was that when both the buffer and the halide showed high concentration, the effect of the first prevails, and the occupancy of the ion appears to decrease.

CHARMM27 force field compatible parameters for both bromide and iodide were determined and validated. The resulting parameters proved to be appropriate to represent these halides and their interactions with proteic residues in biological systems.

Systems including different number of amino acidic residues were optimized both at a MM level, employing the AMBER force field, and at a QM level, using the density functional theory. By doing so, two different objectives were achieved: on the one side, the validation of the AMBER FF to reproduce the desired interactions, and the determination of the region to be studied at a high level of theory during the optimization of the geometries for the protein-halide complexes.

The geometry optimizations of the four complexes considered (protein-bromide and protein iodide complexes, in the two possible conformations for the WPD loop) are being calculated at this time. The size and complexity of the considered systems have not made it possible to obtain conclusions out of these calculations yet.

The collective motions of the PTP1B in complex with the two halides considered were studied by performing Molecular Dynamics simulations. The results were consistent with previous information regarding the flexibility of the WPD loop.

The movement of this flexible loop in both complexes was consistent with the empirical information obtained in this work: while the bromide complex presented a conformational change during the run, this was not observed for the complex with iodide. This supports the previously stated theory regarding the existence of differentiated interactions between both halides and the surrounding protein residues.

5.2- PERSPECTIVES

In order to fully understand the interactions between the halides and the surrounding residues that stabilize the closed conformation of the WPD loop, the optimized geometries of the complexes in both conformations of the loop need to be found and analyzed. Hence, it is necessary to finish the geometry optimizations that are being calculated at this time.

Since both MD simulations achieved stabilization towards the end of the run (during the last 2 ns), it is necessary to carry out a longer run, at least double the time. This would allow observing any interesting conformational changes that may be taking place, which are hinted in the run presented here (specially for the PTP1B-Br^- complex).

To determine the energetic barrier between the open and closed conformation of the WPD loop, a targeted Molecular Dynamics study should be carried out. This particular study would provide an insight into the pathway followed by the conformational change studied.

CHAPTER 6

EXPERIMENTAL DETAILS-

6.1- Expression and purification of the protein

Bacteria containing the cDNA for the PTP1B in pET23b plasmid (Novagen) were grown in *Escherichia coli*BL21(DE3). An overnight culture of the transformed BL21(DE3) cells was diluted 2:100 into 1 liter of LB medium containing 100 µg/mL ampicillin. The culture was grown at 37°C until the absorbance at 600 nm reached 0.6, at which point the cells were induced with 1mM IPTG (Euromedex) for 3 h. The cells were harvested by centrifugation at 4 °C for 20 min at 4000*g*.

The cell pellet was resuspended in 70 mL of 100mM 2-(4-morpholino)-ethane sulfonic acid (MES) pH 6.5 (Buffer A), sonicated 20min and clarified for 1h at 40,000 rpm. The supernatant was then loaded onto a CM-Sephadex column (30-mL bed volume) (Pharmacia) and washed with 10 bed volumes of buffer A. PTP1B was eluted from the column by a linear gradient from 0 to 0.5M NaBr, NaI or NaBr-CH_2-CH_2-SO_3.

Approximately 25mg of PTP1B (95% pure as estimated by SDS-polyacrylamide gel electrophoresis) were recovered from a 3 L culture.

6.2- Co-crystallization

Crystals were grown by hanging-drop vapor diffusion method at 4°C. PTP1B was concentrated at 10 mg/mL in 10mMTris-HCl, pH 7.5, 0.2 mM EDTA, 3.0mM DTT and between 25-200mMNaBr, NaI or NaBr-CH_2-CH_2-SO_3. For crystal growth, a 4µL drop of protein solution was mixed with an equal volume of precipitating solution 100 mMHepes pH 7.5, 14% polyethylene glycol 8000 and between 20-300mM magnesium acetate, and equilibrated against 0.5mL of the precipitating solution.

Typically, rodlike crystals appeared overnight and continued to grow to a maximum size of 0.4 mm x 0.2 mm x 0.7 mm within 10 days. The crystals were cryoprotected by

transferring them first into the crystallization solution plus 15% glycerol than in 30% glycerol. Cryoprotected crystals were flash-cooled in ethane prior to data collection.

6.3- X-ray diffraction data collection

The X-Ray diffraction data were collected at the synchrotron Swiss Light Source (SLS), at the Paul Scherrer Institute, Villigen, Suisse [1].

6.4- X-ray data processing

The processing of the X-ray diffraction data was carried out using the HKL2000 software [2].

6.5- Determination of an initial model and subsequent refinement

The initial phases for the PTP1B-halide complexes were determined employing the *Molecular Replacement* method, using AMoRe[3]. The subsequent optimization of the method was done employing a *Restrained Refinement*, without phase information and using *Maximum Likelihood* as the refinement target, all of which was done with Refmac5[4]. The consecutive Fourier and the difference Fourier electron-density mapswere obtained using thefft[5] software. All of the programs mentioned above are included in the CCP4 program suite [6]. The model visualization, building of the double conformation and some local real space refinement were done with the COOT [7] software. The refinement procedure was iterated until acceptable values for R and R_{free} were obtained [8-10].

Refinement of the occupancies was carried out employing the same methodology used before, but in these cases the occupancy values for both conformations of the loop as well as the corresponding ion were iteratively modified. This procedure was repeated until the B values for all the considered atoms were similar and in the acceptable range (given the relationship between both factors, an acceptable B value can be considered indicative of a correct representation of the presence of those atoms) [11].

REFERENCES

[1] *The Swiss Light Source SLS*; http://www.psi.ch/media/swiss-light-source-sls; August 15th.,2001

[2] Otwinowski, Z.and Minor, W., *Macromolecular Crystallography, part A276*,**1997,** 307-326.

[3] Navaza, J., *Acta Cryst. A50*,**1994,** 157-163.

[4] Murshudov, G.,Vagin, A.and Dodson, E., *Acta Cryst. D53*,**1997,** 240-255.

[5] Read, R. J.and Schierbeek, A. J., *J. Appl. Cryst.21*,**1988,** 490-495.

[6] Collaborative Computational Project, *Acta Cryst. D50*,**1994,** 760-763.

[7] Emsley, P.,Lohkamp, B.,Scott, W. G.and Cowtan, K., *Acta Cryst. D66*,**2010,** 486-501.

[8] *R-factor calculations and their significance*; http://bass.bio.uci.edu/~hudel/mbb254/lecture2/lecture2_1.html; August 20th.,2011

[9] Brünger, A. T.and Rice, L. M. Crystallographic refinement by simulated annealing: methods and applications, in *Methods in Enzymology*; Carter, C. W., Sweet, R. M., Eds.; Academic Press, 1997; Vol. 277; p 243-269.

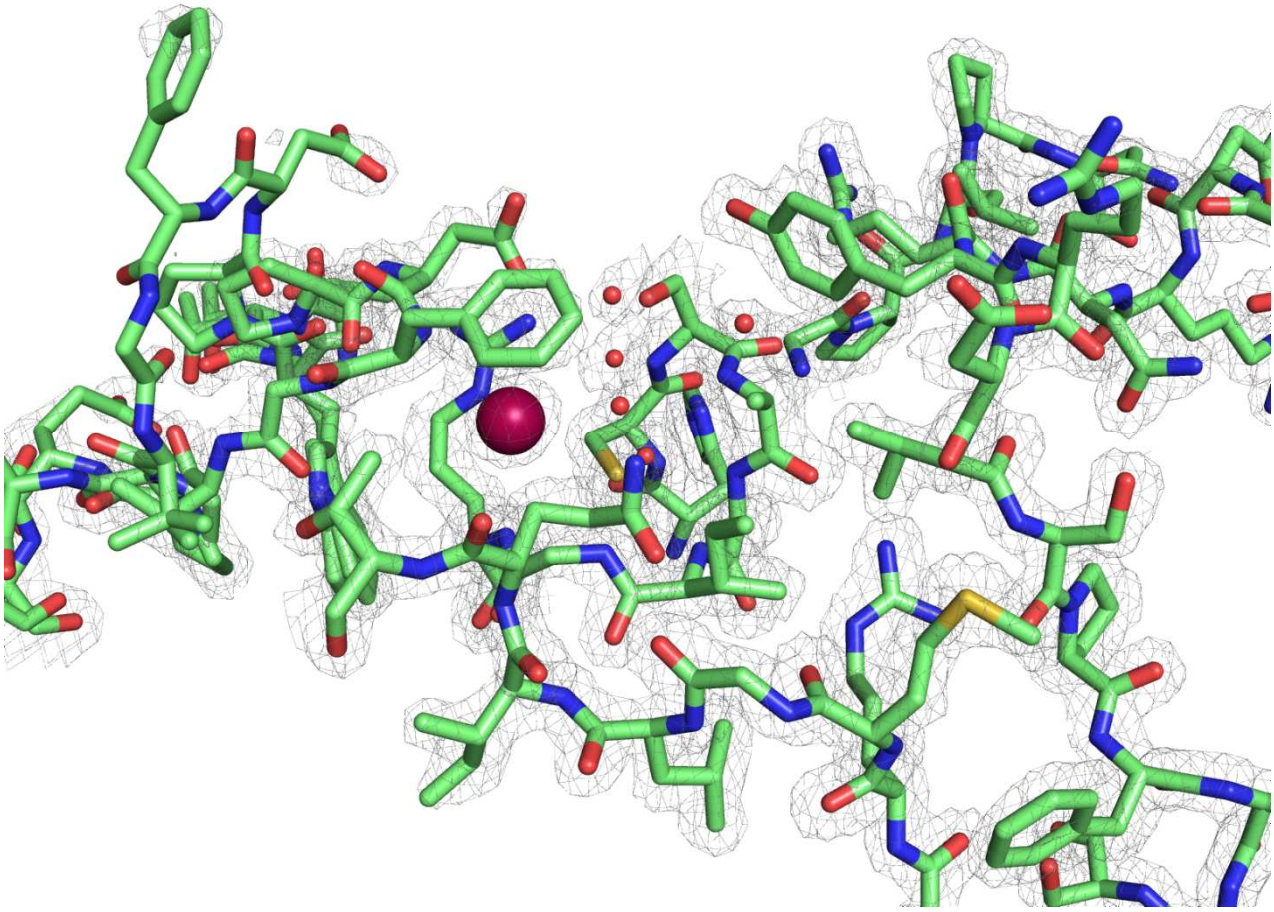
[10] Brünger, A. T. Free R value: cross-validation in crystallography, in *Methods in Enzymology*; Carter, C. W., Sweet, R. M., Eds.; Academic Press, 1997; Vol. 277; p 366-396.

[11] Altman, R. B.,Hughes, C.and Jardetzky, O. "Compositional characteristic of disordered regions in proteins," Stanford University, 1994.

APPENDIX

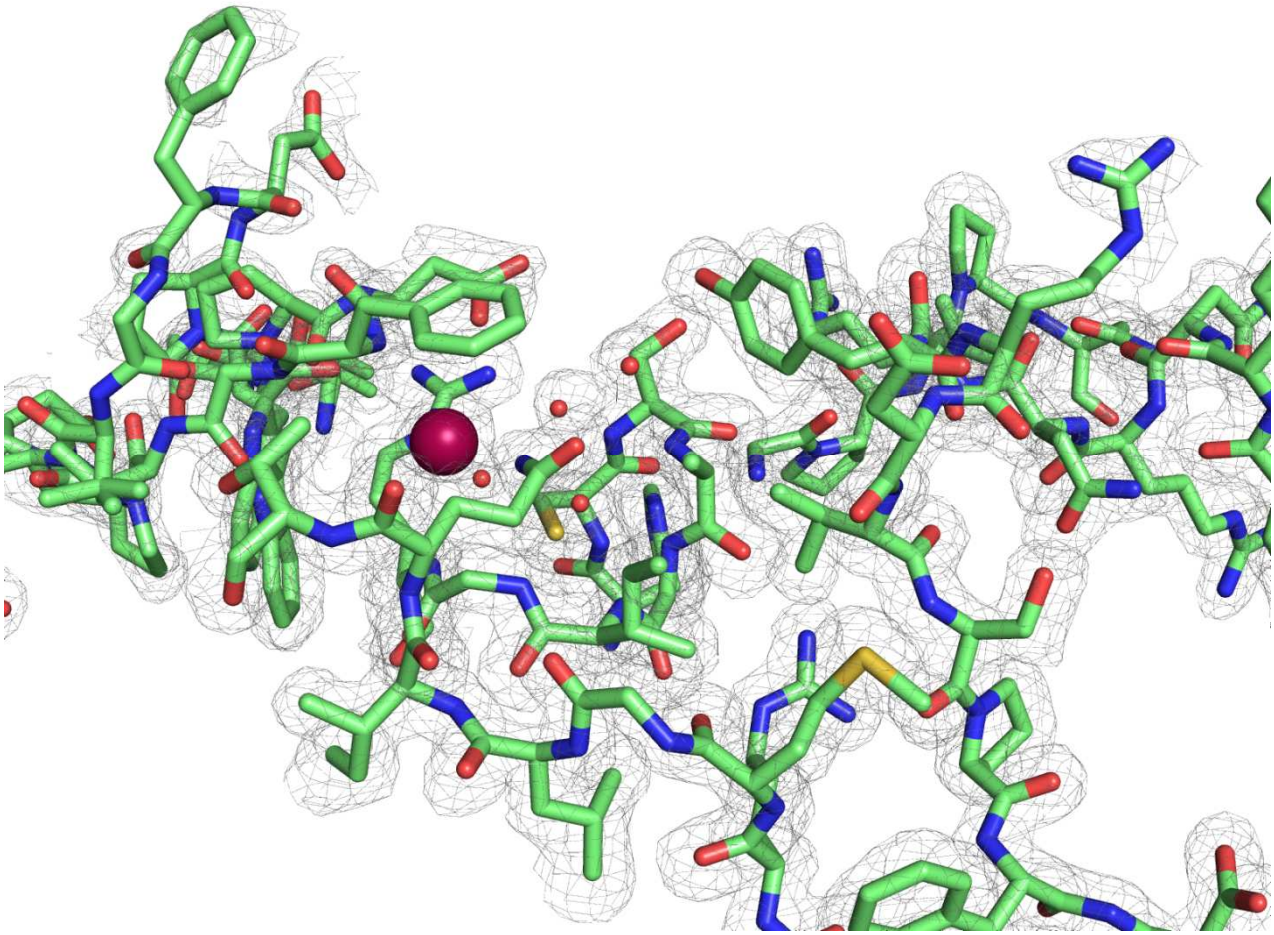
Refined models for the different halide-PTP1B complexes

Data set 1



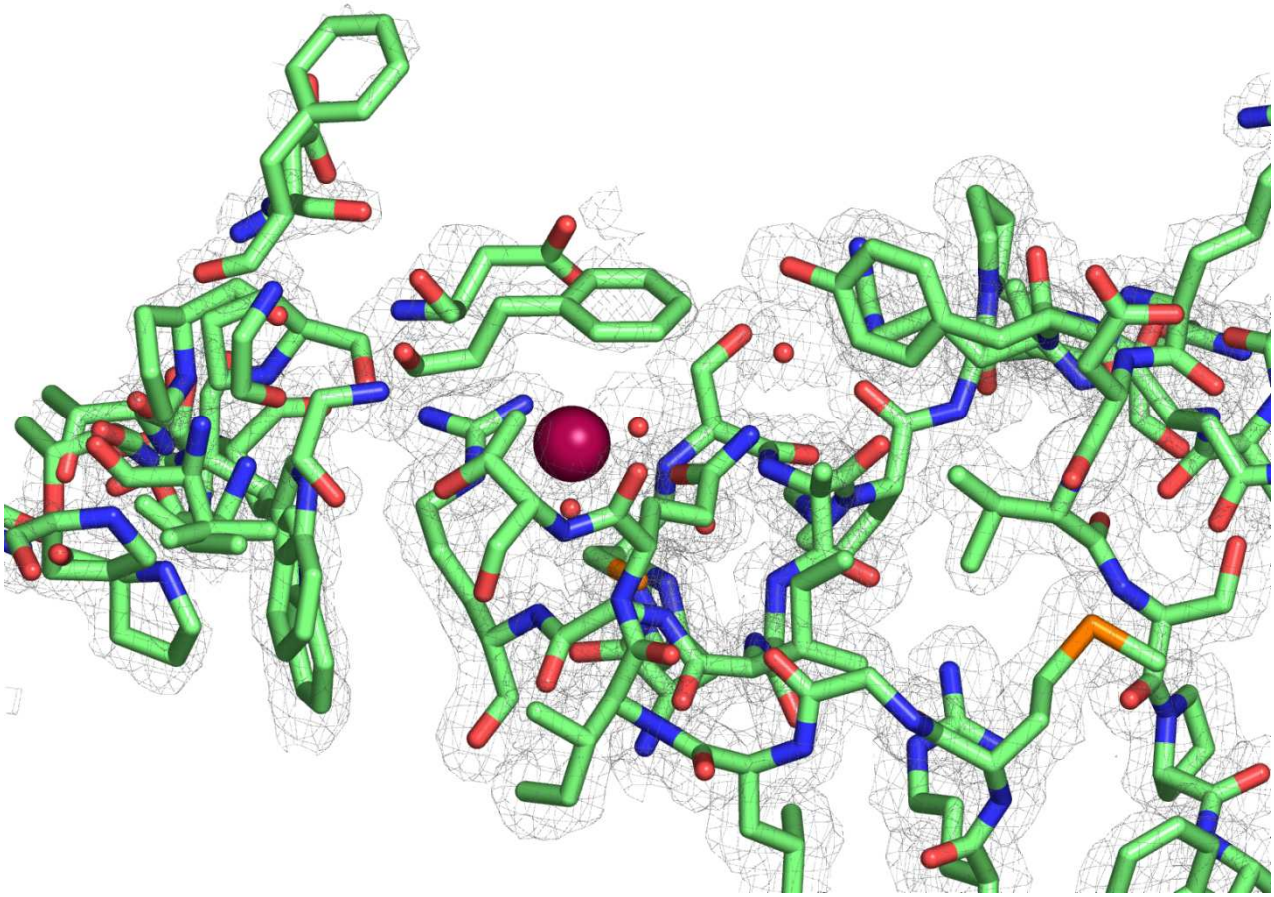
Contour value = 1σ

Data set 2



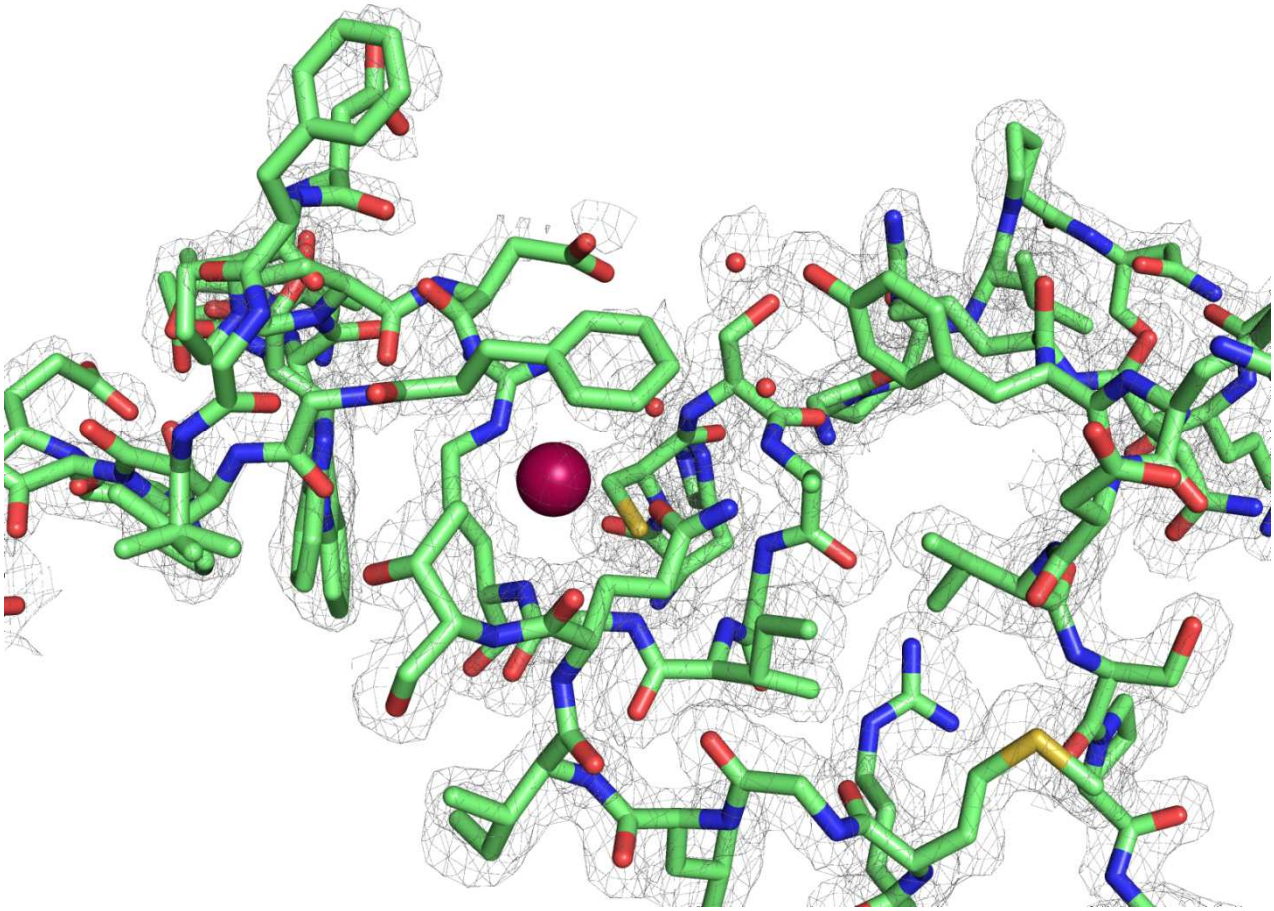
Contour value = 1σ

Data set 3



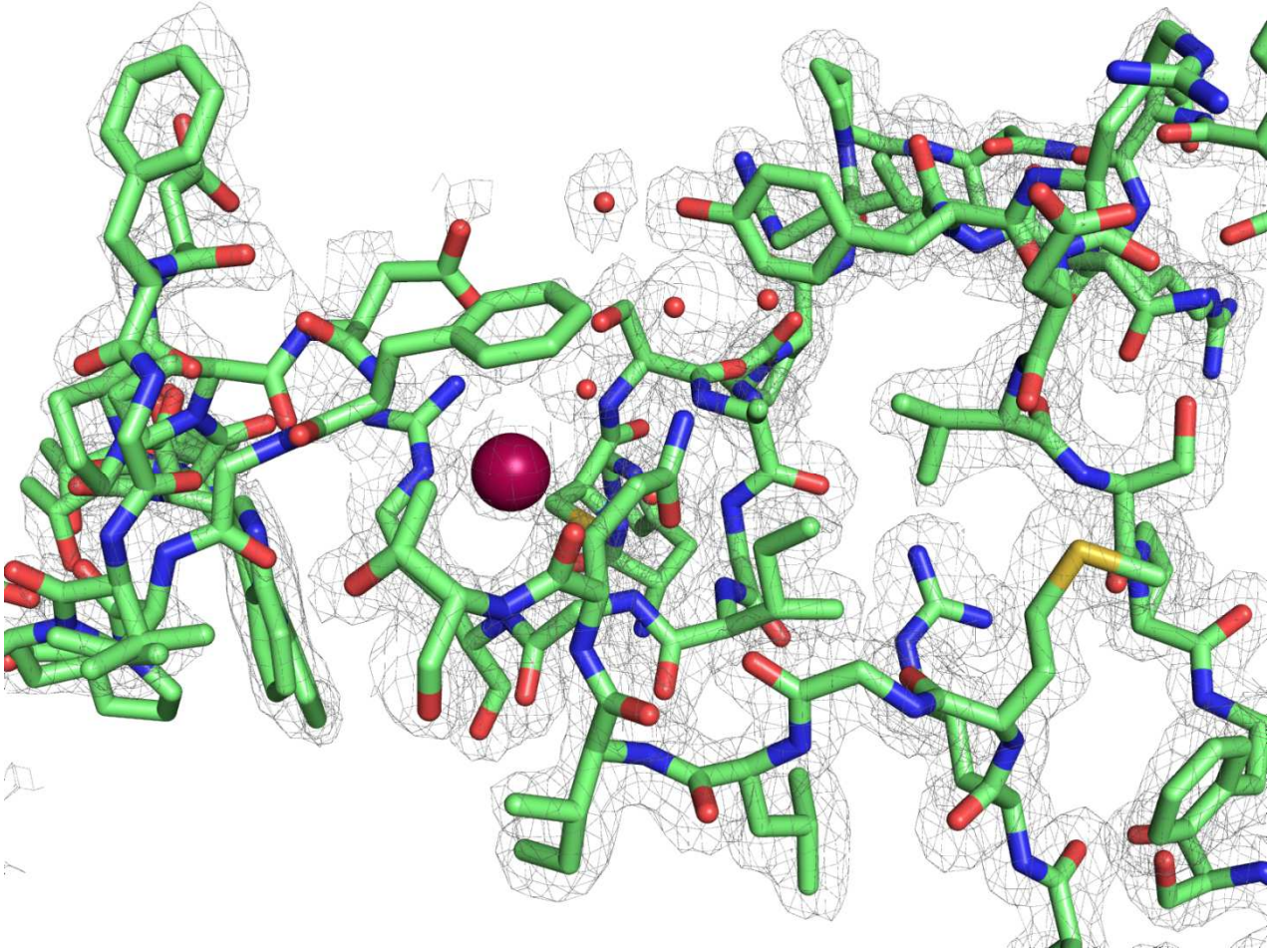
Contour value = 1σ

Data set 4



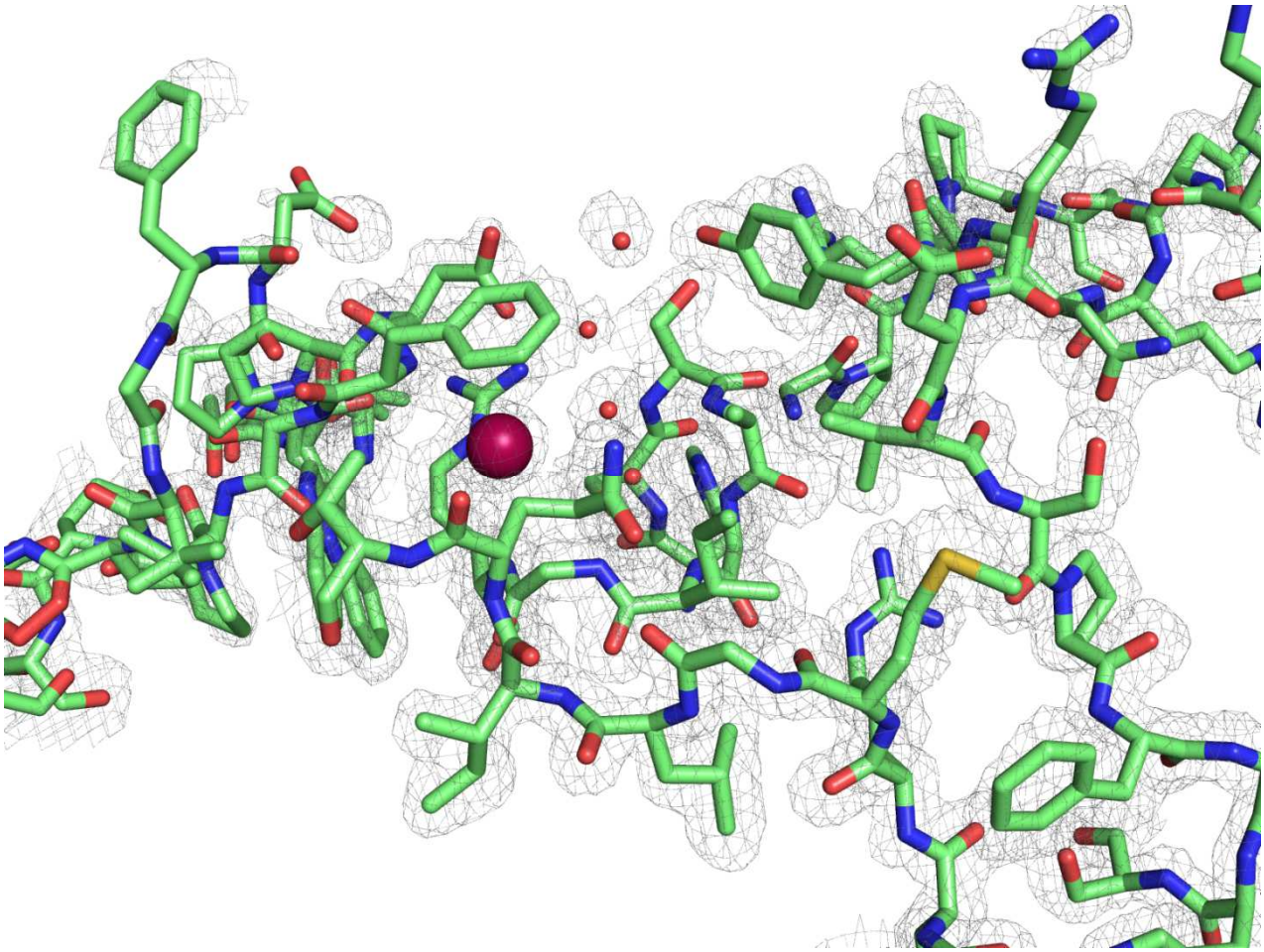
Contour level = 1σ

Data set 5



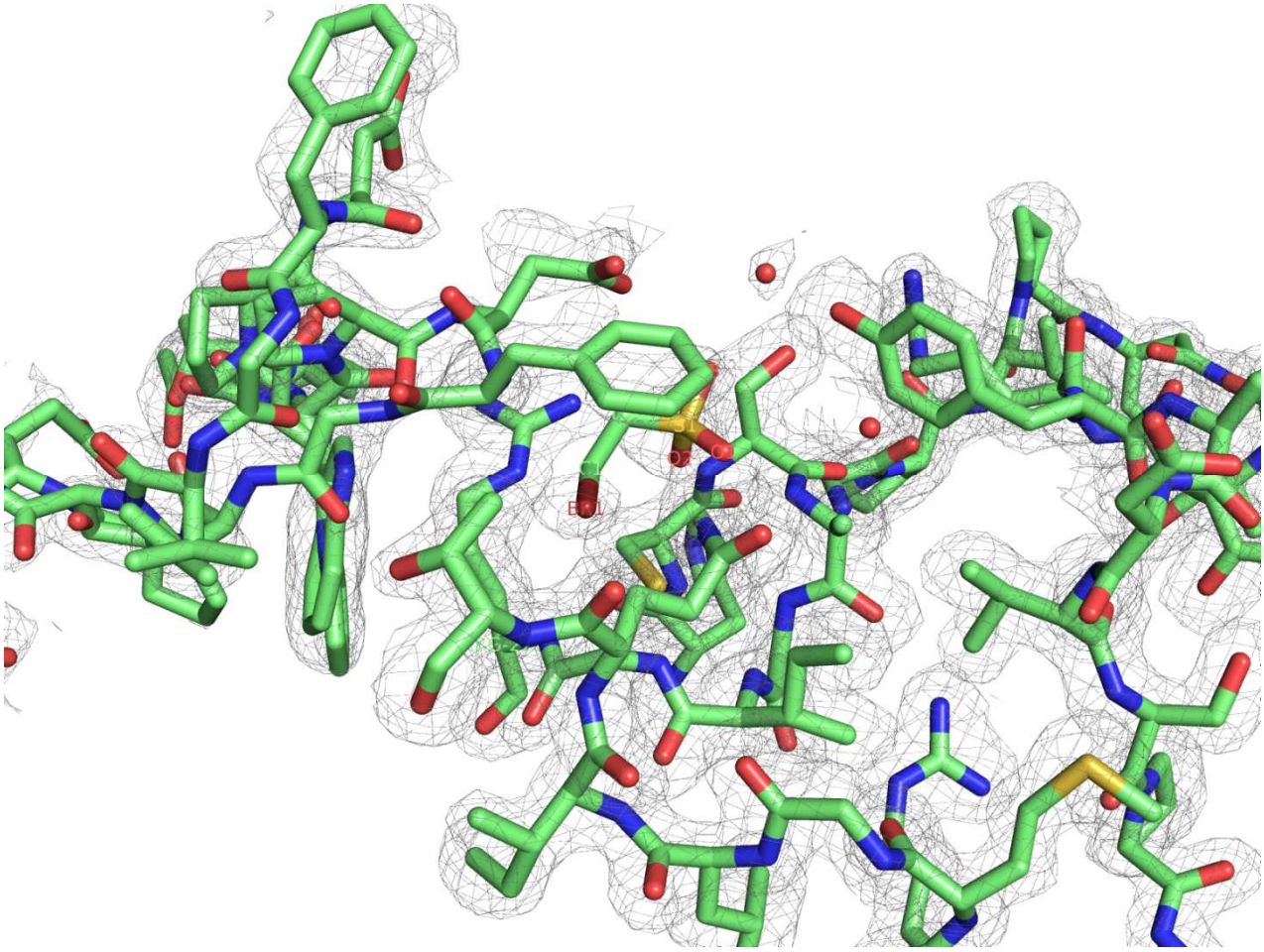
Contour level = 1σ

Data set6



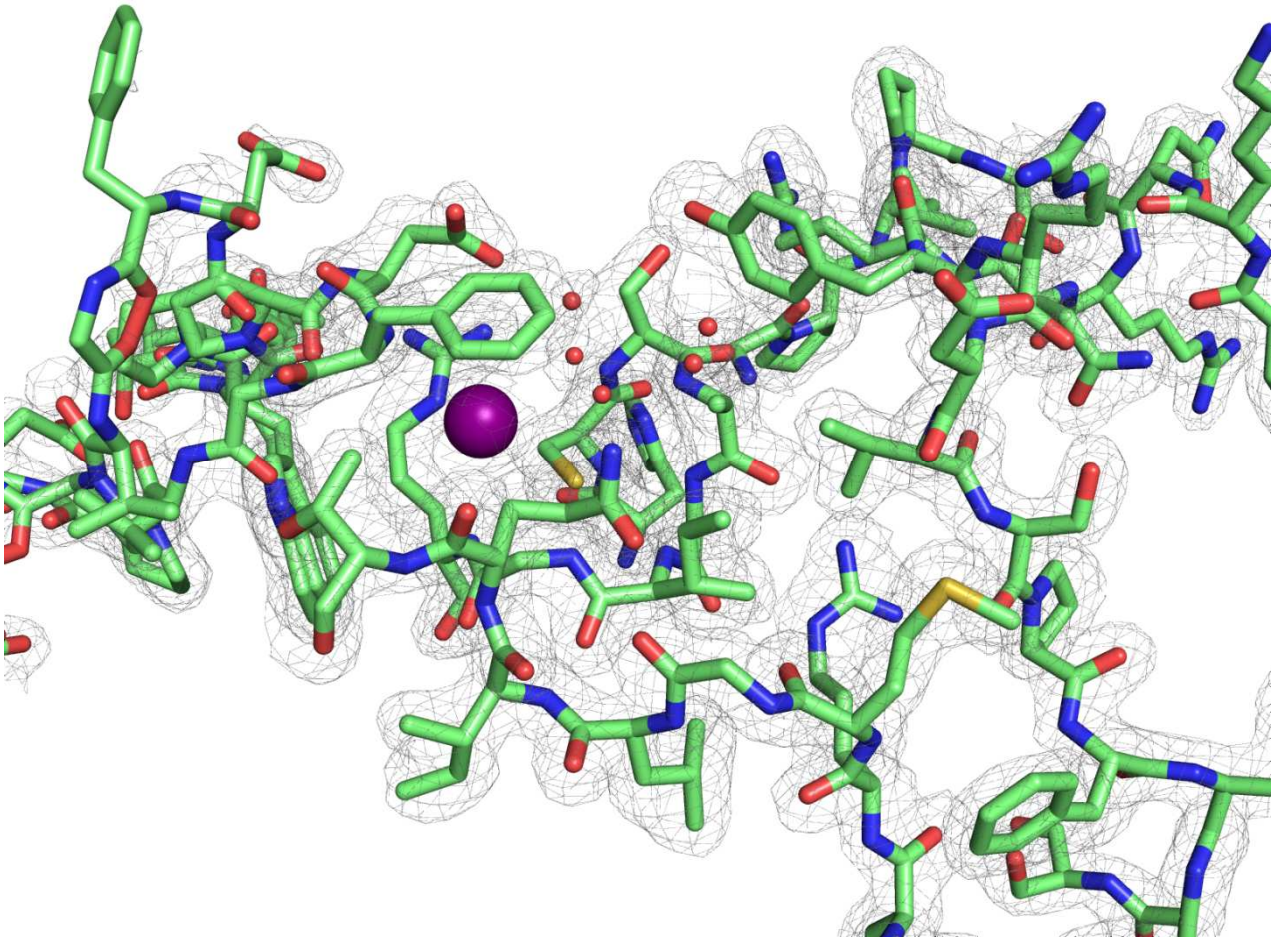
Contour value = 1σ

Data set7



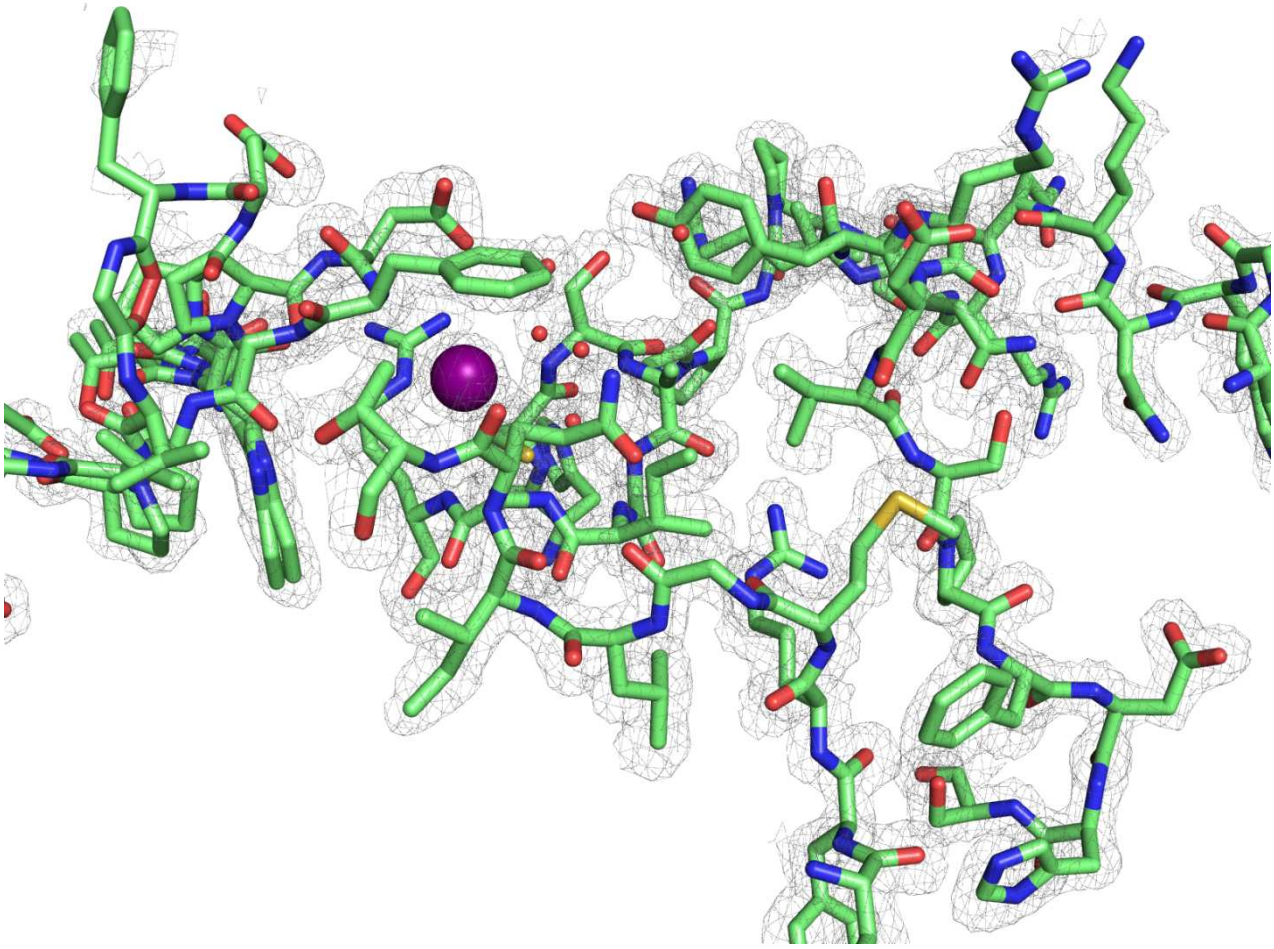
Contour level = 1σ

Data set8



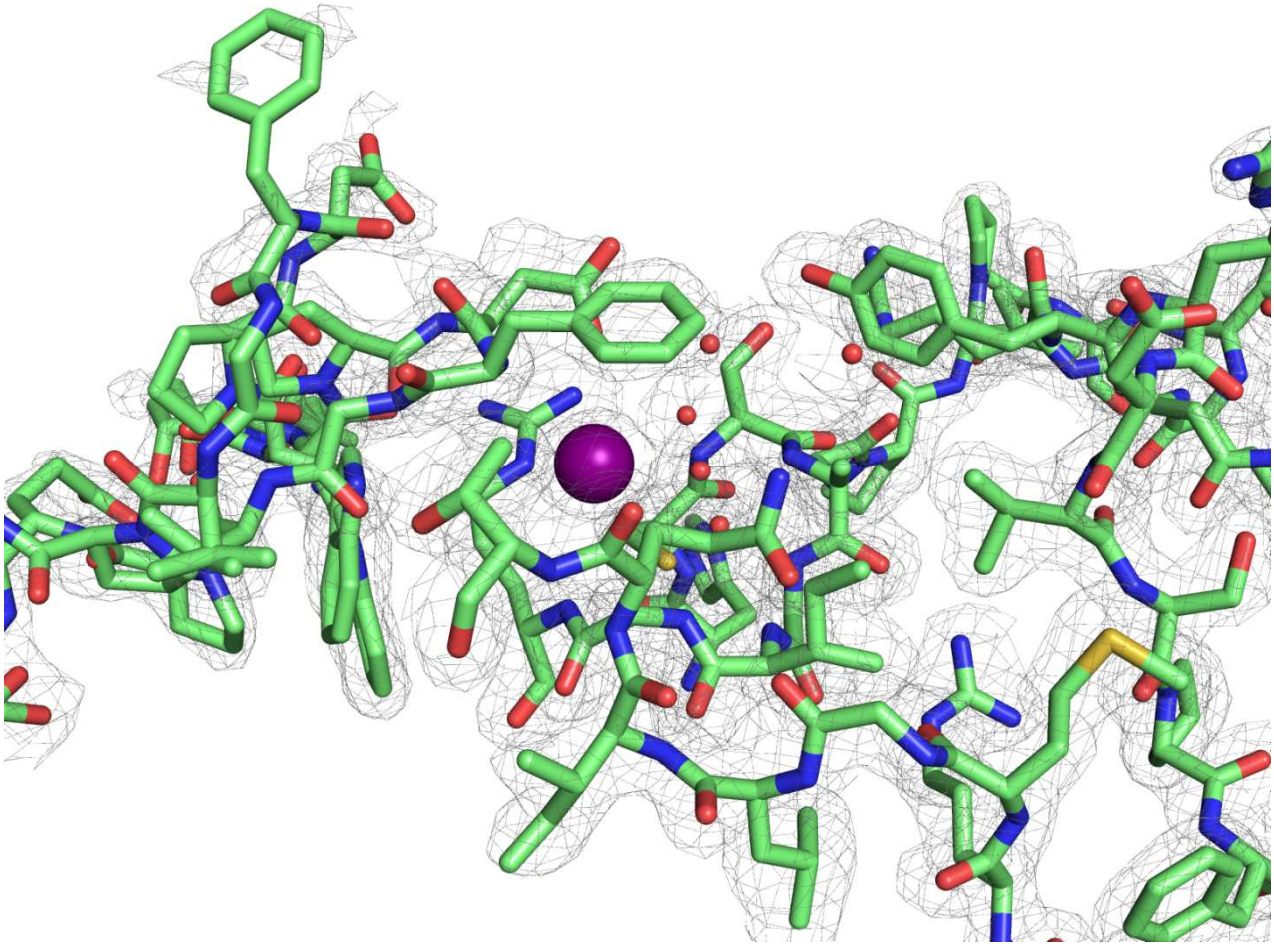
Contour level = 1σ

Data set9



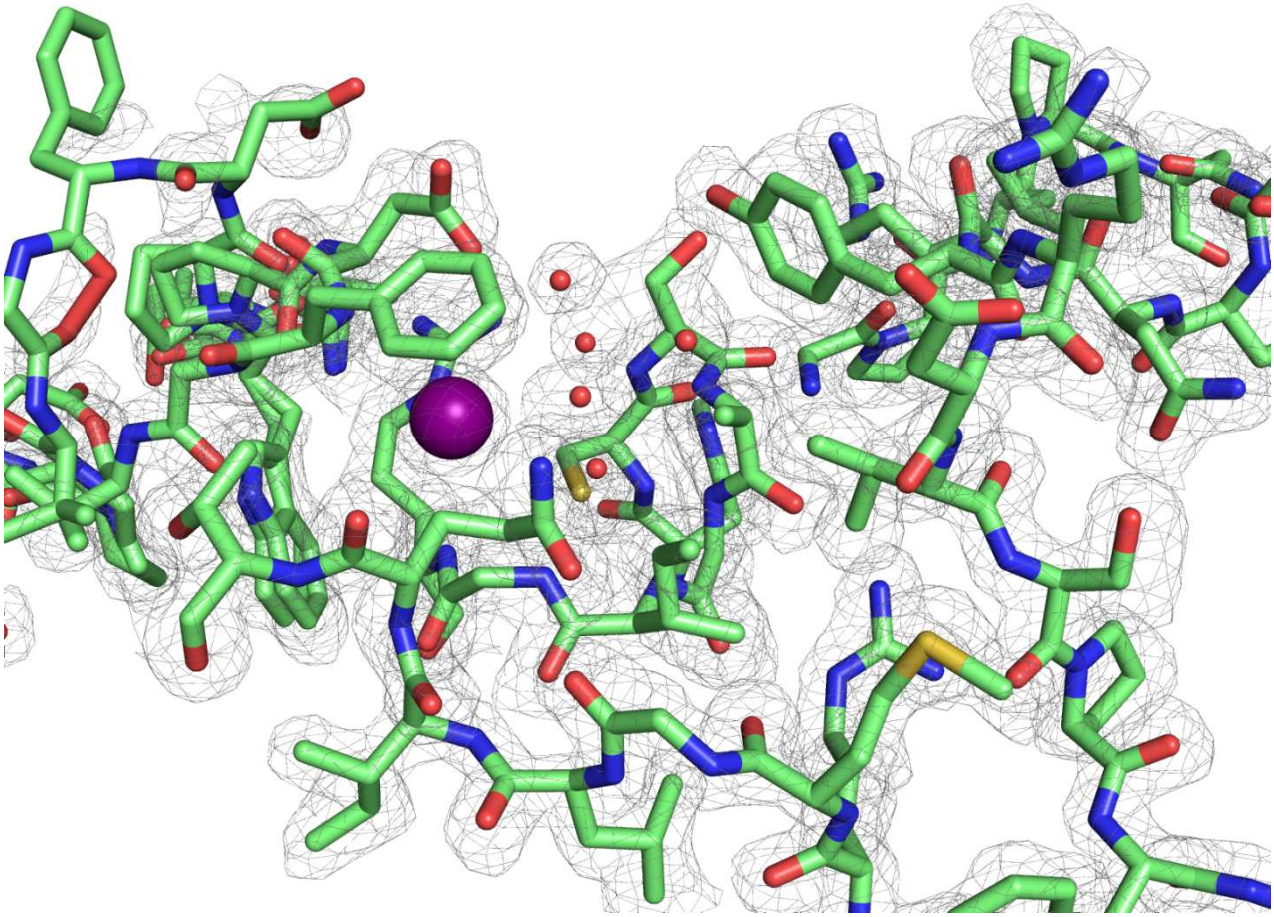
Contour level = 1σ

Data set 10



Contour level = 1σ

Data set 11



Contour level = 1σ