


J. S. Redinha
J. da Providência
A. J. C. Varandas
Editors



Quantal Aspects
in Chemistry and Physics

*A tribute to the memory of
Professor Couceiro da Costa*

8. COMPUTATIONAL PROTEOMICS – FROM METHODOLOGICAL DEVELOPMENTS TO BIOLOGICAL APPLICATIONS

Irina S. Moreira, Natércia F. Bras[†], Alexandra T. P. Carvalho[†], Nuno M. F. S. A. Cerqueira[†], Daniel F. Dourado[†], Marta A. S. Perez[†], Antonio J. M. Ribeiro[†], Sergio F. Sousa[†], Pedro A. Fernandes, and Maria J. Ramos^{*}

REQUIMTE/Departamento de Química, Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre 687, 4169-007 Porto - Portugal

Proteomics, a chimera of proteins and genomics, involves the study of the proteins expressed in a cell, organism, or tissue. Proteins are essential in all aspects of life, and so the computational study of proteomics is becoming a vital element in understanding the underlying concepts. In this review we are going to address some of the challenges and latest developments focusing in four different aspects that are thematic in our group: (a) Molecular Dynamics Simulations; (b) Drug Design; (c) Enzymatic Mechanisms; and (d) Benchmarking of DFT functionals.

8.1 Introduction

Proteomics, a chimera of **proteins** and **genomics**, was invented by Professor Mark Wilkins in the early 1990s and involves the study of the proteins expressed in a cell, organism, or tissue. This includes protein identification and quantification, protein-protein interactions, protein complexes prediction, protein modifications and protein localization in the cell. As proteins are essential for all life, proteomics is crucial in biomedical applications, and although more recent, the computational study of proteomics is becoming a key element in this biological field.

Computational proteomics involves the computational methods, algorithms, databases and methodologies used to model protein structure, dynamics and

^{*}Email address: mjramos@fc.up.pt

[†]Equal participation.

function. In this review we are going to focus on four different aspects that are thematic in our group, and range from methodological developments to their biological application: (a) Molecular Dynamics Simulations; (b) Drug Design; (c) enzymatic Mechanisms; and (d) Benchmarking of DFT functionals.

8.2 Classical Molecular Dynamics Simulations

Classical molecular dynamics (MD) simulations have become during the past two decades, a particularly important discipline within the field of computational biochemistry, allowing the computationally efficient evaluation of a variety of properties in the study of biological molecules for which atomic and molecular motion is vital.

Generically speaking, MD simulations follow the time evolution of a system through the numerical integration of the equations of motion of the corresponding particles. In particular, MD simulations are based on the application of classical mechanics, with intra and inter-molecular interactions described by a sum of different contributions described by mathematical formulations of simple physics phenomena. The corresponding mathematical formulae and the accompanying parameters, which are typically fitted to reproduce experimental data or high-level *ab initio* calculations, are normally described under the generic designation of force field. A variety of different force fields is presently available, differing in features such as scope, accuracy and cost associated.

The range of application of MD simulations is remarkably wide and encompasses the study of phenomena such as protein and/or small molecule conformational changes, molecular association and recognition, folding, ion transport, etc. Over the last few years we have employed MD simulations in the study of several biological systems of interest, often in combination with other computational methods. In this section, we highlight 5 particularly different applications of this very powerful methodology, selected from our own work on biological systems. These include our MD studies on the elusive metalloenzyme farnesyltransferase (FTase), the medically critical HIV reverse transcriptase enzyme (RT), and the economically appealing carbohydrate-binding modules (CBMs)

from family 11.

8.2.1 Farnesyltransferase

Farnesyltransferase is a Zinc metalloenzyme that catalyses the addition of farnesyl groups from farnesyl diphosphate (FPP) to protein cysteine residues present in characteristic carboxyl terminal –CAAX motifs. In this motif C represents the cysteine residue that is farnesylated, A is an aliphatic amino acid, and X represents the terminal amino acid residue [1]. Proteins substrates bearing a CAAX motif include a number of biologically relevant protein targets, most notably the Ras family of proteins known to be implicated in something like 30% of all human cancers [2].

Performing MD simulations on FTase is particularly challenging, compared with the typical enzymes that are comprised simply by standard amino acid residues, because of the presence of a covalently bound Zinc atom with a metal coordination sphere that changes during catalysis. Metal atoms, and the corresponding bonds, angles, dihedrals, charges, and van der Waals parameters are normally absent in the typical biomolecular force fields such as AMBER, CHARMM or OPLS. Their inclusion involves not only the parameterization of the metal atom itself, but also of the directly interacting amino acid residues, and naturally a subsequent process of validation against experimental data. We have parameterized the three different Zn coordination spheres that are formed during the catalytic mechanism of this enzyme using quantum calculations and experimental data and have validated the new parameters against EXAFS and X-Ray crystallographic information [3].

Following this process, we have performed comprehensive MD simulations with the AMBER software package on the several intermediate states formed during the catalytic mechanism of this enzyme, in an attempt to understand the way this enzyme works at a molecular level, and taking into consideration features such as the effect of the solvent and the dynamic effects arising from the interaction of the enzyme, solvent and the substrate/product molecules [4,5]. Starting from extensive 10 ns MD simulations on the enzyme resting state, binary

complex (FTase-FPP), ternary complex (FTase-FPP-CAAX substrate) and product complex (FTase-Product), we have performed comparative analysis of the amino acid flexibility along the FTase sequence, radial distribution functions of water molecules around catalytically relevant atoms, statistic variations on key catalytic distances, detailed analysis on the conformation and orientation adopted by the substrate and product molecules in the presence of the enzyme, and hydrogen bonding analysis on the most important molecular recognition sites.

These results provided very useful information for the subsequent modeling of the catalytic mechanism of this enzyme, guiding and supporting the choice of the models used from QM or QM/MM calculations, and of the several approximations adopted.

8.2.2 Reverse Transcriptase

Reverse transcriptase (RT) is the human immunodeficiency virus (HIV) enzyme whose function is to copy the viral RNA into double-stranded DNA suitable to be integrated in the host cell genome. Several combinations of different RT inhibitors are currently used in antiretroviral therapy. Our study focused on nucleoside reverse transcriptase inhibitors (NRTIs). These are substrate analogues that compete for binding and incorporation into the nascent DNA chain. However, because they lack a 3'OH, after they are incorporated they do not allow the addition of the next incoming nucleoside blocking DNA synthesis. It is presently known that the long-term failure in the treatment of AIDS with the currently available NRTIs is related to the development of resistance by RT at the binding or incorporation level, or subsequent to the nucleotide incorporation (excision).

We have conducted a series of MD simulations of RT with different inhibitors in explicit solvent in order to correlate the structural characteristics of the inhibitors with the stage at which RT resistance emerges. To achieve a greater insight on how RT discrimination gets established we compared incorporation of a normal substrate (dNTP) with incorporation of two very similar inhibitors for which resistance emerges by different mechanisms: phosphorylated zalcitabine,

ddCTP, which is discriminated and phosphorylated stavudine, d4TTP, which is mainly excised [6]. We found that the different resistance profiles arise from the different conformations adopted by the inhibitors at the active site. d4TTP adopts an ideal conformation for catalysis because it forms an ion-dipole intramolecular interaction with the α -phosphate oxygen of the triphosphate, as does the normal substrate. In ddCTP, the lack of this essential interaction results in a different, noncatalytic conformation [6]. To achieve a greater insight on why RT excision occurs we conducted molecular dynamics simulations of complexes of HIV-1 RT with the incorporated substrate and the antiretrovirals AZT and d4T with and without pyrophosphate. For these two inhibitors resistance emerges via the excision mechanism, however they are very different structurally: AZT was a bulky azide group at the 3' position that could impose steric hindrances to translocation and d4T only was an hydrogen at this position [7]. We found that RT preferably excises these inhibitors over the substrate as a consequence of a different pattern of hydrogen bridges they establish with the N site after incorporation. In the complexes with normal nucleotides, the fingers residues K65 and R72 establish hydrogen bonds mainly with the leaving PPI. With the inhibitors, those same residues establish hydrogen bonds primarily with the substituted nucleotides. Consequently, pyrophosphate is eliminated before the opening of the fingers domain for the inhibitors, which allows ATP binding, with subsequent excision and development of drug resistance [7]. Our main conclusion was that although the lack of the 3'OH is the determinant that makes NRTIs inhibitors, it seems that the enzyme is highly specialized in recognizing structures with this group. This seems to be the cause for resistance to NRTIs being so common.

8.3 Carbohydrate Binding-Modules (CBMs)

The conversion of plant cell wall polysaccharides into soluble sugars is one of the most important reactions in nature. This process is of high economical interest as the products obtained (such as glucose derivatives) are very useful in food and pharmaceutical industries. They also have an enormous poten-

tial for the bio-fuel industry, as ethanol can be directly obtained from glucose monomers [7,8]. An efficient degradation of cellulose chains into soluble glucose monomers can be achieved using chemical means or by certain microorganisms such as *Clostridium thermocellum* (*Ct*). The latter method has become the most attractive due to reasons of economy and efficiency [8]. These organisms possess a cluster of enzymes organized in a high-hierarchy multi-subunit complex called cellulosome [9] The enzymes are generally modular proteins that contain non-catalytic carbohydrate-binding modules (CBMs), which increase the activity of the catalytic module and are thus crucial for the efficient degradation of polysaccharides [10].

Since the X-ray structure of *Ct*CBM11 with a bound substrate is not available, we have used MD simulations with both CHARMM and AMBER force fields [11,12], integrated with the recently developed MADAMM docking protocol [13], to determine the molecular recognition of glucose polymers by CBMs from family 11. MD simulations demonstrate that the side chain conformations of some tyrosine residues near the binding pocket (Tyr22, Tyr53, Tyr129 and Tyr152) give rise to a steric obstacle, precluding the efficient binding of the ligands. To overcome this limitation, a novel docking protocol that introduces a certain degree of flexibility to these amino acids in standard docking processes was used. Our results have shown that the binding interface of the *Ct*CBM11 can bind only one single polysaccharide chain, and we have used cellobiose, cellotetraose, cellohexaose, celloctaose and cellotrideose as model substrates. We propose a general mechanism for the interaction between *Ct*CBM11 and cellulose chains, in which four main charged amino acids (Asp99, Arg126, Asp128, and Asp146) have a key role in the interaction with the cellulose chains. Another feature is that a minimum of four glucose monomers in the polymer chains are required for a strong interaction with the central binding site, with the remaining units attached equidistantly to both sides of the *Ct*CBM11 cavity. MD simulations also indicate that the strongest hydrogen interactions occur with the hydroxyl groups attached to C-2 and C-6 of central glucose units of the polymer chain, which is in agreement with STD and line broadening NMR studies [11,12,14].

Furthermore, our data have shown that the three aromatic tyrosine residues (Tyr22, Tyr53, and Tyr129) at the *Ct*CBM11 interface induce a certain distortion in the glucose rings that was found to be important for guiding, reorientation and packing of the polysaccharide chain to the charged region, providing important insights into substrate binding by this important class of proteins. We also suggest that these aromatic residues are crucial to detach the carbohydrate chain from the solvent or other polysaccharide chains, by inducing a reorientation of the hydrogen bonds [11,12].

Considering that the *Ct*CBM11 is topologically similar and structurally homologous to CBMs of families 4, 6, 15, 17, 22, 27 and 29, we suggest that similar molecular determinants drive the binding and recognition of polysaccharides to these CBMs. The knowledge of the interactions that occur at the molecular level between several polysaccharides and the CBMs can be used to improve the efficiency of the linked enzymes and/or possibly of the cellulosome itself.

8.4 Drug Design

The natural tendency of proteins to bind to each other as well as to several small-molecules (ligands), forming stable and specific complexes is essential for all biological processes. The description of the structural and functional properties behind protein–protein or protein–ligand interactions and protein-binding is very important not only to increase the scientific knowledge in basic terms, but also for applied research in biomedical science and industrial pharmaceuticals.

One of the most important fields is Medicinal chemistry, at the interface of chemistry and biology, which has created an important tool in the search for new drug candidates with a combination of good pharmacodynamic and pharmacokinetic properties. Although this study can be carried out by having only an X-ray crystallographic structure of the target, additional structural and functional insights are important for the rational design of more bioactive molecules. The drug design process includes the structural determination of target proteins, hit selection, lead optimization, development of structure-activity relationships and the design of new compounds. The process of drug development is chal-

lenging, time-consuming, labor intensive, and expensive. but has as a final goal to find, develop, and market new chemical entities (NCEs), which can be used against untreatable diseases, or which have superior properties when compared to currently available drugs.

Structure-based drug design was usually a field involving the binding of a small molecule to a biomolecular target that functioned by inhibiting its function. A typical drug-like molecule had to obey the Lipinski's rule of five: no more than five hydrogen-bond acceptors and 10 hydrogen-bond donors, a molecular weight under 500 Dalton, and a partition coefficient $\log P$, a measure of lipophilicity, under 5 [15]. Veber *et al.* added that the candidates should have 10 or fewer rotatable bonds, and polar surface area equal or less than 140 \AA^2 [16]. This ensure that it has the physicochemical and pharmacokinetic properties such as good solubility, a correct balance between lipophilicity and hydrophilicity, metabolic stability (a good absorption, distribution, metabolism, excretion ADME), and bioavailability necessary to inhibit specific interactions. It has to take into account also the toxicity, radical attack (biodegradation), good quantitative structure–property relationships (QSPR) and good quantitative structure–activity relationship (QSAR). Nowadays this concept has enlarged and includes the inhibition of protein–protein interactions (PPI). However, the discovery of molecules capable of selectively inhibiting PPI encounters many obstacles such as the large interfacial areas and the relatively flat topographies of the surface of protein–protein interfaces. Small-peptides capable of disrupting this kind of interactions, usually cyclic peptides and other modified peptides do not necessarily obey to the Lipinski's rule of five. Thus, as peptides present higher number of degrees of freedom than small molecules, we face a crucial challenge of the level of flexibility of the systems under study. So, the rational design of the inhibitors has to take into account the conformational plasticity of the protein and the interplay between different conformations. Modeling a protein–peptide complex allows the determination of the pharmacophore model (geometrical arrangements of chemical features such as hydrogen bonding and electrostatic and hydrophobic interactions) that can be used to design small

molecules capable of mimicking the peptide [17].

In this part of the review we are going to focus on some *in silico* methods used to facilitate the modeling of protein–protein interfaces and protein-binding. Among these methods, we stress the determination of the three-dimensional structures of complexes (protein-protein and protein–ligand) as well as the structural determination of the crucial amino acids residues involved in binding by sequential mutagenesis of the entire protein interface (Alanine Scanning Mutagenesis, ASM), and computational approaches used to design new drugs and/or optimize the lead.

8.4.1 Protein-Protein Studies

Protein-Protein Docking

Protein-protein binding is one of the critical events in biology. It is extremely valuable in obtaining structural information and a complete understanding of both the biochemical nature of the process for which the components come together, and to facilitating the design of compounds that might influence it. However, due to the greater difficulty in crystallizing protein-protein complexes, there is relatively little structural information available about them compared to the proteins that exist as single chains or form permanent oligomers. Hence, experimental studies are faced with remarkable technical difficulties and the number of solved complexes deposited in the Protein Data Bank (PDB; www.rcsb.org/pdb) is still orders of magnitude smaller than those of experimental information on protein interactions and of structures of individual proteins. Nevertheless the practical difficulties for a better understanding of the biological function of a protein, knowledge of its three-dimensional structure is fundamental. Thus, in the past two decades there was an emergence of a large variety of theoretical algorithms designed to predict the structures of protein–protein and protein-ligand complexes – a procedure named docking.

Computational methods, if accurate and reliable, could therefore play an important role, both to infer functional properties and to guide new experiments. The first protein–protein docking algorithm was developed by Janin and Wodak

in 1978 [18-20]. Albeit important successes, docking screens remain hampered by the prediction of false positives and negatives [21]. Because of the complexity of the problem, protein–protein docking is still largely at the theoretical stage and there is still considerable scope for the development of methodology [22,23].

The goal of predictive protein–protein docking [24] is to predict the 3D arrangement of a protein–protein complex from the coordinates of its component molecules, being an accurate prediction the one that will point out most of the residue-residue contacts involved in the target interaction. Usually, this involves an exhaustively searching of the rotational and translational space of one protein with respect to the other, resulting in a six dimensional search. Hence, there are three key ingredients in the docking: representation of the system, conformational space search, and ranking of potential solutions [21]. Although these can vary, the protein–protein docking contains certain problems common to all procedures: “searching and scoring” [25]. Therefore, “searching” is how to accurately describe the energy function of a given protein–protein complex and “scoring” is how to obtain the global minimum energy structure of the complex using the energy function [21]

Protein–protein docking studies originated a very complete review [21] and are a subject of study in your lab. We are currently trying to find new ways to rank the solutions proposed by one of the best softwares in literature (HADDOCK) and achieve a good docking structure [25-30].

Alanine Scanning Mutagenesis (ASM)

Since its initial application to human growth hormone and the growth hormone binding protein, alanine scanning mutagenesis continues to be a valuable procedure for both hot spot detection and analysis of a wide range of protein–protein interfaces. Although slow and labour-intensive, alanine-scanning mutagenesis is the most trendy method for mapping functional epitopes, as alanine substitutions remove side-chain atoms past the β -carbon without introducing additional conformational freedom. With the application of this methodological approach, it has been found that there is a highly uneven distribution of en-

energetic contributions of individual residues across each interface, and that only a few key residues do contribute significantly to the binding free energy of protein–protein complexes: the hot spots [31].

Hot spots have been defined as those sites where alanine mutations cause a significant increase in the binding free energy of at least $2.0 \text{ kcal mol}^{-1}$. To have a strong impact in protein building the binding free energy should be higher than 4 kcal mol^{-1} (3 orders of magnitude in the binding affinity constant). However, residues whose mutation results in such large differences are quite unusual, and the threshold for the hot spots had to be lowered to 2 kcal mol^{-1} in order to get enough data for statistical analysis. Systematic analysis of hot spots has shown a non-random composition: tryptophan (21.0%), arginine (13.3%) and tyrosine (12.3%) [31].

Even though it is very important to develop an accurate, predictive computational methodology for alanine scanning mutagenesis, capable of reproducing and interpreting the experimental mutagenesis values, until recently the success rates had been modest. Two of the major problems were the fact that alanine mutation of charged amino acids usually generates values in disagreement with the experimental ones, and the fact that the computational time involved is much too high to permit a systematic mutagenesis of protein-protein interfaces. Thus, having as a basis the Molecular Mechanics Poisson-Boltzman Surface Area (MM-PBSA) approach, we have focussed our attention in ways to decrease the computational time involved, as well as in techniques that enable the achievement of the chemical accuracy of roughly 1 kcal mol^{-1} [32-39].

So, we developed a fully atomistic computational methodological approach schematized in Figure 8.1 that consists in a computational Molecular Dynamics simulation protocol performed in a continuum medium using the Generalized Born Solvation Model of the wild-type system. The post-processing treatment of the wild-type allows the calculation of the free binding energy of the mutant complex and all the monomers involved. There are 20 alpha amino acids commonly found in proteins and they can be divided into basically four groups according to the structure of the side chain: non-polar and neutral (valine,

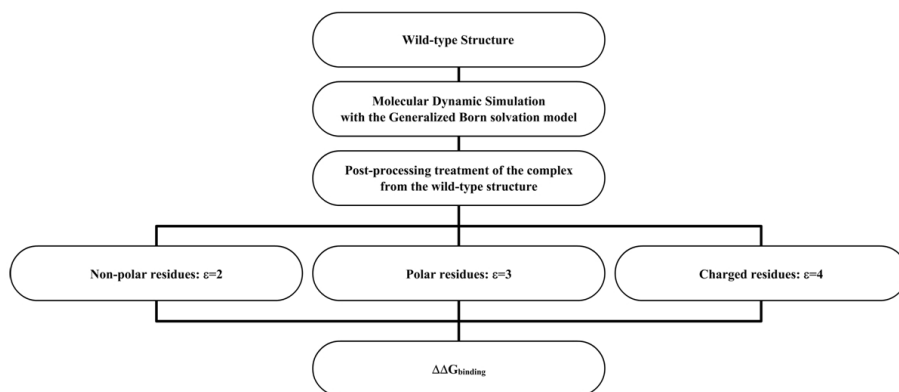


Figure 8.1. Resume of the methodological approach for computational alanine screening mutagenesis.

alanine, leucine, isoleucine, phenylalanine, proline, glycine, methionine and tryptophan), polar and neutral (asparagine, glutamine, cysteine, tyrosine, serine and threonine), acidic and charged (aspartic acid and glutamic acid), basic and charged (lysine, arginine, and histidine). As histidine can be uncharged or charged at physiological pH we have grouped this residue with lysine and arginine at the basic and charged amino acids. Recalling that we used only one trajectory for the computational energy analyses, it is important to highlight that side chain reorientation is not included explicitly in the formalism. As amino acid polarity increases, the structural effect beyond the neighbour residues also increases, and the conformational reorganization after alanine mutagenesis should be more extensive. This reorganization is not explicitly taken into account in the single trajectory protocols but its effect can be implicitly included by raising the internal dielectric constant. It is not possible to know the correct internal dielectric constant value that should be used because it depends on the mutated amino acid and the interacting residues. Nevertheless, we have noticed that by using only an internal dielectric constant set of three different values, exclusively characteristic of the mutated amino acid (2 for the non-polar amino acids, 3 for the polar residues and 4 for the charged amino acids), it was possible to obtain an excellent agreement with the experimental results for the $\Delta\Delta G_{\text{binding}}$ values. If we consider a deviation of $\pm 1.4 \text{ kcal mol}^{-1}$ from the experimental value as an

accurate result, we have an overall success rate of 82%, a 82% success rate for hot-spots [32-39].

8.4.2 Protein-Ligand Studies

Protein-Ligand Docking- MADAMM as an example

Proteins are an essential part of the organisms and participate in virtually every process within cells. A vast number of these processes require that small molecules bind to specific spots of these macromolecules. These molecules can act as switches to turn on or off a protein function, or can be the substrates of a particular chemical reaction that is catalyzed by a specific protein. Understanding the mechanism by which proteins associate and interact with small molecules, has thus become a subject of paramount importance in drug discovery. Conventional experimental techniques for obtaining detailed structural information about protein-ligand complexes are time and resource intensive. To overcome these limitations, several computational methodologies and algorithms have emerged in the last 10 years endeavoring to foresee and improve the understanding of this difficult-to-obtain structural information. These techniques are normally defined as protein-ligand docking and predict as well as rank the structure(s) arising from the association between a given ligand and a target protein of known 3D structure.

Generally speaking, as we have previously stated, all molecular docking methodologies are composed by two different algorithms: the search algorithm and the scoring function. Despite the apparent simplicity of these methodologies, they have several hidden weaknesses and present a number of problems from the computational point of view. One of the major problems arises from the tremendous complexity of the system, from which result hundreds of thousands of degrees of freedom that need to be analyzed, requiring huge computational resources. Furthermore, the combination of the energetic forces acting on the binding process is not-completely-known or/and it is difficult to calculate. Therefore, different simplifications are imposed to these methodologies to turn them fast, accurate, and attractive in different situations.

The degree and number of approximations upon which these theories are based on have thus become the major cornerstone issue in this field. Most of the simplifying assumptions are helpful in order to reduce the computation time, but they can also lead to unusable results [40]. Therefore, the correct balance between these two aspects has become the key of success in this field.

With respect to the dynamics aspects of molecular recognition, most of the molecular docking methodologies lie along a spectrum of models bound by the lock-and-key and induced-fit theories for ligand binding. In such models, the receptor is treated as a rigid body and only in the second one the translation, rotation, and torsion degrees of freedom are calculated. During the last decade, several protein-ligand docking algorithms based on these simplifications have been successfully applied in several problems [41-45]. Despite the breathtaking advances in the field and the widespread application of these methods, several downsides still exist. Particularly, protein flexibility is a major hurdle in current protein-ligand docking efforts that needs to be more efficiently accounted for.

Many programs have been developed in the last 5 years that account for protein flexibility in protein-ligand docking methodologies. All of them have their own merits and shortcomings, and reveal that accounting for protein flexibility in protein-ligand docking algorithms is still challenging. One of the latest is MADAMM, a multi staged Docking with an Automated Molecular Modeling protocol [13]. This program is a new molecular docking protocol that allows introducing a certain degree of flexibility (as much as we want/need) to the receptor and full flexibility to the ligand, without requiring an excessive time of computation in the full process. This docking software has shown excellent results in several studies, in which standard and popular docking software failed to achieve the correct result. To demonstrate the potential and capabilities of the MADAMM protocol, in Figure 8.2 we present our attempts to dock the progesterone to the active site of monoclonal antiprogestosterone antibody DB3. Looking at the active site region of the unbound and bound structures of the receptor (A—dark blue and B—light blue) we can see that Trp100 adopts two distinct conformations. In the unbound form, it is facing to the center of the active site,

whereas in the bound form it is displaced to the right hand side. This small, but crucial conformation rearrangement is sufficient to generate a faulty result when we try to dock the ligand to the active site region of the unbound structure, using standard docking software (A–black). This result is not dependent on the limitations of the search algorithm or scoring function of the docking algorithms. It simply results from the conformation adopted by Trp100, which in the active site of the unbound structure occupies the space that is required for the correct binding of the ligand (as it can be observed analyzing the conformation of Thr100 in the bound structure).

When the MADAMM protocol was applied to this case study, we were able to flexibilize several residues in the region of the active site, including Thr100. After the docking procedure, several complexes were obtained with the ligand bound in different conformations to the active site of the receptor. Each complex was subsequently subjected to a set of minimizations and small dynamics jobs that were recursively ranked and clustered in distinct groups taking into account the protein-ligand affinity. After a set of cycles, the top scored solutions were selected. The best solution is displayed in Figure 8.2 (D–Yellow). This structure almost resembles what is found in the unbound X-ray structure, showing only a small RMSD of 0.13 Å.

This case study has shown that MADAMM can efficiently accomplish a good compromise between what can be predicted and what is obtained experimentally. This means that this protocol can be viewed as a powerful tool to understand protein-ligand interactions, especially on those cases where few or no experimental structures of the complexes are available. The same study also alerts for the limitations of the standard docking software especially on those cases where the orientation of particular residues at the protein-ligand interface is neglected. The results have shown that the conformations adopted by some residues in the region of the active site can have a crucial influence on the way the ligand interacts with the receptor. Disregarding its presence can be responsible for most of the false positives results that are obtained with the available rigid-based docking software.

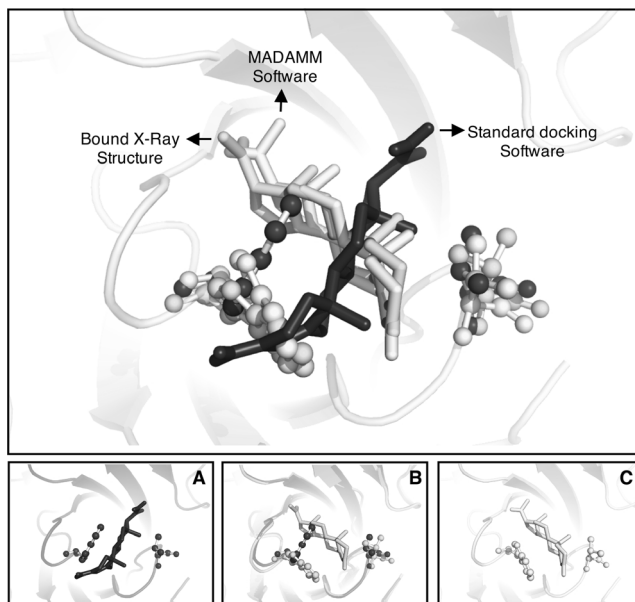


Figure 8.2. Dark blue: X-ray structure of the unbound form of the protein. Light blue: co-crystallized X-ray structure with the bound substrate. Yellow: MADAMM result. Black: standard docking result (using the GOLD software).

Design of Inhibitors and Energy Assessment

Efficient and accurate calculation of protein-ligand affinities is another focus of our group. During the past few years we have studied the ability of the enzymes, namely enzymes involved in lethal diseases, to bind preferentially to a set of ligands (substrates, inhibitors, new inhibitors, proteins). To quantify the preferential binding of a set of ligands, we have calculated values of free energy of binding, using a complex but accurate computational method, called Thermodynamic Integration (TI) [46-49], and/or the more computationally accessible MMPBSA [50] method. TI is a rigorous method for calculating free energies, but requires extremely time consuming simulations, even with a large high-performance computer cluster. The faster and simpler, but still accurate, MMPBSA method is many times used as an alternative to TI.

We have been particularly interested in protein-ligand studies with HIV-1 Protease. All treatments for HIV-1 infected persons include, at least, one Protease inhibitor. Based on Protease-ligand studies we have published a new theory for

HIV-1 Protease recognition [51] and developed new inhibitors for HIV-1 Protease using other inhibitors (Nelfinavir [52], Amprenavir) as leads. The understanding of the mechanism of Substrate recognition by HIV-1 Protease is a key step for drug design targeting the enzyme. In this section we present three new inhibitors, designed with computational tools, with greater affinity for HIV-1 Protease than Nelfinavir itself.

Nelfinavir (Viracept®) is a potent, orally bioavailable inhibitor of the enzyme HIV-1 Protease, which has been developed through structure-based drug design projects and has been approved worldwide for the treatment of HIV infected patients. However, HIV-1 develops drug-resistance and the affinity of Nelfinavir for the binding pocket of Protease is decreased. We have presented three new variants (Figure 8.3) of Nelfinavir, designed with computational tools, with greater affinity for HIV-1 Protease than Nelfinavir itself. In order to increase the inhibitory efficiency, we have introduced rational modifications in Nelfinavir, optimizing its affinity to the most conserved amino acids in Protease. The new inhibitors interact more favourably with well-conserved residues, Leu23, Ala28, Gly49, Arg87, and Asp29, which cannot mutate, as the mutants would render the Protease catalytically inactive [53]. Figure 8.3 shows a schematic representation of the binding region for Nelfinavir with which the substitutions introduced in the inhibitors are meant to interact. The dashed lines give an idea of the location of the empty pockets between the well-conserved amino acids and the inhibitors. Figure 8.3 shows also three examples for which significant increases in affinity can still be achieved without changing the overall structure, molecular mass and hydrophobicity of the inhibitors, thus preserving their very favourable ADME properties. Minimization and molecular dynamics simulations [46] have been carried out on the complexes, HIV-1 Protease with Nelfinavir and subsequently with the new inhibitors, in order to analyze the behavior of the systems. To quantify the affinity of the new inhibitors relatively to Nelfinavir, we have calculated values of free energy of binding, first using the TI approach and subsequently the less computationally demanding MMPBSA [50] methodology. The values for the binding free energy difference presented (Figure 8.3) are val-

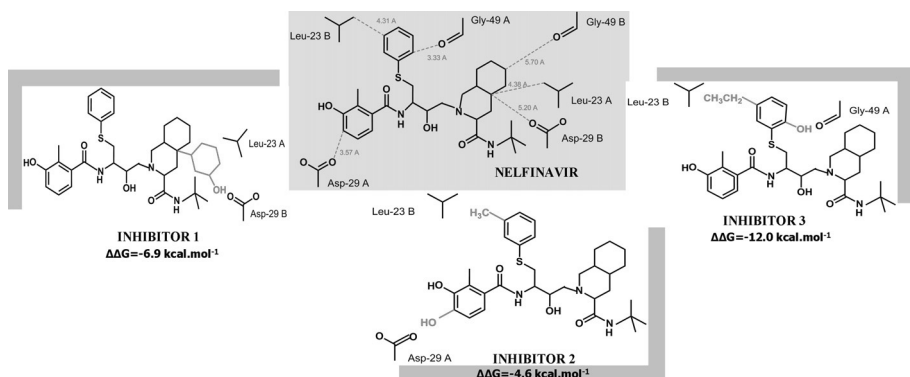


Figure 8.3. Three new inhibitors; based on Nelfinavir as lead, which have a higher affinity *in silico* for Protease than Nelfinavir (the negatives binding free energy difference $\Delta\Delta G$ reflect this).

idated by the correct reproduction of the experimental binding free energy for Nelfinavir.

This example shows how powerful the design of inhibitors can be for rational lead optimisation, avoiding handling dangerous and toxic materials and greatly reducing experimental costs. There have been dramatic successes with drug design.

8.5 Enzymatic Mechanisms

In addition to the several applications outlined in the preceding sections, a particularly strong and promising area in the field of computational proteomics is that of computational enzymology, *i.e.* the use of computational methods to study enzymatic activity, in particular catalytic pathways. In fact, computational methods allow the detailed analysis of important enzymatic reaction mechanisms, providing atomistic insight into very specific processes of highly biological significance, often very difficult to tackle by means of experimental studies alone. While a vast number of experimental techniques are normally employed in the study of reaction mechanism, from spectroscopic experiments and mutagenesis studies to kinetic evaluations, no experimental technique, by itself, can normally give a full view of an entire catalytic pathway. In addition, the interplay between different experimental methodologies is often not peaceful, with differ-

ent techniques frequently pointing to different (at least apparently) conclusions. Computational methods often offer an unbiased way to concert and validate the information obtained from different experimental methods, providing also the missing pieces. So, in a process in which something always seems to be missing computational methods often have the final word, which frequently comes in the form of a structure for the transition state intermediate and of the corresponding energy barrier, indispensable requirements to fully assess the viability of a mechanistic proposal.

During the last decade, we have studied a large number of enzymatic catalytic mechanisms. From these, we would like to highlight 5 particularly important biological systems, in which our effort helped to shape and validate the presently accepted pathway. These are the radical enzyme ribonucleotide reductase, the very important glutathione transferase, thioredoxine, glycosidase, and the Zn metalloenzyme farnesyltransferase.

8.5.1 Ribonuclease Reductase (RNR)

Nowadays, with the exception of viroids and virusoids, all modern organisms have their genetic information encoded into a DNA molecule. The exploit and maintenance of this information depends on the availability of deoxyribonucleotides. These compounds cannot be obtained by external sources and the only way by which they can be synthesized is through the reduction of ribonucleotides into deoxyribonucleotides. This reaction is strictly conserved in all living organisms and is catalyzed by a peculiar enzyme called ribonucleotide reductase (RNR). This key role makes RNR a rate limiting step in DNA replication and repair, turning it into an attractive target for antitumor, antiviral and antibacterial therapies [54,55]

RNRs are mechanistically fascinating proteins because of their free radical chemistry, unusual metallocofactors and complex regulatory mechanisms. In the past 20 years, several studies have been addressed at this enzyme, aiming to understand its peculiar mode of action. However, and despite its success as a target in many cancer therapies and HIV/bacterial treatments, the underlying

mechanisms by which this enzyme is either inhibited or reduces ribonucleotides into deoxyribonucleotides is poorly understood taking into account the information available from the experimental data. Great advances in this field were achieved when RNRs related mechanisms were studied by theoretical and computational methodologies. These methods and particularly the quantum chemical calculations have become an indispensable tool in this field, allowing to solve a puzzle of many unrelated and disconnect pieces that were found out by structural biology studies and biochemical experiments. The main advantage of the computational methods is that they can provide structural information on transition states and snapshots of molecules in the act of reaction, whose direct detection is not possible or is difficult to obtain by normal physical methods. This is particularly important in RNR since it is a radical enzyme, and these methods enable to identify and characterize unstable intermediates at an atomic detail.

During the past 10 years, our group has developed a comprehensive knowledge and thorough understanding of every process involved in the normal RNR functioning and in its inhibition mechanisms, using theoretical and computational means. In this process, we began by exploring different strategies to model the active site of this enzyme. A good balance between the size of the system and accuracy of the results is always difficult to achieve in computational chemistry. This is especially true when we aim to study biological systems where the results can be largely influenced by size of the model that is used. Therefore, we started to study RNR testing *several models of the active site that range from 20 atoms* [56], *300 atoms* [57] *to the full R1 monomer (that contains 30000 atoms)* [58]. *To complete this task, hybrid methodologies, and particularly ONIOM approach*, were used with the larger models. These methodologies allow to divide the model in different layers that can be treated with different theoretical levels. The atoms directly involved in the reaction are calculated with higher theoretical levels and the remaining ones with a lower theoretical level. This approach has proven to be most successful, allowing us to tune the best approach to study this enzyme, from a computational point of view. This information

was then used to explore all relevant chemical pathways that could be involved in the catalytic mechanism [56,59] of RNR as well as in the inhibitory mechanism of several substrate analogues inhibitors that inactivate the function of the enzyme [60-67]. Some of the results obtained through computational methodologies were pioneers in the RNR field and later confirmed and acknowledged by experimental findings.

Theoretical and computational methodologies were therefore a major groundbreaker in this field allowing to unravel the unsolved mysteries around RNR. Without their value contribution it would be very difficult to understand, explain and predict most of the knowledge that is currently available. All the work developed in this area can now be used to understand, which are the most important checkpoints that must be triggered during the inhibitory mechanism in order to enhance and improve the potency of a RNR inhibitor, or develop new ones. These results have thus created a trend in which researchers can now base their studies to conduct a more rational drug design approach in the development of novel drugs against RNR.

8.5.2 Glutathione Transferase

Glutathione transferases (GSTs) have been known as fundamental enzymes of the cell detoxification system for almost fifty years now. The cell detoxification mechanism of xenobiotic and endobiotic compounds follows a series of different steps. In the first step toxic compounds are converted into strong electrophiles by the mixed-function oxidation activity of cytochrome P-450. Those electrophiles are subsequently transformed into more soluble and less toxic substrates, by conjugation with glutathione (GSH) due to the catalytic activity of GST. Finally, these resulting conjugates can be recognized by ATP-dependent transmembrane pumps, such as P-glicoproteins and MRP family proteins, and consequently expelled from the cell. On the other hand, GSTs have also an active role in byosynthesis, cell signaling pathways [68], and are related to human diseases such as Parkinson's [69,70], Alzheimer's [70-74], atherosclerosis [75-77], liver cirrhosis [78,79], aging [80] and cataract formation [81]. The reaction cata-

lyzed by GST consists in the nucleophilic addition of the sulfur thiolate of GSH to a wide range of electrophilic compounds. When GSH binds to the GST G-site active center the pKa of the thiol group drops from 9.1 to about 6.2-6.6 pH units [82], promoting its deprotonation. After GSH activation, the nucleophilic sulphur atom attacks the electrophilic toxic compound present in the H-site active center, producing a less dangerous compound.

Despite cytosolic GST's being vastly studied by the scientific community, the main aspects of the catalytic events are still to be understood. Recently, using as a model the GSTA1-1 enzyme, we proposed a GSH activation mechanism consistent with the experimental data [83,84]. Our studies have demonstrated that a water molecule is able to assist a proton transfer between GSH thiol and alpha carboxylic groups with an activation energy of $13.39 \text{ kcal mol}^{-1}$, after a first conformational rearrangement of GSH ($\Delta G_{\text{conf}} = -1.62 \text{ kcal mol}^{-1}$) that allows the water molecule to interact simultaneously with both the thiol and the glutamyl alpha carboxylate groups. This energy barrier is in agreement with the experimental kinetics for the GST catalyzed GSH-CNDB conjugation, a common electrophilic substrate ($k_{\text{cat}} = 88 \pm 3 \text{ s}^{-1}$, $\Delta G^{\ddagger} = 15.06 \text{ kcal mol}^{-1}$ [85]). Figure 8.4 resumes all the events. We also demonstrated that a catalysed direct proton transfer between the two GSH active groups is very unlikely (energy barrier = $15.88 \text{ kcal mol}^{-1}$) for the GSH conformational rearrangement, plus $19.44 \text{ kcal mol}^{-1}$ for the actual proton transfer). In order to study the free energy associated with the initial GSH conformational rearrangement we calculated its potential of mean force (PMF) using the umbrella sampling method. All the molecular dynamics simulations and subsequent analyses were carried out using the Gromacs software package conjugated with the AMBER99 force field [86-89]. To study the actual proton transfer an ONIOM model of the GSH G-site active center was built. Then we performed a scan of the water proton approach to the most suitable GSH glutamate alpha carboxylate oxygen. With the three stationary points we were able to calculate the proton transfer activation energy, ΔG .

Arg15 is a strictly conserved active site residue in class Alpha GSTs [90],

however very little is known about its role in catalysis. In order to clarify the importance of this conserved residue we analyzed the activation energy barrier and structural details associated with the GSTA1-1 mutants R15A, R15R ϵ , η -c (an Arg residue with the ϵ, η -nitrogens substituted by carbons) and R15Rneutral (a neutral Arg residue due to the addition of a hydride in the ζ -carbon) [91]. A similar mechanism to the one used in our GSH activation proposal was used. The energy barriers associated are in agreement with the experimental values available [90] and can be analyzed in Figure 8.4. The structural analyses of the enzymes allow concluding that in the wild type enzyme GSH binds to the G-site pocket in a specific arrangement not seen in the mutants. The charged Arg15 establishes a strong ion-dipole interaction and a hydrogen bond with the GSH cysteine mainchain, which dictates the arrangement of the substrate. For the R15Rneutral mutant, hydrogen bond interactions are still possible to be established between the residue 15 sidechain ϵ, η nitrogen atoms and the GSH cysteine mainchain carbonyl group. However, without the positive charge, the spatial arrangement of residue 15 changes leading to a new, not catalytically efficient, GSH conformation. In the other mutants this new GSH conformation is also observed and is supported by the experimental data available for the mutant R15A (K_M GSH is 10-fold increased relatively to wildtype enzyme [90]). The volume of Arg15 does not seem to be as catalytically relevant as the charge. The R15Rneutral mutant residue 15 has the same volume as the wildtype Arg15, however this mutant shows an energy barrier similar with the smaller R15A mutant.

8.5.3 Thioredoxins (Trx)

The enzymes of the thioredoxin (Trx) family fulfill a wide range of physiological functions. Although they are structurally similar and have a similar CXYC active site motif, with identical environment and stereochemical properties, where C stands for cysteine and XY for two variable residues, the redox potential and pKa of the cysteine pair varies widely across the family. As a consequence, each family member promotes oxidation or reduction reactions, or

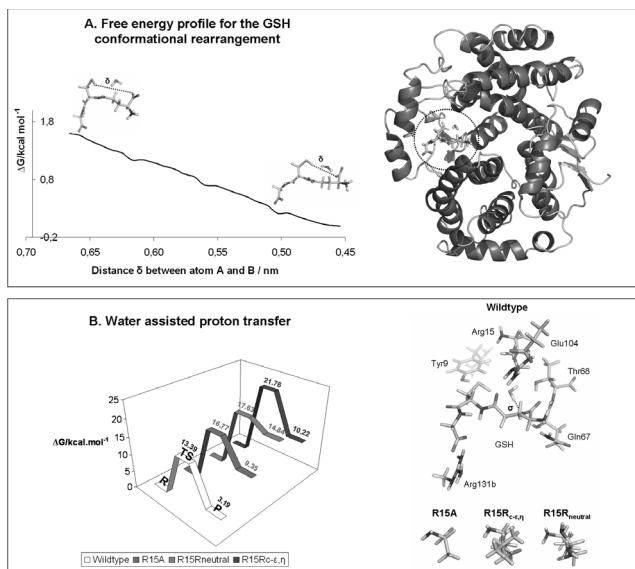


Figure 8.4. GSTA1-1 water assisted GSH activation mechanism.

A – Distance δ was steadily decreased in each PMF window. The curve represents the sum of the entire data obtained from the PMF forward and backward processes. On the right hand side a detailed view of the GSTA1-1 complex, in which GSH has been circled, is represented.

B – Wild type and mutant enzymes water assisted proton transfer Gibbs energies of the three stationary points: Reagent (R), Transition State (TS) and Product (P). Energies calculated with DFT, functional B3LYP and basis set 6-311++G(2d,2p). On the right hand side the G-site model is represented along with distance σ , decreased at each scan point.

even isomerization reactions.

We carried out a set of quantum mechanical calculations in active site models to gain more understanding on the molecular-level origin of the differentiation of the properties across the family. We theoretically explored the reaction mechanisms, both in the gas phase and in water, using density functional theory [92,93]. The mechanism of disulfide reduction involves two consecutive thiol-disulfide exchange reactions, that is, nucleophilic substitutions at sulfur (SN2@S): first, by the nucleophilic cysteine-thiolate group (Cys_{nuc}) at a sulfur atom of the disulfide substrate and, second, by the other cysteine-thiolate group (called buried cysteine, Cys_{bur}) at the sulfur atom of the Cys_{nuc}.

The obtained results, together with earlier QM/MM ONIOM calculations in which absolute and relative pKas of the nucleophilic cysteines for Trx and DsbA

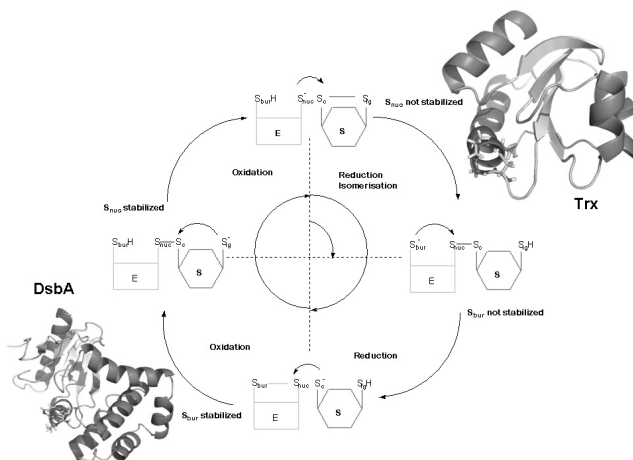


Figure 8.5. DFT calculations on small models and larger ONIOM models gave rise to a consistent line of evidence that thiol-disulfide exchange reactions regulation across the thioredoxin family are promoted by the differential stabilization of both active site cysteines.

were calculated and the possible causes for thiolate stabilization were investigated [94,95], gave rise to a consistent line of evidence, which points to the fact that both active site cysteines play an important role in the differentiation. Contrary to what was assumed, differentiation is not achieved through a different stabilization of the solvent exposed cysteine but, instead, through a fine tuning of the nucleophilicity of both active site cysteines [94,95]. The feasibility of shifting the chemical equilibrium toward oxidation, reduction, or isomerization only through subtle electrostatic effects is quite unusual, and it relies on the inherent thermoneutrality of the catalytic steps carried out by a set of chemically equivalent entities all of which are cysteine thiolates.

8.5.4 Beta-Galactosidase

Carbohydrates are involved in many cellular processes, being crucial to life. Glycosidases constitute a vast family of enzymes that catalyze the breaking and formation of glycosidic bonds. β -Galactosidase is a retaining glycosidase that catalyzes both the hydrolytic breaking of the very stable glycosidic bond of lactose, as well as a series of transglycosylation reactions [96]. It has great bi-

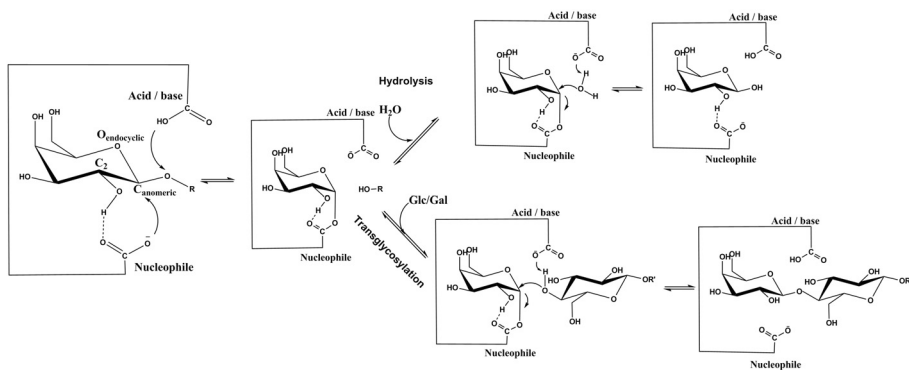


Figure 8.6. Representation of the catalytic mechanism for the hydrolysis/transglycosylation reactions, catalyzed by retaining Glycosidases.

otechnological interest for the food and pharmaceutical industries, where it is used to catalyze the large-scale production of oligosaccharides [97]. To understand the atomic-level factors that determine the outcome of the reaction (hydrolysis/transglycosylation) and the yield of each of the many transglycosylation products, an atomic level study of this catalytic mechanism was performed, using DFT and Molecular Mechanics as theoretical levels [97-100]. In order to shed some light over these topics, we have developed a model system that includes a simplified reaction center and a small substrate molecule to capture the intrinsic reactivity of the active-site motif. A very small enzyme model, containing only the two reactive carboxylic amino acids and a small substrate was used in DFT calculations [99,100]. Our results were able to confirm and provide molecular-level detail to the general mechanism proposed for this family of enzymes (Figure 8.6): a double-displacement mechanism involving a glycosylation and a deglycosylation step, in which one of these key carboxylic acids acts as a nucleophile and the other as an acid/base catalyst, with the reaction proceeding via a covalent intermediate. In the transition states (TSs), a very interesting and short hydrogen bridge is formed between the nucleophilic residue of the enzyme and the HO – C₂ of the sugar ring. Our calculations reveal that the role of this hydrogen interaction is to lower the energy of the TSs by *circa* 5 kcal mol⁻¹, contributing considerably to the stabilization of these states in both steps. A

structural rearrangement of the sugar ring is also observed; therefore, we suggest that the hydrogen bridge facilitates the change from the typical chair form to the half-chair conformation at this stage, which helps to stabilize the nascent oxocarbenium ion.

The performance of a wide variety of DFT functionals was tested in order to choose the one that better describes the thermodynamics and kinetics of our model system. The results have shown that the obtained energies indeed depend on the particular density functional employed. In our case, the correct choice of the functional was crucial as the most widely used have resulted in uncertainty in the activations energies values of over 8 kcal mol^{-1} . Comparison with the very high level calculations (MP2, MP3, MP4 and QCISD(T)) allowed for the identification of the most accurate functionals (BB1K, MPW1K and MPWB1K) for this particular reaction. Based on these conclusions, the ONIOM method (BB1K:AMBER//B3LYP:AMBER calculations) was employed to address such a large enzymatic system [98]. This enzymatic model can efficiently account for the restrained mobility of the reactive residues, as well as the long-range enzyme-substrate interactions. Figure 8.7 shows the enzymatic model studied, which includes a 15 \AA radius of the amino acids around lactose. The high-level layer (treated with quantum mechanics) is also represented. QM/MM calculations demonstrate the crucial importance of the enzyme scaffolding beyond the first-shell amino acids in the stabilization of TSs, indicating the need to include the enzyme explicitly in computational studies. Our results suggest that the role of the magnesium ion in the catalytic reaction is to lower the activation barrier by $14.9 \text{ kcal mol}^{-1}$, contributing considerably to the stabilization of the TS structure. Comparison of the energetic values for the different transglycosylation reactions ($\beta(1-3)$, $\beta(1-4)$ and $\beta(1-6)$) studied shows that these reactions are all very similar from a kinetic perspective, which seems reasonable given the similarity in the bond-breaking/bond-forming processes. However, thermodynamically, they are quite dissimilar: the formation of $\beta(1-3)$ glycosidic linkages is thermodynamically very unfavorable, whilst the formation of $\beta(1-6)$ glycosidic bonds is the most favored, in total agreement with the enantioselectivity

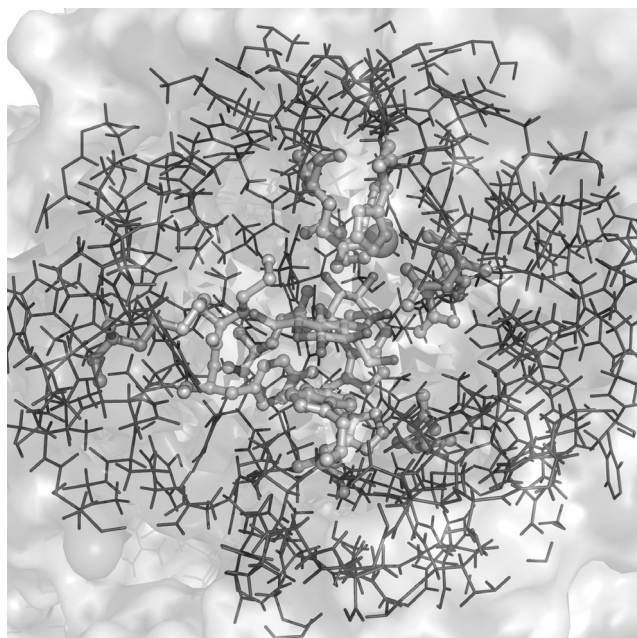


Figure 8.7. Representation of the enzymatic model (colored blue), which includes a 15 Å radius of the amino acids around lactose (shown in green and red). The magnesium ion is the yellow sphere, whereas the high-level layer (treated with quantum mechanics) is colored in orange.

observed experimentally.

As the β -Galactosidase from *E. coli* is an enzyme commonly used in molecular biology research, a complete knowledge of the different reaction pathways is crucial to the development of new chromophore substrates. Furthermore, these results help to improve the efficiency of large-scale industrial design and synthesis of new inhibitors and carbohydrates for both the pharmaceutical and food industries.

8.5.5 Farnesyltransferase

In addition to the enzymatic studies outlined above, computational methods provide particularly important insights into the study of metalloenzymes, acting as a bridge between the several spectroscopic methods normally employed to handle such systems.

In the case of the Zinc metalloenzyme FTase, for example, the large body of

experimental results available from kinetic studies on mutant FTase species and X-ray crystallography, and the information obtained from the relatively limited range of spectroscopic methods directly applicable to biological Zinc complexes, were insufficient to allow an univocal atomic-level interpretation of the catalytic mechanism followed by this enzyme [101-103]. Computational methods can be used in such cases to go beyond the traditional limitations of standard experimental studies, providing the missing pieces [104-106].

We have used first principles quantum mechanics (B3LYP) and ONIOM (B3LYP:PM3) to analyze the several possible Zn coordination modes suggested from the apparently contradicting experimental studies on the FTase resting state. Interestingly, we have found that both coordination proposal discussed in the literature – a tetracoordinated Zn sphere with a monodentate Asp ligand and water molecules versus a bidentate Asp proposal without water – were in agreement with the EXAFS structural data available and could be valid. In addition, our computational studies showed that both alternatives were at a remarkably close energetic proximity and that a conversion between the two could take place with a very small energetic barrier at room temperature. These conclusions allowed us to develop a new and unified paradigm regarding the nature of the Zn coordination sphere in FTase and the identity of the ligands present [107], considering that both alternatives exist at equilibrium, and that such process is achieved through a carboxylate-shift mechanism, where a monodentate to bidentate change (and vice-versa) by the Asp ligand helps to compensate ligand entrance (or exit) processes [108].

This main idea was later applied into the study of the several Zn sphere formed along the FTase catalytic mechanism, also with small model QM and/or QM:QM calculations [109-111], paving the way into an understanding of the more complex catalytic step, which involved a very significant conformational rearrangement of one of the substrate inside the enzyme active-site and which required significantly larger models for accurate computational modeling.

Our mechanistic studies on this enzyme culminated with the finding of the transition state intermediate of this highly concerted step [112], a structure that

could provide a blueprint for the design of more potent and specific FTase inhibitors.

8.6 Density Functional Benchmarking

According to the Hohenberg-Kohn theorems, every electronic property of a ground-state, non degenerate system, can be directly calculated by a functional where the unique variable is the electronic density [113]. Hence, opposed to wave function theories of electronic structure which treat each electron in the chemical system as a single entity, density functional theory works with the electron density as a whole. Here lies the great power of DFT, its simplicity allows its application to very large systems, with dozens or even hundredths of atoms, at much lesser computational effort than other methods where electronic correlation is also accounted for [114]. But what is the problem with DFT? In principle, DFT is exact, as the functional whom relates electron density with energy was proven to exist; in practice, the theory is only an approximation, as the functional is not fully known. For the electronic energy functional, the problematic term is the one used to calculate the exchange-correlation energy, related with the punctual interaction between electrons, either of the same spin (exchange), or opposite spin (correlation). All of the problems with DFT come from this limitation. As this term is not known, dozens of exchange-correlation functionals spawned in the field, along the years, trying to make up for this void. They are distributed among some groups, increasing in complexity: LDA (Local Density Approximation), GGA (Generalized Gradient Approximation), Meta-GGA, Hybrid-GGA and Hybrid-Meta-GGA [3,115,116]. Mostly, the functionals have empirical motivation, as well as empirical parameters, and although as a general trend more complex functionals are better, this is not an exact rule. The empirical character of the functionals make them unfit to apply outside certain chemical problems, to which they were not parameterized. Being so, there is no theoretical way of telling which functional is better for a determined system, and the only way to test it is by doing a functionals benchmarking study. Supporting this consideration one could cite a modest number of benchmarking studies,

each one aiming for certain chemical properties: geometry [117,118], kinetic barriers and thermochemistry [119], binding and dissociation energies [119], non-bonded interactions [120] There are three major concepts in a benchmarking study, each one of them with their proper considerations: (i) The model - The model system chosen to do the benchmarking should be representative of the subject being studied, as usually one is not interested in that system in particular but in a certain detail of it, such as hydrogen bonding or a phosphodiester bond hydrolysis. Furthermore, the model should be small enough, allowing the use of very accurate reference methods, which are, by definition, computationally costly; (ii) The reference value - The reference value should be as accurate as possibly for the system in question. Here, the use of composite methods can be quite resourceful. These methods take advantage of error cancellations by stating, for example, that the difference between the correlation energy at MP2 level and CCSD(T) is independent of the basis-set. Furthermore, there are various methods who emulate the use of a complete basis-set at an acceptable range of error and computer demands. By conjugating these two types of approximations one can proudly state in an abstract an energy reference value calculated at a CCSD(T)/CBS level for a system of twenty atoms; (iii) The functionals - Regarding the choice of the functionals, one could opt by testing them all, although that is hardly a good path as there are a huge number of them. Instead, a representative group should be created, one that includes functionals from all types (GGA, M-GGA) and preferentially the most recent ones. After that, in order to rank the functionals one should compare their results with the results from the reference methods.

At the end of the Benchmarking one should be able to say what functional is more adequate to evaluate a certain chemical property. Provided with this information, another researcher can wisely choose the functional for their study without going for the trendy functional - B3LYP -, the most typical situation, or without the need of doing a functional benchmarking study himself.

Benchmarking of DFT functionals for the hydrolysis of phosphodiester bonds

In our group we work mainly with enzymes and their catalytic mechanisms. These systems, large as they are, tend to be divided into layers, being the outer layer described either by molecular mechanics or semi-empirical methods, and the inner layer by DFT. Most of the times, we stand before the mentioned problem of not knowing the best functional to study the enzymatic reaction. Hence, we usually do a benchmarking study to fix this. In a recent work [121], we take dimethylphosphate hydrolysis as a model for phosphodiester bonds hydrolysis. These kinds of reactions are present on many enzymes involving DNA, RNA and phospholipids. Therefore, this study has great scope and could be applied to many important biological systems. In our particular case, the main motivation behind the study was the 3' end processing reaction of HIV Integrase.

In the work done, we described four reactions paths, all involving dimethylphosphate as the major reactant. We varied the nucleophile, which could be a molecule of water or a hydroxide ion, and the medium that could be either implicit water or vacuum. The potential energy surface for each one of these reactions was obtained at a CCSD(T)/CBS//B3LYP/6311++G(2*d*,2*p*) level, being CCSD(T)/CBS the reference energy. Subsequently, we tested a total of 52 functionals with the obtained structures. Furthermore, the performances of HF, MP2, MP3, MP4 and CCSD were also evaluated. When comparing with the reference energy, the results showed that MPWB1K, MPW1B95 and PBE1PBE are the most accurate functionals for calculating activation and reaction energies, with MUEs (Mean Unsigned Error) below 2 kcal mol⁻¹. Concerning only activation energies, MPWB1K, MPW1B95 and B1B95 give the best results. Furthermore, we take two other important conclusions from this work: the basis-set 6-311+G(2*d*,2*p*) is the most balanced one regarding the relationship between computational time and accuracy; and the inclusion of the triples excitations on CCSD(T) has major implications in the obtained energies.

The necessity for benchmarking studies will be present all the time, as the progress in this field occurs at fast pace and the number of functionals tend to

increase daily. We hope that with this and other works, researchers can always achieve the best outcome from their experiments without being overwhelmed by the enormous amount of DFT functionals available.

References

- [1] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Biol. Inorg. Chem.* **10**, 3 (2005).
- [2] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *Curr. Med. Chem.* **15**, 1478 (2008).
- [3] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *Theor. Chem. Acc.* **117**, 171 (2007).
- [4] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *Bioorg. Med. Chem.* **17**, 3369 (2009).
- [5] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **112**, 8681 (2008).
- [6] A. T. P. Carvalho, P. A. Fernandes, and M. J. Ramos, *J. Med. Chem.* **49**, 7675 (2006).
- [7] A. T. P. Carvalho, P. A. Fernandes, and M. J. Ramos *J. Phys. Chem. B* **111**, 12032 (2007).
- [8] A. L. Demain, M. Newcomb, and J. H. D. Wu, *Microbiol. Mol. Biol. Rev.* **69**, 124 (2005).
- [9] V. M. R. Pires, J. L. Henshaw, J. A. M. Prates, D. N. Bolam, L. M. A. Ferreira, C. Fontes, B. Henrissat, A. Planas, H. J. Gilbert, and M. Czjzek, *J. Biol. Chem.* **279**, 21560 (2004).
- [10] A. B. Boraston, D. N. Bolam, H. J. Gilbert, and G. J. Davies, *Biochem. J.* **382**, 769 (2004).
- [11] A. Viegas, N. F. Bras, N. Cerqueira, P. A. Fernandes, J. A. M. Prates, C. Fontes, M. Bruix, M. J. Romao, A. L. Carvalho, M. J. Ramos, A. L. Macedo, and E. J. Cabrita, *Febs J.* **275**, 2524 (2008).
- [12] N. F. Bras, N. Cerqueira, P. A. Fernandes, and M. J. Ramos, 3rd International Theoretical Biophysics Symposium, Cetraro, ITALY, Jun 16-20, pp 2030

- (2007).
- [13] N. Cerqueira, N. F. Bras, P. A. Fernandes, and M. J. Ramos, *Proteins: Struct. Funct. Bioinf.* **74**, 192 (2009).
- [14] A. L. Carvalho, A. Goyal, J. A. M. Prates, D. N. Bolam, H. J. Gilbert, V. M. R. Pires, L. M. A. Ferreira, A. Planas, M. J. Romao, and C. Fontes, *J. Biol. Chem.* **279**, 34785 (2004).
- [15] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, *Adv. Drug Delivery Rev.* **46**, 3 (2001).
- [16] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, and K. D. Kopple, *J. Med. Chem.* **45**, 2615 (2002).
- [17] T. S. Thorsen, K. L. Madsen, N. Rebola, M. Rathje, V. Anggono, A. Bach, I. S. Moreira, N. Stuhr-Hansen, T. Dyhring, D. Peters, T. Beuming, R. Haganir, H. Weinstein, C. Mulle, K. Stromgaard, L. C. B. Ronn, and U. Gether, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 413 (2010).
- [18] S. J. Wodak and J. Janin, *J. Mol. Biol.* **124**, 323 (1978).
- [19] S. J. Wodak and J. Janin, *Arch. Int. Physiol. Biochim.* **86**, 473 (1978).
- [20] S. J. Wodak and J. Janin, *Acta Crystallogr. A* **34**, S49 (1978).
- [21] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *J. Comp. Chem.* 2010, 31(2), 317.
- [22] J. Janin, K. Henrick, J. Moulton, L. Ten Eyck, M. J. E. Sternberg, S. Vajda, I. Vasker, and S. J. Wodak, *Proteins: Struct. Funct. Bioinf.* **52**, 2 (2003).
- [23] J. Janin and S. J. Wodak, *Protein Modules and Protein-Protein Interactions* **61**, 1 (2003).
- [24] Y. Han, I. S. Moreira, E. Urizar, H. Weinstein, and J. A. Javitch *Nat. Chem. Biol.* **5**, 688 (2009).
- [25] A. M. J. J. Bonvin, *Febs J.* **274**, 67 (2007).
- [26] A. D. J. van Dijk, R. Boelens, and A. M. J. J. Bonvin, *Febs J.* **272**, 249 (2005).
- [27] S. J. De Vries, A. D. J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A. M. J. J. Bonvin, *Proteins: Struct. Funct. Bioinf.* **69**, 726 (2007).

- [28] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin, *J. Am. Chem. Soc.* **125**, 1731 (2003).
- [29] A. D. J. van Dijk and A. M. J. J. Bonvin, *Bioinformatics* **22**, 2340 (2006).
- [30] A. D. J. van Dijk, S. J. De Vries, C. Dominguez, H. Chen, H. X. Zhou, and A. M. J. J. Bonvin, *Proteins: Struct. Funct. Bioinf.* **60**, 232 (2005).
- [31] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *Proteins: Struct. Funct. Bioinf.* **68**, 803 (2007).
- [32] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *Proteins: Struct. Funct. Bioinf.* **63**, 811 (2006).
- [33] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **110**, 10962 (2006).
- [34] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *Theor. Chem. Acc.* **117**, 99 (2007).
- [35] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *Int. J. Quantum Chem.* **107**, 299 (2007).
- [36] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *J. Comp. Chem.* **28**, 644 (2007).
- [37] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *J. Chem. Theory Comput.* **3**, 885 (2007).
- [38] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **111**, 2697 (2007).
- [39] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, *Theor. Chem. Acc.* **120**, 533 (2008).
- [40] D. Goodsell, G. Morris, and A. Olson *J. Mol. Recognit.* **9**, 1 (1996).
- [41] S. Gupta, L. M. Rodrigues, A. P. Esteves, A. M. F. Oliveira-Campos, M. S. J. Nascimento, N. Nazareth, H. Cidade, M. P. Neves, E. Fernandes, M. Pinto, N. M. F. S. A. Cerqueira, and N. Bras, *Eur. J. Med. Chem.* **43**, 771 (2008).
- [42] M. L. Lamb, K. W. Burdick, S. Toba, M. M. Young, K. G. Skillman, X. Q. Zou, J. R. Arnold, and I. D. Kuntz, *Proteins* **42**, 296 (2001).
- [43] C. N. Cavasotto and N. Singh, *Curr. Comput.-Aided Drug Des.* **4**, 221 (2008).

- [44] N. C. J. Strynadka, M. Eisenstein, E. KatchalskiKatzir, B. K. Shoichet, I. D. Kuntz, R. Abagyan, M. Totrov, J. Janin, J. Cherfils, F. Zimmerman, A. Olson, B. Duncan, M. Rao, R. Jackson, M. Sternberg, and M. N. G. James, *Nature Struct. Biol.* **3**, 233 (1996).
- [45] R. T. Kroemer, *Curr. Protein Pept. Sc.* **8**, 312 (2007).
- [46] D. A. Case, T. A. Darden, T. E. Cheatham III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, H. M. Merz, B. Wang, D. A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross, and P. A. Kollman, AMBER 9 (University of California, San Francisco, 2006).
- [47] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande, *J. Chem. Phys.* **119**, 5740 (2003).
- [48] J. W. Pitera and W. F. Van Gunsteren, *Mol. Simulat.* **28**, 45 (2002).
- [49] A. Blondel, *J. Comp. Chem.* **25**, 985 (2004).
- [50] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. H. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case, and T. E. Cheatham, *Accounts Chem. Res.* **33**, 889 (2000).
- [51] M. A. S. Perez, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **114**, 2525 (2010).
- [52] M. A. S. Perez, P. A. Fernandes, and M. J. Ramos, *J. Mol. Graph. Model.* **26**, 634 (2007).
- [53] W. Wang, and P. A. Kollman, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14937 (2001).
- [54] N. M. F. S. A. Cerqueira, P. A. Fernandes, and M. J. Ramos, *Recent Pat. Anti-Canc.* **2**, 11 (2007).
- [55] N. M. F. S. A. Cerqueira, S. Pereira, P. A. Fernandes, and M. J. Ramos, *Curr. Med. Chem.* **12**, 1283 (2005).
- [56] N. M. F. S. A. Cerqueira, P. A. Fernandes, L. A. Eriksson, and M. J. Ramos, *J. Mol. Struct: THEOCHEM* **709**, 53 (2004).
- [57] N. M. F. S. A. Cerqueira, P. A. Fernandes, L. A. Eriksson, and M. J. Ramos, *J. Comput. Chem.* **25**, 2031 (2004).
- [58] N. M. F. S. A. Cerqueira, P. A. Fernandes, L. A. Eriksson, and M. J. Ramos,

- Biophys. J. **90**, 2109 (2006).
- [59] P. A. Fernandes, L. A. Eriksson, and M. J. Ramos, *Theor. Chem. Acc.* **108**, 352 (2002).
- [60] S. Pereira, N. M. F. S. A. Cerqueira, P. A. Fernandes, and M. J. Ramos, *Eur. Biophys. J. Biophys.* **35**, 125 (2006).
- [61] P. A. Fernandes and M. J. Ramos, *Chem.-Eur. J.* **9**, 5916 (2003).
- [62] N. M. F. S. A. Cerqueira, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **110**, 21272 (2006).
- [63] S. Pereira, P. A. Fernandes, and M. J. Ramos, *J. Comput. Chem.* **25**, 227 (2004).
- [64] S. Pereira, P. A. Fernandes, and M. J. Ramos, *J. Comput. Chem.* **25**, 1286 (2004).
- [65] N. M. F. S. A. Cerqueira, P. A. Fernandes, and M. J. Ramos, *Chem.-Eur. J.* **13**, 8507 (2007).
- [66] S. Pereira, P. A. Fernandes, and M. J. Ramos, *J. Am. Chem. Soc.* **127**, 5174 (2005).
- [67] P. A. Fernandes and M. J. Ramos, *J. Am. Chem. Soc.* **125**, 6311 (2003).
- [68] D. F. A. R. Dourado, P. A. Fernandes, B. Mannervik, and M. J. Ramos, *Curr. Protein Pept. Sci.* **9**, 325 (2008).
- [69] A. Yoritaka, N. Hattori, K. Uchida, M. Tanaka, E. R. Stadtman, and Y. Mizuno, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 2696 (1996).
- [70] Y. J. Li, S. A. Oliveira, P. Xu, E. R. Martin, J. E. Stenger, C. R. Scherzer, M. A. Hauser, W. K. Scott, G. W. Small, M. A. Nance, R. L. Watts, J. P. Hubble, W. C. Koller, R. Pahwa, M. B. Stern, B. C. Hiner, J. Jankovic, C. G. Goetz, F. Mastaglia, L. T. Middleton, A. D. Roses, A. M. Saunders, D. E. Schmechel, S. R. Gullans, J. L. Haines, J. R. Gilbert, J. M. Vance, M. A. Pericak-Vance, C. Hulette, and K. A. Welsh-Bohmer, *Hum. Mol. Genet.* **12**, 3259 (2003).
- [71] T. J. Montine, D. Y. Huang, W. M. Valentine, V. Amarnath, A. Saunders, K. H. Weisgraber, D. G. Graham, and W. J. Strittmatter, *J. Neuropath. Exp. Neur.* **55**, 202 (1996).
- [72] R. J. Mark, M. A. Lovell, W. R. Markesbery, K. Uchida, and M. P. Mattson,

- J. Neurochem. **68**, 255 (1997).
- [73] D. A. Butterfield, Free Radical Res. **36**, 1307 (2002).
- [74] H. Kolsch, M. Linnebank, D. Lutjohann, F. Jessen, U. Wullner, U. Harbrecht, K. M. Thelen, M. Kreis, F. Hentschel, A. Schulz, K. von Bergmann, W. Maier, and R. Heun, Neurology **63**, 2255 (2004).
- [75] W. Palinski, M. E. Rosenfeld, S. Yla-Herttuala, G. C. Gurtner, S. S. Socher, S. W. Butler, S. Parthasarathy, T. E. Carew, D. Steinberg, and J. L. Witztum, Proc. Natl. Acad. Sci. U.S.A. **86**, 1372 (1989).
- [76] G. Jurgens, Q. Chen, H. Esterbauer, S. Mair, G. Ledinski, and H. P. Dinges, Arterioscler. Thromb. **13**, 1689 (1993).
- [77] J. H. Brasen, T. Hakkinen, E. Malle, U. Beisiegel, and S. Yla-Herttuala, Atherosclerosis **166**, 13 (2003).
- [78] K. Tsuneyama, K. Harada, N. Kono, M. Sasaki, T. Saito, M. E. Gershwin, M. Ikemoto, H. Arai, and Y. Nakanuma, J. Hepatol. **37**, 176 (2002).
- [79] M. Parola, G. Bellomo, G. Robino, G. Barrera, and M. U. Dianzani, Antioxid. Redox Signal **1**, 255 (1999).
- [80] E. R. Stadtman, Science **257**, 1220 (1992).
- [81] N. H. Ansari, L. Wang, and S. K. Srivastava, Biochem Mol Med **58**, 25 (1996).
- [82] A. M. Caccuri, G. Antonini P. G. Board, M. W. Parker, M. Nicotra, M. Lo Bello, G. Federici, and G. Ricci, Biochem. J. **344** Pt 2, 419 (1999).
- [83] D. F. A. R. Dourado, P. A. Fernandes, B. Mannervik, and M. J. Ramos, Chem.-Eur. J. **14**, 9591 (2008).
- [84] D. F. A. R. Dourado, P. A. Fernandes, and M. J. Ramos, Theor Chem Acc, **124**, 71 (2009).
- [85] M. Widersten, R. Bjornestedt, and B. Mannervik, Biochemistry **35**, 7731 (1996).
- [86] E. Lindahl, B. Hess, and D. Van der Spoel, J. Mol. Model **7**, 306 (2001).
- [87] E. J. Sorin and V. S. Pande, Biophys. J. **88**, 2472 (2005).
- [88] W. D. Cornell, P. Cieplak, C. I. Bayly, I.R. Gould, K.M. Merz Jr., D.M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P.A. Kollman, J.

- Am. Chem. Soc. **117**, 5179 (1995).
- [89] J. M. Wang , P. Cieplak, and P. A. Kollman, *J. Comput. Chem.* **21**, 1049 (2000).
- [90] R. Bjornestedt, G. Stenberg, M. Widersten, P. G. Board, I. Sinning, T. A. Jones, and B. Mannervik, *J Mol Biol* **247**, 765 (1995).
- [91] D. F. A. R. Dourado, P. A. Fernandes, B. Mannervik, and M. J. Ramos, *J. Phys. Chem. B* **114**, 1690 (2010).
- [92] A. T. P. Carvalho, P. A. Fernandes, M. Swart, J. N. P. van Stralen, F. M. Bickelhaupt, and M. J. Ramos, *J. Comp. Chem.* **30**, 710 (2009).
- [93] A. T. P. Carvalho, M. Swart, J. N. P. van Stralen, P. A. Fernandes, M. J. Ramos, and F. M. Bickelhaupt, *J. Phys. Chem. B* **112**, 2511 (2008).
- [94] A. T. P. Carvalho, P. A. Fernandes, and M. J. Ramos, *J. Comp. Chem.* **27**, 966 (2006).
- [95] A. T. P. Carvalho, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **110**, 5758 (2006).
- [96] D. H. Juers, T. D. Heightman, A. Vasella, J. D. McCarter, L. Mackenzie, S. G. Withers, and B. W. Matthews, *Biochemistry* **40**, 14781 (2001).
- [97] N. F. Bras, P. A. Fernandes, and M. J. Ramos, *Theor. Chem. Acc.* **122**, 283 (2009).
- [98] N. F. Bras, P. A. Fernandes, and M. J. Ramos, *J. Chem. Theory Comp.* **6**, 2565 (2010).
- [99] N. F. Bras, P. A. Fernandes, and M. J. Ramos, *J. Mol. Struct.: THEOCHEM* **946**, 125 (2010).
- [100] N. F. Bras, S. A. Moura-Tamames, P. A. Fernandes, and M. J. Ramos, *J. Comp. Chem.* **29**, 2565 (2008).
- [101] D. A. Tobin, J. S. Pickett, H. L. Hartman, C. A. Fierke, and J. E. Penner-Hahn, *J. Am. Chem. Soc.* **125**, 9962 (2003).
- [102] S. B. Long, P. J. Casey, and L. S. Beese, *Nature* **419**, 645 (2002).
- [103] H. W. Park, S. R. Boduluri, J. F. Moomaw, P. J. Casey, and L. S. Beese, *Science* **275**, 1800 (1997).
- [104] M. J. Ramos and P. A. Fernandes, *Accounts Chem. Res.* **41**, 689 (2008).

- [105] M. Leopoldini, T. Marino, M. D. Michelini, I. Rivalta, N. Russo, E. Sicilia, and M. Toscano, *Theor. Chem. Acc.* **117**, 765 (2007).
- [106] F. Himo, *Theor. Chem. Acc.* **116**, 232 (2006).
- [107] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *Biophys. J.* **88**, 483 (2005).
- [108] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Am. Chem. Soc.* **129**, 1378 (2007).
- [109] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Comp. Chem.* **28**, 1160 (2007).
- [110] S. F. Sousa, P. A. Fernandes and M. J. Ramos, *Proteins: Struct. Funct. Bioinf.* **66**, 205 (2007).
- [111] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Mol. Struct.: THEOCHEM* **729**, 125 (2005).
- [112] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *Chem.-Eur. J.* **15**, 4243 (2009).
- [113] W. Kohn, A. D. Becke, and R. G. Parr, *J. Phys. Chem.* **100**, 12974 (1996).
- [114] A. Ghosh, *J. Biol. Inorg. Chem.* **11**, 671 (2006).
- [115] Y. Zhao and D. G. Truhlar, *Accounts Chem. Res.* **41**, 157 (2008).
- [116] N. E. Schultz, Y. Zhao, and D. G. Truhlar, *J. Phys. Chem. A* **109**, 11127 (2005).
- [117] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, *J. Phys. Chem. B* **111**, 9146 (2007).
- [118] P. Jurecka, J. Cerny, P. Hobza, and D. R. Salahub, *J. Comput. Chem.* **28**, 555 (2007).
- [119] A. Karton, A. Tarnopolsky, J. F. Lamere, G. C. Schatz, and J. M. L. Martin, *J. Phys. Chem. A* **112**, 12868 (2008).
- [120] Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.* **3**, 289 (2007).
- [121] A. J. M. Ribeiro, M. J. Ramos, and P. A. Fernandes (submitted).