



Computational Enzymology, a ReaxFF approach

Corozzi, Alessandro; Fristrup, Peter

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Corozzi, A., & Fristrup, P. (2013). Computational Enzymology, a ReaxFF approach. Kgs. Lyngby: Technical University of Denmark (DTU).

DTU Library

Technical Information Center of Denmark

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Front Cover:

Yves Klein, La grande Anthropométrie bleue, ca.1960.
Dry pigment and synthetic resin on paper, mounted on canvas.
Guggenheim Museum Bilbao.

Alessandro Corozzi

Computational Enzymology, a ReaxFF Approach

Computational Enzymology, a ReaxFF approach.

Ph.D. essay



Alessandro Corozzi, June 2013

DTU Kemi
Danmarks Tekniske Universitet

Institut for Kemi, Bygning 201
Kemitorvet 201, Kongens Lyngby
2800 Denmark (DK)
Tel. +4545252419
<http://www.kemi.dtu.dk/>

DTU Chemistry
Department of Chemistry

Preface

This thesis describes a project that I carried out during my enrollment as a Ph.D. student at the Department of Chemistry, Technical University of Denmark, under the supervision of associate Professor Peter Fristrup. The external part of this study has been spent at ICCOM-CNR U.O.S. - Istituto di Chimica dei Composti Organometallici, Pisa, under the supervision of Dr Susanna Monti. The core of this Ph.D. dissertation is the development of Prot-ReaxFF, a new computational reactive force-field to model bio-catalysts in a fast and accurate way enabling us to have some insides that will lead to an increased understanding of the complex reactions that are carried out by enzymes. Prot-ReaxFF was born as a collaboration between Peter Fristrup's group from DTU Kemi, Dr Susanna Monti from DREAMSLAB (Dedicated Research Environment for Advanced Modeling and Simulations) and Dr. Adri van Duin from the Department of Mechanical & Nuclear Engineering of the Pennsylvania State University.

Before the acknowledgments and special thanks to the extraordinary people that I met along the road, I would like to spend some word about my experience as a foreigner here in Denmark. It was not easy to settle because of different customs and way of living, fortunately though I deeply think that children,

artists, and scientists have in common a distorted vision of life. They try to describe it their own way, as a track parallel to reality, way out from public trails. To lead this kind of life it is necessary an uncommon strength, because you need to find elements of balance in unknown environments. All this creates ineluctable strains. Strains sometimes felt as pains. I honestly cannot say that everything was perfect. Sometimes all the difficulties bound each other leading to a personality split, to what is called instability: a fragile equilibrium moving on life's difficulties. Nature did the most hard job, because it was able to give me the peace that I needed. The struggles to overcome problems, the solitude and loneliness were all cuddled by marvelous forests, lakes, and uncontaminated landscapes. I will never forget the violent sky of Denmark, with its rigid winters, the astonishing summers, the heavy rains, the strong winds, and the peaceful snowing when it was the time. Nature helped me in all possible ways, because everything appears to me in the right place. The impression of a sky that is bigger, so big that you can feel the roundness of it, makes me feel in contact with nature. All my problems magically dissolved by mere walking and wandering around. I felt little compare to the vastness of what is out there, and as a consequence little appeared my problems too. I think it would have not been the same in another country, so I am grateful to have had the possibility of living in this gorgeous land. A land that I now consider my second home.

The people did the rest. The feeling of being in a community is strong, it is outstanding their willingness to help you only for the sake of it. Among all the people that I met, I would, first and foremost like to thank Professor Peter Fristrup for the independence that he granted me, and for his constant trust on my own possibilities. A special thanks also to my external advisor Dr Susanna

Monti for the valuable discussions and the sharing of common interests that were sometime far beyond the computational research.

During these long years I began to think about chemical modeling in terms of literary concept. Modeling is cheating, because you create your own world with particular boundary conditions and equations, and only at this point you start the calculations and the process of collecting end results. This is not different from the action of a writer with his romance. Writers create characters, even the rules for them, and a fake environment where they can move. Only after they start the process of writing the romance, so to speak "the simulation". When you read a romance, or when you analyze computational chemistry data, you do almost the same action. You are not strictly reading the reality, but at the same time you can get important things that can be useful keys to get an understanding of the real world.

I dedicate these years of study to my parents. They let me become the man that I am, because of this I love them, though I will never be able to pay them back for what they gave me.

Finally I extend my deepest gratitude towards my dear Carlotta, for her endless love and support. Without her, the end results would never have been possible.

Abstract

This PhD project eassay is about the development of a new method to improve our understanding of enzyme catalysis with atomistic details. Currently the theory able to describe chemical systems and their reactivity is quantum mechanics (QM): electronic structure methods that use approximations of QM theory to describe molecular structure. Modeling enzyme reactions is anyway still inaccessible to these methods because the size of the problem would result in many-particle equations too complicated to be solved even with rather crude approximations such as Hartree-Fock (HF). At the same time there are ordinary classical models - the molecular mechanics (MM) force-fields - that use newtonian mechanics to describe molecular systems. At this level it is possible to include the entire enzyme system still having light equations but renouncing to an easy modeling of chemical transformation during the simulation time. In short: on one hand we have accurate QM methods able to describe reactivity but limited in the size of the system to describe, while on the other hand we have molecular mechanics and ordinary force-fields that are virtually unlimited in size but unable to straightforwardly describe chemical reactivity. A reactive force-field (ReaxFF) is a simplified model that aims to bridge the gap between quantum chemistry methods to the ordinary force-fields of the classical molecular mechanics methods, enabling MM to model chemical reactions as a QM method with bond forming and breaking events during the simulation time. This has been accomplished by simply introducing anharmonicities in the potential energy terms of the force-field. Starting from a published ReaxFF force-field developed for modeling glycine aminoacid a novel ReaxFF force-field, ProtReaxFF, has been developed, optimized and applied to enzyme catalysis reactions.

Abstract

Dette ph.d. projekt omhandler udviklingen af en ny beregningsmetode der har potentiale for at udvide vores forståelser af enzymatisk katalyse med atomar detaljegråd. Nuværende teoretiske metoder der beskriver kemiske reaktioner er primært baseret på kvantemekanik (QM): elektronstrukturberegninger der benytter approksimationer til QM teori til at beskrive molekylær struktur. At modelere enzymatiske reaktioner er udenfor rækkevidde ved brug af disse metoder da størrelsen af det betragtede system resulterer i mange-partikel ligninger der er for svære at løse selv for simple metoder såsom Hartree-Fock (HF). Et andet alternativ er klassiske modeller normalt omtalt som molekyl-mekanik (MM) der bruger Newtons ligninger til at beskrive molekylære systemer. Med disse metoder er det muligt at beskrive et helt enzym med relativt få beregningsmæssige ressourcer men til gengæld er der ingen mulighed for at beskrive kemiske reaktioner. Kort sagt, på den ene side haves QM metoder der kan beskrive kemiske reaktioner men er begrænset i størrelsen af det kemiske system er kan behandles, mens der på den anden side have MM metoder der kan beskrive meget store systemer uden kemiske reaktioner. Det reaktive kraftfelt (ReaxFF) er en simplificeret model der forsøger at danne bro mellem kvantekemiske metoder og de klassiske molekyl-mekaniske metoder med det formål at muliggøre studier af kemiske reaktioner der finder sted undervejs i en simulering. Med udgangspunkt i et publiceret ReaxFF kraftfelt udviklet til beskrivelse af aminosyren glysin er der udviklet, optimeret og anvendt et nyt ReaxFF kraftfelt, ProtReaxFF, til beskrivelse af enzymkatalyse.

Contents

Contents	9
1 Introduction	11
2 Enzyme Catalysis	21
2.1 Biotic Catalysis	21
3 Computational Methods	53
3.1 <i>ab initio</i> Methods	53
3.2 Localization NBO Methods and NBO Steric Analysis	67
3.3 QM-MM Methods	72
4 Development of ReaxFF for Enzyme Catalysis	81
4.1 ReaxFF Potential Energy	85
4.2 ReaxFF Parameterization Techniques	94
4.3 Development of ReaxFF for Glycine	106
4.4 Development of ReaxFF for Protein and Enzymes	114
4.5 Serine Protease catalysis, A ReaxFF study	140
5 Resistance of 5-Thio-Galactoside to α-Galactosidases A	173

5.1	Introduction	173
5.2	Experimental Study	176
5.3	Modeling Study	183
6	Conclusions	199
	Bibliography	203
7	Appendix A - PCA Software	229
7.1	PCA_Matrix_Generation.m	230
7.2	my_write_reaxff.m	241
7.3	my_read_reaxff2.m	244
7.4	my_output_scanner.m	249
7.5	my_output_extractor.m	251

Chapter 1

Introduction

In a world where the depletion of energy resources is becoming an issue, every strategy that pushes towards a more sustainable development is getting more and more attention. There are many aspects of our life that will be affected by future desirable politics where ecology will have a major role and where eco-sustainability will be at the core of every strategy planing. The so called coming age of sustainability [1] will change our customs and our way of thinking in a way we cannot even forecast. Sustainability is expected to become a plus-value for every industrial way of action, and relative issues will be rightly considered from the Research & Development step to the actual production process and relative distribution.

Catalysis is the field of chemistry that traditionally answers the need of having same chemical processes consuming much less energy. A catalyst is involved in the reaction mechanism so to increase chemical reaction rate and therefore the efficiency of a process. Nowadays catalysis for industrial processes is provided by sustainable critical elements. The last decades saw the predominance

of organometallic catalysis carried out mainly by the platinum group metals (Ru, Rh, Pd; Os, Ir, Pt). These are metals so geochemically rare that they are wisely reserved for application of transcendent importance. As a group and in the long term it would be unwise to assume a market's ability to keep pace with consumer's demand [2]. The amount of a resource in Earth's crust is not simply inversely proportional to its concentration. Generally a resource has a bimodal distribution. There is one large peak for the element as a low concentration substituent and a smaller peak for the same element in deposits [3]. In the case of rare elements we are already at the stage of mining from deposits. Since the energy required for extracting something from an ore is inversely proportional to the ore grade (i.e. it takes ten times more energy to process an ore which contains the useful mineral in a ten times lower concentration), therefore the extraction of these elements would require an amount of energy that will be well beyond our capabilities.

Enzyme catalysis can be considered as one of the elective bio-eco-sustainable technologies. An enzyme is a biological nano-machine that is used as catalyst in almost all chemical reactions within biological cells. Enzymes are used since millenia to accelerate and control useful chemical reactions like fermentation processes [4], and it has been recognized more than a century ago the possibility of applied them to useful chemical transformations [5]. By definition we can refer to biotic catalysis as the application of enzymes and microbes to synthetic chemistry, where the use of such catalysts is directed towards substrate and reactions to which enzymes have not evolved for. Many beneficial aspects can be obtained by converting manufacturing unit process to bio-catalysis. Enzymes are infact obtained by renewable sources, they are biodegradable, they

Product	Technology	Company	Year
Simvastatin, a leading drug for treating high cholesterol	Enzyme	Codexis	2012
Buckman's Maximize enzyme to increase paper strength and quality	Enzyme	Buckman International	2012
Succinic acid as chemical feedstock	Fermentation	BioAmber	2011
1,4-butanediol for polymers and chemical feedstock	Fermentation	Genomatica	2011
Higher alcohols as fuels and chemical feedstock	Fermentation	UCLA (Prof. Dr. J. Liao)	2010
Renewable petroleum from fatty-acid metabolism intermediates	Fermentation	LS9	2010
Sitagliptin: a pharmaceutical ingredient for treatment of type 2 diabetes	Enzyme	Merck and Codexis	2010
Esters for cosmetics and personal care products	Enzyme	Eastman Chemical Co.	2009
Atorvastatin intermediate for treatment of high cholesterol	Enzyme	Codexis	2006
Polyhydroxyalkanoates as biodegradable plastics and chemical feedstock	Fermentation	Metabolix	2005
Low trans fats and oils for human nutrition	Enzyme	ADM and Novozymes	2005
Rhamnolipids: biobased, biodegradable industrial surfactants	Fermentation	Jeneil Biosurfactant Company	2004
Taxol for treatment of breast cancer	Fermentation	Bristol Myers Squibb	2004
Improved Paper recycling using enzymes to remove sticky contaminants	Enzyme	Buckman Laboratories International	2004
Polyester synthesis using lipases	Enzyme	Polytechnic University (Prof. Dr R. Gross)	2003
1,3-propanediol for polymers	Fermentation	Dupont	2003
Lactic Acid for poly(lactic acid) polymers	Fermentation	NatureWorks	2002
Removal of natural waxes and oils from cotton before it is made into fabric	Enzyme	Novozymes	2001
Source: United States Environmental Protection Agency EPA (http://www.epa.gov/greenchemistry/pubs/pgcc/presgcc.html)			

Table 1.1: Presidential Green Chemistry Challenge Awards in biocatalysis over the past twelve years.

are non-toxic materials, their selectivity is of great advantage in batch process in chemical industry because it will require much less work to refine the product obtained at the end of the unit process. As an example of the great importance of enzyme catalysis as an eco-friendly technology we report a table of the United States Environmental Protection Agency (see table 1.1). The Presidential Green Chemistry Challenge Awards promote the environmental and economic benefits of novel green chemistry. These prestigious annual awards recognize chemical technologies that incorporate green chemistry into chemical design, manufacture, and use, and it is clear that each year many awards are won by processes heavily using enzyme catalysis.

Recently we are experiencing what is called the third wave of enzyme catalysis as a preliminary result of chemical biology innovations and the modifica-

tion of bio-catalysts via an in vitro version of Darwinian evolution [6]. DNA sequencing and gene synthesis are at the base of tremendous progress in tailoring bio-catalysts by protein engineering [7][8] and protein modeling [9]. The results obtained by the present day enzyme technology enables us to enrich natural enzymes with new and specific characteristic that are tailored at will. Enzymes can be modified to accept previously inert substrates: e.g. the case of Sitagliptin manufacture (see table 1.1 entry# 7), where applying modeling and mutation approach to a transaminase that lacked any activity toward the pro-sitagliptin ketone led to a transaminase with activity for the synthesis of chiral amine. Chiral amines are important key intermediates in the industrial process of sitagliptin, an anti-diabetic drug [10][11]. Of great importance is also the achievement of changing the nature of the product formed as in the case of amino-acid metabolism to obtain alcohols for bio-fuels [12].

In present day technology-market there are requirements of chemical products able to address processing of biomass to bio-fuels [13], materials manufacturing [14], and bio-catalysts for synthetic chemistry [15].

Chemical manufacturing plants need to optimize energy consumption of processes to reduce prices. To achieve the goal there is a need of catalysts that are more stable, more selective and that can operate over a wide range of operative conditions. The field of protein engineering is still young. This means the knowledge is on the form of an ensemble of case studies and therefore there is not a predetermined quantitative approach such that of other engineering fields (e.g. civil, electric, software or chemical engineering). Protein engineering is not a mature field able to address the struggles of the market unless an engineering approach will be developed. There should be intense and sys-

tematic studies to find principles connecting variation of free energies (ΔG) with structural modification. This will enable a systematic project and design of enzyme catalyst to reach a specific required function. According to Bornscheuer et al. [6], it is needed a quantitative measure and a systematic database of free energy changes either for processing design goals: i) higher activity at process condition, ii) increased process stability, iii) increased thermostability to run at higher temperatures, iv) stability to organic solvents, v) absence of substrate and/or product inhibition, vi) increased thermostability for storage and shipping, vii) increased selectivity (enantio- regio- and chemo-), viii) accept new substrates, ix) catalyse new reactions. The same systematic studies should focus as well on topics of more interests for chemical research and for what concerns protein design goals: i) destabilize unfolded enzyme, ii) stabilize folded enzyme, iii) increase substrate binding, iv) hinder unwanted substrate, v) reshape substrate-binding site, vi) add key mechanistic steps, vii) create steric hindrance, viii) add hydrogen bonds and ion-pairs.

Free energy (ΔG) studies for each of the points itemized before should be collected to be able to solve future challenges, to reach design goals leveraging on the required structural changes, and to obtain more focused libraries to test a novel enzyme system.

Computational modeling is a key part of such a development. The *in silico* design and testing of novel enzymes can help in narrowing libraries to be tested experimentally. Computer design of new enzyme activities is not yet accurate [6]. The output of a modeling study still requires an experimental testing of 10-20 case systems, and it usually results in enzyme with low activity that requires substantial further engineering manipulation. Again we take as example the

case of Sitagliptin seen before (see table 1.1 entry# 7); in silico prediction yielded an enzyme able to convert only 0.1 molecule of substrate per day even if the substrate appeared to be well suited for the active site [10] of a model derived from an X-Ray structure (figure 1.0.1).

The aim of this Ph.D. project is to develop a new computational approach to model bio-catalyst in a fast and accurate way enabling us to have some insides that will lead to an increased understanding of the complex reactions that are carried out by enzymes. Quantum Mechanical techniques (QM) cannot be applied so far to biological system because the too many variables involved are beyond our present day computing power. One of the major problem dealing with proteins is calculating the electronic energy for a given nuclear configuration to give a potential energy surface. To deal with large systems semi-empirical methods relying on charge density instead of traditional wave function approaches have been implemented, developed and successfully applied. DFT functionals (see chapter 3 section 3.1.3), especially its B3LYP exchange-correlation functional version [16][17], are accurate enough to perform tasks in many applications. It is worth to notice that even if DFT is a mono-determinantal method, and even if it is able to gain a good portion of electron correlation energy scaling as much as an Hartree Fock Method (see chapter 3 section 3.1.2), it is clear that the computing power should be several orders of magnitude faster than it is now to have it applied to bio-molecules. Present date DFT calculations are limited to single molecules or molecular complexes with sizes of the order of two or three hundreds atoms. In force field methods the troublesome and heavy calculating step is by-passed by writing the electronic energy as function of the nuclear coordinates and fitting parameters to experimental and/or

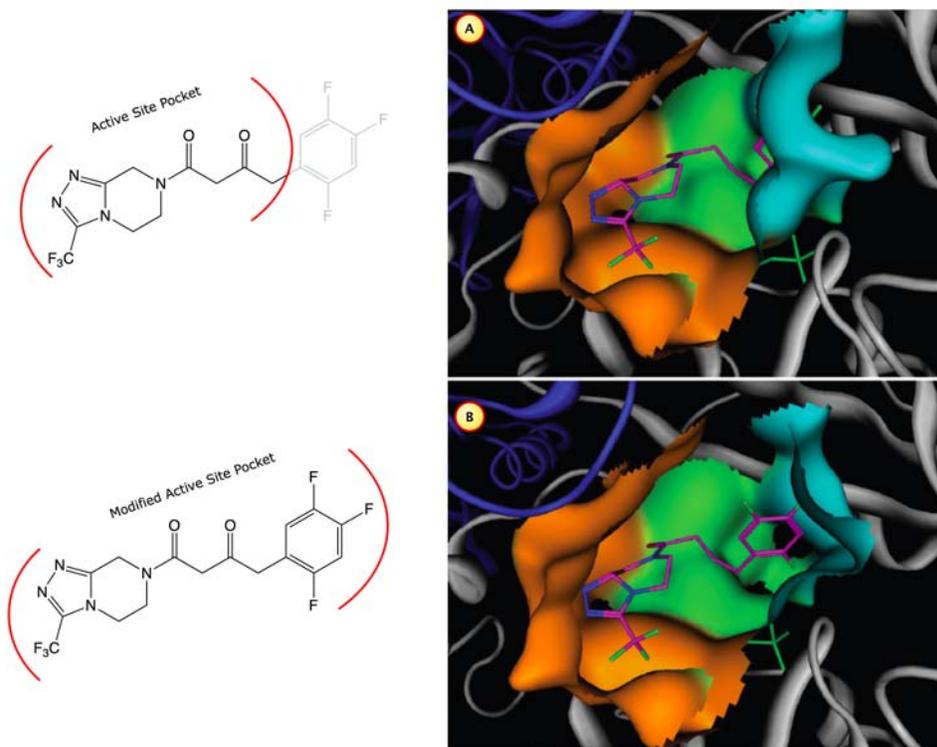


Figure 1.0.1: A pro-sitagliptin ketone analog (Lewis structure on the left) The substrate can be mapped on the binding pockets and docked into the active site of the transaminase enzyme (A). After initial engineering an enzyme variant was generated with activity on the desired substrate by excavating the pocket. “Reprinted with permission from Christopher K. Savile, Jacob M. Janey, Emily C. Mundorff, Jeffrey C. Moore, Sarena Tam, William R. Jarvis, Jeffrey C. Colbeck, Anke Krebber, Fred J. Fleitz, Jos Brands, Paul N. Devine, Gjalte W. Huisman, and Gregory J. Hughes, *Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture* (2010) *Science*, Vol. 329, No. 5989”

computational data. The development of new theoretical methods capable of describing chemical reactions at a fundamental level is a research area which receives immense attention. Development of good, reliable yet fast methods for *in silico* prediction of reactivity could potentially help the protein-engineering development. The molecular mechanics (MM) approach is an extremely fast computational method. Molecular dynamics (MD) calculations can be carried out for systems of more than 100 million atoms. It uses the brute force of modern computer, but its basic algorithms are intrinsically suited to be parallelized, so that it has the required speed for large-scale simulations of bulk materials; however, the inherent fixed description of the bonding pattern does not allow reactions to take place. To overcome this limitation, the Goddard group at the Materials Process and Simulation Center at the California Institute of Technology (MSC, Caltech) has developed the so-called “reactive force-field”, ReaxFF [18], which is a force-field capable of describing chemical reactions. The unique feature of this force-field is that the bond order is obtained from the interatomic distance between two pairs of atoms, i.e. the closer the atoms are to each other, the more strongly bonded they are, a concept which dates back to ideas of Linus Pauling. As direct consequence of this approach the ReaxFF method allows bonds to be formed and broken, thus effectively enabling reactions to occur during simulations. The ReaxFF force-field developed is called Prot-ReaxFF. The field of biotic catalysis is briefly introduced in chapter 2. Computational methods used for modeling studies and for the development of ReaxFF are instead provided in chapter 3. All the insight of the machinery of the algorithm of ReaxFF the parameterization procedure adopted, validation of Prot-ReaxFF and case ReaxFF study of trypsin inhibition is provided in chap-

ter 4. To conclude, in chapter 5 we report a study of the application of QM-MM method to model carbohydrates steric interaction as a corollary study on the Resistance of 5-Thio-Galactoside to α -Galactosidases A action.

Chapter 2

Enzyme Catalysis

2.1 Biotic Catalysis

The biochemical basis of enzymology has been discovered at the beginning of the 20th century and it is fundamental for development and application of enzyme processes. All enzymes are proteins, though the converse is not true. Since enzymes are proteins, they share characteristics with all the other biopolymers. They have a large linear folded structure with α -aminoacids as building blocks that defines their primary structure. The synthesis direction from mRNA is always $NH_3^+ \rightarrow COO^-$. The secondary structure elements are stabilized by hydrogen bonds, or hydrophobic interactions between amino acid residues. Secondary structure elements are collected in domains due to a mix of hydrophobic, charge, and dipole interactions, while tertiary structure consists of more than one domain and is stabilized by aminoacid residues that are far away from each other in the primary structure. A close examination and full details can be read in general biochemistry textbooks [19][20][21]. The main characteristic

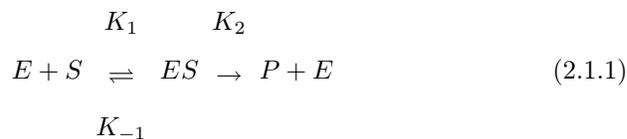
that defines and distinguishes enzymes as a subset of proteins is their catalytic activity. The first enzyme isolated and crystallized is due to a pioneering work of Sumner [5] where the concept of enzyme was established:

I discovered on the 29th of April a means of obtaining from the jack bean a new protein which crystallizes beautifully and whose solutions posses to an extraordinary degree the ability to decompose urea into ammonium carbonate [...] Solutions of the enzyme have an activity of 100'000units/gr of dry material. In other words, 1gr of dry material will produce 100'000mg of ammonia nitrogen from a urea-phosphate solution in 5min at 20°C. (James B. Saumner, 1926)

The active site of an enzyme is defined by all the residues involved in catalysis plus the binding sites that stabilize the complex between receptor and substrate.

Michaelis-Menten Equation

Enzyme catalysis falls under the category of complex reactions. The simplest kinetic to model enzyme reaction has been proposed by Michaelis and Menten. An enzyme (E) acts by forming a complex (ES) with a substrate (S) which can then react to form the product (P). The overall reaction thus involve at least three steps:



The speed with which [S] is consumed is:

$$v = K_2 [ES] \quad (2.1.2)$$

Employing the steady state approximation

$$[ES] = K_1 [E] [S] / (K_{-1} + K_2) \quad (2.1.3)$$

the speed in can be rewritten as

$$v = K_2 \cdot \{K_1 [E] [S] / (K_{-1} + K_2)\} \quad (2.1.4)$$

When the concentration of the substrate is high, the fraction of enzyme in the form of a complex can be very high. The ordinary procedure is to evaluate v in terms of the total enzyme concentration $[E_0]=[E]+[ES]$:

$$[E_0] = [E] + [ES] = [E] \cdot \{1 + K_1 [S] / (K_{-1} + K_2)\} \quad (2.1.5)$$

Collecting the [E] term to the left

$$[E] = [E_0] (K_{-1} + K_2) / \{(K_{-1} + K_2) + K_1 [S]\} \quad (2.1.6)$$

and substituting it in eq. 2.1.4

$$v = \frac{K_1 K_2 [E_0] [S]}{(K_{-1} + K_2) + K_1 [S]} \quad (2.1.7)$$

then inverting eq. 2.1.7

$$\frac{1}{v} = \frac{1}{K_2 [E_0]} + \frac{(K_{-1} + K_2)}{K_2 K_1 [E_0] [S]} \quad (2.1.8)$$

and proceeding with the following substitutions

- $V_{max} = K_2[E_0]$ where K_2 is called K_{cat} or the turnover number.
- $K_m = (K_{-1} + K_2)/K_1$ is the Michaelis-Menten constant.

the Michaelis-Menten equation for velocity is finally obtained:

$$v = \frac{K_{cat} [E_0] [S]}{[S] + K_m} = \frac{V_{max} [S]}{[S] + K_m} \quad (2.1.9)$$

Turnover Number and Specificity

The turnover number, K_{cat} , has the dimension of sec^{-1} . It represents the number of molecules converted per second by the enzyme under the condition of substrate saturation. It is a measure of the activation free energy of the reaction. The Michaelis and Menten constant, K_m , has the dimension of $mol \cdot L^{-1}$. It represents the concentration of the substrate at which 50% of the maximal velocity V_{max} has been reached. It gives an idea of the substrate binding energy. The smaller it is the higher it is the affinity between the substrate and the enzyme. Diagrams that plots the velocity of reaction as a function of substrate concentration are called Michaelis-Menten plots. The ratio K_{cat}/K_m is an apparent constant called specificity constant and it is a quantitative measure of the specificity of the substrate for the enzyme. The specificity constant cannot be larger than the rate of diffusion-controlled bimolecular reaction ($< 10^8 - 10^9 M^{-1}s^{-1}$)[22]. The turnover number can be determined with a titration. For example titration is possible with enzymes such as serine hydrolases, because they have the characteristic of forming an irreversible covalently

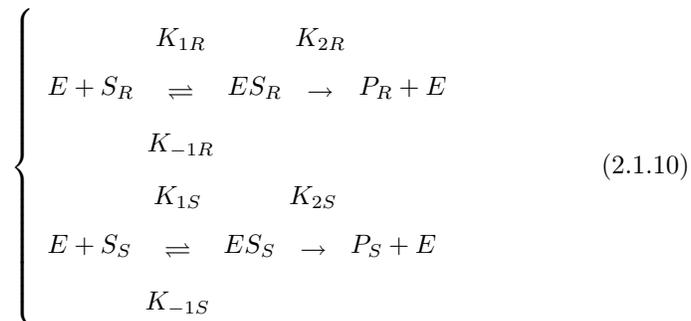
Enzyme	Substrate	T [°C]	pH	K_{cat} [s^{-1}]	K_m [mM]	K_{cat}/K_m [$M^{-1}s^{-1}$]
Trypsin (bovine)	Benzoyl-Arg↓NH ₂	25.5	7.8	27	2.1	13000
	Benzoyl-Arg↓OEt	25.0	8.0	19	0.02	1000000
	Z-Lys↓OMe	30.0	8.2	101	0.23	440000
	Z-Lys↓Ala	40.0	8.2	0		≈ 0
	Z-Lys↓Ala-Ala-Ala	40.0	8.2	4.6	3.7	1200
α-Chymotrypsin (bovine)	Acetyl-Tyr↓OEt	25.0	7.0	160	3.7	42000
	Acetyl-Tyr↓NH ₂	25.0	7.8	0.28	7	40
	Benzoyl-Tyr↓OEt	25.0	7.8	78	4	20000
	Benzoyl-Phe↓OEt	25.0	7.8	37	6	6300
	Benzoyl-Met↓OEt	25.0	7.8	0.77	8.8	1000
	Ac-ProAlaProPhe↓Ala	37.0	8.0	0		≈ 0
	Ac-ProAlaProPhe↓AlaAlaNH ₂	37.0	8.0	37	0.83	44000

Table 2.1: Turnover number K_{cat} and Michaelis-Menten constant K_m for different substrates for two hydrolases. The arrow symbol ↓ indicates the bond that is hydrolyzed.

acylated complex in the active site[22]. Specificity constant can vary several order of magnitude for different substrates. In table 2.1 are represented some turnover number for different substrates of Trypsin and α-chymotrypsin.

Stereoselectivity and Enantiomeric Ratio

When we deal with stereoselectivity we need to evaluate two competitive reactions: one for the (R)-substrate and another one for the (S)-substrate.



Since the kinetic constants are different, than the velocities of the two reactions (v_R and v_S) will be different too:

$$\left\{ \begin{array}{l} v_R = \frac{K_{2R}[S_R][E]}{K_{m,R}} \\ v_S = \frac{K_{2S}[S_S][E]}{K_{m,S}} \end{array} \right. \quad (2.1.11)$$

The enantiomeric ratio, E , is the ratio between the two specificity constants for the (R)-substrate and the (S)-substrate, $(K_{sp})_R$ and $(K_{sp})_S$:

$$\frac{v_R}{v_S} = \frac{[S_R] K_{2R} \cdot K_{m,S}}{[S_S] K_{2S} \cdot K_{m,R}} = \frac{[S_R]}{[S_S]} \cdot \frac{(K_{sp})_R}{(K_{sp})_S} = \frac{[S_R]}{[S_S]} \cdot E \quad (2.1.12)$$

$$\frac{v_R}{v_S} = \frac{d[S_R]}{d[S_S]} = \frac{[S_R]}{[S_S]} \cdot E \quad (2.1.13)$$

Combining equations 2.1.12 and 2.1.13, the enantiomeric ratio can be expressed as:

$$E = \frac{d[S_R]}{[S_R]} \frac{[S_S]}{d[S_S]} = \frac{d \ln [S_R]}{d \ln [S_S]} = \frac{\ln \left(\frac{[S_R]}{[S_R^0]} \right)}{\ln \left(\frac{[S_S]}{[S_S^0]} \right)} \quad (2.1.14)$$

To be measured experimentally, the enantiomeric ratio is expressed in terms of the enantiomeric purity, ee and conversion, C

$$C = \text{conversion} = 1 - \frac{[S_R] + [S_S]}{[S_R^0] + [S_S^0]} \quad (2.1.15)$$

$$ee_P = \text{enantiomeric purity product} = \frac{[P_R] - [P_S]}{[P_R] + [P_S]} \quad (2.1.16)$$

$$ee_S = \text{enantiomeric purity substrate} = \frac{[S_S] - [S_R]}{[S_S] + [S_R]} \quad (2.1.17)$$

So we have the enantiomeric ratio defined in terms of enantiomeric purity and conversion if the experimental setup guarantee conversion data without big errors:

$$E = \frac{\ln [1 - C (1 + ee_P)]}{\ln [1 - C (1 - ee_S)]} \quad (2.1.18)$$

$$E = \frac{\ln [1 - C (1 - ee_S)]}{\ln [1 - C (1 - ee_P)]} \quad (2.1.19)$$

Substrate	K_m [mM]	K_{cat} [s^{-1}]	K_{cat}/K_m [$M^{-1}s^{-1}$]	Stereoselectivity $(K_{cat}/K_m)_S / (K_{cat}/K_m)_R$
(<i>S</i>) Phg ↓ OMe	50	0.46	9.2	14
(<i>R</i>) Phg ↓ OMe	140	0.08	0.57	
<i>N</i> -Acetyl-(<i>S</i>)-Phg ↓ OMe	4	0.8	200	285
<i>N</i> -Acetyl-(<i>R</i>)-Phg ↓ OMe	7	0.005	0.7	

Table 2.2: Stereoselectivity in the hydrolysis with α -chymotrypsin

The expression of the enantiomeric ration in terms of enantiomeric purities can overcome the problem of the usually big uncertainty that comes with the measurement of conversion:

$$E = \frac{\ln \left[\frac{1-ee_S}{1+(ee_S/ee_P)} \right]}{\ln \left[\frac{1+ee_S}{1+(ee_S/ee_P)} \right]} \quad (2.1.20)$$

The enantiomeric ratio is a logarithmic function of ee , therefore even small variation Δee leads to large variation of the enantiomeric ratio.

In figure 2.1.1 are plotted values of enantioselectivity as a function of conversion and enantiomeric purity. Measurements of enantiomeric purity ee_P must be recorded before 50% of conversion, otherwise the abrupt decreases of ee_P at high values of conversion will affect the measure of E . Viceversa, if it is followed the enantiomeric purity ee_S , the measurement must be recorded after 50% of conversion.

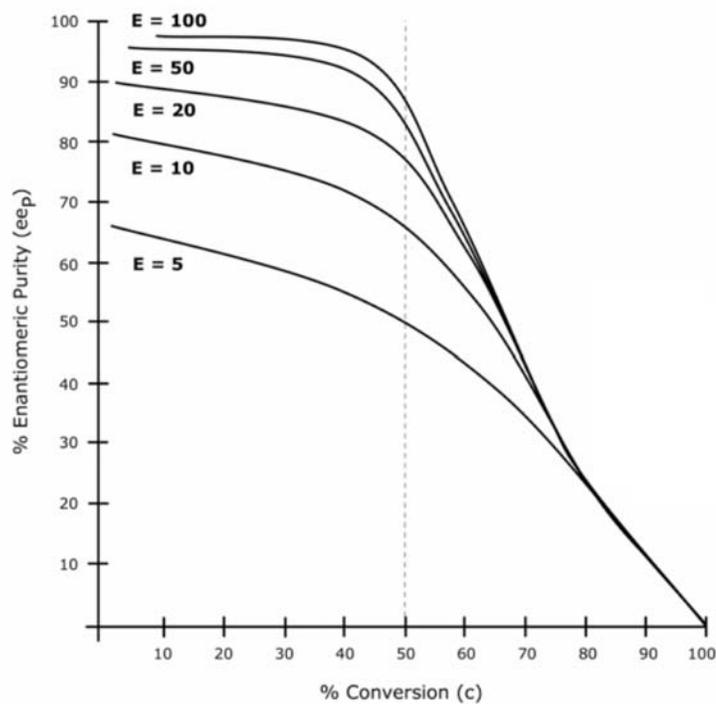


Figure 2.1.1: Enzyme Enantioselectivity

pH dependencies

Previously we have omitted any consideration on the influence of pH on K_{cat} and K_m values of the enzyme. Acidic and basic groups can be found on protein surfaces, but they can also form part of the active site. For α -chymotrypsin the correlation with pH can be analytically explained following the work of Bender et al. [23]. When the histidine is protonated K_{cat} is zero, and the enzyme function is lost. Only the fraction of enzyme with non-protonated histidine is active, K_{cat} becomes a function of pH:

$$K_{cat} = \frac{K_{cat}^0}{1 + \frac{[H^+]}{K_1}} \quad (2.1.21)$$

where:

- K_{cat}^0 is the turnover number without any pH dependence;
- K_1 is the acid dissociation constant for the histidine in the active site;

The same dependence must be introduced also in the Michaelis-Menten constant, if the acidity of another group interferes with the energy of binding:

$$K_m = \frac{K'_m}{\left(1 + \frac{K_2}{[H^+]}\right)} + \frac{K''_m}{\left(1 + \frac{[H^+]}{K_2}\right)} \quad (2.1.22)$$

where:

- K'_m and K''_m are Michaelis Constants without any pH dependence;
- K_2 is an acid dissociation constant.

$$v = \frac{\frac{K_{cat}^0[E_0][S]}{\left(1 + \frac{[H^+]}{K_2}\right)}}{\left[[S] + \frac{K'_m}{\left(1 + \frac{K_2}{[H^+]}\right)} + \frac{K''_m}{\left(1 + \frac{[H^+]}{K_2}\right)} \right]} \quad (2.1.23)$$

The acid dissociation constants of aminoacid in enzymes and in aqueous solutions are similar, so obtaining a diagram such as that one in figure 2.1.2 can give some qualitative information about the aminoacid groups that gives the pH dependence. In the case of α -chymotrypsin the aminoacid that accounts for $pK_1 = 7$ is His57. While K_m value is influenced by a residue with $pK_2 \approx 0$ that should belong to an amino group.

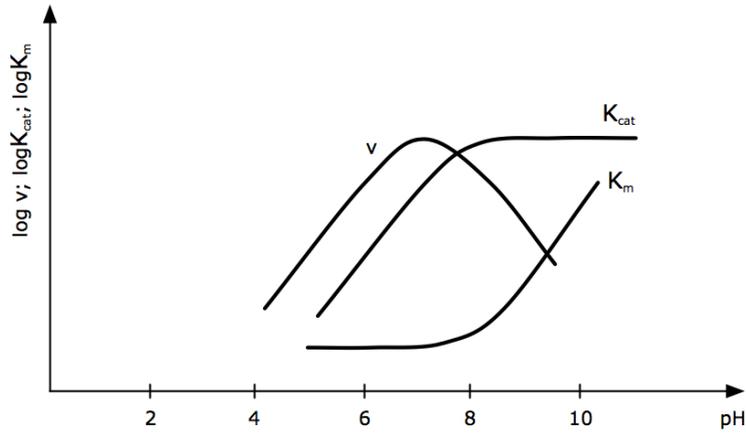
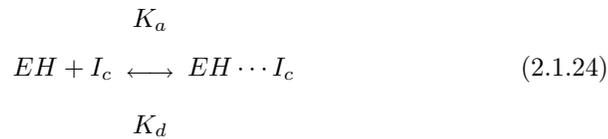


Figure 2.1.2: pH dependence of v , K_{cat} , and K_m for α -chymotrypsin-catalyzed hydrolysis of an uncharged substrate[23].

Binding of an inhibitor Molecules

Structural variation of the active site center of an enzyme will produce changes in the rate of the reaction of equation 2.1.9. An activator is a molecule that increases the rate, on the other hand an inhibitor is a molecule that decreases the rate. When an inhibitor is present we have a competitive reaction of this kind:



The inhibitor constant, K_i , is the ratio between K_a and K_d :

$$K_i = K_d / K_a \quad (2.1.25)$$

When an inhibitor is bound to the enzyme it inhibits the enzyme to perform

its catalytic function. We can have various types of inhibition:

- Product Inhibition: for example in the case of α -chymotrypsin the formation of H^+ during the reaction that can protonate His57;
- Irreversible Acylation: the covalent bond is formed between receptor and inhibitor, such a reaction is irreversible and it prevents further chemical reaction with the substrate;
- Metal Ions Inhibitor: chelating agents such as EDTA can bind metal ions that are essentials for the activity of the enzyme.

Inhibitors or activators can also bind the enzyme in a site that is different from the active site. When such an event happens we need to consider the following equilibrium besides the the enzyme reaction:



The enzyme bound to the inhibitor or activator, $EH \cdots M$, is different from the free enzyme EH . Such an interaction can induce modification of the geometries within the active site. The inhibitor or activator is not bound in the active site. This means that enzyme catalyzed reaction can take place. Anyway the enzyme is modified, which leads to have different K_{cat} and K_m . We have often this case when the activator is a metal ion activator such as calcium ion.

2.1.1 Free Energy Theories of Enzyme Catalysis

As we previously mentioned, enzymes are efficient biological catalysts obtained as a product of life evolution processes. Their strength is to allow organisms to carry out reaction that are compatible with life time-scale. Natural selection created enzyme with a chemical kinetic enhancement of 10^6 to 10^{20} order of magnitude with respect with the uncatalyzed reaction in solution. They are not only remarkably efficient catalysts, but they are also very specific: the reactive complex is only formed between the receptor and the right substrate.

In science we always refer to a reference states to evaluate differences. Enzyme catalysis is not different from the other fields. To understand enzymatic catalysis it is therefore necessary to have a reference uncatalyzed reaction. The reference reaction is ideally the same reaction carried out in solution. The following discussion refers to 2.1.3. The formation of a complex between enzyme and substrate leads to the first stabilization of the energy. The so called Michaelis Complex (MC in what follows) has a free energy of binding defined as a function of the dissociation constant K_m :

$$\Delta_{MC(enzyme)} G^{bind} = e^{\frac{K_m}{RT}} \quad (2.1.27)$$

As a final product of molecular evolution the second step of the reaction, K_{cat} , is so efficient that it can approach the diffusion limit. If we assume i) the validity of Transition State Theory (TST) [24] and ii) the speed of second step below the diffusion limit, than we can also express a variation of Gibbs free energy as a function of the turnover number K_{cat} :

$$\Delta_{TS} G^{cat} = e^{\frac{K_{cat}}{RT}} \quad (2.1.28)$$

At low substrate concentration the initial reaction rate can be expressed as:

$$v_0 = \frac{K_{cat}}{K_m} [E_0] [S_0] \quad (2.1.29)$$

while at high substrate concentration the initial reaction rate depends exclusively from K_{cat} :

$$v_0 = K_{cat} [E_0] \quad (2.1.30)$$

Considering the uncatalyzed reaction in solution with a reaction profile shown in figure 2.1.3, a single rate constant K_{uncat} is associated with $\Delta_{TS}G^{uncat}$. Comparing $\Delta_{TS}G^{uncat}$ with $\Delta_{TS}G^{cat}$ we obtain a ratio between K_{uncat} and K_{cat} that is the definition of catalytic rate enhancement [25]:

$$\text{enzyme rate enhancement} = \frac{K_{cat}}{K_{uncat}} \quad (2.1.31)$$

Examples of enzyme rate enhancement are shown in figure 2.1.4 where are reported K_{cat} and K_{uncat} for reactions at 25°C. The length of each bar represents the rate enhancement by arginine decarboxylase[26], orotidine 5'-phosphate decarboxylase[27], staphylococcal nuclease[28], sweet potato β -amylase[29], fumarase[30], mandelate racemase[31], carboxypeptidase B[32], E. coli cytidine deaminase, ketosteroid isomerase [27], chorismate mutase[33], carbonic anhydrase[27]. The energy gain of the activation free energy can be deduced from figure 2.1.3 with the quantities listed in figure 2.1.5:

$$\Delta_{TS}G^{cat} - \Delta_{TS}G^{uncat} = \Delta_{MC_{(enzyme)}}G^{bind} - \Delta_{TS}G^{bind} \quad (2.1.32)$$

Since left hand member must be positive, than $\Delta_{TS}G^{bind}$ needs to be more negative than $\Delta_{MC(enzyme)}G^{bind}$. This means the affinity of the enzyme is bigger towards the transition state compare to the MC complex. The two free Gibbs energies on the right hand member of figure 2.1.5 represent the main division between the two group of theories that are used to interpret the enzyme ability to speed up chemical reactions:

1. Enzyme Transition State Stabilization Theories (TSS):

$$\Delta_{TS}G^{cat} - \Delta_{TS}G^{uncat} \simeq \Delta_{TS}G^{bind}$$

1. Enzyme MC Theories:

$$\Delta_{TS}G^{cat} - \Delta_{TS}G^{uncat} \simeq \Delta_{MC(enzyme)}G^{bind}$$

2.1.1.1 Transition State Theories (TSS) and electrostatic catalysis

Linus Pauling formulated first the concept of an enzyme environment complementary to the transition state structure[34]. The idea of a transition state stabilization due mainly to electrostatic pre-organization of the active site have been supported by works of Warshel and co-workers[35][36][37][38].

The issue is not the binding itself but rather the change in binding energy by moving from reactant state to the transition state. The first early attempt to quantitatively address catalysis as electrostatic effect that stabilizes transition state was a work by Warshel and Levitt[35] with the study of the stability of the carbonium ion intermediate formed in the cleavage of a glycosidic bond by lysozyme. It was found that electrostatic stabilization is an important factor in increasing the rate of the reaction step that leads to the formation of

the carbonium ion intermediate. Steric factors, such as the strain of the substrate on binding to lysozyme, did not seem to contribute significantly. The definition of “electrostatic catalysis” includes effects of the protein charges, permanent dipoles, induced dipoles effects as polarizability, and solvation by bound water molecules. It does not include van der Waals strain effects, covalently induced charge transfer interactions, orientational entropy, and dynamical effects. Transition State Theory (TST) is based on the concept of the existence of a hyperspace in the phase space with two properties: i) it divides the phase space into a reactant region and a product region, ii) trajectories coming from the reactant region cannot recross before being thermalized in the product region. Transition state is often mistakenly seen as a topological saddle point on the potential energy surface. This is not the case because the so called transition state is an ensemble of structure that implies the existence of a “soft” barrier between reagents and products. Rate constant in the condensed phase can be written as:

$$K = kK_{TST} \quad (2.1.33)$$

where:

- K is the rate constant of the reaction;
- k is the transmission factor that corrects non-recrossing;
- K_{TST} is the TST rate constant

The transmission factor, k , corrects for non-recrossing. It depends on two factors: i) the probability that a trajectory arriving at TS from reactants ends up in product rather than returning to reactant, ii) the average number of time

a trajectory passes back and forth across transition state before it moves to product. The last term of equation 2.1.33, K_{TST} , can be further developed as:

$$K_{TST} = \frac{1}{2} \langle |\dot{x}| \rangle_{TS} \frac{\exp\{-\beta\Delta g^{TS}\}}{\int_{-\infty}^{x^{TS}} \exp\{-\beta\Delta g(x)\} dx} = \frac{1}{2} \langle |\dot{x}| \rangle_{TS} \exp\{-\beta\Delta G^{TS}\} \quad (2.1.34)$$

Where:

- $\langle |\dot{x}| \rangle_{TS}$ represents the average of velocity over transition state configurations;
- $\beta = 1/K_B T$ where K_B is the boltzmann constant and T is the temperature;
- $\Delta g(x)$ is the free Gibbs energy functional that describes the probability of being at different point throughout the reaction coordinate x ;
- Δg^{TS} is the value assumed by $\Delta g(x)$ at the transition state;
- ΔG^{TS} represents the ratio between the probability of being at transition state and the probability of being in the reactant state.

The free energy functional $\Delta g(x)$ describes a potential energy surface that involves a proper coupling between solvent and solute charge distribution. The previous equations are valid both in solution and in protein environment. The folded enzyme provides a dipolar environment already preorganized so to stabilize the charge distribution in TS. In reactions carried out in water, the solvent pays a significant reorientation energy to orient dipoles to stabilize the charge distribution of TS.

When a solute is immersed in a polar solvent half of the energy gained from charge-dipole interaction, $\Delta G_{Q\mu}$, is spent on changing dipole-dipole interaction, $\Delta G_{\mu\mu}$:

$$\Delta G_{sol} \simeq \Delta G_{Q\mu} + \Delta G_{\mu\mu} \simeq \Delta G_{Q\mu} - \frac{1}{2}\Delta G_{Q\mu} = \frac{1}{2}\Delta G_{Q\mu} \quad (2.1.35)$$

In proteins the “reorganization energy” term, $\Delta G_{\mu\mu}$, is smaller than in solution because of the already oriented polar groups, ionized residues and bound water molecules. This means that less free energy is spent on creating oriented dipoles for TS.

The reorganization energy is related to the well-known Marcus reorganization energy. The Marcus reorganization energy is then related to the transfer from the reactant to the product state, even if here we limit ourselves to the TS. Following Marcus theory, the activation energy for chemical reaction can be approximated by:

$$\Delta G^{TS} \approx (\Delta G_0 + \lambda)^2 / 4\lambda - H_{12} + H_{12}^2 / (\lambda + \Delta G_0) \quad (2.1.36)$$

where:

- ΔG_0 is the free energy difference between product and reactants;
- λ is the solvent reorganization energy, the changes in solvent-solvent interactions during the reaction;
- H_{12} is the mixing between reactant and product states.

Both λ and ΔG_0 are leverages to have a reduction of ΔG^{TS} . The reduction of ΔG_0 is accomplished by a preorganized polar environment. Reduction of

ΔG_0 is accompanied by reduction of λ . Two limiting cases can clarify the role of λ , ΔG_0 and their coupled nature: a) when $\Delta G_0 \sim 0$ the catalytic effect is due to variation of Marcus reorganization energy, b) when $\Delta G_0 > 0$ where the catalytic effect is due to reduction of ΔG_0 .

2.1.1.2 MC stabilization and enzyme catalysis:

Enzyme MC Theories focuses the attention on the MC complex formation. In solution water molecules needs to adapt their dipoles from their initial random orientation to a configuration that follows the dynamical charge transformation of the substrate. This re-orientation of water dipoles requires time and energy, while in the enzyme environment the charges are already pre-organized to be suited for the reaction to occur. The term $\Delta_{MC(enzy\text{me})} G^{bind}$ contains contributions related to the substrate rearrangement from a fully solvated molecule to a molecule orientation induced by the enzyme that allows the reaction to properly take place.

To better elaborate MC Theories we refer to the free energy profiles of figure 2.1.3. We consider an imaginary substrate state in solution with the same atom arrangement of the MC complex in the enzyme: $MC_{(aq)}$. Following this line of reasoning $\Delta_{MC(enzy\text{me})} G^{bind}$ can be splitted in two terms:

$$\Delta_{MC(enzy\text{me})} G^{bind} = \Delta_{MC(aq)} G^R + \Delta_{MC(aq)} G^{bind} \quad (2.1.37)$$

Substituting equation 2.1.27 into 2.1.32

$$\Delta_{TS} G^{cat} - \Delta_{TS} G^{uncat} = \Delta_{MC(aq)} G^R + \Delta_{MC(aq)} G^{bind} - \Delta_{TS} G^{bind} \quad (2.1.38)$$

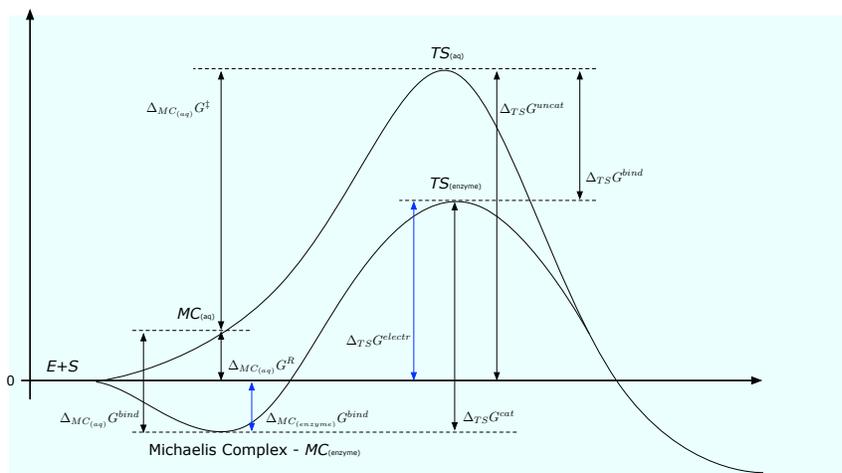


Figure 2.1.3: Reaction energy profile of enzyme catalysis compare to the same reaction in solution.

Since $\Delta_{MC(aq)}G^R > 0$ than a reduction of the activation energy can be obtained with an enzyme that presents a bigger affinity for MC structure ($\Delta_{MC(aq)}G^{bind}$) compare to TS structure ($\Delta_{TS}G^{bind}$). The pre-organization of the substrate that takes place in the enzyme active site and that requires a huge amount of energy when the substrate is solvated by water leads to a better affinity that drags down the energy so to have catalysis. There are some way to account for the pre-organization of the substrate, it depends on how the process is followed. We talk about entropic trap[39] when the MC formation happens with a loss of translational and rotational degrees of freedom so to have a loss of entropy. The loss of entropy is balanced by favourable interactions within the residues of the active site, so to have a negative free energy of binding. Another explanation accounts for both entropic and enthalpic contribution in $\Delta_{MC(aq)}G^R$ term. The cratic enthalpic contribution[40] comes when

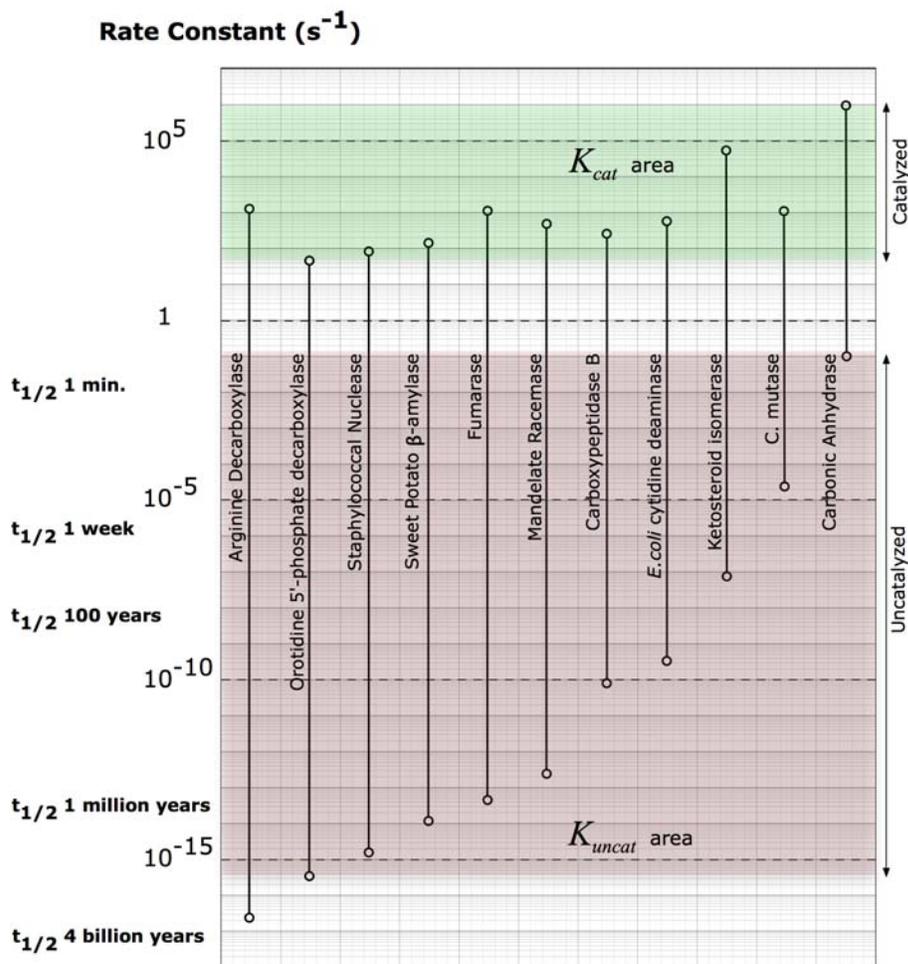


Figure 2.1.4: Logarithmic Values of K_{cat} and K_m at 25°C. The length of each bar represent the rate enhancement of arginine decarboxylase, orotidine 5'-phosphate decarboxylase, staphylococcal nuclease, sweet potato β -amylase, fumarase, mandelate racemase, carboxypeptidase B, *E. coli* cytidine deaminase, ketosteroid isomerase, chorismate mutase, carbonic anhydrase.

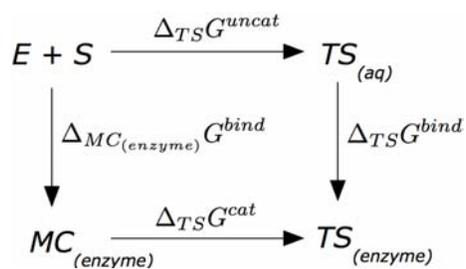


Figure 2.1.5: Free Energy Scheme

the reacting groups of the enzyme are getting nearer the substrate. Orbital steering[41] focuses instead on orbital overlap between reacting groups. At the basis of catalysis is the increased overlap between molecular orbitals involved in the chemical process.

2.1.2 Protein Motion and Hydrogen Transfer in Enzyme Catalysis

Protein fluctuations and the relative conformational relaxation times span a wide range of time scales (from the picoseconds to seconds[42]). What we have in protein is a hierarchy of relaxation times such as that one reported in the scheme of Figure 2.1.6. The motion of the amino acid groups of the protein is lattice-like while the conformational motion is diffusive like with multiple activation barriers along the reaction coordinate.

The mechanism of reaction of Serine Proteases is explained in chapter 4 section 4.5 . From crystallographic data the key distances for the proton transfers involved are too large to have a direct hydrogen transfer [43]. The histidine mobility handled stochastically to resemble the protein dynamics in view of obtaining a good distance asset between donor and acceptor groups, together with the curvature of the potential surfaces of the system are key points of the theory, this can lead us to the idea of the protein environment as an accumulator of the energy liberated after the first proton transfer step of the enzyme. The proton transfer is unlikely to happen even if these “stretching distances” are substantially deformed by a strong donor-acceptor interaction. This strongly suggests that conformational motion must accompany proton transfer. Histidine and serine are attached in folded regions where the fluctuation amplitudes of the individual atom displacement can be of the order of $0.3\text{-}0.4\text{\AA}^2$. It is therefore perfectly feasible that a geometry where the proton transfer is likely to happen can be temporarily reached as a consequence of the protein conformational fluctuations. Hydrogen stretching motions have a frequency of the order of 3000cm^{-1} ($6 \cdot 10^{14}\text{sec}^{-1}$)[43]. This time scale is a lot faster than a typ-

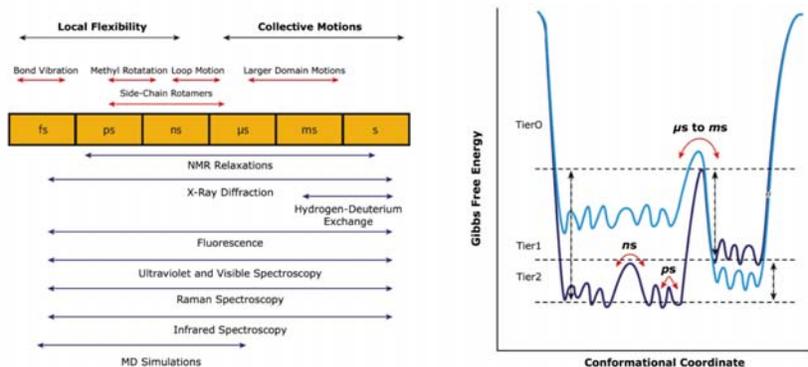


Figure 2.1.6: The energy landscape defines the amplitude and timescale of protein motions: a) Timescale of dynamic processes in proteins and the experimental methods that can detect fluctuations on each timescale; b) One-dimensional cross-section through the high-dimensional energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. Each tier is classified following the description introduced by Frauenfelder and co-workers. A state is defined as a minimum in the energy surface, whereas a transition state is the maximum between the wells. Lower tiers describe faster fluctuations between a large number of closely related sub-states within each tier-0 state. A change in the system will alter the energy landscape (from dark blue to light blue, or viceversa). For example, ligand binding, protein mutation and changes in external conditions shift the equilibrium between states. D. Kern, Nature. 450: 964 (2007). Reproduced by permission of Macmillan.

ical protein conformational relaxation time where frequencies are on the order of 10^6 sec^{-1} . This means that a special favorable proton transfer conformation can last enough so to have the hydrogen transfer accomplished. The energy of a particular conformational state can be transferred to the substrate enabling the system to overcome activation barriers and, only than, to be dissipated. This is the formal frame of the multi-phonon electronic transition theory[44]. Viewing the process as a potential free energy surfaces spanned by the proton coordinate, before the proton transfer the proton motion is represented by a double well

potential. Referring to figure 2.1.7 we describe the system before the proton transfer has been accomplished, with internal coordinate $q = q_i$, where the two minima are vertically separated by a certain amount of energy, and the proton is trapped in the well on the right of the figure. The energies of the two minima coincide when $q = q^*$ and the proton transfer has occurred. It is in this configuration that the vibrational frequency of the hydrogen may change from its initial values, ω , to a lower value ω^* . To reach q^* the energy of the initial well must be raised while the final well energy has to be decreased. What is moving up and down the potential well is the conformational motion of protein. In figure 2.1.7 the energy is presented as a function of the protein conformational coordinate and the potential wells, previously introduced, as a function of an internal coordinate, q . Every conformation of the protein is associated with a local minimum. For each of the value of the protein conformational coordinate, Q , there is a potential energy distribution $V_1(Q)$. Varying our position on $V_1(Q)$ means formation and breaking of bonds, formation and breaking of hydrogen bond networks, perturbations in the protein chain *et cetera*. All the energy of the system accumulated moving in the protein conformational space it is then lost in the local minimum after each transition.

In the case of Serine Proteases the two intermediates that needs a proton transfer to be accomplished are covalent intermediates. After the attack of the substrate and after the first proton has been transferred we have a situation where the curvature of the lower energy potential, $V_1(Q)$, or the potential of the system before the proton is transferred from the serine oxygen to the nitrogen of the imidazole ring of the histidine, is bigger than the potential of the intermediate, $V_m(Q)$. Figure 2.1.8 gives a graphical representation of the

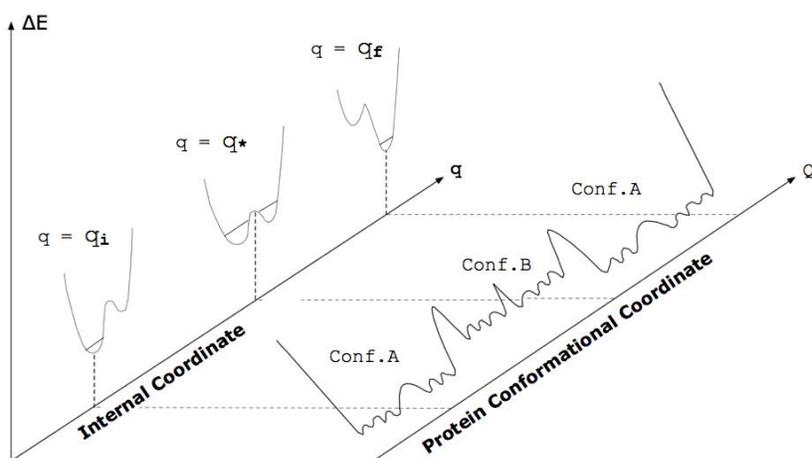


Figure 2.1.7: On the right are shown the potential surfaces along the internal coordinate q , while on the left we have a one-dimensional cross-section through the high-dimensional energy landscape of a protein. Transition between conformation A (Conf. A) and conformation B (Conf. B) have relaxation times of the order of the μs . When the protein passes from the conformation A to the conformation B the structure of the active site changes and the distances reach values suitable for a proton transfer. The conformation B lasts long enough to assure the accomplishment of the proton transfer. When the hydrogen passed from the donor group to the acceptor group another fluctuation of the protein trapped the proton on the acceptor side.

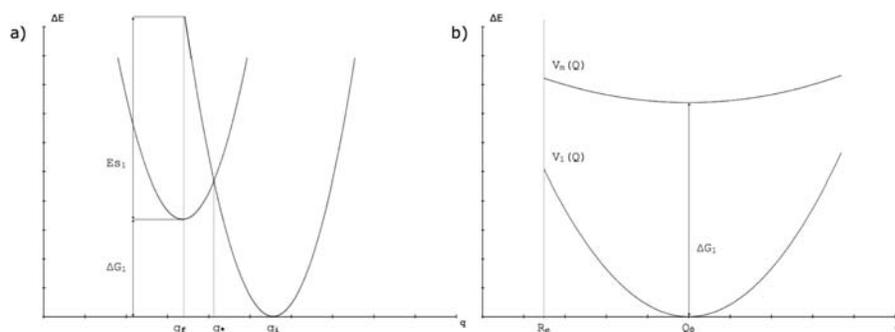


Figure 2.1.8: a) Potential energy surfaces along a molecular coordinate, q , for fixed protein conformational coordinate, Q . b) Potential energy surfaces along the protein conformational coordinate. The lower curve represents the initial state prior proton transfer.

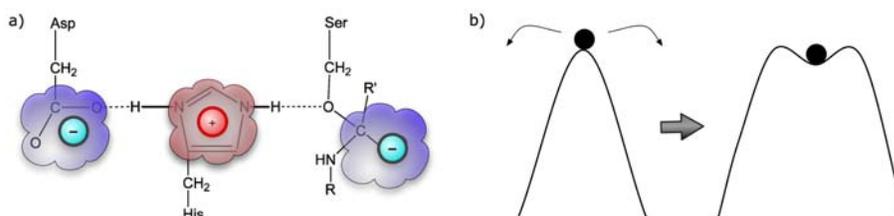
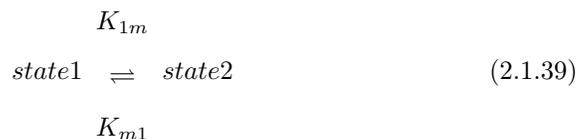


Figure 2.1.9: Unstable equilibrium and system mobility: a) the alignment of the catalytic triad and the relative electronic instability asset of the tetrahedral-intermediate. The motion of the histidine is increase when the adduct is formed. Due to the electronic attraction the histidine can easily reach the aspartate or the adduct part as it would be a ball in a situation of unstable equilibrium; b) Qualitative idea of a protein well soft as a pudding. Deformation of the potential well can lead to a temporary stabilization of unstable equilibrium and to an accumulation of energy.

potential energy surfaces along the protein conformational coordinate.

The difference in energy ΔG_1 between the two curves is that one obtainable with the protein conformational coordinate fixed to the value that is not favorable to have hydrogen transfer. This is the free energy variation of the tetrahedral intermediate state relative to the initial state. The curvature of the

potential shape of the intermediate is due to the charge alignment within the active site. The histidine positively charged is electrostatically unstable, but the potential shape of this intermediate incorporates energy associated with conformational distortion in the protein conformational coordinate Q . This energy contribution is approximately equal to the similar contribution in $V_1(Q)$. It might be the counteraction of this two contributes that gives an overall positive curvature to $V_m(Q)$ though the alignment in the intermediate state structures is maintained. The potential $V_m(Q)$ is shown as the upper curve in figure 2.1.8 b). Proton transfer occurs when the coordinate Q is shifted from Q_0 to R_c due to the thermal fluctuations of Q along the potential surface $V_1(Q)$. Let's say that the process of formation of the tetrahedral intermediate from the Michaelis Complex is characterized by the rate constant K_{1m} and the rate constant K_{m1} for the inverse process:



The two rate constants, K_{1m} and K_{m1} , have the characteristic free energy relationship with the Gibbs free variation of energy between the intermediate-state and the initial state.

$$K_{1m} = K_{m1} \exp\left(\frac{\Delta G_1}{KT}\right) \quad (2.1.40)$$

The two energy surfaces $V_1(Q)$ and $V_m(Q)$ have a qualitative different curvatures as it is shown in Figure 2.1.8. It has been chosen a suitable model fluctuation potential:

$$\begin{cases} V_1(Q) = \frac{1}{2}(Q - Q_0)^2 & \text{for } Q \geq R_c \\ V_1(Q) = \infty & \text{for } Q < R_c \end{cases} \quad (2.1.41)$$

Where:

- $Q = R_c$ represents the most favorable proton transfer geometry for rigid systems.

With such a potential it is possible to obtain the expression of the activation free energy for the proton transfer at $Q = Q_0$.

$$G_{1m}^* = \left[\frac{(E_{s1} - \Delta G_1)^2}{4E_{s1}} \right] + \left[\frac{1}{2} b_1 \gamma_1 (R_c - Q_0)^2 \right] + \left[\frac{1}{2} \hbar (\omega - \omega^*) \right] \quad (2.1.42)$$

where:

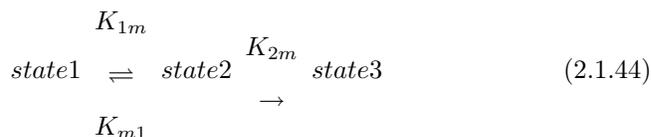
- the first term $\left[\frac{(E_{s1} - \Delta G_1)^2}{4E_{s1}} \right]$ is the activation energy from fluctuations in the molecular modes at $Q = Q_0$;
- the second term $\left[\frac{1}{2} b_1 \gamma_1 (R_c - Q_0)^2 \right]$ represents the free energy variation of the transition state for proton transfer at the most favorable protein conformation $Q = R_c$, where the curvature of the potential shape, b_1 , appear together with γ , a positive factor in the range between 0.0 and 0.5. The last term is the proton zero point energy difference between state1 and the transition state.

A similar equation can be obtained for the acyl intermediate step:

$$G_{1m}^* = \left[\frac{(E_{s2} - \Delta G_2)^2}{4E_{s2}} \right] + \left[\frac{1}{2} b_2 \gamma_2 (R_c - Q_0)^2 \right] + \left[\frac{1}{2} \hbar (\omega - \omega^*) \right] \quad (2.1.43)$$

The terms have the same meaning but ΔG_2 is less than ΔG_1 , and the conformational coordinate, Q' , is indeed different from equation 2.1.42.

If the rate determining step is an hypothetical state3:



The steady state conditions for state1 and state2 are:

$$\left\{ \begin{array}{l} -K_{m1}P_1 + K_{1m}P_m + S_1 = 0 \quad \text{Steady State Condition (state1)} \\ K_{m1}P_1 - K_{1m}P_m - K_{2m}P_m = 0 \quad \text{Steady State Condition (state2)} \end{array} \right. \quad (2.1.45)$$

The obtained catalytic constant for acylation can be regarded as the overall rate constant:

$$S_1 = \frac{K_{m1}K_{2m}}{(K_{1m} + K_{2m})} P_1 = K_{cat} P_1 \quad (2.1.46)$$

The formalism presented is that one of the multiphonon theory [44]. It can be used to evaluate mechanism of action during the elementary steps in serine-proteases enzymes. It is worth spending couple of words about the histidine motion. Hydrogen bonds, free enzyme and substrate-enzyme complex motions determine the shape and the relative characteristic of the potential surfaces. The increased histidine motion as a consequence of the tetrahedral adduct leads to a lower curvature of the first intermediate potential well compare to the well's curvature either of the initial state and of an other intermediate state. The histidine mobility is the results of the interplay between different phenomena: i) after the first hydrogen transfer the hydrogen bond of the histidine got broken

giving more motional freedom to the residue; ii) the alignment of the three residues increases the mobility destabilizing the structure as it is shown in Figure 2.1.9; iii) the conformational motions of the protein acting so to stabilize the structure of the intermediate, acting against the mobility. The potential would be that one of an unstable equilibrium if it would have not been softened by the protein conformational motions as in Figure 2.1.9 b). In this way, part of the high energy of the tetrahedral adduct is stored in the protein system and it may be used further up the road to overcome the activation barrier of the next intermediate.

Chapter 3

Computational Methods

3.1 *ab initio* Methods

An introduction to Quantum Mechanics Theory (QM Theory) is provided in the next subparagraph. All the *ab initio* method used in this Ph.D. thesis, namely Hartree-Fock method, Møller-Plesset method and DFT method, will follow.

3.1.1 Quantum Mechanics Theory

Quantum mechanics enables us to calculate the sizes, shapes, and energies of atoms and molecules. The basis of this understanding is the wave-particle duality in the nature of fundamental entities such as electrons, photons, or nuclei. Such a basis rests on de Broglie relations between the first order momentum (p) of a particle and its wavelength (λ), and between energy (E) of a particle and its frequency (ω)

$$p = \frac{h}{\lambda} \tag{3.1.1}$$

$$E = h\omega \tag{3.1.2}$$

Quantum mechanics is usually introduced by postulates. We have chosen to produce a list of postulates according to Molecular Quantum Mechanics textbook by Peter W. Atkins [45]:

- **Postulate 1.** The state of a quantum mechanical system is completely specified by a wave function $\Psi(\vec{r}, t)$ that depends on the coordinate of particles and on time.
- **Postulate 2.** To every observables in classical mechanics there corresponds a linear, Hermitian operator in quantum mechanics.
- **Postulate 3.** If a system is described by the wave function $\Psi(\vec{r}, t)$, the average observable value Ω of a series of experimental measurements is equals to the expectation value $\langle \Omega \rangle$ of the relative operator $\hat{\Omega}$ as a result of the following integral:

$$\langle \Omega \rangle = \frac{\langle \Psi | \hat{\Omega} | \Psi \rangle}{\langle \Psi | \Psi \rangle}$$

- **Postulate 4.** The probability of finding a particle in an infinitesimal volume $d\tau$ centered at the point \vec{r} is proportional to $\langle \Psi | \Psi \rangle$.
- **Postulate 5.** The exact wave function $\Psi(\vec{r}, t)$ must satisfy the Schrödinger equation:

$$i\hbar \frac{\partial \Psi}{\partial t} = \hat{\mathcal{H}} \Psi$$

- **Postulate 6.** The exact wave function $\Psi(\vec{r}, t)$ must be antisymmetric with respect to the interchange of a coordinate (both space and spin) of any two electrons.

Postulates appears to be a bit cryptic, a pictorial introduction of the concepts of the theory will be briefly presented. Quantum mechanics theory can be applied whenever a microscopic system is confined between barriers of potential energy, in that case all the information about the system is completely specified by a wave function (as postulate 1 says). Everytime a wave motion is confined to a finite region of space, a series of patterns of stationary waves emerge, whose shapes and frequencies are dependent by the nature of the confinement. A confinement could be represented by any kind of field. The emerging patterns that are simpler in shape have lower frequencies, therefore lower energies as stated in equation 3.1.2. Each of these morphic pattern is connected with the existence of discrete quantum states with well defined properties (as a consequence of postulate 5). Since there is a confinement, it follows that in a confined region an electron cannot be completely at rest, because the confined wave cannot have wavelength comparable or smaller than the dimension of the confinement itself. Relation 3.1.2 states that the maximum wave length corresponds to the minimum energy, so the kinetic energy is always present and never null, because it will be at least as big as the one that corresponds to the wave of maximum wavelength (λ_{max}). Following the first de Broglie relation 3.1.1 a qualitative kinetic energy of the particle can be expressed

$$K = \frac{p^2}{2m} \sim \frac{h^2}{2mV^{\frac{2}{3}}} \quad (3.1.3)$$

where:

- m is the mass of the particle;
- $V^{\frac{1}{3}}$ is the linear dimension of the confining region of volume V because of the approximation: $\lambda_{max} \sim V^{\frac{1}{3}}$.

The spontaneous tendency of all systems is to decrease their energy. The minimum kinetic energy of the confined particle is like a minimum pressure that the particle exerts over the walls of the confinement field to expand the confining region, so to decrease the energy. This is the origin of what is called “zero point vibrational energy”. Furthermore *the Pauli’s Principle* (postulate 6) says that if \mathcal{N} equal particles are confined in a volume V , the lowest kinetic energy of each particle is not that one of a single particle 3.1.3, but it is higher:

$$K = \frac{h^2}{2md^2} \quad (3.1.4)$$

where:

- $d = \left(\frac{V}{\mathcal{N}}\right)^{\frac{1}{3}}$ the linear dimension of a volume which is $1/\mathcal{N}$ times the confining volume V .

As can be seen from 3.1.4 in case of several particles the confining region for each particle is not the total volume but the volume divided for the numbers of particles. This means that each particle has a private portion of volume to which it is confined and that determines its minimum energy. The minimum pressure exerts by a system of \mathcal{N} particle is therefore $\mathcal{N}^{\frac{5}{3}}$ larger than the pressure exerts by only one particle

$$P = -\frac{\partial K}{\partial V} \sim \frac{h^2}{m} \left(\frac{\mathcal{N}}{V}\right)^{\frac{5}{3}} \quad (3.1.5)$$

Only particles with opposite spin can share the same private volume of another particle. The exact expression for the pressure of \mathcal{N} particles where $\mathcal{N}/2$ with spin UP and $\mathcal{N}/2$ with spin DOWN is therefore very similar to 3.1.5:

$$P = \frac{1}{5} (3\pi^2)^{\frac{2}{3}} \frac{h^2}{m} \left(\frac{\mathcal{N}}{V}\right)^{\frac{5}{3}} \quad (3.1.6)$$

3.1.2 SCF Hartree-Fock Method (HF Method)

Following the Born Oppenheimer approximation the wave function can be factorized in a electronic part and a nuclear part. The effective Hamilton operator depends on the coordinates of all the N electrons $\{\vec{r}_1 \cdots \vec{r}_i \cdots \vec{r}_N\}$ and it parametrically depends on the n nuclei coordinates $\{\vec{Q}_1 \cdots \vec{Q}_i \cdots \vec{Q}_n\}$. Given a set of k orthonormal spatial orbitals $\{\psi_i\}$ it is possible to write a set of $2k$ spin-orbitals $\{\chi_i\}$ multiplying each ψ_i by the two spin functions $\alpha(\omega)$ and $\beta(\omega)$. The spin-orbitals obtained depend on the spatial coordinates \vec{r} , and on the spin values ω :

$$\begin{aligned} \{\psi_i(\vec{r})\alpha(\omega)\} & \quad i = 1, 2, \dots, k \\ \{\psi_i(\vec{r})\beta(\omega)\} & \quad i = 1, 2, \dots, k \end{aligned} \tag{3.1.7}$$

The eigenstate $|\psi\rangle$ of a system is expressed by a Slater Determinant, so to obtain wave function satisfying the *Pauli exclusion principle*. A Slater determinant is generally indicated through a row vector containing all the diagonal elements of a Slater matrix. In the case of N electrons we have

$$|\psi\rangle = |\chi_1(1)\chi_2(2)\cdots\chi_{2k}(N)\rangle = \left(\frac{1}{N!}\right)^{\frac{1}{2}} \text{Det} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_{2k}(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_{2k}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_{2k}(N) \end{vmatrix} \tag{3.1.8}$$

The energy of the eigenstate $|\psi\rangle$ is the expectation value of the Hamiltonian operator $\hat{\mathcal{H}}$ as follows:

$$\begin{aligned} E_0 = \langle \psi | \hat{\mathcal{H}} | \psi \rangle &= \sum_i^N \langle \chi_i | \hat{h} | \chi_i \rangle + \frac{1}{2} \sum_i^N \sum_j^N \langle \chi_i \chi_j | | \chi_i \chi_j \rangle = \\ &= \sum_i^N h_i + \frac{1}{2} \sum_i^N \sum_j^N (\mathcal{J}_{ij} - \mathcal{K}_{ij}) \end{aligned} \quad (3.1.9)$$

where

- $h_i = \int \chi_i^*(1) \left(-\frac{1}{2} \nabla^2(1) - \sum_Q^{N_{nuclei}} \frac{1}{r_{Q1}} \right) \chi_i(1) d\vec{r}(1)$;
- $\frac{1}{2} \sum_i^N \sum_j^N (\mathcal{J}_{ij} - \mathcal{K}_{ij})$ is Coulombian and Exchange energies.

We are interested in the single Slater determinant that represents that best possible approximation of the ground state of the system described by the Hamiltonian $\hat{\mathcal{H}}$. The spin-orbitals must be chosen so that the determinant obtained can be the state of minimum energy (variational principle). To calculate the energy of a single Slater determinant we proceed to a numerical resolution of the Schrödinger equation through the self consistent field method (SCF method) originally formulated by D. R. Hartree and later perfected by V. Fock and J. C. Slater with the introduction of the electron exchange effect[45][46][47]:

Within this approximation the Coulombic term becomes:

$$\mathcal{J}_{ij} = \int \int \chi_i^*(1) \chi_j^*(2) \frac{1}{r_{12}} \chi_j(2) \chi_i(1) d\vec{r}(1) d\vec{r}(2) \quad (3.1.10)$$

For the exchange part it has been followed an analogues procedure. The substantial difference is that this term does not have an interpretation in terms of classical physics, but it is a consequence of the antisymmetric properties of the wave function

$$\mathcal{K}_{ij} = \int \int \chi_i^*(1)\chi_j^*(2) \frac{1}{r_{12}} \chi_i(2)\chi_j(1) d\vec{r}_{(1)} d\vec{r}_{(2)} \quad (3.1.11)$$

The energy of the system has a functional dependence on the wave function of the system. The variational principle (the minimization of the energy of 3.1.9 with respect of the spin-orbitals) simplifies the original problem of searching a wave function of N electrons to N monoelectronic equations

$$\hat{f}(1) |\chi_i(1)\rangle = \epsilon_i |\chi_i(1)\rangle \quad (3.1.12)$$

where the Fock operator \hat{f} is defined as

$$\hat{f}(1) = \hat{h}(1) + \sum_j^N \hat{\mathcal{J}}_j(1) - \sum_j^N \hat{\mathcal{K}}_j(1) \quad (3.1.13)$$

The algebraic resolution pass through a series expansion of the spatial part of the spin-orbitals using a basis set of M basis functions $\{\phi_M\}$:

$$\chi_i = \sum_{u=1}^M \mathbb{C}_{ui} \phi_u = \mathbb{C}_{1i} \phi_1 + \mathbb{C}_{2i} \phi_2 + \mathbb{C}_{3i} \phi_3 + \dots + \mathbb{C}_{Mi} \phi_M \quad (3.1.14)$$

The set of Roothan equation is obtained substituting the series expansion 3.1.14 in 3.1.12, multiplying by each of the M function $\phi_v^*(1)$ and integrating

$$\sum_{u=1}^M \mathbb{C}_{ui} \langle \phi_v(1) | \hat{f}_i(1) | \phi_u(1) \rangle = \epsilon_i \sum_{u=1}^M \mathbb{C}_{ui} \langle \phi_v(1) | \phi_u(1) \rangle \quad (3.1.15)$$

The set of equation obtained at the end of the procedure can be written in a more compact form:

$$\mathbb{F}\mathbb{C} = \mathbb{S}\mathbb{C}\epsilon \quad (3.1.16)$$

where:

- matrix \mathbb{F} is the sum of the one-electron and two-electrons terms $\mathbb{F}_{vu} = \langle \phi_v(1) | \hat{f} | \phi_u(1) \rangle$;
- matrix \mathbb{S} is the superposition matrix $\mathbb{S}_{vu} = \langle \phi_v | \phi_u \rangle$.

The charge density is expressed as:

$$\rho(\vec{r}) = \sum_i^N |\chi_i(\vec{r})|^2 \quad (3.1.17)$$

and for a closed-shell system it becomes:

$$\rho(\vec{r}) = 2 \sum_i^{N/2} |\psi_i(\vec{r})|^2 \quad (3.1.18)$$

To express the density matrix \mathcal{P} the series expansion 3.1.14 is substituted in 3.1.18:

$$\mathcal{P} = \sum_u^M \sum_v^M \left[2 \sum_i^{N/2} \mathbb{C}_{ui} \mathbb{C}_{vi}^* \right] \phi_u \phi_v^* = \sum_u^M \sum_v^M [\mathcal{P}_{uv}] \phi_u \phi_v^* \quad (3.1.19)$$

With the formalism of the density matrix we can have a better understanding of the element of the \mathbb{F} matrix:

$$\mathbb{F}_{lm} = \langle \phi_l | \hat{h} | \phi_m \rangle + \sum_u^N \sum_v^N \mathcal{P}_{uv} \mathbb{G}_{lmuv} \quad (3.1.20)$$

where:

$$\begin{aligned} \mathbb{G}_{lmuv} = & 2 \int \int \phi_l^*(1) \phi_u^*(2) \frac{1}{r_{12}} \phi_v(2) \phi_v(2) \phi_m(1) d\vec{r}_{(1)} d\vec{r}_{(2)} + \\ & + 2 \int \int \phi_l^*(1) \phi_u^*(2) \frac{1}{r_{12}} \phi_v(2) \phi_m(2) \phi_v(1) d\vec{r}_{(1)} d\vec{r}_{(2)} \end{aligned} \quad (3.1.21)$$

The first term on the right side of equation 3.1.20 is fixed over the iteration cycle procedure, while the second term on the right side of that same equation is updated at each iteration because at each iteration is updated the density matrix \mathcal{P} . At the end of the iterative procedure we obtain M spatial orbitals each one with its relative orbital energy ϵ_a . The first $N/2$ spatial orbitals will be used to build a single Slater Determinant called $|\psi_0^{HF}\rangle$. To obtain the Hartree Fock energy, E_0^{HF} , the Hamiltonian of the system $\hat{\mathcal{H}}$ operates on the determinant $|\psi_0^{HF}\rangle$:

$$\hat{\mathcal{H}} |\psi_0^{HF}\rangle = E_0^{HF} |\psi_0^{HF}\rangle \quad (3.1.22)$$

$$E_0^{HF} = 2 \sum_a^{N/2} \epsilon_a - \sum_a^{N/2} \sum_b^{N/2} (2\mathcal{J}_{ab} - \mathcal{K}_{ab}) \quad (3.1.23)$$

3.1.3 DFT Methods

Density functional theory is based on the two theorems of Hohenberg and Kohn and on the assumption that the electronic energy of the ground state is totally determined by the density of charge [48]. The approach of the density of charge theory with respect to that one of the wave function is sensible when the complexity of the system increases. The density of charge depends on three coordinates only and this is independent from the complexity of the system. Even if the correspondence between density of charge of a system and its energy exists, the functional form that connects the two entities it is not known. If we assume to know such a functional form, the DFT set of equations are similar to the Hartree-Fock method. In particular the KS equations to be solved are similar to 3.1.12:

$$\left\{ -\frac{1}{2}\nabla_1^2 - \sum_Q^{Nuclei} Z_Q \frac{e^2}{r_{Q1}} + \int \frac{\rho(\vec{r}_2)}{r_{12}} d(\vec{r}_2) + V_{\chi C}(\vec{r}_1) \right\} |\varphi_i^{KS}(\vec{r}_1)\rangle = \epsilon_i |\varphi_i^{KS}(\vec{r}_1)\rangle \quad (3.1.24)$$

where:

- The first two terms of the left member are an analogue of the one-electron \hat{h} of the HF method;
- The term $\int \frac{\rho(\vec{r}_2)}{r_{12}} d(\vec{r}_2)$ represents the Coulomb interaction;
- The term $V_{\chi C}$ is the exchange-correlation potential, the functional derivative of the exchange-correlation energy $E_{\chi C}$;
- $|\varphi_i^{KS}\rangle$ is a Kohn-Sham orbital, its relative orbital energy is ϵ_i .

The equivalent of the density of charge 3.1.18 is:

$$\rho(\vec{r}) = 2 \sum_i^{N/2} |\varphi_i^{KS}(\vec{r})|^2 \quad (3.1.25)$$

The of KS equations are solved in a SCF fashion analogously to 3.1.15. To start the calculation it is needed an initial trial function to represent the density of charge, ρ_{guess} and an expression for the term $E_{\chi C}$ of which the first derivative with respect to ρ_{guess} will give the potential of exchange correlation $V_{\chi C}$. The set of equations is then solved to obtain a new set $\{\varphi_i^{KS}\}$ of KS orbitals that allows the calculation of a new density of charge $\rho'(\vec{r})$. The iterative process continues until density of charge and exchange-correlation energy are within a certain threshold. KS orbitals can be written as a series expansion over a certain basis-set, so the resolution of the equations 3.1.24 gives the value of the coefficients of such expansion. Once the final density of charge, $\rho^{final}(\vec{r})$ has been calculated, the energy of such state is calculated by:

$$\begin{aligned} E[\rho] = & -\frac{1}{2} \sum_i^n \int \varphi_i^*(\vec{r}_1) \nabla^2 \varphi_i^*(\vec{r}_1) d(\vec{r}_1) - \sum_Q^{Nuclei} \int \frac{Z_Q \rho(\vec{r}_1)}{|\vec{r}_{Q1}|} d(\vec{r}_1) + \\ & + \frac{1}{2} \int \frac{\rho(\vec{r}_1)\rho(\vec{r}_2)}{|\vec{r}_{12}|} d(\vec{r}_1) d(\vec{r}_2) + E_{\chi C} \end{aligned} \quad (3.1.26)$$

The choice of the right form of the term $E_{\chi C}$ is crucial for the accuracy of the calculation. Within this Ph.D. thesis it has been chosen to adopt the general purpose functional $E_{\chi C}^{B3LYP}$

$$E_{\chi C}^{B3LYP} = A E_{\chi}^{Dirac} + (1 - A) E_{\chi C}^{HF} + B E_{\chi}^{Becke} + (1 - C) E_C^{VWN} + C E_C^{LYP} \quad (3.1.27)$$

where:

- A = 0.2, B=0.72, and C=0.19 are fitted parameters;
- The local functional $E_{\chi}^{Dirac}[\rho] = -C_x \int \rho^{\frac{4}{3}}(\vec{r}) d(\vec{r})$;
- The non-local functional $E_{\chi}^{Becke}[\rho] = \frac{1}{2} \int \rho(\vec{r}) \epsilon_{\chi}^{\beta}[\rho, \nabla\rho, \vec{r}] d(\vec{r})$,
and $\epsilon_{\chi}^{\beta}[\rho, \nabla\rho, \vec{r}] = -\beta\rho^{\frac{1}{3}} \frac{|\nabla\rho|^2}{\rho^{8/3}(1+6\beta\sinh^{-1}x)}$ where the parameter β is obtained by fitting over 6 noble gases.
- E_C^{LYP} is the correlation functional introduced by Lee, Yang, and Parr [49];
- E_C^{VWN} is the correlation functional introduced by Vosko, Wilk, and Nusair [50].

and a functional especially designed for protein application MO62X [51]. The DFT series of MO6 functionals are tailored to have a better description of vdW interactions.

3.1.4 Møller-Plesset Theory

Perturbation theory starts introducing a correction $\hat{\mathcal{V}}$ to an Hamiltonian operator $\hat{\mathcal{H}}_0$. The correction $\hat{\mathcal{V}}$ is called the perturbative correction, while the operator $\hat{\mathcal{H}}_0$ is the unperturbed operator. The addition between these two terms gives the total Hamiltonian of the chemical system:

$$\hat{\mathcal{H}} = \hat{\mathcal{H}}_0 + \lambda\hat{\mathcal{V}} \quad (3.1.28)$$

The sufficient condition to apply a perturbative correction is the asymptotic convergence of the following two series:

$$|\psi_j\rangle = |\psi_j^{(0)}\rangle + \lambda |\psi_j^{(1)}\rangle + \lambda^2 |\psi_j^{(2)}\rangle + \dots + \lambda^n |\psi_j^{(n)}\rangle \quad (3.1.29)$$

$$E_j = E_j^{(0)} + \lambda E_j^{(1)} + \lambda^2 E_j^{(2)} + \dots + \lambda^n E_j^{(n)} \quad (3.1.30)$$

A set of non-homogeneous equations is obtained introducing 3.1.29 and 3.1.30 in $\hat{\mathcal{H}}|\psi_j\rangle = E_j|\psi_j\rangle$ and separating the various orders. It follows the explicit separation until IV order:

	order
$(\hat{\mathcal{H}}_0 - E_j^{(0)}) \psi_j^{(1)}\rangle = E_j^{(0)} \psi_j^{(0)}\rangle$	I
$(\hat{\mathcal{H}}_0 - E_j^{(0)}) \psi_j^{(2)}\rangle = (E_j^{(1)} - \hat{\mathcal{V}}) \psi_j^{(0)}\rangle$	II
$(\hat{\mathcal{H}}_0 - E_j^{(0)}) \psi_j^{(3)}\rangle = (E_j^{(1)} - \hat{\mathcal{V}}) \psi_j^{(1)}\rangle + E_j^{(2)} \psi_j^{(2)}\rangle$	III
$(\hat{\mathcal{H}}_0 - E_j^{(0)}) \psi_j^{(4)}\rangle = (E_j^{(1)} - \hat{\mathcal{V}}) \psi_j^{(2)}\rangle + E_j^{(2)} \psi_j^{(1)}\rangle + E_j^{(3)} \psi_j^{(0)}\rangle$	IV
(.....)	

(3.1.31)

In the particular case of the MP2 method [citazione], the unperturbed Hamiltonina if the Fock operator:

$$\hat{\mathcal{H}}_0 \equiv \hat{F} \quad (3.1.32)$$

so the perturbative correction is the difference between the Hamiltonian of the system and the Fock operator:

$$\hat{\mathcal{V}} = \hat{\mathcal{H}} - \hat{F} \quad (3.1.33)$$

The wave function ψ_{HF} is eigenfunction of the \hat{F} operator, its relative eigenvalue $E^{(0)}$ is given by the summation of all the orbital energy of the occupied

spin-orbitals; the Hartree-Fock energy of the ground state $|\psi_0\rangle$ is instead the expectation value $\langle\psi_{HF}|\hat{\mathcal{H}}|\psi_{HF}\rangle$. If the complete set of eigenfunctions $\{\phi_k^{(0)}\}$ of the operator \hat{F} is known, a linear combination of them, i.e. $|\psi_j^{(0)}\rangle$, remains an eigenfunction of such operator. Using the ground state function $|\psi_0\rangle$ as 0 order term of the perturbative expansion:

$$|\psi_0\rangle = |\psi_0^{(0)}\rangle + \lambda|\psi_0^{(1)}\rangle + \lambda^2|\psi_0^{(2)}\rangle + \dots \quad (3.1.34)$$

Expressing the first order term $|\psi_0^{(1)}\rangle$ as a linear combination of $\{\phi_k^{(0)}\}$:

$$|\psi_0^{(1)}\rangle = \sum_{k \neq 0} c_k |\phi_k^{(0)}\rangle \quad (3.1.35)$$

The expression of the coefficients c_k of 3.1.35 of the unperturbed basis set is obtained substituting 3.1.35 in the II order differential equation of 3.1.31 and multiplying by an orthogonal bra function $\langle\phi_l^{(0)}| \in \{\phi_k\}$:

$$|\psi_0^{(1)}\rangle = \sum_{k \neq 0} \left[\frac{\langle\phi_k^{(0)}|\hat{V}|\phi_0^{(0)}\rangle}{\langle\phi_0^{(0)}|\hat{\mathcal{H}}|\phi_0^{(0)}\rangle - \langle\phi_k^{(0)}|\hat{\mathcal{H}}|\phi_k^{(0)}\rangle} \right] |\phi_k^{(0)}\rangle = \sum_{k \neq 0} \left[\frac{\mathcal{V}_{k0}}{\mathcal{H}_{00}^{(0)} - \mathcal{H}_{kk}^{(0)}} \right] |\phi_k^{(0)}\rangle \quad (3.1.36)$$

With such a term it is possible to express the perturbative correction of the III order:

$$E^{(3)} = \sum_{k \neq 0} \sum_{L \neq 0} \left[\frac{\mathcal{V}_{0k}\mathcal{V}_{kL}\mathcal{V}_{L0}}{(\mathcal{H}_{00}^{(0)} - \mathcal{H}_{kk}^{(0)}) (\mathcal{H}_{00}^{(0)} - \mathcal{H}_{LL}^{(0)})} \right] - \sum_{k \neq 0} \left[\frac{\mathcal{V}_{0k}\mathcal{V}_{00}\mathcal{V}_{k0}}{(\mathcal{H}_{00}^{(0)} - \mathcal{H}_{kk}^{(0)})^2} \right] \quad (3.1.37)$$

The Møller-Plesset method has a one-electron part equals to the Hartree-Fock method, while the description of the two-electrons part is more accurate

because of the perturbative term. When the expansion 3.1.36 is truncated at the II order we have the MP2 method[46].

$$E^{(2)} = \sum_{k \neq 0} \frac{\mathcal{V}_{0k} \mathcal{V}_{k0}}{\left(\mathcal{H}_{00}^{(0)} - \mathcal{H}_{kk}^{(0)} \right)} \quad (3.1.38)$$

The unperturbed wave function is the Hartree-Fock ground state. The II order perturbative energy correction are excited Slater Determinants. Since in MP2 the operator $\hat{\mathcal{V}}$ is two-electrons, all the matrix elements with triple excitation etc. are null integrals. The first excitations do not correct the HF energy (Brillouin theorem), this means that in MP2 the responsible for the subtractive correction term in 3.1.38 are only the double excitations.

3.2 Localization NBO Methods and NBO Steric Analysis

The NBO decomposition of the wave function allows the partition of the wave function into localized bond and lone pair orbitals. One of the results of an ab initio HF calculation is the shape of the charge density, equation 3.2.1. The charge density is the probability of finding an electron in an infinitesimal volume centered on point \vec{r} . Following what previously described for the HF method in section 3.1.2, if the set of orbitals is expanded on a set of basis function $\{\phi(\vec{r})\}$, the density of charge assumes the following form:

$$\rho(\vec{r}) = \sum_u^M \sum_v^M |\mathcal{P}_{uv}| \phi_u(\vec{r}) \phi_v^*(\vec{r}) \quad (3.2.1)$$

where:

- \mathcal{P} is the density matrix defined in equation 3.1.19.

A common approach to gain some insight from the charge density is to visualize it on a topographic map. Apart from the simple graphical visualization of 3.2.1. Among properties obtainable from the charge density, i.e. dipolar moments or quadrupolar moments, it is of particular advantage to the chemical investigation the possibility of assigning charges to atoms or group of atoms. In this way the ordinary pictorial interpretation in terms of functional groups can be obtained directly from the wave function. This technique is the so called population analysis. We give here a brief introduction of Natural Bond Orbital (NBO) analysis as a localization method able to give an estimate of the steric energy of interaction. In a typical NBO analysis the density matrix is analyzed to obtain:

- the shape of atomic orbitals in a particular molecular environment;
- to calculate the charge density between atoms, so to derive chemical bonds.

The matrix \mathcal{P} is considered as formed by different blocks, each one relative to a basis set that belongs to a specific atom:

$$\mathcal{P} = \begin{vmatrix} |\mathcal{P}_{AA}| & |\mathcal{P}_{AB}| & |\mathcal{P}_{AC}| & \cdots \\ |\mathcal{P}_{BA}| & |\mathcal{P}_{BB}| & |\mathcal{P}_{BC}| & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix}$$

Each of the blocks are then diagonalized. The natural atomic orbitals (NAOs) is a set of orbitals obtainable at the end of the process of diagonalization of the various block of the density matrix and after a procedure of orthogonalization of the orbitals weighted by occupancy. For a more exhaustive treatment refer to <http://www.chem.wisc.edu/~nbo5/>.

The density matrix diagonal terms together with a set of NAOs are the population contribution indicated in what follows by \mathcal{POP} . A summation of population contribution of all NAO centered on a certain atomic center, let's say atom A, needs to be done to obtain the charge on that atom

$$\text{charge on } A = \sum_i^{\text{NAOs on } A} \mathcal{POP}_i \quad (3.2.2)$$

Off diagonal blocks represent instead chemical bonds between atoms. To gain insight of the bond orbitals it is needed to distinguish and to eliminate all the terms that do not represent a bond, this is done on the basis of the occupancy of the NAOs involved. After the process of elimination, the off diagonal blocks are diagonalized to obtain natural bond orbitals (NBOs), and their relative eigenvectors of the diagonalization matrix. The eigenvalues represent the occupancy of the NBOs. The orthonormal set of localized high occupancy lewis-like $1 - c$ orbital (lone pair), $2 - c$ bond orbitals ($\bar{\sigma}_i$) as well as residual set of low occupancy non-lewis orbitals valence antibond and rydberg orbitals ($\bar{\sigma}_i^*$).

To evaluate the steric interaction we follow the NBO partitioning of exchange interactions developed by Weinhold and coworkers [52]. Steric repulsion can be associated with the Pauli exclusion principle and the idea of pressure, equation 3.1.6, exerted by particles confined into potential barriers. The antisymmetrization of the wave function leads to an increased of the variational HF energy compare to the non antisymmetrized counterpart.

The antisymmetrization results in the exchange integrals of equation 3.1.11 that is associated with steric interactions, so an antisymmetrized hartree product, $\bar{\Phi}_H$ is raised in energy with respect to the analogous non-antisymmetrized hartree product Φ_H :

$$\langle \bar{E}_H \rangle = \frac{\langle \bar{\Phi}_H | \hat{\mathcal{H}} | \bar{\Phi}_H \rangle}{\langle \bar{\Phi}_H | \bar{\Phi}_H \rangle} > \langle E_H \rangle = \frac{\langle \Phi_H | \hat{\mathcal{H}} | \Phi_H \rangle}{\langle \Phi_H | \Phi_H \rangle} \quad (3.2.3)$$

quantifying the exchange energy as

$$\langle E_{exchange} \rangle = \langle \bar{E}_H \rangle - \langle E_H \rangle \quad (3.2.4)$$

Since the energy difference between the slater determinant $|\psi_0^{HF}\rangle$ and $\bar{\Phi}_H$ is rather small[53]

$$\langle E_0^{HF} \rangle \simeq \langle \bar{E}_H \rangle \quad (3.2.5)$$

this means that the energy difference in equation 3.2.4 is mainly due to orthogonalization rather than the effect of the solely anti-symmetrization so that:

$$\langle E_{orthog} \rangle = \langle \bar{E}_H \rangle - \langle E_H \rangle \simeq \langle E_{exchange} \rangle \quad (3.2.6)$$

and the exchange interaction can be approximated as the energy difference between orthogonal and non-orthogonal orbitals.

The two matrices of the energy of the Roothan equations (see equation 3.1.16) obtained using $\bar{\Phi}_H$ and Φ_H are respectively $\bar{\epsilon}$ and ϵ , and the difference between the summation of their diagonal terms is approximated as the exchange energy:

$$E_{exchange} \simeq \sum_i^n \bar{\epsilon}_i - \epsilon_i \quad (3.2.7)$$

Steric interaction is here approximated as an exchange effect, the Hartree-Fock (HF) method gives a reasonable accurate description of such interaction[54][55].

As NBOs ($\bar{\sigma}_i$ and $\bar{\sigma}_i^*$) are linear combination of the set of orthonormal NAOs, we can derive a set of non-orthogonal Pre-NBOs (σ_i and σ_i^*) that are linear combination of non-orthogonal set of Pre-NAOs. It is using the two set of NBOs and Pre-NBOs that we are going to calculate the NBO estimate of the steric energy:

$$E_{exchange}^{NBO} = \sum_i^n \bar{\epsilon}_i^{NBO} - \epsilon_i^{Pre-NBO} \quad (3.2.8)$$

The localized NBO description allows one to compare steric repulsions between individual groups, making contact with the intuitive picture in a qualitatively satisfactory manner.

3.3 QM-MM Methods

3.3.1 The QM-MM Energy Expression.

The QM-MM formalism can in principle accommodate any combination of QM and MM methods. The QM method must satisfy the condition of being a Self Consistent Field (SCF) method in presence of a point-charge field created by MM charges. The elected method are post-Hartree Fock (post-HF) ab initio methods, such as Møller-Plesset perturbation and Coupled Cluster methods. The superior accuracy of the post-HF methods due to the inclusion of electron correlation and the recent hardware development extended the size of the system to be treated with such methods are appealing for energy calculations with a fixed geometry. In practice biomolecular systems are instead treated using DTF theory and the many developed DFT functionals. The QM-MM potential is given by

$$U_{tot} = \langle \psi(\vec{\mathbf{r}}) | \mathcal{H}_{QM}^0 + \mathcal{H}_{QM/MM} | \psi(\vec{\mathbf{r}}) \rangle + U_{MM} \quad (3.3.1)$$

where:

- \mathcal{H}_{QM}^0 is the Hamiltonian of the QM subsystem of coordinates $\vec{\mathbf{R}}$;
- U_{MM} is the MM potential energy of the rest of the system;
- $\mathcal{H}_{QM/MM}$ is the interaction Hamiltonian between the two regions.

The wave function $\psi(\vec{\mathbf{r}})$ of the QM system at a given position $\vec{\mathbf{R}}$ of the nuclei is polarized by the perturbation Hamiltonian $\mathcal{H}_{QM/MM}$. After integration it is possible to rewrite 3.3.1 as

$$U_{tot} = E_{QM}^0 + \Delta E_{QM/MM} + U_{MM} \quad (3.3.2)$$

where:

- $E_{QM}^0 = \langle \psi(\vec{\mathbf{r}}) | \mathcal{H}_{QM}^0 | \psi(\vec{\mathbf{r}}) \rangle$ is the energy of the QM isolate system;
- $\Delta E_{QM/MM}$ is the interaction energy of equation 3.3.1.

The interaction energy $\Delta E_{QM/MM}$ is the energy required to transfer the QM system from the gas phase into, e.g., the active site of the enzyme. It is like a solvation energy term where the solvent is in this case represented by the protein environment

$$\Delta E_{QM/MM} = \langle \psi(\vec{\mathbf{r}}) | \mathcal{H}_{QM}^0 + \mathcal{H}_{QM/MM} | \psi(\vec{\mathbf{r}}) \rangle - E_{QM}^0 \quad (3.3.3)$$

The MM terms include the normal bonding terms for bond stretching, bond angle bending, and torsions, plus non bonded terms for Coulomb and van der Waals interactions:

$$\begin{aligned} E_{MM} = & \sum_i^{stretches} K_i (r_i - r_0)^2 + \sum_j^{bends} K_j (\theta_j - \theta_0)^2 \\ & + \sum_k^{torsions} \left[\frac{V_{k,1}}{2} (1 + \cos\phi_k) + \right. \\ & \left. + \frac{V_{k,2}}{2} (1 - \cos 2\phi_k) + \frac{V_{k,3}}{2} (1 + \cos 3\phi_k) \right] + \\ & + \sum_{m>n}^{MMatoms} \frac{q_m \cdot q_n}{r_{mn}} + \sum_{m>n}^{MMatoms} 4\epsilon_{mn} \left[\left(\frac{\sigma_{mn}}{r_{mn}} \right)^{12} - \left(\frac{\sigma_{mn}}{r_{mn}} \right)^6 \right] \end{aligned} \quad (3.3.4)$$

In this Ph.D. project was used Qsite software (Schrödinger, LLC, New York, NY, 2011), a QM-MM algorithm that will be described briefly. The interactions

between the QM and MM regions include electrostatic terms between the QM nuclei and electrons, van der Waals terms between MM atoms and QM atoms, and MM-like terms at the boundary of QM and MM regions that account as corrections for the fact that the electrostatic and van der Waals terms in these region are not sufficient to represent differences in energy when the atoms at the boundary move. The first of these correction terms is to prevent short range electrostatic interactions. In practice it neglects electrostatic interactions for atoms connected through 1 bond (1-2 interactions) and two bonds (1-3 interactions). It scales with a factor of 0.5 the 1-4 electrostatic interaction terms. The form of the electrostatic correction is

$$E_{\text{electrostatic correction}} = - \sum_n^{MM_{\text{atoms}}} \sum_i^{QM_{\text{atoms}}} \sigma_{mi} \frac{q_m q_i}{r_{mi}} \quad (3.3.5)$$

where:

- q_m are MM charge assigned through the force-field;
- q_i is instead an MM-like partial charge assigned to the i -th atom that belongs to the QM part;

$$\bullet \sigma_{mn} = \begin{cases} 1 & \text{for interactions } 1-2 \text{ or } 1-3 \\ 0.5 & \text{for interactions } 1-4 \\ 0 & \text{for interactions } \geq 1-5 \end{cases}$$

Additional correction terms are for bond terms, stretch terms, bend terms, and torsion terms. Parameters for correction terms are calculated fitting QM datas. In figure 3.3.1 we have an example of truncation between MM and QM regions. We use the previously mentioned figured to explain the stretching correction term. Bend and torsion terms will follow. The QM system is here hydrogen

capped (further info in the following Link Atom section). The the four bonds around the QM atom at the boundary are the four stretching terms that will undergo to the stretching correction.

Stretching correction term is

$$E_{str. corr.} = \sum_i K_i (r_i - r_{i0})^2 \quad (3.3.6)$$

where:

- K_i, r_{i0} are fit parameters.

Fit parameters are obtained fitting the following equation against QM data:

$$E'_{str. corr.} = \sum_i \left[\alpha_i (r_i - r_0) + K'_i (r_i - r_0)^2 \right] \quad (3.3.7)$$

where:

- r_0 is the reference bond length;
- α_i is a fit parameter linked to K_i and r_{i0} of eq.3.3.6 through $\alpha_i = -2K_i (r_{i0} - r_0)$;
- K'_i is a fit parameter and it is equal to K_i fit parameter of eq.3.3.6.

An analogous couple of equations account for the bending correction terms, while for the torsion terms the parameterization proceeds by performing a weighted linear least-squares fit using singular value decomposition. The torsional expression is

$$E_{tor corr.} = \sum_K \left[\frac{V_{k,1}}{2} (1 + \cos\phi_k) + \frac{V_{k,2}}{2} (1 + \cos 2\phi_k) + \frac{V_{k,3}}{2} (1 + \cos 3\phi_k) \right] \quad (3.3.8)$$

where:

- $V_{k,1}, V_{k,2},$ and $V_{k,3}$ are fit parameters.

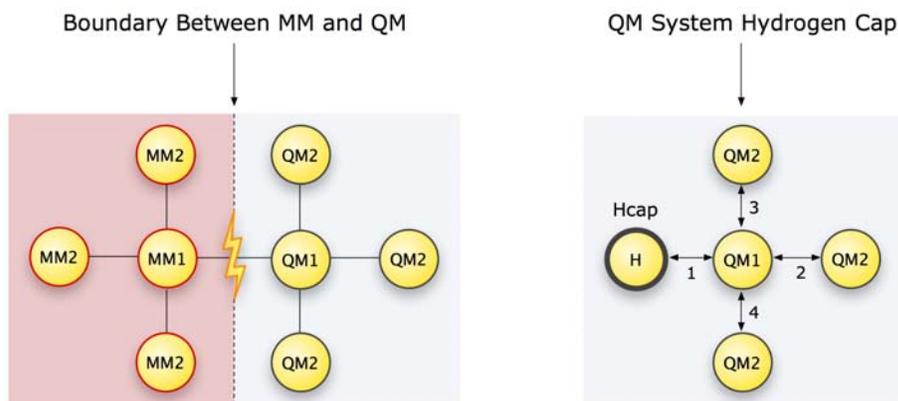


Figure 3.3.1: Boundary between QM and MM regions and the relative hydrogen capped QM system. The four stretching internal coordinates are numbered and graphically represented in the right hand picture. The four stretchings the QM atoms at the boundary, QM_1 , will have stretching correction terms.

The master equation for the total energy can be presented as follows:

$$\begin{aligned}
 E_{Qsite}^{tot} = & \left[\sum_{\mu\nu} P_{\mu\nu} H_{\mu\nu}^{core} + \frac{1}{2} \sum_{\mu\nu} P_{\mu\nu} (2J_{\mu\nu} - K_{\mu\nu}) \right] + \\
 & + E_{nuc}^{QM} + E_{MM} + E_{coulomb}^{QM/MM} + E_{vdW}^{QM/MM} + \\
 & + (correction\ charges) + (correction\ stretches, bendings, torsions)
 \end{aligned}
 \tag{3.3.9}$$

It is necessary to spend some word on the boundary schemes and the relative arising problems. Apart from studies where MM system is represented exclusively by solvent molecules, in enzymology it is quite unavoidable to have QM - MM boundary cuts through a covalent bond. The issues arising are i) the need of introducing a cap because it would not be realistic to treat the system as it would be homolytically or heterolytically cleaved, ii) link atoms introduces an overpolarization of the QM density from the MM charges.

There are fundamentally three kinds of truncation:

- link atom scheme: this is the easiest for what concerns the implementation, it simply caps the QM and MM systems with a heteroatom (usually hydrogens) not present in the real system. The so called link atoms saturate the free valency upon the truncation site;
- boundary-atom scheme: a special boundary atom replaces an MM atom. This atom is present in both MM and QM part. On the QM side it mimics the cut bond and the electrostatic moiety attached to QM1. In the MM part it behaves instead as a normal MM atom;
- frozen orbital scheme: hybrid orbitals are placed at the boundary. All the frozen orbitals cap the QM region replacing the cut bond.

In picture 3.3.2 are graphically visualized the three possible boundary truncation schemes explained before.

There should be a right balance between computational effort and minimization of possible artifacts due to the choice of the boundary truncation site. Usually it is the QM region that contains the interesting part of the system where reactions take place and where properties are observed, i.e. in enzymology the QM region is usually the active site of the enzyme. A good choice would be to move the boundary truncation far away from this region of interest. This would lead to bigger size QM parts so to reduce the artifacts, but all this is at the expense of a bigger computational effort. Unpolar bonds not involved in any reaction and not having conjugative interactions are the elective site for truncation. Also a cut through a MM charge groups must be avoided because of issues of strong overpolarization of QM charge density. We will give more de-

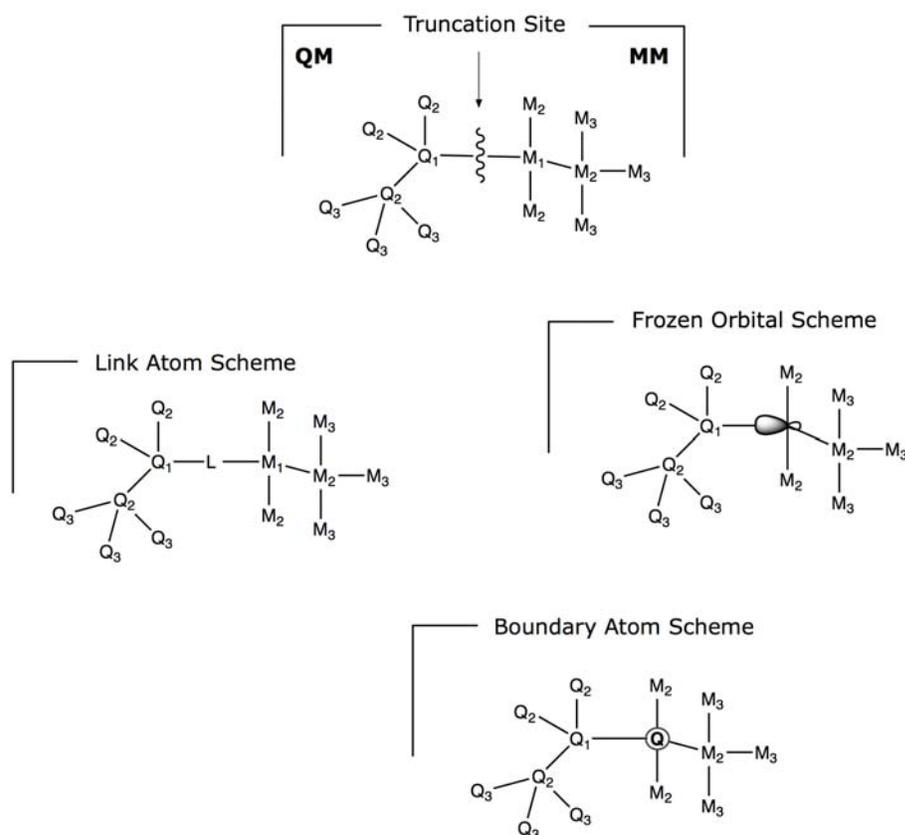


Figure 3.3.2: Truncation Site and relative Boundary Schemes generally adopted in QM-MM techniques.

tails of the link atom scheme only because it has been used in this PhD project when QM-MM has been adopted (chapter 5).

Link Atom Scheme

It is the oldest coupling scheme adopted in QM-MM [56][57]. The introduction of a link atom, usually a hydrogen, creates the following issues: i) three additional degrees of freedom not present in the original system, ii) the introduced

link atom near the MM frontier atom at the boundary will get overpolarization by it, iii) the link atom is different from the group it is substituted with, both chemically and electronically. Of particular importance is the issue of the overpolarization. Atoms that are 0.5\AA away from the boundary will tend to be overpolarized by the rigid MM point charges. The more the QM charge density is polarizable and the more the overpolarization will cause errors. To reduce such an error it is convenient to not introduce polarization and diffuse functions in the basis-set. Polarization and diffuse function will cause the QM charge density of being more elastic and prone to overpolarization [58]. To overcome the problem in Qsite software it has been implemented the so called “smearing” function for the MM point charges near the boundary. The charges close to the QM region are re-distributed on a grid around the QM-MM truncation site with a Gaussian distributions manner so to represent the potential of the atoms near the cap, the MM ordinary point charges are instead used for the rest of the MM region.

Chapter 4

Development of ReaxFF for Enzyme Catalysis

The vision below this PhD project is to seek and improve the understanding of enzyme catalysis with atomistic details. Currently the theory able to describe chemical systems and their reactivity is quantum mechanics (QM) and its relative QM methods that uses approximation of the theory for the description of the molecular structure. To the present day modeling enzyme reactions is inaccessible to QM methods because the size of the problem would result in many-particle equations too complicated to be solved even with rather crude approximations such as Hartree-Fock (HF). Classical models such as the ordinary molecular mechanics (MM) force-fields use newtonian mechanics to describe molecular systems. At this level it is possible to include the entire enzyme system and still have light equations. On one hand we have accurate QM methods able to describe reactivity but limited in the size of the system to describe, while on the other hand we have molecular mechanics and ordinary force-fields that

are virtually unlimited in size but unable to straightforwardly model chemical reactivity. A reactive force-field (ReaxFF) is a provisional simplified model that bridges the gap between quantum chemical domain to the ordinary force-fields of the molecular mechanics domain. ReaxFF method is fast as MM methods, but at the same able to model chemical reactions as a QM method. In figure 4.0.1 we see an illustration with MM domain at the left-top and QM domain at the right bottom and the gap between the two methods that should be bridged by ReaxFF.

ReaxFF was originally developed by Adri C. T. van Duin and co-workers, to describe hydrocarbon reactions[18]. It was found that the force field, was able to describe simple reactions and geometries at an accuracy similar or better than the semi-empirical PM3 method, but about 100 times faster. To describe hydrocarbons only hydrogen and carbon were included, but now the most of the periodic table has been described. To the contrary of ordinary MM force-fields, ReaxFF uses only one “atom type” per element. This means that a sp^3 , sp^2 or sp hybridised, aromatic carbons etc. are described in the same way. ReaxFF is a reactive force-field that uses the concept of bond order to model interactions within a chemical system. The anharmonicities to describe chemical reactions are included directly in ReaxFF potential terms where each term of the potential has been modified to be a function of bond order. As an example, the function to calculate bond order for a carbon-carbon interaction with distance is represented in figure 4.0.2. By making each atomic interaction bond order dependent, we can attain a dynamic description of each atomic and molecular interaction that does not depend on predefined reactivity as with ordinary MM potentials. This is done by parameterization against quantum

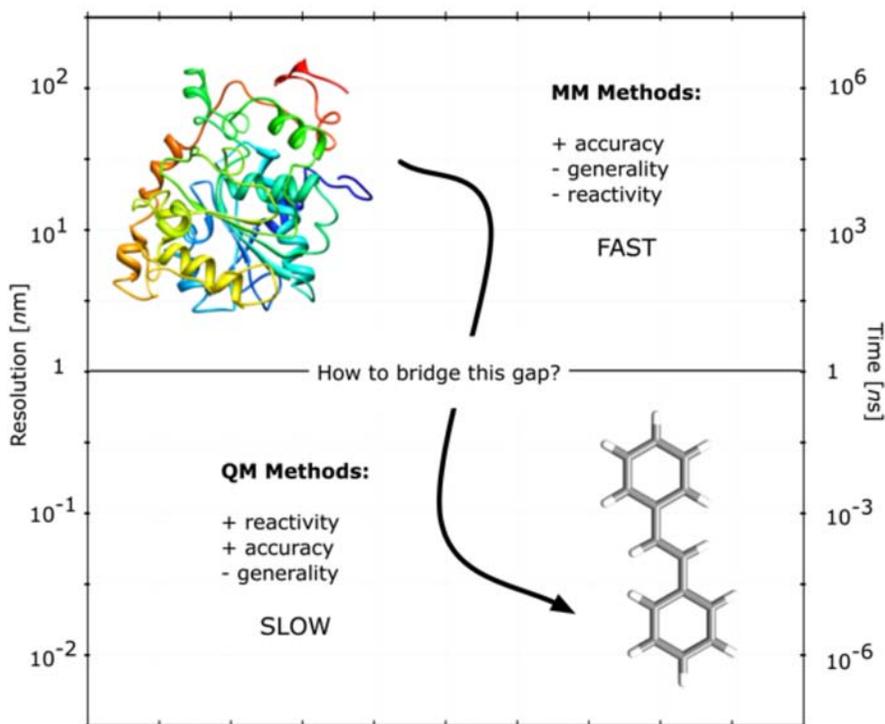


Figure 4.0.1: ReaxFF is an attempt to bridge the gap between QM domain and MM domain enabling an MM force-field to describe chemical reactions. ReaxFF method should describe chemical reaction accurately and significantly faster than QM methods.

and experimental data.

Right now the force field successfully describes chemistry that spans from hydrocarbon oxidation [59], transition metal catalysed carbon nanotube formation [60], explosives [61], and water adsorption on ZnO surfaces [62]. Applications to biomolecules include the adsorption of aminoacid on TiO₂ surfaces [63], interactions with silica surfaces [64]. ReaxFF can accurately reproduce the geometries and stabilities of several nonconjugated, conjugated and radical-containing

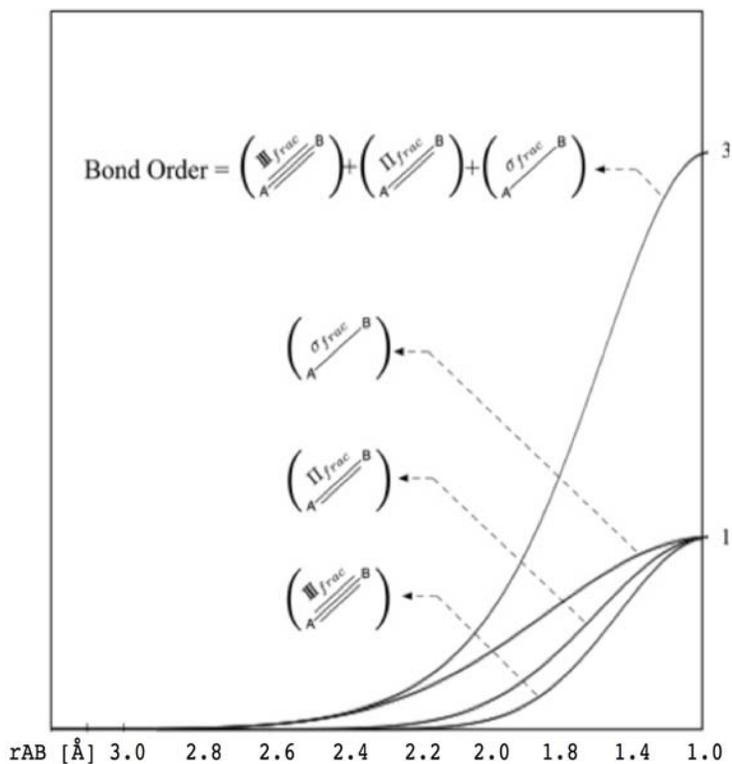


Figure 4.0.2: The bond order of a carbon-carbon interaction with respect to distance.

compounds. Several MD simulations were performed obtaining valuable insights about the complex reaction mechanism. ReaxFF correctly predicts the reaction rate and relative barrier heights. These studies are the building block to simulate complex reaction mechanism in organic and biological molecules. The first application of ReaxFF in enzyme catalysis was to study a yeast polymerase, even though the description was affected by an improper optimization of the interactions around a magnesium ion [65]. A ReaxFF force-field has lately been

developed in 2011, to have an accurate description of proton transfer in glycine in order to provide a methodology for simulating proton transfer in molecules in the aqueous phase [66]. The first application on biocatalysis is a study of the proline catalyzed aldol reaction [67], but only recently it has been developed by us a proper parameterization to treat real enzyme systems. In the following paragraphs we briefly get some insight of the machinery of ReaxFF (paragraph 4.1 and paragraph 4.2). After that we get into an overview of the force field developed to study glycine proton transfer [66] (paragraph 4.3) and then we describe how it has been developed the state of the art of Prot-ReaxFF force-fields for protein systems and enzyme catalysis (paragraph 4.4). We conclude with a study about serine protease and in particular a ProtReaxFF study of trypsin inhibition (paragraph 4.5) .

4.1 ReaxFF Potential Energy

As previously mentioned, ReaxFF uses the concept of bond order to describe how atoms interact with one another [18][68][69][70]. The total potential energy is given by the following equation:

$$E_{system} = E_{bond} + E_{over} + E_{under} + E_{lp} + E_{val} + E_{tor} + E_{vdW} + E_{coul} \quad (4.1.1)$$

A differentiation between non-bonded and covalent interactions. Contributions to the total energy are the covalent interaction energy terms for the bond (E_{bond}), the over-coordination penalty (E_{over}), under-coordination stability (E_{under}), lone-pair (E_{lp}), valence angle (E_{val}), and valence torsion (E_{tor}). The non-bonding energy contributions are instead the coulombic (E_{coul}), and

the van der Waals term (E_{vdW}). A general scheme of each of computation iteration during a ReaxFF step can be seen in figure 4.1.1.

E_{bond} term consists of several terms:

$$E_{bond} = E_{str} + E_{bend} + E_{coa} + E_{C2} + E_{triple} + E_{conj} + E_{hb} \quad (4.1.2)$$

where we have stretching energy (E_{str}), bending energy (E_{bend}), three-body energy (E_{coa}), a correction for the description of C_2 molecule (E_{C2}), triple bond energy correction (E_{triple}), stabilization term for aromatic systems (E_{conj}), hydrogen bond energy (E_{hb}).

Additional terms can be added to account for special phenomena of the system under investigation.

4.1.1 Covalent Terms

After the initial position of each atom of the system are recorded, the first step of a ReaxFF is the calculation of the bond order between each pair of atoms. An example of bond order calculation for two pairs of carbon atoms is calculated with the following equation:

$$BO'_{ij} = \exp \left[p_{bo,1} \cdot \left(\frac{r_{ij}}{r_o} \right)^{p_{bo,2}} \right] + \exp \left[p_{bo,3} \cdot \left(\frac{r_{ij}}{r_o} \right)^{p_{bo,4}} \right] + \exp \left[p_{bo,5} \cdot \left(\frac{r_{ij}}{r_o \pi} \right)^{p_{bo,6}} \right] \quad (4.1.3)$$

where:

- BO'_{ij} represents the bond order between atom i and atom j ;
- $\exp \left[p_{bo,1} \cdot \left(\frac{r_{ij}}{r_o} \right)^{p_{bo,2}} \right]$ represents the single bond contribution.

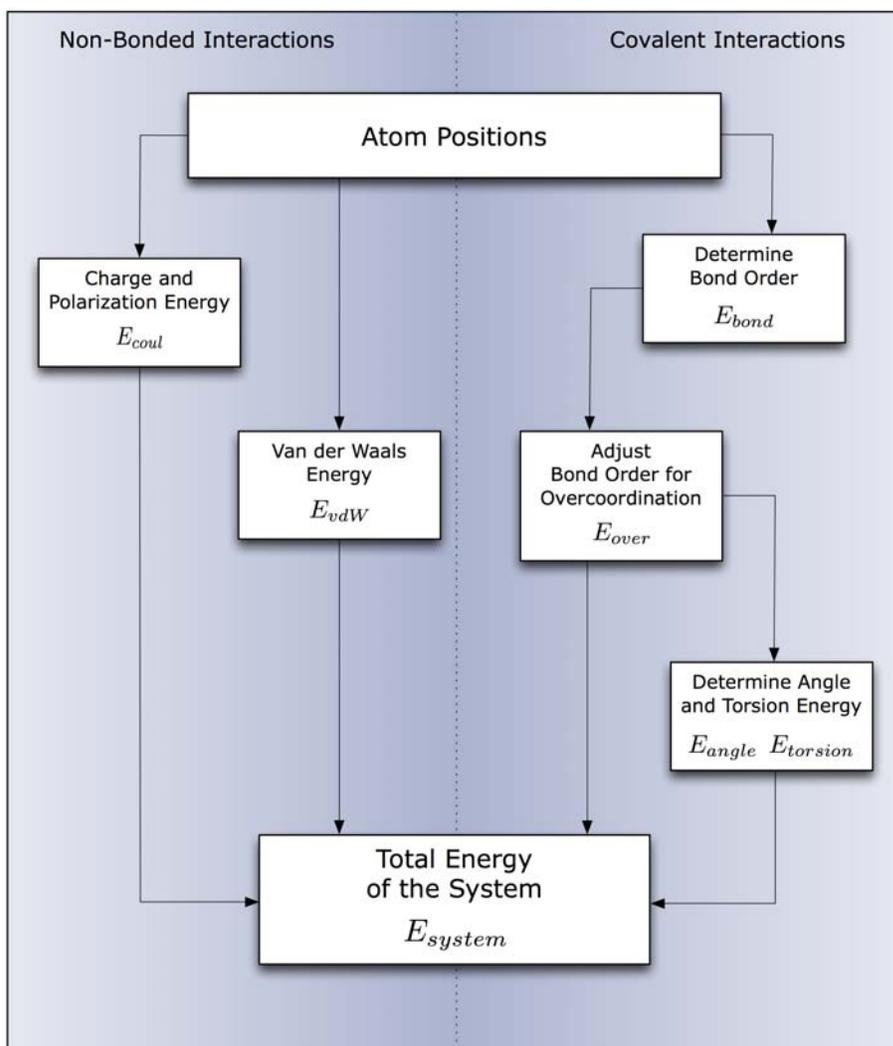


Figure 4.1.1: Scheme of a ReaxFF iteration. Non-Bonded iteration on the left and the covalent/bonded interactions are on the right.

- $p_{bo,1}, r_0^\sigma, p_{bo,2}$ are fitting parameters for the single bond contribution;
- $\exp \left[p_{bo,3} \cdot \left(\frac{r_{ij}}{r_0^\sigma} \right)^{p_{bo,4}} \right]$ represents the double bond contribution;
- $p_{bo,3}, r_0^\pi, p_{bo,4}$ are fitting parameters for double bond contribution;
- $\exp \left[p_{bo,5} \cdot \left(\frac{r_{ij}}{r_0^{\pi\pi}} \right)^{p_{bo,6}} \right]$ represents the triple bond contribution;
- $p_{bo,5}, r_0^{\pi\pi}, p_{bo,6}$ are fitting parameters for the triple bond contribution.

Each bonding parameter term, p , and each bonding equilibrium distance, r_σ , have been parameterized to have bond strengths and distances that agree with quantum mechanically obtained values. The graphical representation of equation 4.1.3 is instead shown in figure 4.0.2. From figure 4.0.2 it can be evinced the continous dependence on the distance for each of the three bond types (single, double, and triple), plus the continuum transition of the total bond order from a non-bonded situation to a triple bond state. As it can be seen interactions begin to take place at large distances, so that is possible to model long range, partially bonded states. The overcoordination is a problem arising due to long range interactions with second nearest neighbour. An example of this kind are weak bond interactions between carbons and hydrogen bonded to the nearest neighbour carbon atom.

The bond order must be corrected to avoid overcoordination as follows:

$$BO_{ij} = BO'_{ij} \cdot \mathcal{F} [\Delta'_i, \Delta'_j] \quad (4.1.4)$$

where:

- BO_{ij} is the corrected bond order by the correctio function $\mathcal{F} [\Delta'_i, \Delta'_j]$;

- $\Delta'_i = \sum_{j=1}^{nbond} (BO_{ij} - Val_i)$

- Val_i is the valence of atom i

- $\Delta'_j = \sum_{i=1}^{nbond} (BO_{ij} - Val_j)$

- Val_j is the valence of atom j

- BO'_{ij} is the uncorrected bond order.

An example of such a correction in ethane molecule is shown in figure 4.1.2.

Once the bond order BO'_{ij} has been corrected in BO_{ij} , if atom i is still overcoordinated, the interaction between i and j gets a penalty energy just to prevent an erroneous modeling:

$$E_{over} = p_{over} \cdot \Delta_i \cdot \left[\frac{1}{1 + \exp(\lambda_6 \cdot)} \right] \quad (4.1.5)$$

where:

- E_{over} is the penalty energy for overcoordination;
- p_{over} is a fitted parameter;

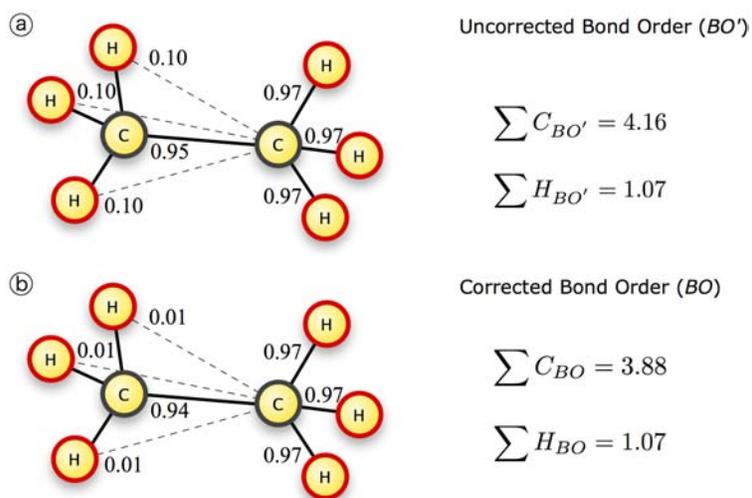


Figure 4.1.2: (a) Uncorrected bond orders for carbon and hydrogen in ethane molecule. (b) Corrected bond orders with weak interactions with hydrogen atoms neglected.

- $\Delta_i = \sum_{j=1}^{nbond} (BO_{ij} - Val_i)$;
 - BO_{ij} is the corrected bond order by equation 4.1.4;
 - Val_i is the valence of atom i ;
 - $\Delta_i > 0$ in case of overcoordination;
 - $\Delta_i < 0$ in case of undercoordination;
- λ_6 is a fitted parameter.

The bond energy, E_{bond} , is finally calculated as a function of 4.1.4:

$$E_{bond} = -D_e BO_{ij} \exp [p_{be,1} (1 - BO_{ij}^{p_{be,1}})] \quad (4.1.6)$$

where:

- E_{bond} is the bond energy contribution to the total potential energy;
- D_e and $p_{be,1}$ are fitting parameters
- BO_{ij} is the bond order equation 4.1.3

Upon dissociation BO_{ij} goes to zero, so the bond term for the couple of atoms (i, j) goes to zero too. The same dependence on bond order is present also in ReaxFF description of valence angles and torsions. These interactions are usually described with a simple harmonic relationship regardless of how strong and weak a bond gets. The potential term for a bond angle is:

$$E_{angle} = [1 - \exp(\lambda \cdot BO_{ij}^3)] [1 - \exp(\lambda \cdot BO_{jk}^3)] \cdot \left\{ k_a - k_b \exp \left[-k_b (\phi - \phi_0)^2 \right] \right\} \quad (4.1.7)$$

where:

- E_{angle} is the covalent angle potential term;
- λ is a fitted parameter;
- BO_{ij} and BO_{jk} are the bond order of atoms (i, j, k) involved in the covalent angle.
- k_a and k_b are fitting parameters;
- ϕ is the measure of the covalent angle.

4.1.2 Non-Covalent Interactions

Among non bonded interactions we have charge polarization, Coulombian electrostatic interactions, and vdW forces.

4.1.2.1 Charge Polarization

The method with which ReaxFF is capable to polarize charges within molecules is based on the EEM and Qeq methods [71][72][73]. The polarization is calculated with the following system of equations:

$$\begin{cases} \frac{\partial E}{\partial q_2} = \chi_2 + 2q_2\eta_2 + C \cdot \sum_{j=1}^n \frac{q_j}{\{r_{2,j}^3 + (1/\gamma_{2,j})^3\}^{1/3}} \\ \vdots \\ \frac{\partial E}{\partial q_i} = \chi_i + 2q_i\eta_i + C \cdot \sum_{j=1}^n \frac{q_j}{\{r_{i,j}^3 + (1/\gamma_{i,j})^3\}^{1/3}} \\ \vdots \\ \frac{\partial E}{\partial q_n} = \chi_n + 2q_n\eta_n + C \cdot \sum_{j=1}^n \frac{q_j}{\{r_{n,j}^3 + (1/\gamma_{n,j})^3\}^{1/3}} \end{cases} \quad (4.1.8)$$

where:

- $\frac{\partial E}{\partial q_i}$ is the chemical potential defined as the derivative of the energy, E , with respect to the charge q_i of i^{th} atom;
- χ_i is the electronegativity of the i^{th} atom;
- η_i is the hardness of the i^{th} atom;
- C and $\gamma_{i,j}$ are fitted parameters;

plus imposing the electroneutrality condition:

$$\sum_{j=1}^n q_j = 0 \quad (4.1.9)$$

The set of equations 4.1.8 together with the condition of electroneutrality expressed in equation 4.1.9 is solved to obtain the n charges.

4.1.2.2 Coulomb and van der Waals forces

In ReaxFF the Coulomb and van der Waals forces are calculated between all atom pairs, irrespective of their connectivity. A shielding term is put in front of potential terms for Coulomb and van der Waals energy terms to avoid excessive repulsive or attractive interactions between bonded atoms.

The Coulomb potential term is:

$$E_{coul} = C \left\{ \frac{q_i q_j}{\left[r_{ij}^3 + (1/\gamma_{ij})^3 \right]^{1/3}} \right\} \quad (4.1.10)$$

where:

- E_{coul} is the coulomb energy term;
- C is a fitted parameter;
- q_i and q_j are the charges of the pair of atoms i and j ;
- γ_{ij} is the shielding parameter previously mentioned.

While the van der Waals term it is used a distance corrected Morse Potential:

$$E_{vdW} = D_{ij} \left\{ \exp \left[\alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{vdW}} \right) \right] - 2 \exp \left[\frac{1}{2} \alpha_{ij} \left(1 - \frac{f_{13}(r_{ij})}{r_{vdW}} \right) \right] \right\} \quad (4.1.11)$$

where:

- E_{vdW} is the van der Waals potential term;
- α_{ij} is a fitted parameters;
- r_{vdW} is the van der Waals radius
- $f_{13}(r_{ij}) = \left[r_{ij}^{\lambda_1} + \left(\frac{1}{\lambda_2} \right)^{\lambda_3} \right]^{1/\lambda_4}$ is the distance correction term where:
 - $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are fitted parameter.

4.2 ReaxFF Parameterization Techniques

4.2.1 Standard Parameterization Machinery

In a force-field like ReaxFF the less computational power needed compare to QM is obtained at a cost of a larger numbers of parameters with respect to ordinary force-fields. The parameters needed to describe the potential curve of ReaxFF are many [18]. General and specific ReaxFF parameters, both bonded and non-bonded for each of the element described, form what is called a “parameter space”. These parameters must be defined before one can apply the force field to the desired chemical problem. This is a general challenge for all force fields, and different methods have been developed to optimize the parameters to give the best description of different systems. Before one can start optimizing the parameters, some starting values need to be guessed, which is mostly done by combining experimental data or physical constants. To be able to tell how well a set of parameters perform, an error function, defining the total error in the force field with a given training set, is constructed.

ReaxFF parameterization procedure [74] uses the following equation:

$$\text{sum of squares} = \sum_i^n \left[\frac{(\chi_{i,ref} - \chi_{i,calc})}{\sigma} \right]^2 \quad (4.2.1)$$

where:

- *sum of squares* is the definition of error;
- the sum is done over all the n reference data of a defined data set (training set);
- $\chi_{i,ref}$ is the reference value of the training set;
- $\chi_{i,calc}$ is the calculated value;
- σ is the acceptance value.

The acceptance criterion σ in this equation is used as a tool to impose the accuracy with which the force field reproduce the reference data of the training set. A deviation between the calculated and the experimental data of more than this acceptance criterion has a relatively large influence upon the total sum of squares. When the experimental error is large, the acceptance value increases. On the other hand, when an experimental observation was of particular importance for the final shape of the force field, the acceptance value is reduced. Table shows some example of σ values as used during ReaxFF parameterization:

The goal in any force field development is to minimize function 4.2.1 by changing, or optimising the parameters. The choice of which parameters to optimize requires intimate knowledge of the chemical systems and the functional form of the force field. It would be desirable to have a computer-aided selection of parameters so to have a more rigorous approach to the task. We thought

Data	Acceptance Criterion
valency angle	1.0°
bond length	0.005Å
torsion angle	2.0°
heat of formation	2.09 KJ/mol*
	*0.5 Kcal/mol

Table 4.1: Acceptance criteria

Principal Component Analysis would be of advantage to tackle the problem of choosing the right parameters to adjust as briefly reported in appendix #, matlab software produced is given in the same appendix. PCA strategy has not been used for the parameterization of force-field for protein. The optimization strategy used for the development of the Prot-ReaxFF for biocatlaylsis application uses an optimisation algorithm written by A. C. T. van Duin. It is a parameterization algorithm that optimizes one parameter at a time. At first, the value of the error function is calculated for the initial parameter, then the error function is successively calculated for other two points defined with a biased step size from the initial parameter. Such a step size can defined differently for each paramter. Using the three points a parabola is defined. The minimum of this parabola is taken to be the new value for the parameter. If equation 4.2.1 has several minima, one cannot be sure to get to the global minimum, if a too small or too large stepsize is chosen. A large value will not assure a good sampling of the parameter space, while a too small step-size would increased the parameterization time with the risk of remaining in a local minima. This is known, though, and an “accepted increase in error” control is actually available in the program. The risk of parameterizing using a one paramter at a time procedure is that it does not take into consideration correlation between pa-

rameters. A more sophisticated method that uses the entire parameter space is the Particle Swarm Optimization (PSO). This will be briefly reported in section 4.2.3, even if it has not been employed for the force-field development object of this thesis.

4.2.2 Computer Aided Selection of Parameters

The main difficulty in a parameterization procedure is typically the choice of a proper training-set. In a force-field like ReaxFF the less computational power needed with respect to QM is obtained with a larger numbers of parameters compared to ordinary force-fields. The parameters needed to describe the potential curve of ReaxFF are many [18]. General and specific ReaxFF parameters (both bonded and non-bonded for each of the elements described) form a “parameter space”. We use as an example the trivial approach to optimize only the torsion part of the potential. In fact, all the energy potential contributions have an interplay in determining the energy of a system (e.g. the torsion description needs not only a well parameterized torsion potential, but also a proper description of non-bonded interactions). This section describes a procedure on how to reduce the parameter space using Principal Component Analysis (PCA) [75][76]. The matlab scripts to accomplish the task are provided in Appendix A. The project is still under development, right now it uses the Hydrocarbon version of ReaxFF force-field [18]. The goal is to obtain a more efficient parameter optimization step, where only key parameters are included. The current ReaxFF parameterization procedure optimizes one parameter at time [74][18]. This procedure is not optimal because on one hand it is a time consuming procedure to obtain new parameters, and on the other hand the parameters that

are strictly correlated can be unbalanced because they should be optimized at the same time. The PCA might help to discriminate which parameters are correlated, and in this way it reveals hidden relationship where the intuition of a chemist is not enough. The PCA might act reducing the computational time for the optimization without solving any issue regarding the “unbalancement” risk of considering variation of only one parameter at the time. The last consideration is that PCA might consider a correlation between parameters that is an underestimation of the real correlation between parameters.

Matrix Formulation. The object of such a sensitivity analysis is a sensitivity matrix. This matrix should represent the internal response of the system resulting from the variation of an i^{th} parameter value. The “samples” ensemble (where samples has the ordinary chemometrics meaning) can be thought as the union of the various terms of the energy calculated by equation 4.1.1 together with the ensemble of geometrical descriptors for the system under investigation (bond length, angles, torsions etc.). Let’s call each element of this ensemble s_i and the total ensemble of m elements $\{s_i\}$. We collect also all the parameters of ReaxFF creating the ensemble of n parameters $\{p_i\}$. Since we are interested in “variations” of the samples we will build the ensemble $\{ds_i\}$ where each variation has been obtained as a consequence of the variation of each parameter in $\{dp_i\}$. Each value of $\{ds_i\}$ has to be normalized (apart from the portion of energies because since they are portion of the error function they are already pure number) as well the the variation of parameters. At the end we build a vector of m elements for the variation of samples \vec{dS} :

$$\vec{dS} = \begin{pmatrix} Ebond_{errf} \\ Eatom_{errf} \\ \vdots \\ Angle_{H-C-H} \\ Length_{C-H} \\ \vdots \end{pmatrix} \quad (4.2.2)$$

And a vector $\frac{\vec{dp}}{p}$ of n elements for the variation of the parameters:

$$\frac{\vec{dp}}{p} = \begin{pmatrix} \frac{dp_1}{p_1} \\ \frac{dp_2}{p_2} \\ \vdots \\ \frac{dp_i}{p_i} \\ \vdots \\ \frac{dp_n}{p_n} \end{pmatrix} \quad (4.2.3)$$

The tensor that governs the interplay between vector \vec{dS} and vector $\frac{\vec{dp}}{p}$ is a tensor $m \times n$, our sensitivity matrix \mathbb{X} :

$$\mathbb{X} = \begin{pmatrix} \frac{\partial s_1}{\partial \ln(p_1)} & \frac{\partial s_1}{\partial \ln(p_2)} & \frac{\partial s_1}{\partial \ln(p_3)} & \frac{\partial s_1}{\partial \ln(p_4)} & \dots & \frac{\partial s_1}{\partial \ln(p_j)} & \dots & \frac{\partial s_1}{\partial \ln(p_n)} \\ \frac{\partial s_2}{\partial \ln(p_1)} & \frac{\partial s_2}{\partial \ln(p_2)} & \frac{\partial s_2}{\partial \ln(p_3)} & \frac{\partial s_2}{\partial \ln(p_4)} & \dots & \frac{\partial s_2}{\partial \ln(p_j)} & \dots & \frac{\partial s_2}{\partial \ln(p_n)} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ \frac{\partial s_i}{\partial \ln(p_1)} & \frac{\partial s_i}{\partial \ln(p_2)} & \frac{\partial s_i}{\partial \ln(p_3)} & \frac{\partial s_i}{\partial \ln(p_4)} & \dots & \frac{\partial s_i}{\partial \ln(p_j)} & \dots & \frac{\partial s_i}{\partial \ln(p_n)} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ \frac{\partial s_m}{\partial \ln(p_1)} & \frac{\partial s_m}{\partial \ln(p_2)} & \frac{\partial s_m}{\partial \ln(p_3)} & \frac{\partial s_m}{\partial \ln(p_4)} & \dots & \frac{\partial s_m}{\partial \ln(p_j)} & \dots & \frac{\partial s_m}{\partial \ln(p_n)} \end{pmatrix} \quad (4.2.4)$$

Where:

- $\mathbb{X}_{ij} = \frac{\partial s_i}{\partial \ln(p_j)}$

In matrix notation the problems is simply:

$$\vec{dS} = \mathbb{X} \frac{\vec{dp}}{p} \quad (4.2.5)$$

Principal Component Analysis. Following the formalism in a singular value decomposition fashion. The variance σ^2 is obtained as the norm of the sample vector:

$$\sigma^2 = [\vec{dS}^T] \cdot \vec{dS} = \begin{bmatrix} \vec{dp}^T & \mathbb{X}^T \end{bmatrix} \mathbb{X} \frac{\vec{dp}}{p} = \frac{\vec{dp}^T}{p} (\mathbb{X}^T \mathbb{X}) \frac{\vec{dp}}{p} \quad (4.2.6)$$

where the matrix $(\mathbb{X}^T \mathbb{X})$ is symmetric $n \times n$ and it can be diagonalized with a proper choice of a basis-set rotation:

$$(\mathbb{X}^T \mathbb{X}) = \mathbb{U} \mathbb{S} \mathbb{U}^T \quad (4.2.7)$$

The variace can be obtained substituting the factorization 4.2.7 in 4.2.6 so to obtain:

$$\sigma^2 = \frac{\vec{dp}^T}{p} \mathbb{U} \mathbb{S} \mathbb{U}^T \frac{\vec{dp}}{p} \quad (4.2.8)$$

Where:

- \mathbb{S} is a diagonal matrix having as elements the n eigenvalues;
- \mathbb{U} is an $n \times n$ matrix of eigenvectors.

The scalar σ^2 is calculated as:

$$\sigma^2 = \sum_i \lambda_{ii} \left[\frac{\vec{d}p}{p} \cdot \mathbb{U} \right]^2 = \sum_i \lambda_{ii} P_i \quad (4.2.9)$$

where:

- P_i is the element i-th of the transpose vector $\vec{P}^T = \left[\frac{\vec{d}p}{p} \cdot \mathbb{U} \right]^2$

The columns of \mathbb{U} are eigenvectors of the eigenvalue in \mathbb{S} . These eigenvectors represent the Principal Components of our parameter space: each element of each eigenvector is associated with one of the n parameters.

At the end of the analysis we have a diagonal matrix of eigenvalues \mathbb{S} and a matrix of eigenvectors \mathbb{U} . The value of each eigenvalue determine the more or less contribution of a certain eigenvector in varying the the system under investigation (\vec{dS}). Based on a threshold value ϵ it might be possible to divide the eigenvalues in 2 groups:

- the main-param-space: $\{\lambda_{ii}\} \geq \epsilon$;
- the null-param-space: $\{\lambda_{ii}\} < \epsilon$.

$$\mathbb{S} = \left(\begin{array}{ccccc} U_{11} & U_{12} & U_{13} & U_{14} & U_{15} \\ U_{21} & U_{22} & U_{23} & U_{24} & U_{25} \\ U_{31} & U_{32} & U_{33} & U_{34} & U_{35} \\ U_{41} & U_{42} & U_{43} & U_{44} & U_{45} \\ U_{51} & U_{52} & U_{53} & U_{54} & U_{55} \\ U_{61} & U_{62} & U_{63} & U_{64} & U_{65} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{i1} & U_{i2} & U_{i3} & U_{i4} & U_{i5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{n1} & U_{n2} & U_{n3} & U_{n4} & U_{n5} \end{array} \right) \left\| \left(\begin{array}{ccccc} U_{16} & \cdots & U_{1i} & \cdots & U_{1n} \\ U_{26} & \cdots & U_{2i} & \cdots & U_{2n} \\ U_{36} & \cdots & U_{3i} & \cdots & U_{3n} \\ U_{46} & \cdots & U_{4i} & \cdots & U_{4n} \\ U_{56} & \cdots & U_{5i} & \cdots & U_{5n} \\ U_{66} & \cdots & U_{6i} & \cdots & U_{6n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{i6} & \cdots & U_{ij} & \cdots & U_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ U_{n6} & \cdots & U_{ni} & \cdots & U_{nn} \end{array} \right)$$

\uparrow \uparrow
main – param – space *null – param – space*

Each element of each eigenvector is associated with a parameter. Again a threshold γ might be chosen to consider an elements of an eigenvector as significantly contributing to the relative eigenvalue or not. If the element is less than γ the relative parameter can be eliminated from the parameterization step because its variation does not influence $d\vec{S}$. Entire parameter families might be eliminated analyzing the elements of the eigenvectors belonging to the *main – param – space*. Once parameters are eliminated another PCA analysis might be carried out to see if the null space dimension decreases or not. Even if the procedure will end up with trivial results this would mean that the force-field description is consistent.

Illustrative Example. after doing a PCA on a parameter space of 5 parameters a diagonal matrix \mathbb{S} (5x5) was obtained where only the first two eigenvalues are greater than the threshold ϵ :

$$\mathbb{S} = \begin{vmatrix} \lambda_{11} & & & & \\ & \lambda_{22} & & & \\ & & \sim 0 & & \\ & & & \sim 0 & \\ & & & & \sim 0 \\ & & & & & \sim 0 \end{vmatrix}$$

The eigenvectors to be analyzed are the first two columns of \mathbb{U} . If an element of these columns is greater than γ its value is equal to 1, while if the element is less than γ , its value would be equal to 0. The hypothetical matrix will now appear like a logical matrix of 0 and 1:

$$\begin{array}{l} \textit{parameter } p_1 \rightarrow \\ \textit{parameter } p_2 \rightarrow \\ \textit{parameter } p_3 \rightarrow \\ \textit{parameter } p_4 \rightarrow \\ \textit{parameter } p_5 \rightarrow \end{array} \begin{vmatrix} 1 & 0 & \dots & \dots & \dots \\ 1 & 1 & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \dots \\ 1 & 1 & \dots & \dots & \dots \end{vmatrix} = \mathbb{U}$$

In this case the parameter p_1 , p_2 and p_5 will be the new parameter space and eventually they will be parameterized, while p_3 and p_4 will be left unchanged.

4.2.3 Other Parameterization Strategy

Particle Swarn Optimization (PSO) technique [77] has been applied to optimize parameters of ReaxFF force-field. The error function 4.2.1 is the objective function to be optimized. The standard parameterization machinery described previously is able to optimize one parameter at time and it is implemented as a serial code. PSO technique is instead a parallel code able to optimize more than one parameter at time. The search space of all the possible solution of the objective function is a multidimensional space of all the parameters to be optimized. The PSO engine works by maintaining several possible candidate solutions in such a search space. Each solution is a set of parameters and can be pictorially imagined as particles on a potential surface. At the beginning the positions of the particles are randomly chosen. The objective function is then calculated for each of the particle. Depending on the value obtained, each of the particle will be assigned with a value and its relative velocity. The best value is stored as “individual best candidate solution” for the objective funtion. After many iterations the “individual best candidate solution” form the so called array of “global best candidate solution”. At the end of PSO calculation the best value obtained will eventually represent the global best position for the particle, the best optimized set of parameters. The positions of particles are updated at each iteration step by being attracted by the partial global best positions. A certain amount of random movements are also allowed, so to avoid particles stuck in local minima and have a better sampling of the search surface. Everytime an objective function needs to be calculated it uses a ReaxFF instance. Since the algorithm developed uses the serial version of ReaxFF software, the rate limiting step of the process is the calculation of the objective function. The

#cores	Comp. Time [hh:mm]	Compared to 1 core [%]
1	24:34	100.00
2	13:33	55.16
4	09:56	40.43
8	03:48	15.47

Table 4.2: Parameterization of 30 particles over 100 runs PSO ReaxFF with 1, 2, 4, and 8 nodes.

maximum number of ReaxFF instances needs to be equal to the nodes available in the hardware, as well as the so called worker thread of PSO. If the number of ReaxFF instances would not be the same as PSO worker threads we would end up in a situation of overhead. On the other end one of the node needs to be reserved to the control thread, i.e. the thread that initializes each job and it waits for the results at the end. The parameterization algorithm has been developed at the department of mechanical engineering Results obtained with 30 particles over 100 iteration of PSO ReaxFF parameterization suggests a great deal of parallelization, even though some overhead is present because the scaling is not the ideal. In table 4.2.

4.3 Development of ReaxFF for Glycine

The first attempt of a parameterization of ReaxFF to study bio-molecules has been done by van Duin and coworkers in 2011 [66]. The paper describes the development of the ReaxFF potential for glycine, and in particular the direct tautomerization mechanism from the neutral to the zwitterionic through a proton transfer in water. We will give an overview of the parameterization and discussion of glycine force-field because it has been used as starting point for the development of ReaxFF for application to protein and enzymes [78]. This

force-field has been trained against binding energies of water clusters, concerted proton transfer reactions in neutral water, water self-ionization reactions, and density and cohesive energies for ice crystals water. This water description was validated for bulk water by comparison against experimental data on water cohesive energy, diffusion constant and structure. Some properties of this water potential are available in the literature [79]. Oxygen and Hydrogen parameters from this ReaxFF water description have been added to the existing ReaxFF potential for hydrocarbon oxydation [80]. A training set of glycine conformers and glycine water complexes were optimized against QM geometries and energies. QM calculations were performed using the density functional theory with functional B3LYP [17][16] and a split valence basis set 6-311G**++ basis set [81]. Gaussian suite 03 was used for all the energy, geometry and frequency calculations. Angular distortion and rotational energy barriers for torsion angles were calculated fixing the angle of interest.

4.3.1 Glycine valence and dihedral angle energy

QM calculations have been performed to determine valence angle energies such as H-N-C, H-N-H, C-C-N, C-C=O, and H-C-H. The agreement of the parameterized force-field with QM data is shown in figure 4.3.1. ReaxFF reproduces QM distortion energies with an rmsd of 0.79 Kcal/mol. Consistently with previous QM studies [82] the energy barrier of C-C-O-H torsion is higher compare to the other two examples shown in figure 4.3.1. QM geometries favor a symmetric structure containing two intramolecular N-H—O hydrogen bonds [83][84] over the asymmetric structure favored by ReaxFF with only one N-H—O hydrogen bond.

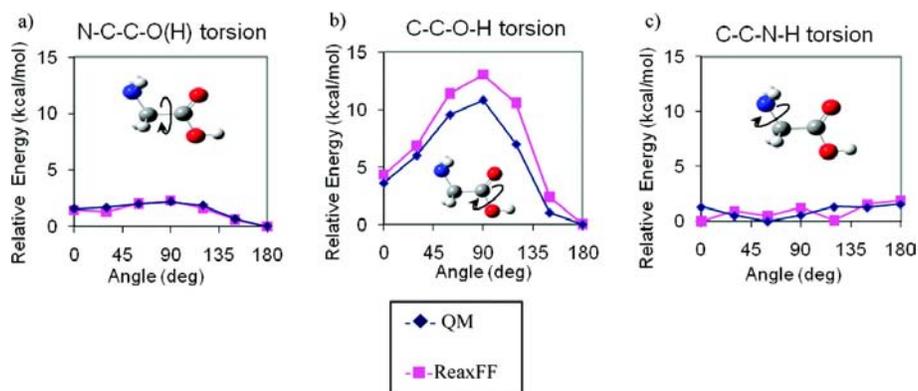


Figure 4.3.1: QM and ReaxFF energies for different dihedral angles of glycine in gas phase. "Reprinted with permission from O. Rahaman, A. C. T. van Duin, W. A. Goddard, III, and D. J. Doren, *Development of a ReaxFF Reactive Force Field for Glycine and Application to Solvent Effect and Tautomerization* (2011) The Journal of Physical Chemistry B, Vol. 115, No. 2, Copyright (2011) American Chemical Society."

4.3.2 Glycine gas-phase conformers

Glycine is predicted to favor a neutral conformation in gas phase [83][85][84][86][87][88][82]. The five lowest energy conformations of previous studies [83][88] have optimized and named nA, nB, nC, nD, and nE. Additional experimental proof of at least the first three conformers using matrix-IR spectroscopy [89][90][91][92] and microwave spectroscopy [93][94][95], and electron diffraction [96]. ReaxFF behaves well compare to QM, as it can be seen in figure 4.3.2. The lowest energy conformer nA has been correctly predicted by ReaxFF as the global minimum. All the glycine conformers in gas phase are predicted as neutral by ReaxFF, this is in accordance with literature [83][85][84][86][87][89][82][90][91][92].

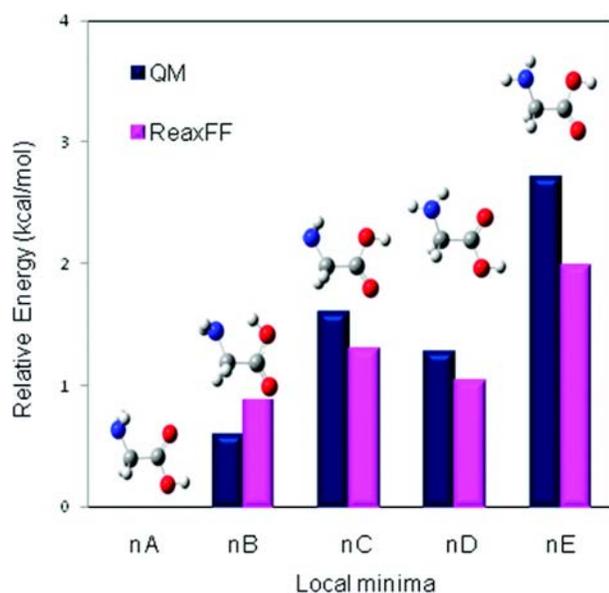


Figure 4.3.2: QM and ReaxFF energies for five glycine minima in the gas phase. "Reprinted with permission from O. Rahaman, A. C. T. van Duin, W. A. Goddard, III, and D. J. Doren, *Development of a ReaxFF Reactive Force Field for Glycine and Application to Solvent Effect and Tautomerization* (2011) The Journal of Physical Chemistry B, Vol. 115, No. 2, Copyright (2011) American Chemical Society."

4.3.3 Glycine-water complexes

To have a good behaviour of water molecules interactions with glycine QM calculations have been done on various conformers of glycine coordinated to water molecules. In particular complexes of glycine with one and two water molecules, respectively Gly-1H₂O and Gly-2H₂O in what follows, have been included in the training set. Initial configurations for the geometry optimization of Gly-1H₂O were partly available in literature [97][98][99][100] or manually built from gas phase conformers placing a water molecule in a way to have hydrogen bond between glycine and water. Configurations of 12 Gly-1H₂O

complexes have been studied. They are shown in the first column of table 4.3 together with a comparison of the energy obtained via QM and via ReaxFF, the average unsigned error is of 2.82 Kcal/mol. The last column of the table reports the root mean square deviation between QM and ReaxFF geometries. The average rmsd is 0.15Å.

Structure a) of table 4.3 is the QM global minimum as reported also in previous literature computational studies [97] [98][100] and a microwave spectroscopy study [101]. The planar ring structure formed by the glycine carboxyl termination and a water molecule is probably the reason of its stability. ReaxFF does not describe conformer a) as global minimum, though its ReaxFF energy is only 0.1 Kcal/mol higher than complex i). Apart from the zwitterionic complexes m) and n), all structures are optimized without any restraint. At least two water molecules are needed to stabilize the zwitterionic form [100]. To obtain zwitterionic complexes the 3 amide protons have been restrained at a distance of 1.08Å from the nitrogen. The relative instabilities of zwitterionic complexes compare to the other neutral conformers are well described by ReaxFF.

For Gly-2H₂O complexes the procedure to build the input was the same as Gly-1H₂O with two water molecules placed in a favorable position around neutral and zwitterionic glycine. The conformations obtained are 24. With two water molecules there was no need of imposing constraints to have minimization of zwitterionic complexes. Due to lack of space we don't report the structure of the set Gly-2H₂O. Detailed informations about the structures of Gly-2H₂O, the comparison of ReaxFF, and QM data are available in the literature [66]. The comparison between QM and ReaxFF minimization give an average unsigned error of 3.5 Kcal/mol, and an average rmsd of 0.16Å, so there is a reasonable

conformer	relative energy [Kcal/mol]		rmsd (Å)
	QM	ReaxFF	
a 	0.00	0.13	0.07
b 	6.67	2.97	0.06
c 	3.41	2.62	0.11
d 	3.41	2.29	0.12
e 	5.15	1.08	0.27
f 	4.91	0.88	0.16
g 	5.09	4.21	0.13
h 	5.99	1.92	0.16
i 	1.21	0.00	0.12
l 	4.27	2.55	0.09
m 	19.86	14.61	0.18
n 	22.53	15.68	0.37

Table 4.3: QM and ReaxFF Relative Energies of Glycine-Water complexes in the gas phase

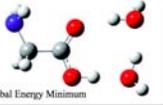
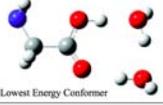
Conformer	relative energies [Kcal/mol]		
	QM	ReaxFF	rmsd [Å]
 Global Energy Minimum	0.00	0.00	0.10
 2 nd Lowest Energy Conformer	1.09	0.39	0.09

Table 4.4: QM and ReaxFF Relative Energies of the global energy minimum and the second lowest energy minimum of Gly-2H₂O complexes in gas phase.

match between QM and ReaxFF results. The global minimum is again stabilized by a planar structure formed between the carboxyl moiety of glycine and the two water molecules. The second lowest energy conformer is only 1.1 Kcal/mol higher in energy than the local minimum. The same result is found in other computational studies [100]. ReaxFF successfully predict the same QM global energy minimum plus the same second lowest energy conformer with a 0.4 Kcal/mol higher energy.

The part of the training set dedicated to the hydrogen transfer to model the gas phase tautomeric transformation from zwitterionic form to neutral form have been parameterized against several glycine conformers along the reaction coordinate. The zwitterionic form is unstable in gas phase, so there is no energy barrier [100]. Both QM and ReaxFF predicted a neutral form down in energy with respect of the zwitterionic form as well as it is correct also the energy trend along the reaction coordinate, though the energy barrier predicted by ReaxFF underestimates of ~ 5 Kcal/mol the energy difference between the two tautomeric forms. The Glycine ReaxFF potential has been successfully applied

to investigate the effect of solvation on the conformational equilibrium of neutral form in water and to obtain energy profiles of the proton transfer reaction in water with a direct mechanism and by one or two water molecules mediated mechanism. The one water mediated mechanism had the lowest energy barrier, and this suggests that the neutral form to zwitterionic tautomerization in water might mediated by a single water molecule [66].

The potential developed can be applied to characterize biologically important phenomenon. It represents a first step towards the development of a realistic model to describe complex biological phenomena. Such a development is the subject of the next paragraph of this manuscript.

4.4 Development of ReaxFF for Protein and Enzymes

The subject of this paragraph is the development of a reactive force-field, Prot-ReaxFF, especially designed to treat biological systems. This work extends the the previously mentioned ReaxFF force-field for glycine [66] to protein structures and the reaction mechanism of biomolecular systems. Prot-ReaxFF is potentially a fast, simple, and efficient method which is capable of differentiating reactive and non-reactive species. The current more credited QM-MM methods to deal with biological systems in their entirely atomic details. QM-MM description is based on the association between QM methods and a force-field based approach. This hybrid technology is able provide accurate results with a low-demanding computational power even for system with more than one hundred thousands of atoms, however it is characterized by very well known problems: i) the level of theory, ii) the basis set, and iii) the definition of the boundary between QM part and the remaining part treated with a force-field based approach. The level of theory and the basis set influences significantly simulation trends and values. The boundary definition usually adds additional atoms which are not part of the system. It is also probable that polarization problems arise at the boundary region. This is due to the mutual interaction between QM density of charge and the charges associated to each atom type of the force-field description. Prot-ReaxFF does not have any problem bound to the choice of the basis set, it can treats systems of the same size of QM-MM methods and it treats all the atoms of the system with the same description, so there is no boundary region and all the problems associated with it. In what follows it will be described the parameterization procedure of the state of the art of Prot-ReaxFF. A comparison between Amber force-field and Prot-ReaxFF

has been carried out on a data set of small proteins and oligopeptides, to determine the performances of Prot-ReaxFF compared to a non-reactive force-field. The reactive performances are instead presented as a mechanistic study on the hydrolysis of β -lactam antibiotics and a chymotrypsin enzyme.

4.4.1 Parameterization

Starting from the parameters developed for glycine-water complexes by Adri Van Duin and coworkers [66] molecular models of increasing complexity has been added to the training set to improve ReaxFF capability of describing biomolecular systems. The parameterization follows a three stage approach (see top flow chart of table 4.5). The first stage is the starting force-field developed for glycine-water complexes [66] to which has been added the first ten low energy conformation of each aminoacid. QM data of the conformation analysis of each aminoacid has been provided by Jijing and coworkers. The last stage is instead the improvement of the training set for aminoacid dipeptides, tripeptides and tetrapeptides. QM optimized geometries were available in the literature as supplementary material [102][103][104] and well organized database [105]. In table 4.5 it is graphically represented the adopted parameterization three-stage approach. Not all the structure were actually added to the training set. Some of them have been used as a validation test for comparing Prot-ReaxFF and QM results.

Glycine, alanine, serine, and cysteine dipeptides (respectively entry number 2, 3, 4, and 5 of table 4.5), side chains corresponding to -H, -CH₃, CH₂OH, and -CH₂SH, respectively, have been taken from a study of Kaminsky and Jensen [102]. N-Acetyl and N-methylamide functional groups are added to the C- and

N-terminus to mimic the environment in longer peptide chains. The procedure through which the conformational study has been carried out proceeds with trial structures generated by varying the ϕ and ψ torsion angles in steps of 30° , while the χ_1 torsional angle was varied in steps of 120° . The torsion angle χ_2 was present at a value of 180° , while the torsional angles ω_1 and ω_2 associated with the terminal amide bonds were given initial values of 180° , corresponding to trans junctions. This produces a total of 144 starting geometries for glycine and alanine and 432 for serine and cysteine. Additional trial conformations for the serine and cysteine systems have obtained by varying the χ_2 torsional angle. All trial structures were optimized without any restraints at the MP2/ 6-31G(d,p) level and characterized as true minima by frequency calculations, and the unique structures were reoptimized with the aug-cc-pVDZ basis set. Improved estimates of the relative conformational energies were obtained by single point MP2 and B3LYP calculations with an augmented double- ζ basis set (cc-pVDZ, cc-pVTZ, cc-pVQZ, and aug-cc-pVTZ), and method dependency was tested using CCSD(T)/cc-pVDZ calculations. The best estimate of relative energies was obtained by separately extrapolating the Hartree-Fock(cit HF) and MP2 (Cit. MP2) energies to the basis set limit and adding the CCSD(T) correction. The selected structures consist on 2 conformers of Acetyl-Glycine-N-methylamide, 7 conformers of Acetyl-Alanine-N-methylamide, 47 conformers of Acetyl-Serine-N-methylamide, and 38 conformers of Acetyl-Cysteine-N-methylamide. These conformers have been used as a validation set for Prot-ReaxFF parameters obtained, just to asses the reliability of it with a consistent number of QM datas.

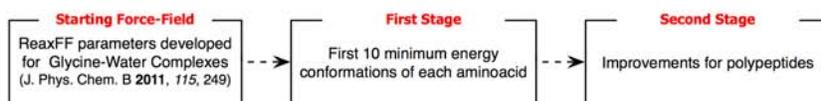
Part of the dataset of Hobza and coworkers [106][104] have been directly included in the training set (respectively entry number 6, 7, 8 of table 4.5), while

the rest of the structures have been used to assess the validity of Prot-ReaxFF (respectively entry number 9, 10, 11 of table 4.5). Structures are short named as follows: Phe-Gly-Phe = FGF, Trp-Gly = WG, Trp-Gly-Gly = WGG, Phe-Gly-Gly = FGG, Gly-Gly-Phe = GGF, and Gly-Phe-Ala = GFA. Geometry optimizations were performed at the RI-MP2/cc-pVTZ, and also by means of Amber FF99SB force-field with HF/6-31G* RESP charges.

A series of alanine tetrapeptide conformations have been studied by Martin Head-Gordon and coworkers [103]. They sampled 24 conformations of alanine tetrapeptide. All the conformations were optimized at HF/6-31G** level. In all MP2 calculations performed on the already optimized HF structures, the frozen-core approximation was introduced to reduce computational timings and resources. Single point calculations have been done with cc-pVTZ and cc-PVQZ basis sets and then extrapolation to the cc-pV(TQ)Z limit. Data on Tetra Alanine peptide were not part of the trainset (entry number 12 of table 4.5), though they are used instead as a validation set to test Prot-ReaxFF results.

From figure 4.4.1 it can be noted that both AMBER and Prot-ReaxFF force field show oscillations around the reference ab initio data. The data set for Tetra-Alanine peptide shows bigger oscillation even though the description is not significantly different from the other cases. In figure 4.4.2 a regression analysis of the various data sets is presented. The scattered data suggest a satisfactory and some of the time (67% of the cases) superior description of Prot-ReaxFF compare to AMBER force fields. Tetra Alanine conformations, as well as the Hobza's peptides used as validation set only, are described equally well even though the description is slightly worse than that one of AMBER FF99SB. Ab initio geometries have been compared with AMBER and ReaxFF geometries.

Parameterization Strategy



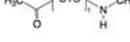
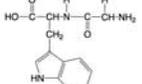
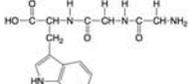
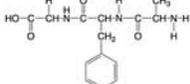
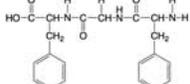
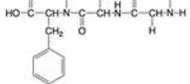
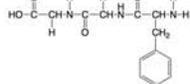
Training and Validation Sets		
	Used as Training set	Used as Validation set
21 Aminoacids 10 low energy conformations	✓	
 (2 conformations)		✓
 (7 conformations)		✓
 (47 conformations)		✓
 (38 conformations)		✓
 (15 conformations) WG	✓	
 (15 conformations) WGG	✓	
 (16 conformations) GFA	✓	
 (15 conformations) FGF		✓
 (15 conformations) FGG		✓
 (15 conformations) GGF		✓
Tetra-Alanine (27 conformations)		✓

Table 4.5: QM Training Set and QM validation Set

The average root mean square deviation (RMSD) of ReaxFF compare to *ab initio* is about 0.08Å (with a maximum of 0.2Å). For the same observable the comparison of ordinary force fields with *ab initio* data show larger oscillations -errors in the 0.1-0.3Å range- compare to ReaxFF. Superimposing Prot-ReaxFF with *ab initio* structures revealed that the main difference is how side chains are oriented with respect to the backbone. The comparison with *ab initio* data suggests the possibility of using Prot-ReaxFF to obtain a description that is equivalent to high level QM data. At the same time both energetic and geometric results of Prot-ReaxFF suggest an accuracy that is at least comparable with that one of other force fields commonly used to model bio-molecules.

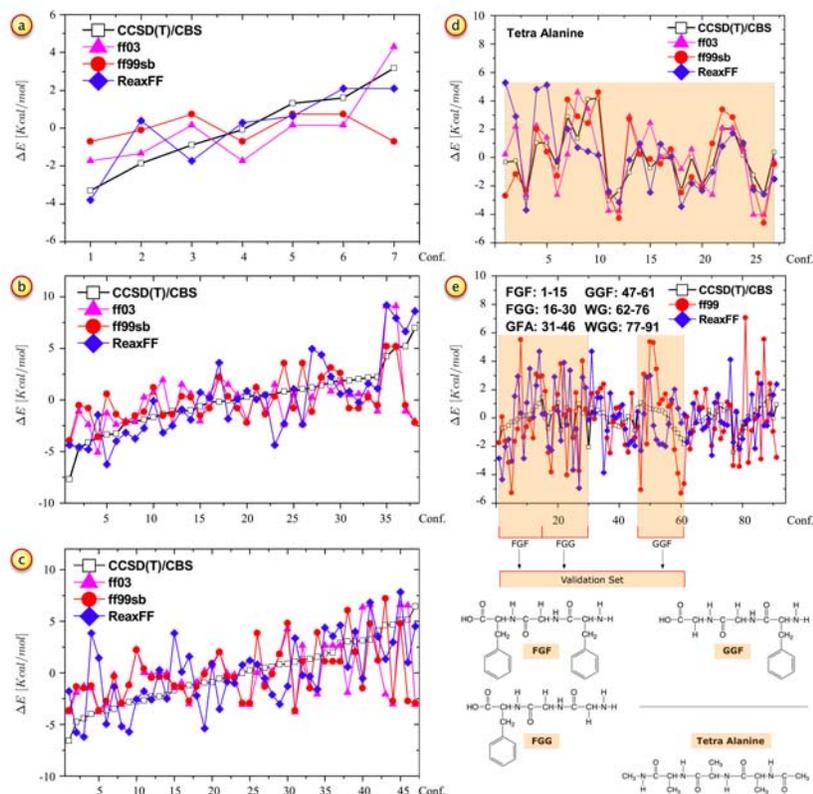


Figure 4.4.1: Energy Differences of the various conformers with amber, ReaxFF FFIELDS, and the ab-initio CCSD(T)/CBS level with respect to the corresponding minimum energy structure: a) Alanine Conformers; b) Serine Conformers; c) Cysteine Conformers; d) Tetra Alanine Conformers; e) FGF (Phe-Gly-Phe), FGG (Phe-Gly-Gly), GFA (Gly-Phe-Ala), GGF (Gly-Gly-Phe), WG (TRP-Gly), WGG (Trp-Gly-Gly) conformers. Data sets used as validation set only have been highlighted.

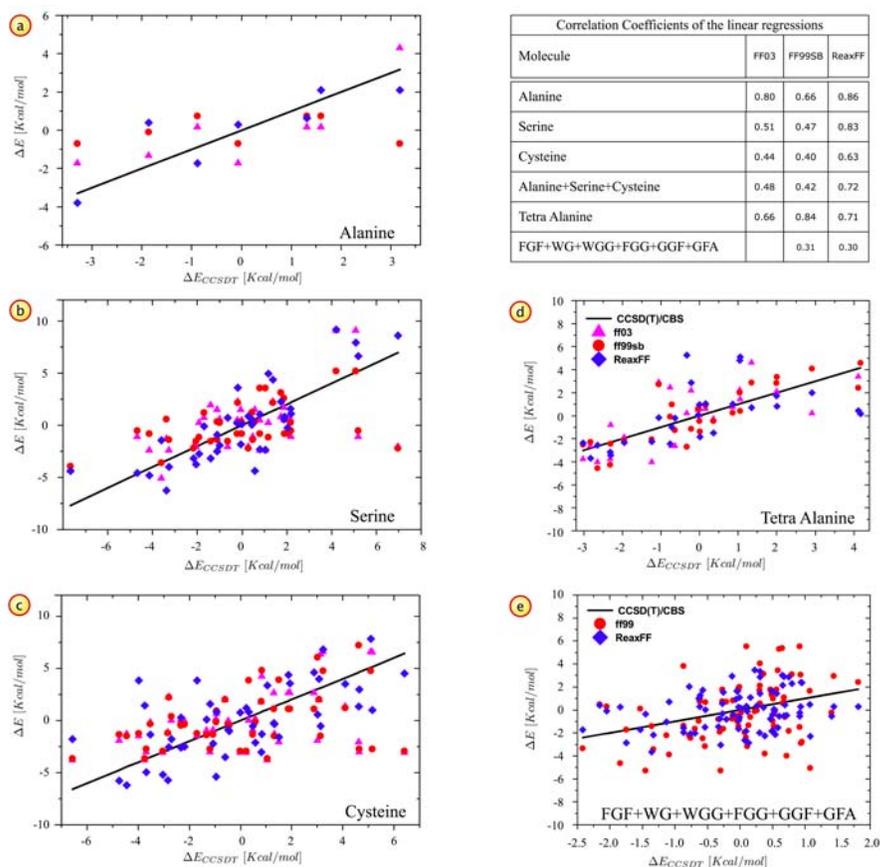


Figure 4.4.2: Linear regression of the amber, ReaxFF FFfields, with the energy differences calculated with ab-initio CCSD(T)/CBS level: a) Alanine Conformers; b) Serine Conformers; c) Cysteine Conformers; d) Tetra Alanine Conformers; e) FGF (Phe-Gly-Phe), FGG (Phe-Gly-Gly), GFA (Gly-Phe-Ala), GGF (Gly-Gly-Phe), WG (TRP-Gly), WGG (Trp-Gly-Gly) conformers. On the top-right it is reported a table with the relative correlation coefficients of the linear regression calculated.

4.4.2 Validation of ProtReaxFF non-Reactive Performances

The first question to be solved is whether ReaxFF can predict biological structure with the same accuracy as another ordinary force field. To answer the question a comparison between AMBER FF99SB and Prot-ReaxFF force fields performances have been carried out with small proteins and oligopeptides. We decided to focus the attention of 6 typical models having a definite secondary structure motifs:

- Crambin protein (**Crambin**);
- Tryptophan Cage (**TrpCage**);
- Oligopeptides:
 - C-Terminal β -hairpin of G protein (**ProtG**);
 - the synthetic α_R -helix (**EK**);
 - the helical C peptide of ribonuclease A (**Baldwin**);
 - the synthetic trpzip2 hairpin (**trpzip2**).

All the calculations done with force fields can be found in table 4.6. We go briefly over the data set of small proteins and oligopeptides to emphasize their characteristics.

Crambin Protein (Crambin). It is a small molecule consisting of 46 residues [107] belonging to the thionin family. Even though it is small it contains several secondary structure motifs such as two antiparallel α -helices, two and one and a half strands β -sheet, five turns among which a classical β -turn, an extended region, hydrophilic and hydrophobic residues, a salt bridge between Arg10 and Asn46, the structure is cross-linked by three disulfide bridges: one of

them between the two α -helices, the second between β -strands, and the third between second β -strand and a mobile floppy loop. Through a recent synchrotron X-ray study, an high-resolution structure (0.48 Å) of Crambin has been determined [107]. The protein crystallized readily and its crystal structure has been known for two decades [108][109]. Crambin does not contain many residues that are capable of chemical activity or a noticeable active site, suggesting that its natural role is instead accomplished by its structure, shape or surface properties. However, its potential binding partners and interactions with the environment remain a mystery. Despite the hydrophobic nature of Crambin the accessible surface area in the crystal is close to that one of water-soluble proteins like mioglobin and carboxypeptidase A.

Tryptophan Cage (TRP-cage). Neidigh and coworkers [110] have designed a 20-residue sequence (NLYIQ WLKDG GPSSG RPPPS) that folds spontaneously and cooperatively to a “Trp-cage”, a globular fold with a combination of secondary structure and tertiary contacts more typical of a larger, more complex molecule. This synthetic miniprotein has a folded tertiary structure, a globular shape and a combination of secondary motifs, tertiary contacts typical of more complex proteins [107][110]. Folding is cooperative, fast (in about 4 μ s), and hydrophobically driven by the encapsulation of a tryptophan side chain (residue 6) in a sheath of proline rings shielding it from solvent contact. A short α -helix (residues 2-9), a 3_10 helix (residues 11-14) and a polyproline II helix at the C-terminus are present. TRP-cage has been studied both from experimental and theoretical point of view especially for its very fast folding feature accessible to time scale of ordinary force-field techniques. In many cases solvent effect plays a crucial role in conformation dynamics and molec-

ular stability, so an explicit solvent description is necessary to get a realistic representative model.

C-Terminal β -hairpin of G protein (ProtG). A β -hairpine structure consists of a β -turn connecting two strands of antiparallel β -sheets. C-terminal β -hairpin of G rprotein is a 16 residue peptide (Gly-Glu-Trp-Thr-Tyr-Asp-Asp-Ala-Thr-Lys-Thr-Phe-Thr-Val-Thr-Glu) that is the first reported case of a short linear peptide taht adopts β -hairpin conformation in water solution [111]. ProtG has 5 threonine residues out of 16. Threonine residues han an intrinsic β -sheet forming tendencies [112]. Fundamental contribution to the stability of the β -hairpin conformation are either hydrophobic interactions between Tyrosine/Phenilalanine and Tryptophan/Valine residue pairs and four salt bridges between charged residues.

Synthetic EK α_R -helix. It is a peptide with the repeating sequence of $Ac - Tyr - (Ala - Glu - Ala - Ala - Lys - Ala)_2 - Phe - NH_2$ that shows the capability of folding as an α -helix structure [113].

The helical C peptide of ribonuclease A (Baldwin). It is a short C peptide lactone with the following sequence: NH_2 -Lys-Glu-Thr-Ala-Ala-Ala-Lys-Phe-Glu-Arg-Gln-His-HSer(lactone). The helix is formed intramolecularly without the need for tertiary interactions. Specific side-chain interactions in the helix are responsible for its stability. The pH titration indicate that a salt bridge stabilizes the C-peptide helix, because both a negatively charged glutamate-9 and the positively charged form of histidine-12 (or possibly the a-NH' of lysine-1) are required for stable helix formation [114].

The synthetic trpzip2 hairpin (trpzip2). Tryptophan zipper (trpzip) is structural motif that greatly stabilizes the β -hairpin conformation in short

peptides. Folding free energies of the trpzip2 exceed substantially those of all previously reported β -hairpins and even those of some larger designed proteins. The sequence of aminoacids of trpzip2 is Ser-Trp-Thr-Trp-Glu-Asn-Gly-Lys-Trp-Thr-Trp-Lys. The three-dimensional structures of trpzip2 were determined by NMR [115]. The interactions between side chains of its tryptophan residues and the mutual interaction of a salt bridge between Glu5 and Lys8 side chains are responsible for its stability.

4.4.2.1 Computational Details

AMBER 10[116] suite of programs have been used for all non-reactive simulations. with FF99SB [117] force-field. The simulation systems were all neutral. Periodic boundary conditions and explicit TIP3P [118] solvent model has been used. The models have been minimized for 5000 steps restraining heavy atoms. The systems have been gradually heated to 300K with an NVT simulation of 50ps with restraints on heavy atom positions. Such a restraints ($50 \text{ Kcal}\cdot\text{mol}^{-1}\cdot\text{\AA}^2$) have been gradually released until the beginning of the production runs without any restraint. A 12\AA cutoff for the short-range nonbonded interactions was employed in combination with the particle Ewald method [119] to treat long-range electrostatic interactions. The simulations were performed with time step of 1 fs. The relaxation protocol adopted consisted of two stages: i) 200 ps simulation in NVT ensemble, ii) 300 ps simulation in NPT ensemble with a weak coupling scheme to adjust solvent density to the bulk value. Berendsen [120] thermostat and barostat have been used to control temperature and pressure. Productive runs were carried out in NVT ensemble for 20ns saving snapshots every picosecond using Berendsen thermostat [120].

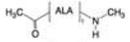
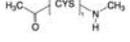
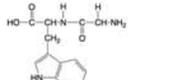
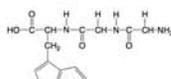
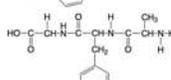
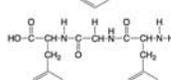
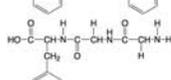
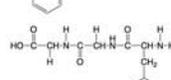
Amber Validation Set		
	Ffield FF03	Ffield FF99SB
Crambin		✓
TRP-Cage	Tryptophan Cage	✓
GProt	C-terminal β -hairpin of G protein	✓
EK	Synthetic EK α_R -Helix	✓
Baldwin	Helical C Peptide of ribonuclease A	✓
TRPzip2	Synthetic trpzip2 Hairpin	✓
21 Aminoacids		
10 low energy conformations	✓	✓
 (2 conformations)	✓	✓
 (7 conformations)	✓	✓
 (47 conformations)	✓	✓
 (38 conformations)	✓	✓
 (15 conformations) WG		✓
 (15 conformations) WGG		✓
 (16 conformations) GFA		✓
 (15 conformations) FGF		✓
 (15 conformations) FGG		✓
 (15 conformations) GGF		✓
Tetra-Alanine Peptide (27 conformations)	✓	✓

Table 4.6: Amber Validation Set

The geometries obtained at the end of the equilibration phase of the AMBER molecular dynamics were used as starting structures of the reactive simulations with Prot-ReaxFF. The protein/peptide conformations were close to the corresponding experimental geometries because of the restraints applied during the non-reactive equilibration. Moreover, the solvent density was close to the bulk value. Reactive simulations with a simulation time of 500ps were performed in the NVT ensemble using the Berendsen thermostat [120] with a relaxation constant of 0.1 ps. The equations of motion were solved with the velocity-Verlet algorithm [121], temperature was set to 300K, and the time step was set to 0.25 fs. Before printing production MD-NVT simulation, every system was energy minimized at T=0 K. Then, the system temperature was gradually increased from 0K to 300K in 12.5 ps.

4.4.2.2 Prot-ReaxFF validation with small proteins and oligopeptides

The validation procedure consisted of checking if the reactive force field can describe protein and oligopeptide structures, and their dynamics in solution. This is fundamental for using ReaxFF-based MD programs in the simulation of enzymatic reactions. The involved biomolecules could experience conformational changes that influence their catalytic power and, as a consequence, the reaction rates. Thus, it is necessary to explore the wide range of internal motions through extensive reaction-pathway calculations and configurational sampling. Representative geometries were extracted and should be then used to clarify the effects that the conformational mobility have on the enzymatic reactions. The structural reorganization of the molecules was monitored by examining the

root mean square deviations (RMSDs) of C_α from the initial arrangements together with the radius of gyration. Their average values, measured during the last 200 ps of the simulation time, are reported in table 4.8. From table 4.7 we see that in all cases, except for Crambin and trpzip2, Prot-ReaxFF structures are, on average, closer to the starting reference geometries than are the FF99SB ones with RMSD differences lower than 0.8\AA relatively to the other force field. Furthermore, all the RMSDs are within 1.7\AA with quite small standard deviations, suggesting that the two trajectories (that is Prot-ReaxFF and FF99SB) explore similar regions of the C_α conformational space in proximity to the starting arrangement. Positional fluctuations of C_α were also analyzed and plotted as a function of the residue number in figure 4.4.3 and figure 4.4.4. These descriptors reflect all together the degree of packing of the molecules and their propensity to unfold. No significant movements of the various domains (terminal regions excluded) are observed (Figure 4.4.5 and 4.4.6). However, in the case of Crambin (Figure 4.4.5 a), even if ReaxFF simulations capture the dynamics of the backbone atoms in the two closely aligned α -helices whose thermal fluctuations are small, the overall structure moves from a more compact arrangement (low RMSD - FF99SB) to a less packed state (higher RMSD - ReaxFF) where secondary structure elements are preserved but the terminal unstructured regions are visibly reorganized. This general trend can be readily interpreted in terms of the topology of the molecule. Indeed, residues involved in disulfide bridges (or in α -helices) have relatively smaller displacements from the reference structure and have lower mobility than the other regions. The three disulfide bridges (between the two helices - Cys16-Cys26, between the beta strands - Cys4-Cys32, and between a loop and the second beta strand -

Molecule	$RMSD$ (Å)			
	FF99SB		ReaxFF	
	$\langle RMSD \rangle$	σ_{RMSD}	$\langle RMSD \rangle$	σ_{RMSD}
Crambin	0.67	0.10	1.29	0.22
TRP-Cage	1.55	0.52	1.36	0.24
Baldwin	1.62	0.58	0.85	0.20
EK	1.42	0.55	1.02	0.35
ProtG	1.68	0.43	1.54	0.33
trpzip2	0.51	0.11	1.01	0.33

Table 4.7: Root mean square deviation of C_α 's relative to the starting conformation of the production run.

Molecule	R_{gyr} (Å)			
	FF99SB		ReaxFF	
	$\langle R_{gyr} \rangle$	$\sigma_{R_{gyr}}$	$\langle R_{gyr} \rangle$	$\sigma_{R_{gyr}}$
Crambin	9.78	0.06	9.84	0.10
TRP-Cage	6.40	0.08	6.43	0.20
Baldwin	5.40	0.13	5.07	0.06
EK	5.88	0.18	5.82	0.30
ProtG	7.63	0.22	7.22	0.21
trpzip2	5.54	0.06	5.41	0.08

Table 4.8: Radius of gyration of C_α 's relative to the starting conformation of the production run.

Cys3-Cys40 are correctly described by both force fields, the average distance between sulfur atoms is 2.1Å and the main chains of the two models are closely superimposed in figure 4.4.7. Notwithstanding individual changes are contained, as a whole they are responsible for the increase in the C_α RMS deviation.

A similar, but less pronounced, effect is observed instead in the case of **trpzip2**, where the higher value of Prot-ReaxFF RMSD, in relation to the FF99SB one, seems to be due to the lengthening of the intramolecular hydrogen bonds connecting the donor and acceptor groups on the opposite sides of the hairpin. This is especially evident in the terminus regions where no intrachain

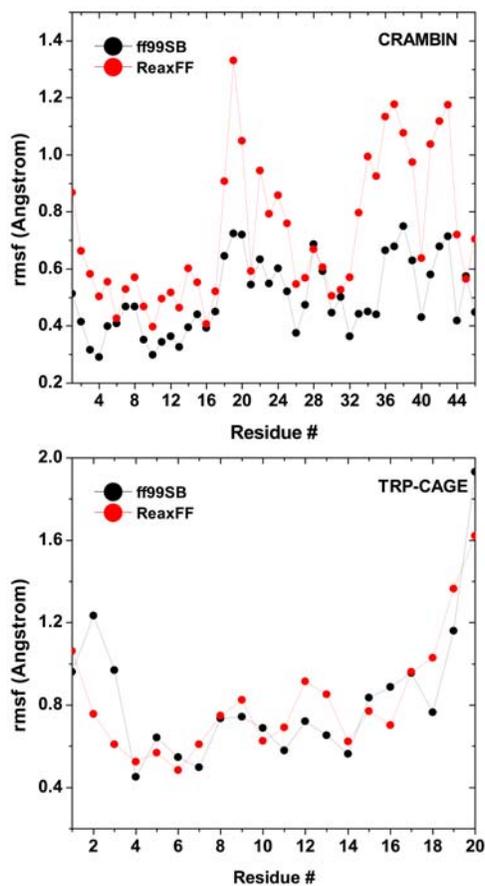


Figure 4.4.3: Root mean square fluctuation of the α carbons of Crambin (top), and TRP-cage (bottom) about their average coordinates as a function of the residue number.

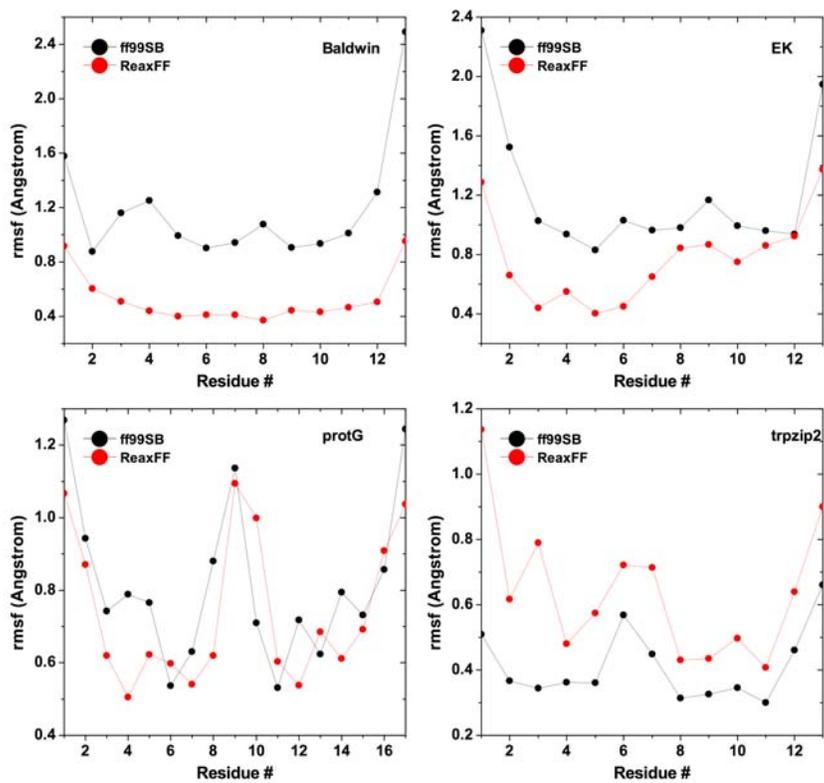


Figure 4.4.4: Root mean square fluctuation of the α carbons of the four oligopeptides: Baldwin (top left), EK (top right), ProtG (bottom left), and trpzip2 (bottom right) about their average coordinates as a function of the residue number.

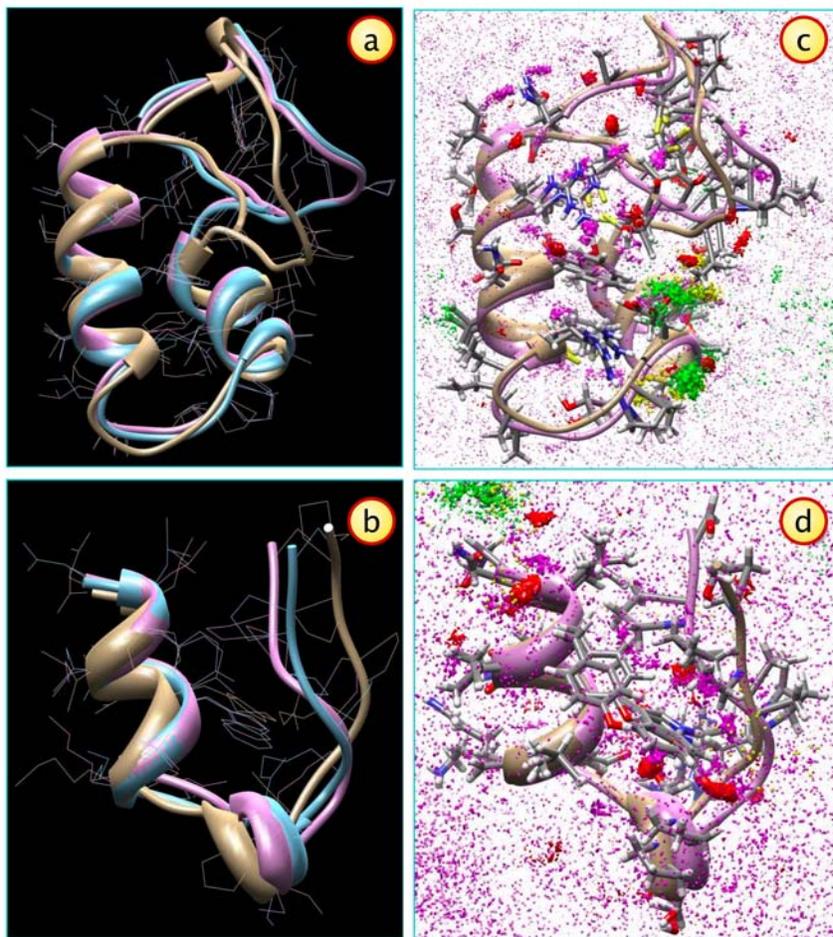


Figure 4.4.5: On the left hand side there are representative Structures extracted from ReaxFF and FF99SB simulations of a) Crambin, and b) TRP-cage. These geometries have been superimposed on the α carbons of the longer helical portion. Backbones are represented through tan, pink, and cyan ribbons which identify ReaxFF, FF99SB and native structures, respectively. On the right hand side the average structures of c) Crambin, and d) TRP-cage surrounded by 3D-density contour of water oxygen (red=ReaxFF, magenta=FF99SB), and Cl^- ions (green=ReaxFF, yellow=FF99SB).

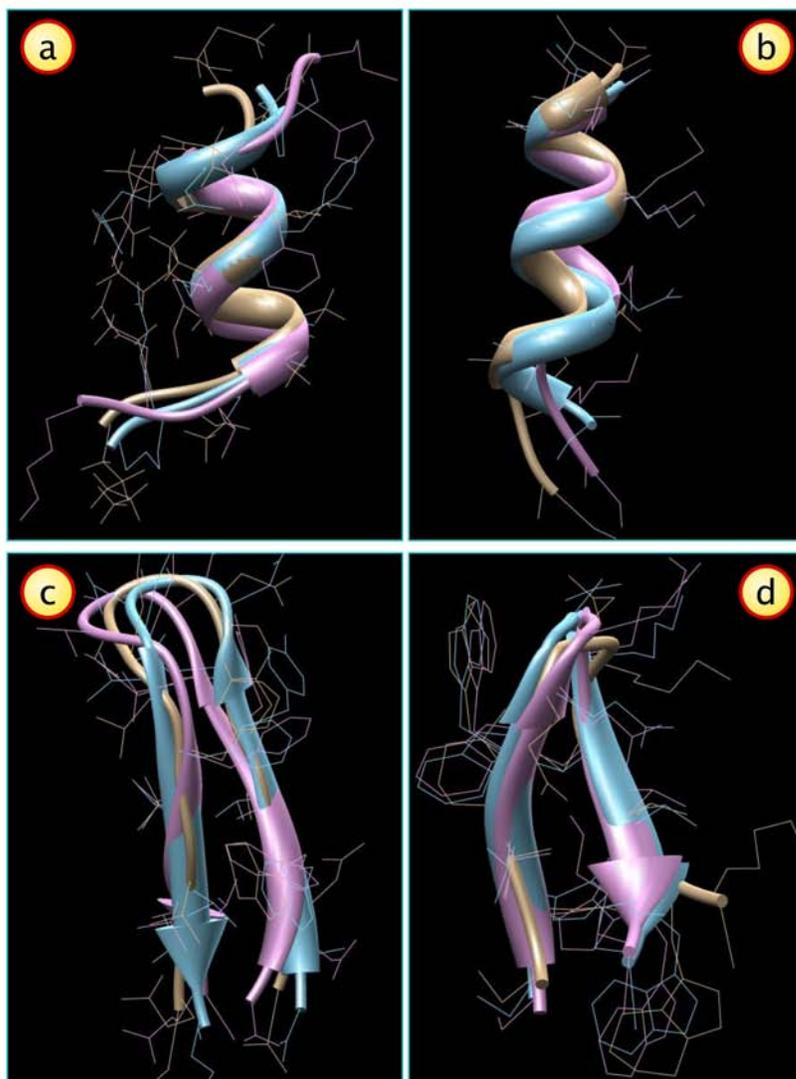


Figure 4.4.6: Representative Structures extracted from ReaxFF and FF99SB simulations of a) Baldwin, b) EK, c) ProtG, and d) trpzip2. These geometries have been superimposed on the α carbons of the longer helical portion. Backbones are represented through tan, pink, and cyan ribbons which identify ReaxFF, FF99SB and native structures, respectively.

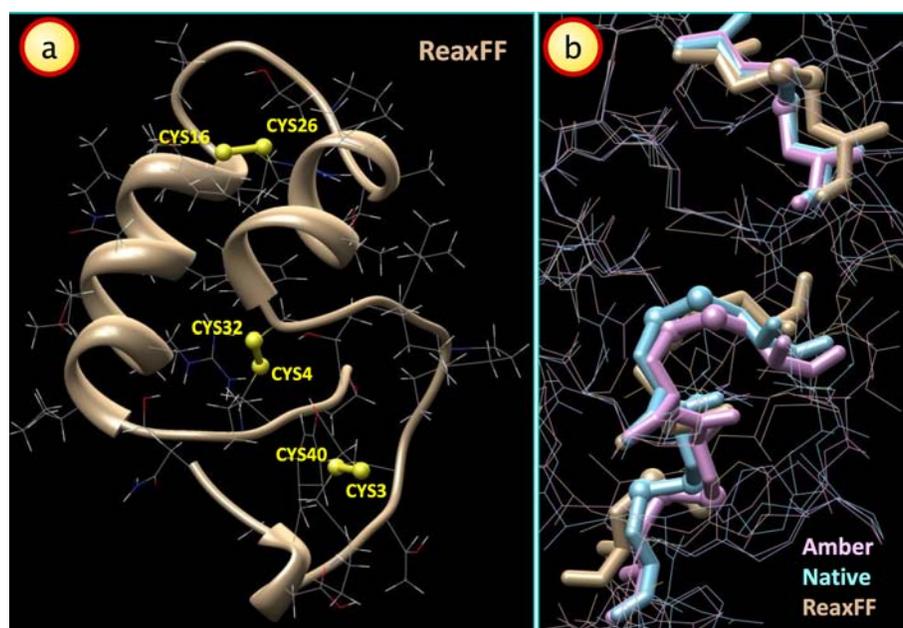


Figure 4.4.7: Representative Structures extracted from ReaxFF and FF99SB simulations of a) Crambin where the disulfide bridges are highlighted (yellow ball and stick models). On the right hand side b) the superimposed conformations of the native (cyan), ReaxFF (tan), and FF99SB (pink) structures in the region of the three cross-linkers are shown to underline the compactness of the chain arrangements and the coherent description of the two different force fields.

connections are present, figure 4.4.6 d).

Another descriptor that can give an idea of the overall dimension of the molecules is the radius of gyration R_{gyr} , which is defined here as the mass-weight root mean square distance of C_{α} from their common center of mass. The comparison of the values obtained with the two different force fields, Table 4.8, suggests that all the models are stable and do not undergo significant structural transitions which take them to other regions of the conformational space. However, through a quantitative analysis of the flexibility of the differ-

ent portions of both the proteins and the peptides, obtained by means of the root mean square fluctuations (RMSFs) of C_α atoms in relation to their time averaged positions (Figure 4.4.3 and 4.4.4), it can be noticed that the termini are more mobile (Figure 4.4.3). Even though the two trends are similar in each case, the ReaxFF description of Crambin presents greater fluctuations. As already mentioned, this is particularly evident in the range 18-20, which is the turn region connecting the two helices, and in the portions defined by residues 18-20, 22-25, 33-39, 41-43 which correspond to turn and β -sheet segments. On the contrary, both ReaxFF and FF99SB predicted similar the behavior for the **TRP-cage**. As far as the peptides are concerned, ReaxFF has the tendency to preserve helical arrangements (**Baldwin** and **EK** peptides), which are less stable in the FF99SB description, but shows major fluctuations when applied to the hairpin (**trpzip2**) due to the partial loss of intramolecular hydrogen bonds. This was perhaps induced by a different motion of the side chains of terminal residues which perturbed the backbone arrangement and as a consequence the intra-backbone connections.

For the secondary structure content, which was obtained using the DSSP method of Kabsch and Sander [122] and reported as percentage of each residue to adopt α -helix, 3_10 -helix, π -helix, parallel/antiparallel β -sheet and turn arrangements during the trajectories (Figures 4.4.8 and 4.4.9) it was observed that all models preserved their characteristic arrangements and the overall fold was well maintained. α -helices were adequately conserved by the two methods even though ReaxFF was more efficient than FF99SB in preserving the typical helical organization. On the contrary the percentage of anti parallel β -sheets was reduced and secondary structure elements belonging to different structural

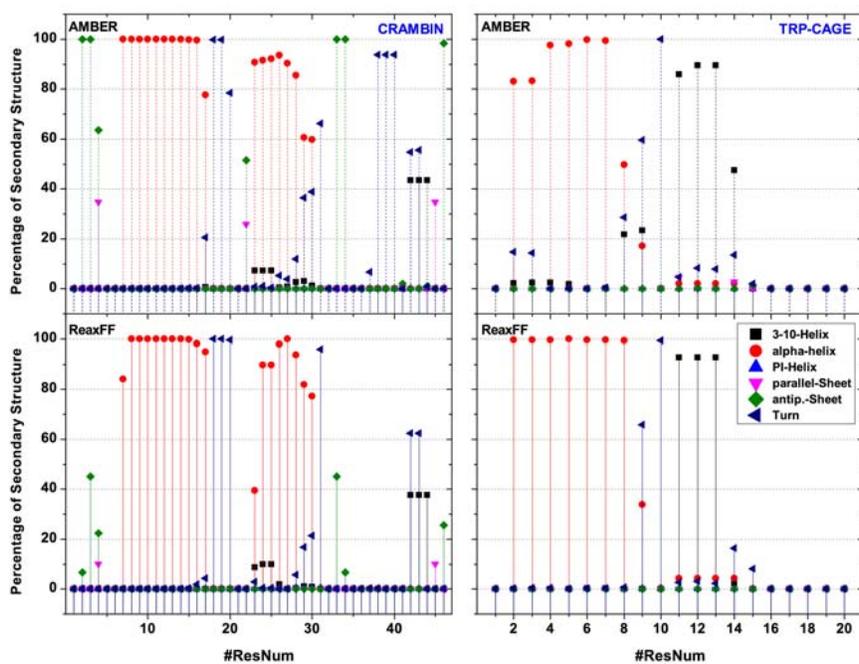


Figure 4.4.8: Percentage for each residue to adopt one of the secondary structure types over the course of the simulation. Crambin and TRP-cage proteins. Amber99SB (top) and ReaxFF (bottom) datas.

domains appeared. Although most of the β -sheet content remained in both models during the entire simulation time, FF99SB tends to conserve these secondary structure elements more than ReaxFF. A comparison of the number of persistent hydrogen bonds in **Crambin** simulations has been done. Persistent hydrogen bonds are defined as those having an occupancy (%occ) greater than 80% (where occupancy is defined as the number of frames with the hydrogen bond present divided by the total number of frames used for the analysis). In addition the two following conditions have been fulfilled: 1) a donor-acceptor distance lower than 3.0\AA , and 2) a H-donor-acceptor angle lower than 20 deg

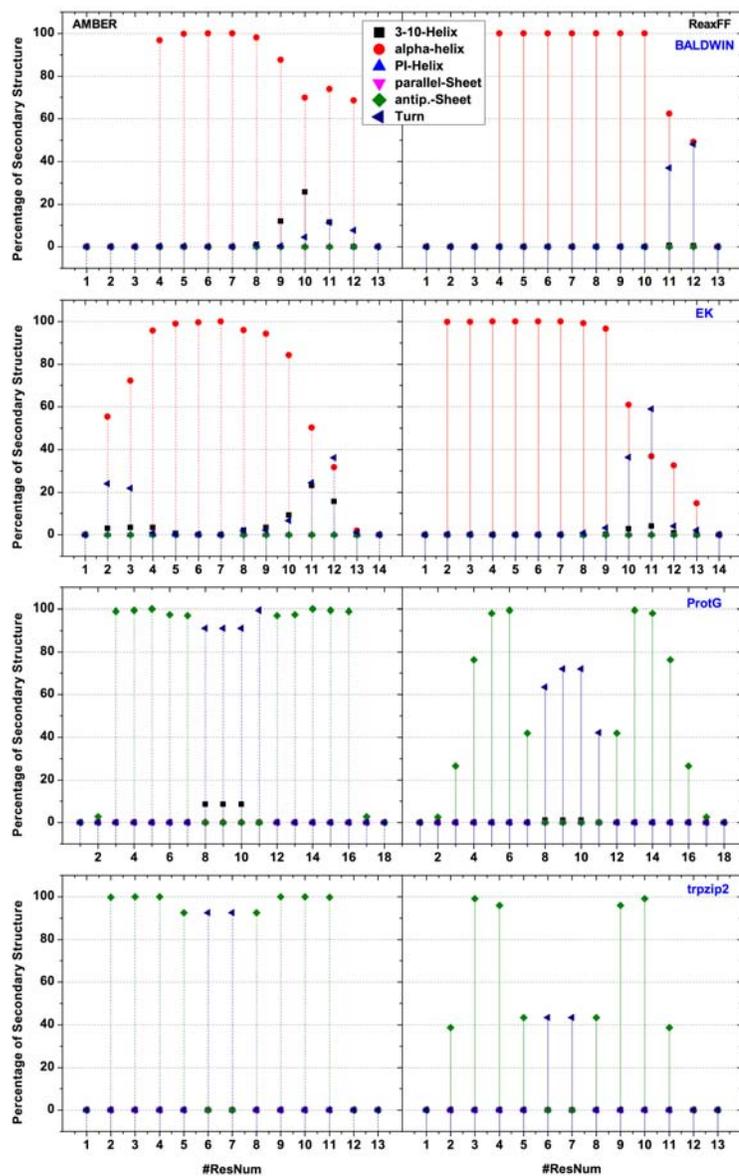


Figure 4.4.9: Percentage for each residue to adopt one of the secondary structure types over the course of the simulation for Baldwin, EK, ProtG, and TRP-cage oligopeptides. Amber99SB (left hand side) and ReaxFF (right hand side) datas.

[123]. The analysis shows that persistent hydrogen bonds are highly conserved even though FF99SB suffer from a reduction in the occupancy (about 1%) and a lengthening of the donor acceptor distances (of about 3Å) with respect to ReaxFF (where the average values are 95.3% and 2.67Å, respectively). In particular, the salt bridge between the guanidinium group of Arg10 and the carboxylate of Asn46 at the C terminus, which is important for the formation of the correct disulfide bonds and contributes to the compactness and stability of the structure [124][125], is present during the entire simulation time.

The TRP-cage exhibits a lower content of persistent hydrogen bonds in relation to **Crambin**. The H-bond between the Trp6 indole NH ϵ 1 and the Arg16 backbone carbonyl is established and well conserved in both models (%occ=86), whereas the salt bridge Asp9-Arg16, which provides an additional stabilization to the cage is formed only in the FF99SB description where its occupancy is about 92%, on average. On the contrary, in ReaxFF the percentage of occupancy obtained is lower than 50%, but the H-bond seems to be shifted to the nearby Ser14 residue, which becomes engaged in a more persistent interactions with both Asp9 and Arg16 residues (%occ>65%). For the four oligopeptides with potential salt bridges these interactions were established in all cases but they were less populated in the non-helical structures (**ProtG** and **trpzip2**). Even if less populated, these peptide conformations were maintained, suggesting that the association of the charged side chains were not crucial for the stability of the molecule, in agreement with Shell and co-workers [126]. Given the importance of water and counterions in stabilizing protein structures directly or through mediated interactions, it was interesting to analyze water and ions distributions obtained with ReaxFF for both **Crambin** and **TRP-cage** and

compare them with the other description (FF99SB). In line with data in the literature [125] it was found that water could link polar side chains at the protein surface and interconnect, indirectly, hydrophilic residues. A clear view of the hydration sites was obtained by calculating and visualizing, as three dimensional (3D) contour plots around an average geometry of the proteins, the spatial distribution functions (SDFs) of the water oxygen atoms. The isosurfaces, right hand side of Figure 4.4.8 d) and e), are the probability regions corresponding to different densities relative to the average bulk value. The red and green contours correspond to the highest probability of finding water oxygens and chlorine ions close to the protein according to the ReaxFF description, whereas magenta and yellow surfaces represent FF99SB data. The protein appears to be uniformly covered with discrete localized regions of higher water density. These hydration sites are similarly placed by the two force fields and seem located in proximity of Glu23, Arg10, Arg17 and Asn14 side chains. However, water molecules are also found engaged in hydrogen bonding interactions with the backbone atoms of the turn regions. Moreover, the possibility of deeper water insertion between the chains is further justified by the presence of small dense areas in the central portion of the protein. Instead, the chloride 3D-density areas appear located mainly in proximity to the regions explored by Glu23 and Arg17 side chains. TRP-cage solvation is described similarly by the two force fields. A detailed analysis indicates that the original Asp9-Arg16 salt bridge actively stabilize the structure and as a consequence its solvation is reduced. However, water molecules penetrate the cavity between the tryptophan side chain and protein backbone (red and magenta areas close to the center of the molecule) and solvate the core. Chloride 3D-density contours are predominantly positioned near

the N-terminus portion of the protein with the α -helix dipole moment creating a positive partial charge in that region of the helix.

4.5 Serine Protease catalysis, A ReaxFF study

4.5.1 Introduction

Hydrolases are a group of enzymes that catalyze cleavage of chemical bonds by reaction with water (i.e. hydrolysis). Since they are involved in many biological processes their function is of importance in a number of pathology including blood clotting [127], digestion [128], nervous system signaling [129], inflammation [130], and cancer [131][132][133] diseases.

Recently, they have also become useful tools in synthetic organic chemistry where they have been shown to act also on a wide range of synthetic esters and amides [134]. Their success in this field stems from the fact that they have a broad substrate specificity, including various synthetic intermediates, and that they show high stereoselectivity towards both natural and non-natural substrate. The proteases are a subgroup of hydrolases, they are involved in peptide-bond cleavage by reaction with water so to hydrolyze proteins to smaller peptides and then to aminoacids. Among all the proteases the serine proteases forms the largest group. A serine residue in the active site gives the name to these enzymes. This serine residue plays the important role of carrying out the nucleophilic attack onto the carbonyl group of the incoming substrate. The same residue is part of one of the main characteristics of this enzyme class, namely the catalytic triad [135]. The catalytic triad consists of a set of three residues that have developed specifically for the hydrolysis of peptide bonds.

The catalytic triad is formed by an aspartate, an histidine, and the serine residue previously mentioned. These three residues have been present since two billions years [136] with the most primitive organism appeared on earth to provide functionality for digestion and metabolism of their own proteins. From this common ancestor, the catalytic triad machinery has been preserved and adapted into four evolutionary branches of serine proteases: Chymotrypsin, Subtilisin, Carboxypeptidases, and CLP Protease.

Chymotrypsin-like proteases are abundant in nature and very well studied in literature. For these reasons we selected fusarium oxysporum trypsin to test the capability of the developed Prot-ReaxFF force-field. As well as hydrolases trypsins are of fundamental importance in many of physiological processes mentioned previously for the hydrolase family. Lack of expression or inhibition of such enzyme has dramatic consequences on digestion , hemostasis, apoptosis, signal transduction, reproduction, and immune response systems [137][138][139][140] [141]. Protease cascades are mainly involved in blood coagulation and fibrinolysis [142][143][144], while other protein cascades are responsible for development, matrix remodeling [145], adipocyte differentiation [146], and wound healing [147].

Following the classification by Kraut et al. [148], the general motives of chymotrypsin-like proteases are the presence of a catalytic site, a substrate recognition site (the active site cleft) and the zymogen activation domain. The core of the catalysis machinery is carried out by the interplay of the three residues of the catalytic triad: Ser195, His57 and Asp102 (residues's numbering according to X-Ray structure of Fusarium Oxysporum Trypsin [149]). The three residues are located on one side of the active site cleft as it can be seen on figure

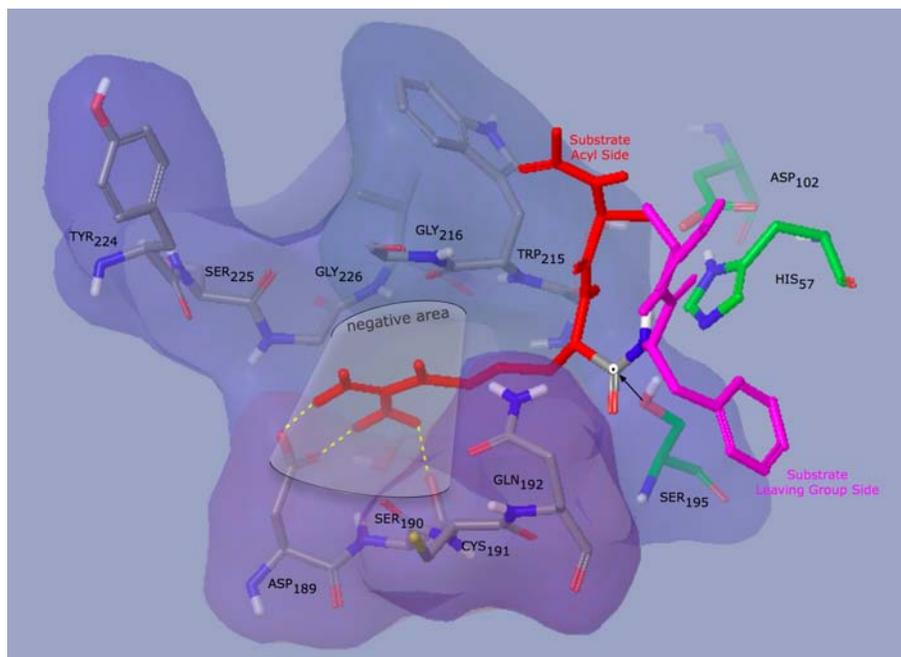


Figure 4.5.1: Graphical representation of S1 site of *Fusarium Oxysporum* Trypsin. The catalytic triad Asp102, His57 and Ser195 is represented by green colored carbon. Red atoms forms the acylic branch of the substrate, while magenta colored atoms form the leaving group branch. The nucleophilic attack by Ser195 towards the carbonylic carbon of the arginine residue of the substrate is indicated by an arrow. The S1 site is formed by residues 189-192, 214-216, 224-228. The hydrophobic pocket negatively charged that accounts for the recognition of the substrate is formed by Asp189, Ser190, Gly226, and Gly216.

4.5.1. The structure and the relative orientation of the triad is shown in figure 4.5.2. His57 is in the less favorable $N\delta 1$ -H tautomer due to the strong hydrogen bond to Asp102. Hydrogen bonds are observed between $N\delta 1$ -H of His57 and $O\delta 1$ of Asp102, and between the OH of Ser195 and the $N\epsilon 2$ -H of His57. The latter hydrogen bond is connected with reactivity [150]. The His57-Asp102 hydrogen bond has the syn orientation [151] relative to the carboxylate, the hydrogen bond is formed with the more basic electron pair. There is also an hydrogen bond formed by the side chain hydroxy group of Ser214 and $O\delta 1$ of Asp102. Hydrogen bonds are also observed between $O\delta 2$ of Asp102 and the main chain NHs of His57. The hydrogen bond network that couples His57 and Asp102 is believed to be the reason of their relative orientation in the active enzyme. An interaction between the $C\epsilon 1$ -H of His57 and the main chain carbonyl of Ser214 is also observed. The Hydrogen bond network of the trypsin object of this study, the X-Ray structure of *Fusarium Oxysporum* Trypsin [149], is reported as an example in figure 4.5.2. The backbone of Ser195 together with the main chain NH group of Asp194 forms what is called the oxyanion hole: a pocket lined with positive charge that activates the carbonyl group for nucleophilic attack. Furthermore, it stabilizes the negative charge of lone pair on the carbonyl oxygen of the substrate once the nucleophile attack has been carried out by Ser195. The oxyanion hole of *Fusarium Oxysporum* Trypsin [149] is shown in figure 4.5.3.

The specificity of trypsin proteases are usually interpreted in terms of substrate recognition sites. Following the nomenclature of Schetcher and Berger [152] we would say in terms of S1-P1 interactions. The S1 site is around Ser195 and it forms a scabbard for the acyl side of the substrate. It is formed by residue 189-192, 214-216, and 224-228 and it is shown in figure 4.5.1. The four

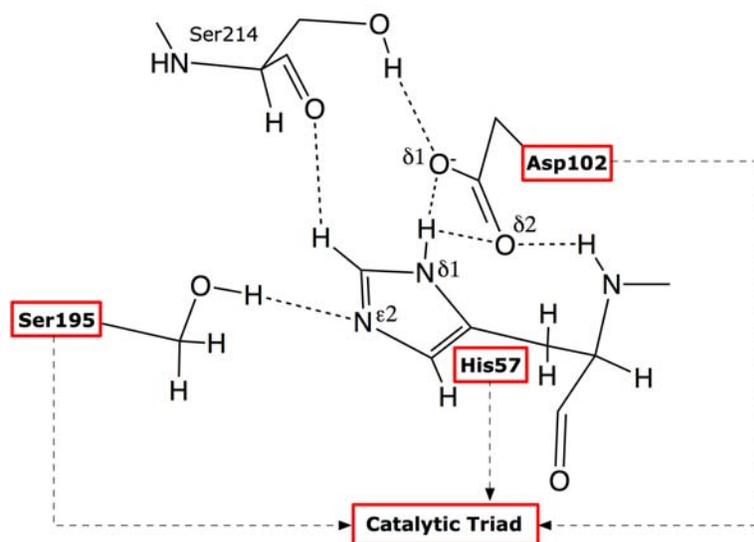


Figure 4.5.2: Hydrogen Bonding in the catalytic triad of the Michaelis Complex of Fusarium Oxysporum Trypsin.

residues Asp189, Ser190, Gly216, and Gly226 forms the negatively charged hydrophobic pocket that accounts for specific interaction with Arginine or Lysine at P1. The predilection of trypsin for cleaving the carboxyl side of Arginine and Lysine residues of the substrate is a very well known feature you can find in biochemistry textbooks [19][20] and articles of scientific journals [153][154]. Residues 214-216 are part of an antiparallel β -sheet that has interaction with P1, P2, and P3 sites. It is worth noticing that a different conformation of the Gly216 and the relative hydrogen bond strength in that area differentiates chymotrypsin, trypsin, and elastase [155]. It has been observed that in many trypsin inhibitor that the P1 to P3 area assumes an antiparallel β -sheet called canonical conformation [156]. It has been conjectured that the canonical conformation of the inhibitor is somewhat related to the lack of trypsin catalysis

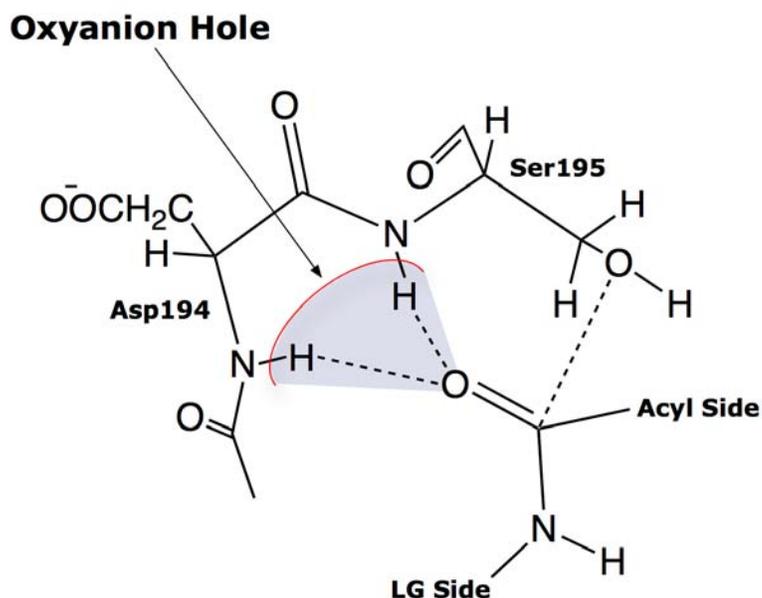


Figure 4.5.3: Oxyanion hole of the Michaelis Complex of Fusarium Oxysporum Trypsin.

because of the heavy hydrogen bond network formed between S1 to S3 area and P1 to P3 area. On the other hand it has been observed also that disrupting the hydrogen bond network between S1 to S3 area and P1 to P3 area has little effect on catalysis [157]. In addition values of K_d of inhibitor and K_{cat}/K_m for hydrolysis of analogues substrate show a strong similarity, so it means that presumably also the substrate assumes the canonical conformation as well as the inhibitor [158][159].

Through natural selection the serine proteases have evolved to have a chemical kinetic enhancement of 10^{10} order of magnitude with respect to the analogous reaction in solution. Peptide bonds are highly stable in solution due to the resonance obtained as a result of electron donation from the amide nitrogen

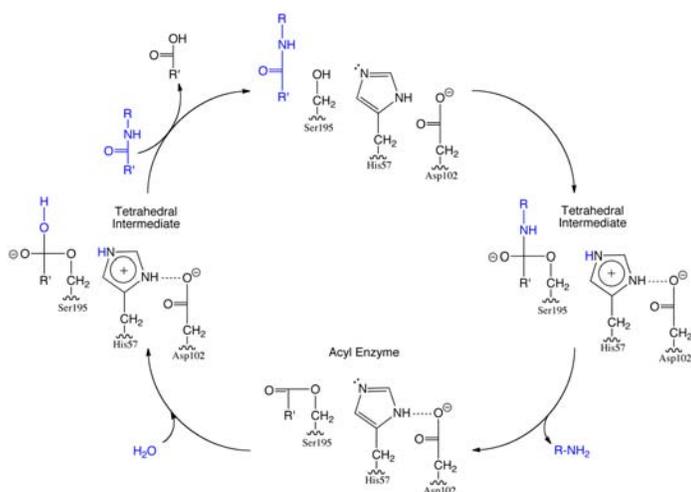


Figure 4.5.4: Catalytic cycle suggested for serine proteases.

to the carbonyl. Proteases interacts with the amide bond as a result of the amide carbonyl carbon interaction with Ser195 hydroxyl acting as an acid. The resonance that stabilizes the bond is perturbed by the distortion induced, so the bond becomes weaker. The hydrogen transfer to the amide nitrogen leads to the expulsion because it creates a better leaving group. The accepted catalytic cycle for the reaction mechanism of Serine Proteases is illustrated in figure 4.5.4 and it has been discussed in many articles in the literature [148][160][161]. The first step is represented by the formation of a Michaelis-Menten complex formed between Trypsin and the polypeptide substrate. Then, the hydroxyl group of Ser195 undergoes a nucleophilic attack onto the carbonyl carbon of the carboxyl part of an arginine (or lysine) residues while at the same time an hydrogen is transferred from the hydroxyl group of Ser195 to His57 N ϵ 2. After the nucleophilic attack a covalent bond is formed between protein and substrate.

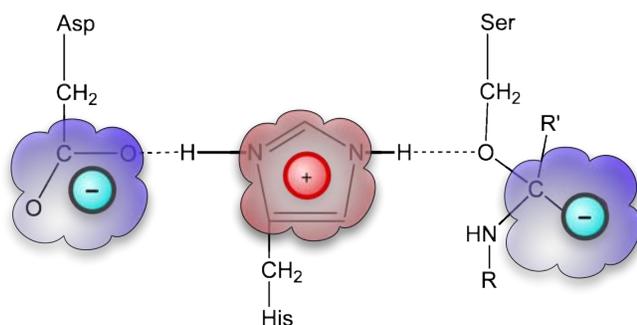


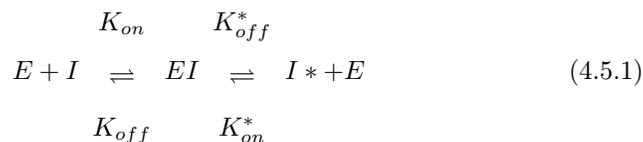
Figure 4.5.5: Unstable equilibrium and system mobility: the alignment of the catalytic triad and the relative electronic instability asset of the tetrahedral-intermediate. The tendency to motion of the histidine increases when the adduct is formed. Due to the electronic attraction the histidine can easily reach the aspartate or the adduct part as in a situation of unstable equilibrium.

The carbonyl oxygen of the substrate assumes a tetrahedral geometry with a negative charge located on the carbonyl oxygen of the substrate, which is stabilized by the positive charge of the oxyanion hole. The hydrogen transferred to His57 charges positively the 5 membered ring. The increased histidine motion as a consequence of the tetrahedral adduct leads to a lower curvature of the tetrahedral-intermediate potential well compare to the well's curvature either of the initial state and of the acyl-intermediate state. The histidine mobility is the results of the interplay between different phenomena: i) after the first hydrogen transfer the hydrogen bond of the histidine got broken giving more motional freedom to the residue, ii) the alignment of the three residues increases the mobility destabilizing the structure. The unstable equilibrium as a consequence of the charge separation is emphasized in figure 4.5.5. This is the so called one-proton transfer mechanism, 1H NMR gives experimental proof of it [162][163][164] as well as neutron diffraction experiments [165].

Proteases carry out a large number of hydrolytic reactions both intra- and

extracellularly [166], so they need to be properly controlled because their expression may be the cause of large damages in presence of proteins. There two leverages through which a living organism can control proteases action: regulation of the expression [167], and by degradation the enzymes when it is not anymore needed. Another kind of regulation -usefull also in pharmacological science- is the inhibition of the proteolytic activity through inhibition by other protein [168]. The largest group of protein inhibitors are canonical inhibitors and their standard mechanism of inhibition is well known since more than twenty years [169]. BPTI are a class of canonical inhibitors of the protein family Kunitz-type serine protease inhibitors [170]. The inhibitor forms a tight non-covalent interaction similar to the enzyme-substrate Michaelis complex [171][172]. The solvent exposed loop of the inhibitor contains the reactive site $P_1 - P_1'$ [173] and it is complementary to the active site pocket of the enzyme. The formation of the stable complex avoids any further possibility of proteolytic capability of the trypsin. The reactive site of the inhibitor can be selectively hydrolyzed by the enzyme, but the equilibrium value of this cleavage is often close to 1 at neutral pH [174].

The interaction between enzyme and inhibitor can be presented in a simplified form as a hydrolysis/resynthesis reaction of the $P_1 - P_1'$ reactive-site peptide bond:



where

- E is the enzyme,
- I is the inhibitor,
- I^* is the cleaved inhibitor,
- EI is the stable complex,
- K_{on} and K_{on}^* are association rate constants,
- K_{off} and K_{off}^* are respective the dissociation rate constants of the complex.

In the following study ReaxFF with Prot-ReaxFF force field has been applied to a catalytic active substrate (NMA-ALA-PHE-ARG-ALA-ACE) [175] and a Kunitz-BPTI inhibitor [176] using *Fusarium Oxysporum* Trypsin as receptor. The aim of the study is to check whether the ReaxFF method is properly parameterized and able to give a qualitative different description of the two different cases. To our knowledge this is the first real application of ReaxFF in enzyme catalysis modeling.

4.5.2 Computational Details

AMBER MD Simulation. In our simulation the initial geometry for Trypsin was obtained from the X-Ray crystal structure (pdb code 1FN8) of *Fusarium Oxysporum* Trypsin [149]. The structure was initially determined at 1.07Å resolution and 283K and successively refined at a resolution of 0.81Å at 100K. The initial coordinates have been downloaded from the RCSB Protein Data-bank (<http://www.rcsb.org/>) [177]. The trypsin substrate found in the X-Ray structure is a tripeptide (Gly-Ala-Arg) with arginine residue correctly put in its pocket with relative hydrogen bond interactions with S1 residues, namely Ser190 and Asp189. This tripeptide has been modified to obtain a tetrapeptide with N-Methylacetamide and acetyl terminations. In what follows this tetrapeptide will be called simply “substrate”. Its residue sequence is: N-Methylacetamide (NMA), Alanine, Phenilanine, Arginine, Alanine, and the acetyl termination (ACE). This model substrate mimicks a similar catalitically active substrate of Trypsin ($K_{cat} = 52.7 s^{-1}$ and $K_{cat}/K_M = 2.4 \cdot 10^5 L/M \cdot s$) [175].

Additional residues have been added to the tripeptide of the X-Ray structure. This has been done adding residues one by one freezing the entire structure while leaving only the added residues free to reorient itself by means of OPLS-2005 force field [178] and macromodel software package (MacroModel, version 9.9, Schrödinger, LLC, New York, NY, 2012.).

A simulation box of TIP3P water molecules [118] has been created around the receptor+substrate system. The cubic box is of $274'625 \text{Å}^3$ of volume and it has been filled with 7951 water molecules. Five chloride ions have been added to neutralize the charge of the entire system.

The inhibitor protein BPTI has been instead obtained by an X-Ray structure

of Marquart et al. [176] determined to 1.9Å resolution (pdb code 2PTC). The inhibitor consists of 58 aminoacid residues. A lysine residue is in the right position for the nucleophilic attack. Lysine carbonyl carbon has a distance of only 2.675Å from the oxygen of the catalysis serine of the receptor. Catalytic triad residues of 2PTC and 1FN8 X-Ray structure do not show large differences in their positions. Superimposing C_α and side chains carbons of the catalytic triad of 1FN8 and 2PTC, it accounts for an RMSD of only 0.17Å with a maximum difference of 0.22Å. The shapes of the two receptors after superposition of the triad show an high graphical similarity. The inhibitor has been cut from 2PTC X-Ray structure and superimposed and placed instead of the substrate of the 1FN8 X-Ray structure. Steric clashes have been removed using the same strategy previously described for the substrate. A simulation cubic box of 305'620Å³ of volume with 8815 TIP3P water molecules [118] has been created around the complex between the protein and the inhibitor; ten chloride ions have been added to neutralize the system. MD simulations were performed by AMBER 10 [116] suite of programs with FF99SB [117] force-field. The final trajectories were analyzed and visualized by means of VMD [179].

After a minimization of 5'000 steps with :i) a short range 12Å cutoff for non bonded interaction, ii) restraints on heavy atoms (force constant 100Kcal/mol) to allow a re-orientation of hydrogen atoms only, a relaxation protocol has been carried out before starting the productive MD simulations. The energy minimized models have been gradually heated from 0 K to 300K along a 100ps restrained dynamic. The initial restraint on heavy atoms has been reduced from 100Kcal/mol to 25Kcal/mol and then decreased further in the production runs. Details of the four steps MD simulations of the relaxation protocol are presented

Substrate MD Relaxation Protocol	Ensemble	Simulation Time [ps]	Simulation steps	Force Constants [Kcal/mol] ^[1]	Temperature [K]	τ_p [ps] ^[2]	Cut-off [Å] ^[3]
step 2	NVT	100	50'000	100.0	10 to 300	-	12
step 3	NPT	100	50'000	100.0	300	0.01	12
step 4	NPT	100	50'000	100.0	300	1.0	12
step 5	NPT	100	50'000	25.0	300	1.0	12

^[1]Constraints are only on the heavy atoms of the receptor. The substrate structure is kept without any constraints. ^[2]Barostat Relaxation Time Constant. ^[3]Short range Cut-off for non-bonded interactions.

Inhibitor MD Relaxation Protocol	Ensemble	Simulation Time [ps]	Simulation steps	Force Constants [Kcal/mol]	Temperature [K]	τ_p [ps] ^[1]	Cut-off [Å] ^[2]
step 2	NVT	100	50'000	100 ^[3]	10 to 300	-	12
step 3	NPT	100	50'000	100 ^[3]	300	0.01	12
step 4	NPT	100	50'000	100 ^[3]	300	1.0	12
step 5	NPT	100	50'000	25 ^[4]	300	1.0	12

^[1]Barostat Relaxation Time Constant. ^[2]Short range Cut-off for non-bonded interactions. ^[3]Force Constant Constraints on the heavy atoms. ^[4]Force Constant Constraints on the heavy atoms of the receptor.

Table 4.9: Relaxation protocol adopted to prepare the system to the productive AMBER MD simulation runs. Top: details of the relaxation protocol for receptor+substrate; bottom: details of the relaxation protocol for receptor+inhibitor.

in table 4.9.

Productive simulations of 2ns (10^6 steps) have been carried out with weak force constant on heavy atoms (5Kcal/mol). System configurations were sampled every 1ps over the last 1000ps of the simulation trajectory. A 12Å cutoff for short range non-bonded interactions was employed in combination with particle mesh Ewald method [119] to treat long-range electrostatic interactions. To adjust solvent density to the bulk value a weak pressure coupling scheme [120] has been used, while the temperature is instead controlled by Andersen method [180]. Three dimensional periodic boundary conditions were applied for all the simulations.

ReaxFF MD Simulation. The geometries obtained at the end of the equilibration phase carried out with non-reactive AMBER molecular dynamics have

Receptor + Inhibitor Frame #	RMSD* C _α backbone	Receptor + Inhibitor Frame #	RMSD* C _α backbone
999	0.2650	982	0.2647
998	0.2689	981	0.2607
997	0.2791	980	0.2804
996	0.2678	979	0.2601
995	0.2691	978	0.2631
994	0.2572	977	0.2607
993	0.2653	976	0.2725
992	0.2610	975	0.2694
991	0.2642	974	0.2650
990	0.2592	973	0.2617
989	0.2711	972	0.2623
988	0.2651	971	0.2633
987	0.2637	970	0.2632
986	0.2646	969	0.2636
985	0.2546	968	0.2573
984	0.2656	967	0.2696
983	0.2714	966	0.2688

*Reference structure of the last frame of Amber Simulation with receptor with substrate Nma-Ala-Phe-Arg-Ala-Ace

Table 4.10: Table Amber Inhibitor comparison with substrate last frame

been used as starting structures of the reactive simulations. Protein conformations used for the Prot-ReaxFF study are close to the X-Ray geometries because of the restraints applied during the non reactive simulations. The receptor X-Ray structure is the same for the two cases studied (receptor+substrate and receptor+inhibitor). A comparison with backbone C_α RMSD is here presented to validate the choice of the last frame of the two AMBER MD productive simulations to start the ReaxFF study. As it can be seen from table 4.10 the last 34 frames of the receptor+inhibitor AMBER MD simulation are all similar for what concerns their RMSD compare to the last frame of the receptor+substrate AMBER MD simulation. The average RMSD is only 0.26Å and the variations are centered around it.

ReaxFF simulations have been performed in the NVT ensemble using Berendsen thermostat [120] with a relaxation constant of 0.1ps. Equations of motions

Receptor + Inhibitor Frame #	Amber ΔE_{pot} [Kcal/mol]	ProtReaxFF ΔE_{pot} [Kcal/mol]	Amber Backbone RMSD	ProtReaxFF Backbone RMSD
999	0 ^[1]	0 ^[1]	0 ^[1]	0 ^[1]
998	-77.13	14.62	0.3304	0.2912
997	-23.31	83.33	0.3687	0.3328
995	-157.17	-237.74	0.3641	0.3244
991	-184.10	-320.97	0.4354	0.4086
988	-199.61	-281.82	0.4653	0.4370
976	32.70	-290.81	0.6718	0.6545
971	-224.37	-298.96	0.6833	0.6679
Receptor + Inhibitor Frame #	Amber vs ProtReaxFF Backbone RMSD	Amber Inhibitor ^[2] Backbone RMSD	ProtReaxFF Inhibitor ^[2] Backbone RMSD	Amber vs ProtReaxFF Inhibitor ^[2] Backbone RMSD
999	0.1687	0 ^[1]	0 ^[1]	0.1715
998	0.1667	0.6821	0.7111	0.1712
997	0.1655	0.6905	0.7139	0.1699
995	0.1608	0.7612	0.7754	0.1689
991	0.1648	0.9660	0.9847	0.1725
988	0.1639	1.0038	1.0201	0.1670
976	0.1676	1.2796	1.2875	0.1763
971	0.1664	1.3039	1.3097	0.1659

^[1] The Reference is frame #999, the last frame of the Amber simulation of receptor + inhibitor; ^[2] Only residues of the inhibitor are included in the calculation of backbone RMSD

Table 4.11: Table Amber simulation Inhibitor vs ReaxFF simulation

were solved with leap-frog Verlet algorithm [121] and the time step was set to 0.25fs. The ReaxFF version incorporated into the Amsterdam Density Functional (ADF) program [181][182] has been used for all the reactive dynamics simulations.

before performing NVT-MD simulations the two systems have been energy minimized. These energy minimizations consist of 1000 steps simulations at T=0K. They have been carried out to adapt the system to the new Prot-ReaxFF parameters. Some frame of the inhibitor+receptor AMBER simulation have been minimized with this procedure. From table 4.11 it can be evinced that Prot-ReaxFF new parameters do not change the information contained in the AMBER trajectory. Comparing directly AMBER and ReaxFF backbone of the receptor we have pretty much constant values with average RMSD of 0.16Å and a maximum difference between RMSD of only 0.04Å.

In the AMBER simulation the backbone of the inhibitor has been left wi-

without any constraints on the heavy atoms. As it can be seen from table 4.11 the larger movements of the inhibitor backbone is also caught by Prot-ReaxFF description. From a direct comparison between inhibitor backbone atoms of AMBER and Prot-ReaxFF force fields, we can see a satisfactorily similar descriptions with an average RMSD of of 0.17\AA and data centered around the average value.

The system temperature was raised gradually with 12.5ps MD-NVT simulations, and successively equilibrated at $T=300\text{K}$ with 3.75ps MD simulations. Scan MD-NVT simulations have been instead carried out to study the reaction that goes from the Complex formed (Michaelis complex in the case of the substrate) to the First Tetrahedral intermediate of the catalytic cycle. To reduce either the computational effort of the scan simulations and energy artifacts due to the large number of water molecules we decided to freeze all the atoms that are 5\AA away from the catalytic triad and substrate residues. Residues not frozen are presented in figure 4.5.6. The shell of not frozen water within 5\AA from the active site consists of 69 water molecules for receptor+substrate MD-NVT simulations, and of 55 water molecules for the inhibitor+receptor MD-NVT simulations.

Before to carry out studies with fixed water, MD-NVT simulations of 1.25ps at $T=50\text{K}$ have been carried out freezing all the residues just to optimize water molecules positions.

The partly frozen MD-NVT simulations have been done at $T=50\text{K}$ with distance constraints. As a matter of convenience more detailed informations of the calculations regarding this part of the study are given in the Results and

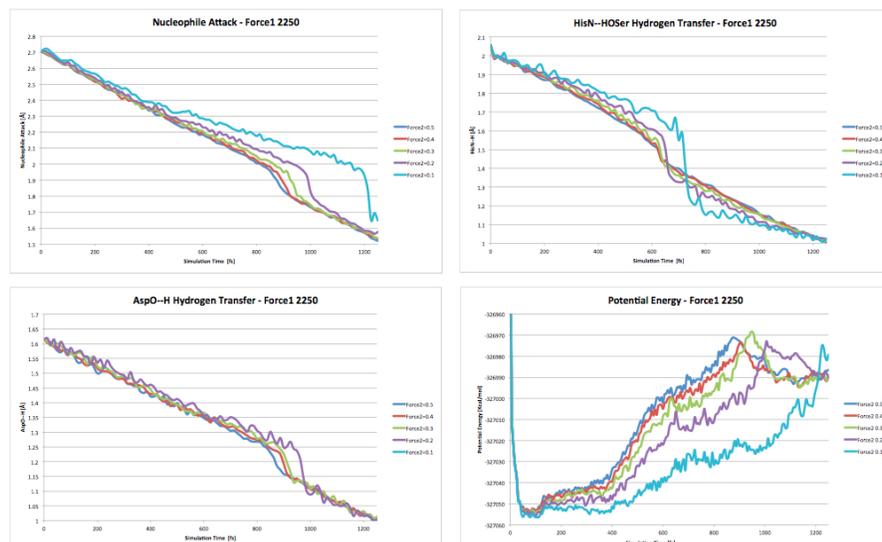


Figure 4.5.8: Force Constant Study: Force1 = 2250 and Force2 in the range 0.5→0.1. Top left: Nucelophile attack of the Serine with simulation time (driver1); Top Right: Hydrogen Transfer between Serine and Histidine (driver2) with simulation time; Bottom left: Hydrogen Transfer between Histidine and Aspartate (driver3) with simulation time; Bottom right: Potential Energy Profiles with simulation time.

with Force2 fixed to 0.5 and a range of Force1 between 2250 and 500 (Results shown in figure 4.5.8), the case with a fixed Force1 to 2250 and a variable Force2 in a range between 0.5 and 0.1 (shown in figure 4.5.9).

There is a correlation between low force constant and the reaction profile. With Force1 below 1750 and Force2 below 0.3 the atoms are jumping from one position to another and the transition state area is becoming less and less described because we have a shift of the transition state peaks towards the

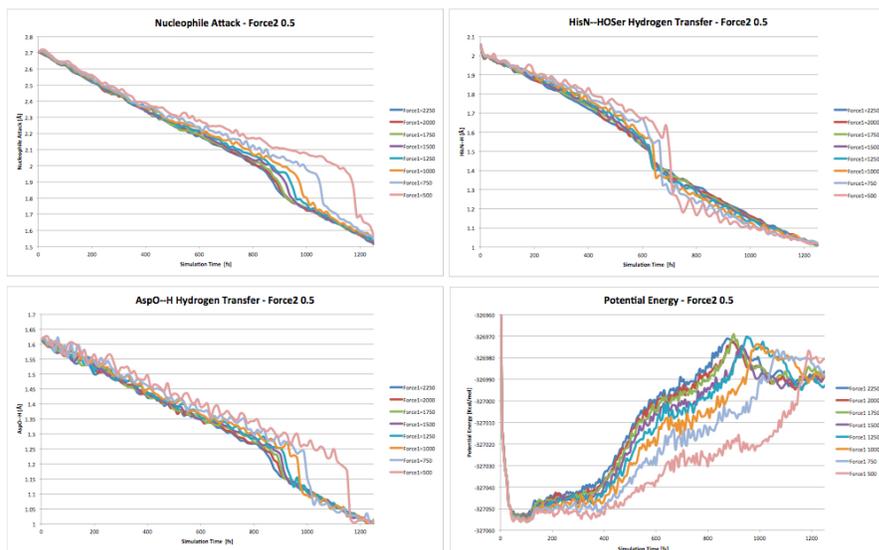


Figure 4.5.9: Force Constant Study: Force2 = 0:5 and Force1 in the range 2250→500. Top left: Nucelophile attack of the Serine with simulation time (driver1); Top Right: Hydrogen Transfer between Serine and Histidine (driver2) with simulation time; Bottom left: Hydrogen Transfer between Histidine and Aspartate (driver3) with simulation time; Bottom right: Potential Energy Profiles with simulation time.

end of the simulation. Apart from the atom jumps, the three drivers have not the same oscillating behavior in the two figures, in particular the hydrogen transfer between Histidine and Aspartate. To avoid both excessive oscillation and discontinuity we think the right bias for Force1 is between 2500 and 2000, while Force2 - the most sensible term - must be put it at 0.5 in all atom transfers.

4.5.3 Results and Discussion

Prot-ReaxFF force field has been used to study the different behavior of the catalytic active substrate (NMA-ALA-PHE-ARG-ALA-ACE) [175] and a Kunitz-BPTI inhibitor [176] both at the level of the formation of the complex and while following a reaction pathway to the first tetrahedral intermediate, the step that presumably is the rate determining of the catalysis [148][160][161]. Both systems have been partly frozen as described in the Computational Details section. Reactions have a different time scale compare to protein movements, therefore all the MD simulations described are carried out at $T=50\text{K}$. Structures of the complex guesses for the two systems have been isolated through a 3.75ps constrained simulation. The distance between Ser190 oxygen and the carbonyl carbon of either substrate (main chain carbonyl carbon of Arg227) and inhibitor (main chain carbonyl carbon of Lys239) have been constrained to 3.0\AA . The structures represent the isolated Michaelis complex in the case of the substrate, and a complex between inhibitor and receptor in the case of the inhibitor. Angle constraints have been used to create a correct orientation of the three triad residues for the hydrogen transfers between i) Ser190 and His54 and ii) His54 and Asp84. In figure 4.5.10 it is shown the numbering of the atoms of the triad as it is in the input files used by ReaxFF (files in biograften format are reported in appendix B). Angles $(C_{555} - N_{554} - H_{2456})$, $(C_{552} - N_{554} - H_{2456})$, $(C_{552} - N_{550} - H_{551})$, $(C_{549} - N_{550} - H_{551})$ are put to 130° . Angles $(C_{554} - N_{2456} - H_{2455})$, $(C_{550} - N_{551} - H_{1155})$ are put to 175° .

The result of such simulations have been used for a fast 1.25ps NVT-MD scan simulation acting simultaneously on the three scan drivers, namely the nucleophilic attack distance, the distance between N_{554} and H_{2456} (hydrogen transfer

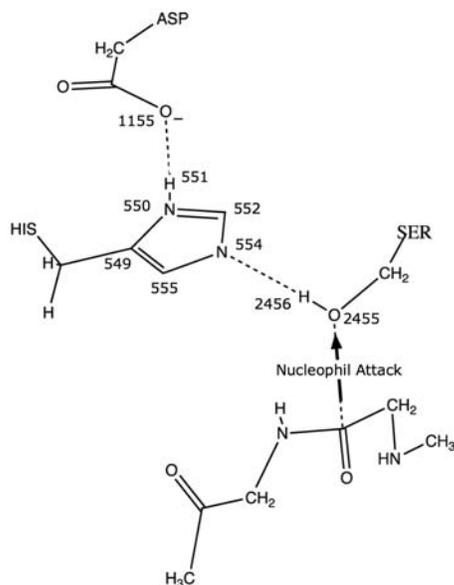


Figure 4.5.10: Triad residues numbering as in the input files of ReaxFF simulations (geo files in Appendix B).

from the serine to the histidine $N\epsilon_2$ nitrogen), and distance between O_{1155} and H_{551} (hydrogen transfer from histidine to the $O\delta_1$ oxygen). The result is shown as energetic profiles in figure 4.5.11. The energy of the two diagrams reported are plotted against the nucleophile attack distance for both substrate+receptor scan (labelled in the diagram as substrate ReaxFF Reactive Simulation), and inhibitor+receptor scan (labelled in the diagram as inhibitor ReaxFF Reactive Simulation). The last frames of the two simulations (nucleophilic attack equals to 1.5\AA) have been further equilibrated with a 3.75ps MD-NVT ($T=50\text{K}$) simulation. The equilibrated structures obtained at the end mimick what we call the tetrahedral intermediates. The variation of the energy with respect of the input structure of the two scan simulations are respectively 26.99Kcal/mol and

-16.08 Kcal/mol for receptor+substrate and receptor + inhibitor. The shape of the energy profile of a ReaxFF scan MD simulation is not a proper way to model and isolate transition state structures, anyway it is worth to notice that the trajectory profile of the receptor+substrate (diagram a) of figure 4.5.11), around the area where presumably the transition states probably (in the range 2.2/2.0Å) have average energies that are 26.8 Kcal/mol higher compare to the inhibitor+receptor simulation (diagram b) of figure 4.5.11). This suggests a larger stabilization of the energy due to inhibitor receptor interactions throughout the reaction coordinate chosen.

Inhibitors Kunitz-BPTI form a strong complex with the receptor. Such a complex have K_M values $10^6/10^9$ fold lower than a Michaelis Complex formed with a catalytic active substrate [171][172]. This feature is qualitatively caught by our ReaxFF model. Focusing on the nucleophilic attack range between 3.0Å and 2.4Å (area where the non covalent complex is formed) we appreciate a different response of the two systems in terms of the first derivative of the energy plots.

The energy profile for the substrate increases more sharply than in the inhibitor case as it is shown in figure 4.5.12. This can be better observed in figure 4.5.12 where the angular coefficients of the lines for each point of the two series of data have been plotted together with a calculated R^2 linear regressions.

Linear regression R^2 coefficients indicate a large correlation: R^2 values are respectively 92% and 79% for the substrate and the inhibitor case. From the regression analysis of the inhibitor case we see initial negative values of the angular coefficient that after 2.85Å become positive but still lower than the substrate case. This is an indication of a larger stabilization effect due to

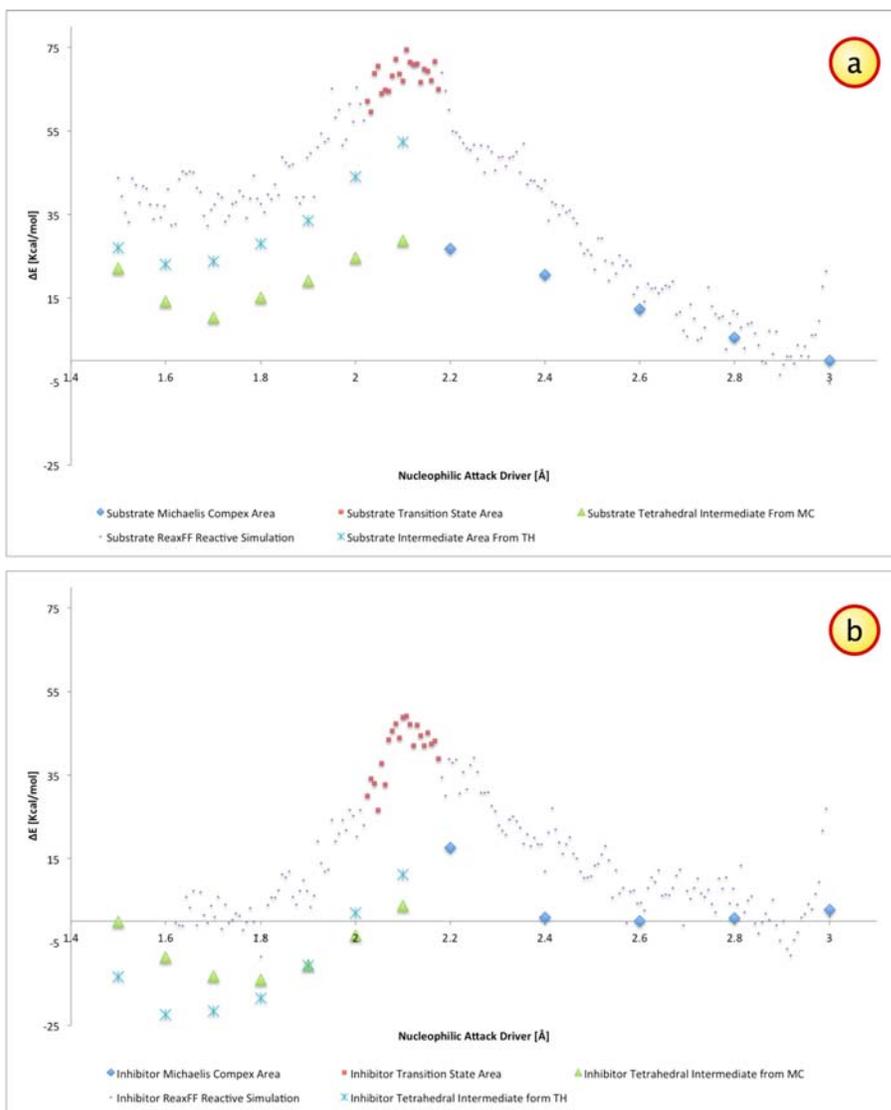


Figure 4.5.11: ProtReaxFF scan diagram. The horizontal axis represents the driver of the nucleophilic attack of the Serine residue. The direction of the reaction is from right to left. On the right there is the Michaelis Complex, while on the left there is the tetrahedral intermediate area. Each gross dot has been minimized from the ReaxFF simulation trajectory. Two cases are reported: a) simulations of receptor + substrate; b) simulations of receptor + inhibitor.

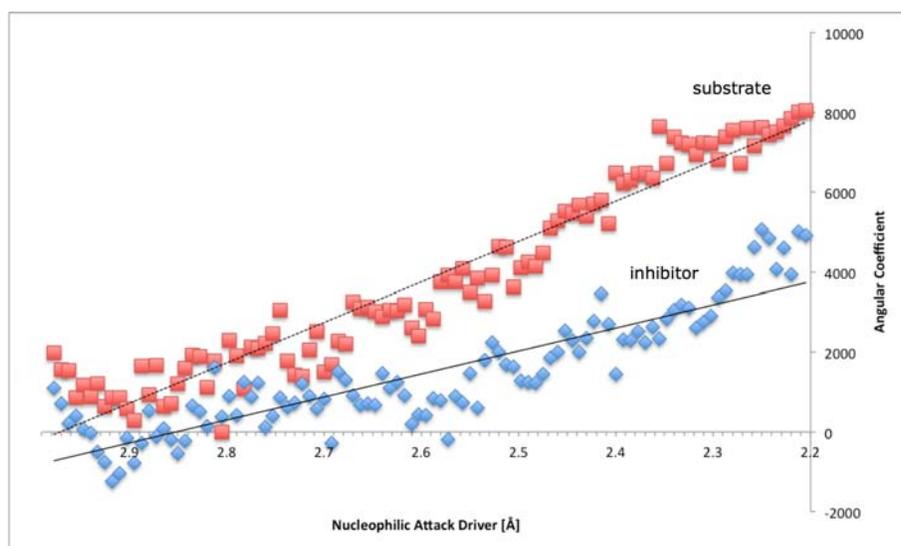


Figure 4.5.12: Linear regression analysis of angular coefficients of lines between the first point and each point of the MD-NVT scan simulation in the range between 2.98Å and 2.4Å for receptor+substrate case and receptor+inhibitor case.

interaction between inhibitor and receptor compare to the substrate case.

To investigate further the hydrogen transfer over this first catalytic step we decided to use three dimensional scans. A set of scans of 1.25ps MD-NVT simulation constraining only on the nucleophilic attack drivers have been performed starting from the complex structures. The same set simulations have been carried out also starting from the tetrahedral intermediate structures. Energies obtained are shown in table 4.12.

Substrate nucleophilic attack [Å]	E_{pot} [Kcal/mol]	ΔE_{pot} [Kcal/mol] ^[1]	Standard Deviation ^[2]				Number of Steps ^[3]	
3.0	-2420763.34	0.00	4.87				10'000	
2.8	-2420757.80	5.55	4.87				10.000	
2.6	-2420751.05	12.29	4.87				10.000	
2.4	-2420742.84	20.50	4.87				10.000	
2.2	-2420736.54	26.80	5.93				10.000	
	Forward Scan ^[4]	Backward Scan ^[5]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]
2.1	-2420734.60 ↓	-2420710.87 ↑	28.74 ↓	52.47 ↑	4.46 ↓	4.29 ↑	5'000 ↓	5'000 ↑
2.0	-2420738.65 ↓	-2420710.87 ↑	24.69 ↓	44.08 ↑	4.14 ↓	4.94 ↑	5'000 ↓	5'000 ↑
1.9	-2420744.15 ↓	-2420729.80 ↑	19.19 ↓	33.54 ↑	5.10 ↓	5.14 ↑	5'000 ↓	5'000 ↑
1.8	-2420748.19 ↓	-2420735.29 ↑	15.15 ↓	28.06 ↑	5.09 ↓	4.88 ↑	5'000 ↓	5'000 ↑
1.7	-2420753.03 ↓	-2420739.55 ↑	10.31 ↓	23.79 ↑	4.00 ↓	5.01 ↑	5'000 ↓	5'000 ↑
1.6	-2420749.22 ↓	-2420740.21 ↑	14.12 ↓	23.13 ↑	4.31 ↓	4.92 ↑	5'000 ↓	5'000 ↑
1.5	-2420741.26 ↓	-2420736.35 ↑	22.08 ↓	26.99 ↑	5.05 ↓	4.92 ↑	5'000 ↓	15'000 ↑
	Forward Scan ^[4]	Backward Scan ^[5]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]
3.0	-2737982.85	2.74	4.27				10'000	
2.8	-2737984.84	0.75	4.49				10.000	
2.6	-2737985.59	0.00	4.44				10.000	
2.4	-2737984.76	0.84	4.52				10.000	
2.2	-2737967.97	17.63	6.31				10.000	
	Forward Scan ^[4]	Backward Scan ^[5]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]	Forward Scan ^[6]	Backward Scan ^[6]
2.1	-2737981.97 ↓	-2737974.41 ↑	3.62 ↓	11.19 ↑	5.56 ↓	9.31 ↑	5'000 ↓	5'000 ↑
2.0	-2737988.92 ↓	-2737983.53 ↑	-3.33 ↓	2.06 ↑	5.54 ↓	8.54 ↑	5'000 ↓	5'000 ↑
1.9	-2737996.00 ↓	-2737996.21 ↑	-10.41 ↓	-10.62 ↑	5.42 ↓	6.55 ↑	5'000 ↓	5'000 ↑
1.8	-2737999.56 ↓	-2738004.65 ↑	-13.97 ↓	-18.46 ↑	4.96 ↓	5.69 ↑	5'000 ↓	5'000 ↑
1.7	-2737998.75 ↓	-2738007.04 ↑	-13.16 ↓	-21.45 ↑	5.66 ↓	5.49 ↑	5'000 ↓	5'000 ↑
1.6	-2737994.23 ↓	-2738007.89 ↑	-8.64 ↓	-22.30 ↑	5.53 ↓	5.28 ↑	5'000 ↓	5'000 ↑
1.5	-2737985.66 ↓	-2737998.93 ↑	-0.07 ↓	-13.34 ↑	5.97 ↓	5.31 ↑	5'000 ↓	15'000 ↑

^[1] The Reference is the average value over 9000 steps simulations with nucleophilic attack driver equals to 3 Å. ^[2] The first 1'000 steps are not included in the standard deviation formula. ^[3] simulation time step 0.25fs, temperature 300, every details can be found in the computational section. ^[4] scan in the forward direction means from Michaelis complex to the tetrahedral intermediate. ^[5] scan in the backward direction means from the tetrahedral intermediate to the Michaelis Complex. ^[6] The Reference is the average value over 9000 steps simulations with nucleophilic attack driver equals to 2.6 Å.

Table 4.12: ProtReaxFF table of the energies and standard deviation of datas reported in figure 4.5.11. The top table describes the simulations of receptor + substrate, while the bottom table reports data of the simulations of receptor + inhibitor.

Constraining only the nucleophilic attack distance we did not appreciate any hydrogen transfers, therefore from table 4.12 and from the diagram a) and b) of figure 4.5.11 arises a wrong description of the potential energy surface of the developed force-field because a protonated Ser190 appears to be energetically stable even when the covalent bond is formed with the carbonyl carbon of both substrate and inhibitor. In the case of the substrate, in the range 1.5/2.1 Å, scan points are even more stable (forward scan of table 4.12) than the structure where hydrogen transfers have been accomplished (backward scan of table 4.12).

The forward scan points have been used as starting structures of a set of other 1.25ps MD-NVT scan simulations with constraints on the hydrogens to be transferred between the triad residues. The two contour maps of figure 4.5.13 represent the PES obtained for substrate and inhibitor case. The reaction co-

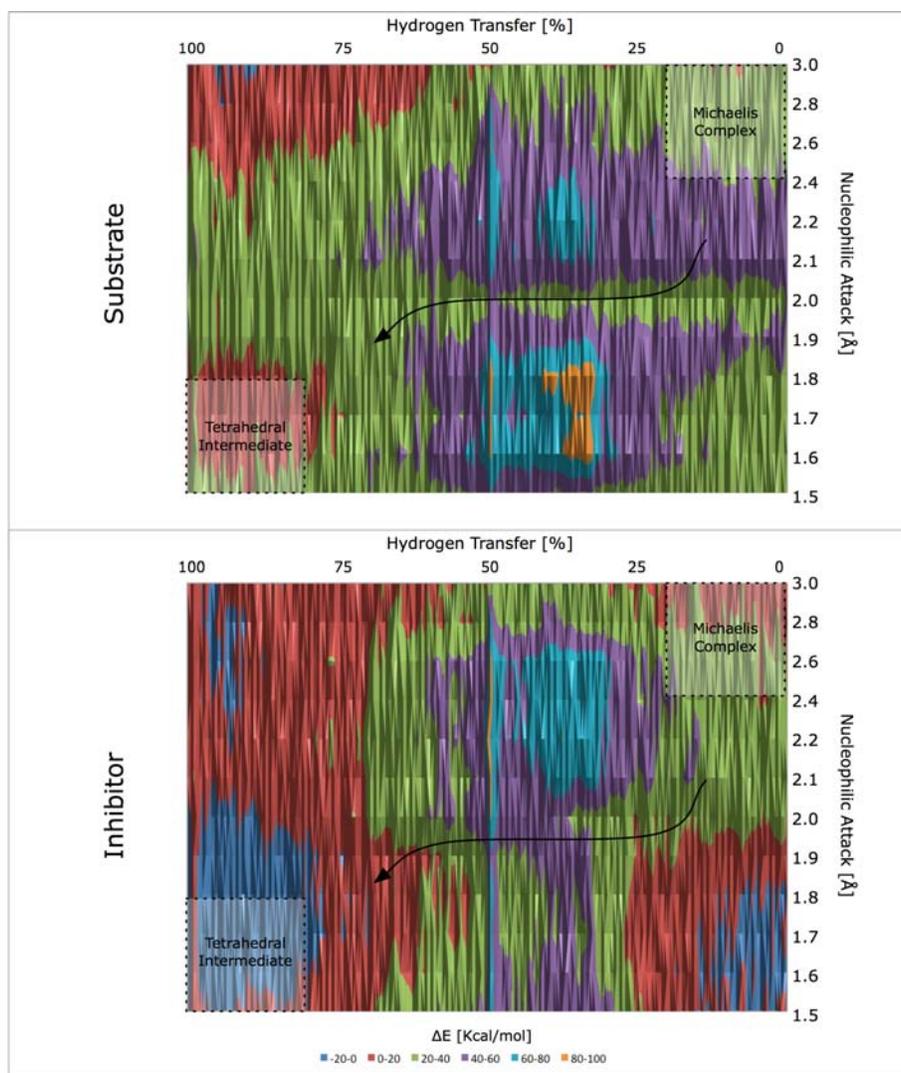


Figure 4.5.13: Contour maps for substrate and inhibitor case.

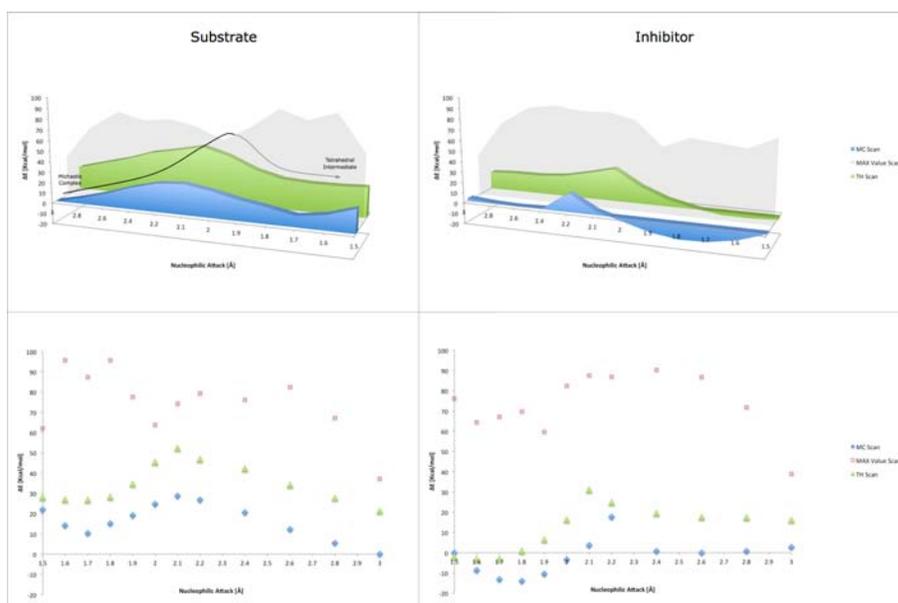


Figure 4.5.14: Nucleophilic attack scans. Optimized structures at the beginning and at the end of ReaxFF simulation scan. The max value scan is a plot of the maximum value from each of hydrogen transfers scan simulations. The same values are presented at the bottom as a 2D diagrams

ordinate hydrogen transfer on the x-axis is reported as a percentage, it consists of a combination of the two hydrogen distances: N_{554}/H_{2456} (hydrogen transfer from the serine to the histidine $N\epsilon_2$ nitrogen), and O_{1155}/H_{551} (hydrogen transfer from histidine to the $O\delta_1$ oxygen). On the y-axis we have the nucleophilic attack distance, the colour represents the variation of the energy.

The scan points with nucleophilic attack distance equal to 1.5\AA have been furtherly equilibrated with 1.25ps simulations. As illustrative example in the diagrams of 4.5.14 minimized structure at the beginning and the end of the hydrogen trasfer scan simulations are plotted (they are called respectively MC scan and TH scan). Values of the TH scan are not that different from the

backward scan reported in table 4.12 and figure 4.5.11. We see room for further refinement of the Prot-ReaxFF because in the TH scan we see stable structure with deprotonated Ser190 and nucleophilic attack large distances for both substrate and inhibitor case, the same problem shown by the MC scan.

The starting structures of the three dimensional scan simulations are those one obtained constraining only the nucleophilic attack distances. We notice a different asset of the catalytic triad varying the nucleophilic attack distance. In table 4.13 and table 4.14 are collected the distances between i) Serine oxygen and $N\epsilon 2$ nitrogen of the histidine, and ii) $N\delta 1$ nitrogen of the histidine and $O\delta 1$ oxygen of the aspartate. The distance between $N\delta 1$ nitrogen of the histidine and $O\delta 1$ oxygen of the aspartate from the MC scan is pretty much constant at 2.7\AA in both substrate and inhibitor case, while the distance between serine oxygen and $N\epsilon 2$ nitrogen of the histidine of the same scan in the range between 2.4\AA to 1.8\AA is larger in the substrate case compare to the inhibitor one. Such a distance in the equilibrated structure of the TH scan in the range of nucleophilic attack between 1.5\AA and 1.7\AA appears to be again shorter for the substrate case compare to the inhibitor one.

This suggests us to carry out a study upon the hydrogen transfer at various distances between Serine oxygen and histidine $N\epsilon 2$. Such a distance has been fixed at three values: 3.0\AA , 2.8\AA , and 2.6\AA . The nucleophilic attack distances have been instead fixed at 2.2\AA , 2.0\AA , and 1.8\AA , while the distance between $N\delta 1$ nitrogen of the histidine and $O\delta 1$ oxygen of the aspartate fixed to 2.7\AA . Equilibration simulation of 1.25ps has been carried out to obtain the starting structure for the study, then a series of 1.25ps MD-NVT scan simulations with constraints on the previously mentioned distances plus the hydrogen transfer

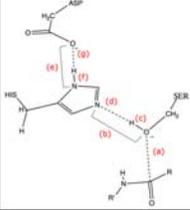
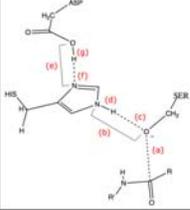
Substrate MC scan ^[1]		[1] Scan structures with de-protonated aspartate and protonated serine.									
	Nucleophilic Attack Distance (a) [Å]	2.4	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.5	
	SerO—N _{α2} (His) Distance (b) [Å]	2.6	2.6	2.7	2.8	2.8	2.9	3.0	3.0	3.0	
	SerO—H Distance (c) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	H—N _{α2} (His) Distance (d) [Å]	1.6	1.6	1.7	1.8	1.8	1.9	2.0	2.0	2.0	
	(His)N _{β1} —O(Asp) Distance (e) [Å]	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	
	(His)N _{β1} —H Distance (f) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	H—O(Asp) Distance (g) [Å]	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	
Substrate TH scan ^[2]		[2] Scan structures with protonated aspartate and protonated histidine.									
	Nucleophilic Attack Distance (a) [Å]	2.4	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.5	
	SerO—N _{α2} (His) Distance (b) [Å]	2.8	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	
	SerO—H Distance (c) [Å]	1.8	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	
	H—N _{α2} (His) Distance (d) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	(His)N _{β1} —O(Asp) Distance (e) [Å]	3.0	2.9	2.8	2.7	2.7	2.7	2.6	2.9	2.9	
	(His)N _{β1} —H Distance (f) [Å]	2.0	1.9	1.8	1.8	1.7	1.7	1.6	1.9	1.9	
	H—O(Asp) Distance (g) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	

Table 4.13: Distances Scan Substrate

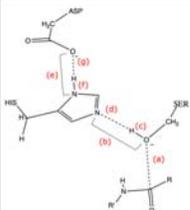
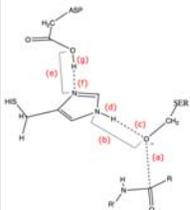
Inhibitor MC scan ^[1]		[1] Scan structures with de-protonated aspartate and protonated serine.									
	Nucleophilic Attack Distance (a) [Å]	2.4	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.5	
	SerO—N _{α2} (His) Distance (b) [Å]	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.1	3.1	
	SerO—H Distance (c) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	H—N _{α2} (His) Distance (d) [Å]	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.1	2.1	
	(His)N _{β1} —O(Asp) Distance (e) [Å]	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	2.7	
	(His)N _{β1} —H Distance (f) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	H—O(Asp) Distance (g) [Å]	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	
Inhibitor TH scan ^[2]		[2] Scan structures with protonated aspartate and protonated histidine.									
	Nucleophilic Attack Distance (a) [Å]	2.4	2.2	2.1	2.0	1.9	1.8	1.7	1.6	1.5	
	SerO—N _{α2} (His) Distance (b) [Å]	2.7	2.7	2.7	2.7	2.7	2.7	2.8	2.8	2.9	
	SerO—H Distance (c) [Å]	1.7	1.7	1.7	1.7	1.7	1.7	1.8	1.8	1.9	
	H—N _{α2} (His) Distance (d) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	(His)N _{β1} —O(Asp) Distance (e) [Å]	2.6	2.6	2.7	2.6	2.6	2.6	2.6	2.6	2.6	
	(His)N _{β1} —H Distance (f) [Å]	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	
	H—O(Asp) Distance (g) [Å]	1.6	1.6	1.7	1.6	1.6	1.6	1.6	1.6	1.6	

Table 4.14: Distance Scan Inhibitor

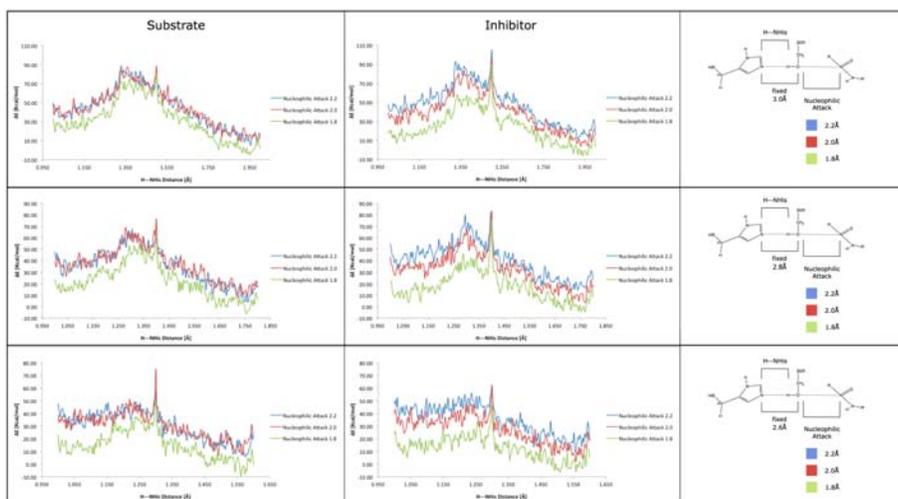


Figure 4.5.15: Comparison SerO–NHis Fixed Nuc Attack variable

N_{554}/H_{2456} (hydrogen transfer from the serine to the histidine $N\epsilon_2$ nitrogen), and O_{1155}/H_{551} (hydrogen transfer from histidine to the $O\delta_1$ oxygen) have been carried out. Energy plots of such a scan with fixed distances between Serine oxygen and histidine $N\epsilon_2$ with a range of nucleophilic attack distances are plotted in figure 4.5.15, while the same scans but with fixed nucleophilic attack distances varying Serine oxygen and histidine $N\epsilon_2$ distances is plotted in figure 4.5.16. In all the diagrams the reaction coordinate is represented by the hydrogen transfer between Serine and Histidine (the distance between N_{554} and H_{2456}).

We do not appreciate differences between substrate and inhibitor. According with the intuition the picture that arises is an easier hydrogen transfer when the distance between hydrogen donor and acceptor is shorter. As previously mentioned, in table 4.13 and 4.14 it has been noted that the inhibitor is characterized by longer distance between hydrogen donor and acceptor. This may be

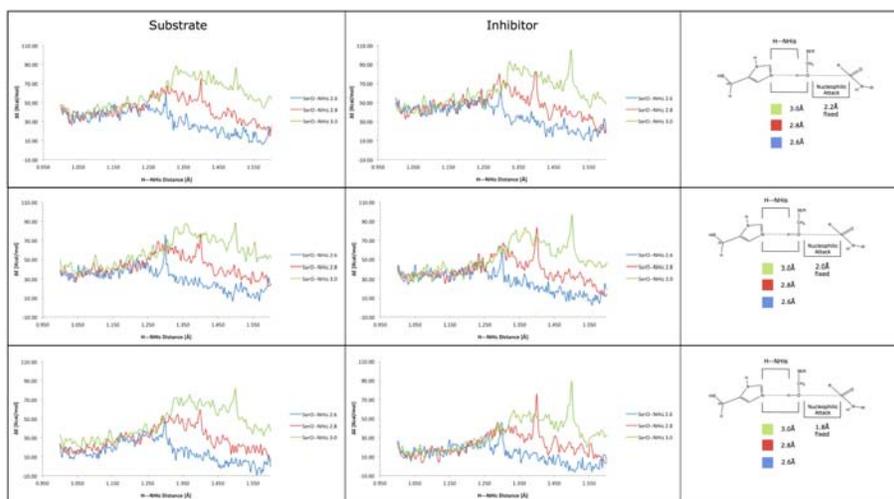


Figure 4.5.16: Comparison Nuc Attack Fixed SerO—NHis variable

a not perfect situation for the hydrogen transfer reaction because of a not well oriented catalytic triad to accomplish it. Such an observation is not a conclusion because it might be a simple artifact of the large error of ReaxFF method. The average value of the standard deviation is 5.6Kcal/mol, as it can be seen from table 4.12. Even though with the model hereby proposed it has been possible to describe extreme situations, i.e. the differences between a catalytic substrate and an inhibitor, the quite high value of the error make impossible to draw any conclusion for applications such as screening of different substrates of pharmaceutical or industrial interests. Additional refinements of Prot-ReaxFF force field together with longer simulation time with optimized simulation parameters will be the object of further studies.

Chapter 5

Resistance of 5-Thio-Galactoside to α -Galactosidases A

5.1 Introduction

The enzyme α -galactosidase (α -GAL, also known as α -GAL A; E.C. 3.2.1.22) is responsible for the breakdown of α -galactoside in the lysosome. Defects in human α -galactosidase lead to the development of Fabry disease, which is a lysosomal storage disorder characterized by the accumulation of α -galactosylated substrates in tissues. This enzymatic defect leads to a progressive accumulation of globotriaosylceramide (GL-3) and its related glycosphingolipids (GLS), in the blood vessels of the skin, kidney, heart, and brain. Glycosphingolipids (GSLs) are ubiquitous, carbohydrate containing lipids that are especially abundant in cells of the nervous system. Within cells they undergo repetitive cycles of synthesis and degradation[183]. Synthesis takes place in the endoplasmic reticulum (ER) and Golgi apparatus, where glycosyltransferases add sugar residues to ce-

ramide in a stepwise manner. Mature GSLs are then transported to the plasma membrane surface where they have roles in many cellular functions including in cell signaling events[183]. After a period of exposure on cell surfaces, GSLs are internalized to be partially or completely degraded by glycosidases and other hydrolases present in the endosomal/lysosomal system. Degradation products are then reused in synthesis. GSL turnover can be rather fast; half lives from two to five hours have been reported[184]. The GSL composition of cells is characteristic for a given cell type and dramatic changes may occur during cell differentiation[184] and upon malignant transformation[185].

A point mutation in which a single nucleotide is changed (missense mutation) is the cause of Fabry disease in a significant number of cases[186]. Indeed, most of these genetic mutations lead to disruption of the hydrophobic core of the protein, Fabry disease is primarily due to protein unfolding, but leave the catalytic activity[187]. Inhibiting human α -GAL is a promising technique to develop such a new pharmacological chaperone therapy[188]. Promising candidates for such a therapy are GSLs containing a terminal sugar in which the ring oxygen has been replaced by sulfur. Figure 5.2.2 shows the structure of UDP-5'-thio-galactose[189]. The sulfur atom is larger and more polarizable than oxygen. Moreover, the carbon-sulfur bond is longer, weaker and less polar than the carbon-oxygen bond and the C-S-C angle is more acute than the C-O-C angle, 95-100° compared to 110-113°. As a consequence, 5-thio-galactose shows differences in the anomeric effect, conformational behavior and chemical reactivity[190][191][192]. Through a collaboration with Monica Palcic and co-workers it was born a computational study trying to get some insight on their experimental observation that glycosides containing terminal α -linked 5-thio-

galactosyl residues are resistant to hydrolysis by mammalian α -galactosidases in vivo and in vitro[193]. Computational modeling is here useful employed to get a deeper understanding of the structural and electronic properties that cause such glycosylation inhibition. The glycosylation reaction comes from an interplay between two aspartate residues present in the protein, namely ASP231 and ASP170. ASP231 acts as a proton donor and protonates the glycosidic oxygen while ASP170 plays the role of a nucleophile by attacking the anomeric center and forming a glycosyl–enzyme intermediate[194]. The active site of α -GAL A together with the ASP231 and ASP170 are shown in figure 5.1.1. The results of the computational study suggest that CYS142-CYS172 disulfide bridge (as well shown in Figure 5.1.1) has an inhibiting role. More specifically the steric interaction of such a bridge with the sulfur atom of 5-thio-galactosides hinders the conformational change of the substrate that is necessary for the hydrolysis. The disulfide bridge CYS142-CYS172 is located within the active site of the protein and interacts with the oxygen and the sulfur atom, respectively for galactoside and 5-thio-galactoside substrate. In the rest of the article the oxygen and sulfur previously mentioned will be named as the oxygen or sulfur heteroatom of the substrate. Combination of classical MD simulations, hybrid QM-MM methods and QM steric analysis is here successfully used to strengthen the aforementioned hypothesis.

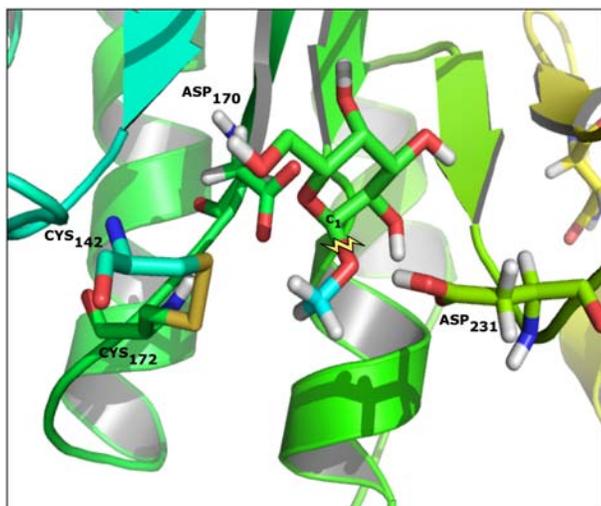


Figure 5.1.1: Substrate and residues involved in the catalysis of the glycosidic breakage: i) ASP170 is the nucleophile that attacks C1 carbon of the substrate, ii) the methoxy leaving group is attached to C1, iii) the ASP231 that assists the reaction with acid-base catalysis, and iii) the disulfide bridge between CYS172 and CYS142

5.2 Experimental Study

5.2.1 Experimental Evidence of Resistance of

5-Thio-Galactoside to α -Galactosidases A action

Monica Palcic et al. synthesized eight fluorescently labeled glycosides containing either terminal O-galactose or the corresponding 5-thio-galactose residue. These were prepared using $\alpha(1\rightarrow3)$ - and $\alpha(1\rightarrow4)$ -galactosyltransferases, UDP-galactose (UDP-Gal) or UDP-5-thio-galactose as the donors, and fluorescently tagged Lac-OR (2a and 2b, Figure 5.2.1) as acceptors. These were tested to see whether they are resistant to hydrolysis by two different types of galactosidase enzymes, human α -galactosidase A (CAZy GH27, a retaining glycosidase)[195]

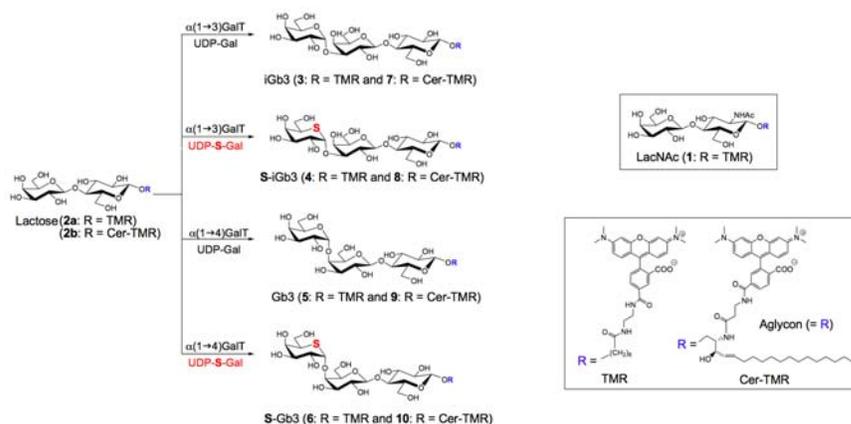


Figure 5.2.1: TMR labeled compounds (1-10) and their enzymatic synthesis.

and *Bacteroides fragilis* α -1,3-galactosidase (FragB, CAZy GH110, an inverting glycosidase) and enzymes present in mammalian cell extracts which are all retaining α -galactosidases. In addition to the in vitro assays, hydrolytic resistance of 5-thio-galactoside was explored in vivo.

5.2.2 Experimental results and discussion

Enzymatic synthesis of TMR labeled galactosides

Eight TMR labeled galactosides (compounds 3 to 10, Figure 5.2.1) containing either galactose or 5-thio-galactose as the terminal sugar were synthesized with specific galactosyltransferases, that catalyze the transfer of galactose from UDP-Gal or UDP-5-thio-Gal to Lac-O-TMR and Lac-Cer-TMR. Both α -1,3-galactosyltransferase and α -1,4-galactosyltransferase readily catalyzed the desired reactions to produce derivatives of iGb3 and Gb3 respectively[196]. The reactions with UDP-5-thio-galactose were slower than those with the natural

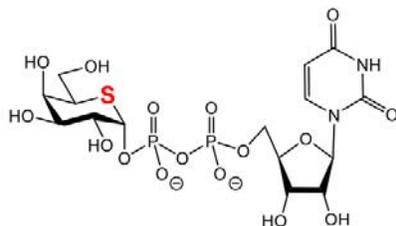


Figure 5.2.2: Structure of UDP-5-thio-galactose.

UDP-Gal donor. Reactions with UDP-5-thio-galactose required a greater excess of donor due to enzymatic hydrolysis which is considerably faster than the enzymatic hydrolysis of UDP-Gal. Conversions above 90%, as estimated by TLC analysis, were seen for all reactions and products were isolated by preparative TLC. The various products differed with respect to aglycon structure, nature of the terminal glycosidic linkage, and the heteroatom in the ring of the terminal sugar. The structural details are provided in 5.2.1.

Hydrolytic resistance assay of TMR-labeled galactosides to human α -galactosidase A and *B. fragilis* α -1,3-galactosidase

Compounds 1-10 were first tested for resistance to hydrolysis by human α -galactosidase A. Human α -galactosidase A cleaves terminal α -galactosyl residues from a variety of oligosaccharides with retention of configuration, with preferred substrates being trihexosylceramides such as Gb3 and iGb3 (figure 5.2.1).

The presence of iGb3 in humans is currently a matter of debate[197]. Human α -galactosidase A is used as an enzyme replacement therapy for the treatment of Fabry disease, an X-linked inherited disorder caused by a defect in the α -galactosidase encoding gene. The enzyme used in this study is from and mar-

Compound	α -GAL A ($\text{pmol} \cdot \text{min}^{-1} \mu\text{g}^{-1}$)	B. fragilis α -1,3-GAL ($\text{pmol} \cdot \text{min}^{-1} \mu\text{g}^{-1}$)
1 (LacNAc-O-TMR)	n.h.[a]	n.h.
2a (Lac-O-TMR)	n.h.	n.h.
3 (iGb3-O-TMR)	87	6400
4 (S-iGb3-O-TMR)	n.h.	69
5 (Gb3-O-TMR)	15	n.h.
6 (S-Gb3-O-TMR)	n.h.	n.h.
7 (iGb3-Cer-TMR)	17	1200
8 (S-iGb3-Cer-TMR)	n.h.	36
9 (Gb-Cer-TMR)	33	n.h.
10 (S-Gb3-Cer-TMR)	n.h.	n.h.
[a] n.h. = no hydrolysis, the reaction rate is less than $0.02 \text{ pmol} \cdot \text{min}^{-1} \mu\text{g}^{-1}$.		

Table 5.1: Rate of enzyme-catalyzed hydrolysis of ten TMR-labeled compounds, using human α -galactosidase A (retaining enzyme) and B. fragilis α -1,3-galactosidase (invertin enzyme)

keted by Shire. The hydrolysis rate and the products were determined by CE (Table 5.1).

Compounds containing natural α -galactose as the terminal sugar (3, 5, 7 and 9) were readily hydrolyzed to yield the corresponding lactosides. The reaction rate depended on the nature of the substrate. Generally O-TMR was preferred as the aglycon compared to the ceramide and the $\alpha\text{Gal}(1 \rightarrow 3)\text{Gal}$ linkage was hydrolyzed more rapidly than the $\alpha\text{Gal}(1 \rightarrow 4)\text{Gal}$ linkage. LacNAc-O-TMR (1) and Lac-O-TMR (2a), containing terminal β -galactose, were not hydrolyzed. On the other hand, galactosides containing terminal 5-thio-galactose (4, 6, 8 and 10) were resistant to hydrolysis by α -galactosidase A. Very slow degradation at a rate less than $0.02 \text{ pmol} \cdot \text{min}^{-1} \mu\text{g}^{-1}$ enzyme was observed only after four days of incubation. Guce et al. have reported X-ray crystal structures of human α -galactosidase in complex with substrates which showed that ring oxygen of galactoside has a van der Waals interaction with a cysteine residue

on α -galactosidase[198]. The sulfur substitution for the oxygen atom might cause electron repulsion and destabilize the α -galactosidase-substrate complex. The panel of substrates were next evaluated with *B. fragilis* α -1,3-galactosidase (FragB) which is specific for α -1,3 linkages including blood group B antigens.[21] This enzyme catalyzes hydrolysis with inversion of configuration. To date the FragB gene family has only been found in prokaryotes and no homologues or similar inverting enzymes have yet been found in eukaryotes[195]. As expected, α -1,3-galactosidase did not act on compounds lacking a α -1,3 galactoside linkage (1, 2a, 5, 6, 9 and 10) (Table 5.2.3, right). However, all iGb3 derivatives (3, 4, 7 and 8) including thio-galactoside compounds were hydrolyzed by the enzyme. Even though the hydrolysis rates of thio-galactosides 4 and 8 were very slow (1 or 3% of the O-galactosides), complete hydrolyses was observed in all iGb3 derivatives after long incubation times (data not shown). Inverting glycosidase reactions are thought to occur via a single-displacement mechanism with an oxocarbenium ion-like transition state. Most retaining glycosidase reactions are believed to proceed via a double displacement mechanism (two transition states) with formation of a covalent enzyme-substrate intermediate[199]. Replacement of oxygen with a sulfur atom in the sugar ring might impair inverting galactosidases less than retaining enzymes if the transition state activation barrier is greater for thio-galactosides.

Hydrolytic resistance assay of TMR-labeled galactosides to enzymes from cell extracts

Since numerous α -galactosidases in cells can potentially hydrolyze Gb3 and iGb3, we therefore examined TMR-labeled compound (1-10) in extracts from

MCF-7 and NG108-15 cells. The enzyme solution was prepared by lysis of the cell extracts and the protein concentrations were determined by the Bradford method; 3.8 mg/ml from MCF-7 and 5.0 mg/ml from NG108-15, respectively. By using the extracted enzyme solution, the hydrolytic stabilities of TMR-labeled galactosides 1-10 were evaluated as was done for human α -galactosidase A and *B. fragilis* α -1,3-galactosidase. The corresponding compounds containing natural galactose (3, 5, 7 and 9) were readily degraded. The substrate preference observed was similar to that of α -galactosidase A, suggesting that this enzyme is also present in the extracts, despite the fact that a distinct band corresponding to human α -galactosidase A was not evident from SDS page analysis (data not shown). In this assay, LacNAc-O-TMR (1) and Lac-O-TMR (2a) were both hydrolyzed. The degradation pattern was different depending on whether the pure enzyme or an enzyme extract was used. While α -galactosidase cleaved only the terminal sugar, incubation with either cellular extract yielded further hydrolysis products as well, due to the action of β -galactosidases in the extracts. Figure 3 presents the hydrolysis products from Lac-O-TMR (2a) and iGb3-O-TMR (3) after incubation with each cell extract. The β -linked galactoside in Lac-O-TMR was cleaved within 3 hours giving Glc-O-TMR and aglycon O-TMR as degradation products. Gb3-O-TMR was hydrolyzed more slowly than Lac-O-TMR, however, more than 20% hydrolysis occurred after 24 h incubation in both cell extracts. MCF-7 cellular extracts exhibited higher hydrolysis activity compared to NG 108-15 extracts. Compounds (4, 6, 8 and 10) containing terminal 5-thio-galactose were resistant to degradation by enzymes present in the cell extracts. The thio-galactose therefore appears resistant to the action of cellular α -galactosidases.

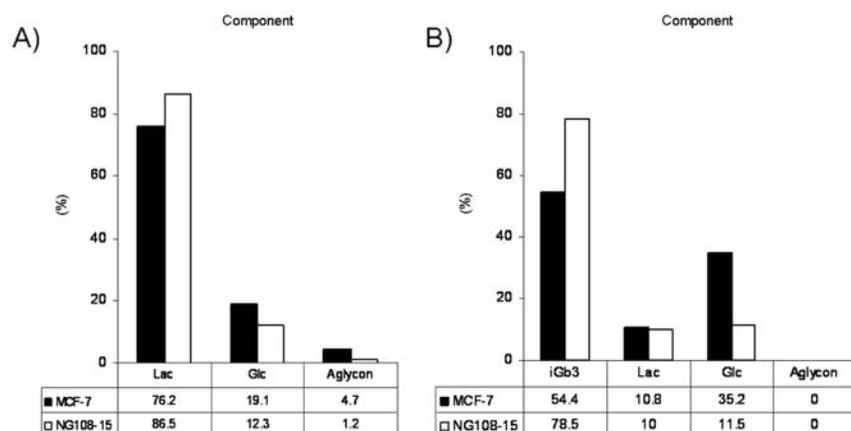


Figure 5.2.3: Composition of TMR-compounds after hydrolysis reaction using cell extracts. Substrates were (a) Lac-O-TMR (3h reaction) and (b) iGb3-O-TMR (24 h reaction). Each bar (black; MCF-7 and white; NG108-15) indicates the percentage of TMR-labeled compounds. The ratios on table were determined by capillary electrophoresis. Aglycon = O-TMR.

Compound	MCF-7 (pmol*h-1)	NG108-15 (pmol*h-1)
1 (LacNAc-O-TMR)	82.5	18.5
2a (Lac-O-TMR)	17.5	14.5
3 (iGb3-O-TMR)	10.5	3.5
4 (S-iGb3-O-TMR)	n.h.	n.h.
5 (Gb3-O-TMR)	1.4	0.3
6 (S-Gb3-O-TMR)	n.h.	n.h.
7 (iGb3-Cer-TMR)	1.0	1.6
8 (S-iGb3-Cer-TMR)	n.h.	n.h.
9 (Gb-Cer-TMR)	0.5	0.5
10 (S-Gb3-Cer-TMR)	n.h.	n.h.

[a] n.h. = no hydrolysis, the reaction rate is less than 0.02 pmol*min⁻¹μg⁻¹.

Table 5.2: Rate in enzyme-catalyzed hydrolysis of ten TMR-labeled compounds, using cell extracts.

In vivo stability of 5-thio-galactoside using NG108-15 cells

On the basis of the in vitro assay, we tested the hydrolytic resistance of 5-thio-galactosides in vivo with NG 108-15 cells. The cells were grown at 37 °C for 24 h with S-iGb3-O-TMR (4) added to the medium and then the intracellular contents in cell homogenate were analyzed by CE. Lac-O-TMR (2a) (7%) was present as a contaminant in this preparation of 4 due to incomplete enzymatic synthesis. If the two TMR-labeled compounds are taken up by the cells and show the hydrolysis properties observed in the in vitro assay (Table 5.2), only Lac-O-TMR should be digested by intracellular galactosidases. Gb3-O-TMR (3) was assayed in the same way as S-iGb3-O-TMR. Figure 5.2.4 shows the electropherogram of NG 108-15 cell contents after incubation with (a) S-iGb3-O-TMR (containing 7% Lac-O-TMR) or (b) iGb3-O-TMR. S-iGb3-O-TMR was not hydrolyzed in vivo; while the peak corresponding to Lac-O-TMR was greatly reduced with conversion to Glc-O-TMR. In contrast, iGb3-O-TMR was hydrolyzed to Lac-O-TMR and Glc-O-TMR. Anabolic glycosylated products were not generated from either iGb3 substrate. However, these results revealed that 5-thio-galactoside is stable towards hydrolysis in vivo as well as in vitro.

5.3 Modeling Study

5.3.1 The role of a disulfide Bridge in the Resistance of 5-Thio-Galactoside to α -Galactosidases A

The starting structures of α -galactosidase A were downloaded from the RCSB Protein Databank (<http://www.rcsb.org/>)[177]. Human α -Gal A is a homod-

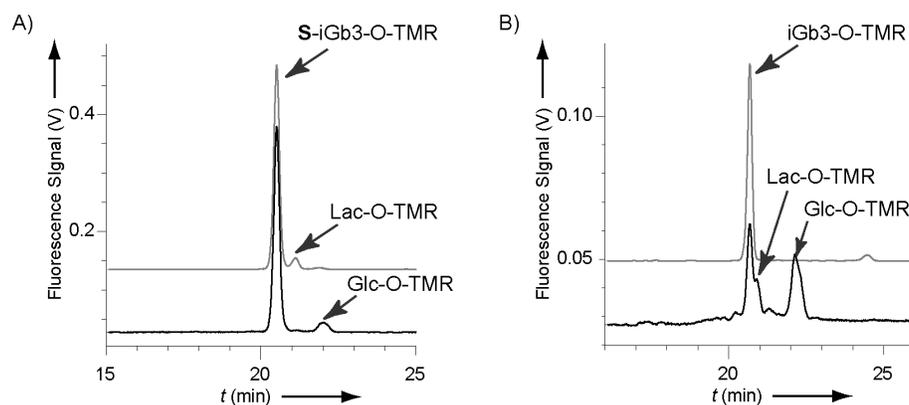


Figure 5.2.4: CE analyses of NG108-15 homogenates after incubation with (a) S-iGb3-O-TMR and (b) iGb3-O-TMR. Gray and black lines represent before and after incubation, respectively. The S-iGb3-O-TMR in panel (a) contains Lac-O-TMR (7%) due to incomplete enzymatic synthesis.

imer, where each monomer consists of two domains, an N-terminal barrel containing the active site, and an antiparallel C-Terminal domain. Each of the monomer has an active site. The coordinates were obtained by means of an X-Ray diffraction mechanistic study[198]. There is no substantial difference between the two monomers and no evidence of cooperativity, therefore only a of the monomers has been selected to as starting structure of our computational investigation . Pdb codes fo the structures are:

- 3HG3 (DOI:10.2210/pdb3hg3/pdb);
- 3HG4 (DOI:10.2210/pdb3hg4/pdb).

They respectively represent the Michaelis complex and the covalent complex formed between receptor and substrate. Molecular dynamics simulations have been performed using Desmond Molecular Dynamics System, version 2.2 (D. E. Shaw Research, New York, NY, 2008)[200] and OPLS-2005 force-field[178]. The

final trajectories were analyzed and visualized using Maestro v9.3 (Schrödinger, LLC, New York, NY, 2012). 3HG4 consists of the substrate 2-deoxy-2,2-difluoro-beta-D-lyxo-hexopyranose covalently bound to the enzyme. The fluorinated hexopyranose substrate has been replaced with 5'-galactose. The Melibiose substrate of the 3HG3 X-Ray structure has instead been left unchanged. The Schrodinger protein preparation wizard[201] has been used to prepare the protein for the analysis, to optimize the hydrogen bond network and to model moderate low pH condition. The structure used can be found in the supporting information. A simulation cubic box of length 120Å filled with more than 53000 TIP3P water molecules has been created around the protein. Sodium ions were added to neutralize the entire system, while the physiologic condition of the ionic strength is assured by a salt concentration of 0.15 mol/L. Simulations were carried out in the NPT regime; time is set to 12ns and 10ns, respectively for covalent complex (3HG4) and Michaelis Complex (3HG3). A pressure of 1.01325 bar and a cryogenic temperature of 100K have been set using the Nose-Hoover chain thermostat method[202] and the Martyna-Tobias Klein Barostat method[203]. Pre-equilibration steps before the productive simulations follow the Desmond default relaxation protocol. Calculations using the hybrid QM/MM (quantum mechanics/molecular mechanics) as implemented in Qsite v5.7 (Schrödinger, LLC, New York, NY, 2011)[204][205][206] has been carried out too. In a Qsite calculation the QM part of the system is calculated through jaguar v7.8 while the MM part is carried out by Impact v57111. The QM part is treated using Density functional theory (DFT) framework with functional B3LYP[16][17] and MO6_2X[51]. The basis set used is a split valence 6-31G*[81]. We do not introduce diffuse functions and further polarization

functions, since in presence of link atoms it would result in a more flexible wave function, which is more prone to over polarization at the boundaries[58]. The MM part is described by the OPLS2005 force-field[178]. The last frame of the trajectory of the MD simulation of 3HG4 has been used as starting structure for the QM/MM analysis. The oxygen atom of the galactose has been exchanged with sulfur to model 5-thio-galactosides. Truncation between the QM and MM part have been done using hydrogen cap atoms. The complete substrate has been included in the QM part together with the two catalytic Asp231 and Asp170 residues for a total of 40 atoms (including two hydrogen atom caps). The chosen truncation site for both aspartate residues is between the C α and C β carbons. Structure 3HG4 has 6136 atoms in total, so the MM part consists of 6100 atoms (including two hydrogen caps). We are not interested in the motion of the protein backbone but only in the electrostatic interaction of the medium with the QM part, so we decided to model the protein environment with a shielding dielectric constant equals to 15. This dielectric constant has been extrapolated from two experimental dispersions spectra of lysozyme with longitudinal relaxation times of 650 ps[207]. In the real system the charged side chain on the protein surface are solvated by water molecules. Following the standard approach used extensively by Friesner et al. [206] we avoided the introduction of a solvent description neutralizing the charges of solvent-exposed side chains (with the only exception of residues of opposite charge with a net charge equals to zero and charged residues near the active site) and freezing residues that are distant from the active site of the protein. Positive charges neutralized are those one of lysine (#Res no: 237, 393, 314, 82, 127, 130, 326, , 213) and arginine residues (#Res no: 118, 196, 38, 220, 392, 404, 402, 332,

252, 193). To neutralize arginine residues the sidechain group $\text{NH}_2\text{-C}=\text{NH}$ has been replaced by a methyl group. The neutralized negative charges are instead those one of aspartate (#Res no: 175, 182, 153, 161, 33, 83, 55, 234, 233, 313) and glutamine residues (#Res no: 178, 103, 87, 74, 71, 59, 58, 48, 338, 418, 251, 398). We froze all the residues having a distance of more than 10\AA from the QM region. The MM part within 10\AA from the QM region consists of #Res no: 5, 46, 47, 48, 49, 50, 51, 52, 53, 54, 64, 69, 90, 91, 92, 93, 94, 95, 96, 106, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 152, 155, 159, 166, 167, 168, 169, 170, 171, 172, 173, 174, 179, 180, 181, 183, 184, 187, 191, 200, 201, 202, 203, 204, 205, 206, 207, 208, 223, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 237, 238, 239, 240, 241, 242, 245, 264, 265, 266, 267, 268, 269, 283, 287, 295, 297. The frozen atoms are 4620 in total. A sketch of the atoms involved in the catalysis in the QM part can be found in figure 5.3.2, while free and frozen part of the system are shown in figure 5.3.1. To evaluate steric interaction at QM level we have used the NBO steric method as implemented in Jaguar NBO 5.0[208]. Steric interaction is here approximated as an exchange effect, therefore Hartree-Fock (HF) method gives a reasonable accurate description of such interaction[54][55]. NBO steric analysis have been performed at HF level with 6-31G** basis set doing single point calculations on fragments cut from QM-MM optimized structures.

5.3.2 Computational Results and Discussion

The two simulation trajectories have been analyzed in order to examine the position and orientation of the CYS142-CYS172. disulfide bridge. As it can be seen in Figure 5.1.1, this disulfide bridge is situated in the vicinity of the

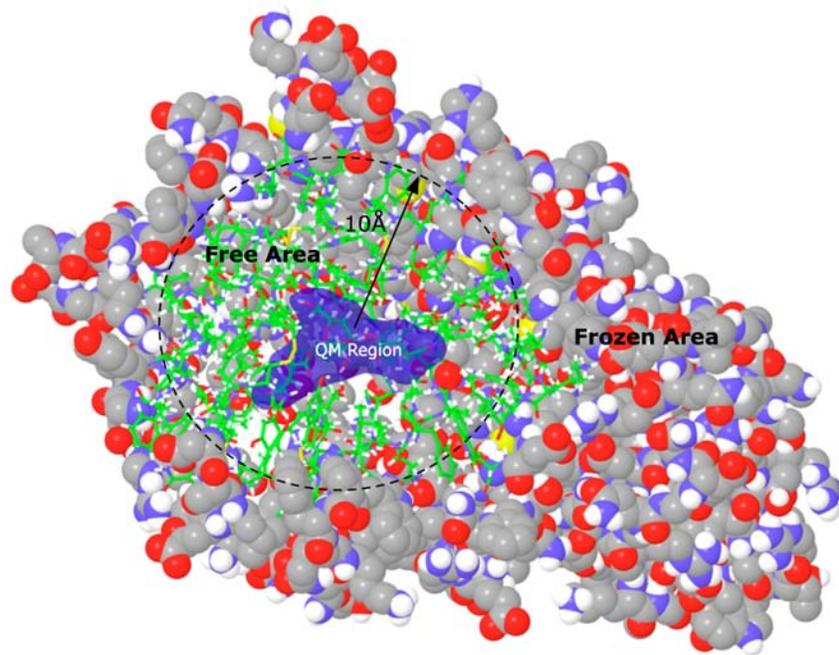


Figure 5.3.1: The figure shows the free to move area (both QM and MM) and the frozen MM region consisting of residues above 10Å from QM area.

ring-oxygen of the substrate. In the following discussion the simulation of the Covalent Complex will be 3HG4, while that one of the Michaelis Complex will be 3HG3. From the simulation event analysis the average distance between the ring-oxygen of the substrate and the sulfur atom of CYS142 side chain assumes bigger values in the 3HG3 compare to 3HG4 simulation. In 3HG4 such a distance has a range of value between 3.3Å and 3.5Å, while in the 3HG3 case the range is between 3.9Å and 4.2Å. A histogram of the two trajectory distributions of distances between ring-oxygen of the substrate and the sulfur atom of CYS142 is reported in Figure 5.3.3. Blue bars and red bars respectively

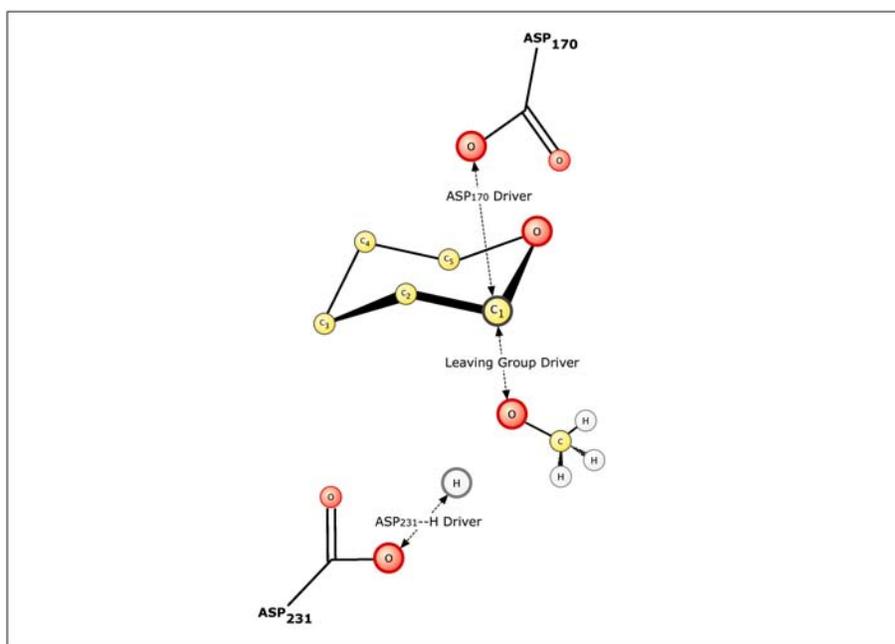


Figure 5.3.2: Scan Drivers.

represent the frequency of such a distance values for 3HG4 and 3HG3 simulation. Further analysis on the orientation of the disulfide bridge dihedral angles shows that they are responsible for the longer. The orientation found throughout the 3HG4 trajectory is named as UP \uparrow (because the disulfide bridge sulfur is above the puckering plane C1-O-C5). Such orientation is can be seen in the sketched scheme on the top right corner of Figure 5.3.3. The orientation of the same disulfide bridge in 3HG3 simulation is different and it is named DOWN \downarrow (because the disulfide bridge sulfur is below the puckering plane C1-O-C5). The UP \uparrow orientation has backbone Φ angles indicatively in the range 160°/170° and 180° respectively for CYS172 and CYS142, while the DOWN \downarrow orientation has the same backbone Φ angles in the 90°/100° range and centered around

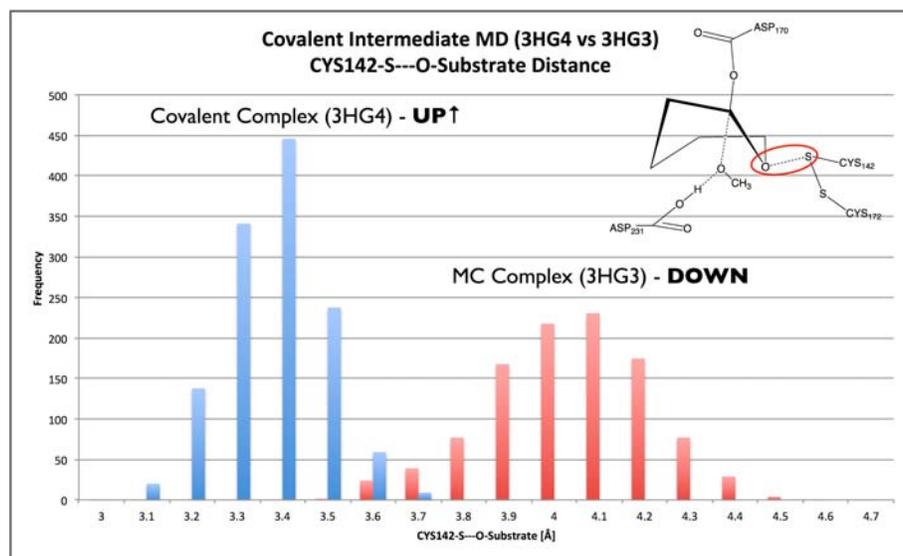


Figure 5.3.3: Ring-Heteroatom distances from sulfur of residue CYS142

110° respectively for CYS172 and CYS142. Histograms with frequency of Φ dihedral angles are shown in Figure 5.3.4.

In particular during the MD simulation of 3HG4 the disulfide bridge does not change the orientation, and as a consequence the bridge is UP \uparrow , as in the X-Ray crystallographic structures of Garman et al. [198], for each step of the reaction. In this configuration the interaction between ring-oxygen of the substrate and the disulfide bridge should be stronger because the distance between such oxygen and the sulfur of CYS142 is much shorter than in the DOWN \downarrow case. The last frame of the 3HG4 MD trajectory has been used as starting structure of QM-MM analysis. To adapt the system to the QM-MM description the protein has been treated as illustrated in the previous paragraph and it has been optimized without any constraints with QM-MM (B3LYP-OPLS2005). The structure obtained represents the covalent complex and it has been used as

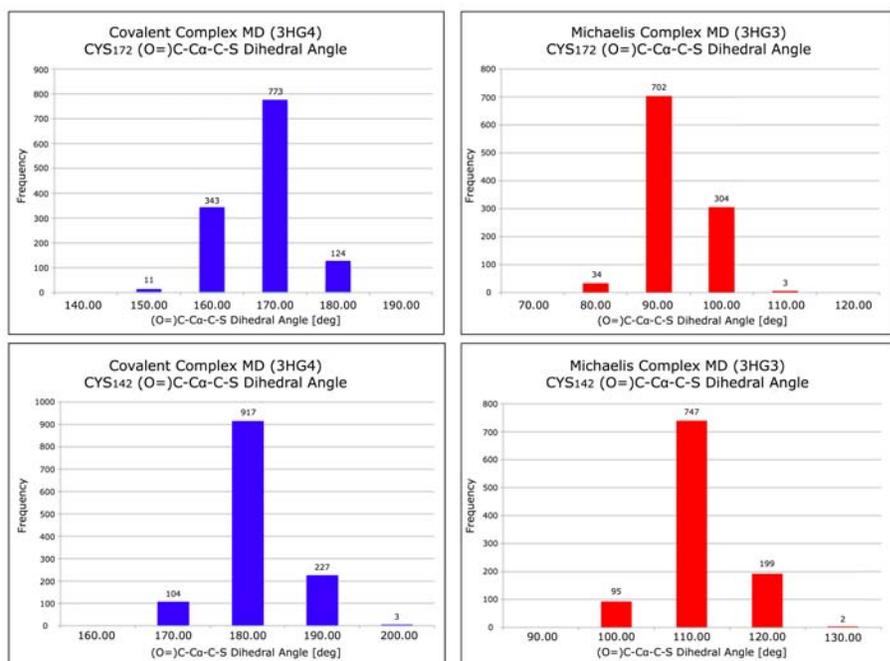


Figure 5.3.4: CYS172 and CYS142 backbone Φ angles distributions.

starting structure of our preliminary QM-MM (B3LYP-OPLS2005) scan calculation. The constraints on the distances between carbon C1 of the substrate and the oxygen of the methoxy leaving group (Leaving Group distance), plus the distance between Asp231 oxygen and an hydrogen atom (Asp231-H distance) serve as a way of sampling a PES to get an understanding of a possible reaction path. From these calculations it has been observed that during the reaction the conformation of the ring changes from Boat to Chair (from Covalent complex to Michaelis Complex), and the disulfide bridge between changes its orientation too. Such a change of orientation gives a huge stabilization of the energy (around 100 KJ/mol). The stabilizing configuration of the CYS142-CYS172

disulfide bridge is the DOWN \downarrow orientation. To perform a glycosidic reaction, the substrate undergoes a change of its conformation as shown in figure 5.3.6. The larger vdW radius of sulfur compared to oxygen suggests that this change of conformation would require more space for 5-thio-galactose compare to galactose. To discuss further and to get an understanding of the results that will follow, we list here our main assumptions:

i) the enzyme catalysis effect is due to a stabilization of the transition state structure by the enzyme environment;

ii) reaction time scales are on the order of the femtoseconds, while the protein motions belong to time scales with higher order of magnitude. Following our assumption we believe that the environment cannot drastically change during the course of a reaction.

Our working hypotheses are instead the following:

1) the disulfide bridge orientation must be UP \uparrow throughout all the reaction steps,

2) the UP \uparrow disulfide bridge orientation is more suited for galactose transition states compare to 5-thio-galactose transition states, and in particular the UP \uparrow disulfide bridge orientation it is an impediment for the 5-thio-galactose reactions to occur. Garman et al. investigated the X-Ray structures of the wild-type enzyme (3HG2), the Michaelis Complex (3HG3), the Covalent Complex (3HG4), and the product of the reaction (3HG5). As previously mentioned one of these X-Ray structures has been used to start the computational analysis. In all their X-Ray structures the CYS142-CYS172 disulfide bridge does not change its orientation as it can be seen in Table 5.3.

The information obtained from X-Ray structures is used as a support for

Dihedral Φ [deg]	3HG2	3HG3	3HG4	3HG5
CYS142	168.1	176.6	173.7	169.4
CYS172	152.4	164.0	159.8	154.9

Table 5.3: . In all the X-Ray structures obtained by Garman et al. the disulfide bridge between CYS142 and CYS172 has the orientation the we call UP \uparrow . Such orientation doesn't change throughout the reaction path.

the hypothesis 1) that implies that the orientation of the disulfide should not change along the reaction path, it is all the time UP \uparrow . The UP \uparrow orientation is stable throughout the MD trajectory of 3HG4. Since 3HG4 represents the covalent intermediate of the reaction, we can infer that its geometry should be similar to the Transition State geometry. Following this line of reasoning the CYS142-CYS172 disulfide bridge has been fixed in its UP \uparrow orientation with a force constant of 100 KJ/mol in the following QM-MM (MO62x-OPLS2005) analysis. The distances kept constrained are still those one illustrated in figure 5.3.2, while the results are instead presented as 3D PES in Figure 5.3.5. From top to down of the figure we have the transition state area PES for galactose, 5-thio-galactose and the difference between the two of them. The two surfaces look similar in shape even though the 5-thio-galactose one has more steep gradients. The covalent complex area is at the right edge pointing outside the plane of the paper, with a Leaving Group distance of 3.5Å and a distance between ASP170 and C1 equals to 1.5Å. Checking only the points in the transition state area, as shown in Figure 5.3.7, it can be seen that the majority of the transition state area scan points of 5-thio-galactose are higher in energy compare to 5'-galactose. The majority of the points of 5-thio-galactose are more than 10 KJ/mol higher in energy than the galactose.

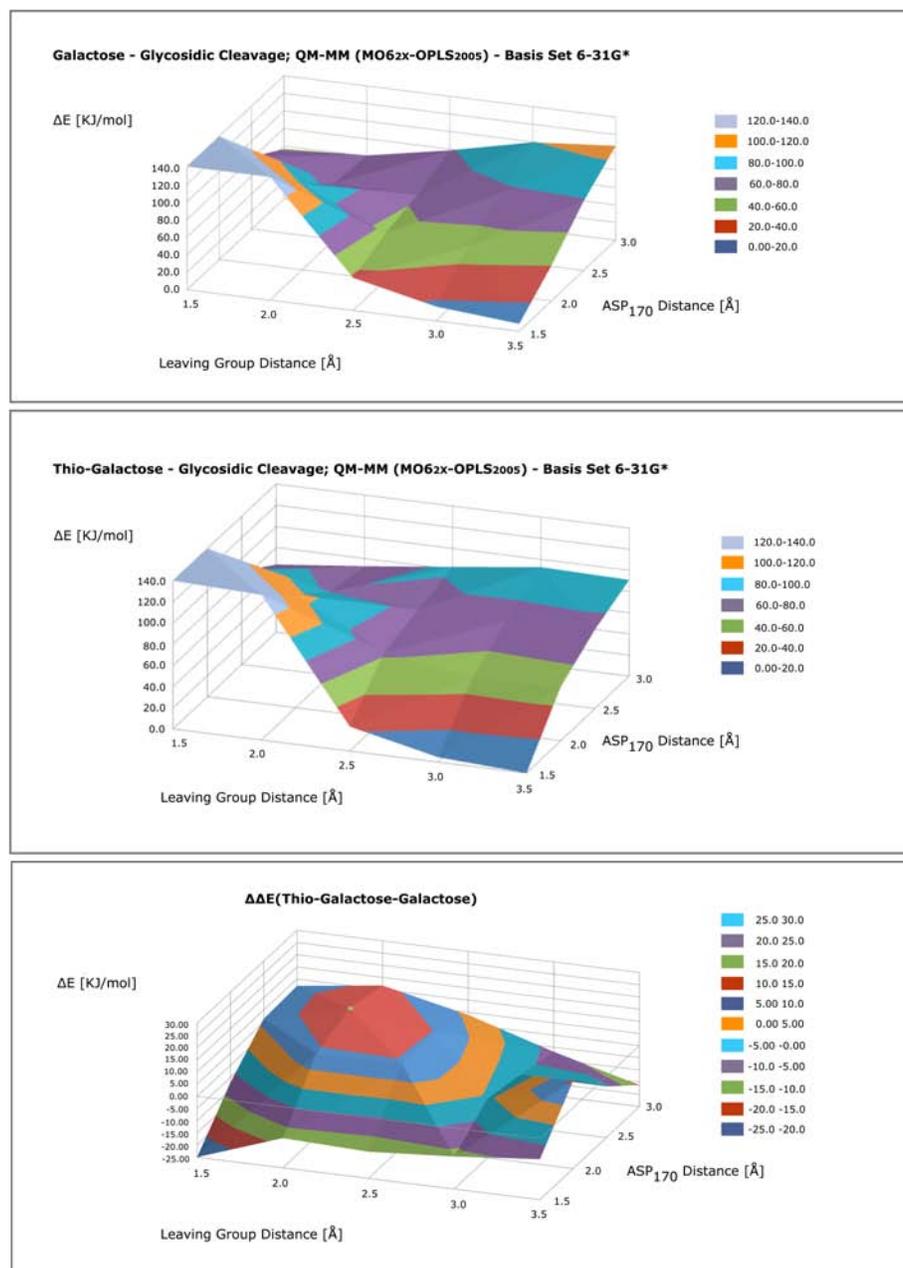


Figure 5.3.5: QM-MM (MO62x-OPLS2005) PES of the Glycosidic Cleavage. Top: galactose; middle: 5-thio-galactose; bottom: Difference between 5-thio-Galactose and galactose.

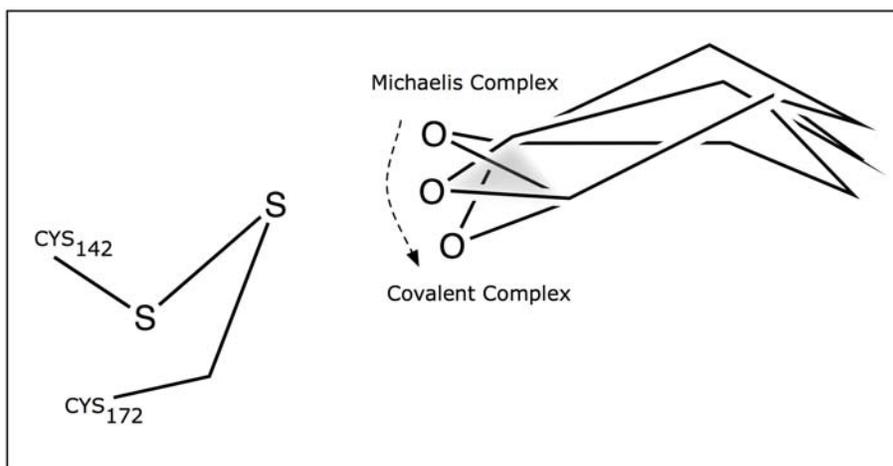


Figure 5.3.6: Conformational changes along the reaction path.

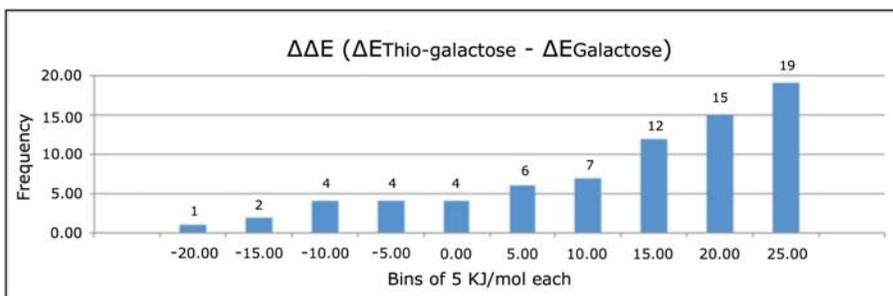


Figure 5.3.7: Distribution of $\Delta\Delta E$ within the transition state area. The majority of the points of the 5-thio-galactose PES is higher in energy with respect to the galactose case.

We believe that such a different behavior of α -galactosidase A with respect to the two substrates can be explained in terms of steric interaction between the ring heteroatom (oxygen or sulfur for galactose and 5-thio-galactose respectively) and the active site CYS142 residues. The following steric analysis focuses on the steric interaction between the ring heteroatom and the sulfur of CYS142.

From a theoretical point of view, steric interactions are due to the Pauli exclusion principle and the antisymmetry of wave functions[209]. As previously mentioned, the NBO decomposition of the wave function into localized bond and lone pair orbitals corresponding to the Lewis structure of the molecule allows the steric repulsion to be expressed in terms of relative changes of individual NBOs upon orthogonalization. The total exchange repulsion can be approximately separated into pair-wise “local” NBO-NBO interactions by means of local partial deorthogonalization[52]. NBO analysis can be applied to any wave function for which a Fock matrix is available.

Transition state guess and Michaelis complex structures have been selected to evaluate the entity of steric interaction between CYS142 sulfur and ring heteroatom of both substrates. Transition state structure guess is characterized by a distance between methoxide leaving group and C1 of the substrate equals to 2.0Å, a distance between the nucleophilic oxygen of the Asp170 and the same C1 atom equals to 2.0Å and the hydrogen atom transferred on Asp231. The Michaelis Complex has instead a leaving group distance of 1.5Å (bound to C1 of the substrate) and the nucleophilic Asp170 oxygen distant 2.7Å from the substrate. The steric interaction is evaluated between two fragments. For what concerns the transition state guess structure, the covalent complex formed between Asp170 and the substrate forms the first fragment, while CYS142-

TS Guess	Galactose		5-thio-galactose	
NBO Analysis	Distance [Å]	Steric Int. [Kcal/mol]	Distance [Å]	Steric Int. [Kcal/mol]
	3.7	-22.93	3.73[a]	-1.95
	3.6	-22.12	3.7	0.18
	3.5	-20.17	3.6	1.23
	3.43[a]	-17.36	3.45	2.91
	3.4	-19.0	3.4	6.56
	3.3	-15.93	3.3	10.92
	3.2	-11.36	3.2	17.12
[a] Distance as it comes from QM-MM output.				

Table 5.4: Transition state guess steric interaction with distance between ring-heteroatom and Disulfide Bridge. [1] Methoxide Leaving Group distance from substrate C1 equals to 2.0Å; nucleophile Asp170 oxygen distance from substrate C1 equals to 2.0Å.

CYS172 disulfide bridge forms the second fragment. For the Michaelis Complex we cut only the substrate and the CYS142-CYS172 disulfide bridge, so that they form respectively the first and second fragment. Steric interactions have been evaluated with different distances between ring-heteroatom and disulfide bridge sulfur of CYS142. Results obtained are shown in Table 5.4 and Table 5.5. The NBO steric repulsion analysis shows a different trend for the TS guess and the MC guess case. In TS guess case, as long as the disulfide bridge is getting near the heteroatom, steric repulsion grows. In the galactose case such interaction remains attractive while for the 5-thio-galactose the sign reveals a repulsive steric interaction from 3.7Å of distance to 3.2Å. In the Michaelis Complex structure the steric interaction is always repulsive, but the modulus of such an interaction is always higher for 5-thio-galactose compare to galactose.

MC Guess	Galactose		5-thio-galactose	
NBO Analysis	Distance [Å]	Steric Int. [Kcal/mol]	Distance [Å]	Steric Int. [Kcal/mol]
	3.7	0.98	3.93[a]	4.65
	3.6	2.01	3.7	11.10
	3.5	3.18	3.6	14.82
	3.43[a]	6.16	3.5	20.03
	3.4	5.98	3.4	25.61
	3.3	9.87	3.3	34.82
	3.2	14.58	3.2	43.39
[a] Distance as it comes from QM-MM output.				

Table 5.5: Michaelis Complex steric interaction with distance between ring-heteroatom and Disulfide Bridge. [1] Methoxide Leaving Group distance from substrate C1 equals to 1.5Å; nucleophile Asp170 oxygen distance from substrate C1 equals to 2.7Å.

Chapter 6

Conclusions

Conclusions

As it happens with releases of new force fields, the conclusions are partly future perspectives because there is always room for further improvements, for more comparisons with reference data, other fitting strategies, and other fitting procedures. Anyway the picture that comes out from the results of chapter 4 points in the good direction. The current version of Prot-ReaxFF can be used to perform molecular dynamics simulations of biomolecules. The results show that the information obtained with non-reactive force field of the AMBER family is conserved when the system is passed to Prot-ReaxFF parameters. This means that biomolecules main features are described by Prot-ReaxFF in a comparable way of the ordinary force fields adopted in the field of biophysics. Root mean square deviations of backbone atoms, shapes of the proteins, secondary structure motifs, disulfide bridges, and hydrogen bonds were retained while switching from AMBER parameters to Prot-ReaxFF parameters. Prot-ReaxFF has also been applied in a catalysis study with short dynamics simulations. There is

still space for improvement because with the current description the error is not yet contained enough to apply ReaxFF in pharmacological and/or industrial substrate screening. The study of chymotrypsin inhibition has been selected to check whether the force-field is well trained to discriminate the different reaction pathways of Kunitz-BPTI inhibitor and a catalytic active substrate. The results indicate a larger stabilization of the system in the case of the complex between inhibitor and receptor. Variations of the energy between starting structures and the first intermediate step of the catalytic cycle point in the same direction of a larger stabilization in case of inhibition. It would be auspicious to have more tools to investigate biochemical reactions, e.g. the approximation of freezing all the residues and water molecules 5\AA away from the active residues is rather crude and not even efficient because the frozen atoms are anyway included in each step of the calculation even though they have been fixed. This means that an additional tool for selecting different regions of the system to be treated in a different way, i.e. reactive and non-reactive regions, would be of great advantage, because it would enable us to have a better model of protein system dynamics and at the same time the possibility of following reactions only where they are supposed to take place. As well as ReaxFF, QM-MM is another method that attempts to bridge the gap between QM and MM. It is the present day more credited method to deal with biological systems in their entirely atomic details. This level of technology is already able to provide accurate results with a low-demanding computational power. The weak points are the level of theory used in the QM part, the basis set dependency, and the definition of the boundary between QM part and the remaining part treated with a force-field based approach. The improvement of QM-MM is far beyond the pur-

pose of this essay. QM-MM method has been only used here to investigate the resistance of 5-thio-galactoside to galactosidase Action (chapter 5). From QM-MM (M062X-OPLS2005) potential surfaces we have obtained a 5-thio-galactose transition state area higher in energy compare to the 5'-galactose. By means of *ab initio* NBO steric analysis we hypothesized that part of the reason of such a difference may be due to steric repulsion between the ring heteroatom and a disulfide bridge within the active site of the protein. Such a steric repulsion maybe an impediment to the change of conformation of the ring that occurs during the reaction. Experimental investigation should be designed to rule out the hypothesis supported by our computational results. As it was pointed out in the introduction of this essay, computational chemistry will have a fundamental role in the future development of the field of enzyme technology. This role is not limited to an increased capacity of computational methods to address chemical problems more accurately and with a higher precision. This is, so to speak, only a part of the future developments needed by the community. The role of information technology has to be intended here in a broader sense. It is worth spending a couple of words in the revolution that is going to take place in the next decades due to computational science and advances in information technology. What is going to change is the way scientific information will be available to the community. For example what in the present day we are accustomed to do it is just the reading about models using search engines and electronic databases of literature. In the near future, instead of just reading about a model we will be able to access the model through network services, verify the data even at the stage of peer-review processes, use the models for our own purpose and update instantaneously the databases with our findings.

In the coming future descriptions of materials will be semantically more rich, with a greater level of structural, graphical and functional details. This will make easier a systematic, disciplined, and quantifiable approach to the design and development of the foundations of -but not only- enzyme technology. This represents a giant pace for the community. It is for sure far away from the present day line drawing of chemical structures and scientific journals reading. As Martin Rosvall and Carl Bergstrom correctly say: “science is not merely a set of ideas but also the flow of these ideas through a multipartite and highly differentiated social system”. We think that information technology and computational science will be able to address the management of this important flow of ideas, so to move a step forward our technology just to prepare ourselves for the harsh coming era of eco-sustainability.

Bibliography

- [1] Alvin M. Weinberg H. E. Goeller. The age of sustainability. *Science*, 191(4228):683–689, February 1976.
- [2] William F. Pickard. Geochemical constraints on sustainable development: Can an advanced global economy achieve long-term stability? *Global Change and Planetary Change*, 61:285–299, 2008.
- [3] Brian J. Skinner. Earth resources. *Proceedings of the National Academy of Sciences of the United States of America*, 76(9):4212–4217, September 1979.
- [4] Peter S. J. Cheetham. The use of biotransformations for the production of flavours and fragrances. *Trends in Biotechnology*, 11(11):478–488, 11 1993.
- [5] James B. Sumner. The isolation and crystallization of the enzyme urease. *Journal of Biological Chemistry*, 69(2):435–441, 1926.
- [6] R. J. Kazlauskas S. Lutz J. C. Moore K. Robins U. T. Bornscheuer, G. W. Huisman. Engineering the thirdwave of biocatalysis. *Nature*, 4 8 5:185–194, May 2012.
- [7] U. T. Bornschauer R. J. Kazlauskas. Finding better protein engineering strategies. *Nature chemical biology*, 5:526–529, 2009.
- [8] N. J. Turner. Directed evolution drives the next generation of biocatalyst. *Nature chemical biology*, 5:567–573, 2009.

- [9] Andrew M. Wollacott Lin Jiang Jason DeChancie Jamie Betker Jasmine L. Galaher Eric A. Althoff Alexandre Zanghellini Orly Dym Shira Albeck Kendall N. Houk Dan S. Tawfik David Baker Daniela Rothlisberger, Olga Khersonsky. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, 2008.
- [10] Christopher K. Savile, Jacob M. Janey, Emily C. Mundorff, Jeffrey C. Moore, Sarena Tam, William R. Jarvis, Jeffrey C. Colbeck, Anke Krebber, Fred J. Fleitz, Jos Brands, Paul N. Devine, Gjalb W. Huisman, and Gregory J. Hughes. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science*, 329(5989):305–309, 07 2010.
- [11] Aman A. Desai. Sitagliptin manufacture: A compelling tale of green chemistry, process intensification, and industrial asymmetric catalysis. *Angewandte Chemie International Edition*, 50(9):1974–1976, 2011.
- [12] James C. Liao Shota Atsumi, Taizo Hanai. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature*, 451(7174):86–89, 2008.
- [13] Sung Kuk Lee, Howard Chou, Timothy S Ham, Taek Soon Lee, and Jay D Keasling. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current Opinion in Biotechnology*, 19(6):556–563, 12 2008.
- [14] Sean Daughtry Oliver P. Peoples Kristi D. Snell Karen Bohmert-Tatarev, Susan McAvoy. High levels of bioplastic are produced in fertile transplastomic tobacco plants engineered with a synthetic operon for the production of polyhydroxybutyrate. *Plant Physiology*, 155:1690–1708, 2011.
- [15] Rebekah McKenna and David R. Nielsen. Styrene biosynthesis from glucose by engineered e. coli. *Metabolic Engineering*, 13(5):544–554, 9 2011.

-
- [16] Axel D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993.
- [17] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the collesalvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988.
- [18] Adri C. T. van Duin, Siddharth Dasgupta, Francois Lorant, and William A. Goddard. Reaxff, a reactive force field for hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, 2001.
- [19] Judith G. Voet Donald Voet. *Biochemistry*. J. Wiley and Sons, 4th edition, 2011.
- [20] Lubert Stryer, Jeremy Berg, and John Tymoczko. *Biochemistry, 5th edition*. W. H. Freeman and Co., 5th edition, 2002.
- [21] Charles R. Cantor and Paul R. Schimmel. *Biophysical Chemistry*. W. H. Freeman and Co., 2nd edition, 1980.
- [22] A. Fersht. *Enzyme Structure and Mechanism*. Freeman, 1977.
- [23] F. J. Kézdy Heck D'A Heck M. L. Bender, G. E. Clement. The correlation of the ph dependence and the stepwise mechanism of alpha-chymotrypsin-catalyzed reactions. *Journal of the American Chemical Society*, 86:3680–3689, 1964.
- [24] Henry Eyring. The activated complex in chemical reactions. *Journal of Chemical Physics*, 3:107, 1935.
- [25] Richard Wolfenden and Mark J. Snider. The depth of chemical time and the power of enzymes as catalysts. *Accounts of Chemical Research*, 34(12):938–945, 2001.

- [26] Mark J. Snider and Richard Wolfenden. The rate of spontaneous decarboxylation of amino acids. *Journal of the American Chemical Society*, 122(46):11507–11508, 2000.
- [27] Anna Radzicka and Richard Wolfenden. A proficient enzyme. *Science*, 267(5194):90–93, 01 1995.
- [28] Gregory Young Richard Wolfenden, Caroline Ridgway. Spontaneous hydrolysis of ionized phosphate monoesters and diesters and the proficiencies of phosphatases and phosphodiesterases as catalysts. *Journal of the American Chemical Society*, 120:833–834, 1998.
- [29] Richard Wolfenden, Xiangdong Lu, and Gregory Young. Spontaneous hydrolysis of glycosides. *Journal of the American Chemical Society*, 120(27):6814–6815, 1998.
- [30] Stephen L. Bearne and Richard Wolfenden. Enzymic hydration of an olefin: the burden borne by fumarase. *Journal of the American Chemical Society*, 117(37):9588–9589, 1995.
- [31] Stephen L. Bearne and Richard Wolfenden. Mandelate racemase in pieces effective concentrations of enzyme functional groups in the transition state. *Biochemistry*, 36(7):1646–1656, 1997.
- [32] Richard Wolfenden and Anna Radzicka. Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *Journal of the American Chemical Society*, 118:6105–6109, 1996.
- [33] Ben E. Evans and Richard V. Wolfenden. Catalysis of the covalent hydration of pteridine by adenosine aminohydrolase. *Biochemistry*, 12(3):392–398, 1973.
- [34] Linus Pauling. Molecular architecture and biological reactions. *Chemical Engineering News*, 24(10):1375–1377, 1946.

-
- [35] M. Levitt A. Warshel. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology*, 103(2):227 – 249, 1976.
- [36] Arieh Warshel. Energetics of enzyme catalysis. *Proceedings of the National Academy of Sciences USA*, 75(11):5250–5254, 1978.
- [37] Arieh Warshel. Electrostatic basis of structure-function correlation in proteins. *Accounts of Chemical Research*, 14(9):284–290, 1981.
- [38] Arieh Warshel, Pankaz K. Sharma, Mitsunori Kato, Yun Xiang, Hanbin Liu, and Mats H. M. Olsson. Electrostatic basis for enzyme catalysis. *Chemical Reviews*, 106(8):3210–3235, 2006.
- [39] William P. Jenks Michael I. Page. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proceedings of the National Academy of Sciences USA*, 68(8):1678–1683, August 1971.
- [40] Peter A. Kollman, Bernd Kuhn, Oreola Donini, Mikael Perakyla, Rob Stanton, and Dirk Bakowies. Elucidating the nature of enzyme catalysis utilizing a new twist on an old methodology: Quantum mechanical free energy calculations on chemical reactions in enzymes and in aqueous solution. *Accounts of Chemical Research*, 34(1):72–79, 2001. PMID: 11170358.
- [41] Daniel E. Koshland Andrew D Mesecar, Barry L. Stoddard. Orbital steering in the catalytic power of enzymes: Small structural changes with large catalytic consequences. *Science*, 277(5323):202–206, 1997.
- [42] Dorothee Kern Katherine Henzler-Wildman. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.
- [43] Jens Ulstrup Hitoshi Sumi. Dynamics of protein conformational fluctuation in enzyme catalysis with special attention to proton transfer in serine proteinases. *Biochimica et Biophysica Acta*, 995(1):26–42, 1988.

- [44] R.R. Dogonadze, A.M. Kuznetsov, and J. Ulstrup. Conformational dynamics in biological electron and atom transfer reactions. *Journal of Theoretical Biology*, 69(2):239 – 263, 1977.
- [45] P. W. Atkins; R. S. Friedman. *Molecular Quantum Mechanics*. third edition edition, 1997.
- [46] A. Szabo; N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. 1989.
- [47] Frank Jensen. *Introduction to Computational Chemistry*. 1999.
- [48] Max C. Holthausen Wolfram Koch; Max C. Holthausen Wolfram Koch. *A Chemist's Guide to Density Functional Theory*. 2nd edition edition, 2001.
- [49] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 01 1988.
- [50] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Canadian Journal of Physics*, 58(8):1200–1211, 2013/06/16 1980.
- [51] Donald G. Truhlar Yan Zhao. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts Theory, Computation, and Modeling*, 120:215–241, 2008.
- [52] J. K. Badenhoop and F. Weinhold. Natural bond orbital analysis of steric interactions. *The Journal of Chemical Physics*, 107(14):5406–5421, 1997.
- [53] Phillip A. Christiansen and William E. Palke. Effects of exchange energy and orbital orthogonality on barriers to internal rotation. *The Journal of Chemical Physics*, 67(1):57–63, 1977.

-
- [54] J. Peter Toennies. On the validity of a modified buckingham potential for the rare gas dimers at intermediate distances. *Chemical Physics Letters*, 20(3):238–241, 1973.
- [55] J. Peter Toennies, Wolfgang Welz, and Gunther Wolf. Molecular beam scattering studies of orbiting resonances and the determination of van der waals potentials for rare gas dimers. *The Journal of Chemical Physics*, 71(2):614–642, 1979.
- [56] U. Chandra Singh and Peter A. Kollman. A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the exchange reaction and gas phase protonation of polyethers. *Journal of Computational Chemistry*, 7(6):718–730, 1986.
- [57] Martin J. Field, Paul A. Bash, and Martin Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry*, 11(6):700–733, 1990.
- [58] Hans Martin Senn and Walter Thiel. Qm/mm methods for biomolecular systems. *Angewandte Chemie International Edition*, 48(7):1198–1229, 2009.
- [59] Kimberly Chenoweth, Adri C. T. van Duin, and William A. Goddard. Reaxff reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *The Journal of Physical Chemistry A*, 112(5):1040–1053, 2008. PMID: 18197648.
- [60] Kevin D. Nielson, Adri C. T. van Duin, Jonas Oxgaard, Wei-Qiao Deng, and William A. Goddard. Development of the reaxff reactive force field for describing transition metal catalyzed reactions, with application to the initial stages of the catalytic formation of carbon nanotubes. *The Journal of Physical Chemistry A*, 109(3):493–499, 2005.

- [61] Alejandro Strachan, Edward M. Kober, Adri C. T. van Duin, Jonas Oxgaard, and William A. Goddard III. Thermal decomposition of rdx from reactive molecular dynamics. *The Journal of Chemical Physics*, 122(5):054502, 2005.
- [62] David Raymond, Adri C. T. van Duin, Daniel Spångberg, William A. Goddard III, and Kersti Hermansson. Water adsorption on stepped zno surfaces from md simulation. *Surface Science*, 604(9–10):741–752, 5 2010.
- [63] Susanna Monti, Adri C. T. van Duin, Sung-Yup Kim, and Vincenzo Barone. Exploration of the conformational and reactive dynamics of glycine and diglycine on tio2: Computational investigations in the gas phase and in solution. *The Journal of Physical Chemistry C*, 116(8):5141–5150, 2012.
- [64] Anant D. Kulkarni, Donald G. Truhlar, Sriram Goverapet Srinivasan, Adri C. T. van Duin, Paul Norman, and Thomas E. Schwartzentruber. Oxygen interactions with silica surfaces: Coupled cluster and density functional investigation and the development of a new reaxff potential. *The Journal of Physical Chemistry C*, 117(1):258–269, 2013/01/21 2012.
- [65] Yue Zhang Adri C. T. van Duin William A. Goddard III Dennis R. Salahub Rui Zhu, Florian Janetzko. Characterization of the active site of yeast rna polymerase ii by dft and reaxff calculations. *Theoretical Chemical Account*, 120:479–489, 2008.
- [66] Obaidur Rahaman, Adri C. T. van Duin, William A. Goddard, and Douglas J. Doren. Development of a reaxff reactive force field for glycine and application to solvent effect and tautomerization. *The Journal of Physical Chemistry B*, 115(2):249–261, 2011.
- [67] PierreO. Hubin, Denis Jacquemin, Laurence Leherte, Jean-Marie André, AdriC. T. Duin, and DanielP. Vercauteren. Ab initio quantum chemical

- and reaxff-based study of the intramolecular iminium–enamine conversion in a proline-catalyzed reaction. 131(8), 2012.
- [68] Adri C. T. van Duin and Jaap S. Sinninghe Damsté. Computational chemical investigation into isorenieratene cyclisation. *Organic Geochemistry*, 34(4):515–526, 4 2003.
- [69] Nan Chen, Mark T. Lusk, Adri C. T. van Duin, and III Goddard, William A. Mechanical properties of connected carbon nanorings via molecular dynamics simulation. *Physical Review B*, 72(8):085416–, 08 2005.
- [70] Sang Soo Han, Jeung Ku Kang, Hyuck Mo Lee, Adri C. T. van Duin William A. Goddard, and III. Liquefaction of hydrogen molecules upon exterior surfaces of carbon nanotube bundles. *Applied Physics Letters*, 86(20):203108–3, 05 2005.
- [71] Wilfried J. Mortier, Swapan K. Ghosh, and S. Shankar. Electronegativity-equalization method for the calculation of atomic charges in molecules. *Journal of the American Chemical Society*, 108(15):4315–4320, 1986.
- [72] Geert O. A. Janssens, Bart G. Baekelandt, Helge Toufar, Wilfried J. Mortier, and Robert A. Schoonheydt. Comparison of cluster and infinite crystal calculations on zeolites with the electronegativity equalization method (eem). *The Journal of Physical Chemistry*, 99(10):3251–3258, 1995.
- [73] Anthony K. Rappe and William A. Goddard. Charge equilibration for molecular dynamics simulations. *The Journal of Physical Chemistry*, 95(8):3358–3363, 1991.
- [74] Adri C. T. van Duin, Jan M. A. Baas, and Bastiaan van de Graaf. Delft molecular mechanics: a new approach to hydrocarbon force fields. inclusion of a geometry-dependent charge calculation. *Journal Chemical Society Faraday Transactions*, 90:2881–2895, 1994.

- [75] Clarence Schutt Herschel Rabitz Roberta Susnow, Robert B. Nachbar. Sensitivity of molecular structure to intramolecular potentials. *Journal of Physical Chemistry*, 95:8585–8597, 1991.
- [76] Herschel Rabitz Roberta Susnow, Clarence Schutt. Principal component analysis of dipeptides. *Journal of Computational Chemistry*, 15(9):963–980, 1994.
- [77] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948, 1995.
- [78] Yun Kyung Shin Peter Oelschlaeger Adri C. T. van Duin Vincenzo Barone Susanna Monti, Alessandro Corozzi; Peter Frstrup; Kaushik L. Joshi. Development of a reaxff reactive force field for molecular dynamics simulations of amino acids, peptides, proteins and enzymes. *The Journal of Physical Chemistry C*, none(9):none, 2013.
- [79] Joseph C. Fogarty, Hasan Metin Aktulga, Ananth Y. Grama, Adri C. T. van Duin, and Sagar A. Pandit. A reactive molecular dynamics simulation of the silica-water interface. *The Journal of Chemical Physics*, 132(17):174704–10, 05 2010.
- [80] Kimberly Chenoweth, Adri C. T. van Duin, and William A. Goddard. Reaxff reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *The Journal of Physical Chemistry A*, 112(5):1040–1053, 2008. PMID: 18197648.
- [81] R. Ditchfield, W. J. Hehre, and J. A. Pople. Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics*, 54(2):724–728, 1971.
- [82] P. Selvarengan and P. Kolandaivel. Potential energy surface study on glycine, alanine and their zwitterionic forms. *Journal of Molecular Structure: THEOCHEM*, 671(1–3):77–86, 2 2004.

-
- [83] Attila G. Csaszar. Conformers of gaseous glycine. *Journal of the American Chemical Society*, 114(24):9568–9575, 1992.
- [84] Ching Han Hu, Mingzuo Shen, and Henry F. Schaefer. Glycine conformational analysis. *Journal of the American Chemical Society*, 115(7):2923–2929, 1993.
- [85] Jan H. Jensen and Mark S. Gordon. Conformational potential energy surface of glycine: a theoretical study. *Journal of the American Chemical Society*, 113(21):7917–7924, 1991.
- [86] Kui Zhang and Alice Chung-Phillips. Gas-phase basicity of glycine, a comprehensive ab initio study. *The Journal of Physical Chemistry A*, 102(20):3625–3634, 1998.
- [87] S. J. McGlone, P. S. Elmes, R. D. Brown, and P. D. Godfrey. Molecular structure of a conformer of glycine by microwave spectroscopy. *Journal of Molecular Structure*, 485–486(0):225–238, 8 1999.
- [88] Thomas F. Miller, III and David C. Clary. Quantum free energies of the conformers of glycine on an ab initio potential energy surface. *Phys. Chem. Chem. Phys.*, 6:2563–2571, 2004.
- [89] Igor D. Reva, Alexander M. Plokhotnichenko, Stepan G. Stepanian, Alexander Yu. Ivanov, Eugeni D. Radchenko, Galina G. Sheina, and Yuri P. Blagoi. The rotamerization of conformers of glycine isolated in inert gas matrices. an infrared spectroscopic study. *Chemical Physics Letters*, 232(1–2):141–148, 1 1995.
- [90] A. Yu. Ivanov, A. M. Plokhotnichenko, V. Izvekov, G. G. Sheina, and Yu. P. Blagoi. Ftir investigation of the effect of matrices (kr, ar, ne) on the uv-induced isomerization of the monomeric links of biopolymers. *Journal of Molecular Structure*, 408–409(0):459–462, 6 1997.
- [91] S. G. Stepanian, I. D. Reva, E. D. Radchenko, M. T. S. Rosado, M. L. T. S. Duarte, R. Fausto, and L. Adamowicz. Matrix-isolation infrared and theoret-

- ical studies of the glycine conformers. *The Journal of Physical Chemistry A*, 102(6):1041–1054, 1998.
- [92] A. Yu Ivanov, G Sheina, and Yu. P Blagoi. Ftir spectroscopic study of the uv-induced rotamerization of glycine in the low temperature matrices (kr, ar, ne). *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 55(1):219–228, 12 1998.
- [93] Harrell L. Sellers and Lothar Schafer. Investigations concerning the apparent contradiction between the microwave structure and the ab initio calculations of glycine. *Journal of the American Chemical Society*, 100(24):7728–7729, 1978.
- [94] Lothar Schaefer, H. L. Sellers, F. J. Lovas, and R. D. Suenram. Theory versus experiment: the case of glycine. *Journal of the American Chemical Society*, 102(21):6566–6568, 1980.
- [95] R. D. Suenram and F. J. Lovas. Millimeter wave spectrum of glycine. a new conformer. *Journal of the American Chemical Society*, 102(24):7180–7184, 1980.
- [96] Kinya Iijima, Kumiko Tanaka, and Shigeki Onuma. Main conformer of gaseous glycine: molecular structure and rotational barrier from electron diffraction data and rotational constants. *Journal of Molecular Structure*, 246(3–4):257–266, 6 1991.
- [97] Galina M. Chaban and R. Benny Gerber. Anharmonic vibrational spectroscopy of the glycine–water complex: Calculations for ab initio, empirical, and hybrid quantum mechanics/ molecular mechanics potentials. *The Journal of Chemical Physics*, 115(3):1340–1348, 2001.
- [98] Weizhou Wang, Wenxu Zheng, Xuemei Pu, Ning-Bew Wong, and Anmin Tian. The 1:1 glycine–water complex: some theoretical observations. *Journal of Molecular Structure: THEOCHEM*, 618(3):235–244, 11 2002.

- [99] Yanbo Ding and Karsten Krogh-Jespersen. The 1:1 glycine zwitterion-water complex: An ab initio electronic structure study. *Journal of Computational Chemistry*, 17(3):338–349, 1996.
- [100] Yanbo Ding and Karsten Krogh-Jespersen. The 1:1 glycine zwitterion-water complex: An ab initio electronic structure study. *Journal of Computational Chemistry*, 17(3):338–349, 1996.
- [101] JoséL. Alonso, Emilio J. Cocinero, Alberto Lesarri, M. Eugenia Sanz, and Juan C. López. The glycine–water complex. *Angewandte Chemie International Edition*, 45(21):3471–3474, 2006.
- [102] Jakub Kaminský and Frank Jensen. Force field modeling of amino acid conformational energies. *Journal of Chemical Theory and Computation*, 3(5):1774–1788, 2007.
- [103] Robert A. DiStasio, Yousung Jung, and Martin Head-Gordon. A resolution of the identity implementation of the local triatomics in molecules model for second order moller plesset perturbation theory with application to alanine tetrapeptide conformational energies. *Journal of Chemical Theory and Computation*, 1(5):862–876, 2005.
- [104] H. Valdes, K. Pluhackova, and P. Hobza. Phenylalanyl-glycyl-phenylalanine tripeptide: A model system for aromatic-aromatic side chain interactions in proteins. *Journal of Chemical Theory and Computation*, 5(9):2248–2256, 2009.
- [105] P. Hobza H. Valdes, K. Pluhackova. *Phys. Chem. Chem. Phys.*, 10:2747, 2008.
- [106] D. Řeha, H. Valdés, J. Vondrášek, P. Hobza, Ali Abu-Riziq, Bridgit Crews, and Mattanjah S. de Vries. Structure and ir spectrum of phenylalanyl–glycyl–glycine tripeptide in the gas-phase: Ir/uv experiments, ab initio quantum chemical calculations, and molecular dynamic simulations. *Chemistry – A European Journal*, 11(23):6803–6817, 2005.

- [107] Andrea Schmidt, Martha Teeter, Edgar Weckert, and Victor S. Lamzin. Crystal structure of small protein crambin at 0.48Å resolution. *Acta Crystallographica Section F*, 67(4):424–428, Apr 2011.
- [108] Martha M. Teeter and Wayne A. Hendrickson. Highly ordered crystals of the plant seed protein crambin. *Journal of Molecular Biology*, 127(2):219–223, 1 1979.
- [109] Martha M. Teeter, S. Mark Roe, and Nam Ho Heo. Atomic resolution (0.83 Å) crystal structure of the hydrophobic protein crambin at 130 k. *Journal of Molecular Biology*, 230(1):292–311, 3 1993.
- [110] Jonathan W. Neidigh, R. Matthew Fesinmeyer, and Niels H. Andersen. Designing a 20-residue protein. *Nat Struct Mol Biol*, 9(6):425–430, 06 2002.
- [111] Francisco J. Blanco, German Rivas, and Luis Serrano. A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nat Struct Mol Biol*, 1(9):584–590, 09 1994.
- [112] P. S. Kim D. L. J. Minor. Measurement of the beta-sheet propensity of aminoacids. *Nature*, 367:660.663, 1994.
- [113] J. Martin Scholtz, Doug Barrick, Eunice J. York, John M. Stewart, and Robert L. Baldwin. Urea unfolding of peptide helices as a model for interpreting protein unfolding. *Proceedings of the National Academy of Sciences of the United States of America*, 92(1):185–189, 01 1995.
- [114] Andrzej Bierzynski, Peter S. Kim, and Robert L. Baldwin. A salt bridge stabilizes the helix formed by isolated c-peptide of rnae a. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8):2470–2474, 04 1982.
- [115] Andrea G. Cochran, Nicholas J. Skelton, and Melissa A. Starovasnik. Tryptophan zippers: Stable, monomeric beta-hairpins. *Proceedings of the National Academy of Sciences*, 98(10):5578–5583, 2001.

- [116] D. A. Case et al. Amber 10, university of california, san francisco, 2008.
- [117] C. C. Huang G. S. Couch D. M. Greenblatt E. C. Meng T. E. Ferrin E. F. Pettersen, T. D. Goddard. *Journal of Computational Chemistry*, 25:1605, 2004.
- [118] W. L. Jorgensen. *Journal of American Chemical Society*, 103:335–350, 1981.
- [119] P. P. Ewald. *Annal of Physics*, 64:253, 1921.
- [120] W. F. V. Gunsteren A. DiNola J. R. Haak H. J. C. Berendsen, J. P. M. Postma. *Journal of Chemical Physics*, 81:3684, 1984.
- [121] M. P. Allen. *Computer Simulation of liquids*. Clarendon Press, 1989.
- [122] C. Sander W. Kabsch. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.
- [123] A. Korostelev M. S. Chapman F. Fabiola, R. Bertram. *Protein Science*, 11:1415–1423, 2002.
- [124] Duhee Bang, Valentina Tereshko, Anthony A. Kossiakoff, and Stephen B. H. Kent. Role of a salt bridge in the model protein crambin explored by chemical protein synthesis: X-ray structure of a unique protein analogue, [v15a]crambin-[small alpha]-carboxamide. *Mol. BioSyst.*, 5:750–756, 2009.
- [125] Frances H. Arnold. Protein design for non-aqueous solvents. *Protein Engineering*, 2(1):21–25, 1988.
- [126] M. Scott Shell, Ryan Ritterson, and Ken A. Dill. A test on peptide stability of amber force fields with implicit solvation. *The Journal of Physical Chemistry B*, 112(22):6878–6886, 2008. PMID: 18471007.
- [127] Earl W. Davie and Oscar D. Ratnoff. Waterfall sequence for intrinsic blood clotting. *Science*, 145(3638):1310–1312, 09 1964.

- [128] David C. Whitcomb and Mark E. Lowe. Human pancreatic digestive enzymes. 52(1), 2007.
- [129] R. M. Lane; S. G. Potkin; A. Enz. Targeting acetylcholinesterase and butyrylcholinesterase in dementia. *International Journal of Neuropsychopharmacology*, 9:101–124, 2006.
- [130] Joseph V. Bonventre, Zhihong Huang, M. Reza Taheri, Eileen O’Leary, En Li, Michael A. Moskowitz, and Adam Sapirstein. Reduced fertility and postischaemic brain injury in mice deficient in cytosolic phospholipase a2. *Nature*, 390(6660):622–625, 12 1997.
- [131] Javier A. Menendez and Ruth Lupu. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer*, 7(10):763–777, 10 2007.
- [132] Gabriel M. Simon and Benjamin F. Cravatt. Activity-based proteomics of enzyme superfamilies: Serine hydrolases as a case study. *Journal of Biological Chemistry*, 285(15):11051–11055, 04 2010.
- [133] Daniel K. Nomura, Jonathan Z. Long, Sherry Niessen, Heather S. Hoover, Shu-Wing Ng, and Benjamin F. Cravatt. Monoacylglycerol lipase regulates a fatty acid network that promotes cancer pathogenesis. *Cell*, 140(1):49–61, 1 2010.
- [134] Uwe Theo Bornscheuer; Romas Joseph Kazlauskas. *Hydrolases in Organic Synthesis*. Wiley, 2006.
- [135] David M. Blow. The tortuous story of asp...his...ser: Structural analysis of alpha-chymotrypsin. *Trends in Biochemical Sciences*, 22(10):405–408, 10 1997.
- [136] M. N. G. James. *Canadian Journal of Biochemistry*, 58:251, 1980.
- [137] Hans Neurath. Evolution of proteolytic enzymes. *Science*, 224(4647):350–357, 04 1984.

-
- [138] D. E. Johnson. Noncaspase proteases in apoptosis. *Leukemia*, 14:1695–1703, 2000.
- [139] Kaplan Allen P. Joseph Kusuman, Ghebrehiwet Berhane. Activation of the kinin-forming cascade on the surface of endothelial cells. *Biological Chemistry*, 382:71–75, 2001.
- [140] Shaun R. Coughlin. Thrombin signalling and protease-activated receptors. *Nature*, 407(6801):258–264, 2000.
- [141] C. Barros, J. A. Crosby, and R. D. Moreno. Early steps of sperm–egg interactions during mammalian fertilization. *Cell Biology International*, 20(1):33–39, 1 1996.
- [142] Donald J. Davidson, Deborah L. Higgins, and Francis J. Castellino. Plasminogen activator activities of equimolar complexes of streptokinase with variant recombinant plasminogens. *Biochemistry*, 29(14):3585–3590, 1990.
- [143] A. Laich R. B. Sim. Serine proteases of the complement system serine proteases of the complement system serine proteases of the complement system serine proteases of the complement system. *Biochemical Society Transactions*, 28:545–550, 2000.
- [144] H. R. Lijnen D. Collen. *CRC Crit. Rev. Oncol. Hematol.*, 4:249–301, 1986.
- [145] Mark R. Wormald Raymond A. Dwek Pauline M. Rudd Philippe E. Van den Steen, Ghislain Opendakker. Matrix remodelling enzymes, the protease cascade and glycosylation. *Biochimica et Biophysica Acta*, 1528:61–73, 2001.
- [146] Toshihiko Takeuchi Charles S. Craik Zena Werb Sushma Selvarajan, Leif R. Lund. A plasma kallikrein-dependent plasminogen cascade required for adipocyte differentiation. *Nature Cell Biology*, 3:267–275, 2001.
- [147] Ellen K LeMosy, Charles C Hong, and Carl Hashimoto. Signal transduction by a protease cascade. *Trends in Cell Biology*, 9(3):102–107, 3 1999.

- [148] Joseph Kraut. Serine proteases: Structure and mechanism of catalysis. *Annual Review of Biochemistry*, 46:331, 1977.
- [149] Wojciech R. Rypniewski, Peter R. Østergaard, Mads Nørregaard-Madsen, Mirosława Dauter, and Keith S. Wilson. *Fusarium oxysporum* trypsin at atomic resolution at 100 and 283K: a study of ligand binding. *Acta Crystallographica Section D*, 57(1):8–19, Jan 2001.
- [150] Z. S. Derewenda, U. Derewenda, and P. M. Kobos. Hydrogen bond in the active sites of serine hydrolases. *Journal of Molecular Biology*, 241(1):83–93, 8 1994.
- [151] Richard David Gandour. On the importance of orientation in general base catalysis by carboxylate. *Bioorganic Chemistry*, 10(2):169–176, 6 1981.
- [152] Israel Schechter and Arieh Berger. On the size of the active site in proteases. i. papain. *Biochemical and Biophysical Research Communications*, 27(2):157–162, 4 1967.
- [153] R Huber, D Kukla, W Bode, P Schwager, K Bartels, J Deisenhofer, and W Steigemann. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. ii. crystallographic refinement at 1.9 a resolution. *J Mol Biol*, 89(1):73–101, Oct 1974.
- [154] M. Krieger, L. M. Kay, and R. M. Stroud. Structure and specific binding of trypsin: Comparison of inhibited derivatives and a model for substrate binding. *Journal of Molecular Biology*, 83(2):209–230, 2 1974.
- [155] John J. Perona, Lizbeth Hedstrom, William J. Rutter, and Robert J. Fletterick. Structural origins of substrate discrimination in trypsin and chymotrypsin. *Biochemistry*, 34(5):1489–1499, 1995.
- [156] Robert Huber and Wolfram Bode. Structural basis of the activation and action of trypsin. *Accounts of Chemical Research*, 11(3):114–122, 1978.

- [157] Gary S. Coombs, Mohan S. Rao, Arthur J. Olson, Philip E. Dawson, and Edwin L. Madison. Revisiting catalysis by chymotrypsin family serine proteases using peptide substrates and inhibitors with unnatural main chains. *Journal of Biological Chemistry*, 274(34):24074–24079, 1999.
- [158] Charles A. Kettner, Roger Bone, David A. Agard, and William W. Bachovchin. Kinetic properties of the binding of .alpha.-lytic protease to peptide boronic acids. *Biochemistry*, 27(20):7682–7688, 1988.
- [159] Robert C. Thompson. Use of peptide aldehydes to generate transition-state analogs of elastase. *Biochemistry*, 12(1):47–51, 1973.
- [160] J. V. Killheffer M. Bender. Chymotrypsins. *Critical Reviews in Biochemistry*, 1:149, 1973.
- [161] T A Steitz and R G Shulman. Crystallographic and nmr studies of the serine proteases. *Annual Review of Biophysics and Bioengineering*, 11(1):419–444, 2013/02/27 1982.
- [162] William W. Bachovchin, Winona Y. L. Wong, Shauna Farr-Jones, Ashok B. Shenvi, and Charles A. Kettner. Nitrogen-15 nmr spectroscopy of the catalytic triad histidine of a serine protease in peptide boronic acid inhibitor complexes. *Biochemistry*, 27(20):7689–7697, 1988.
- [163] Tzyy Chyau Liang and Robert H. Abeles. Complex of .alpha.-chymotrypsin and n-acetyl-l-leucyl-l-phenylalanyl trifluoromethyl ketone: structural studies with nmr spectroscopy. *Biochemistry*, 26(24):7603–7608, 1987.
- [164] Jing Lin, William M. Westler, W. Wallace Cleland, John L. Markley, and Perry A. Frey. Fractionation factors and activation energies for exchange of the low barrier hydrogen bonding proton in peptidyl trifluoromethyl ketone complexes of chymotrypsin. *Proceedings of the National Academy of Sciences*, 95(25):14664–14668, 1998.

- [165] S. A. Spencer A. A. Kossiakoff. Neutron diffraction identifies his 57 as the catalytic base in trypsin ti - neutron diffraction identifies his 57 as the catalytic base in trypsinneutron diffraction identifies his 57 as the catalytic base in trypsin. *Nature*, 288(5789):414–416, 1980.
- [166] Hans Neurath. Proteolytic processing and physiological regulation. *Trends in Biochemical Sciences*, 14(7):268–271, 7 1989.
- [167] Amir R. Khan and Micael N. G. James. Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Science*, 7(4):815–836, 1998.
- [168] Wolfram Bode and Robert Huber. Structural basis of the endoproteinase–protein inhibitor interaction. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1477(1–2):241–252, 3 2000.
- [169] M Laskowski and I Kato. Protein inhibitors of proteinases. *Annual Review of Biochemistry*, 49(1):593–626, 2013/06/15 1980.
- [170] D. Krowarsch, T. Cierpicki, F. Jelen, and J. Otlewski. Canonical protein inhibitors of serine proteases. 60(11), 2003.
- [171] S Jones and J M Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, 93(1):13–20, 01 1996.
- [172] Włodzimierz Apostoluk and Jacek Otlewski. Variability of the canonical loop conformations in serine proteinases inhibitors and other proteins. *Proteins: Structure, Function, and Bioinformatics*, 32(4):459–474, 1998.
- [173] Richard M Jackson and Robert B Russell. The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *Journal of Molecular Biology*, 296(2):325–334, 2 2000.
- [174] David A. Estell, Karl A. Wilson, and Michael Laskowski. Thermodynamics and kinetics of the hydrolysis of the reactive-site peptide bond in pancreatic

- trypsin inhibitor (kunitz) by *dermasterias imbricata* trypsin 1. *Biochemistry*, 19(1):131–137, 2013/06/15 1980.
- [175] Spartaco A. Bizzozero and Hans Dutler. Comparative specificity of porcine pancreatic kallikrein and bovine pancreatic trypsin: Importance of interactions n-terminal to the scissible bond. *Archives of Biochemistry and Biophysics*, 256(2):662–676, 8 1987.
- [176] M. Marquart, J. Walter, J. Deisenhofer, W. Bode, and R. Huber. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. *Acta Crystallographica Section B*, 39(4):480–490, Aug 1983.
- [177] Frances C. Bernstein, Thomas F. Koetzle, Grahame J.B. Williams, Edgar F. Meyer Jr., Michael D. Brice, John R. Rodgers, Olga Kennard, Takehiko Shimanouchi, and Mitsuo Tasumi. The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535 – 542, 1977.
- [178] Jay L. Banks, Hege S. Beard, Yixiang Cao, Art E. Cho, Wolfgang Damm, Ramy Farid, Anthony K. Felts, Thomas A. Halgren, Daniel T. Mainz, Jon R. Maple, Robert Murphy, Dean M. Philipp, Matthew P. Repasky, Linda Y. Zhang, Bruce J. Berne, Richard A. Friesner, Emilio Gallicchio, and Ronald M. Levy. Integrated modeling program, applied chemical theory (impact). *Journal of Computational Chemistry*, 26(16):1752–1780, 2005.
- [179] William Humphrey, Andrew Dalke, and Klaus Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 2 1996.
- [180] Tariq A. Andrea, William C. Swope, and Hans C. Andersen. The role of long ranged forces in determining the structure and properties of liquid water. *The Journal of Chemical Physics*, 79(9):4576–4584, 1983.

- [181] C. Fonseca Guerra, J. G. Snijders, G. te Velde, and E. J. Baerends. Towards an order-n dft method. *Theoretical Chemistry Accounts*, 99(6):391–403, 1998.
- [182] G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, and T. Ziegler. Chemistry with adf. *Journal of Computational Chemistry*, 22(9):931–967, 2001.
- [183] Tom Wennekes; Richard J.B.H.N. van den Berg; Rolf G. Boot; Gijsbert A. van der Marel; Herman S. Overkleeft; Johannes M.F.G. Aerts. Glycosphingolipids: Nature, function, and pharmacological modulation. *Angewandte Chemie International Edition*, 48(47):8848–8869, 2009.
- [184] G. Tettamanti. Ganglioside/glycosphingolipid turnover: New concepts. *Glycoconjugate Journal*, 20(5):301–317, 2003.
- [185] Sen itiroh Hakomori. Tumor malignancy defined by aberrant glycosylation and sphingo(glyco)lipid metabolism. *Perspectives in Cancer Research*, 56(5309-5318), 1996.
- [186] Scott C. Garman and David N. Garboczi. The molecular defect leading to fabry disease: Structure of human alpha-galactosidase. *Journal of Molecular Biology*, 337(2):319–335, 3 2004.
- [187] Scott C Garman. Structure–function relationships in alpha-galactosidase a. *Acta Pædiatrica*, 96:6–16, 2007.
- [188] Dominique P Germain. Fabry disease. *Orphanet Journal of Rare Diseases*, 5(30), 2010.
- [189] Hideya Yuasa, Ole Hindsgaul, and Monica M. Palcic. Chemical-enzymic synthesis of 5'-thio-n-acetyllactosamine: the first disaccharide with sulfur in the ring of the non-reducing sugar. *Journal of the American Chemical Society*, 114(14):5891–5892, 1992.

- [190] Jose G. Fernández-Bolaños, Najim A. L. Al-Masoudi, and Inés and Maya. Sugar derivatives having sulfur in the ring. *Advances in Carbohydrate Chemistry and Biochemistry*, Volume 57:21–98, 2001.
- [191] Zbigniew J. Witczak Inmaculada Robina, Pierre Vogel. Synthesis and biological properties of monothiosaccharides. *Current Organic Chemistry*, 5(12):1177–1214, 2001.
- [192] Alonso Aguirre-Valderrama and Jose A. Dobado. Conformational analysis of thiosugars: Theoretical nmr chemical shifts and $^3J_{H,H}$ coupling constants of 5-thio-pyranose monosaccharides. *Journal of Carbohydrate Chemistry*, 25(7):557–594, 2006.
- [193] Dietlind Adlercreutz, Yayoi Yoshimura, Karin Mannerstedt, Warren W. Wakarchuk, Eric P. Bennett, Norman J. Dovichi, Ole Hindsgaul, and Monica M. Palcic. Thiogalactopyranosides are resistant to hydrolysis by alpha-galactosidases. *ChemBioChem*, 13(11):1673–1679, 2012.
- [194] Xiao-Liang Pan, Wei Liu, and Jing-Yao Liu. Mechanism of the glycosylation step catalyzed by human alpha-galactosidase: A qm/mm metadynamics study. *The Journal of Physical Chemistry B*, 117(2):484–489, 2013.
- [195] Qiyong P. Liu, Huaiping Yuan, Eric P. Bennett, Steven B. Levery, Edward Nudelman, Jean Spence, Greg Pietz, Kristen Saunders, Thayer White, Martin L. Olsson, Bernard Henrissat, Gerlind Sulzenbacher, and Henrik Clausen. Identification of a gh110 subfamily of alpha-1,3-galactosidases: Novel enzyme removal of the alpha-3gal xenotransplantation antigen. *Journal of Biological Chemistry*, 283(13):8545–8554, 2008.
- [196] Dietlind Adlercreutz, Joel T. Weadge, Bent O. Petersen, Jens Ø. Duus, Norman J. Dovichi, and Monica M. Palcic. Enzymatic synthesis of gb3 and igb3 ceramides. *Carbohydrate Research*, 345(10):1384–1388, 7 2010.

- [197] Sharon Ahmad. igb3: to be or not to be? *Nature Reviews Immunology*, 7(5):325, 2007.
- [198] Abigail I. Guce, Nathaniel E. Clark, Eric N. Salgado, Dina R. Ivanen, Anna A. Kulminskaya, Harry Brumer, and Scott C. Garman. Catalytic mechanism of human alpha-galactosidase. *Journal of Biological Chemistry*, 285(6):3625–3632, 2010.
- [199] Brian P Rempel and Stephen G Withers. Covalent inhibitors of glycosidases and their applications in biochemistry and biology. *Glycobiology*, 18(8):570–586, 2008.
- [200] K.J. Bowers, E. Chow, Huafeng Xu, R.O. Dror, M.P. Eastwood, B.A. Gregersen, J.L. Klepeis, I. Kolossvary, M.A. Moraes, F.D. Sacerdoti, J.K. Salmon, Yibing Shan, and D.E. Shaw. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC 2006 Conference, Proceedings of the ACM/IEEE*, page 43, nov. 2006.
- [201] Suite 2012: Schrödinger suite 2011 protein preparation wizard; epik version 2.3, schrödinger, llc, new york, ny, 2012; impact version 5.8, schrödinger, llc, new york, ny, 2012; prime version 3.1, schrödinger, llc, new york, ny, 2012.
- [202] D. J. Evans and B. L. Holian. The Nose–Hoover thermostat. *The Journal of Chemical Physics*, 83(8):4069–4074, 1985.
- [203] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. Constant pressure molecular dynamics algorithms. *The Journal of Chemical Physics*, 101(5):4177–4189, 1994.
- [204] Qsite, version 5.7, schrödinger, llc, new york, ny. qsite, version 5.7, schrödinger, llc, new york, ny, 2011., 2011.
- [205] R. B. Murphy, D. M. Philipp, and R. A. Friesner. A mixed quantum mechanics/molecular mechanics (qm/mm) method for large-scale modeling of chemistry

- in protein environments. *Journal of Computational Chemistry*, 21(16):1442–1457, 2000.
- [206] Richard A. Friesner and Victor Guallar. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (qm/mm) methods for studying enzymatic catalysis. *Annual Review of Physical Chemistry*, 56(1):389–427, 2005.
- [207] S. C. Harvey and P. Hoekstra. Dielectric relaxation spectra of water adsorbed on lysozyme. *The Journal of Physical Chemistry*, 76(21):2987–2994, 1972.
- [208] Nbo 5.0. e.d. glendening, j. k. badenhoop, a. e. reed, j. e. carpenter, j. a. bohmann, c.m. morales, and f. weibold (theoretical chemistry institute, university of wisconsin, madison, wi, 2001; <http://www.chem.wisc.edu/nbo5>).
- [209] Victor F. Weisskopf. Of atoms, mountains, and stars: A study in qualitative physics. *Science*, 187(4177):605–612, 02 1975.

Chapter 7

Appendix A - PCA Software

The software uses some home made function that reads, extract, manipulate and write a new ReaxFF force-field. LAMMPS (citation) software engine is then used to calculate observables. What follows are the various part of the code:

1. PCA_Matrix_Generation.m
2. my_read_reaxff2.m
3. my_write_reaxff.m
4. my_output_scanner.m
5. my_output_extractor.m

7.1 PCA_Matrix_Generation.m

```

%% READING THE FFIELD
%
% This part of the script simply read and store datas from the
% file ffield.reax.cho (ffield limited to C, H and O)
%
% FORCE FIELD PARAMETERS
%
% *****
% * Array   * ffield section           * cell structure   *
% *****
% * D3      * Atom type parameters     * atom_type        *
% * D6      * Bond type parameters     * bond_type        *
% * D3off   * Off diagonal bond type parameters * bond_type_off_diagonal *
% * D18     * Angle parameters         * angle_type       *
% * D26     * Torsion parameters       * torsion_type     *
% * DHbond  * Hydrogen bond parameters * Hbond_type       *
% *****
[ D3, D6, D3off, ...
  D18, D26, DHbond, ...
  X, atom_type, bond_type, ...
  bond_type_off_diagonal, angle_type, ...
  torsion_type, Hbond_type] = my_read_reaxff2('ffield.reax.cho');

%% REFERENCE DATAS
%
% This section of the script should be modified to
% automatize the storage of reference datas.
% It is simply a single run of Lammps using the original
% force field file and then manually edited output
% store in self exhaustive variables
%
% REFERENCE DATAS
%
% *****
% * variables * meaning

```

```
% *****
% * v_ea      * bond energy
% * v_eb      * atom energy
% * v_elp     * lone-pair energy
% * v_emol    * molecule energy (always 0.0)
% * v_ev      * valence angle energy
% * v_epen    * double-bond valence angle penalty
% * v_ecoa    * valence angle conjugation energy
% * v_ehb     * hydrogen bond energy
% * v_et      * torsion energy
% * v_eco     * conjugation energy
% * v_ew      * van der Waals energy
% * v_ep      * Coulomb energy
% * v_efi     * electric field energy (always 0.0)
% * v_eqeq    * charge equilibration energy
% *****

tot_energy = -328.10223069999;
v_ea = 160.4940239953;
v_eb = -761.24961526814;
v_elp = -3.5122068975296e-06;
v_emol = 0;
v_ev = 95.43838743188;
v_epen = 9.8170258341819e-07;
v_ecoa = 0;
v_ehb = 0;
v_et = -1.0309604475806;
v_eco = -0.025709308579911;
v_ew = 188.64158632336;
v_ep = -30.383895529237;
v_efi = 0;
v_eqeq = 20.013954633522;

%% Sensitivity matrix bond type
%
% This loop should run over all the parameters that need to be varied
% for i=1:(numero di parametri da cambiare)
% right now the script is limited to only one section of the force field
```

```
%  
  
% m: rows of the matrix D6 (keep in mind that two rows contains  
%   contains parameters)  
%  
% example: the following parameters are registered in two lines of D6  
%  
% 1 1 156.5953 100.0397 80.0000 -0.8157 -0.4591 1.0000 37.7369 0.4235  
%   0.4527 -0.1000 9.2605 1.0000 -0.0750 6.8316 1.0000 0.0000  
%  
%  
[m, n] = size(D6);  
bip = 0;  
% This is a trick to go over the all rows and columns  
%  
% it runs on each rows of D6 and then run over the columns of D6 to build  
% the sensitivity matrix. For each parameter a single run  
% of lammps is performed using the following system call:  
%  
% [status, output] = system('./lmp_4proc');  
%  
% the script lmp_4proc must be placed within the same directory  
% and it should have the following content:  
%  
% *****  
% *  
% * #!/bin/sh  
% *  
% * /usr/lib64/openmpi/bin/mpirun -np 4 ...  
% * /opt/lammps_openmpi/src/lmp_parallel_2 -in ...  
% * in.formaldehyde > formaldehyde_output  
% *  
% *****  
%  
% the output files is read by the the function my_output_extractor  
% so pay attention to the file namings  
%  
% and various call to functions have been made (off course they need to be
```

```
% placed withing the same directory of the main script):
%
% my_write_reaxff
% my_output_extractor
% my_output_scanner
%
% the Sensitivity matrix is instead built in loco (temporarily)
Sensitivity_Element = 0;
for i=1:m
  for j=1:n
    % *****
    % * back up of the matrix D6 to D6_new
    % *****
    D6_new = D6;
    % *****
    % * parameter of D6_new doubled (when it is not zero)
    % *****
    D6_new(i,j) = D6(i,j)+D6(i,j)*2;
    % *****
    % * the logarythm of the difference between
    % * parameter values before and after the parameter variation
    % *****
    dP_6(i,j) = log(D6_new(i,j) - D6(i,j));
    % *****
    % * Write the modified matrix to the file ffield.reax.cho.new
    % * calling my_write_reaxff function
    % * done = okay if the call goes through rightly
    % * done = [] if the call doesn't go through rightly
    % * all the parameter matrix are transferred, also those one
    % * that weren't modified
    % *****
    [done] = my_write_reaxff(D3, D6_new, D3off, D18, D26, DHbond, X,...
      atom_type, bond_type, bond_type_off_diagonal,...
      angle_type, torsion_type, Hbond_type)
    % *****
    % * system call to run the bast script (see at the beginning of
    % * the section if you need more information about the script)
    % *****
```

```

[status, output] = system('./lmp_4proc');
% *****
% * system call to run the bast script (see at the beginning of
% * the section if you need more information about the script)
% *****
%
[POT] = my_output_extractor('formaldehyde_output');
% *****
% * in the v_reax are stored all the variables
% * the section if you need more information about the script)
% *****
[v_reax, energy_reax] = my_output_scanner(POT);
Sensitivity_Element = Sensitivity_Element + 1;
% building each element of the sensitivity matrix
Sensitivity_Matrix_bond(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_6(i,j);
Sensitivity_Matrix_bond(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_6(i,j);
Sensitivity_Matrix_bond(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_6(i,j);
Sensitivity_Matrix_bond(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_6(i,j);
Sensitivity_Matrix_bond(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_6(i,j);
Sensitivity_Matrix_bond(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_6(i,j);
Sensitivity_Matrix_bond(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_6(i,j);
Sensitivity_Matrix_bond(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_6(i,j);
Sensitivity_Matrix_bond(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_6(i,j);
Sensitivity_Matrix_bond(10,Sensitivity_Element) = (v_reax(15) - v_eqeq)/dP_6(i,j);
bip = bip + 1;
end;
end;

%% Sensitivity matrix off diagonal bond-type
%
[m, n] = size(D3off);
bip = 0;
Sensitivity_Element = 0;
for i=1:m
    for j=1:n
        D3off_new = D3off;
        D3off_new(i,j) = D3off(i,j)+D3off(i,j)*2;
        dP_3off(i,j) = log(D3off_new(i,j) - D3off(i,j));
    end
end

```

```

[done] = my_write_reaxff(D3, D6, D3off_new, D18, D26, DHbond, X,...
    atom_type, bond_type, bond_type_off_diagonal,...
    angle_type, torsion_type, Hbond_type);
[status, output] = system('./lmp_4proc');
[POT] = my_output_extractor('formaldehyde_output');
[v_reax, energy_reax] = my_output_scanner(POT);
Sensitivity_Element = Sensitivity_Element + 1;
Sensitivity_Matrix_bond_off(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_3off(i,j);
Sensitivity_Matrix_bond_off(10,Sensitivity_Element) = (v_reax(15) - v_eqeq)/dP_3off(i,j);
bip = bip + 1;

end;
end;

%% Sensitivity matrix angle type
%

[m, n] = size(D18);
bip = 0;
Sensitivity_Element = 0;
for i=1:m
    for j=1:n
        D18_new = D18;
        D18_new(i,j) = D18(i,j)+D18(i,j)*2;
        dP_18(i,j) = log(D18_new(i,j) - D18(i,j));
        [done] = my_write_reaxff(D3, D6, D3off, D18_new, D26, DHbond, X,...
            atom_type, bond_type, bond_type_off_diagonal,...
            angle_type, torsion_type, Hbond_type);
        [status, output] = system('./lmp_4proc');
        [POT] = my_output_extractor('formaldehyde_output');

```

```

[v_reax, energy_reax] = my_output_scanner(POT)
Sensitivity_Element = Sensitivity_Element + 1;
Sensitivity_Matrix_angle(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_18(i,j);
Sensitivity_Matrix_angle(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_18(i,j);
Sensitivity_Matrix_angle(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_18(i,j);
Sensitivity_Matrix_angle(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_18(i,j);
Sensitivity_Matrix_angle(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_18(i,j);
Sensitivity_Matrix_angle(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_18(i,j);
Sensitivity_Matrix_angle(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_18(i,j);
Sensitivity_Matrix_angle(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_18(i,j);
Sensitivity_Matrix_angle(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_18(i,j);
Sensitivity_Matrix_angle(10,Sensitivity_Element) = (v_reax(15) - v_eqeq)/dP_18(i,j);
bip = bip + 1;
end;
end;

%% Sensitivity matrix torsion type
%
[m, n] = size(D26);
bip = 0;
Sensitivity_Element = 0;
for i=1:m
    for j=1:n
        D26_new = D26;
        D26_new(i,j) = D26(i,j)+D26(i,j)*2;
        dP_26(i,j) = log(D26_new(i,j) - D26(i,j));
        [done] = my_write_reaxff(D3, D6, D3off, D18, D26_new, DHbond, X,...
            atom_type, bond_type, bond_type_off_diagonal,...
            angle_type, torsion_type, Hbond_type);
        [status, output] = system('./lmp_4proc');
        [POT] = my_output_extractor('formaldehyde_output');
        [v_reax, energy_reax] = my_output_scanner(POT);
        Sensitivity_Element = Sensitivity_Element + 1;
        Sensitivity_Matrix_torsion(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_26(i,j);
        Sensitivity_Matrix_torsion(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_26(i,j);
        Sensitivity_Matrix_torsion(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_26(i,j);
        Sensitivity_Matrix_torsion(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_26(i,j);
        Sensitivity_Matrix_torsion(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_26(i,j);
    end
end

```

```

Sensitivity_Matrix_torsion(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_26(i,j);
Sensitivity_Matrix_torsion(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_26(i,j);
Sensitivity_Matrix_torsion(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_26(i,j);
Sensitivity_Matrix_torsion(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_26(i,j);
Sensitivity_Matrix_torsion(10,Sensitivity_Element) = (v_reax(15) - v_eqeq)/dP_26(i,j);
bip = bip + 1;
end;
end;

%% Sensitivity matrix atom type
% it is reported at the end because this part doesn't need to be changed
[m, n] = size(D3);
bip = 0;
Sensitivity_Element = 0;
for i=1:m
    for j=1:n
        D3_new = D3;
        D3_new(i,j) = D3(i,j)+D3(i,j)*2;
        dP_3(i,j) = log(D3_new(i,j) - D3(i,j));
        [done] = my_write_reaxff(D3_new, D6, D3off, D18, D26, DHbond, X,...
            atom_type, bond_type, bond_type_off_diagonal,...
            angle_type, torsion_type, Hbond_type);
        [status, output] = system('./lmp_4proc');
        [POT] = my_output_extractor('formaldehyde_output');
        [v_reax, energy_reax] = my_output_scanner(POT);
        Sensitivity_Element = Sensitivity_Element + 1;
        Sensitivity_Matrix_atom(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_3(i,j);
        Sensitivity_Matrix_atom(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_3(i,j);
        Sensitivity_Matrix_atom(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_3(i,j);
        Sensitivity_Matrix_atom(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_3(i,j);
        Sensitivity_Matrix_atom(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_3(i,j);
        Sensitivity_Matrix_atom(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_3(i,j);
        Sensitivity_Matrix_atom(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_3(i,j);
        Sensitivity_Matrix_atom(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_3(i,j);
        Sensitivity_Matrix_atom(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_3(i,j);
        Sensitivity_Matrix_atom(10,Sensitivity_Element) = (v_reax(15) - v_eqeq)/dP_3(i,j);
        bip = bip + 1;
    end;
end;

```

```
end;

%% Sensitivity matrix Hydrogen Bond Type
% it is reported at the end because this part doesn't need to be changed
[m, n] = size(DHbond);
bip = 0;
Sensitivity_Element = 0;
for i=1:m
    for j=1:n
        DHbond_new = DHbond;
        DHbond_new(i,j) = DHbond(i,j)+DHbond(i,j)*2;
        dP_Hbond(i,j) = log(DHbond_new(i,j) - DHbond(i,j));
        [done] = my_write_reaxff(D3, D6, D3off, D18, D26, DHbond_new, X,...
            atom_type, bond_type, bond_type_off_diagonal,...
            angle_type, torsion_type, Hbond_type);
        [status, output] = system('./lmp_4proc');
        [POT] = my_output_extractor('formaldehyde_output');
        [v_reax, energy_reax] = my_output_scanner(POT);
        Sensitivity_Element = Sensitivity_Element + 1;
        Sensitivity_Matrix_hbond(1,Sensitivity_Element) = (v_reax(3) - v_eb)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(2,Sensitivity_Element) = (v_reax(4) - v_ea)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(3,Sensitivity_Element) = (v_reax(5) - v_elp)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(4,Sensitivity_Element) = (v_reax(7) - v_ev)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(5,Sensitivity_Element) = (v_reax(8) - v_epen)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(6,Sensitivity_Element) = (v_reax(11) - v_et)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(7,Sensitivity_Element) = (v_reax(12) - v_eco)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(8,Sensitivity_Element) = (v_reax(13) - v_ew)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(9,Sensitivity_Element) = (v_reax(13) - v_ep)/dP_Hbond(i,j);
        Sensitivity_Matrix_hbond(10,Sensitivity_Element) = (v_reax(15) - v_eqq)/dP_Hbond(i,j);
        bip = bip + 1;
    end;
end;

%% Building the total Sensitivity_Matrix

Sensitivity_Matrix = [Sensitivity_Matrix_atom ...
    Sensitivity_Matrix_bond ...
```

```

        Sensitivity_Matrix_bond_off ...
        Sensitivity_Matrix_angle ...
        Sensitivity_Matrix_torsion ...
        Sensitivity_Matrix_hbond];

%% Saving the matrices generated
% *****
% * The matrix PCA is the symmetric matrix obtained
% * as the product of the sensitivity matrix
% *****
PCA = Sensitivity_Matrix'*Sensitivity_Matrix;
% *****
% * open the file named PCA where the matrix will be stores
% *****
fid = fopen('PCA', 'w');
% *****
% * Writing the PCA matrix,
% *
% * the inner loop is closed by new lineb at the
% * end of each row
% * fprintf(fid,'\n')
% *****
[m, n] = size(PCA);
for i=1:m
    for j=1:n
        fprintf(fid, '%13.8f',PCA(i,j));
    end;
    fprintf(fid,'\n');
end
fclose(fid);
% *****
% * the same operaions are done with the sensitivity matrix
% *****
fid = fopen('Sensitivity_Matrix', 'w');
[m, n] = size(Sensitivity_Matrix);
for i=1:m
    for j=1:n
        fprintf(fid, '%13.8f',Sensitivity_Matrix(i,j));
    end;
    fprintf(fid,'\n');
end
fclose(fid);

```

```
    end;
    fprintf(fid,'\n');
end
fclose(fid);
```

7.2 my_write_reaxff.m

```
function [done] = my_write_reaxff(D3, D6, D3off, D18, D26, DHbond, X, atom_type, bond_type,
    bond_type_off_diagonal, angle_type, torsion_type, Hbond_type)
lineend_header = 45;
linestart_section3 = 46;
lineend_section3 = 57;
linestart_section6 = 60;
lineend_section6 = 71;
linestart_section_offdiagonal = 73;
lineend_section_offdiagonal = 75;
linestart_section18 = 77;
lineend_section18 = 94;
linestart_section26 = 96;
lineend_section26 = 121;
linestart_sectionHbond = 123;

fid = fopen('ffield.reax.cho.new', 'w');

% Writing the headers

for a=1:lineend_header
    fprintf(fid, '%s\n', X{a});
end

% section3
[m, n] = size(D3);
for i=1:m
    fprintf(fid, '%s', atom_type{i});
    fprintf(fid, ' %8.4f', D3(i,1));
    for j=2:n-1
        fprintf(fid, ' %8.4f', D3(i,j));
    end;
    fprintf(fid, ' %8.4f', D3(i,n));
    fprintf(fid, '\n');
end
```

```
% section6
[m, n] = size(D6);
for a=lineend_section3+1:lineend_section3+2
    fprintf(fid, '%s\n', X{a});
end;
for i=1:m
    fprintf(fid, '%s', bond_type{i});
    fprintf(fid, '%8.4f', D6(i,1));
    for j=2:n-1
        fprintf(fid, ' %8.4f', D6(i,j));
    end;
    fprintf(fid, ' %8.4f', D6(i,n));
    fprintf(fid, '\n');
end

%section D3off
fprintf(fid, '%s\n', X{lineend_section6+1});
[m, n] = size(D3off);
for i=1:m
    fprintf(fid, '%s', bond_type_off_diagonal{i});
    fprintf(fid, '%8.4f', D3off(i,1));
    for j=2:n-1
        fprintf(fid, ' %8.4f', D3off(i,j));
    end;
    fprintf(fid, ' %8.4f', D3off(i,n));
    fprintf(fid, '\n');
end

% section 18
fprintf(fid, '%s\n', X{lineend_section_offdiagonal+1});
[m, n] = size(D18);
for i=1:m
    fprintf(fid, '%s', angle_type{i});
    fprintf(fid, '%8.4f', D18(i,1));
    for j=2:n-1
        fprintf(fid, ' %8.4f', D18(i,j));
    end;
end;
```

```
fprintf(fid, '%8.4f', D18(i,n));
fprintf(fid, '\n');
end

%section 26
fprintf(fid, '%s\n', X{lineend_section18+1});
[m, n] = size(D26);
for i=1:m
    fprintf(fid, '%s', torsion_type{i});
    fprintf(fid, '%8.4f', D26(i,1));
    for j=2:n-1
        fprintf(fid, '%8.4f', D26(i,j));
    end;
    fprintf(fid, '%8.4f', D26(i,n));
    fprintf(fid, '\n');
end

%section Hbond
fprintf(fid, '%s\n', X{lineend_section26+1});
fprintf(fid, '%s', Hbond_type{1});
fprintf(fid, '%8.4f', DHbond(1))
for j=2:length(DHbond)
    fprintf(fid, '%8.4f', DHbond(j));
end;
fprintf(fid, '\n')

fclose(fid);

done='okay';
end
```

7.3 my_read_reaxff2.m

```
function [D3 D6 D3off D18 D26 DHbond X atom_type bond_type bond_type_off_diagonal
        angle_type torsion_type Hbond_type ] = my_read_reaxff2(filename)

disp(['read_output_file, reading: ' filename]);
fid = fopen(filename,'r');
if ~fid
    error('ERROR: file not found');
end
counter = 1;
% feof(fid)=1 quando e' alla fine del file
% feof(fid)=0 quando NON e' alla fine del file.
while ~feof(fid)
% cell structure, it is a structure where to put datas
    X{counter} = fgetl(fid);
    counter = counter + 1;
end
fclose(fid);

%Force field unchanged parts
lineend_header = 45;
linestart_section3 = 46;
lineend_section3 = 57;
linestart_section6 = 60;
lineend_section6 = 71;
linestart_section_offdiagonal = 73;
lineend_section_offdiagonal = 75;
linestart_section18 = 77;
lineend_section18 = 94;
linestart_section26 = 96;
lineend_section26 = 121;
linestart_sectionHbond = 123;

% first + section 3
for i=1:lineend_header
D.Header{i} = X{i}(1:end);
end;
```

```
% section6
for i=lineend_section3+1:lineend_section3+2
D.Header{i} = X{i}(1:end);
end;
% section3 off diagonal
D.Header{lineend_section6+1} = X{lineend_section6+1}(1:end);
% section 18
D.Header{lineend_section_offdiagonal+1} = X{lineend_section_offdiagonal+1}(1:end);
%section 26
D.Header{lineend_section18+1} = X{lineend_section18+1}(1:end);
%section 1
D.Header{lineend_section26+1} = X{lineend_section26+1}(1:end);

%for a=1:length(D.Header)
%   disp(D.Header{a});
%end

%for a=1:length(D.Header)
%   disp(D.Header{a});
%end

% Section 3

% this read each line of section 3 and store the parameters in a matrix D3
% a test is done to know if there is an element as header, if it is it is
% stored in a cell structure called atom_type

k = 1;
for a=lineend_header+1:lineend_section3
    temp01 = X{a};
    temp02 = double(temp01);
    index_C = find(temp02==67);
    index_H = find(temp02==72);
    index_O = find(temp02==79);
    index_space = find(temp02==32);
    index_line_loop = length(index_space)-1;
    j = 1;
```

```
if index_C == 0 & index_H == 0 & index_O == 0
    for i=1:index_line_loop;
        if index_space(i+1)-index_space(i) > 1
            D3(a-45,j) = str2num(temp01(index_space(i):index_space(i+1)));
            j = j + 1;
        end;
    end;
else
    atom_type{k} = temp01(1:3);
    k = k + 1;
    for z=3:index_line_loop;
        if index_space(z+1)-index_space(z) > 1
            D3(a-45,j) = str2num(temp01(index_space(z):index_space(z+1)));
            j = j + 1;
        end;
    end;
end;
end;

% section 6 Bond Parameter
k =1;
for b=linestart_section6:lineend_section6
    temp01 = X{b};
    temp02 = double(temp01);
    index_space = find(temp02==32);
    index_line_loop = length(index_space)-1;
    bond_type{k} = temp01(1:7);
    k = k + 1;
    j = 1;
    for z=5:index_line_loop;
        if index_space(z+1)-index_space(z) > 1
            D6(b-linestart_section6+1,j) = str2num(temp01(index_space(z):index_space(z+1)));
            j = j + 1;
        end;
    end;
end
end
```

```
% section 3 Off-Diagonal Terms

k = 1;
for c=linestart_section_offdiagonal:lineend_section_offdiagonal
    temp01 = X{c};
    temp02 = double(temp01);
    index_space = find(temp02==32);
    index_line_loop = length(index_space)-1;
    bond_type_off_diagonal{k} = temp01(1:7);
    k = k + 1;
    j = 1;
    for z=5:index_line_loop;
        if index_space(z+1)-index_space(z) > 1
            D3off(c-linestart_section_offdiagonal+1,j) = str2num(temp01(index_space(z):index_space(z+1)));
            j = j + 1;
        end;
    end;
end

% Section 18 Angle Parameter

k = 1;
for d=linestart_section18:lineend_section18
    temp01 = X{d};
    temp02 = double(temp01);
    index_space = find(temp02==32);
    index_line_loop = length(index_space)-1;
    angle_type{k} = temp01(1:10);
    k = k + 1;
    j = 1;
    for z=8:index_line_loop;
        if index_space(z+1)-index_space(z) > 1
            D18(d-linestart_section18+1,j) = str2num(temp01(index_space(z):index_space(z+1)));
            j = j + 1;
        end;
    end;
end;
```

```
end

% Section 26 Torsion Parameter

k = 1;
for e=linestart_section26:lineend_section26
    temp01 = X{e};
    temp02 = double(temp01);
    index_space = find(temp02==32);
    index_line_loop = length(index_space)-1;
    torsion_type{k} = temp01(1:13);
    k = k + 1;
    j = 1;
    for z=9:index_line_loop;
        if index_space(z+1)-index_space(z) > 1
            D26(e-linestart_section26+1,j) = str2num(temp01(index_space(z):index_space(z+1)));
            j = j + 1;
        end;
    end;
end

% Hydrogen Parameter

temp01 = X{linestart_sectionHbond};
temp02 = double(temp01);
index_space = find(temp02==32);
Hbond_type{1} = temp01(1:10);
j = 1;
for z=8:index_line_loop;
    if index_space(z+1)-index_space(z) > 1
        DHbond(1,j) = str2num(temp01(index_space(z):index_space(z+1)));
        j = j + 1;
    end;
end;
```

7.4 my_output_scanner.m

```
function [v_reax energy_reax] = my_output_scanner(POT)

[m, n] = size(POT)
step = 0;
for a=1:n
temp01 = POT{a};
temp02 = double(temp01);
index_space = find(temp02==32);
% to check wether there are 6 spaces
% I use a for loop and a spaces variable to
% count the number of spaces
spaces = 0;
if length(index_space) >= 6
    for i=1:6
        if index_space(i)==i
            spaces = spaces + 1;
        end;
    end;
end;
if spaces == 6
%
% look for the labels of the energy stored in temp03
%
% temp_energy = POT{a-1}
%
    index_line_loop = length(index_space)-1;
    j = 1;
    step = step + 1;

% This loop will give us the labels extracted
%
% if step == 1
%     temp_labels = POT{a-1};
%     temp_labels_ascii = double(temp_labels)
%     index_labels_ascii = find(temp_labels_ascii==32)
%     for i=1:index_line_loop;
```

```
%         if index_labels_ascii(i+1)-index_labels_ascii(i) > 1
%         label{j} = temp_labels(index_labels_ascii(i):index_labels_ascii(i+1))
%         j = j + 1;
%         end;
%     end;
% end;

% This instead the loop to build the matrix of the observables
for i=1:index_line_loop;
    for i=1:index_line_loop;
        if index_space(i+1)-index_space(i) > 1
            number(step, j) = str2num(temp01(index_space(i):index_space(i+1)));
            j = j + 1;
        end;
    end;
end;
end;
v_reax = number(step-1, :);
energy_reax = number(step, :);
end
```

7.5 my_output_extractor.m

```
function [POT] = my_output_extractor(filename)

disp(['read_output_file, reading: ' filename]);
fid = fopen(filename,'r');
if ~fid
    error('ERROR: file not found');
end
counter = 1;
% feof(fid)=1 quando e' alla fine del file
% feof(fid)=0 quando NON e' alla fine del file.
while ~feof(fid)
% cell structure, it is a structure where to put datas
    POT{counter} = fgetl(fid);
    counter = counter + 1;
end
fclose(fid);
end
```