Ab initio study of alanine-based polypeptides secondary-structure motifs in the gas phase

Vorgelegt von
M. Sc.
Mariana Rossi Carvalho
aus Berlin

Von der Fakultät II – Mathematik und Naturwissenschaften
der Technische Universität Berlin
zur Erlangung des akademischen Grades
Doktorin der Naturwissenschaften
- Dr. rer. nat. -

Vorgelegte Dissertation

Promotionsausschuss:
Vorsitzender: Prof. Dr. Thomas Möller
Berichter: Prof. Dr. Andreas Knorr
Berichter: Prof. Dr. Matthias Scheffler
Berichter: Prof. Dr. Roland Netz
Tag der Wissenchaftlichen Aussprache: 06.12.2011

Berlin - 2011
D-83

Ao meu pai e à minha mãe

Contents

Abstract

There is a variety of theoretical descriptions today regarding peptides and proteins. Ideally, one would like to simply resort to the full quantum mechanics of the many-body Schrödinger's equation. However, for systems with several hundreds of atoms, and accuracy requirements of perhaps a few tens of meV in relative energies, the use of these "first principles" methods remain an active challenge. In this work, systematic steps towards a fully "first principles" description of the secondary structure of polypeptides are presented. The typical secondary structure elements (the structural building blocks) are helices, sheets, and turns. Here, we focus on helical motifs formed by the alanine-based polypeptides Ac-Ala$_n$-LysH$^+$, n=4-15, in the gas-phase. The gas-phase provides a "clean-room" environment which allows a direct assessment of the intramolecular forces that stabilize secondary structure, as well as an unambiguous comparison between theory and experiment.

The challenge of exploring the large conformational space is addressed for n=4-8. Starting from a conformational pre-screening in the OPLS-AA force field, thousands of conformers' relaxations are performed, using density-functional theory (DFT), as implemented in the all-electron/localized basis code FHI-aims [1]. The PBE generalized gradient approximation, augmented with a C$_6$/R^6 van der Waals correction[2] (PBE+vdW), was employed. The α-helical structure is found *not* to be the lowest energy conformer for small n, being preferred on the PES only for n $\geq$7, but only barely. Upon inclusion of vibrational free energy (harmonic approximation), these helices are further stabilized at finite temperatures over globular structures. For $n \geq 8$, the α-helix is predicted to be the only accessible conformer in the free-energy surface at 300 K, in agreement with indirect experimental data [3]. The inclusion of vdW effects are essential for obtaining reliable energy hierarchies and for the stabilization of the helical motif over other structures at n=8.

The structure predictions are corroborated by the calculation of ion-mobility cross sections and the comparison of computed IR spectra with room temperature gas-phase IRMPD spectra for n=5, 10, and 15. A quantitative validation based on Pendry's reliability factor is presented. The inclusion of anharmonic effects into calculated spectra, by means of *ab initio* molecular dynamics, improves substantially the agreement between experiment and theory. The longer molecules (n=10,15) are predicted to be firmly α-helical, as expected. For n=5, a mix of conformers is predicted to be present in experiment, including the lowest-energy PBE+vdW conformer, which is not a simple helix [4].

The experimentally observed [5] high temperature (≈ 700 K) α-helical stability of Ac-Ala$_{15}$-LysH$^+$ in the gas-phase is studied by trying to unfold the molecule, using extended-time Born-Oppenheimer *ab initio* molecular dynamics simulations. The molecule unfolds rapidly (within a few ps) at T=800 K and 1000 K, but not at 500 K or 700 K. Most importantly, the observed stability depends critically not just on a correct inclusion of H-bonds and the designed termination effects, but also on vdW interactions. If vdW forces are not properly included, the helix unfolds already at 700K, and the structural stability at 500K is mostly 3$_{10}$-helical, whereas including vdW effects the α-helical structure is favored [6].

Finally, benchmark data for the energy hierarchies were obtained from higher level methods, as, e.g., exact exchange plus correlation treated in the random-phase approximation (EX+cRPA) and Møller-Plesset perturbation theory (MP2). Additional numeric atom-centered orbital (NAO) basis sets were specifically developed for converging relative energies within these methods. EX+cRPA based on PBE orbitals agrees with the energetic hierarchy found with the PBE+vdW functional for n=5. It is observed, however, that for these methods (EX+cRPA and MP2) there is a strong dependence of the predicted relative energies on the starting point, i.e. the orbitals.

Zusammenfassung

Es gibt heute eine Vielzahl von Möglichkeiten zur theoretischen Beschreibung von Peptiden und Proteinen. Ideal wäre es, auf die volle quantenmechanische Vielteilchen Schrödinger-Gleichung zurückzugreifen; allerdings stellen solche “first principles” Berechnungen für Systeme mit mehreren hundert Atomen und Genauigkeitsanforderungen an die relative Energie von einigen Zehntel meV eine große Herausforderung dar. Diese Arbeit nähert sich in systematischer Weise dieser exakten Beschreibung der Sekundärstruktur von Peptiden mittels “ersten Prinzipien”. Die typischen Sekundärstrukturelemente (die Strukturbausteine) sind Helices, Faltblätter und Turns. Das Augenmerk der vorliegenden Arbeit liegt auf helikalen Strukturmotiven in Alanin-basierten Polypeptiden (Ac-Ala$_n$-LysH$^+$, n=4-15) in der Gasphase. Derartige Studien bieten Reinraumbedingungen zur unmittelbaren Untersuchung der Sekundärstruktur-stabilisierenden intramolekularen Kräfte und erlauben den eindeutigen Vergleich zwischen Theorie und Experiment.

Für Modellpeptide mit n=4-8 wurde der riesige Konformationsraum untersucht. Ausgehend von einer Struktursuche mit dem OPLS-AA Kraftfeld wurden tausende Geometrieoptimierungen auf Basis von Dichtefunktionaltheorie (DFT), in einer Allelektronen-Implementierung (FHI-aims Programmpaket [1]), durchgeführt. Es wurde das durch eine C$_6$/R^6 van der Waals (vdW) Korrektur erweiterte PBE Funktional (PBE+vdW) benutzt [2]. Bei kurzen Peptiden wird die α-Helix nicht als niederenergetischstes Konformer gefunden, erst ab $n \geq$7 ist sie leicht bevorzugt. Die Berücksichtigung von vibronischen Freien Energien (harmonische Näherung) stabilisiert die α-Helix weiter und bevorzugt sie gegenüber globulären Strukturen. Für $n \geq$8 stellt die α-Helix bei 300 K, im Einklang mit experimentellen Daten [3], die einzige zugängliche Struktur in der “Landschaft” der Freien Energie dar. Um zuverlässige Energiehierarchien zu erhalten, z. B. für den “Kreuzungspunkt” der Stabilität der α-Helix gegenüber anderen Konformeren bei n=8, ist die Berücksichtigung von vdW-Effekten essentiell.

Diese Strukturvorhersagen werden durch die Berechnung von Ionenmobilitätsquerschnitten und den Vergleich von simulierten IR-Spektren mit Raumtemperatur IRMPD-Spektren in der Gasphase für n=5, 10 und 15 untermauert. Eine Quantifizierung der Übereinstimmung erfolgt auf Basis des Pendry R-Faktors. Die Berücksichtigung anharmonischer Effekte in den berechneten IR-Spektren durch *ab initio* Molekulardynamiksimulation verbessert die Übereinstimmung von Theorie und Experiment deutlich. Wie erwartet werden die längeren Moleküle (n=10, 15) als eindeutig α-helikal vorhergesagt. Für n=5 wird ein Ensemble von Konformeren vorhergesagt, wobei der stabilste PBE+vdW Konformer nicht α-helikal ist [4].

Um die experimentell beobachtete hohe Temperaturstabilität ($\approx$ 700 K) [5] der α-Helix von Ac-Ala$_{15}$-LysH$^+$ in der Gasphase zu untersuchen, wurden lange Born-Oppenheimer *ab initio* Moleküldynamiksimulationen durchgeführt. Das Molekül entfaltet schnell (innerhalb weniger ps) bei Temperaturen von 800 K oder 1000 K, aber nicht bei 500 K oder 700 K. Hervorzuheben ist, dass die α-Helix Stabilität, auch hinsichtlich der Dynamik, nicht nur von Wasserstoffbrücken und den Effekten der Terminierung abhängt, sondern auch von vdW-Wechselwirkungen. Sind diese Effekte nicht korrekt berücksichtigt, entfaltet sich die Helix bereits bei 700 K und bei 500 K wird vor allem die 3$_{10}$-helikale Struktur bevorzugt [6].

Es wurden Referenzdaten für Exakten Austausch mit *random-phase approximation* Korrelationen (EX+cRPA) und Møller-Plesset Störungstheorie (MP2) berechnet. Zusätliche *numeric atom-centered orbital* (NAO) Basissätze wurden entwickelt, um relative Energien mit diesen Methoden zu konvergieren. Auf PBE-Orbitalen basierende EX+cRPA Energiehierarchien stimmen mit PBE+vdW bei n=5 überein. Für diese Methoden (EX+cRPA und MP2) stellt man eine starke Abhängigkeit der relativen Energien vom Startpunkt, d.h. den zugrundeliegenden Orbitalen, fest.

Chapter 1

Introduction

Life as we know it is based on processes happening at the molecular scale. These processes involve the interactions of small molecules, like water, ions, etc., as well as of potentially extremely large biomolecules (e.g. DNA, lipids, sugars, or proteins). These interactions give rise to an enormous variety of complex processes that, taken together, allow ”life” to take place. Many details about these processes are not known, making them one of the most daunting remaining mysteries of bio-science. A few of the unanswered questions are the following: What are the rules that govern gene expression? How did chemical compounds come together to form bio-molecules (DNA, RNA), that eventually led to the formation of living organisms? How do proteins fold to their native structure, allowing them to perform their unique function?; What determines the stability and kinetics of a protein? Why do some proteins misfold permanently, provoking diseases (e.g. Creutzfeldt-Jakob, BSE a.k.a. mad cow, Alzheimer, Diabetes Type 2, etc.)?

The answers to these questions do not lie in a single field of science, but in many: physics, chemistry, and biology meet to approach these problems from the atomistic, chemical, statistical mechanics, and phenomenological sides. Proteins, specifically, are one of the most studied biomolecules, due to the prominent role they play in living organisms, performing the tasks encoded in the genes. For example, proteins serve as “messengers” (e.g. hormones), carrying information between different cells, or ”antibodies”, protecting the body against external agents, or “enzymes” helping the chemical reactions to take place, or a storage place, binding to ions or other small molecules, as well as carriers of these molecules between different structures, etc. The first measurement of a protein structure has been performed more than 60 years ago, in the late 1950s [7]. However, even nowadays a fully predictive theoretical framework, that can describe the stability and dynamics of a protein *quantitatively* (and not only qualitatively), with quantum-mechanics accuracy (without evoking empirical parameters), is still missing.

The non-empirical description of protein structure and dynamics is a great challenge. Although its folded (native) form, in a given environment, is proposed to be encoded in its amino acid (building units) sequence [8], the size of its conformational space is huge. The famous Levinthal paradox [9, 10] states that if even a small protein of $\approx$100 amino acids would probe all its possible configurations very fast ($\approx 10^{13}$ configurations per second), in order to arrive to its native form, it would take longer than the age of Earth. Obviously proteins in living organisms fold much faster than that (we exist, after all). In fact, they do so in a matter of seconds or less[1] [10–13].

[1] This does not exclude the possibility of a class of proteins having a longer folding time. These were probably simply dumped by natural evolution, thus not being found in living organisms.

The energy landscapes explored by proteins to find their respective native structures (fold), are, potentially, of extreme complexity, involving interactions with the environment and other molecules (chaperones or regulating enzymes that control possible folding errors). These interactions are determined by the electronic structure of the molecules involved. The energy landscape actually reflects the statistical average associated with finite-temperature thermodynamics in the physiological environment. Ideally, thus, "all" one would have to do, in order to make a fully quantitative and accurate prediction, would be to solve the Schrödinger equation for the (thousands or millions of) electrons of a protein interacting with its environment and find the structure of minimum free energy (connecting to statistical mechanics), at certain thermodynamic conditions. Unfortunately, this task is not feasible as stated, involving a very complicated quantum many-body problem. Moreover, the time-scale is an issue: 1 second (or even 1ms or μs) seems very little to our daily life, but not if you plan to describe all those thousands of electrons outrightly and evolve them in time. This is related also to the combinatorial explosion of possible conformations to be explored. Only recently, in the past one or two decades, computers have had enough power to aid scientists in solving at least the electronic structure of large molecules (but not full proteins!), based on the law of quantum mechanics, even if only in an approximate way. Nowadays, we have powerful computers, state-of-the-art theories, and approximations that allow the quantum-mechanical many-body problem to be solved with a reasonable computational cost (*first-principles* methods). Nevertheless, we are still far away from the "holy-grail" that would be having a predictive and quantitative theory that predicts the structure and dynamics of physiologically relevant biomolecules.

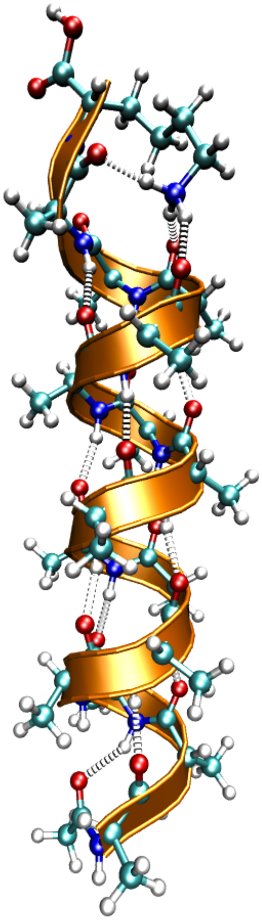


Figure 1.1: Representation of the Ac-Ala$_{15}$-LysH^{+} polypeptide in an α-helical configuration, corresponding to the theoretical (DFT-PBE+vdW) ground state. The acetate group is on the bottom, the helix formed by the alanine amino-acids is in the middle and the charged lysine residue (containing the NH$_3^+$) group on the top. H-bonds are drawn in grey. The polypeptide series Ac-Ala$_n$-LysH$^+$, $n = 4 - 15$, is the model system of this thesis.

The work presented in this thesis attempts to make a theoretical, quantum-mechanics based, quantitative contribution towards describing small, basic pieces of the problems listed above, with sufficient accuracy. Even taking a reductionist approach here, this is a huge challenge in itself. In fact, also the applicable "first-principles" methods available today require further research, making it necessary to compare to model, benchmark studies in order to assess their quality and predictive power. For example, even the so-called "weak", long-range, non-covalent, van der Waals interactions are challenging to be described, but of great relevance to biomolecules [14, 15].

The model systems of choice in this thesis are alanine-based polypeptides in the gas-phase (i.e. in the absence of solvent). A polypeptide is a small protein, and alanine is an amino acid. Amino acids are molecules that have in common a sequence of nitrogen, carbon and oxygen atoms (to be further explained in Section 2.1), differing from one another by the so-called side chain. Alanine is a relatively simple amino acid, having its side chain formed by the small CH$_3$ group. Alanine-based polypeptides (or polyalanine) are an established model system, both experimentally [16–24] and theoretically [25–36], for at least 40 years. The alanine amino acid has a very strong tendency to form helices, as, e.g., shown in Figure 1.1. Helices are ubiquitous structures in proteins, consisting of one of the most important protein building blocks, referred to as "secondary structure". The polypeptides studied here have between 5 and 16 amino acids. They have the proper length for studying helix formation, while still being small enough to allow a first-principles treatment.

Alanine-based polypeptides have been shown to form helices not only in solvents, but also in the gas-phase [37], so that clean and precise experimental data exists in the literature [3, 5, 37–39]. Studying these systems in the gas-phase presents an unique chance to study these molecules in a clean-room environment, where the *intra*molecular stabilizing interactions can be separated from those of the environment (solvent, membrane, protein interior). Only after the intramolecular interactions are understood and well described, the next steps (e.g. interaction with environment, larger molecules) can be safely tackled.

In this thesis the series of polypeptides first studied in 1998 by R. Hudgins, M. Ratner, and M. Jarrold [37], called Ac-Ala$_n$-LysH$^+$ $(n > 4)$, will be treated. The prefix "Ac" stands for the acetate group, "Ala" for the alanine amino acid, and "LysH$^+$" for the protonated lysine amino acid. The n=15 member of this series is pictured in Figure 1.1, in its theoretical ground state (α-helix) configuration. Part I of this thesis provides a comprehensive introduction on: (i) the properties of polypeptides and their energy landscapes; (ii) methods to describe the potential energy surface (PES), in particular density-functional theory (DFT); (iii) methods to describe the movement of the nuclei on the PES; and (iv) some details about the electronic structure code used (FHI-aims), to which development the author of the thesis has also significantly contributed. Part II of the thesis focuses on: (i) the characterization of polyalanine as a model system for studying secondary-structure formation, including a summary of the experiments that will be directly relevant and compared to in this work; (ii) the challenge of exploring the huge conformational space for the smaller members of the Ac-Ala$_n$-LysH$^+$ series (n=4-8, 70 to 110 atoms); (iii) studying the reasons for helical structure stabilization (vdW and free energies); (iv) and direct comparisons of the first principles predictions to experiment, by calculating ion-mobility cross-sections, IR spectra, including dynamic, configurational, temperature, and anharmonic contributions, and calculating the high-temperature dynamical stability of Ac-Ala$_{15}$-LysH$^+$ (from first-principles). Finally, Part III is dedicated to obtaining benchmark data from high-level first-principle methods. The development of additional numeric atom-centered basis sets that allow relative energy convergence within electronic-structure methods including non local correlation is described, and comparison of energy hierarchies of Ac-Ala$_5$-LysH$^+$ to such methods is shown.

The polypeptides studied here are much smaller than real-life proteins, but secondary structure motifs often have a length similar to what will be regarded in this work. These molecules are, thus, of the ideal size to study helical secondary structure in depth, while still retaining high accuracy with state-of-the-art first-principles methods.

Part I

Polypeptides and the potential-energy surface

Chapter 2

Peptides, proteins, and their energy landscapes

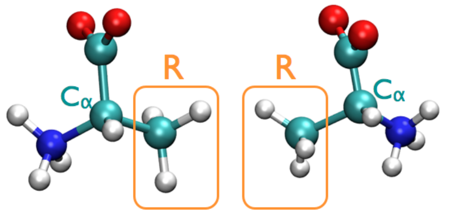
2.1 Peptide and protein formation

Polypeptides and proteins are macromolecules of sizes ranging from 20 to $\approx$ 500 000 atoms [40, 41]. They perform many different functions in a living organism, being the real “workers” maintaining life. To name only a few functions, they can mediate chemical processes like photosynthesis, they serve as carriers of other substances, like in blood (hemoglobin), they help carrying ions in and out of the cell, etc. All known proteins are built from only 22 different molecules called (natural) amino acids, schematically depicted in Figures 2.1 and 2.2. They are organic molecules presenting a carboxyl group (COOH) and an amino group (NH_2), hence the name. They differ from one another only by the so-called side-chain, marked with the letter “R” in 2.1. Additionally, natural amino acids usually exhibit one chiral center, which is a carbon atom commonly called C_α, labeled in Figure 2.1. Exceptionally, glycine has no chiral center, since its side chain is only a hydrogen atom, and some amino-acids have more than one (in their side chain). The orientation of the connecting groups (hydrogen, carboxyl, amino, and side-chain) to this carbon defines the two possible optically active isomers [L- (levo) or D- (dextro)], as can be seen in Figure 2.1. The color scheme chosen to represent these molecules in Figure 2.1 will be the same throughout this thesis: carbon atoms are colored cyan, hydrogen atoms are colored white, oxygen atoms are colored red, and nitrogen atoms are colored blue. This color scheme, with atoms and bonds represented as balls and sticks, respectively, is called CPK, after Corey, Pauling, and Koltun [42]. The color of the carbon atom varies (sometimes it is grey), though, in different visualization programs.

Almost all amino acids forming proteins are of the L type. The reason for that remains unknown, being an active field of research [43]. The different side-chains of 20 protein-forming amino acids are shown in Figure 2.2, along with their names, one letter code and three letters code used to refer to them. [1] In solution (depending on the pH value), the amino acid usually assumes a zwitterion configuration, where its terminations become charged by the formation of a COO^- and a NH_3^+ group.

In Figure 2.2, the 5 amino acids of the first row contain a polar side chain with a high proton affinity (being charged under physiological pH conditions), the first 4 amino acids of the second row contain a polar but uncharged side chain under normal pH conditions, the next 3 in the second row are considered

[1] The two amino acids that are not shown are called selenocysteine and pyrrolysine. Pyrrolysine is only found in a small number of methanogenic bacteria. Selenocysteine is found in a small quantity in all living organisms, being important for the catalytic activity of certain proteins.



(a) L-amino acid (b)D-amino acid

Figure 2.1: Schematic representation of the Alanine amino acid in its zwitterion configuration. Red atoms are oxygen, blue atoms are nitrogen, white atoms are hydrogen and cyan atoms are carbon. The R symbol stands for the side-chain, here represented by the CH_3 group, which differs for each amino acid. In (a) L-amino acid, with respect to the central C_α carbon and in (b) a D-amino acid.

"special cases" (contain sulphur, or no chiral center, or no amino group), and finally, in the last row all amino acids have a hydrophobic side chain.

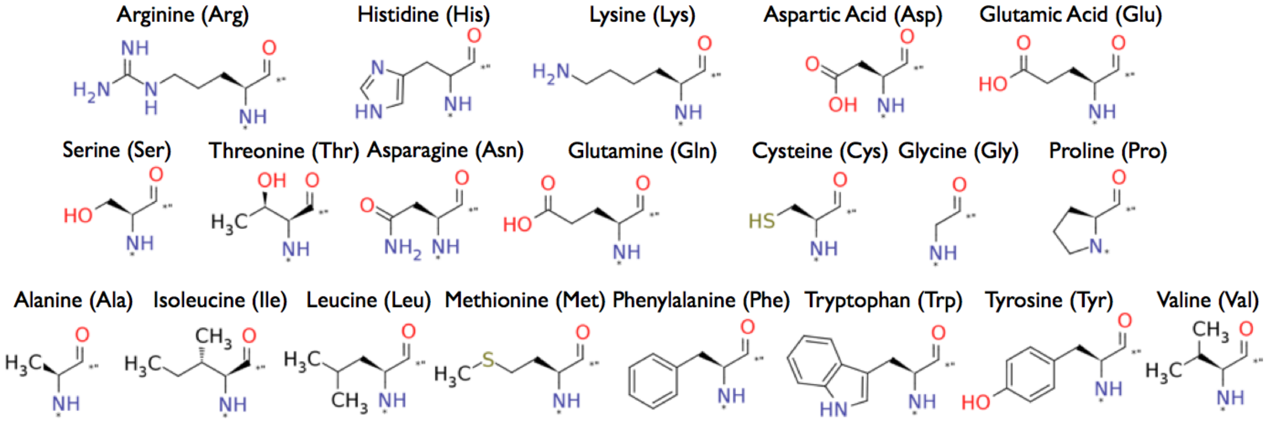


Figure 2.2: Table with 20 (most common) natural amino acids. Blue "N" are nitrogens, red "O" are oxygens, black "C" are carbons, and light-brown "S" is sulphur. Hydrogens, where explicitly shown, are represented by the letter "H". Each "kink" represents one carbon atom (with its connecting hydrogens) that is not shown explicitly. Three letter name code for each of them are also shown.

The biosynthesis of amino acids starts in the nucleus of a cell. Only the essential features of this process are summarized here; detailed accounts can be found in textbooks, for example the book by Voet and Voet [44]. In the nucleus, the DNA is transcribed into RNA, which leaves the nucleus and goes into the cytoplasm of the cell. This RNA is then translated, amino acid by amino acid, inside ribosomes and with the help of enzymes, to form the protein. Two amino acids bind to each other through a bond known as the peptide bond. This bond takes place between the carboxyl group of one amino acid and the amino group of another, with formation of a water molecule (the process is schematically represented in Figure 2.3). The name "polypeptide" is then given to short chains of amino acids. Longer chains are called *proteins*. The definition of the size when a polypeptide becomes a protein is not well established, but usually polypeptides of more than 50 amino acids are considered proteins [44]. The repeating sequence (N-C_α-C(O))$_n$ in a protein or polypeptide is called the *backbone*. After formation of the peptide bonds, the end of the peptide where the NH_2 group is located is called the N-terminus, while that where the COOH group is located is called the C-terminus, both labeled in Figure 2.4.

The peptide bond lies in a plane comprising the C^i_α-C(O)-N-C^{i+1}_α backbone atoms. The dihedral angle[2] defined precisely by these atoms is called ω and has the value of $\approx 0^o$ for what is called the *cis* configuration and $\approx 180^o$ for what is called the *trans* configuration. Due to the L chirality of the amino

[2] A dihedral angle is an angle between two planes. It needs at least 4 points to be defined, so that 2 intercepting planes can be drawn.

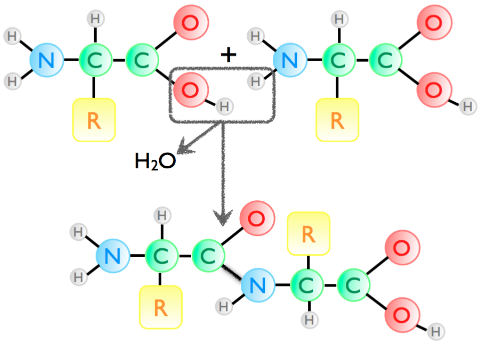


Figure 2.3: Schematic representation of the formation of the so-called peptide bond: Two amino-acids react to form a peptide via the production of a water molecule.

acids and the position of the side chains, in folded proteins the *trans* configuration is usually preferred for steric reasons. The backbone configuration can then be characterized by other 2 dihedral angles: the ϕ angle defined by the sequence C(O)-N-C_α-C(O); and the ψ angle defined by the sequence N-C_α-C(O)-N. These angles adopt preferred values already for the isolated amino-acids[45], which are then reflected in the peptide backbone, as will be further discussed in Section 2.2.3. All these angles are schematically shown in Figure 2.4.

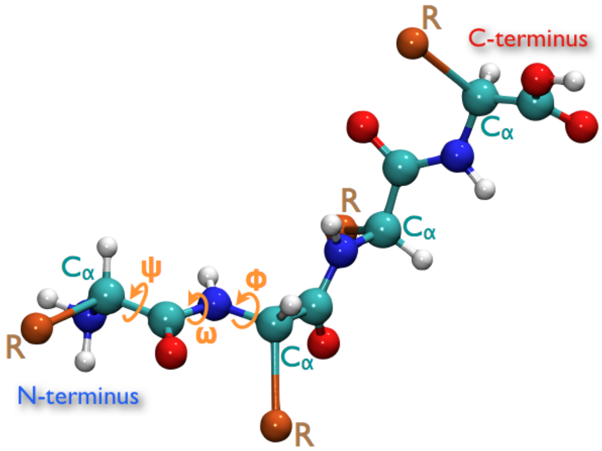


Figure 2.4: Schematic representation of a generic polypeptide, with the C_α carbons and the side-chains R labeled. The dihedral angles ω, ϕ, and ψ are specified.

Each polypeptide or protein forms an unique three-dimensional structure (which still depends on environmental factors, such as nature of solvent, pH value, specific ions etc.), that finally defines its function in an organism. Since each protein is formed by a different sequence of amino acids, it has been proposed [8] that the final native structure of a protein in a given environment should be encoded in its amino acid sequence alone. This will be discussed in more detail in Section 2.3, but before going deeper into that subject, a better characterization of the structure of proteins and polypeptides will be given.

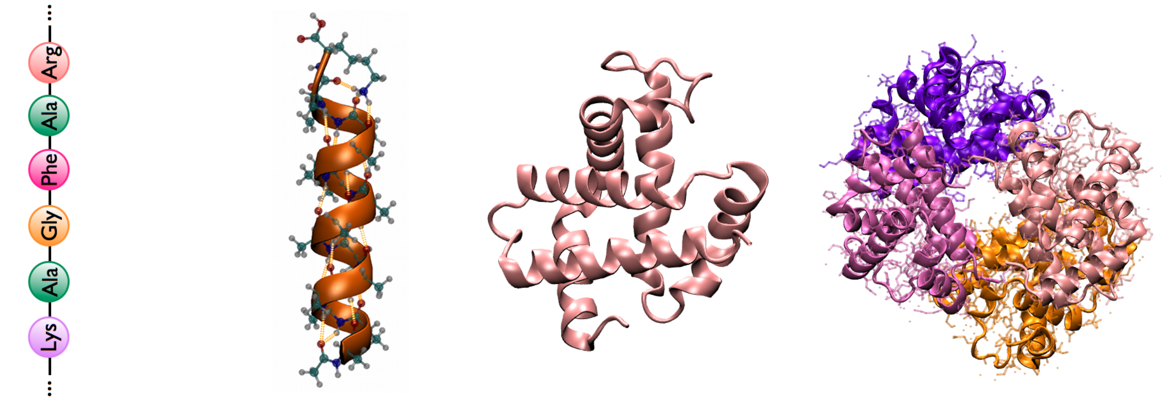
2.2 Primary, secondary, tertiary, quaternary structure

The first protein to have its 3D structure revealed was myoglobin, in 1958, via X-Ray analysis [7]. Although the experiment was of low resolution, different “macro” features could already be identified in different parts of the protein.

In fact, the existence of contrasting structural features in proteins had been proposed already in 1951, in a series of landmark papers published by Linus Pauling and Robert Corey. They correctly proposed, based on studies of the crystal structure of peptides and geometrical considerations, the existence of what is known today as the two main structural elements of proteins: the α-helix and the β-sheet [46–48].

The classification of protein structures into “classes” appeared also in 1951, coined by K. Linderstrøm-Lang [49]. The proposed classification, used until today, is:

- Primary structure: the sequence of amino acids composing a protein [Figure 2.5(a)].
- Secondary structure: geometric form of localized segments of amino acids, “the building blocks” (e.g. α-helices, β-sheets, turns)[Figure 2.5(b)].
- Tertiary structure: full 3D structure of the protein, composed of various connected secondary structure elements [Figure 2.5(c)].
- Quaternary structure: aggregation of several proteins [Figure 2.5(d)].



(a) Primary structure (b) Secondary structure (c) Tertiary structure (d) Quaternary structure

Figure 2.5: Examples of protein’s (a) primary structure, (b) secondary structure (α-helical Ac-Ala$_{15}$-LysH^{+}), (c) tertiary structure (human hemoglobine α-chain), and (d) quaternary structure (human hemoglobine).

2.2.1 The fundamental interactions that shape proteins

The primary structure of peptides and proteins is thus defined as a covalently bonded chain - i.e., a generally strong chemical bond. Electrostatic interactions are always present as well, in qualitatively different forms. For example, in the presence of charged residues (ions or polar side chains), ionic bonds take place [3], or if there are permanent dipoles present, there will be dipole-dipole and dipole-induced dipole interactions. However, the actual three-dimensional structure of the protein is also a product of much more subtle qualitatively different “interaction” types [50, 51], defined as:

- van der Waals bond (London dispersion force [52]): weak bond between instantaneous dipoles in atoms or parts of the molecules.
- Hydrogen bonds: medium-strength bond between a hydrogen and two other species, in such a way that it forms a covalent bond with one of them and an electrostatic bond with the other (AH$\cdots$B where A is considered an electron donor and B an electron acceptor).

[3] The “salt-bridge”, for example, is the ionic interaction between a positively charged NH_3^+ group in one amino acid with a negatively charged COO^- group in another.

The van der Waals interactions[4] (vdW) are now commonly used to describe a type of “so-called” weak interactions, arising from instantaneous induced dipoles, that are ubiquitous in nature. Some textbooks refer to van der Waals forces as all intramolecular forces (including repulsion and electrostatic forces), but this is not what is meant through this thesis. The nomenclature used here refers mainly to the contributions arising from dispersion (also called London [52] forces). A classical example of this type of van der Waals bond is that between two rare-gas atoms. Details about its (complex) nature will be further discussed in Chapter 3. The characterization of its importance in determining the structure of folding motifs will be a central piece in this thesis, since it is presently established that these interactions are of very important for protein structure[50].

The hydrogen bonds (H-bonds), extensively characterized in the book of L. Pauling [51], are also very relevant. A quote from the book says (p. 450): “It has been recognized that hydrogen bonds restrain protein molecules to their native configuration, and I believe that (...) it will be found that the significance of the hydrogen bond for physiology is greater than that of any other single structure feature”. The nature of the H-bond has been recently discussed in a review article (Ref. [53]), in which it is argued that this interaction arises from a complex balance of electrostatic, charge transfer and dispersion (van der Waals) interactions. In amino-acids, the H-bond appears usually between the CO and NH groups of different residues. These bonds are also known to be subject to a “cooperativity” effect, when a chain of interlinked H-bonds is formed[54–60]. This effect causes the strength of the H-bonds within an H-bond chain to increase due to the cooperative interaction between themselves. In essence, the more H-bonds are formed, the stronger they are, up to a certain limit, that can be calculated for infinite systems. This effect strongly affects the stability of complexes that exhibit a series of hydrogen-bonds (e.g., polypeptide helices, water). The nature of this effect cannot be explained by a sum of pairwise electrostatic interactions between each H-bond unit [57, 58], being generated by non-additive interactions arriving from polarization and induced polarization (polarization due to the polarization) of the electron clouds [61], and thus being a many-body effect.

Additionally, steric hindrance constraints, due to the side-chains and shape of the amino-acids forming the protein, also influence the protein structure. Furthermore, in a biological environment, depending on the polar or apolar nature of the solvent where the protein is located, hydrophobic and hydrophilic interactions will affect the final 3D structure.

Although the covalent bonds are essential for the final shape of polypeptides and proteins, the overall structure of the folding motifs would not exist as we know if not for the *non-covalent* interactions (electrostatics, H-bonds, and vdW) [62]. This is why it is very important to describe such interactions (as well as the peptide bond) very accurately and if possible, on the same footing.

2.2.2 Secondary structure main motifs

Identifying and understanding secondary structure motifs will be a central subject in this thesis. Therefore, a detailed classification of various types of known motifs follow.

There are three main elements of secondary structure, namely, helices, pleated-sheets, and turns, all of which are shown in Figure 2.6, where ribbons are drawn through the backbone atoms to make the secondary structure pattern clearer. The turns are regarded as non-periodic motifs, while helices and sheets are regarded as periodic, in the sense that a repeating unit can be defined, allowing for a characterization based on pairs of torsional angles (like the ϕ and ψ defined above, and represented in Figure 2.4).

[4] Named after the Dutch physicist and Nobel laureate Johannes Diderik van der Waals.

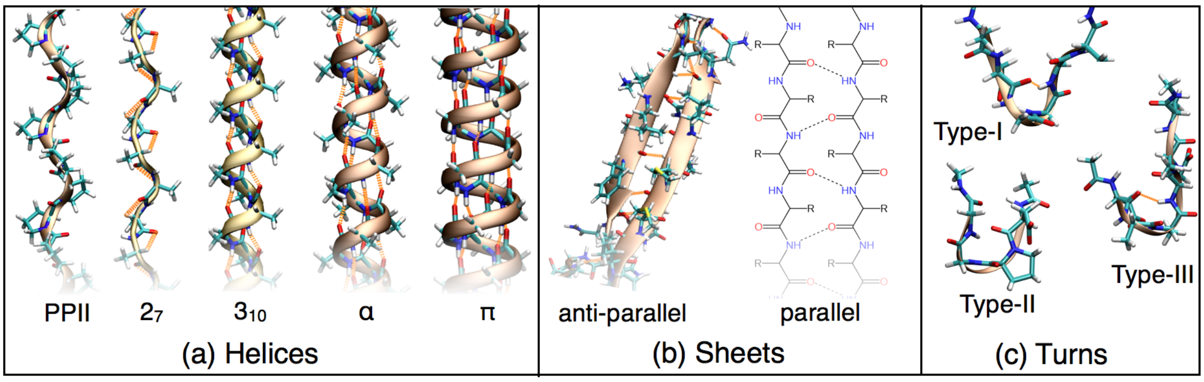


Figure 2.6: Schematic examples of different types of secondary structures: (a) PPII, 2_7, 3_{10}, α, and π helices; (b) parallel and anti-parallel β-sheets; and (c) β-turns of types I, II, and III.

The helices can assume several configurations, which are essentially characterized by the H-bond pattern they form between the respective residues. The L-amino acids prefer to form right-handed helices due to steric clashes between side chain and backbone atoms. The most famous is the α-helix, proposed by Pauling in 1951 [46] and schematically represented in Figure 2.6(a). It is characterized by H-bonds between residues i and residues $i + 4$. In order to form this bond, the residues adopt (ϕ,ψ) angles around (-60^o,-45^o), so that there are 3.6 residues per turn and 13 atoms separating the oxygen and the hydrogen involved in the H-bond, if counting only backbone ones. This type of helix is the most common helical motif found in known protein structures [40]. The so called 3_{10} helix has H-bonds forming between residues i and $i + 3$, and (ϕ,ψ) angles around (-50^o,-25^o). It has 3.0 residues per loop and is less compact than the α-helix. The subscript 10 relates to the fact that there are 10 atoms separating the ones forming the H-bond. This helix is also observed in proteins, usually at the ends of α-helices. The π-helix is formed by H-bonds between residues i and $i+5$, and although it is hard to find a perfect π-helix in nature, its structure would ideally have (ϕ,ψ) angles around (-55^o,-70^o), and 4.4 residues per turn, being more compact than the α-helix. In principle, there is also a very extended helix that is called 2_7. This structure presents very weak hydrogen bonds between the residues i and $i + 2$, and the subscript 7 also relates to the number of backbone atoms separating the ones forming the H-bond. It is characterized by (ϕ,ψ) angles around (-78^o, 59^o), with 2.2 residues per turn. The π and 2_7 helices are rarely observed. In addition, proline amino acids often form a helical conformation without H-bonds. This helix is called the PPII helix, and has (ϕ, ψ) angles of (-75^o,150^o). All these helices' types are shown in Figure 2.6(a).

The pleated sheets, also known as β-sheets, were also predicted by Pauling in 1951 [47]. They are formed by two extended polypeptide strands that form H-bonds between them. The preferred (ϕ,ψ) angles for these structures lie around (-135$\pm 15^o$, 135$\pm 15^o$). When they are oriented in the same direction with respect to the C and N-terminus the sheet is called *parallel*. When they are oriented in opposite directions it is called *anti-parallel*. These examples are shown in Figure 2.6(b).

Turns are necessary motifs to reverse the propagation of sheets and helices, so that compact structures can be formed. Turns are defined by having two C_α carbons separated by a few peptide bonds (more than 2) that come in close contact [63, 64], causing a change of orientation in the peptide chain. It is not necessary for a H-bond to form in order for the motif to be characterized as a turn, but many do form through the formation of hydrogen bonds. The most common type is known as the β-turn, which causes a 180^o change in the propagation direction, and has been characterized by Venkatachalam in 1968 [65]. This turn is characterized by a H-bond forming between residues $i + 3 \rightarrow i$. It can be further

Table 2.1: Idealized backbone torsion angles of the β-turn types.[64, 66]

Turn	ϕ_{i+1}	ψ_{i+1}	ω	ϕ_{i+2}	ψ_{i+2}
βI	-60	-30	180	-90	0
βI'	60	30	180	90	0
βII	-60	120	180	80	0
βII'	60	-120	180	-80	0
βIII	-60	-30	180	-60	-30
βIII'	60	30	180	60	30
βVIII	-60	-30	180	-120	120
βVIa1	-60	120	0	-90	0
βVIa2	-120	120	0	-60	0
βVIb	-120 (-135)	120 (135)	0	-60 (-75)	150 (160)

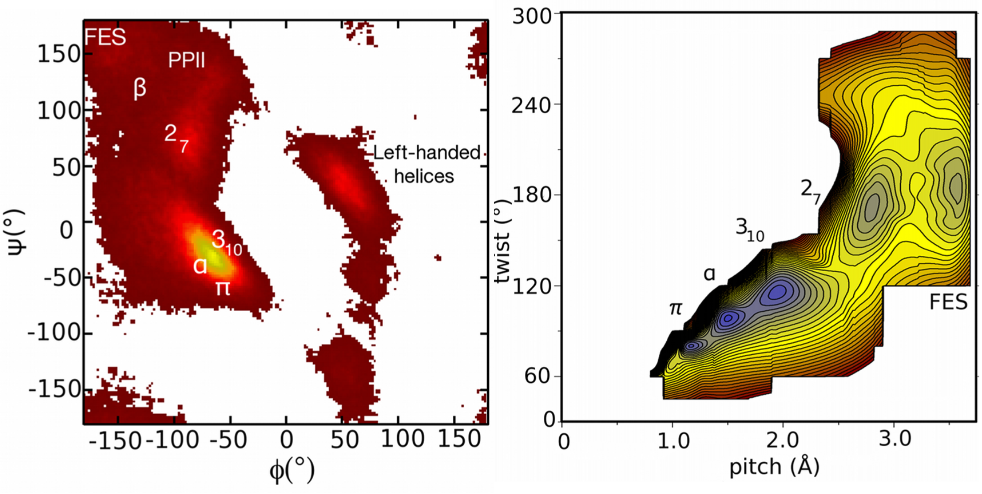
divided in subcategories given by the ϕ and ψ angles' range assumed by residues $i+1$ and $i+2$. The classification proposed by Hutchinson and Thornton [66] in eight well-defined subtypes detailed in Table 2.1. Of these subtypes, the most common are the type I and type II turn and their inverse counterparts I' and II'. Types VIa and VIb are "special", in the sense that they exhibit a rare *cis* peptide bond between the central residues. A few examples of β-turns of types I, II, and III can be seen in Figure 2.6(c). Other types of turn include the γ-turn, involving residues $i+2 \rightarrow i$, the α-turn (residues $i+4 \rightarrow i$), and the π-turns (residues $i+5 \rightarrow i$).

A special case of turn is the structure called "hairpin", in which, while the direction of propagation is reversed, the adjacent secondary structure elements interact. For example, the β-hairpin involves two β-sheets oriented in an antiparallel arrangement.

2.2.3 Characterization of secondary structure motifs

The ϕ and ψ backbone angles defined in Figure 2.4 can be used to unambiguously define the backbone structure of a protein. The famous Ramachandran diagram [67, 68] is drawn with respect to ϕ versus ψ angles for each residue in a protein. To the approximation of ideal covalent geometry, this plot gives a concise and intuitive information on the backbone conformation. Therefore, it has been widely used to characterize protein structure. An example of a Ramachandran plot for various conformations of each alanine residue in a polyalanine peptide, taken from simulations performed in this work, is shown in Figure 2.7(a). The regions visited by the amino acid in this plot are observed to be quite similar for all amino acids configurations in known proteins, except proline and glycine, as was observed by Lovell *et al.* [68]. The white areas in the Ramachandran plot of Figure 2.7(a) are "forbidden" regions. In the original publication of Ramachandran in 1963, the forbidden regions were understood in terms of backbone atom's hard-sphere repulsions and steric clashes only. This would render a plot with more extended forbidden regions, although the main features [i.e. most likely (ϕ, ψ) configurations] were correctly described. This has been revised in Refs. [68, 69], and it was found that a few steric clashes of the original hard-sphere model can be ignored (although most remain and are indeed the most determining factor for the strictly forbidden regions), and that backbone H-bonding influences the shape of the regions.

Secondary structure motifs, like helices and sheets, can be characterized in the Ramachandran plot, since they are composed by a sequence of amino acids all having the same (or approximately the same) backbone torsional angles. The α-helix, as mentioned above, presents (ϕ,ψ) angles around (-60^o,-45^o), for example. Specifically for the characterization of helices, however, other coordinates than the (ϕ,ψ)



(a) Ramachandran plot (Φ, ψ)

(b) Pitch-twist plot (L, θ)

Figure 2.7: (a) Example of a Ramachandran plot (ϕ, ψ) with secondary structure types labeled. α, 3_{10}, 2_7, π, and PPII correspond to different helical types, and FES corresponds to the "fully extended structure", with (ϕ, ψ) angles around (-180^o, 180^o). Figure (a) is a frequency probability plot, obtained from *ab initio* molecular dynamics simulations of alanine-based polypeptides. Yellow regions correspond to high probabilities and white regions correspond to forbidden regions. (b) Example of pitch-twist (L, θ) plot with different helical secondary structure types labeled. This figure was provided by J. Ireta, and appears in Ref. [70]. Figure (b) is an actual potential energy surface obtained for polyalanine helices. Purple/blue regions correspond to minima.

coordinates prove to be more useful. When plotting the (ϕ,ψ) angles corresponding to helices, they fall in regions of the Ramachandran plot that are very close to each other, without the appearance of noticeable barriers between them, since all values in that region are accessible to all configurations (see Figure 2.7(a), where the labels α, 3_{10}, and π appear). One set of coordinates that can differentiate better between different types of helices are the cylindrical coordinates (L, θ), where L is called the helix pitch, i.e., the increment per residue along the helix' axis, and θ is the helix twist, i.e. the angle a helix turns between one residue and the next [55]. These coordinates are schematically drawn in Figure 2.8. The regions corresponding to each kind of helix in these coordinates are drawn in Figure 2.7(b).

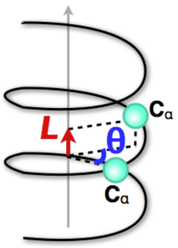


Figure 2.8: Schematic representation of a helix and two amino acid's C$_\alpha$ carbons, with the L (pitch) and θ (twist) coordinates in evidence.

Other types of coordinates, e.g. distances between atoms, have been used in the literature in order to characterize peptide structures. The end-to-end distances (distance from an atom in one terminus to another on the other terminus, usually the N-terminal nitrogen to the C-terminal oxygen), for example [71, 72], are frequently used to characterize groups of conformations. This coordinate does not define the structure of every single amino acid, though.

In this work, the specific H-bond connection, as well as the number of H-bonds in one peptide will be used to characterize and group structures. The idea is to define to exactly which NH group each CO group of the molecule is bonded. Using the H-bond connection as a structure characterizer is also often seen in the literature [31, 73]. The classification used here is such that a H-bond is considered to form whenever the O$\cdots$H distance between a (C)O and a (N)H atom is smaller than or equal to 2.5 Å . Typical H-bond distances lie around 2 Å [74], so that the value chosen here considers also "stretched" H-bonds. By not taking into account the angle between the atoms involved in the H-bond, and using this upper limit for the distance, this classification can give false positives, in the sense that it can measure a H-bond where it actually does not exist. More important for this work, though, is that this classification does not give any false negatives (i.e., no H-bond is identified where one is actually present), as will become clear in Chapters 7 through 10. All possible H-bonds of the polypeptides are considered for this classification, and therefore polypeptide conformers having the same H-bond pattern can differ by slight bends of the backbone atoms.

Here, the conformations of the polypeptides will also be characterized by their qualitative character, by the terms helical, non-helical, or globular and compact. The definition is the following: the term "helical" is used to designate structures where the C-terminus and the N-terminus are not interconnected and are separated by a well-defined (and reasonably repeating) pattern of H-bonds (see Figure 2.5(b) or Figure 2.6(a), for example.). Conformers that have bonds joining the two terminations but still present a helical loop (this can happen for small molecules) will be often labeled "not pure" or "not simple" helix. Finally, the conformers that have the two terminations joined by a H-bond and present no repeating pattern or clear loop will be called globular or compact.

2.3 Potential energy/free energy landscapes

The definition of a potential-energy landscape [13] is that of the multi-dimensional surface that describes how the energy of a system changes with geometry. More specifically, the "potential-energy surface" (PES) is the energy function $E(\vec{R}^N)$, where $\vec{R}^N$ are the atomic positions of N atoms. It is worth noting already here that this first definition is essentially classical, since, for this definition to work, the nuclear coordinates do not have any uncertainty associated with them. For the purposes of this thesis, a quantum-mechanically more rigorous derivation, in the context of the Born-Oppenheimer approximation, will be introduced in Chapter 3. For the present discussion, the existence of $E(\vec{R}^N)$ is taken for granted.

If the system is in a local or global minimum of the PES, any small displacement in any direction will increase the potential energy, i.e. the derivative of $E(\vec{R}^N)$ with respect to $\vec{R}^N$ is zero, and the second derivative is positive for displacements in any direction. There can be many local minima in the energy surface of a complex system like a polypeptide or protein. The minimum with the lowest potential energy is called the global minimum of the PES.

To describe the correct thermodynamics of a real system, though, the energy surface that has to be considered is the free energy surface, that is a function of of standard thermodynamic variables (temperature, pressure, entropy, volume, etc.) or some other, generic variables (order parameters, in a sense), that characterize the state of the system. For instance, these could be the pitch-twist or ϕ-ψ coordinates, but conceivably also a more generic measure of "helicity", H-bond network, etc. Due to the necessarily statistical, averaging nature of experiments, the free energy is the quantity that governs the behavior of the real system in experiment. Moreover, this is the surface explored by proteins at biological conditions, and therefore the one that really rules folding and structure formation.

The Helmholtz free energy, for example, obtained as a function of the temperature and the volume of the system, can be written as:

$$F(T, V) = U - TS, \tag{2.1}$$

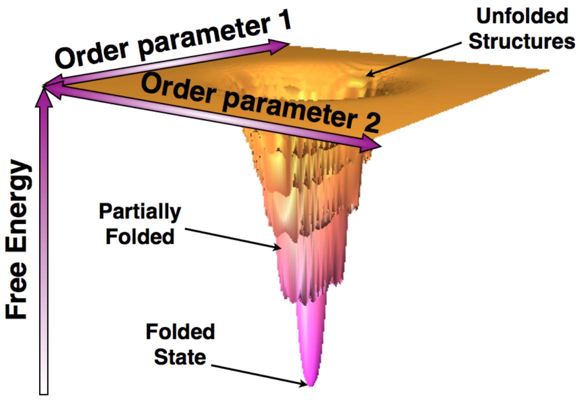


Figure 2.9: (Very) schematic example of a funnel-shaped folding free-energy landscape, as a function of two generic order parameters. Energy landscapes of different proteins differ vastly, and could have many other additional local minima.

where U is the internal energy, T the temperature and S the entropy. The internal energy U is the ensemble average of the energies intrinsic to the system, as ,e.g., the potential plus the kinetic energy, in a classical picture. The degrees of freedom of the system are thus averaged out in the thermodynamic variables. The Gibbs free energy:

$$G(T,p) = U - TS + pV, \tag{2.2}$$

is a function of temperature and pressure, and is commonly computed in order to compare to experimental data obtained under certain values of these two variables. Free energy landscapes can be constructed, for example, by projecting all coordinates of the system into one or two “order-parameters” that describe a particular feature and averaging over all the rest, for certain values of the variables. It thus connects the microstates of the system to thermodynamic meaningful macrostates.

In this thesis, the Helmholtz free energy will be used, since the experiments involving polypeptides in the gas-phase (described further below in Chapter 6), are done at essentially zero pressure. In any case, when comparing energy differences between different conformers of the same polypeptide (as will be done in this work), the pV term would cancel making $\Delta G = \Delta F$.

How a protein navigates its landscape and folds to its native state is a very active area of research nowadays, which makes large use of computers (see the Folding@home [75] project, for example). The question of how a protein folds has been posed more than 40 years ago. In 1968-1969 C. Levinthal proposed a paradox [9, 10]: how could proteins fold to the native structure so fast (less than seconds) in living organisms if even a short protein of 100 amino acids would have so many different states (degrees of freedom), that searching randomly through all the configurations would take an enormous amount of time [5] [76]? Levinthal himself proposed a solution to this paradox, in which proteins would have a well defined pathway to go to the folded state, making it a kinetically favorable process. A few years later, in 1973, C. Anfinsen [8] proposed the famous hypothesis “*that the native conformation is determined by the totality of inter-atomic interactions and hence the amino acid sequence, in a given environment*”. This idea led to the proposition of funnels in the energy landscapes, connected to specific interactions between the amino acids that would lead the protein faster to the native structure [11]. Today the comprehension of this problem has grown and it is generally accepted that [10, 12, 13]: (i) Proteins fold following a not-so-smooth energy funnel; (ii) proteins can fold by way of many different paths in a given environment, but (iii) there seem to be intermediate configurations in the process of folding that are visited and are as important as the final native structure for the protein function [77]. Knowing

[5] In Ref. [76], an estimation for this time is given. Considering three possible states for each bond in a protein of ≈100 amino acids, the number of possible configurations is 3^{100}. Even if the protein could be extremely fast and sample 10^{13} configurations per second, it would take 10^{27} years!

exactly why and how a protein folds and samples its landscape is not only an academic exercise. It is of enormous importance for understanding and perhaps curing diseases that are caused by proteins that fold "incorrectly" (misfold). Genetic diseases like Parkinson's, Alzheimer's, and Huntington's, for example, are connected to the aggregation of misfolded proteins [78, 79]. Also contagious diseases like the mad-cow (BSE) and Creutzfeldt-Jakob disease happen through the so-called *prion* protein PrP^{Sc} [80], which is a misfolded form of PrP^{C}, and is able to "infect" other proteins, causing new misfoldings (amyloid formation) [80, 81].

It is now generally accepted [12] that the unfolded state of the protein occupies a part of the free energy landscape defined by a relatively high internal energy, with respect to the folded state. Thus, in order to minimize the free energy, the unfolded state must have a high overall entropy (encompassing configurational, vibrational, and any other hypothetical source of entropy). The unfolded state is often modeled by a *random-coil* statistical ensemble, with a gaussian distribution of end-to-end distances [71, 82]. The random-coil idea implies that there are no strongly preferred backbone conformations in that state, such that energy differences among different backbone conformations should be small, of the order of $k_B T$. In the folded state, on the other hand, the proteins have lower energy, but also lower configurational entropy (the folded state is defined by specific constraints on the structure, thus occupying a much smaller volume of phase space overall) [6]. The relation between the internal energy and entropy mentioned above leads to a rough funnel in the free energy landscape that can have many intermediate minima, as shown schematically in Figure 2.9. A major thermodynamic factor that opposes the folding process is, therefore, the loss of *configurational* (or backbone) entropy [50]. On the other hand, it will be seen in this thesis that *vibrational* entropy (i.e. the vibrations of the molecule within one specific free-energy basin) actually stabilizes secondary structure motifs like helices. The first first-principles evidence for this stabilization mechanism will be presented in Chapter 8. Thinking about a vibrational free energy landscape, this implies that the minima lying at higher values of free energy have narrow funnels, while the native structure has a lower energy and broader funnel.

The picture discussed above is further complicated by the environment the protein is in. The various interactions with the environment (covalent, H-bonds, vdW, electrostatics), as well as the free energy of the solvent (or membrane, lipid, etc.) have to be taken into account. The structure that a protein adopts under certain physiological conditions will, thus, depend on a balance between all the intra- and inter-molecular factors mentioned so far. [7]

In all these considerations regarding the free energy landscape, it is important to remember that the quantity that free energies, entropy terms, etc. are ultimately derived from (via statistics and, more rigorously, the partition function) is still the potential energy surface of all individual micro-states. While it is still a long way from the microscopic potential energy surface to the correct thermodynamic behavior of a real system [8], an accurate potential-energy surface remains the essential ingredient to any meaningful computational predictions of larger-scale behavior from first principles. This thesis explores and aims to extend the reach of first-principles potential-energy surfaces for peptides and proteins, by means of computer simulations based on electronic structure methods.

[6] Temperature plays a role in this picture, favoring the unfolded state, the higher it gets. To make matters more complicated, proteins also unfold at very low temperatures [83, 84] (even below the freezing point of water).

[7] It seems to be quite a delicate balance, as Dobson and coworkers, for example, have observed that inducing only a small structural change can lead the whole protein to misfold, even if the environment itself (solvent, temperature, pH, etc.) is not changed [79].

[8] A comprehensive account on how to sample and characterize energy landscapes of biomolecules is given in the book of D. Wales [85].

Chapter 3

Theoretical methods to describe the potential-energy surface

This chapter focuses on explaining the theoretical methods used in this thesis to describe the potential energy surface (PES). It starts with the description of an empirical method, namely, an empirical classical Hamiltonian referred to as "force field". Then, the focus is moved to the description of "first-principles" electronic structure methods. The term "first-principles" (or *ab initio*) relates to the fact that such methods solve the electronic many-body problem based solely on the laws of quantum mechanics, including approximations, but without resorting to empirical parameters. From a practical point of view, they have the advantage to give quantitative accuracy and predictive power in a wide range of problems (transferability) and the disadvantage to be computationally significantly more expensive than empirical methods. For the interested reader, more details about the methods discussed here and many others can be found in the textbooks, as, e.g., *Modern Quantum Chemistry* by A. Szabo and N. Ostlund [86], *Density Functional Theory* by E. K. U. Gross and R. Dreizler [87], and others [88–90].

3.1 Force fields

The qualitative insights about the interactions shaping proteins, that was given in the previous chapter, can be used to design an empirical approximation for the estimation of the potential energy surface, namely, the "force fields" commonly used in molecular modeling. Force fields consist of empirical potentials that are divided in different terms corresponding to qualitatively different interactions, each of which are based on several parameters. The parameters can be obtained from experiments or from quantum mechanical calculations. As a more detailed example of such a force field, the functional form of the (popular) OPLS-AA (Optimized Potentials for Liquid Simulations - All Atoms[91]) force field, proposed by Jorgensen and coworkers, is presented below. This force field will be used in this thesis.

In OPLS-AA, the calculated total energy is divided into several contributions:

$$E^{OPLS-AA}(\vec{R}^N) = E_{bond} + E_{angle} + E_{torsion} + E_{nb} \tag{3.1}$$

The first three terms in Eq. 3.1 are related to "bonded" interactions between atoms, and have the following analytical form:

$$\begin{aligned} E_{bond} &= \sum_{bonds} K_r (R - R_0)^2 && (3.2) \\ E_{angle} &= \sum_{angles} K_\theta (\theta - \theta_0)^2 && (3.3) \\ E_{torsion} &= \sum_i \frac{V_1^i}{2}[1+\cos(\phi_i)] + \frac{V_2^i}{2}[1+\cos(2\phi_i)] + \frac{V_3^i}{2}[1+\cos(3\phi_i)], && (3.4) \end{aligned}$$

where R corresponds to the distance between pairs of atoms, θ corresponds to the angle between three atoms, ϕ corresponds to the dihedral angle (or torsion angle) between four atoms, and the subscript 0 refers to the equilibrium quantities. The sums run over all (covalent-)bonds, all angles and all dihedral angles respectively for each term. The constants K_r, K_θ, are empirical parameters which were taken from a previous (Amber) force field [91]. V_1^i, V_2^i, V_3^i have been fitted to first-principles data [30].

The remaining term E_{nb} corresponds to the non-bonded interactions between atoms, and is given by the sum of Coulomb and Lennard-Jones (LJ) pairwise contributions:

$$E_{nb} = \sum_{i,j;i<j} \left[\frac{q_i q_j e^2}{R_{ij}} + 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{R_{ij}^{12}} - \frac{\sigma_{ij}^{6}}{R_{ij}^{6}} \right) \right] f_{ij} \qquad (3.5)$$

The charges q_i and q_j, and LJ parameters ϵ_{ij} and σ_{ij} are empirically fitted to reproduce properties of organic liquids [91]. The parameter f_{ij} is 0 for $i - j$ pairs that are separated by 1 or 2 bonds, 0.5 for atoms separated by three bonds, and 1.0 for all other cases. In order to have a better description of atoms that are differently hybridized in a molecule, usually in a force field there are many more “types” of atoms than the existing number of chemical elements (e.g. carbon can be C, CH, CH3, etc.). Each one of them has its own set of fitted parameters.

Other popular force fields are the Amber [92] force fields, from Kollmann and coworkers, and the CHARMM [93] force fields, from Karplus and coworkers. More recently, “polarizable” force fields have been proposed, for example the AMOEBA [94] force field, by Ponder and coworkers. Polarizable force fields include the contribution coming from induced dipoles at the atomic positions due to the presence of the charges sitting on each atom of a molecule, representing a step forward over classical force fields.

Force fields do not treat explicitly the electrons of the system, whereas the first-principles methods (which will be discussed in the next sections) do. Nevertheless, they are clearly useful to calculate statistical averages and sample the (large) conformational space of biomolecules, especially where first-principles methods become unfeasible, due to the heavy computational cost and time demand. Although force fields provide the advantage of allowing to simulate full proteins, solvent, etc., they are not strictly transferable. They often perform well for the molecules and configurations they were trained on, but the reliability of their predictions for new molecules or geometries cannot be directly assessed, and different force fields give different answers for the same problem [32, 95]. It is possible to improve them by taking data from high level *ab initio* calculations, to which they can be benchmarked and from which more accurate parameters for specific interactions can be taken [30, 96]. In order to make a quantitative and reliable prediction with quantum-mechanical (QM) accuracy, though, the use of first-principles QM methods is necessary.

3.2 The many-body Hamiltonian in quantum mechanics

A polyatomic system can be described by a many-body Hamiltonian. This Hamiltonian, in its non-relativistic form[1] can be written as,

$$\begin{aligned}\hat{H} &= -\frac{\hbar^2}{2}\sum_{I=1}^{M}\frac{\nabla_I^2}{M_I} - \frac{\hbar^2}{2}\sum_{i=1}^{N}\frac{\nabla_i^2}{m_i} - \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\sum_{I=1}^{M}\frac{Z_I e}{|\vec{r}_i - \vec{R}_I|} + \\ &\quad + \frac{1}{4\pi\epsilon_0}\sum_{i=1}^{N}\sum_{j<i}^{N}\frac{e^2}{|\vec{r}_i - \vec{r}_j|} + \frac{1}{4\pi\epsilon_0}\sum_{I=1}^{M}\sum_{J<I}^{M}\frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|} \qquad (3.6)\end{aligned}$$

where i, j are indexes for the N electrons, I, J are indexes for the M nuclei, Z and e are the charges of nuclei and electron respectively, M and m the masses of the nuclei and electron respectively, ϵ_0 the vacuum dielectric constant, and $\hbar$ the reduced Planck constant. In the rest of this thesis (except where specified) the natural units $m = e = 1/4\pi\epsilon_0 = \hbar = 1$ will be used. In short, this Hamiltonian is re-written in the following form:

$$\hat{H} = \hat{T}_{nuc} + \hat{T}_e + \hat{V}_{nuc-e} + \hat{V}_{e-e} + \hat{V}_{nuc-nuc} \qquad (3.7)$$

where $\hat{T}_{nuc}$ and $\hat{T}_e$ are the kinetic-energy operators related to the nuclei and the electrons respectively, and $\hat{V}_{nuc-e}$, $\hat{V}_{e-e}$, and $\hat{V}_{nuc-nuc}$ the terms related to the electrostatic interaction between two electrons, a electron and a nucleus, and two nuclei[2] respectively. These terms are precisely the ones written explicitly in Eq. 3.6, in the same order.

Solving this hamiltonian in a (non-relativistic, time-independent) quantum-mechanical framework means to solve the time-independent Schrödinger equation[3]:

$$\hat{H}\psi(\vec{r}, \vec{R}) = E\psi(\vec{r}, \vec{R}) \qquad (3.8)$$

where E is the total energy of the system and ψ is the many-body wave function. Since each electron and each nucleus can move in the x, y and z coordinates, solving this equation involves a problem of $3N + 3M$ degrees of freedom, in which all particles are coupled. In order to turn such a problem into a feasible enterprise, approximations need to be made. The first approximation taken is usually to decouple the movement of the electron and nuclei. This is known as the Born-Oppenheimer (or adiabatic) approximation, which will be discussed in more detail in the next section.

3.3 The Born-Oppenheimer approximation

This thesis is (in parts) also on nuclear motion, which will be explained in Chapter 4. Therefore, in this section, the derivation of the BO approximation will be given, which clarifies how it is possible to define the movement of the nuclei (decoupled from the electrons). The guiding idea of the Born-Oppenheimer (BO) approximation [97] is that the ratio between electron mass and nucleus mass (m/M) is very small,

[1] Here also the electron spin dependency has been dropped for simplicity.

[2] Here the frozen nuclei expression, valid when the nuclei are distant enough so that their (positive) charge densities do not overlap.

[3] For the relativistic case, the equation to be solved would be the Dirac equation.

so that the nuclei do not follow the fast movement of the electrons. The derivation also briefly illuminates the limits of this approximation, which can be very relevant in biological processes involving peptides.

Mathematically one can separate the Hamiltonian of Equation 3.7 into an "electronic-only" part $\hat{H}_e$ consisting of

$$\hat{H}_e = \hat{T}_e + \hat{V}_{nuc-e} + \hat{V}_{e-e}, \tag{3.9}$$

$$\hat{H} = \hat{H}_e + \hat{T}_{nuc} + \hat{V}_{nuc-nuc}. \tag{3.10}$$

One can then solve the time-independent Schrödinger equation for $\hat{H}_e$ in order to use its eigenvectors as a basis to expand the eigenstates of the full Hamiltonian, as

$$\hat{H}_e \phi_\nu(\vec{r}, \vec{R}) = E^e_\nu(\vec{R}) \phi_\nu(\vec{r}, \vec{R}), \tag{3.11}$$

where $E^e_\nu(\vec{R})$ is the electronic energy for a given configuration of the nuclei and the ϕ_ν are assumed to be orthonormalized. The eigenstate ψ of the many-body hamiltonian can then be expanded as

$$\psi = \sum_\nu \lambda_\nu(\vec{R}) \phi_\nu(\vec{r}, \vec{R}). \tag{3.12}$$

If one writes ψ as 3.12 in Equation 3.8 and multiplies by $\langle\phi_\mu|$, the expression becomes:

$$\begin{aligned}
\langle\phi_\mu|\hat{H}|\sum_\nu \lambda_\nu\phi_\nu\rangle &= E\lambda_\mu \Rightarrow & (3.13)\\
E\lambda_\mu &= \left[E^e_\mu + \hat{T}_{nuc} + \hat{V}_{nuc-nuc}\right]\lambda_\mu + & (3.14)\\
&+ \sum_\nu \sum_I \frac{\hbar^2}{2M_I}\left[\langle\phi_\mu|\nabla^2_I|\phi_\nu\rangle\lambda_\nu + 2\langle\phi_\mu|\nabla_I|\phi_\nu\rangle\nabla_I\lambda_\nu\right]. & (3.15)
\end{aligned}$$

The off-diagonal elements of the two terms appearing in 3.15 are called non-adiabatic, referring to the fact that they involve the interaction between two different electronic states. The ones lying on the diagonal are called adiabatic[4]. Up to here, no approximations were introduced. However, if one could neglect the terms in 3.15, by defining $\hat{H}_{nuc} = \hat{T}_{nuc} + E^e_\mu + \hat{V}_{nuc-nuc}$ it would be possible to write:

$$\hat{H}_{nuc}\lambda_\mu = E\lambda_\mu. \tag{3.16}$$

For that to be possible, approximations need to be introduced. One approximation is that it is necessary to assume an adiabatic system (i.e. the atomic motion does not induce electronic excitations), so that:

$$\langle\phi_\mu|\nabla_I|\phi_\nu\rangle = \langle\phi_\mu|\nabla^2_I|\phi_\nu\rangle = 0 \text{ for } \mu \neq \nu. \tag{3.17}$$

The other approximation is that the diagonal elements $\langle\phi_\mu|\nabla^2_I|\phi_\mu\rangle$ are very small if compared to their electronic counterpart, meaning that:

$$|\langle\phi_\mu|\nabla^2_I|\phi_\mu\rangle| \leq |\langle\phi_\mu|\nabla^2_i|\phi_\mu\rangle| \tag{3.18}$$

[4] Actually, since $\langle\phi_\mu|\nabla_I|\phi_\nu\rangle$ is anti-symmetric, its diagonal elements ($\mu = \nu$) are always zero.

It is possible to show that this is an acceptable approximation by multiplying by $\hbar^2/2M_I$ each side of Equation 3.18, and the right side by $m_\mu/m_\mu = 1$, arriving at:

$$|\frac{\hbar^2}{2M_I}\langle\phi_\mu|\nabla_I^2|\phi_\mu\rangle| \leq |\frac{m_\mu}{M_I}\frac{\hbar^2}{2m_\mu}\langle\phi_\mu|\nabla_i^2|\phi_\mu\rangle|, \tag{3.19}$$

where, finally, knowing that m_μ/M_I is at least of the order of 10^{-4} (electron to proton mass ratio is 5×10^{-4}), the term in Equation 3.19 can also be neglected.

Making all these assumptions, the Born-Oppenheimer potential energy surface, where the nuclei move, is then defined for the electronic ground state μ=0 as

$$\hat{V}_{BO} = E_0^e + \hat{V}_{nuc-nuc}, \tag{3.20}$$

where E_0^e is given by Eq. 3.11. Knowing all the approximations involved, it is straightforward to see in which cases this approximation is not valid. If there is a non-adiabatic process going on (e.g. crossing between two electronic states) the approximation breaks down. Electron-phonon coupling, for example, can only be described when the non-adiabatic terms (Eq. 3.17) are taken into account. In the biological realm, photosynthesis is one (very important) example where the system is definitely non adiabatic. The Born-Oppenheimer (BO) potential provides, nevertheless, a good approximation for the for the potential energy surface explored by the nuclei in most situations. The electronic structure methods discussed in the following section all assume the BO approximation.

3.4 The Hartree-Fock method

Due to its conceptual importance, the first method discussed here used to solve the electronic Hamiltonian, defined in Eq. 3.9, is the Hartree-Fock method. In the electronic Hamiltonian, the most complicated term is the $\hat{V}_{e-e}$ term, which couples all electrons. The Hartree-Fock method in an improvement upon Hartree theory [98]. In Hartree theory, the *ansatz* wave function is a product of single particle orbitals: $\Psi_0^H(\vec{r}_N) = \phi_1(\vec{r}_1)\phi_2(\vec{r}_2)\cdots\phi_N(\vec{r}_N)$. In this way, the term $\hat{V}_{e-e}$ appearing in Equation 3.9 becomes separable and can be written as a sum of single particle contributions, in the following way:

$$\langle\Psi_0^H|\hat{V}_{e-e}|\Psi_0^H\rangle = \sum_{\substack{k,k'=1\\k\neq k'}}^{N}\int\frac{n(\vec{r_k})n(\vec{r_{k'}})}{|\vec{r_k}-\vec{r_{k'}}|}d^3r_k d^3r_{k'} \tag{3.21}$$

$$\text{where } n(\vec{r_k}) = \sum_{k=1}^{N}|\phi_k(\vec{r_k})|^2, \tag{3.22}$$

and $n(\vec{r}')$ is the electronic density. The quantity

$$v^H(\vec{r_k}) = \int\frac{n(\vec{r})}{|\vec{r_k}-\vec{r}|}d^3r \tag{3.23}$$

is called the Hartree potential.

One can interpret this approximation in the sense that one electron, at any given time, can be considered as moving in an average potential given by all the other electrons present in the system, in a classical mean-field sense. This greatly simplifies the problem at hand, because the particles are now decoupled and the problem reduces to solving a single particle Hamiltonian subject to an effective potential. The expectation value of $\hat{H}_e$ satisfies a variational principle, in the sense that it has to be bounded *below* by the exact energy (given the limitation of the wave function) in the Born-Oppenheimer surface. In the Hartree theory, the *ansatz* wave function ignores the Pauli principle. This omission is corrected by Hartree-Fock[99] theory. Here the wave function is explicitly anti-symmetric, by assuming the form of what is known as a *Slater determinant*, defined as

$$\Psi^{HF} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\vec{r}_1) & \phi_2(\vec{r}_1) & \phi_3(\vec{r}_1) & \cdots & \phi_N(\vec{r}_1) \\ \phi_1(\vec{r}_2) & \phi_2(\vec{r}_2) & \phi_3(\vec{r}_2) & \cdots & \phi_N(\vec{r}_2) \\ \phi_1(\vec{r}_3) & \phi_2(\vec{r}_3) & \phi_3(\vec{r}_3) & \cdots & \phi_N(\vec{r}_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_1(\vec{r}_N) & \phi_2(\vec{r}_N) & \phi_3(\vec{r}_N) & \cdots & \phi_N(\vec{r}_N) \end{vmatrix}, \tag{3.24}$$

where $\phi_i(\vec{r}_j)$ is the single-particle wave function of electron j for state i [5].

The idea is then to find the function Ψ_0^{HF} that minimizes the expression $\langle\Psi^{HF}|\hat{H}_e|\Psi^{HF}\rangle$ subject to the condition that $\langle\Psi^{HF}|\Psi^{HF}\rangle = 1$. The Hartree-Fock energy for the ground-state can then be written as

$$\begin{aligned} E^e_{HF} = \langle\Psi_0^{HF}|\hat{H}_e|\Psi_0^{HF}\rangle = \sum_{k=0}^{N}\left[\int \phi_k^*(\vec{r}_k)\left(-\frac{1}{2}\nabla_k^2 + \hat{V}_{nuc-e}\right)\phi_k(\vec{r}_k)d^3r_k\right] + \\ + \underbrace{\frac{1}{2}\sum_{k=0}^{N}\sum_{k'\neq k}^{N}\iint \phi_k^*(\vec{r}_k)\phi_{k'}^*(\vec{r}_{k'})\frac{1}{|\vec{r}_k-\vec{r}_{k'}|}\phi_k(\vec{r}_k)\phi_{k'}(\vec{r}_{k'})d^3r_kd^3r_{k'} -}_{E_{Hartree}} \\ \underbrace{-\frac{1}{2}\sum_{k=0}^{N}\sum_{k'\neq k}^{N}\iint \phi_k^*(\vec{r}_{k'})\phi_{k'}^*(\vec{r}_k)\frac{1}{|\vec{r}_k-\vec{r}_{k'}|}\phi_k(\vec{r}_k)\phi_{k'}(\vec{r}_{k'})d^3r_kd^3r_{k'}\,.}_{E_x} \end{aligned} \tag{3.25}$$

The term labeled E_x in 3.25 is called the Hartree-Fock *exchange* energy and the one labeled $E_{Hartree}$ is the called the Hartree energy. If $k = k'$ the exchange term cancels exactly with the Hartree term so that the spurious interaction of the electron with itself (self-interaction) is automatically removed. Although the spin of the electrons was not explicitly written, this exchange term only acts on electrons of the same spin. This means that the movement of electrons of same spin in the HF approximation is correlated, in the sense that the so-called “exchange-hole” is formed around the position of electron k. The expression of 3.25 allows to define the exchange potential for the electron k as

$$\kappa(\vec{r_k}) = \int \frac{\phi_k^*(\vec{r}_{k'})\wp_{\vec{r}_k,\vec{r}_{k'}}\phi_k(\vec{r}_{k'})}{|\vec{r}_k-\vec{r}_{k'}|}d^3r_{k'}, \tag{3.26}$$

where $\wp_{\vec{r}_k,\vec{r}_{k'}}$ is an operator that acts on $\phi_k(\vec{r}_{k'})$ to change $\vec{r}_{k'}$ to $\vec{r}_k$. What is known as the Fock operator for one electron is defined as

$$\hat{F}_k = \nabla_k^2 + \hat{V}_{nuc-e}(\vec{r_k}) + \upsilon^H(\vec{r_k}) - \kappa(\vec{r_k}), \tag{3.27}$$

and the Hartree-Fock single particle equations read

[5] The spin component has been again neglected here, but should not be forgotten. Since the component is factorizable $\Phi(\vec{r}_i\sigma) = \phi(\vec{r}_j)\chi(\sigma)$, it would yield twice as many components in 3.24 for a closed shell atom, one for each spin.

$$\hat{F}_i\phi_i = \epsilon_i\phi_i, \tag{3.28}$$

where ϵ_i are the Lagrange multipliers used to constrain the normalization of the orbitals, and minimize Eq. 3.25. It can be shown, by Koopmans' theorem [100], that the eigenvalues ϵ_i that correspond to occupied orbitals are equivalent to negative ionization potentials, as long as there is no relaxation of the orbitals upon removal of one electron (which is an approximation that is seldom valid). Commonly ϵ_i are just interpreted as orbital energies, but one must keep in mind that this interpretation would only be strictly true if electrons were really independent effective single particles.

The Hartree-Fock method has to be solved self-consistently, since the Hartree potential and the exchange term require as inputs the electronic orbitals, that are only known after solving the set of Eq. 3.28. Therefore, starting from an initial guess, one solves the Hamiltonian to get new orbitals, builds new potentials and solves again the equations. This process is repeated until the initial guess and the solution of the equations become the same, within a certain threshold.

Due to the integral appearing in Eq. 3.25 for the exchange term, that requires the explicit calculation of contributions of four different orbitals at a time, the scaling of this method is formally of N^4 (where N is a measure of the size of the system, e.g. electrons or basis functions), although with some manipulation it is possible to reduce the computation of the matrices to linear scaling, but only for very large systems[101].

3.5 Møller-Plesset perturbation theory

Although the exchange term presented above is a form of correlation between electrons (also called "Pauli correlation"), what is usually coined as "correlation energy" is everything that is missing from the Hartree-Fock energy:

$$E^e_{corr} = E^e - E^e_{HF}, \tag{3.29}$$

where E^e was defined in Eq. 3.11. The difference between E^e and E^e_{HF} is that the former is (theoretically) calculated with the true many-body wave function, while the later is calculated with the approximated HF wave function. A natural way to include such correlations is by adding a perturbation to the Hamiltonian. The derivation of the first and second order corrections for the energy in such a theory is straightforward, and can be found in the textbook of Szabo and Ostlund [86], for example. The quantum-chemistry method known as Møller Plesset second order perturbation theory (MP2) [102] is a particular case of many-body perturbation theory where the unperturbed Hamiltonian is taken to be the Hartree-Fock one:

$$\hat{H}_0 = \sum_i^N \hat{F}_i, \tag{3.30}$$

and the perturbation is given by the difference of the true many-body electronic Coulomb interaction and what is already included in Hartree-Fock:

$$\hat{H}' = \hat{H}_e - \sum_i^N \hat{F}_i = \sum_{i,j;i<j} \frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_i [\upsilon^H(\vec{r}_i) - \kappa(\vec{r}_i)] \tag{3.31}$$

For the Hartree-Fock Hamiltonian all single Slater-determinant wave functions that satisfy $\hat{H}_0|\Psi\rangle = E^e|\Psi\rangle$ can be calculated, where the HF label has been dropped for simplicity. These form a complete, orthonormal set of functions that can be used as a starting point for perturbation theory. The differences between the ground-state determinant $|\Psi_0\rangle$ and other possible solutions are interpreted as electronic excitations, since, they differ by interchanging one or more rows of the determinants. If the electrons could be seen as effective single particles, this interchanging could be understood as promoting one or more electrons from a occupied state to a non occupied one in the Hartree-Fock ground state. The nomenclature used here will be that $|\Psi_i^a\rangle$ corresponds to a single excited electronic state, $|\Psi_{ij}^{ab}\rangle$ to double excited electronic states and so on, where $i, j, k, ...$ denote occupied states and $a, b, c, ...$ unoccupied ones. These determinants are schematically drawn in Figure 3.1.

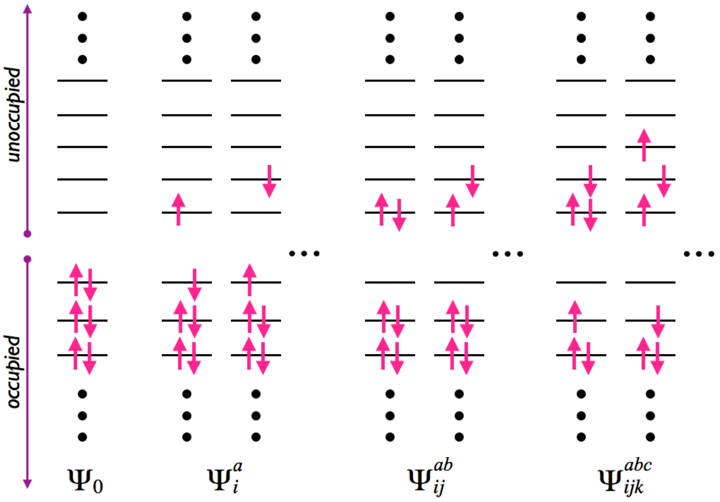


Figure 3.1: Schematic representation of a electronic ground-state and a few possible single, double and triple excited electronic states, assuming the interpretation of the HF orbital energies as actual single particle energy levels. The bars correspond to electronic energy levels and the arrows to electrons.

MP2 considers the time-independent perturbation expansion of the energy only up to second order. The first order correction to the energy in this basis, with the perturbation given by 3.31, yields the Hartree-Fock energy itself, making HF exact to first order. For the second-order perturbation term of the energy there would be matrix elements involving the ground state plus single excitations and double excitations, but not higher order excitations. The lack of contribution from higher order excitations is due to the fact that the perturbation $\hat{H}'$ is a two-particle operator and the orbitals are orthonormal, such that the result of the matrix element will be zero unless they are connected by a non-trivial operator. However, according to Brillouin's theorem [86], singly excited HF states also do not contribute. For the specific single-determinant wave function generated by Hartree-Fock, this means that singly-excited states do not interact directly with the HF ground state (they can, indirectly, through higher order perturbation terms), so that any matrix element involving these two orbitals is zero. The second-order energy correction in MP2 is thus given by:

$$\xi_n^{(2)} = \sum_{i\leq j}^{occ.} \sum_{a\leq b}^{unocc.} \frac{\langle \Psi_0|\hat{H}'|\Psi_{ij}^{ab}\rangle\langle\Psi_{ij}^{ab}|\hat{H}'|\Psi_0\rangle}{E_0^e - E_{ij}^{ab}} \tag{3.32}$$

where i, j are occupied orbitals, a, b are empty orbitals, E_0^e is the HF ground-state energy and E_{ij}^{ab} the energy corresponding to a particular doubly excited determinant. The matrix elements appearing in the numerator of 3.32 can be written as two-electron integrals over molecular orbitals ϕ, and the energy difference in the denominator can be written as a difference between molecular orbital energies, since the wave function is a Slater determinant. The expression for the MP2 total energy is written as:

$$E^{MP2} = E^{HF} + \frac{1}{4}\sum_{ijab} \frac{|\,(ij||ab)\,|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}, \tag{3.33}$$

where ϵ_i is the HF molecular orbital eigenvalue for state i and

$$(ij||ab) = (ij|ab) - (ij|ba) \text{ with} \tag{3.34}$$

$$(ij|ab) = \iint \frac{\phi_i^*(\vec{r})\phi_j^*(\vec{r}')\phi_a(\vec{r})\phi_b(\vec{r}')}{|\vec{r}-\vec{r}'|} d^3r d^3r'. \tag{3.35}$$

MP2 has a “modest” scaling (formally N^5) if compared, for example, with coupled cluster theory including single, double, and triple excitations (explained in the next Section). This is one of the main reasons for its popularity in the quantum-chemistry community, as a first approximation to include explicitly the many-body correlation effects. The truncation of the expansion on the second order in the perturbation series, though, neglects higher order energy terms that could (and do) also have contributions coming from double excitations. The correlation energy estimated by MP2 is also found to be usually over-estimated (too negative), due to the fact that higher order terms in the expansion would have contributions with a positive sign, which are not being accounted for. Finally, MP2 clearly diverges for metallic systems (there would be a division by zero in 3.33 due to the nonexistence of a gap), making it not suitable to be applied in metallic solids, for example.

3.6 Coupled cluster theory: the “gold-standard”

Formally, the correct solution to the many-body problem would be full configuration interaction (CI) [6]. This formalism, though, is extremely computationally demanding, and truncating the CI method to include excitations only up to a certain order causes a size-extensive problem, i.e. the energy in this method does not scale correctly (linearly) with the number of electrons. Coupled cluster theory fixes this problem, which is an advantage. Moreover, when considering single, double, and perturbative triple excitations (further explained below), this theory gives very accurate (often better than 1 kcal/mol or 43 meV) for ground-state properties of molecules [14]. It is sometimes referred to as the “gold-standard” of quantum chemistry, being the most accurate and computationally affordable method to solve the many-body problem[103].

Coupled cluster theory was initially proposed in 1960, in the context of nuclear physics [104]. The equations for electrons were proposed in 1966 [105] and have since become a very popular method in quantum chemistry [103]. In this thesis, this method was not directly used, but for benchmarking purposes, results were compared to coupled-cluster (CC) results. Therefore, a brief overview of this method is given below.

The wave function *ansatz* for CC theory in quantum-chemistry is written as:

$$\Psi_{cc} = e^{\hat{T}}\Psi_0 = (1 + \hat{T} + \frac{\hat{T}^2}{2!} + \frac{\hat{T}^3}{3!} + ...)\Psi_0, \tag{3.36}$$

where Ψ_0 is the HF ground-state slater determinant and $\hat{T}$ is an operator that can be expanded in order to “cluster” contributions coming from single, double, triple, etc. excitations, in the following way:

[6] The reader is referred to the textbook of Ref. [86] for details about this method and its derivation.

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + ... \tag{3.37}$$

$$\hat{T}_1\Psi_0 = \sum_i^{occ.} \sum_a^{unocc.} t_i^a \Psi_i^a \tag{3.38}$$

$$\hat{T}_2\Psi_0 = \sum_{ij}^{occ.} \sum_{ab}^{unocc.} t_{ij}^{ab} \Psi_{ij}^{ab} \tag{3.39}$$

$$\vdots \tag{3.40}$$

where t_i^a, t_{ij}^{ab} are the excitation amplitudes. In addition, the expansion of the exponential in powers of $\hat{T}$ given in 3.36 gives rise to cross terms and higher power terms, like $\hat{T}_1^2, \hat{T}_1\hat{T}_2$, etc., so that grouping the expansion by excitation order yields:

$$e^{\hat{T}} = 1 + \hat{T}_1 + (\hat{T}_2 + \frac{\hat{T}_1^2}{2}) + (\hat{T}_3 + \hat{T}_1\hat{T}_2 + \frac{\hat{T}_1^3}{6}) + \cdots , \tag{3.41}$$

where the first term on the right corresponds to the Hartree-Fock system, the second term produces all single excitations, the third all double excitation, etc. The coupled-cluster energy is then obtained by minimizing, as a function of the amplitudes, the following expression:

$$E^{CC} = \langle\Psi_0|e^{-\hat{T}}\hat{H}e^{\hat{T}}|\Psi_0\rangle. \tag{3.42}$$

The remaining problem is to which order the cluster operator $\hat{T}$ itself should be considered. The coupled cluster singles and doubles (CCSD) theory is the one obtained when $\hat{T}$ is written as $\hat{T}_1 + \hat{T}_2$, yielding all possible single and double excitations and their corresponding correlation (see eq. 3.41), plus disconnected contributions of higher orders ($\hat{T}_1\hat{T}_2$, $\hat{T}_2^2$, etc.). These "disconnected" terms render the theory size-extensive, which was not the case for MP2, for example (or truncated CI, briefly mentioned above). Both MP2 and CC (considering a finite number of excitations) are not variational, though, meaning that it is in principle possible to find a total energy for the electronic system that is lower than the true many-body energy. For detailed formulas of the energy and wave function of CCSD the reader is referred to Refs. [86, 103]. The scaling of CCSD is already of N^6 (N=size of the system, like number of electrons or basis-functions).

Coupled cluster singles, doubles and triples (CCSDT) theory would consider $\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3$, but that is already extremely computationally expensive. When the triple excitations are considered only in a perturbative manner, the method is called CCSD(T) and has a O(N^7) scaling.

3.7 Density-functional theory

Density-functional theory (DFT) is the electronic structure method most used through this work. The guiding principle of this theory is to reduce the "size" of the many-body problem by substituting the 3N dimensional wave function by the electronic density. The theory incorporates ideas from the work of L. Thomas and E. Fermi in the 1920's [106, 107], as well as the work of Slater [108],and many others. In 1964, with the work of P. Hohenberg and W. Kohn [109], density-functional theory was placed on a rigorous foundation. The theory is founded on two simple but fundamental theorems, that provide a rigorous proof that the all the observables of the system can be written as a function of the electronic density. These theorems state:

1. There is a one to one correspondence between the external potential $\upsilon_{ext}(\vec{r})$ (equivalent to $\hat{V}_{nuc-e}$) and the density of the ground-state $n_0(\vec{r})$: $\upsilon_{ext}(\vec{r}) = \upsilon_{ext}[n_0](\vec{r})$ and $n_0(\vec{r}) = n_0[\upsilon_{ext}](\vec{r})$, which means that the external potential is fully determined by the density as well as the other way around. This allows one to write any observables as a functional.

2. By defining the *universal* functional $F[n] = T[n] + E_{e-e}[n]$, where $T[n]$ is the kinetic-energy functional and $E_{e-e}[n]$ is the electron-electron interaction functional, and the total energy as $E[n] = \int \upsilon_{ext}(\vec{r})n(\vec{r})d\vec{r} + F[n]$, it is possible to show that the ground-state energy $E_0[n]$ is minimum for the exact (Born-Oppenheimer) ground-state density n_0: $E_0[n_0] \leq E_0[n]$.

The original proofs for these theorems are very simple and easy to follow, and are textbook material today (e.g., Ref. [87]). In addition, in 1979 M. Levy [110] showed a more elegant proof for the theorems which validates them also for degenerate ground-states. It also means that no matter how one gets the electronic density, it is theoretically possible to use it to (re-)construct the corresponding potential, under specific constrains [110–112].

These theorems, however, do not provide a practical way to solve the equations and obtain the densities, or how to build the external potential from them. The most popular scheme that provided a practical way to solve these equations is known as the Kohn-Sham scheme, proposed by W. Kohn and L. Sham in 1965 [113].

3.7.1 The Kohn-Sham equations

The idea of the Kohn-Sham scheme [113] is to map the system of interacting electrons into another fictitious one of *non-interacting* electrons, but that happens to have the exact same density $n(\vec{r})$ of the interacting system. One addresses the kinetic-energy operator by splitting it in two parts, one corresponding to a (yet unknown) system of non-interacting electrons $T_s[n]$ and another corresponding to the remaining part that accounts for the correlations $T_c[n]$:

$$T[n] = T_s[n] + T_c[n]. \tag{3.43}$$

This simplifies the problem greatly, as the kinetic-density operator of a non-interacting system can be written as the Laplacian of single particle orbitals. Similarly, the potential energy related to the interaction between electrons is also separated in the following way:

$$E_{e-e}[n] = E_H[n] + E_{xc}[n] \tag{3.44}$$

where $E_H[n]$ is the Hartree term $\int \upsilon_H(\vec{r})n(\vec{r})d\vec{r}$, with $\upsilon_H(\vec{r})$ given by 3.23, and $E_{xc}[n]$ corresponds to the exchange and correlation between the interacting electrons. In this way, the energy functional can be written as:

$$E_{KS}[n] \quad = \quad T_s[n] + E_H[n] + E_{ext}[n] + E_{xc}[n], . \tag{3.45}$$

$E_{xc}[n]$ is then called the exchange-correlation energy functional, and contains basically all quantum-mechanical many-body effects. Minimizing Eq. 3.45 with respect to the density, subject to the condition

$\int n(\vec{r})d^3r = N$ (where N is the number of electrons in the system), one obtains:

$$\frac{\delta}{\delta n(\vec{r})}\left(E_{KS}[n] - \mu\left[\int n(\vec{r})d^3r - N\right]\right) = 0 \Rightarrow \tag{3.46}$$

$$\Rightarrow \frac{\delta E_{KS}[n]}{\delta n} = \mu \Rightarrow \tag{3.47}$$

$$\Rightarrow \frac{\delta T_s[n]}{\delta n} + \upsilon_H(\vec{r}) + \upsilon_{ext} + \frac{\delta E_{xc}[n]}{\delta n} = \mu, \tag{3.48}$$

where μ is the Lagrange multiplier used to minimize the expression and the term given by:

$$\upsilon_{xc} = \frac{\delta E_{xc}[n]}{\delta n}, \tag{3.49}$$

is called the exchange-correlation potential. At this point, we can define an effective single-particle potential $\upsilon_{eff} = \upsilon_H + \upsilon_{xc} + \upsilon_{ext}$, so that Eq. 3.48 becomes:

$$\frac{\delta T_s[n]}{\delta n} + \upsilon_{eff} = \mu, \tag{3.50}$$

which describes a system of non-interacting particles moving in the the effective potential υ_{eff}. For such a system, we know how to define the kinetic-energy and the density (n_s), in terms of single-particle orbitals:

$$T_s[n] = \langle\phi_i|\nabla^2|\phi_i\rangle \tag{3.51}$$

$$n_s(\vec{r}) = \sum_{i=1}^{N}|\phi_i|^2. \tag{3.52}$$

Since the non-interacting and interacting system are equivalent, $n(\vec{r}) = n_s(\vec{r})$, so that in order to find the density of the *interacting* system, it is only necessary to solve the following set of N single particle equations:

$$\left[-\frac{1}{2}\nabla^2 + \upsilon_{eff}(\vec{r})\right]\phi_i = \epsilon_i\phi_i, \tag{3.53}$$

$$\tag{3.54}$$

in a very similar manner as what one needs to do for the Hartree/Hartree-Fock scheme, self-consistently. The total energy expression as a function of the eigenvalues ϵ_i reads:

$$E_{KS}[n] = \sum_{i}^{N}\epsilon_i - \frac{1}{2}\int \upsilon_H(\vec{r})n(\vec{r})d^3r - \int \upsilon_{xc}(\vec{r})n(\vec{r})d^3r + E_{xc}[n], \tag{3.55}$$

where a double-counting term needs to be subtracted from the sum of eigenvalues. The expression in 3.55 only strictly holds after achievement of self-consistency in the KS scheme. Up to this point, the theory is exact in the limit of the non-relativistic Born-Oppenheimer approximation. If all the functional forms appearing in 3.55 were known, they would give an exact expression for the ground-state energy. However, the functional form of $E_{xc}[n]$ is unknown and an approximation is necessary for this term.

These approximations for the exchange-correlation potential, discussed in detail in the next section, have made DFT quite a success, being essentially the most used theory for electronic structure in the field of condensed-matter theory, and also very prominent in the field of quantum chemistry [114]. The reader is referred to excellent reviews found in Refs. [87, 90, 115, 116] for more details. The great advantage of the formalism is that it allows one to approximate (in practice, by postulating E_{xc}) any system of interacting particles by solving a set of single particle equations. In its Kohn-Sham formulation, it formally scales as N^3 (not for all functional forms, though), although it can also be reformulated to scale linearly. In the next Section, the most common approximations to $E_{xc}[n]$ are discussed.

3.8 Exchange-correlation approximations in DFT

In this section the most popular exchange-correlation functional forms are presented in a “Jacob's Ladder” going from the Hartree approximation ($E_{xc}[n] = 0$) to the “exact” heaven, as proposed by J. Perdew [117]. The steps in the ladder correspond to adding more complex elements to the exchange-correlation functionals. One must keep in mind, though, that these steps do not represent an actual monotonic improvement on the performance of the functionals. In fact, there are quite a few known cases where a functional in a lower rung of the ladder will perform better for specific systems and properties than others higher up [116]. Nowadays there is an enormous zoo of exchange-correlation functionals available, and there are new ones coming out essentially every year [118]. Since part of the work of the author of this thesis was to implement a few GGA and meta-GGA functionals into the FHI-aims code (discussed further in Chapter 5), that were necessary for the work presented here, some more details about the ideas on which the functional forms are based will be given. For a very detailed review on the current status and development of exchange-correlation functionals the reader is referred to Ref. [118], for example.

3.8.1 Local (spin) density approximations

The local (spin) density approximation (LDA or LSDA) assumes that the exchange-correlation potential v_{xc} depends only on the electronic (spin) density and is exact for an homogeneous electron gas (HEG). This approximation was actually already proposed with the original paper of Kohn and Sham [113]. The exchange-correlation energy functional has the form of:

$$E_{xc}^{LDA} = \int n(\vec{r})\epsilon_{xc}^{HEG}(n(\vec{r}))d\vec{r} \tag{3.56}$$

The exchange-energy of the the HEG is known analytically:

$$E_{x}^{HEG} = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int n^{4/3}(\vec{r})d\vec{r}, \tag{3.57}$$

and therefore used to construct E_x^{LDA}[116]. The correlation of the HEG has no known analytical form, and only limits for very high or low densities are known[119]. Very accurate Quantum Monte-Carlo (QMC) calculations have been performed for intermediate densities of ϵ_{xc}^{HEG} [120]. These values have been interpolated for the correlation part in several ways by different authors [121–123], but producing only slightly different results.

The LDA is a very good approximation for systems where the density varies slowly and profits from some error cancellations in describing the exchange and correlation “holes”, which explains to some extent its surprisingly good description of many solids [90]. This approximation is, however, strictly local, depending on the density of the system only at point $\vec{r}$. What LDA clearly does not describe are situations where the density varies more rapidly (i.e., is strongly inhomogeneous), for example in single atoms or small/medium molecules. A common feature of LDA is also that it over-binds systems, predicting too strong binding and cohesive energies and too small lattice constants [124, 125]. There are also other problems with the LDA functional e.g. the lack of the derivative discontinuity (the LDA exchange-correlation potential varies continuously for fractional occupation numbers when it should not [126]), non-exactness for the one electron limit, etc. In particular, the problem that LDA does not include any inhomogeneity of the electron gas, is what the generalized gradient approximation tries to improve, by adding gradient corrections.

3.8.2 Generalized gradient approximations

The idea to include gradients of the density in the exchange-correlation potential came from the early realization that LDA is not a good approximation for systems where the density varies rapidly. Including the gradient by considering a Taylor expansion of ϵ_{xc} in terms of the density, would take into account some non-locality and therefore larger variations of the electron gas density. Using just a Taylor expansion, however, would lead to unphysical effects where the density varies too rapidly (correlation energies may actually become positive [127]) which makes the approximation fail dramatically for finite systems [116].

It was found that more general functions of $n(\vec{r})$ and $\nabla n(\vec{r})$ (instead of a power-series) work much more satisfactorily [116]. These approximations are called the generalized gradient approximations (GGA). The form of the xc energy functional is:

$$E_{xc}^{GGA} = \int n(\vec{r})\epsilon_{xc}^{GGA}(n(\vec{r}), \nabla n(\vec{r}))d\vec{r}. \tag{3.58}$$

For GGA functionals $\epsilon_{xc}^{GGA}(n(\vec{r}), \nabla n(\vec{r}))$ is often written in terms of an enhancement factor F_{xc} multiplied by the exchange-density of the homogeneous electron gas,

$$E_{xc}^{GGA} = \int \epsilon_{x}^{HEG} F_{xc}(n(\vec{r}), \nabla n(\vec{r}))d\vec{r}. \tag{3.59}$$

GGAs often show a better performance, as compared to LDA, for geometries and ground-state energy of molecules, especially when dealing with covalent bonded and weakly bonded systems [128–130]. The functional form of F_{xc} for GGAs in not fixed, though, which gave rise to many different GGA functionals that have been proposed over the years.

One of the most popular GGA functionals, and also the most used in this thesis is the one proposed by Perdew, Burke, and Ernzerhof [131][7] (PBE). This functional was a follow-up of other GGAs previously proposed by J. Perdew and co. (PW86 and PW91)[132, 133]. Although, as mentioned above, the form of F_{xc} for GGAs is not fixed, PBE is often referred to as a non-empirical GGA in the sense that its parameters are obtained from considering exact limits, like the homogeneous electron gas and sum rules. This functional describes well lattice constants, but exhibits a general tendency (opposed to LDA) to under-bind systems [124, 125]. Changes to PBE (still remaining in the GGA class of functionals) have been proposed over the years. Their differences lie mainly in the exchange enhancement factor F_x. In Appendix A the explicit enhancement factors of various GGA functionals are given and plotted. To cite here only a few variations: revPBE [134] changes one parameter on F_x by one obtained from fitting exchange only total energies to reference data of atoms and molecules RPBE [135] changes the form of the exchange enhancement factor F_x so that properties of chemisorbed atoms/molecules are better described than PBE; and PBEsol [136] that recovers the exact density gradient expansion for the HEG for slowly varying densities in the exchange term, describing better solids and surfaces (but not atoms and molecules).

[7] The original PBE reference has actually been cited more than 17000 times according to ISI Web of Science!

Besides the Perdew school of GGA functionals, another important line of development has been pursued by Becke and coworkers [137, 138]. By realizing that molecules and atoms bear very little resemblance to the HEG, Becke proposed an exchange functional that was based on fitting two parameters to bond and dissociation energies of a set of diatomic molecules. This exchange functional (see Appendix A for its explicit form) can be connected to a correlation term e.g. the PW86 [132], giving what is known as the BP86 functional. Functionals with the Becke exchange are popular in the chemistry community, due to their fairly good description of atoms and molecules coming from the fit used for its construction. The most popular of them is the BLYP functional, which uses a correlation functional proposed by Lee, Yang, Parr[139], containing an empirical parameter in the correlation part. This functional, however, has been shown to perform poorly for extended systems [140], with the "non-empirical" GGAs of Perdew and co. performing on average better when considering molecular and extended systems.

More recently, Armiento and Mattson [141, 142] have proposed a GGA functional for systems containing electronic surfaces. They take a subsystem approach: The exchange for surface regions is taken from the Airy gas, which is a model for an edge electron gas where the electrons are subject to a linear effective potential [143]. For the bulk regions, the exchange is taken from LDA. There is an "on the fly" interpolation between these two descriptions depending on values of the reduced density gradient s (ratio between density gradient and density, given explicitly in Appendix A).

3.8.3 Hybrid functionals

Hybrid or hybrid-exchange functionals include a fraction of exact exchange in the Hartree-Fock (Eq. 3.26) exchange sense, but using orbitals different from the HF ones to compute it. This type of functionals aim to reduce the self-interaction error present in local or semi-local functionals The price to pay is that one needs to evaluate the non-local exchange operator and the calculations tend to become more expensive.

In particular the functional known as PBE0[144, 145], that will be used in this thesis, mixes $a_0 = 1/4$ of exact exchange (E_{EX}) to the PBE functional, having the form:

$$E_{xc}^{PBE0} = a_0 E_{EX} + (1 - a_0) E_x^{PBE} + E_c^{PBE}, \tag{3.60}$$

The value $1/4$ is was chosen based on considerations from fourth order many-body perturbation theory [146].

Another hybrid functional that will be used in this work is called B3LYP[147]. This functional is perhaps the most popular in the quantum-chemistry community due to its good description of molecules and vibrational frequencies. The functional contains three empirical parameters a_0, a_1, a_2 fitted to reproduce atomization energies, ionization potentials, and proton affinities. It has the following form:

$$E_{xc}^{B3LYP} = a_0 E_{EX} + (1 - a_0) E_x^{LDA} + a_1 \Delta E_x^B + (1 - a_2) E_c^{VWN} + a_2 E_c^{LYP} \tag{3.61}$$

where ΔE_x^B corresponds to only the gradient correction to the exchange energy given by Becke[137], E_x^{LDA} is the LDA exchange functional, E_c^{VWN} is the LDA correlation functional of Vosko, Wilk, and Nusair [122], and E_c^{LYP} the GGA correlation functional of Lee, Yang, and Parr. The parameters are set to $a_0 = 0.20, a_1 = 0.72$, and $a_2 = 0.81$. In the first implementation of this functional in the Gaussian code [148], an old version of the VWN correlation [149] functional was used, which was not the one the authors of the original reference [147] had intended. The implemented version became the final definition of B3LYP used up to now, and surprisingly was even proven to be more accurate than the original version [149].

The performance of these hybrid functionals (as well as the others mentioned above) for weakly bonded complexes (vdW, H-bonds) that are of relevance to this thesis will be discussed in Section 3.10.

3.8.4 Meta-generalized gradient gpproximations

The exchange-correlation functionals known as meta-GGAs (mGGA) consider, additionally, the Laplacian of the density ($\nabla^2 n$) or [8] the orbital (spin) kinetic-density:

$$\tau(\vec{r}) = \frac{1}{2}\sum_i |\nabla\phi_i|^2, \tag{3.62}$$

where ϕ_i are the Kohn-Sham orbitals.

These two quantities are related by

$$\tau(\vec{r}) = \frac{1}{4}\nabla^2 n(\vec{r}) - \frac{1}{2}\sum_i \phi_i^*(\vec{r})\nabla^2\phi_i(\vec{r}). \tag{3.63}$$

In practice most functionals use only τ because it is strictly positive and avoids the singularities of the Laplacian close to the nuclei [9]. Perdew and Constantin [151] have studied the differences of building a mGGA functional based solely on $\nabla^2 n$ or on $\tau(\vec{r})$, and arrived to the conclusion that both quantities carry the same information beyond what is contained in n and ∇n. The general form of a (τ-dependent) mGGA functional is:

$$E_{xc}^{MGGA} = \int n(\vec{r})\epsilon_{xc}^{MGGA}(n(\vec{r}), \nabla n(\vec{r}), \tau(\vec{r}))d\vec{r}. \tag{3.64}$$

The kinetic-energy density depends explicitly on the Kohn-Sham orbitals and only indirectly on the density. Still, it is possible, in principle, to build a potential that is only functional of the density. This, however, requires the evaluation of an additional functional derivative, namely:

$$\frac{\delta\tau(\vec{r})}{\delta n(\vec{r})}. \tag{3.65}$$

This term is hard to be evaluated since τ is not an explicit function of n (see Eq. 3.63). Most self-consistent implementations of meta-GGAs do not evaluate the functional derivative with respect to the density but with respect to the orbitals [152]. There is no rigorous proof, though, that orbitals derived from orbital self-consistency are the same as, or similar, to orbitals derived from the density self consistency. For many considerations concerning energetics, though, it is sufficient to compute the meta-GGA energy after achievement of self-consistency with another functional (e.g. a GGA functional) [116, 153] (post-processing).

Examples of mGGAs are the PKZB [154] and TPSS [155] functionals, from the "Perdew branch", and the M05 [156], M06 [157, 158], and M08 [159] suite of functionals by Zhao and Truhlar. The latter suites of functionals can have more than 30 parameters fitted to give good descriptions of barrier heights of chemical reactions and non-covalent interactions (including vdW interactions). In the M06 suite, M06, M06-2X, and M06-HF include some portion of exact exchange (0.27, 0.54, and 1 respectively), while M06L contains no contribution from exact exchange.

[8] There are mGGA that use the Laplacian and others (most common) that use only the kinetic-density (see Ref. [150] and references therein).

[9] The last term on the right of Eq. 3.63 is not strictly positive, and thus generally not used to develop mGGA funcionals.

3.8.5 Van der Waals corrections to DFT exchange-correlation functionals

Nowadays, van der Waals interactions are accepted to be absolutely necessary for the description of biomolecules [14, 15, 160]. It is customary to refer to van der Waals (vdW) forces as the London quantum "dispersion" forces proposed by F. London in 1930 [52], when studying the attraction between two neutral noble-gas atoms. These forces arise from the electronic *correlation*, that takes place due to the formation of instantaneous induced dipole - induced dipole interactions between two polarizable atoms. In a classical picture, one can understand this interaction as the $1/R^6$ attractive electrostatic energy that appears when one atom induces a temporary dipole in a second atom at distance R. The quantum dispersion force comes from quantum fluctuations of the electronic density [89] that gives rise to instantaneous dipoles. Casimir and Polder [161] formulated the expressions for the van der Waals forces in terms of quantum electrodynamics (QED), having the same origins as the famous Casimir effect [161, 162]. The idea is that the instantaneous fluctuating polarization of the electrons in the atoms interact with the zero-point vibration of the electromagnetic field permeating the system, which in turn interacts with other possible electrons. This causes the polarizability of each atom to be frequency dependent. Even though the fluctuating dipole of one atom averages to zero over time, all atoms oscillate in (opposite-)phase, such that on average there is an attractive force between them.

Van der Waals interactions are highly non-local and dominate when there is no overlap between charge densities, for systems with no permanent electrostatic multipole moments. This presents an obvious problem for the local and semi-local functionals of DFT that take into consideration only the electronic density at point $\vec{r}$ (and its immediate vicinities). The density and its gradient expansion have no knowledge of variations which arise more than 3-4 Å away from the point where they are evaluated (see Eq. 3.58), and that is where attractive vdW interactions arise. In fact, it can be shown that with standard LDA and GGA functionals, the asymptotic tail of the energy, for a large separation R between atoms, approaches zero exponentially, while a proper theory that takes vdW interactions into account should have the characteristic $1/R^6$ tail. It has been proven by a thought-experiment presented in Ref. [163] that these interactions are missing in any local or semi-local DFT functional and shown to be clearly missing for LDA and GGA calculations of rare-gas dimers binding energies [164]. Semi-local functionals that claim to include vdW effects like the M06 and M08 suite of meta-GGAs mimic locally (for short separation between atoms) this interaction. These interactions are included (not always fully [165]) in theories like MP2 and CC presented above, for example. In a many-body picture, the interaction of the ground state of one atom with the possible excited states is the source of the dispersion interaction.

One straightforward way to include vdW interactions in DFT is via empirical or semi-empirical corrections. In terms of the so-called Casimir-Polder integral, the leading $1/R^6$ term for the dispersion at long ranges can be written with respect to the (imaginary) frequency dependent polarizability $\alpha(i\omega)$ of two atoms A and B as:

$$E_{disp} = -\frac{1}{R_{AB}^6}\frac{3}{\pi}\int \alpha_A(i\omega)\alpha_B(i\omega)d\omega = -\frac{1}{R_{AB}^6}C_6^{AB} \tag{3.66}$$

which gives an expression for the heteronuclear C_6^{AB} coefficient. Tang [166] derived an expression for the heteronuclear coefficient in terms of the homonuclear coefficients C_6^{AA} and C_6^{BB} and their static polarizabilities (α_0^A, α_0^B), which is:

$$C_6^{AB} = \frac{2C_6^{AA}C_6^{BB}}{\frac{\alpha_0^B}{\alpha_0^A}C_6^{AA} + \frac{\alpha_0^A}{\alpha_0^B}C_6^{BB}} \tag{3.67}$$

The idea of the semi-empirical corrections is to estimate the C_6 coefficients of equation 3.66 and add a term of the form $-\sum_A \sum_{B>A} C_6^{AB}/R_{AB}^6$ to the exchange-correlation energy, multiplied by a damping function so that the divergence for $R=0$ is removed and no double-counting for short distances, where the exchange-correlation functionals are already accurate, occur. The general form of this type of correction to the DFT energy is:

$$E_{DFT+vdW} = E_{DFT} - \sum_A \sum_{B>A} f_{damp}(R_{AB}) \frac{C_6^{AB}}{R_{AB}^6}, \tag{3.68}$$

where f_{damp} is the damping function, which is arbitrary (to a certain extent), so that it remains an intrinsic "empiricity" for all these methods. There have been several schemes along the lines of Eq. 3.68 proposed [14, 15, 167–175] and they differ in the way they determine the C_6 coefficients and in the form of the damping function, including, or not, extra parameters.

Most schemes [167–172] use C_6 coefficients that are empirically evaluated and fixed for each atom. Real dispersion coefficients, though, depend on the molecular environment of each atom [168, 175]. The use of fixed coefficients introduces errors to the estimation of the dispersion correction, since an atom like carbon, for example, can exhibit a variation of more than 50% in its C_6 coefficient, depending on the "environment" (e.g. diamond, graphene, or CH_3). The popular scheme of Grimme [170] also proposes a global scaling parameter, that is different for each DFT-functional, so that the functional dependence of the scheme is reduced.

One of the most non-empirical (if such a definition exists) schemes of this type, is the one proposed in 2005 by Becke and Johnson [174, 175], where the dipersion C_6 coefficients are calculated from the exchange-hole dipole moment, such that the inputs needed are the KS-orbitals and the density of a system (plus the polarizabilities of free-atoms). The damping function in this scheme remains empirical, though.

The scheme used in this thesis, proposed by A. Tkatchenko and M. Scheffler [2] (TS-vdW), incorporates ideas from both strategies listed above. The C_6 coefficients depend on the environment, via an explicit dependence on the ratio of Hirshfeld volumes of the atom in the molecule and the free atom. The Hirshfeld partition[176] of the electronic density for one atom in a molecule ($n_A(\vec{r})$) is

$$n_A(\vec{r}) = n(\vec{r}) \frac{n_A^0(\vec{r})}{\sum_X n_X^0(\vec{r})}, \tag{3.69}$$

where $n(\vec{r})$ is the electronic density of the molecule, $n_A^0(\vec{r})$ is the density of the free atom, and the sum over X runs over all atoms of the molecule, taken as free atoms, but in the position they would be found in the molecule. The number of electrons in the atom is obtained by integrating $n_A(\vec{r})$ over all space, and the volume of the atom in the molecule is obtained by assuming spherical atoms with a constant density. The coefficients for the free atoms can be calculated very accurately, being taken from Ref. [177][10] in the TS-vdW scheme. In this way, the C_6 coefficients become also a functional of the density n, but still retain some empiricity introduced by the Hirshfeld partition. This correction has the following form:

[10] Where the values were calculated using time-dependent density-functional theory with an LDA kernel (TDLDA).

$$
\begin{aligned}
E_{TS-vdW} &= -\sum_{A}\sum_{B>A} f_{damp}(R_{AB}, R^0_A, R^0_B)\frac{C_6^{AB}}{R^6_{AB}} && (3.70)\\
f_{damp}(R_{AB}, R^0_A, R^0_B) &= \frac{1}{1+\exp[-d(\frac{R_{AB}}{s_R R^0_{AB}}-1)]} && (3.71)\\
C^{eff}_{6AA} &= \left(\frac{V^{eff}_A}{V^{free}_A}\right)^2 C^{free}_{6AA} && (3.72)
\end{aligned}
$$

where C_6^{AB} is given by Eq. 3.67. The term $R^0_{AB} = R^0_A + R^0_B$ in 3.71 is the sum of effective van der Waals radii and is taken directly from the Hirshfeld volumes V^{eff}. The scaling parameter s_R in the damping function is empirical and functional dependent, determining the onset of the correction [11]. The parameter s_R is obtained for different functionals, by fitting to the S22 data base [178], which will be discussed in Section 3.10. For most functionals, the value of s_R is close to one. The value of this parameter for different functionals (PBE0, B3LYP, PBE, BLYP, revPBE, AM05, M06L, and M06), is shown in Table 3.1.

Functional	s_R
M06L	1.26
M06	1.16
PBE0	0.96
PBE	0.94
AM05	0.84
B3LYP	0.84
BLYP	0.62
revPBE	0.60

Table 3.1: Values for the parameter s_R of the TS-vdW scheme, for different functionals.

As a last remark in this section, it is worth pointing out that a description of dispersion interactions within DFT can also be achieved by explicitly nonlocal correlation functionals. Such a functional was devised by Langreth, Lundqvist and coworkers[179, 180]. It adds a non-local term to a "standard" GGA exchange correlation energy based on an approximation for the dieletric function between two fragments. This functional was originally coupled to revPBE exchange, which is not a rigorously motivated choice and has been subject to debate [181]. A newer version of the functional, the vdW-DF2 [182], seems to improve the accuracy of the results.

3.9 Random-phase approximation

The method called the random-phase approximation[12] (RPA) [119, 183–187], is a DFT functional that takes into account the many-body correlation (and therefore vdW dispersion). RPA is also a perturbation theory, but it is formulated in terms of the screened Coulomb interaction. This interaction differs from the bare Coulomb interaction (used in Hartree-Fock or MP2, for example), by considering not only the bare electron, but the electron plus its polarization cloud, which is commonly called the "quasi-particle". RPA "dresses" the Coulomb interaction by taking into account the non-interacting density response function χ^0 of the system (which is related to its dielectric function), through the framework of linear response theory. The screened Coulomb interaction, contains not only the response of the non-pertubed system to a

[11] The parameter d is chosen to be equal to 20, with results changing negligibly for $12 < d < 45$[2].

[12] The origin of this name is explained in Ref. [183].

perturbation but also the response of the response, and so on, giving rise to an infinite geometric series. The RPA correlation energy sums all contributions from second-order electron-hole interaction, up to infinite order in the perturbation [88] (in diagramatic terms, they are called bubble diagrams). In contrast to finite order perturbation theory, as e.g. MP2, the divergence for metallic systems is avoided in RPA.

The RPA correlation energy E_c^{RPA} can be written as follows, in imaginary frequency ($i\omega$):

$$\begin{aligned} E_c^{RPA} &= \frac{1}{2\pi}\int_0^\infty d\omega \mathrm{Tr}\left[\ln\left(\varepsilon(i\omega)\right) + \left(1-\varepsilon(i\omega)\right)\right] \\ &= \frac{1}{2\pi}\int_0^\infty d\omega \mathrm{Tr}\left[\ln\left(1-\chi^0(i\omega)v\right) + \chi^0(i\omega)v\right], \end{aligned} \tag{3.73}$$

where v is the Coulomb interaction, ε is the frequency-dependent dielectric constant, and χ^0 (the response function) is given by [188, 189]:

$$\chi^0(\vec{r},\vec{r}',i\omega) = \sum_\sigma \sum_j^{\mathrm{occ}} \sum_a^{\mathrm{unocc}} \frac{\phi_{j\sigma}^*(\vec{r})\phi_{a\sigma}(\vec{r})\phi_{a\sigma}^*(\vec{r'})\phi_{j\sigma}(\vec{r'})}{i\omega - \epsilon_{a\sigma} + \epsilon_{j\sigma}} + \mathrm{c.c.}, \tag{3.74}$$

where c.c. denotes “complex conjugate”, $\phi_n(\vec{r})$ and ϵ_n are single-particle orbitals and orbital energies, and σ denotes the spin components.

RPA is most often used in a post-processing fashion, so that the molecular orbitals and energies appearing in 3.74 are taken to be that of DFT or HF. The exchange part for the RPA total energy is evaluated as the exact exchange, which is defined by the same expression as in HF (Eq. 3.27) but using non Hartree-Fock orbitals. From here on this method will be referred to as EX+cRPA. EX+cRPA evaluated for DFT-PBE Kohn-Sham orbitals (EX+cRPA@PBE) orbitals has been shown to work very well for extended systems (including metallic ones), providing very good lattice constants, bulk moduli, heats of formation, adsorption energies and surface energies [125].

When the Hamiltonian H_0 taken to optimize the orbitals entering the EX+cRPA expression is *not* the HF Hamiltonian, the Brillouin theorem, mentioned in the discussion about MP2, is not valid anymore. It means that there will be a contribution from singly excited configurations interacting with the ground-state that is not zero [190]. Including these contributions in the expression for the RPA total energy is what will be named EX+cRPA+SE (SE for single excitations).

3.10 Performance of functionals for H-bonded and vdW systems

In order to categorize the “zoo” of competing electronic structure methods of the preceding sections, their performance is here addressed, especially for phenomena relevant for this thesis. As mentioned in the introduction, two of the most important non-convalent interaction shaping a polypeptide are H-bonds and van der Waals (vdW) dispersion interactions. While H-bonds form mainly due to electrostatic interactions and charge transfer, vdW forces arises mainly from nonlocal electronic correlation effects. The evaluation of the H-bonding energy contribution is straightforward, and any first-principles electronic structure method describes H-bonding, although at different levels of accuracy (as will be discussed in further details in the next paragraphs). The evaluation of the vdW contribution is usually much more involved, and in fact also H-bonds are affected by vdW interactions [191], such that often the two cannot be regarded separately.

The Hartree-Fock method already describes reasonably well H-bonding interactions. Hartree-Fock has been applied to the molecule formed by one alanine amino acid, capped by the acetate group on the N-terminus and the N-methyl amide group on the C-terminus (Ac-Ala-NMe), known as alanine dipeptide [192], for example. The energetic ordering of H-bonded and non-H-bonded structures [192] was well described, even if the absolute energetics did not agree with methods that include correlation. VdW forces, on the other hand, are completely missing from Hartree-Fock, such that it predicts a completely repulsive binding-energy curve for rare-gas dimers, for example. HF by itself is, thus, not suitable for the treatment of biomolecules.

Conformer Nr.	CCSD(T)	MP2
1	-0.138	-0.139
2	-0.218	-0.218
3	-0.807	-0.807
4	-0.692	-0.688
5	-0.888	-0.886
6	-0.725	-0.753
7	-0.710	-0.717
8	-0.023	-0.022
9	-0.066	-0.070
10	-0.065	-0.081
11	-0.118	-0.215
12	-0.192	-0.299
13	-0.428	-0.484
14	-0.226	-0.352
15	-0.530	-0.647
16	-0.066	-0.073
17	-0.142	-0.157
18	-0.102	-0.118
19	-0.193	-0.224
20	-0.119	-0.157
21	-0.249	-0.305
22	-0.306	-0.337

Table 3.2: Binding energies, in eV, of the S22 set of molecular dimers, calculated with CCSD(T) and MP2 (extrapolated to the CBS limit), taken from Ref. [178].

MP2 gives a satisfactory description of both H-bonds and dispersion interactions. However, it has a tendency to overestimate the dispersion interaction for non-covalent bonded molecules, as has been shown, for example, in Refs. [165, 178] [13]. A particularly well-studied set of weakly bonded model systems is the so-called S22 database: 22 non-covalently bonded dimers of up to 30 atoms, introduced in in Ref. [178]. This set is used also in this work as a benchmark for testing the accuracy of different electronic-structure methods for non-covalent interactions. The 22 molecular complexes are shown explicitly in Appendix E. In the original publication, the molecules are divided in three groups, according to the predominant character of the non-covalent bond present: numbers 1–7 are hydrogen bonded; numbers 8–15 are dispersion bonded; and numbers 16–22 are "mixed" complexes. CCSD(T) binding energy extrapolated to the complete basis set (CBS) limit are also reported in Ref. [178] for all the 22 complexes. The performance of MP2, compared with the CCSD(T) values for the binding energy of these

[13] Recently, in 2010, MP2 has been seen to have the opposite effect for extended weakly polarizable systems[193], i.e., in that case the correlation is understimated.

complexes is rather good, as can be seen in Table 3.2, where the values were taken from Ref. [178]. The absolute values of the binding energies of these complexes are rather small, such that high accuracy methods are required for a good description of the systems.

In the realm of DFT, different functionals exhibit very different performances for the treatment of both interactions.

For H-bonds, extensive work regarding different DFT-functionals has been done by Santra *et al.* on the geometry of small water clusters [194–196]. The PBE0 functional shows the best performance for the energetics of the H-bonds considered, coming to within 5-10 meV deviation for the dissociation energies of the water clusters (up to the pentamer), taking MP2 as a reference. BLYP and B3LYP functionals consistently underestimate the strength of the H-bond, and PBE has a variable performance with cluster size [194]. When adding vdW interactions, the energetics of the water clusters are seen to be improved for several GGA functionals (BLYP, PBE0, PBE)[195]. Staroverov *et al.* [197] studied several hydrogen-bonded complexes, concluding that the “semi-empirical” functionals (e.g. BLYP, B3LYP) did not present a consistent performance with respect to system size: they perform on average good for small molecules, but bad as molecular size increases. The PBE functional has been seen to yield a reasonable description of H-bonds over several model H-bonded biomolecules (with up to 24 atoms) by Ireta *et al.*, exhibiting errors around 1 kcal/mol (43 meV) per H-bond, but this accuracy seems to depend on the directionality of the H-bonds [198].

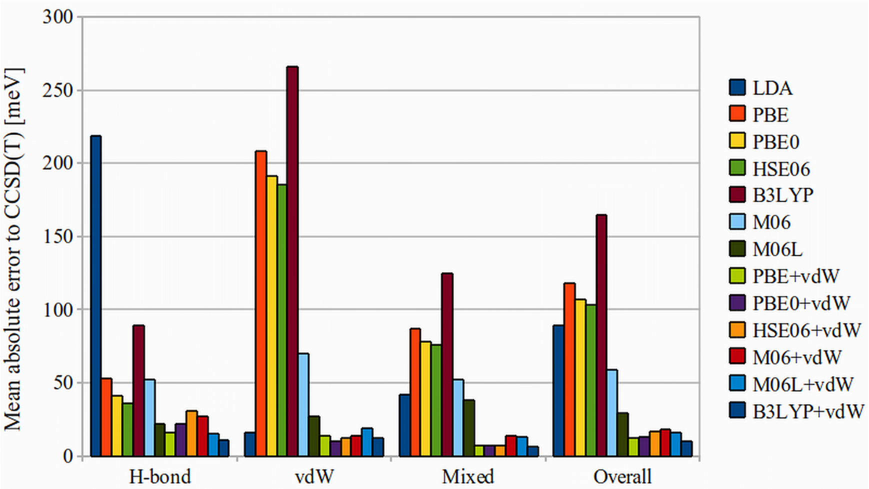


Figure 3.2: Reproduced from Ref. [199], with permission from the other authors: “Mean absolute errors of different functionals with and without the TS-vdW correction with respect to CCSD(T) reference values for the binding energies of the S22 data set”

More specifically regarding vdW interactions, recent work from Marom *et al.* (in which the author of this thesis collaborated) has assessed the performance of several GGA, mGGA and hybrid-GGA/mGGAs, with and without including the TS-vdW [2] correction for the binding energies of the S22 set of molecules. The result of this study is shown in Figure 3.2, reproduced from [199]. Among the plain GGA and hybrid-GGA functionals, PBE and PBE0 perform better than B3LYP for all types of complexes in the set (H-bonded, vdW, and mixed). The meta-GGA M06L [157], which was designed to mimic vdW interactions locally, shows the best performance among the bare functionals. When including vdW interactions, the performance is substantially improved for *all* functionals. PBE+vdW and PBE0+vdW show very similar performance for all types of complexes in the set, with PBE+vdW even being a bit better than PBE0+vdW for the H-bonded ones. B3LYP+vdW shows the best performance overall, although similar to PBE+vdW

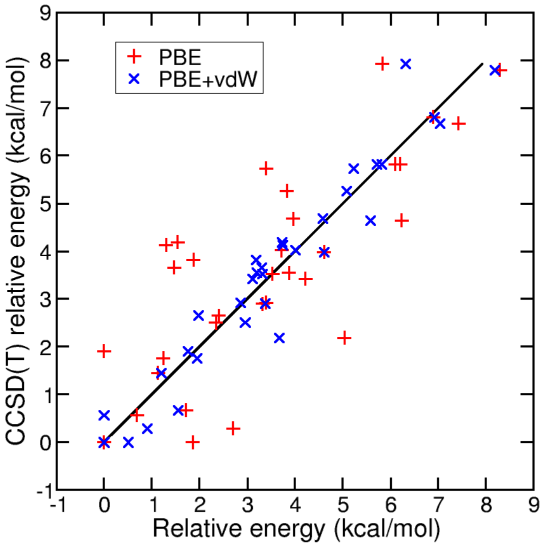


Figure 3.3: Stabilization energies of alanine dipeptide and tetrapeptide conformers as obtained by PBE and PBE+(TS-)vdW methods in comparison to CCSD(T) results. For every method, the energy zero was taken to be the lowest energy CCSD(T) structure of the tetrapeptide and dipeptide.

and PBE0+vdW. The meta-GGAs M06L and M06 also benefit from the inclusion of the C_6/R^6 term, underlining the fact that the asymptotic correlation corrections also affect the energetics in the equilibrium distance range.

The performance of the TS-vdW scheme connected to the PBE functional (PBE+vdW), was also tested for the relative stabilization energies of 32 different conformations of the alanine dipeptide and tetrapeptide [6]. The reference was taken to be the CCSD(T) relative stabilization energies and the results are shown in Figure 3.3. This data is published and detailed in the supplemental material of Ref. [6]. In Figure 3.3 the PBE+vdW functional is also compared to the standard PBE functional, highlighting the remarkable improvement caused by the inclusion of vdW interactions. The mean absolute error of PBE+vdW in comparison to CCSD(T) for these conformers is of only 18 meV. This accuracy is meaningful when it comes to vdW interactions in these systems, as well as larger ones, as the term is already definitely large.

The accuracy of EX+cRPA and EX+cRPA+SE was also studied for the binding energies of the S22 set, by Ren and coworkers [190]. While EX+cRPA with PBE orbitals (EX+cRPA@PBE) systematically underbinds with respect to CCSD(T) reference data, EX+cRPA+SE@PBE performs better overall, and shows a significant improvement also over MP2 for dispersion bonded systems.

From the DFT functionals discussed, PBE+vdW, PBE0+vdW and EX+cRPA+SE@PBE, thus emerge as the best choices for the description of the polypeptides discussed in this thesis. From these three, PBE+vdW is the computationally cheapest one, therefore becoming the natural choice for the production calculations. Its accuracy for the specific systems studied in this thesis will be further investigated in the chapters to come. Data from other functionals, as well as MP2 and EX+cRPA(+SE) will also be shown in this thesis.

Chapter 4

Molecules in motion: vibrational spectroscopy and molecular dynamics

Having characterized methods to describe the potential-energy surfaces in Chapter 3, this chapter describes a few theoretical techniques to treat the movement of the nuclei on these surfaces. A detailed account of the methods outlined here can be found in textbooks like *Statistical Mechanics* by D. McQuarrie [200], *Understanding Molecular Simulation* by D. Frenkel and B. Smit [201], and others [202].

4.1 Harmonic approximation for vibrations of the nuclei

In order to understand the harmonic approximation for the vibration of the nuclei, we start from a classical derivation (the quantum-mechanical one will then follow). Assuming the Born-Oppenheimer approximation discussed in Section 3.3, the degrees of freedom of the nuclei can be separated from that of the electrons. If the nuclei are classic particles, the following Hamiltonian describes their motion:

$$\mathcal{H}(\vec{R}_I, \vec{p}_I) = \sum_I \frac{p_I^2}{2M_I} + V_{BO}(\vec{R}), \tag{4.1}$$

where p_I is the momenta associated with the nuclei and

$$V_{BO}(\vec{R}) = E_0^e(\vec{R}) + V_{nn}(\vec{R}), \tag{4.2}$$

where, in turn, $E_0^e(\vec{R}_I)$ is the total electronic energy and $V_{nn} = \frac{1}{2}\sum_{I,J} \frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|}$ is the Coulomb interaction between the nuclei. We here follow the derivation using V_{BO}, but such potential energy can, in principle, also be obtained from a force-field or any other empirical potentials.

Considering small displacements close to the bottom of the potential well, the BO potential can be expanded in a Taylor series

$$V_{BO}(\vec{R}) = V_{BO}^0 + \sum_I (\vec{R}_I - \vec{R}_I^0)\left(\frac{\partial V_{BO}}{\partial \vec{R}_I}\right)_0 + \frac{1}{2}\sum_{I,J}(\vec{R}_I - \vec{R}_I^0)(\vec{R}_J - \vec{R}_J^0)\left(\frac{\partial^2 V_{BO}}{\partial \vec{R}_I \partial \vec{R}_J}\right)_0 + ... \tag{4.3}$$

If one takes $\vec{R}_I^0$ to be the equilibrium geometry, the linear term vanishes. Truncating the expansion at second order, only the value of potential at the equilibrium position plus the term depending on its second derivatives (the curvature at the minimum) are left. These second derivatives are the so-called force-constants, and the name "harmonic" approximation comes from truncating this series after second order. Assuming a spherically-symmetric potential (for small displacements), the Hamiltonian can be separated in a translational Hamiltonian and another that describes the relative motion of the bodies with respect to the center of mass of the system. Rotations also become separable, taking the system as a rigid-rotor [200]. Here, we will focus only on the relative vibrations of the nuclei, for reasons that will be explained further down.

In order to continue, we perform a change in variables, of the form

$$\vec{q}_I = \sqrt{M_I}(\vec{R}_I - \vec{R}_I^0), \tag{4.4}$$

so that the kinetic-energy and potential energy (where the $_{BO}$ label will be dropped for simplicity) become:

$$T = \frac{1}{2}\sum_I^{3N} \dot{q}_I^2 \qquad V = \frac{1}{2}\sum_{I,J}^{3N}\left(\frac{\partial^2 V}{\partial q_I \partial q_J}\right)_0 q_I q_J, \tag{4.5}$$

where N here means the number of atoms in the system, and $\dot{q}_I$ is the derivative of q_I with respect to time. The Newton equation of motion, in these coordinates, can be written as:

$$\begin{aligned} \frac{d}{dt}\frac{\partial T}{\partial \dot{q}_J} + \frac{\partial V}{\partial q_J} &= 0 && (4.6)\\ \ddot{q}_J + \sum_I^{3N}\left(\frac{\partial V}{\partial q_I \partial q_J}\right) q_J &= 0 && (4.7) \end{aligned}$$

Assuming the *ansatz* that $q_I = A_I \cos(\omega t + \phi)$ (A_I corresponds to an amplitude, ω to a frequency and ϕ to a phase) and substituting into 4.7, one obtains

$$\sum_I^{3N}\left[\left(\frac{\partial V}{\partial q_I \partial q_J}\right) - \omega^2 \delta_{IJ}\right] A_I = 0, \tag{4.8}$$

such that, in order to determine ω and A_I, it is necessary to solve the secular equation:

$$|\mathbb{H}_{ess} - \omega^2 \mathbb{I}| = 0 \tag{4.9}$$

$$\text{with} \qquad \mathbb{H}_{ess} = \begin{pmatrix} \frac{\partial V}{\partial q_1^2} & \frac{\partial V}{\partial q_1 \partial q_2} & \cdots & \frac{\partial V}{\partial q_1 \partial q_{3N}} \\ \frac{\partial V}{\partial q_2 \partial q_1} & \frac{\partial V}{\partial q_2^2} & \cdots & \frac{\partial V}{\partial q_2 \partial q_{3N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial V}{\partial q_{3N} \partial q_1} & \frac{\partial V}{\partial q_{3N} \partial q_2} & \cdots & \frac{\partial V}{\partial q_{3N}^2} \end{pmatrix}, \tag{4.10}$$

and $\mathbb{I}$ the identity matrix. The matrix $\mathbb{H}_{ess}$ is called the (mass-weighted) Hessian matrix. There are $3N$ coordinates needed to specify a polyatomic system of N atoms, with 3 needed to describe the center of mass, and other 3 (or 2 for linear molecules) needed to specify the global orientation of the molecule. The remaining $3N - 6$ ($3N - 5$ for linear molecules) coordinates are the ones necessary to describe internal motions of the atoms, i.e. vibrations. For this reason, 4.9 will have 6 eigenvalues that will be zero, with the corresponding eigenmodes describing the rotations and translations. The other eigenmodes describe the vibration of each atom. If the molecule is in a minimum of the PES, the eigenvalues corresponding to the vibrational modes will all be positive. If it is in a saddle point, there will be a number of negative eigenvalues (i.e. imaginary frequencies, since the eigenvalues are ω^2), related to normal-mode directions with negative curvatures in the PES.

Before moving to the quantum formulation, it is necessary to introduce the normal coordinates Q_k, given by

$$Q_k = \sum_I l_{kI} q_I, \tag{4.11}$$

where $l_{kI} = A_{kI}/[\sum_I A_{kI}^2]^{1/2}$ and $\sum_I l_{kI}^2 = 1$. For these coordinates, the potential energy becomes separable, assuming the form

$$V = \frac{1}{2} \sum_k^{3N} \omega_k^2 Q_k^2, \tag{4.12}$$

where Q_k are, thus, the normal modes of vibration of the molecule, which diagonalize the (harmonic approximated) potential. Since Q_k are related to q_I through 4.11, they can be used to describe simultaneous displacements of all atoms in the molecule in the direction of the eigenvectors A_{kl}, with the same frequency, but different amplitudes.

Moving now to a quantum description of the nuclei, from the BO approximation and the harmonic approximation for the potential where the nuclei move, the wave function of the full system can be factorized in an electronic and a nuclear part, that includes vibrations, rotations, and translations:

$$\psi = \psi_e(\vec{r}; \vec{R})\psi_t(\vec{R})\psi_v(\vec{R})\psi_r(\vec{R}), \tag{4.13}$$

where ψ_e is the electronic wave function that depends on the position of the electrons $\vec{r}$ and also parametrically on the position of the nuclei $\vec{R}$, and $\psi_t(\vec{R})\psi_v(\vec{R})\psi_r(\vec{R})$ is the translational, vibrational, and rotational wave function related to the movement of the nuclei[1].

The Schrödinger equation for the nuclear vibrational wave function, in the harmonic approximation and as a function of Q_k reads (here explicitly showing the $\hbar$ factor):

$$-\frac{\hbar^2}{2} \sum_k^{3N-6} \frac{\partial^2 \psi_v}{\partial Q_k^2} + \frac{1}{2} \sum_k^{3N-6} \omega_k^2 Q_k^2 \psi_v = E_v \psi_v \tag{4.14}$$

In this form, ψ_v itself becomes separable in the Q_k components, i.e. $\psi_v = \psi_v(Q_1)\psi_v(Q_2)\cdots\psi_v(Q_{3N-6})$, and Eq. 4.14 is a set of $3N-6$ coupled quantum harmonic oscillators. The energies of each mode will be $E_{(k,n)} = \hbar\omega_k(n_k + 1/2)$ where n_k is the quantum number of the modes (see schematic representation in Figure 4.1). The total vibrational energy will be given by $E_v = \sum_k E_{(k,n)}$. One refers to fundamental levels when only one $n_k = 1$ and all the others are zero.

To obtain the IR intensities of each mode of vibration, one can assume a small electric field perturbation interacting with the dipole of the molecule. The dipole moment can be expanded with respect to the normal modes of vibration as:

[1] We here assume that separating the vibrational and rotational contributions is a valid approximation. For more details see Ref. [203]

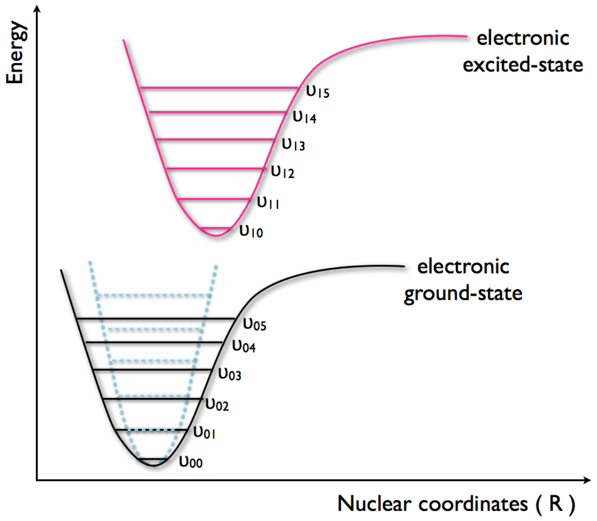


Figure 4.1: Schematic drawing of electronic and vibrational states of a hypothetical molecule. The Morse-like [204] curves in black and pink represent the electronic ground-state and the first excited state, respectively. The levels drawn in each of them correspond to the possible vibrational frequencies v_{0n}, v_{1n}. The dashed parabola-like blue line represents the harmonic approximation, with equally spaced vibrational levels.

$$\vec{\mu} = \vec{\mu}_0 + \sum_{k=1}^{3N-6} \frac{\partial \vec{\mu}}{\partial \vec{Q}_k} \vec{Q}_k + ... \tag{4.15}$$

where $\vec{\mu}$ is the dipole moment of the molecule and $\vec{Q}_k$ is the kth normal mode of vibration. Assuming that only terms up to first-order in 4.15 contribute, then one can write the matrix elements that represent the transition between two vibrational states (when no transition between *electronic* states take place, i.e., the molecule remains in its electronic ground state) via the operator $\vec{\mu}$ as:

$$\int \psi_{v'} \vec{\mu} \psi_{v''} d\Omega_v = \vec{\mu}_0 \int \psi_{v'} \psi_{v''} d\Omega_v + \sum_{k=1}^{3N-6} \frac{d\vec{\mu}}{d\vec{Q}_k} \int \psi_{v'} \vec{Q}_k \psi_{v''} d\Omega_v \tag{4.16}$$

where ψ_v is the vibrational wave function and $d\Omega_v$ is a volume element in configurational space associated to the vibrations. Since the vibrational wave functions for each vibrational state are orthogonal, the first term in the left side of 4.16 vanishes unless $v' = v''$ (i.e. no vibrational transition), which means that this term does not affect the intensity of vibrational transitions. μ_0 is actually the permanent dipole moment of a molecule (when the molecule has one), being, thus, irrelevant for the transition intensities. For the second term on the right of 4.16, if one assumes that the vibrational wave function ψ_v can be written as a product of harmonic oscillator functions (again a harmonic approximation), all terms will vanish unless $v_i'' = v_i' \pm 1$[2], giving rise to a selection rule. When considering IR intensities of vibration in the harmonic approximation, thus, the only transitions observed will be the ones where the vibrational quantum number changes by 1, and $\frac{d\vec{\mu}}{d\vec{Q}_k} \neq 0$. Only fundamental frequencies associated with a change in dipole moment can be present in the harmonic IR spectrum estimation. In "real life", though, combinations and overtones are also observed, but these can only be calculated if one goes away from the harmonic approximation.

The radiation interacting with the atoms can be assumed to be of the form of a plane wave electric field of a certain frequency. The IR intensity of vibrations, in this case, is obtained through time-dependent perturbation theory. The transition matrix elements involved (for the harmonic approximation), are the ones discussed in the paragraph above, since the perturbation is considered to be the interaction of the electric field with the dipole of the molecules [203]. The derivation of the IR intensity in this approximation is a bit tedious (see e.g. Ref. [205]), but gives as a final expression [205–207]:

[2] That happens because for the wave functions of the harmonic oscillator $\psi_{v'} \vec{Q}_k \propto \psi_{v'+1}$.

$$I_i^{IR} = \frac{N_A \pi}{3c} \left| \frac{d\vec{\mu}}{d\vec{Q}_i} \right|^2 , \tag{4.17}$$

where N_A is Avogadro's number and c the velocity of light. This is often called the "double-harmonic" approximation [207].

There are limitations for the use of this approximation, as expected. At high temperatures, for example, this is not a good approximation because when the atoms start to explore higher regions of the potential well, it cannot be anymore approximated anymore as a parabola. The potential well looks more like a Morse potential [204], than like a parabola. The Morse potential has a shape similar to that drawn in Figure 4.1, given, e. g. in the simple case of a diatomic molecule, by $V(d) = D_0(1 - e^{\alpha(d-d_0)^2})$, where d is the distance between two atoms, D_0 is the depth of the potential well, and α is a parameter controlling the width of the well. Therefore, the "real" anharmonic modes are more closely spaced than what is estimated by the harmonic picture, as schematically drawn in Figure 4.1. Disregarding the anharmonic terms in this approximation leads to sizable shifts in the frequencies, already due to anharmonicities in the QM expectation value of the zero-point wave function, for real systems [208–210]. Methods that include anharmonicities in a QM framework will not be discussed in this thesis, but the reader is referred to Ref. [210] for more details. In short, due to the multidimensional nature of the problem, these methods lead to incredibly expensive calculations for the size of systems treated in this thesis, especially when connected with *ab initio* methods. A way to include anharmonicities in the estimation of the spectra, although in a classical picture, will be discussed in Section 4.3.

The harmonic approximation for the analysis of normal modes is perhaps the most used first approximation for the study of vibrational states of a system. In recent years, many techniques to improve its efficiency for the calculation of (very) large molecules have been explored, but will not be used in this work. The reader is referred to Refs. [211–213] for more details on this subject.

4.1.1 Vibrational free energy in the harmonic approximation

The energy surface which a molecule explores at finite temperature is that shaped by the internal energy, the entropy and the temperature, i.e. the free energy surface. This is the quantity of fundamental interest for comparison with experiments and the one that rules all dynamics of the system. We here focus on the Helmholtz free energy, for reasons already explained in Section 2.3. With respect to the partition function $Z(T)$ of a system in the canonical ensemble, the Helmholtz free energy can be written as[200]:

$$F(T) = -k_B T \ln[Z(T)], \tag{4.18}$$

where k_B is the Boltzmann constant and T is the temperature.

Formally the free energy of a molecule with many atoms depends on a partition function that is composed of several (not necessarily separable) terms, corresponding to the number of configurations, translations, rotations, and vibrations available, and to the electronic degrees of freedom. As long as the Born-Oppenheimer approximation is assumed, the contribution from electronic excited states becomes clearly separable and is usually negligible for the free energies of systems of non-degenerate ground-states and if the first electronic excited state lies electron-volts away from the ground state (see ref. [200], chapter 5). If this assumption holds, the remaining problem is calculating the vibrational, rotational and configurational components. We here focus only on the vibration contributions to the free energy because: (i) any energy term coming from the nuclear translational wave function will depend only on the mass of the system [200], such that for energy differences between distinct conformers of the same molecule, they will always cancel; (ii) the rotational motions will have energy levels that depend on the shape of the molecule through the moments of inertia (see, e.g. Refs. [200, 203]). For the molecules studied in this thesis, this term is very small, representing only 5 meV of the energy differences.

Here, we focus on the vibrational contributions. Knowing the frequencies (energies) of vibrations in the harmonic approximation and assuming that the system obeys Bose-Einstein statistics and can be described by Boltzmann weights, the partition function can be written as a product of the several Boltzmann-weighted vibrational energy levels:

$$Z_{vib}(T) = \prod_{i=1}^{3N-6} \frac{e^{-\frac{\hbar\omega_i}{2k_BT}}}{1-e^{-\frac{\hbar\omega_i}{k_BT}}}, \tag{4.19}$$

where ω_i are the normal modes of vibration of the molecule and the product runs over all modes except the ones corresponding to translations and rotations [3]. Substituting 4.19 in 4.18, we get the following expression for the harmonic free energy:

$$F_{vib}(T) = V_{BO} + \sum_{i=1}^{3N-6}\left[\frac{\hbar\omega_i}{2} + k_BT\ln\left(1-\exp^{-\frac{\hbar\omega_i}{k_BT}}\right)\right]. \tag{4.20}$$

where $\hbar\omega_i/2$ term in equation 4.20 is called the zero-point energy of vibration, because it contributes even when $T=0$ K.

The vibrational contribution to the internal energy U, in the harmonic approximation, can also be calculated from $Z_{vib}(T)$, ($U = -\frac{\partial \ln Z(T)}{\partial\beta}$) and is given by:

$$U_{vib}(T) = \sum_{i}^{3N-6}\left[\frac{\hbar\omega_i}{2} + \frac{\hbar\omega_i}{\exp^{\frac{\hbar\omega_i}{k_BT}}-1}\right], \tag{4.21}$$

such that there is a temperature dependence of the internal energy that differs from the classical picture, where it is only linearly dependent on the temperature, by the energy equipartition theorem.

4.2 *Ab initio* molecular dynamics

The method of Born-Oppenheimer molecular dynamics assumes classical Newton mechanics for the movement of the nuclei, subject to the BO potential-energy. The classical Hamiltonian $\mathcal{H}$ and equations of motion for the nuclei are

$$\mathcal{H}(\vec{R},\vec{p}) = \sum_I \frac{p_I^2}{2M_I} + V_{BO}(\vec{R}) \tag{4.22}$$

$$\frac{d\vec{R}_I}{dt} = \frac{\partial\mathcal{H}}{\partial\vec{p}_I} \tag{4.23}$$

$$\frac{d\vec{p}_I}{dt} = -\frac{\partial\mathcal{H}}{\partial\vec{R}_I} = -\frac{\partial V_{BO}(\vec{R})}{\partial\vec{R}_I} = \vec{F}_I(\vec{R}_I) \tag{4.24}$$

$$\vec{F}_I = M_I\frac{d^2\vec{R}_I}{dt^2}, \tag{4.25}$$

[3] The harmonic approximation for the partition function presents inaccuracies at high temperatures due to the harmonic energy levels, as mentioned previously, but also at low temperatures. The partition function cannot be separated in a product of single components (failure of Boltzmann premise) and its derivation becomes very involved (see Ref. [200], Chapter 4).

where t is the time, $\vec{p}_I$ is the classical moment of the nuclei, $\vec{R}_I$ the coordinates of the nuclei, M_I the masses, $\vec{F}_I$ the forces, and V_{BO} the Born-Oppenheimer potential, given by solving the electronic-structure method of choice (DFT, HF, etc.). The total energy of the system is defined as the value of the Hamiltonian for specific values of the coordinates and momenta. This Hamiltonian gives rise to the classical Newton equation of motion given in 4.25, such that the forces are calculated as the negative gradient of $V_{BO}(\vec{r};\vec{R})$. The nuclei are thus treated classically, and for each new configuration of the nuclei the electronic density is converged self-consistently, and the forces are evaluated. An alternative (and popular) type of *ab initio* molecular dynamics (AIMD), not used in this work, is known as Car-Parrinello molecular dynamics [214]. In this case, the Hamiltonian/Langrangian that describes the system is extended by an extra term that takes into account the electronic degrees of freedom. The equations of motion of the ions and the electrons are then coupled, leading to advantages that will be explained below (no energy drift, even if self consistency is not achieved, which allows for computationally faster simulations).

The fact that Newton's equation of motion are used presumes that the system can be treated classically. Assuming one wants to describe the position of the particle at time $t+\Delta t$, where Δt is a small quantity, the position $r(t+\Delta t)$ can be expanded in a Taylor series of the form:

$$r(t+\Delta t) = r(t) + \underbrace{\frac{d}{dt}r(t)}_{v(t)}\Delta t + \frac{1}{2}\underbrace{\frac{d^2}{dt^2}r(t)}_{F(t)/m}\Delta t^2 + \frac{1}{6}\frac{d^3}{dt^3}r(t)\Delta t^3 + O(t^4) + ..., \tag{4.26}$$

where $v(t)$ is the velocity of the particle. Provided that the initial conditions at $t=0$ for the velocities and forces are known, in principle Eq. 4.26 could be truncated after second order and the positions at time $t+\Delta t$ could be readily calculated. This is known as the Euler algorithm, and it is not commonly used because it leads to large inaccuracies, since the error is $O(t^3)$. A better approximation is obtained by writing this Taylor expansion for $r(t+\Delta t)$ and $r(t-\Delta t)$ and summing them, obtaining:

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \tfrac{1}{2}\tfrac{F(t)}{m}\Delta t^2 + \tfrac{1}{6}\tfrac{d^3}{dt^3}r(t)\Delta t^3 + O(t^4) + ... \tag{4.27}$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \tfrac{1}{2}\tfrac{F(t)}{m}\Delta t^2 - \tfrac{1}{6}\tfrac{d^3}{dt^3}r(t)\Delta t^3 + O(t^4) + ... \tag{4.28}$$

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \tfrac{F(t)}{2m}\Delta t^2 + O(t^4) + ... \Rightarrow \tag{4.29}$$

$$\Rightarrow r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \tfrac{F(t)}{2m}\Delta t^2 + O(t^4) \approx 2r(t) - r(t-\Delta t) + \tfrac{F(t)}{2m}\Delta t^2 \tag{4.30}$$

The truncation of the expansion in Eq. 4.30 after second order is what is known as the Verlet algorithm. Knowing the forces at time t it is possible to calculate the positions at time $t+\Delta t$. The remaining error is now only of the fourth order in the positions because the odd terms of the expansion cancel, rendering the Verlet algorithm rather accurate in relation to its computational cost. [4]. However, the knowledge of the positions at $t-\Delta t$ is required, and this information is not available for $t=0$. This is what motivates the use of the so-called Velocity-Verlet algorithm. In this algorithm the positions are calculated as in the Euler algorithm but the velocities are calculated as:

$$v(t+\Delta t) = v(t) + \frac{F(t+\Delta t)+F(t)}{2m}\Delta t. \tag{4.31}$$

The Velocity-Verlet algorithm of Eq. 4.31 yields the same expression for the update of the positions as the Verlet algorithm (Eq. 4.30) [201]. However, in this algorithm only the knowledge of the velocities at

[4] The Verlet algorithm also has the advantage of being a *symplectic* algorithm, i.e. the "coordinate transformation" given by the time evolution is such that it is canonical (its Jacobian is 1), and thus the volume element in phase space (p, r) does not change during the time evolution.[201]

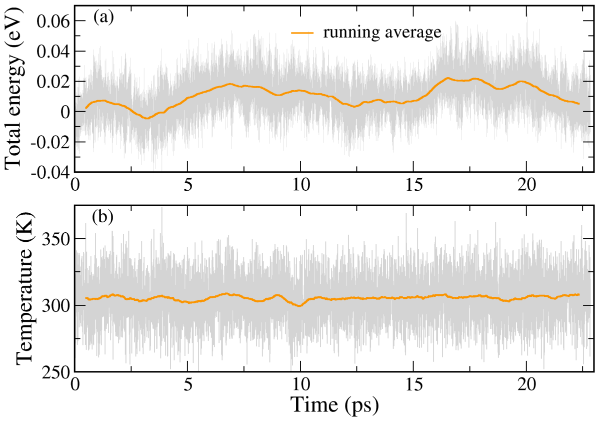


Figure 4.2: Microcanonical simulation of Ac-Ala$_{10}$-LysH$^+$ (130 atoms) using BO-MD as implemented in FHI-aims, with 1fs time step. Panel (a) shows the total energy at each time step of the simulation with its running average (over 100 fs). Panel (b) shows the temperature at each time step. Since the simulation started from a previous one thermalized at 300K, this is approximately the value the temperature assumes through the simulation, but with large oscillations.

time t and forces at times t and $t + \Delta t$ are required, which is available at all points of the simulation.

There are two important points that have to be taken into consideration for performing an accurate AIMD simulation: the size of the time step, that determines the accuracy of the integration, and the quality of the self consistency convergence, that determines the quality of the calculated forces.

Too large time steps can lead to inaccuracies so great in the integration that the molecule falls apart after just a few MD steps. The maximum Δt value possible to be used for a system depends on the smallest period of vibration, so that molecules containing light atoms require smaller time steps, and molecules containing heavier atoms can tolerate larger time steps. The time step used in this thesis is of $\Delta t = 1fs$. Tests for the accuracy of this time step can be found in Appendix D, for the NH$_3$ molecule. $\Delta t = 1fs$ proved to induce small enough energy oscillations such that the simulation is stable. As an example for the molecules studied in this thesis, Figure 4.2(a) shows the total energy at each time step of the simulation of Ac-Ala$_{10}$-LysH$^+$ (130 atoms). Moreover, this time step size is small enough so that specific properties of the system interesting for this work (e.g. IR spectra, discussed in the next section) can be successfully evaluated. Smaller fluctuations of the energy than the one seen in Figure 4.2 can be achieved either using smaller time steps or higher order algorithms [215, 216], that consider more terms in the Taylor expansion of the positions. Different integration algorithms with different time steps are also tested in Appendix D. The higher order algorithms are seen to require more force-evaluations, so that the calculations become much more expensive for a modest gain in accuracy. The Verlet algorithm turns-out to be the most efficient in the region of interest (see Figure D.4).

Since in Born-Oppenheimer molecular dynamics the forces are calculated as the gradient of the BO-potential at electronic self-consistency, the accuracy of the convergence of the self-consistent cycle is of extreme importance. If the convergence is not satisfactory, the calculated forces will not be consistent with the BO surface, and there will be a progressive drift of the energy with time of simulation [217]. A small drift will always happen due to the numeric nature of the simulation, but this drift should not exceed a few meV during the whole time of simulation, lest unphysical effects happen. Tests for the convergence parameters can be found in Appendix D. From those tests, it was found that for the simulation not to exhibit a substantial energy drift, the electronic density has to be converged up to 10^{-5} electrons[5], the eigenvalues to 10^{-4} eV, and the total energy to 10^{-6} eV. This produces the drift exemplified in Figure 4.2(a) for a ≈23 ps simulation of Ac-Ala$_{10}$-LysH$^+$ (130 atoms). Although the energy oscillates with ≈30 meV amplitude, the average drift in the whole trajectory does not considerably exceed 10 meV. The drift can also be mitigated by extrapolating the wave function from the previous MD step to the next one, for

[5] Convergence, in the self-consistency cycle, is measured as the change in the input quantity (e.g. density). and the one obtained after solving the KS or HF equations.

which several schemes have been proposed in the literature [217–219].

Molecular dynamics is generally used in connection with statistical mechanics, which allows to connect the microscopic details of a system with physical observables (thermodynamic properties, diffusion coefficients, spectra). This is possible because the states sampled at each time step of a MD simulation belong to the same ensemble, having, thus, the same partition function. By evaluating averages in this ensemble, one can connect to the thermodynamic properties. In particular, through the calculation of auto-correlation functions, it is possible to obtain vibrational spectra with anharmonic contributions, as has been explained in Section 4.3. By solving the Hamiltonian of Eq. 4.22, the energy and momenta of the system are conserved, thus defining a simulation in the microcanonical ensemble (also the number of particles do not change and the volume, if possible to define one, should not change). Other statistical ensembles are usually characterized by fixed values of other thermodynamical variables: e.g. canonical (NVT), isothermal-isobaric (NPT), and grand canonical (μVT, where μ is the chemical potential). In order to simulate these other ensembles, the Hamiltonian of the isolated system has to be brought into contact with a reservoir. In the next section, simulations in the canonical ensemble will be discussed.

It is worth noting, as well, that MD can be connected with other schemes (not used in this thesis), in order to calculate free energy differences [201], as e.g. replica exchange (parallel tempering), transition path sampling, or thermodynamic integration. These schemes require a lot of sampling of the conformational space, which motivated techniques that enhance the sampling speed to be developed.

Additionally, the full quantum-mechanical nature of the nuclei can be accessed by performing path-integral molecular dynamics [220, 221]. These effects are particularly important when hydrogen atoms (protons) play a big role, for example in the description of proton transfer reactions. With increasing temperature, anharmonic effects will dominate over the quantum ones, and the quantum nature of the nuclei will not be treated in this thesis. That said, quantum effects at room temperature are apparently quantitatively important for H-bonded molecules and water, even if their effect is not so drastic. These effects seem to be necessary, for example, to reproduce the correct radial distribution functions of water [221], as well as its heat capacity at room temperature [222]. With respect to H-bonds, a recent study by Li, Walker, and Michaelides [223], has shown that the quantum nuclear effects slightly strengthen strong H-bonds and weaken weak ones [6]. Therefore, we cannot expect to be exactly quantitative with respect to the temperature-dependent stability of molecules presented here. For physiological environements, where T changes of around 8 K may represent the difference between life and death, the correct inclusion of these effects could be important. This will be the subject of future work in our group. Nevertheless, the molecular dynamics simulations presented here differ from the usual force-field treatment ("molecular mechanics") for polypeptides, in the fact that the electronic contributions to the Born-Oppenheimer potential-energy are treated quantum-mechanically.

4.2.1 Thermostats

Most "real-life" experiments cannot be done in the micro canonical ensemble, i.e. one where the energy is kept strictly constant. Instead, the temperature and/or pressure of a system are usually the variables that can be controlled. In order to simulate such a system, one has to bring the system into thermal contact with a heat bath. From a statistical mechanics point of view, in the canonical ensemble the energy equipartition holds. Equipartition means that the kinetic-energy is equally distributed for each degree of freedom of the system, assuming that they can be treated independently. If this is the case,

[6] The differences in equilibrium geometry of the "strong" H-bonds are reported to be of around 0.04 Å. For the α-helices studied here, these bonds should become slightly stronger.

the momenta $\vec{p}_I = M_I\vec{v}_I$ follow the Maxwell-Boltzmann (MB) distribution:

$$f(p_I) = \left(\frac{\beta}{2\pi M_I}\right)^{3/2} \exp^{\left(-\beta p_I^2/(2M_I)\right)}, \tag{4.32}$$

where $\beta = 1/k_BT$. The instantaneous temperature is given by the relation $T = \frac{2E_{kin}}{3Nk_B}$, where N is the number atoms, and $E_{kin} = \sum_I M_I\vec{v}_I^2$ the kinetic-energy of the system (previously called T, but here changed to avoid confusion with the temperature). Since E_{kin} depends on the instantaneous velocities of each particle at each time t, the temperature is not strictly constant but can (and should) fluctuate around the average value. Assuming the central limit theorem, the kinetic-energy distribution (over the time of simulation) will be approximately Gaussian, and thus, the distribution of the temperature fluctuations can be written as:

$$P(T - T_0) = C \exp^{-\frac{(T-T_0)^2}{2\sigma^2}}, \tag{4.33}$$

where $\sigma^2 = \frac{2T^2}{3N}$, with N is the number of atoms, and C is a constant.

The coupling of the system with a heat bath is achieved through the use of so-called "thermostats" - modifications of the Hamiltonian Eq. 4.22, that enforce consistency with a given ensemble. An early proposed thermostat was the Andersen thermostat [201, 224], which makes use of stochastic processes to bring the velocities to sample the canonical ensemble. This thermostat disrupts the trajectory of the molecules, such that dynamical quantities (which will be important for this thesis, as e.g. autocorrelation functions) are not reliable. Simple velocity-rescaling thermostats (e.g. Berendsen [201, 225]), have also been proposed, but if done naively, the system does not sample the canonical ensemble [226, 227], rendering the trajectories unreliable.

The thermostat used in the constant-temperature simulations presented in Chapter 10 is the Nosé-Hoover thermostat, which belongs to a class of thermostats called the "extended Lagrangian". The idea (like in Car-Parrinello molecular dynamics) is to add fictitious degrees of freedom, such that the overall total energy is conserved but the atomic subsystem can span ensembles other than microcanonical. The Lagrangian proposed by Nosé [228] with the equations proposed by Hoover [229] read as follows

$$\dot{\vec{R}}_I = \vec{p}_I/M_I \tag{4.34}$$

$$\dot{\vec{p}}_I = -\frac{\partial V\left(\vec{R}\right)}{\partial \vec{R}_I} - \frac{\Pi\vec{p}_I}{Q} \tag{4.35}$$

$$\dot{\eta} = \frac{\Pi}{Q} \tag{4.36}$$

$$\dot{\Pi} = \left(\sum_I \frac{\vec{p}_I^2}{M_I} - \frac{g}{\beta}\right) \tag{4.37}$$

where g is the number of degrees of freedom of the system, η is a fictitious extra coordinate, Π is its conjugated momentum, Q a variable that can be related to a fictitious so-called thermostat mass (it can be understood as the mass of an harmonic oscillator coupled to the system), and the dot denotes a time derivative. The conjugated momentum Π of the extra coordinate η acts as a fluctuating drag parameter to the atomic subsystem. The conserved energy associated with the equations of motion is:

$$\mathscr{E} = \sum_I \frac{\vec{p}_I^2}{2M_I} + V\left(\vec{R}\right) + \frac{1}{2}\frac{\Pi^2}{Q} + g\frac{\eta}{\beta}. \tag{4.38}$$

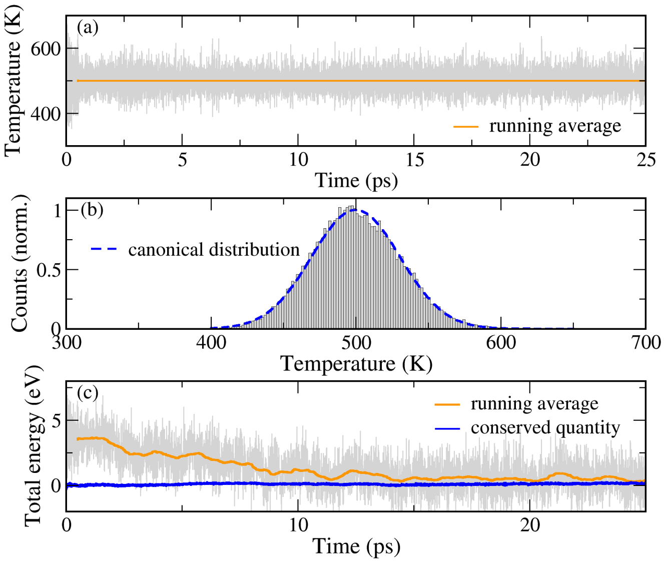


Figure 4.3: Constant T simulation of Ac-Ala$_{15}$-LysH^{+} (180 atoms) using the Nosé-Hoover thermostat, with 1fs time step, target temperature 500K, thermostat mass $Q = 1700$cm^{-1}. Panel (a) shows the temperature at each time step of the simulation with its running average. Panel (b) shows the temperature distribution compared to the ideal canonical distribution (Eq. 4.33). Panel (c) shows the conserved quantity of the thermostat (Eq. 4.38), where both quantities had their average value shifted to zero for better visualization (in reality they do not converge to the same value).

It is necessary to choose a value for the "thermostat mass" Q such that the coupling with the system is effective. Since the thermostat can be regarded as a harmonic oscillator with mass Q, the value of this variable should be chosen so that there is some overlap between the frequencies of vibration of the system (preferably delocalized ones) and the thermostat frequency [230]. In this way, the thermostat can couple to the movement of the atoms, such that the desired temperature is efficiently achieved. If the thermostat couples only to a very localized and harmonic vibrational mode of the system, the trajectory becomes biased by the initial conditions, and the system gets trapped in a small region of the phase space (i.e. the trajectory is non-ergodic). Applying a thermostat to the system and getting the system to the desired temperature is often referred to as "thermalization". For systems that are large and anharmonic enough, the thermalization is rapidly achieved and the trajectory is ergodic (i.e. does not depend on the initial conditions). For very harmonic or very small systems alternative methods like the Nosé-Hoover chains, i.e. a series of coupled Nos-Hoover thermostats, have to be used [201, 230].

Since this thermostat was widely used in this work, in Figure 4.3 a thermostatted simulation of Ac-Ala$_{15}$-LysH^{+} (introduced in Chapter 1) is shown. The time step used was 1fs, the target temperature 500K, and the thermostat mass Q was 1700cm^{-1} (amide-I region). One can see that the temperature distribution is centered on the target temperature, it is consistent with the canonical ensemble, and the conserved quantity of Eq. 4.38 is conserved within certain accuracy limits.

It is worth to mention, as well, a recently developed thermostat, proposed by Bussi, Donadio, and Parrinello [231]. This thermostat goes beyond the thermostats mentioned so far, such that the target temperature follows a differential equation that involves a velocity rescaling term and a (time-dependent) stochastic white noise. This thermostat formally samples the canonical ensemble, has a conserved quantity ("effective energy"), does not suffer from ergodicity problems and seems to maintain the time correlation of the system [232]. Although very promising, this thermostat is very recent, and its capabilities and limitations are still being explored.

4.3 IR spectra from spectral functions

A way to go beyond the harmonic approximation on the calculation of the vibrational spectrum of a molecule is through the evaluation of time auto-correlation functions of the form $C(t) = \langle A(0)A(t)\rangle_t$, where $A(t)$ is the quantity of interest (see Appendix C for details on these functions). In this approach, finite temperature effects and the various configurations a molecule adopts over a finite period of time are also naturally incorporated. This approach also brings the advantage that there is no need to rely on the harmonic approximation, such that IR spectra naturally include anharmonic effects. On the other hand, it is generally assumed that the nuclei can be treated classically. Recent studies show that this formalism can also be applied for quantum nuclei, although many technical issues still need to be solved [233–235]. As will be seen later in this chapter, the inclusion of quantum effects for the nuclei would automatically fulfill the quantum detailed balance relation for the lineshape of spectra, desymmetrizing the peaks and changing the intensities. Peak shifts are also expected to happen, but they are of a smaller magnitude than the shifts due to anharmonicities at finite temperatures. In fact, quantum effects become more important at lower temperatures, since equipartition becomes strongly unfulfilled for the internal normal modes of the molecules. In this thesis, the nuclei will only be treated classically, and a quantum correction factor will be applied to account for the detailed balance.

A few relevant steps of the derivation follows, but a much more detailed derivation can be found in the book by McQuarrie [200]. Consider the time dependent Schrödinger equation

$$\hat{H}\psi(\vec{r},t,\vec{R}) = i\hbar\frac{\partial}{\partial t}\psi(\vec{r},t,\vec{R}), \tag{4.39}$$

where H is the total Hamiltonian of the system, $\hbar$ is the reduced Planck constant, and Ψ is the wavefunction of the system. Equation 4.39 has as a formal solution:

$$\psi(\vec{r},t;\vec{R}) = e^{-\frac{iEt}{\hbar}}\psi(\vec{r};\vec{R}), \tag{4.40}$$

where E refers to the eigenvalue energies of the system obtained from the time-independent Schrödinger equation.

Since we are interested in the changes of the dipole moment of the system, it is convenient to consider an external electromagnetic field and treat it as a perturbation to the original Hamiltonian. Since spin does not play a role for the vibrations of the system (only the spatial coordinates do), we are left only with the electric field. We can then write the Hamiltonian of the system as an unperturbed one ($\hat{H}_0$) plus a perturbation ($\hat{H}_1$) given by a weak electric field [200],

$$\hat{H} = \hat{H}_0 + \hat{H}_1 = \hat{H}_0 - \hat{\mu}\cdot\vec{\epsilon}, \tag{4.41}$$

where $\hat{\mu}$ the dipole moment and $\vec{\epsilon}$ the electric field, assumed to be a plane wave of the form $\epsilon = \vec{\epsilon_0}(e^{i\omega t} + e^{-i\omega t})$.

In this context, we can apply Fermi's golden rule for the perturbation $-\hat{\mu}\cdot\vec{\epsilon}$ in order to find the transition probability between two states of the system, under the constrain that the transitions happen in resonances with the external field. The intensity will then be proportional to the transition probability w_{fi} (taken here as a combination of the absorption and emission probabilities):

$$w_{fi} = \frac{2\pi}{\hbar}|\langle f|\hat{\mu}\cdot\vec{\epsilon}|i\rangle|^2\{\delta(E_f - E_i - \hbar\omega) + \delta(E_f - E_i + \hbar\omega)\} \tag{4.42}$$

$$w_{fi} = \frac{2\pi}{\hbar}\langle i|\hat{\mu}\cdot\vec{\epsilon}|f\rangle\langle f|\hat{\mu}\cdot\hat{\epsilon}|i\rangle\{\delta(E_f - E_i - \hbar\omega) + \delta(E_f - E_i + \hbar\omega)\}, \tag{4.43}$$

where f and i denote the final and initial states, respectively, and E_f corresponds to the energy associated with the final state of the system, as given by the unperturbed Hamiltonian, while E_i corresponds to the initial one, and ω is the frequency of oscillation of the external field ($\hbar\omega$ is the energy of the photon). In a statistical ensemble framework, the rate of energy loss from the radiation (E_{rad}) to the system can be written as:

$$-\frac{d}{dt}E_{rad} = \sum_i \sum_f p_i \hbar\omega w_{fi}, \tag{4.44}$$

where, assuming Boltzmann statistics, $p_i = e^{-\beta E_i}/Z$ are Boltzmann statistical weights, with $\beta = 1/(k_B T)$ and Z the partition function, and ω the frequency related with the energy loss from going to state i to state f.

Upon manipulation of 4.44 (see the book by McQuarrie [200] for a detailed derivation) and the consideration of the Fourier transform representation of the δ function, it is possible to show that the absorption cross-section that gives the line shape for the absorption of radiation by the system is proportional to

$$I(\omega) \propto \sum_i p_i(1 - e^{-\beta\hbar\omega})\omega \int_{-\infty}^{\infty} e^{-i\omega t}\langle i|\hat{\mu}(0)\cdot\hat{\mu}(t)|i\rangle dt, \tag{4.45}$$

where only the factors depending on ω were explicitly written.

In order to arrive at the equation used for the actual calculations, one more point needs to be taken into account, and that is that since the system should be ergodic, according to equation C.3 the ensemble average can be written as a time average of the *classical* autocorrelation function, and the final expression becomes [7]:

$$I(\omega) \propto (1 - e^{-\beta\hbar\omega})\omega \underbrace{\int_{0}^{\infty} e^{-i\omega t}\langle \mu(0)\cdot\mu(t)\rangle dt}_{G(\omega)}. \tag{4.46}$$

Eq. 4.46 final result connecting the intensity of vibrations active in the IR with the Fourier transform of the dipole auto correlation function. This formalism gives a way to estimate the frequencies of vibration of a molecule without relying in any way on the harmonic approximation and with temperature and dynamical effects automatically included by the dynamical simulation required to evaluate the dipole autocorrelation function. It is, therefore, a powerful method to estimate IR spectra at finite temperatures including anharmonic effects. The drawback, if compared to the harmonic approximation, introduced previously, is that the dipole autocorrelation integral $G(\omega)$, though, is evaluated classically, for classic nuclei, by means of AIMD, such that the quantum distribution of the nuclear wave-function is disregarded. Auto-correlation functions in the realm of quantum mechanics are defined by the Kubo transform (see Refs. [234, 236] for a detailed discussion and derivation of these quantities). The quantum nature of the “oscillators” (pairs of atoms) in the molecule would cause a desymmetrization of the peaks [237, 238], due to the necessity to obey the principle of detailed balance [$G(-\omega) = \exp(-\beta\hbar\omega)G(\omega)$]. In Refs. [237, 238] this effect is extensively studied and several corrections to the line-shape are proposed. The one that is known as the quantum harmonic correction [238, 239], of the form $D(\omega) \propto \omega/(1 - e^{-\beta\hbar\omega})$, has been shown to give the best agreement with existing experimental data [238–241] [8]. After multiplying into 4.46,

[7] In Eq. 4.46, it was taken into account that the autocorrelation function is real, which allows to change the lower integration limit to 0, upon multiplication of a factor 2 (not explicitly written).

[8] Multiplying by this factor is actually equivalent to performing a Kubo-transform on the autocorrelation function, although the function itself is still a classic object here.[236]

this factor produces:

$$\tilde{I}(\omega) \propto \omega^2 \int_0^\infty e^{-i\omega t} \langle \mu(0) \cdot \mu(t) \rangle dt, \tag{4.47}$$

which is the equation for the calculation of the spectra shown in this work. The evaluation of the IR spectra in this way allows the simulation of non-fundamental vibrations (and anharmonicities), which goes beyond the harmonic picture. An example of the spectrum of formamide ($HCONH_2$) obtained with this formalism, from an NVE AIMD run with $\langle T \rangle$=300 K, compared to the harmonic approximation, is shown in Figure 4.4. In Appendix D, an example with the ammonia molecule (NH_3) is shown, where peaks corresponding to non-fundamental vibrations (overtones and combinations) appear.

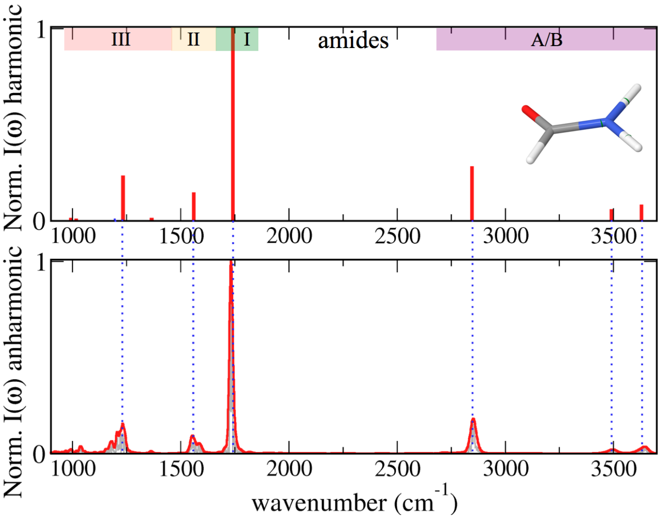


Figure 4.4: Gas-phase spectrum of the formamide molecule. Top: harmonic approximation (DFT-PBE). Bottom: from AIMD dipole autocorrelation, with $\langle T \rangle$=300 K - grey lines corresponds to the raw output, red line corresponds to a convolution with a very narrow Gaussian function, for better visualization. Blue dotted lines serve just as a guide to the eye, and are centered at the harmonic frequencies of vibration. Amides I, II, III, and A/B are marked on the top.

This method has been applied successfully in the literature, in connection with *ab initio* methods, (see, for example Refs. [36, 239–246]), in order to calculate IR spectra of molecules and liquids. This will be the method also used in this work, in order to obtain spectra with anharmonic contributions. Even though this approximation disregards the quantum-mechanical nature of the nuclei, it averages over the anharmonicity in the nuclear trajectory at finite temperatures, which gives the correct direction of the anharmonic shifts and better frequencies, even if the expectation value of the shift is now computed classically.

4.4 Spectroscopy as a tool to identify peptide structures

In order to experimentally characterize the structural motifs of proteins, one needs to probe the interactions between the atoms. Vibrational spectroscopy, for which theoretical methods have been outlined in the previous sections, is an appealing method to probe the structure of the molecules, since the vibrational modes depend on the composition, geometry, and chemical bonds of the molecule.

Vibrational modes can be excited by absorption of photons in the infrared (IR) range of the electromagnetic spectrum, i.e. wavelengths ranging from 2.5 μm to 250 μm (5 meV to 0.5 eV). In IR spectroscopy the most commonly used measure is the inverse of the wavelength, i.e. the wavenumber ($\tilde{\nu} = 1/\lambda$). This measure is directly proportional to the energy and the frequency, and the most common unit used is cm^{-1}. The IR range in these units goes from 40 to 4000 cm^{-1}.

Localized vibrations in a molecule (involving a few atoms) can be of different types, and are classified as following:

- Stretching vibrations (antisymmetric and symmetric): changes in bond lengths
- Bending vibrations (scissoring, rocking): “in-plane” change in bond angles
- Wagging and twisting vibrations (also umbrella): “out-of-plane” change in bond angles

In proteins and polypeptides the vibrations of the amide group are of particular interest. The amide group is composed by a carbonyl group connected to a nitrogen atom, generally represented as RC(O)NHR′ (where R and R′ re generic groups), being the structure formed by the peptide bond. A very simple molecule containing this group is formamide (HCONH$_2$), for which the theoretical IR-spectrum, obtained from the harmonic approximation and from AIMD-derived dipole autocorrelation function, was shown in Figure 4.4. For an experimentally measured spectra, one can consult the NIST database[9] or Ref. [247], where the spectrum appears as continuous lines, with broadened peaks. The broadening of the peaks in an experiment can be influenced by many effects, which may reflect a physical feature of the system (very flexible structures, more than one conformation, modes coupling, etc.), or can be an artifact of the experimental setup (e.g. non-linear absorption of photons, width of the laser, etc.). These broadening mechanisms (related to temperature, conformational freedom, and the experimental technique to some extent), render measured spectra often continuous lines.

The vibrational normal modes ($\vec{Q}_k$) of formamide are shown in Figure 4.5. One thing that becomes clear from plotting the normal modes is that although a mode is named by its “main character”, even in this small molecule, the vibration is never completely decoupled from the vibration of other atoms.

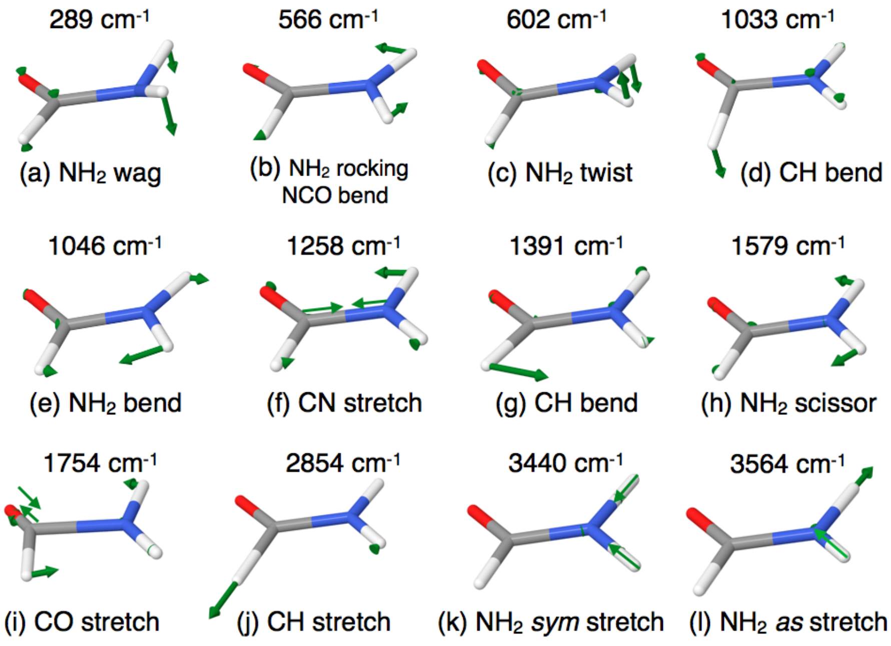


Figure 4.5: Vibrational modes of the formamide (CH$_3$NO) molecule, calculated with the harmonic approximation, using density-functional theory and the PBE functional (see Chapter 3 for details on these methods).

[9] http://webbook.nist.gov/cgi/cbook.cgi?ID=C75127&Type=IR-SPEC&Index=0#IR-SPEC

The spectrum shown in Figure 4.4 shows characteristic structures of the amide group. There is the most intense peak, at around 1700cm^{-1}, which is called the amide-I band. The peak close to it, to the left, around 1500cm^{-1} is called the amide-II band. Then, there is a group of structures between 900-1300cm^{-1}, which are named amide-III band. Finally, there is a gap between 1800cm^{-1} and 2800cm^{-1}, and above 2800cm^{-1} again structures can be found, which are called amide-A/-B bands. These bands can be connected to specific atoms, by analysing the normal modes shown in Figure 4.5. In polypeptides and proteins exactly these peaks (or peak-families) are seen, although the average positions are shifted, since the atoms (or groups) connecting to the amide group are usually heavier than hydrogen, and the interactions between the various atoms change the force-constants related to the vibrations. For polypeptide chains, the characteristic regions and the vibrations involved in each one of them are the following [248]:

- Amide-A and -B($\approx$ 3170 - 3300 cm^{-1}): Localized N-H stretching vibration modes. The band positions in this range are not very sensitive to the backbone conformation but are very sensitive to H-bonds.
- Amide-I mode ($\approx$ 1650 cm^{-1}): collective C=O stretching vibrations coupled to out-of-phase C-N stretching vibrations, the C-C-N deformations, and in-plane N-H bending vibrations. The position of these vibrations are more sensitive to backbone conformations.
- Amide-II mode ($\approx$ 1550 cm^{-1}): collective out-of-phase in-plane N-H bending vibrations, also with contributions from C-N stretching vibrations. Also sensitive to backbone conformation but harder to correlate.
- Amide-III mode ($\approx$ 1200 - 1400 cm^{-1}): Collective and localized in-phase N-H and C-H bending vibrations and C-N stretching vibrations. High structure sensitivity, even upon small conformational changes[249, 250].

Vibrations can bring a significant amount of information about the character and the geometry of molecules [248, 251]. The chemical composition can be probed by exploring the relation between the frequency of vibrations with the mass of the atoms. By inducing protonation/deprotonation, or by performing isotopic substitutions, peak shifts that can be assigned to specific structures are observed. The bond lengths and strengths can be related to the force-constants connecting the atoms, allowing, again, for bond-strength/length assignment depending on shifts of the peaks. For the specific example of Hydrogen bonding, the frequency of vibration for a free CO is higher than for a H-bonded CO. This is illustrated by the fact that the (free) CO stretch vibration of the formamide is reported to be at 1754cm^{-1}, and the H-bonded CO vibrations in polypeptides (in the amide-I peak), are found at around 1650cm^{-1}. The amide-II (NH bends) peak, on the other hand, has a tendency to shift to higher frequencies, although the shift for this peak is not so pronounced. The coupling between two or more frequencies of vibrations also induces relative shifts and/or peak splittings. This is the mechanism that renders the amide-I/amide-II peaks sensitive to backbone secondary structure: each structure will present different coupling strengths of the collective vibrations, thus shifting these peaks. Lastly, the broadening of the peaks can also give information about the conformational freedom of a molecule at a certain temperature, with broader peaks indicating more flexible structures.

Techniques used to measure vibrational spectra will be briefly described in Chapter 6, in connection to an overview of experimental and theoretical work that have been performed on alanine-based polypeptides, the central application-system of this thesis.

Chapter 5

Implementation details

5.1 Solving, computationally, the electronic structure methods

The code most (and almost solely) used throughout this thesis is the Fritz Haber Institute "*ab initio* molecular simulations" (FHI-aims) computer program package[1]. This code is actively developed in the Theory Department of the Fritz Haber Institute and the author of this thesis is one of its developers, having programmed several GGA functionals, the kinetic energy density and the M06 suite of mGGAs (non self-consistent) in the code [1]. In addition, the author has also been involved in several minor parts of the code, as well as more prominent parts related to the molecular dynamics. FHI-aims is an all-electron, localized numeric atom-centerd orbital (NAO) basis code that is able to perform cluster and periodic calculations with DFT and HF-based methods on the same footing. Of course, there are many other packages that implement the basics of DFT and beyond in various different ways (e.g. Gaussian03 [148], NWChem [252], VASP [253], Siesta [254], and many others). While FHI-aims provided the functionality needed for the present thesis in an efficient and sufficient way, others that were used for comparison data (high level QC methods) include, for example Gaussian03. In other codes, other choices (besides NAO) for the basis sets are used, for example: gaussian orbitals, that have an analytical gaussian form and are also centered in each atom, or plane waves that span the whole space.

It is a common practice in electronic structure codes to expand the single particle orbitals ϕ_l into basis functions that have a known "form", in the following way:

$$\phi_l(\vec{r}) = \sum_{i=1}^{N_b} c_{il}\varphi_i(\vec{r}), \tag{5.1}$$

such that the eigenvalue problem to be solved for DFT or HF becomes discretized into a generalized eigenvalue problem [255]:

$$\sum_j h_{ij}c_{jl} = \epsilon_l \sum_j s_{ij}c_{jl}, \tag{5.2}$$

where $h_{ij} = \langle\varphi_i|\hat{h}|\varphi_j\rangle$ is the matrix element of the Kohn-Sham (or Hartree-Fock) Hamiltonian, and $s_{ij} = \langle\varphi_i|\varphi_j\rangle$ is the overlap matrix element.

As mentioned above, FHI-aims uses numeric atom-centered orbitals (NAOs) as basis sets. These, as the name says, are numeric orbitals centered at each atom composing the system being studied. The

[1] Results from this implementation were published in a collaboration work in Ref. [199].

Dunning "augmented correlation-consistent" (aug-cc-pVNZ) gaussian basis sets, will also be used in this thesis, because: (i) they can be used to compare FHI-aims to reference codes that use Gaussian basis functions, and (ii) explicitly correlated calculations are almost exclusively done in these basis sets. These basis sets are based on configuration interaction and coupled cluster calculations, such that they describe the non-local correlation systematically better as the size of the basis set increases. The V in the name of these basis sets means that they were optimized considering valence orbitals, the p means that there is the addition of the "polarization functions", which are functions with angular momenta higher than the valence orbital of the atom, and NZ stands for the multiple number ($N = D, T, Q, 5, 6$) of functions added to each orbital. The word "augmented" means that diffuse functions are added to describe long-range dispersion interactions.

The NAOs in FHI-aims are atom-centered basis functions of the form:

$$\varphi_i(\vec{r}) = \frac{u_i(r)}{r} Y_{lm}(\Omega), \tag{5.3}$$

where the radial shape of $u_i(r)$ is numerically tabulated and fully flexible, and Y_{lm} denotes the spherical harmonics. For all-electron codes[2], NAOs have some important advantages. One is that by using a minimal NAO basis (consisting of the core and valence functions of spherically symmetric free atoms) that is exact for free atoms, the shape of the orbitals close to the nuclei, where the external potential is deep and dominated by the partially screened nucleus, is automatically well described also for bonded atoms. For a DFT-derived minimal basis used in a DFT calculation, this feature rapidly reduces the so-called basis set superposition error (discussed in detail in the next section) with increasing basis set size. Another advantage is that, since the radial functions can be localized by a confining potential, such a scheme allows for an almost $O(N)$ scaling of the code. The radial functions $u_i(r)$ obey the Schrödinger-like equation given by :

$$\left[-\frac{1}{2}\frac{d^2}{dr^2} + \frac{l(l+1)}{r^2} + v_i(r) + v_{cut}(r) \right] u_i(r) = \epsilon_i u_i(r), \tag{5.4}$$

where $v_i(r)$ is the potential that sets the shape of the radial function and $v_{cut}(r)$ is the confining potential, which in FHI-aims has the following form:

$$v_{cut} = \begin{cases} 0 & r \leq r_{onset} \\ s \cdot \exp(\frac{r_{cut}-r_{onset}}{r-r_{onset}}) \cdot \frac{1}{(r-r_{cut})^2} & r_{onset} < r < r_{cut} \\ \infty & r \geq r_{cut} \end{cases} \tag{5.5}$$

in which s is a global scaling parameter. This confining potential ensures a smooth decay of all basis functions and their derivatives to zero, and the radial functions are evaluated on a dense logarithmic grid [$r(i) = r_0 \exp[(i-1)\alpha], i = 1, ..., N_{log}$], that has the convenient features of being dense close to the nucleus and coarse far away. In the default settings for each atomic species of FHI-aims, the value of r_{onset} is chosen conservatively, so as not to influence significantly the shape of the radial functions. For DFT calculations $r_{onset} \approx 4$ Å gives converged results [1], but for explicitly correlated calculations this value may have to be larger, as will be shown and discussed in Chapter 11. The choice for the value of this parameter is an explicit keyword in the code, so that it can and should be tested by the users explicitly.

Pre-constructed basis sets for all elements of the periodic table are distributed with the code. The

[2] DFT codes that are not "all-electron" often use so-called pseudopotentials or projector formalisms to effectively describe the core electrons, such that only valence electrons are explicitly treated.

strategy to derive the basis sets is explained in Ref. [1]. They are obtained based on DFT-LDA calculations of dimers of each element and represent a hierarchical improvement on the calculated average total energies. A very similar procedure will be followed and explained in Chapter 11, when describing basis-sets developments performed in this work. The basis functions chosen by the optimization procedure come from a defined pool of possible basis functions with two different shapes. The potential $v_i(r)$ in equation 5.4 is set to be either that of the hydrogen atom, with an effective charge, or that of doubly positively charged ionic species. Others can be or are implemented, but the listed classes are sufficient for the purpose of creating a flexible, generic basis set library. These basis functions are organized in *tiers* (ranks, classes). which are ordered by the amount of improvement each basis function brings to the total energy of the dimers, with *tier1* containing the functions that bring the largest improvement, down to *tier4* where the functions that bring the smallest but still noticeable improvements are located. These functions, in the order they appear, come automatically from the optimization procedure.

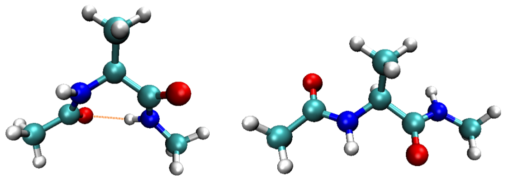
The standard basis sets used for the atoms studied in this thesis (carbon, oxygen, nitrogen, and hydrogen) are given in Appendix B.

The integrations evaluated in FHI-aims are done on a real-space grid composed by overlapping, atom-centered grids described in Refs. [1, 256]. The integrands are localized on top of each atom by use of atom-centered partition functions and each single-atom integrands are then computed in a Lebedev grid [257] of spherical integration shells.

The Hartree potential in FHI-aims is calculated from a multipole decomposition of the electron density. The highest angular momentum for this decomposition can be chosen explicitly, such that all higher angular momenta terms are discarded. For tightly converged calculations, a value of l=6 is used (see Ref. [1] and references therein).

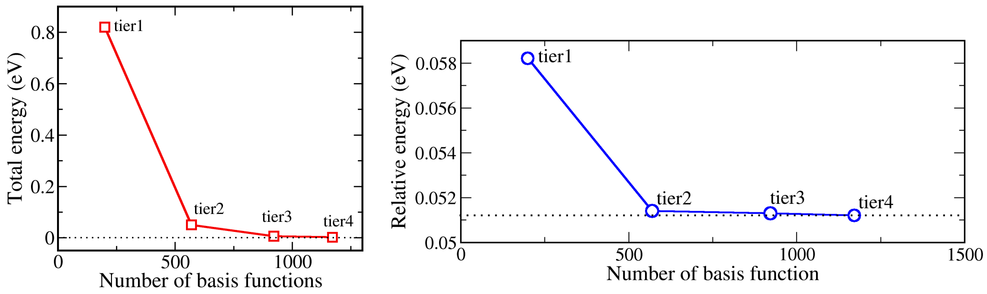
There are two distinct sets of numerical defaults for each atomic species in FHI-aims that will be used in this thesis. They are called *light* and *tight* settings. In the *light* settings, when considering light ($Z = 1 - 10$) elements, the basis sets are taken to be *tier1* and the integration grids are not so dense. They are well suited for pre-relaxations and initial energy and geometry estimates. The *tight* settings employ the *tier2* basis sets for the light elements and dense integration grids, converged hartree-potential multipole expansion, and the cutoff potential onsets at large distances (6 Å). These settings are recommended for obtaining “final” results. For convergence purposes, the specification of the basis set itself (tier 1, tier 2, etc.) may still be increased or decreased as needed. For details of which radial functions are included for each $tier$, for the elements relevant to this thesis, see Appendix B. As an example, the DFT the total energy of an alanine dipeptide conformational minimum (22 atoms), as well as relative energy between two different conformational minima of this molecule (see Figure 5.1) are shown in Figure 5.2, with respect to the basis set size in FHI-aims. The basis used were $tier1 \rightarrow tier2 \rightarrow tier3 \rightarrow tier4$, otherwise keeping the *tight* settings for the cutoff potential, integration grids and Hartree potential. The calculation was performed with the PBE+vdW functional and for the relative energies, the conformer that has one H-bond (Fig. 5.1(a)) was always taken to be the zero in energy.

The convergence of the relative energies is much faster than the convergence of the total energy, as expected. For relative energies, in the example above, the $tier2$ basis set (which will be used throughout this thesis), is already only 0.2 meV away from the value predicted by the very best basis set ($tier4$). A convergence of 0.2 meV for energy differences is not typical, but for DFT (LDA, GGA) calculations sub-meV convergence is usually achieved at $tier2$ and tight settings, as shown in Ref. [1]. The molecules studied in this work present small energy separations between different conformations, typically of less than 0.1 eV. Therefore, it is important that meV level convergence in energy differences is achieved.



(a) (b)

Figure 5.1: The two conformations of the alanine dipeptide used here. (a) presents an H-bond and was taken to be the reference.



(a) (b)

Figure 5.2: (a) Convergence of the total energy of the alanine dipeptide conformer shown in Figure 5.1(a), with respect to the FHI-aims basis-set size ($tiers1-4$, PBE+vdW generalized gradient approximation, and *tight* settings for numerical grids); the total energies have been shifted so that zero corresponds to the $tier4$ value. (b) Convergence of the relative energy between the two alanine dipeptide conformers shown in Figure 5.1, with the same settings and basis sets as in (a).

Forces are obtained through the evaluation of the negative gradient of the total energy with respect to the nuclear coordinates:

$$\vec{F}_I = -\frac{dV_{BO}}{d\vec{R}_I} = -\frac{dE_{tot}}{d\vec{R}_I}. \tag{5.6}$$

At present, only forces for the DFT total energies (Eq. 3.55) with LDA and GGAs functionals are implemented in FHI-aims. Beyond the terms recognized as the Hellmann-Feynman [258, 259] terms, the so-called Pulay terms [260] arise, because the NAO basis functions present a non-zero gradient with respect to the atomic coordinates. Another term appearing in the force evaluation comes from the multipole expansion of the Hartree potential, and finally, in case of the GGA functionals, there is an extra term related to the variation of the density gradient with respect to the atomic coordinates that has to be evaluated. Detailed expressions for all these terms, as coded in FHI-aims, can be found in Ref. [261].

Much work has been done on the eigenvalue solver in order to replace the conventional one found in the open-source ScaLAPACK library, by one that scales better and more efficiently with the number of processors [262]. With these improvements, allied to optimized algorithms and parallelizations in all other parts of the code, FHI-aims scales efficiently for DFT calculations using thousands of processors. The largest calculation done in this thesis was for Ac-Ala$_{15}$-LysH$^+$ (180 atoms, 672 electrons), using 512 processors (SUN cluster), with the PBE+vdW functional and *tight* settings for basis sets and species defaults (4812 basis functions). Each SCF cycle including force evaluation (one molecular dynamics step), took about 100 seconds. The time taken to evaluate 10 molecular dynamics steps (PBE+vdW functional) for this molecule, varying the number of processors and using the *light* (1692 basis functions) and *tight* (4812 basis functions) settings are shown in Figure 5.3.

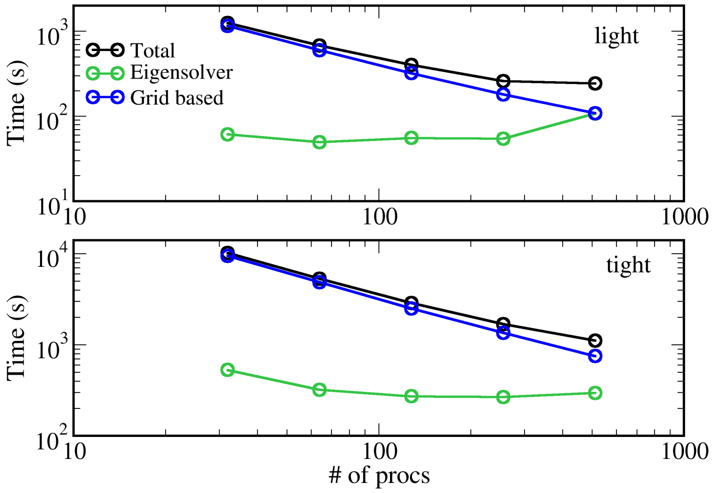


Figure 5.3: FHI-aims timings for 10 force evaluations of Ac-Ala$_{15}$-LysH$^+$ (180 atoms, 672 electrons) with the PBE+vdW functional. Top: *light* settings (1692 basis functions). Bottom: *tight* settings (4812 basis functions). The grid-based operations consist of the evaluation of the electrostatic potential, the electron density, and the integration of h_{ij}.

For the evaluation of the harmonic normal modes of vibrations, FHI-aims [1] uses a finite differences technique. Each atom is displaced in the $x, y,$ and z directions starting from the equilibrium geometry and the forces are calculated at each displacement. The derivatives used to build the Hessian matrix (Eq. 4.10) are then simply calculated as:

$$\frac{\partial E}{\partial x_1^2} = \frac{\partial F_{x1}}{\partial x_1} = \frac{F_{x1}(x_1+\Delta) - F_{x1}(x_1-\Delta)}{2\Delta} \tag{5.7}$$

$$\frac{\partial E}{\partial x_1 \partial x_2} = \frac{\partial F_{x1}}{\partial x_2} = \frac{F_{x1}(x_2+\Delta) - F_{x1}(x_2-\Delta)}{2\Delta}, \tag{5.8}$$

$$\vdots \tag{5.9}$$

where the coordinates run from x_1 to x_{3N}, F_{x1} is the x component of the force of atom 1, F_{x2} is the y component of the force of atom 1, etc., and Δ is a small displacement in the respective direction. This requires $6N+1$ (with N the number of atoms of the system) single point calculations, though, and can get very expensive for very large molecules. Displacements of $\Delta = 5 \times 10^{-3}$ Å with a force convergence criterion of 1×10^{-3} eV/Å have proven to give reliable results (see [261] and Appendix D for details) and are the parameters used in this thesis.

The dipole moment is calculated as the first moment of the density distribution:

$$\vec{\mu}(\vec{r}) = \int n(\vec{r}_0)(\vec{r} - \vec{r}_0) d^3r. \tag{5.10}$$

This quantity is well defined only for neutral molecules. For charged ones, it depends on the choice of the origin. Its derivative as a function of the normal modes of vibration (Eq. 4.17), though, which enters the calculation of the intensities of the harmonic IR vibrational spectra, is independent on the choice of the origin even for charged systems.

As a last remark, it is important to mention that in FHI-aims, calculations that involve the explicit evaluation of non-local exchange and correlation energies (HF, MP2, RPA, ...) are done using the so-called “resolution of identity” method [263–266]. In this method the four-center integrals:

$$(ij|ab) = \sum_{mnkl} c_m^{i*} c_n^{j*} c_a^k c_b^{l*} \iint \frac{\varphi_m(\vec{r})\varphi_n(\vec{r}')\varphi_k(\vec{r})\varphi_l(\vec{r}')}{|\vec{r} - \vec{r}'|} d^3r d^3r' \tag{5.11}$$

are simplified by expanding the products of basis functions in an auxiliary basis:

$$\varphi_m(\vec{r})\varphi_k(\vec{r}) \approx \sum_{\mu} C^{\mu}_{mk} P_{\mu}(\vec{r}), \tag{5.12}$$

where $P_{\mu}(\vec{r})$ are the auxiliary basis functions and C^{μ}_{mk} are the coefficients of the expansion. The four-center integrals can be, thus, rewritten as:

$$\iint \frac{\varphi_m(\vec{r})\varphi_n(\vec{r}')\varphi_k(\vec{r})\varphi_l(\vec{r}')}{|\vec{r}-\vec{r}'|} d^3r d^3r' \approx \sum_{\mu\nu} \iint \frac{C^{\mu}_{mk} P_{\mu}(\vec{r}) P_{\nu}(\vec{r}') C^{\nu}_{nl}}{|\vec{r}-\vec{r}'|} d^3r d^3r' \tag{5.13}$$

The evaluation of the expansion coefficients C^{μ}_{mk} involves 3-center integrals, such that the 4-center integrals are factorized in 3-centers and 2-centers integrals (Eq. 5.13), and the pre-factor involved in these calculations is reduced. The use of this method represents an enormous speed-up for such calculations.

5.2 Basis set superposition error

The basis set superposition error is an error that arises due to the incompleteness of the basis sets used in an electronic structure calculation [267–269]. When using localized basis sets, each atom has its set of basis functions. When the atoms are bonded together in a molecule, they effectively have available the basis sets of all other atoms, beyond their own basis set. If binding energies or atomization energies need to be calculated, the monomers or atoms have much fewer basis functions available than the full complex. Since the basis sets are finite and incomplete, this leads to binding (or atomization) energies that are much too negative. Also when comparing different conformations of the same molecule, the basis functions overlap differently, which leads to different BSSEs for different conformers, making the comparison of relative energetics also subject to this error [14, 270–272]. For example, consider an "extended" and a "globular" form of the same molecule. In the globular form, the density of basis functions per volume is simply higher. This leads to an increased resolution of the expansion of any object in that volume in these basis functions [3]. For codes where plane-waves are used, this error does not strike, since the basis functions do not depend on the placement or density of atoms in the system

A (powerful but complicated) way proposed to correct for this error is called the Chemical Hamiltonian Approach [273, 274]. This is an *a priori* correction, where the error coming from the incompleteness of the basis sets is included through projection operators that are used to produce a BSSE-free wave-function. Even though this method will not be used in this work, the idea behind this correction is useful because it helps to understand where the BSSE comes from. To exemplify (as explained in Ref. [274]), we can take a molecular complex and consider only the one electron operators of the "intramolecular" Hamiltonian of a single monomer A ($\hat{h}_A$). When this Hamiltonian acts on the molecular orbital ϕ^A_i, that belongs to monomer A, the resulting function can be written as:

$$\hat{h}_A|\phi^A_i\rangle = \hat{P}_A\hat{h}_A|\phi^A_i\rangle + (1-\hat{P}_A)\hat{h}_A|\phi^A_i\rangle \tag{5.14}$$

$$\hat{P}_A = \sum_{\mu\nu} |\phi^A_{\mu}\rangle S^{-1}_{(A)\mu\nu}\langle\phi^A_{\nu}|, \tag{5.15}$$

where $\hat{P}_A$ is the projector on the subspace of the molecular basis of monomer A, and $S_{(A)}$ is the *molecular*

[3] This effect has been illustrated in Ref. [1] for large polyalanine molecules.

overlap matrix. The term $(1-\hat{P}_A)\hat{h}_A|\phi_i^A\rangle$ in Eq. 5.14 denotes components in the orthogonal complement to the subspace spanned by the basis orbitals of monomer A. If the calculation of the isolated monomer is performed, only the first term on the right side of Eq. 5.14 appears, because only the basis of that monomer are available. However, when considering the molecular complex with finite basis sets the second term will appear, and this is what gives rise to BSSE. This error decreases if the molecular basis increases and vanishes in the limit of the infinite basis set, because $\hat{P}_A$ becomes the identity operator. A similar reasoning can be followed for the two-electron operators.

In order to compensate the error, the CHA method modifies the terms appearing in the Hamiltonian of the molecular complex, such that consistency is kept with the free monomer calculations by omitting the terms in the orthogonal complement. For example: $\hat{h}_A|\phi_i^A\rangle$ is substituted by $\hat{P}_A\hat{h}_A|\phi_i^A\rangle$, and similarly for the two-electron operators (see Refs. [273, 274]). By making these substitutions in the evaluation of the Hamiltonian matrix elements, it is possible to obtain a BSSE-free wave-function, which is then used to calculate the total energy with the conventional Hamiltonian (see the work by Mayer in Refs. [273, 274]).

Alternatively, the BSSE error can be removed *a posteriori*, by correcting the energies. The method proposed by Boys and Bernardi [268, 275], often referred to as the counterpoise (CP) correction, is very popular and consists of calculating binding energies by computing the energy of the monomers in the presence of the atom-centered basis sets of the full molecule. The binding energies E_b in this case are defined as follows:

$$E_b = E^{sys}(sys) - \sum_f E^f(f) \tag{5.16}$$

$$E_b^{BSSE} = E_b + \Delta_{CP} \tag{5.17}$$

$$\Delta_{CP} = \sum_f [E^f(f) - E^f(sys)], \tag{5.18}$$

where $E^x(y)$ means the total energy of system x in the basis set of y, and sys stands for the whole system while f stands for its fragments. Even though conceptually different, the two methods discussed above converge to the same limit with increasing basis set [274, 276].

In the all-electron calculations performed with FHI-aims and the basis sets discussed in the previous session, the BSSE error is reduced efficiently and rapidly for standard DFT(LDA, GGA) and HF calculations (see Ref. [1]). The NAO minimal basis is exact for the (DFT) free atom, and is already nearly optimal for the description of the DFT Kohn-Sham core orbitals in any other chemical environment, so that very little BSSE is expected from the core orbitals in a DFT calculation. Only the contribution coming from the valence orbitals that are involved in the bonds has to be compensated, and thus already using the *tier2* basis sets for the light atoms produces very little BSSE (i.e., BSSE does not affect AIMD and relaxations with GGA functionals shown in this thesis).

However, for explicitly correlated methods, like RPA or MP2, energy differences converge much slower and the BSSE is much larger. An example for the case of the water dimer is shown in Figure 5.4, comparing DFT-PBE, MP2, EX+cRPA@PBE (RPA with PBE orbitals) and EX+cRPA+SE@PBE. In the figure, the binding energy of the dimer without BSSE correction (E_b) and including it through the counterpoise method [268, 275, 278] (E_b^{BSSE}, Eq. 5.18) are shown [4]. The basis sets used are the aug-cc-pVNZ basis sets [279]. The reason here for not using the NAO basis sets is to perform a comparison to a benchmark case, and these basis-sets are the most established benchmark sets in

[4] The geometry for the dimer and the monomers were taken to be the MP2 relaxed ones, the same as in the S22 set[178] that will be used in other parts of this thesis. For simplicity no relaxation of the monomers were allowed for all other methods.

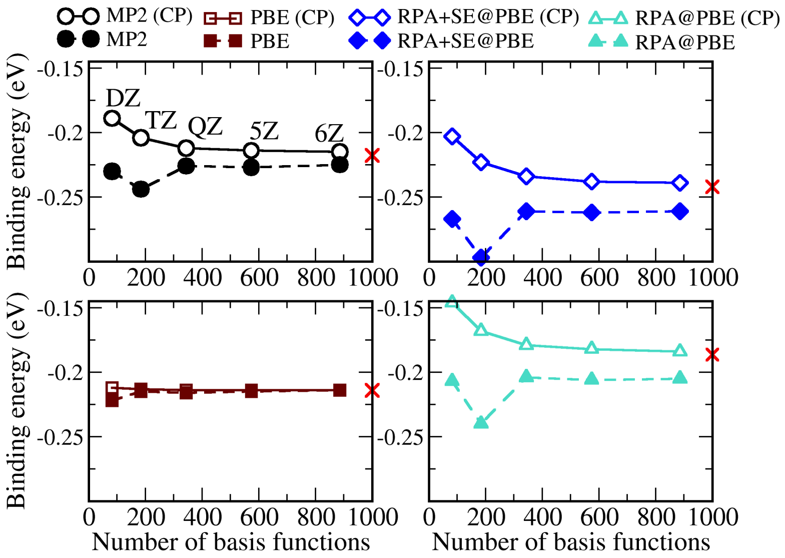


Figure 5.4: Convergence of the binding energy of the water dimer (frozen at the relaxed MP2 geometry) with basis set size (gaussians, aug-cc-pv$N$Z, N=D, T, Q, 5, 6). Filled symbols correspond to binding energies without BSSE correction (Eq.5.16) and open symbols to counterpoise corrected energies (Eq.5.17). Black circles, bordeaux squares and blue diamonds correspond to MP2, PBE, EX+cRPA@PBE, and EX+cRPA+SE@PBE methods respectively. In all plots, the red crosses mark the extrapolation to the complete basis set limit (CBS), from N=5 to 6, following the extrapolation proposed by Halkier *et al.* [277].

quantum chemistry. Moreover, due to their systematic convergence for the correlation energy, these basis sets also allow an analytic extrapolation of total energies to the complete basis-set (CBS) limit [277, 280, 281]. In Figure 5.4, the red crosses show the CBS limit value for the binding energy, as obtained with the extrapolation scheme proposed by Halkier *et al.* [277], using the aug-cc-pV5Z and aug-cc-pV6Z values for the total energies.

The corrected and uncorrected curves for PBE are seen to converge to almost the same value already at the aug-cc-pVTZ basis sets. For the explicitly correlated methods, on the other hand, even for aug-cc-pV6Z there is a difference between the two curves, although this basis sets already contain more than 900 basis functions (for comparison, FHI-aims $tier2$ basis sets for the water dimer have 138 basis functions). The reason why it is possible to get rid of this error efficiently in DFT (as well as HF, although not shown explicitly here) but not in the methods that include non-local correlation, lies in the origin of this error. In DFT and HF, the total energies depend only on the occupied orbitals and its eigenvalues. The number of these orbitals are finite for each system. Since these states are finite, the projection operator shown in Eq. 5.14 quickly approaches identity within the volume spanned by the occupied orbitals, so that BSSE is rapidly diminished. For the case of explicitly correlated methods there is a problem, related to the fact that the total energy in this case depends also on the interaction of the occupied orbitals with the unoccupied ones. This can be seen in the expression of the MP2 energy explicitly (Eq. 3.33) and in the expression for the RPA correlation (Eq. 3.73, where they enter through the polarizability), in which sums over unoccupied orbitals are present. The unoccupied orbitals should be in principle summed up to the continuum, but there is no continuum if the volume of the extended states is explicitly restricted by the restriction of the basis set. The projection operator in Eq. 5.14, thus, cannot be really approximated by the identity operator, which causes BSSE to strongly strike these methods. Moreover, also the core electrons interact with the unoccupied orbitals, which introduces terms that do not exist in a representation based on the occupied orbitals of DFT (which is the case of the *tiers* distributed with

FHI-aims), even if those occupied orbitals were exact. Core energies are large, so that the corresponding terms will also be large, in absolute terms.

Another fact that can be seen in Fig. 5.4 is that even at the largest basis set tested (aug-cc-pV6Z), which should be practically BSSE free, there is still a difference between the corrected and uncorrected values: ≈10meV for MP2 and ≈20meV for EX+cRPA+SE [5]. This fact has been observed before, e.g., in Refs. [278, 282], and has also generated some debate about the counterpoise method of Eq. 5.18 over-correcting BSSE. This argument, however, has been discarded [274, 278] for a number of reasons, including that the CHA method agrees with the CP one. The current understanding of this subject, well discussed in the book of I. Kaplan [89], is that both the corrected and uncorrected curves approach the infinite basis set limit but from different sides. The approach is very slow, such that the slope can be barely seen in Figure 5.4, but one can see that the red crosses corresponding to the CBS limit lie between the corrected and not corrected curves, although they are closer to the CP corrected value for aug-cc-pV6Z. [6]

The counterpoise correction described in Eq. 5.18 is correcting essentially for *inter*molecular BSSE, by the definition of molecular fragments. Nevertheless, when dealing with energy differences between distinct conformations of molecules, there is also an *intra*molecular component of this BSSE that is extremely relevant, as has already been discussed above. A possible *a posteriori* cure for this problem, in the lines of the counterpoise correction, is to perform a BSSE correction of atomization energies, and then take differences only between the (better converged) corrected total energies[271, 282]. This method will be extensively discussed in Chapter 11. In the same Chapter, it will be addressed how to develop basis sets that take care of most of the BSSE in explicitly correlated methods and can be used to converge energy hierarchies, also in connection to atomization BSSE corrections.

[5] For MP2, this difference has been checked against the Gaussian03[148] code by Xinguo Ren in our group, for the same basis sets, obtaining the same 10meV difference for $N = 6$.

[6] This difference (between the CP corrected and not corrected curves) is also observed to be larger when the monomers present a permanent dipole moment. The dipole moment of the monomer calculated only with its own basis and the one from that calculated in the presence of the basis sets of the other monomer are different. This is exactly the case for the water dimer and its monomers shown here, where even at the aug-cc-pV6Z basis set in MP2(HF) the dipole moments exhibit a 30% difference in magnitude for the CP-corrected monomer and the uncorrected one.

Part II

Helical secondary-structure in alanine-based polypeptides

Chapter 6

Alanine-based polypeptides

The primary model systems investigated in this work are alanine-based polypeptides. These polypeptides are considered a paradigm to understand the formation of specific secondary structure elements, especially helices, as the one shown in Figure 6.1. Moreover, there are very good (high-quality and clean) experiments regarding these polypeptides in the gas phase. In this section, an overview of experimental and theoretical work concerning polyalanine peptides in solution and in the gas phase will be given. In the last part of this chapter, a detailed account of a few experiments, that will be directly relevant to the work presented in this thesis, is given.

6.1 Polyalanine as a model system

Proteins, with their composing peptide chains, have evolved over 1 billion years subject to natural selection, so that they could fold (efficiently) into bioactive conformations, allowing them to bind to other specific molecules and perform specific tasks. The efforts to understand protein structure, and with that perhaps elucidate its function, have focused, from the chemical-physical point of view, on understanding the contributing factors by analyzing well-defined parts of the whole problem - a reductionist approach. One appealing idea, pursued in the 60's by various authors [16–18, 283–285], is to identify isolated secondary structure elements (especially helices) in existing proteins. The corresponding amino acid sequence of the relevant piece could then be reproduced and studied separately, in solution. For example, Epand and Scheraga [283] and Crumpton and Small [284], studied, using circular dichroism (CD) spectroscopy [1], parts of myoglobin (or sperm-whale myoglobin), which is a protein known to have a high content of helices. Remarkably, the results did not show high helical content for the isolated sequences, though. Klee and co-workers [17, 285] were partially successful upon studying ribonuclease-based peptide sequences, where some content of helical structure was indicated to be present by CD measurements.

An alternative approach to obtain an ideal system where isolated helical formation can be studied, is to fabricate artificial peptide sequences, often referred to as a *de novo* design. The first evidence for helical secondary structure in a designed polypeptide was reported by Marqusee and Baldwin in 1987 [19]. The peptides studied were 16 and 17 residues long, containing almost solely alanine residues, with

[1] Circular dichroism is a spectroscopy technique that explores the fact that the amino acids are optically active, because they have a chiral center (C_α). In this way, by using left and right polarized lights on a sample, it is possible to measure the differences in attenuation of left and right circularly polarized light. A more modern variant, often referred to as vibrational circular dichroism (VCD) uses infra-red light to probe the sample, which then also couples with the IR-active vibrational modes. For a review of the application of this technique in polypeptides and proteins, the reader is referred to the work by Keiderling [286].

glutamic acid (Glu) and lysine (Lys) residues intercalated at certain intervals. The peptides showed up to 75% helical content upon analyzing CD spectra. The charged amino-acids (Glu and Lys) are necessary to avoid aggregation of the alanine polypeptides (since alanine has a hydrophobic side chain), but may also have an effect on the helix formation. When synthesizing a peptide containing mostly Glu, or mostly Lys in 1989, Lyu, Marky and Kallenbach [287] found that the helical content was not high, pointing to the fact that the presence of the alanine amino acid is necessary for the helical formation in those peptides.

In 1989, Marqusee and Baldwin [288] provided a direct proof of the intrinsic high-helical propensity of alanine by studying short (16-residue) helix-forming alanine-based polypeptides where no side-chain stabilization could occur (e.g., by including only Lys residues in a few positions). This observation was confirmed by subsequent studies [20, 21, 50, 289–295], many of them from the group Baldwin and coworkers. Alanine was found to be the only amino acid to form helices in water without the aid of additional amino acids that provide stabilizing interactions, being a truly intrinsic helix stabilizer [296, 297]. Nevertheless, at specific positions, these charged amino acids have been shown to have a stabilizing/destabilizing effect on the helix. Of particular interest here, is the effect of a charged side-chain interaction with the helix macro-dipole [290]. Due to the fixed orientation of the of the CO and NH groups in a helix (see Figure 6.1), there is a concentration of positive charge in the N-terminus and negative charge in the C-terminus [298], defining a macro-dipole, estimated to be around 3.2 Debye per residue for the neutral α-helix [20]. Already more than 20 years ago, Ooi and coworkers [299, 300] studied the effect of having a polyalanine helix attached to 20 positively charged Lys residues either on the C-terminus or on the N-terminus (Ala$_{20}$Lys$_{20}$Phe and Lys$_{20}$Ala$_{20}$Phe) [300]. A helix stabilization was found for opposite-charge interactions (i.e. Lys on the C-terminus), while a helix destabilization was found for the other case (Lys on the N-terminus). This suggests that a favorable interaction of a charged residue with the macro-dipole of the helix stabilizes helical formation.

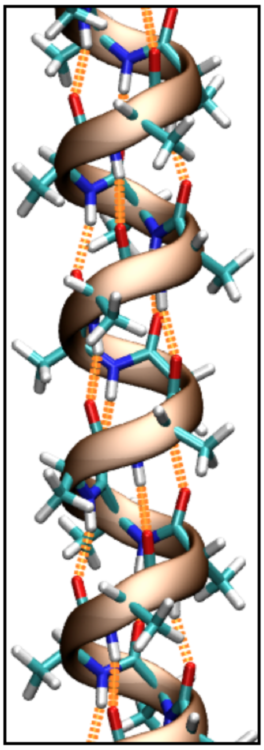


Figure 6.1: An α-helix composed by alanine amino acids. The connecting chains of H-bonds are highlighted in orange.

The actual nature of the helix formed by such Ala-rich peptides in solution was studied by Millhauser *et al.* using electron-spin resonance (ESR) and nuclear-magnetic resonance (NMR) [22, 301][2] , concluding that there were considerable fractions of 3_{10}- as well as α-helical loops. In any case, due to the small side-chain of alanine, the stability of the α-helices formed by these peptides was interpreted to come mainly from intra-molecular H-bond interactions [20, 302].

The helix propensity (i.e. the tendency of a particular amino acid to form a helix) of all amino acids were extensively studied experimentally by host-substitution experiments, where specific amino acids are substituted in a polypeptide chain of otherwise fixed composition [20, 292, 296, 302–307][3]. Experimentally, helix propensities can be measured as the relative gain in (Gibbs) free energy differences ($\Delta\Delta G$) with respect to a reference, usually taken to be the ΔG value for the Ala amino acid [307]. The helix propensity was rationalized in terms of the entropy of the side-chains, which opposes helix

[2] The use of the NMR technique to obtain 3D images of proteins in solution and polypeptides was awarded the Chemistry Nobel prize of 2002, given to K. Wüthrich.

[3] Some of these studies use also statistical theories of helical nucleation, like the Zimm-Bragg [308] and the Lifson-Roig [309] model. These models describe the helix-coil transition of polypeptides, based on statistical mechanics. Statistical weights for nucleating and propagating a helix (plus forming a hydrogen-bond in the Lifson-Roig model) are assigned to each residue in the polypeptide, depending on their state plus that of their nearest neighbors, so that a configurational partition function can be built.

formation because there is entropy loss upon forming the helix [20, 305, 306]. Alanine, with the small CH_3 side-chain, was found to have the highest helical propensity. Gly, that has only a hydrogen atom as a "side-chain" is the amino acid with the lowest helix propensity, at least 1 kcal/mol less than Ala [307]. This, however, is understood by the fact that the high conformational freedom of Gly's backbone atoms themselves favor the non-helical state. In the past decade, Kemp and coworkers [293, 310–313] studied the effects of side-chains and different cappings on helical propensities of alanine polypeptides, via CD and NMR, as well as statistical modeling. In these studies, helical propensity (in water) was found to decrease with increasing temperature, increase with increasing length of the polypeptide, and to have a complex dependence with residue sequence [310].

Also small alanine polypeptides have been the subject of experiments. Polyalanine containing 3 or 4 residues were investigated using VCD [314], for example. With additional theoretical analysis, it was found that these small polypeptides have a high tendency to form a polyproline-II (PPII) helix (presented in Figure 2.6) in solution at room temperature. This PPII tendency was also observed for a 7-residue polyalanine peptide, but at 2^oC, studied through CD and NMR [302].

The simplicity of the alanine amino acid, with its small CH_3 side-chain, as well as its high helix propensity with intrinsic stabilizing interactions, and the great wealth of existing experimental data, were also appealing to theoretical work. Properties of folding, temperature stability, and conformational preferences in solution have been studied mainly employing force-fields and statistical mechanics models as, e.g., the Lifson-Roig [309] and the Zimm-Bragg [308] models. D. Wales [25–27, 85] provided important theoretical insights on the energy landscape and thermodynamic properties of polyalanine peptides [25] and biomolecules in general [26, 85].

Several characteristics of solvated polyalanine were investigated by A. Garcia and coworkers [28, 29, 315–317], using force-fields. Their work illustrates the complexity of studying these systems in water, and using force-fields. In Ref. [315], the effect of the sequence variation on helix stability on alanine peptides was studied with a modified version of the Amber force-field, and evidence for stabilization via shielding of the backbone H-bonds from the water molecules was found. In Ref. [316], an Ala_{21} peptide was studied with the same force-field, and a preference for PPII helices was found for room-temperature. In Ref. [317], the pressure and temperature dependence of chemical shifts in the IR spectra of polyalanine in water was studied, with the conclusion that chemical shifts seen for varying pressure could be due to change in coordination of the water molecules close to the peptides, and not to loss of helicity. In Ref. [29], the shifts of the amide-I band of a 21-alanine peptide in a water-methanol mixture were investigated (still with the same force-field) concluding that the shifts are not only sensitive to the water coordination but also to the local composition of the solvent close to the peptide. In Ref. [28], the modified Amber force-field was compared to another force-field parametrization (OPLS [91]), and quite different structures were predicted for the folded and unfolded structures of an alanine-based polypeptides by each force-field. The OPLS force-field was also applied to the alanine dipeptide and the undeca-alanine peptide by Jorgensen and coworkers [30, 31]. The results obtained were good for conformation energetics and preferences, in comparison to experiments and quantum chemical calculations. It is clear, though, that in these cases there are uncertainties arising due to the parametrization of the force-fields. Several different force-fields (AMBER99, AMBER99SB, CHARMM27, OPLS-AA, and AMOEBA) have been compared to DFT (PBE functional) calculations in a systematic way for infinite helices by Penev, Ireta, and Shea [32]. Only AMBER99SB and OPLS-AA predict the three minima for π, α, and 3_{10} helices (the others miss the 3_{10} minimum), present in the PBE calculation for the PES, with AMBER99SB presenting a closer quantitative agreement.

First-principles simulations are not as numerous as force-field ones for polypeptides in solution, mainly due to the heavy computational costs and the poor description of biomolecules by standard DFT functionals (without the inclusion of vdW interactions). DFT simulations including implicit solvent [4] have been performed by Keiderling and coworkers [34, 286, 319–321], focusing on calculating and reproducing VCD spectral data in order to study the conformation of alanine polypeptides in water. In [34], static calculations are presented, including explicit water for peptides containing 4 to 6 alanine residues. These short explicitly solvated peptides are used to obtain DFT-derived parameters for the implicit solvation of larger ones containing up to 21 alanine residues, and a satisfactory agreement to VCD experimental data is achieved. The DFT calculations did not contain vdW interactions, though. DFT-B3LYP (no vdW corrections) was used by Dannenberg and coworkers [33, 322, 323] in connection with a semi-empirical Hamiltonian to describe the solvent. In Ref. [33], the gas phase and solvation phase minima of neutral polyalanine peptides with 2 to 18 alanine residues were investigated. The preferred unfolded structure in the gas phase was reported to be a β-strand, while in the solution phase it was a PPII helix. The absence of vdW interactions in the calculations might influence such results, though. Finally, in [322], Dannenberg and coworkers have investigated protonation sites of an uncapped 17-Ala peptide, concluding that while in the gas phase the proton can go to any of the three CO dangling bonds on the C-terminus, in solution it prefers the COO(H) group.

Infinite polyalanine helices (no solvent) have been used as a model to probe the stability of different motifs, with respect to application of strain, H-bond cooperativity, and temperature, using only DFT-PBE (no vdW corrections), by J. Ireta and coworkers [35, 55, 70, 324, 325]. In Ref. [55] the importance of H-bond cooperativity on the energetic stability were probed, concluding that within an infinite chain, the cooperativity effect strengthens the H-bonds by more than a factor of two. In Refs. [324, 325], phonon spectra of several infinite helical-motifs were calculated in order to study thermodynamic properties. All helices were found to be destabilized over the fully extended structure when vibrational free energies were calculated, with the π-helix being the most destabilized.

The studies mentioned above, although *extremely* valuable, contain two large uncertainties: Experimentally, it is unclear how to separate the intrinsic helical formation properties of the polypeptides from those induced by the solvent; Theoretically, benchmark studies for geometries and energetics are lacking, where theory can be accurately compared to experiment. It is, thus, unclear whether all interactions are understood and represented with sufficient accuracy, which (if any) empirical force field can be used with confidence, and what is the accuracy, power, and limitations of first-principles methods (DFT or higher-level quantum-chemical methods) to describe such systems. As will be seen in the next section, these questions can be adressed by performing experiments and accurate calculations in the absence of solvent, fully in the gas phase.

6.2 Alanine-based polypeptides in the gas phase

The study of (bio)molecules in the gas phase has become, over the past decades, an increasingly refined way of obtaining general, precise insights [23, 326]. This popularity is due to the development of experimental techniques in the late eighties, that can gently transfer intact biomolecules to the gas phase, like MALDI (Matrix Assisted Laser Desorption Ionization [327]) and ESI (Electrospray Ionization [328]), in combination with high accuracy mass spectrometers [329, 330][5].

[4] "Implicit solvent" refer to models that treat the solvent, e.g. water, as a continuous dielectric medium, instead of treating explicitly each molecule. See Ref. [318] for different models of implicit solvation.

[5] In fact, J. Fenn has been awarded the Chemistry Nobel Prize in 2002 for his development of the ESI technique. It was the same year of the NMR technique, mentioned above. The general subject of the 2002 Chemistry Prize was "for the development of methods for identification and structure analyses of biological macromolecules".

The gas phase environment allows to isolate secondary structure motifs, so that their "unperturbed" energy landscape and stabilizing intermolecular interactions can be carefully studied. Then, the "environmental" effects can be added in a controlled way, for example by the stepwise addition of water molecules to the polypeptide [331] or by adding metal ions to the complexes [332, 333]. At the same time, "clean" experiments in the gas phase allow to benchmark theoretical methods, at system sizes that can be treated in a fully first-principles manner [334].

The overview that will be given in this section focuses again on the study of alanine-based polypeptides in the gas phase, but studies of bio-molecules in the gas phase have encompassed much more than only polyalanine. Polypeptides of several types, sugars and full proteins have also been transferred and measured in the gas phase. Good reviews can be found in Refs. [23, 24, 242, 326, 330, 331, 334–336], and in 2004, the Phys. Chem. Chem. Phys. journal has dedicated an entire issue exclusively to bio-molecules in the gas phase [337].

The relevant questions to be addressed by studying alanine-based peptides in the gas phase are similar to the ones that are posed in solution-phase studies, namely: How does the intrinsic secondary structure stability develops as a function of peptide length and composition? What is the real intrinsic secondary structure propensity of such peptides, in the absence of solvent interactions? What is the role of *intra*-molecular and *inter*-molecular interactions in secondary-structure stabilization? How stable and robust are such structures by themselves, and how much does the structure adopted depend on the environment?

6.2.1 Experimental techniques

In this section, a very brief description of gas phase spectroscopy techniques, that will be relevant for this thesis and the remaining discussion in this chapter, will be given.

Ion-mobility spectroscopy

In ion mobility spectroscopy of polypeptides [338, 339], a small amount of ions, selected from a mass spectrometer, is pumped into a chamber (drift-tube) where there is a non-interacting buffer gas (usually helium) and a weak electric field ($\mathcal{E}$). The arrival time of the ions on the other side of the drift-tube is then recorded. This arrival time, when a sufficiently weak field is applied, and a sufficiently small amount of ions is pumped into the tube, is related only to the geometry of the molecules. It will be, therefore, directly proportional to the collision cross section of the ions with the buffer gas, and usually this cross section is what is reported. The formula used to calculate the collision cross section Ω_m is the following [340]:

$$K = v_D \mathcal{E} \tag{6.1}$$

$$K_0 = \frac{PT_0}{P_0 T} K \tag{6.2}$$

$$\Omega_m(T) = \frac{\sqrt{18\pi}}{16} \frac{Ze}{N_0\sqrt{k_B T}} \left[\frac{1}{m_I} + \frac{1}{m_B}\right]^{1/2} \frac{1}{K_0}, \tag{6.3}$$

where K is the mobility, v_D the drift velocity, $\mathcal{E}$ the electric field, K_0 is the "reduced mobility", defined for a reference pressure and temperature P_0 and T_0, P the pressure, T the temperature, Ze the ion charge, N_0 density of the buffer gas at standard temperature and pressure, m_I the mass of the ion, and m_B the mass of the buffer gas.

This type of spectroscopy is sensitive to overall structural differences, but not detailed ones.

Infrared multiple photon dissociation (IRMPD) spectroscopy

The IRMPD [341] technique is an example of what is known as “action”-spectroscopy. The idea is to have a very low concentration of *mass-selected* (with a mass-spectrometer) ions in the gas phase and a very powerful laser. The laser is tunable to a range of IR frequencies (e.g. 1000 - 1800 cm^{-1}), such that when there is a resonance with an IR-active vibrational mode of the molecule, a fragmentation is induced through absorption of several (tens to hundreds) photons. The depletion in the signal of the parent ion beam (or the appearance of fragment ions) is monitored with respect to the laser frequency, allowing to reconstruct the IR spectrum. This technique is sensitive to the detailed structure of polypeptides, and is applicable to a wide range of molecules. The spectra obtained, however, tend not to be extremely sharp, probably due to the multiple photon absorption. Other details about this technique, relevant to the work in this thesis, will be discussed in Section 6.3.4.

IR-UV double resonance spectroscopy

In IR-UV double resonance spectroscopy (see e.g. Ref. [326], and references therein), the first step is to shoot an UV laser on the mass-selected ions, which (may) cause them to fragment. The fragmentation happens if the energy absorbed is sufficiently high to cause a transition to an excited *electronic* state, that is dissociative in some coordinate. This fragmentation signal is recorded, giving information about the electronic excitations of the molecule. The obtained spectrum may allow to differentiate between different conformations of the ion that are present in the beam [39, 342]. The infrared spectrum is recorded by monitoring the fragment signal induced by the UV laser while turning on an IR laser a short time earlier ($\approx$ 100ns). If the IR laser is in resonance with a normal mode of vibration of the ion, the vibrational ground-state gets depopulated, which is detected as a depletion in the UV fragmentation signal, allowing the reconstruction of the IR spectrum. Specific conformations can be chosen by tuning the frequency of the UV light. This technique supposes that the molecule in the vibrational excited states does not absorb UV light of the same frequency as it does in the ground-state. The technique is also sensitive to the detailed structure of the molecules, but requires them to have an aromatic chromophore (usually a benzene ring), able to be excited by the UV light. The spectra obtained are of very high resolution.

6.2.2 Findings

Alanine-based polypeptides in the gas phase have been extensively studied experimentally by M. Jarrold and coworkers [3, 5, 37, 333, 335, 343–346], employing ion mobility spectroscopy of mass-selected ions [338, 339]. In a groundbreaking experiment from 1998 [37] (also discussed in more detail in Section 6.3.1), Jarrold and coworkers found evidence that designed alanine-based polypeptides could form helices in the gas phase. These designed peptides contained, besides a series of alanine residues, one charged, protonated lysine residue on the C-terminus (Ac-Ala$_n$-LysH$^+$, with Ac standing for acetate), an architecture that had already been seen to stabilize helices in the solution phase [300]. The protonated polyalanine (no Lys termination) was subject to subsequent studies [343], in which some helical conformers were also observed in the ion-mobility experiments, but in a much lower concentration than for the LysH$^+$ terminated molecules, being mostly globular. The interpretation was that the proton in these peptides might be mobile, being able to “walk” from one end to the other of the molecule, inducing a mix of conformations. In a follow-up work [344], the conformations of alanine-based peptides with the charged lysine residue now on the N-terminus (Ac-LysH$^+$-Ala$_n$) were studied, concluding that the monomers of this molecules did not form helices, but helical dimers were observed. Moreover, recently, McLean and coworkers [347] performed ion-mobility experiments on polyalanine peptides with the Lys residue inserted in the middle of the chain, observing low helical content. These results point to the fact that also in the gas phase, alanine polypeptides are stabilized through a favorable interaction of the charge with the helical dipole, and destabilized by an unfavorable one, but that more than the charge (proton) is needed to fully explain the stability.

Water adsorption experiments involving ion-mobility have been performed by several authors [331, 345, 348]. The adsorption of one or a few water molecules by biomolecules is usually referred to as "microsolvation". Such experiments may elucidate details of how solvation takes place and how water affects the gas phase structure of the polypeptide. The Ac-LysH$^+$-Ala$_n$ and Ac-Ala$_n$-LysH$^+$ molecules, with $n =$15 and 20, were the subject of a water adsorption experiment by Jarrold and coworkers [345]. The likelihood of the molecules to adsorb one water molecule (microsolvation) were measured. The globular Ac-LysH$^+$-Ala$_n$ conformers were found to adsorb one water molecule with a binding energy of 0.5 eV, while the helical Ac-Ala$_n$-LysH$^+$ conformers were found not to bind one water molecule at all. Based on this observation, the size (number of alanine residues in the backbone) of Ac-Ala$_n$-LysH$^+$ for which the helical preference onset would occur was probed by a water adsorption experiment [3]. This study will be discussed in more detail in Section 6.3.2. The conclusion is that this helical onset would happen for $n = 8$. M. Bowers and coworkers, having a great experience on ion-mobility experiments of biomolecules in the gas phase [331, 339, 348–350], also studied the microsolvation of these exact same polypeptides [331, 348]. Independently, the same trend observed by Jarrold and coworkers was observed in Bowers and co. experiments: that alanine-based helical conformers in the gas phase do not absorb water molecules, while globular ones do. Specifically, the helical onset observed in Ref. [3] for the Ac-Ala$_n$-LysH$^+$ at n=8 was also observed [348].

Helical propensities for the amino acids in the gas phase were investigated in Ref. [351]. The results indicate that the intrinsic helical propensity of the amino acids cannot be rationalized in terms of the side-chain entropy, given the fact that in the gas phase Valine and Leucine seem to have an even higher helical propensity than Alanine. Steric hindrance due to the size of the non-polar side chains to form globular structures might be behind [335] the observed propensities.

Regarding the thermal stability of helical structures as a function of temperatures, gas phase experiments provide unique insights. The temperature stability of the helix formed by Ac-Ala$_{15}$-LysH$^+$ was probed also by Jarrold *et al.* [5], through the measurement of the ion mobility cross-sections at various temperatures. It was found that this helix was remarkably stable up to 700K, denoting that there are extremely strong *intra*molecular forces stabilizing this motif. This work will also be further discussed in Section 6.3.3. Upon studying alanine/glycine peptides, Jarrold and coworkers [352] have probed the temperature dependence of the conformations adopted by these molecules, finding that the polypeptide Ac-Ala$_4$-Gly$_7$-Ala$_4$ is globular at room temperature but actually becomes helical as the temperature is raised to 400K! Since globular conformations are expected to have higher configurational entropy, the stabilization of the helix with the temperature increase was rationalized in terms of the helix having a higher vibrational entropy than the globular conformation. This fact will be important for the work in this thesis [6].

The ion-mobility experiments mentioned above are very insightful, but they produce no detailed

[6] This interpretation was also backed-up by force-field simulations of the normal modes of vibrations and corresponding free energies of a 21-residue alanine-based peptide, as well as other 60 short peptides [353], but we will come back to this in Chapter 8 of this thesis.

information about the geometry of the molecules, since only global cross-sections can be measured. Such information can be retrieved from IR spectroscopy measurements in the gas phase, as has been pointed out in Section 4.4.

Perhaps the most precise measurements of IR spectra of alanine-based helices in the gas phase come from the group of T. Rizzo [24, 39, 342], from IR-UV double-resonance spectroscopy experiments. These experiments are done at very low temperatures, of around 10 K. Ac-Phe-Ala$_5$-LysH$^+$ (the aromatic ring is in the Phe amino acid), Ac-Phe-Ala$_{10}$-LysH$^+$, and Ac-LysH$^+$-Phe-Ala$_{10}$ were measured using this technique [39, 342], in the amide-A/amide-B region, obtaining very sharp and well resolved spectra. Isotopic shifts were also measured, which allowed the assignment of each peak to a different residue, profiting from the fact that in the amide-A/amide-B regions the hydrogen-stretch vibrations are very localized. For Ac-Phe-Ala$_5$-LysH$^+$, the spectra of four different conformations that were the ones with highest probability in the experimental beam, were compared to DFT-B3LYP harmonic spectra of several conformers. Helical conformations, exhibiting 3_{10} and α helical H-bonds, were consistent with the spectra. For the larger molecule, where a firmly α-helical motif is expected to be preferred, no theoretical spectra were presented.

In order to probe the sensitivity of IR spectra, obtained from the IR-UV double resonance technique to conformations and amino-acid sequence, M. Mons and coworkers studied different conformations of alanine-based polypeptides in the gas phase, also in the amide-A/B region, at low temperatures [354–357]. Spectra obtained for capped tetra-peptides containing two Ala and one Phe amino acid intercalated at different positions [354, 356] show sensitivity to the amino acid sequence. Comparison to DFT-B3LYP harmonic spectra of selected conformations reveal that when Phe was on the N-terminus, a 3_{10}-type H-bond was formed, when it was in the middle, 2_7-type H-bonds were formed, and when it was in the C-terminus, both types of H-bonds were present. The preferred conformations of the benchmark series of protonated polyalanine peptide (Ala)$_{n=2-8}$H$^+$ were studied with IRMPD spectroscopy, at room temperature, also by M. Mons and coworkers [358]. Harmonic spectra calculated with DFT-B3LYP showed that the measured spectra were consistent with compact/globular conformations protonated at the N-terminus, for $n > 3$. Computing the spectra for these conformers, using the dipole auto-correlation function from *ab initio* Molecular Dynamics (DFT-BLYP), in collaboration with the group of M.-P. Gaigeot [36, 242], has produced a much better theory-experiment match, indicating that extended as well as compact families contributed to the measured spectra at 300K.

IR spectra of several biomolecules in region of the amide-I/amide-II/amide-III bands have been measured in the group of G. von Helden [4, 359–361], using the IRMPD technique in connection with the FELIX free-electron laser [362]. Information about the secondary structure of peptides, contained in this region of the spectrum, is of extreme value. Part of the work of this thesis was done in collaboration with G. von Helden and his group. IRMPD spectra of the Ac-Ala$_n$-Lys$^+$ conformers, with n=5, 10, 15 (the same molecules studied by M. Jarrold *et al.* in the above mentioned paper of 1998 [37]) were measured at room temperature. Theoretical analysis of these spectra is the subject of Chapter 9.2 of this thesis. Brief details about the experiment will be given in Section 6.3.4. For a full description, the reader is referred to the Ph.D. thesis of P. Kupser [38].

A great wealth of studies on conformations of peptides and other biomolecules using first-principles methods have been published in recent years (see reviews by Hobza and coworkers [14, 334] and references therein, for example). The conformational analysis of small peptides in the gas phase has been of special interest for the use of high-level quantum chemistry methods (MP2, coupled cluster), focusing on the characterization and accurate description of the non-covalent intramolecular interactions (see work by Hobza *et al.* [14, 169, 178, 334, 363], or Grimme *et al.* [15, 364, 365], for instance).

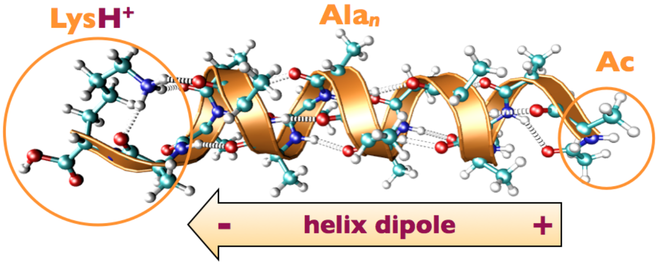


Figure 6.2: Schematic representation of the molecule Ac-Ala$_{15}$-LysH$^+$ in a helical configuration. Each termination is labeled and the direction of the helix-dipole is specified.

Although gas phase experiments, together with quantum mechanical calculations, are seen to reveal details of structure, interactions, and dynamics at the molecular level, the theoretical treatment of biomolecules of increasing size requires the development more powerful computational strategies. The challenges to be met by theory include large conformational landscapes and proper description of dispersive forces. These two specific challenges will be studied and discussed in this thesis, in the next chapters.

6.3 Summary of experimental studies directly relevant to this work

Following, detailed information on particular aspects of experiments that will be directly relevant to work in this thesis is presented.

6.3.1 "Design of Helices that are Stable in Vacuo" (1998)

As has been extensively discussed in the previous sections, alanine based peptides were known since the 80's [19] to form helices in solution. The (charged) lysine and glutamic-acid added to the helices, initially in order to improve solubility [21, 293–295], also affect the structure. Specifically, it was observed that positively charged Lys residues would stabilize a helix (in solution) when added to the (negative, in a macrodipole sense) C-terminus [300]. In 1998, R. Hudgins, M. Ratner, and M. Jarrold [37] used this concept (although it is not clear if there was knowledge about previous work) to design polyalanine peptides capped with a protonated lysine on the C-terminus (Ac-Ala$_n$-LysH$^+$, schematically shown in Figure 6.2 for n=15), that were then brought into the gas phase. Since the N-terminus is capped by an acetate (Ac) termination, the proton is expected to be in the lysine residue. In Figure 6.2 an schematic representation of the direction of the dipole of the helix is shown. For a α or 3_{10} helix (the most common types), 4 or 3 CO groups, respectively, are left dangling on the C-terminus. The LysH$^+$ can, thus, not only stabilize the helix through a favorable interaction with the macro-dipole, but also by saturating the dangling H-bonds of the C-terminus, as shown in Figure 6.2. The "idea" was, thus, to see if this design would actually stabilize helices in the gas phase.

The experiment of Ref. [37] used ion-mobility spectroscopy in order to measure the ion-mobility cross section of these molecules in the gas phase. A detailed explanation of the experimental setup used in the group of M. Jarrold can be found in Reference [366].

Figure 2 of Ref. [37], reproduced here (with permission from M. Jarrold) in Figure 6.3, reports the relative ion-mobility cross section for Ac-Ala$_n$-LysH$^+$, with n ranging from 4 to 20. This relative cross section is calculated with respect to a perfect α-helix for each n, according to the following equation:

$$\Omega_{rel} = \Omega_m - 14.5n\,\text{Å}^2, \tag{6.4}$$

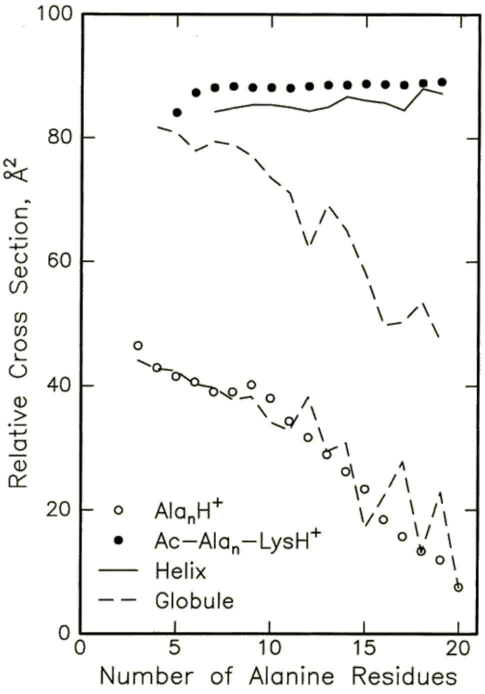


Figure 6.3: Reproduced from Ref. [37], with permission from M. Jarrold: “Plot of the relative collision cross section against the number of alanine residues for Ala$_n$H$^+$ (○) and Ac-Ala$_n$-LysH$^+$ monomers (•). The dashed lines show relative collision cross sections calculated for globular structures from MD simulations. The solid line shows cross sections calculated for helical conformations from MD simulations.”

where Ω_m is the measured cross-section and 14.5 Å^2 is the reported average cross-section of the perfect α-helix[7]. Plotting the cross-sections in this way has the advantage that if a molecule is helical, one should get a constant straight line with increasing n. This happens to be exactly what is seen for the Ac-Ala$_n$-LysH$^+$ series.

In order to check if only the charge would produce the same effect, the ion mobility cross-sections of the Ac-Ala$_n$-H$^+$ series, with n=4-20, was measured. In this case, the “constant line” behavior for the relative cross sections is not observed, instead they decrease with increasing length of the peptide. It is unclear which is the preferential binding site for the proton in this molecule. The cross sections proved consistent with globular conformations taken from force-field simulations, providing indirect proof that the saturation of the dangling CO groups by the Lys side-chain also plays a role on the structure stabilization.

The conclusion of this work is that helical-secondary structure can be observed in the gas phase.

6.3.2 “Water Molecule Adsorption on Short Alanine Peptides” (2004)

In this paper from 2004, [3], the authors M. Kohtani and M. Jarrold performed a microhydration experiment (addition of 1 water molecule) on four different polypeptide series, including the Ac-Ala$_n$-LysH$^+$ with n=4-8. This experiment was devised to probe the existence of short helices, since ion-mobility cross-sections for small molecules are very similar for helical and compact/globular conformers.

In the experiment described in [3] the ions are subject to a constant pressure of water vapor in the drift tube. An equilibrium constant (K_{eq})for water adsorption is calculated, in the following way:

$$K_{eq} = \frac{I_{p+w}}{I_p P_w}, \tag{6.5}$$

where I_{p+w} and I_p are the integrated intensities of the peptide and peptide-water complex peaks in the mass spectrum, and P_w is the partial pressure of water vapor in the drift tube. These measurements had to be done at a temperature of 277K in order to enhance water adsorption by the molecules.

[7] Obtained from classical force-field simulations, as explained in Ref. [367]

From the four series of peptides studied, two of them were alanine-based: Ac-Ala$_n$-LysH$^+$ (previously known to be helical for large n) and Ac-LysH$^+$-Ala$_n$ (previously known to be globular for large n)[344]. Analysis of the results reveals that the smallest n where the equilibrium constant differs significantly between these two series is $n = 8$, allowing the authors to infer that the smallest helix would be formed by Ac-Ala$_8$-LysH$^+$. These results are summarized in Figure 6.4, reproduced from Ref. [3], with permission of M. Jarrold.

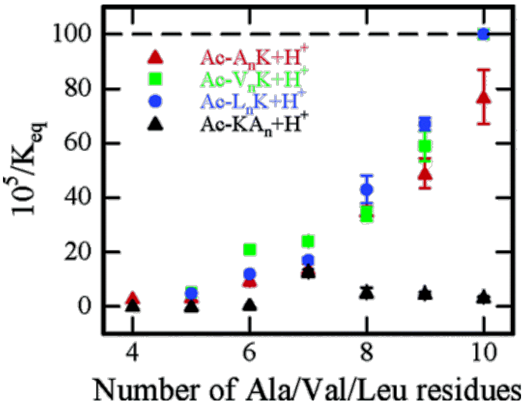


Figure 6.4: Reproduced from Ref. [3], with permission from M. Jarrold: "Plot of $10^5/K_{eq}$ (where K_{eq} is the measured equilibrium constant for adsorption of the first water molecule onto the peptide in question) against the number of nonpolar residues for Ac-KA$_n$+H$^+$, Ac-A$_n$K+H$^+$, Ac-V$_n$K+H$^+$, and Ac-L$_n$K+H$^+$. The error bars are the standard deviation of the mean."

As mentioned before, the same observation (same experiment too) was made by M. Bowers and coworkers [348].

It is worth pointing out that the predicted water adsorption geometry (and adsorption energy at T=0) do not differ between the preferred conformations of Ac-Ala$_5$LysH$^+$ and Ac-Ala$_8$LysH$^+$[368]. The direct connection between helical structure and the discrepancy in Fig. 6.4 is thus somewhat tenuous. Some more direct evidence for this helical onset based on purely theoretical grounds (no water adsorption) will be provided in Chapter 8.

6.3.3 "Extreme Stability of an Unsolvated α-Helix" (2004)

The experiment presented in the paper from M. Jarrold and co. in 2004[5] reports a measurement of the temperature stability of Ac-Ala$_{15}$-LysH$^+$ and Ac-LysH$^+$-Ala$_{15}$, via ion-mobility spectroscopy. The authors measured the ion-mobility cross sections of these molecules in temperatures varying from 300K to 800K and compared to cross sections for perfect helices at these temperatures.

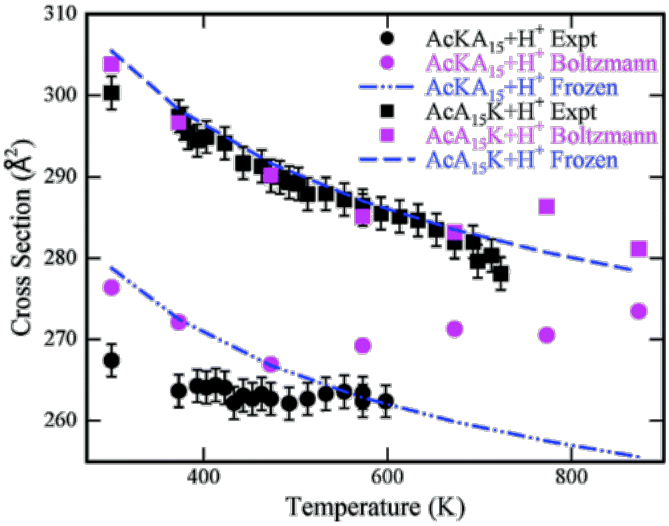


Figure 6.5: Reproduced from Ref. [5], with permission from M. Jarrold:"Measured and calculated collision cross sections for Ac-A$_{15}$K+H$^+$ and Ac-KA$_{15}$+H$^+$. The solid black points are the measured values. The purple points are Boltzmann-weighted average calculated cross sections derived from MD simulation results. The dashed blue lines show the calculated temperature dependence of the collision cross sections for a rigid α-helix (top) and rigid globule (bottom)"

The main finding, summarized in Figure 1 of Ref. [5] and reproduced here in Figure 6.5 with permission from M. Jarrold, is that the helical Ac-Ala$_{15}$-LysH$^+$ seems to maintain its secondary structure, consistent with a helical one, up to around 700K. At $\approx$650K the cross-sections start to deviate from the ones predicted for an ideal helix, and above 725K the signal disappears, indicating dissociation of the molecule. They also find that the helical conformer is more stable than the globular one, that dissociates already at $\sim$575K. The remarkable high-temperature stability of Ac-Ala$_{15}$-LysH$^+$ indicates that the intramolecular forces dictating its secondary structure build an extremely stable conformation, very robust against unfolding.

This experiment will be connected to in Chapter 10 of this thesis, where the dynamics of the unfolding of this molecules will be studied from *ab initio* molecular dynamics (AIMD), at several temperatures.

6.3.4 "Infrared spectroscopic characterization of secondary structure elements of gas-phase biomolecules", Ph.D. Thesis of P. Kupser (2011)

In Chapter 4 of the Ph.D. Thesis of P. Kupser [38], Infrared Multiple-Photon Dissociation (IRMPD) [341], gas phase data for n=5, 10, 15, and 19 of the Ac-Ala$_n$-LysH$^+$ series is presented. The data was taken at room temperature (≈ 300K) in the FELIX free-electron laser facility, in the Netherlands [362]. These spectra were taken at the amide-I/amide-II/amide-III region, considered very sensitive to backbone conformational changes.

A very detailed description of the experiment can be found in P. Kupser's thesis [38]. Here just a summary of the results directly relevant to this work will be given.

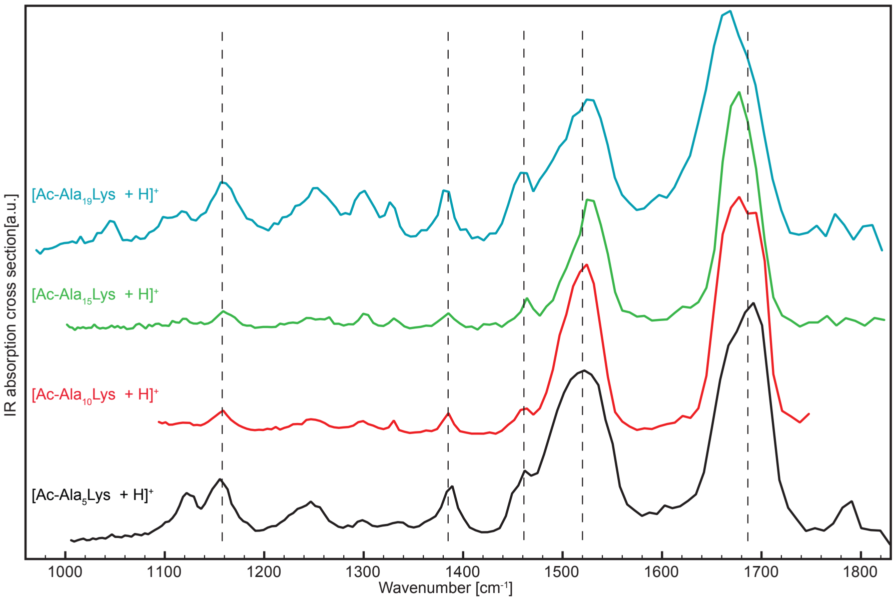


Figure 6.6: Reproduced from Ref. [38], with permission from P. Kupser:"Infrared spectra of polyalanine based peptides with lysine at the C-terminus. Ac-Ala$_5$-LysH$^+$ (black, bottom), Ac-Ala$_{10}$-LysH$^+$ (red, center), Ac-Ala$_{15}$-LysH$^+$ (green, center), and Ac-Ala$_{19}$-LysH$^+$ (light blue, top) are investigated. All spectra are normalized with respect to the highest peak intensity and the spectra are shifted upwards for better readability. Dotted lines are guidelines for the eye and emphasize differences and common features between the spectra. Further explanations are given in the text."

Figures 6.6 and 6.7 contain the measured IRMPD spectra for Ac-Ala$_n$-LysH$^+$, n=5, 10, 15, and 19. The IR-spectrum of n=19 is determined by depletion of the parent ion signal intensity and the spectra

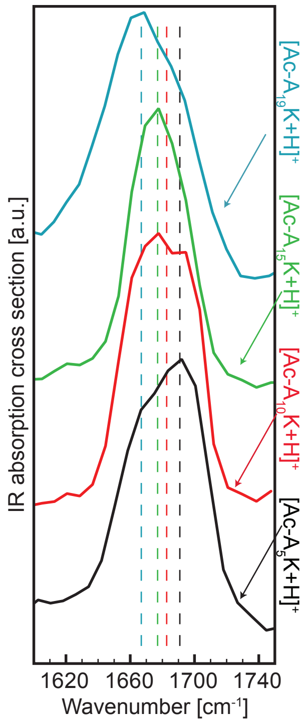


Figure 6.7: Reproduced from Ref. [38], with permission from P. Kupser:"Infrared spectra of Ac-Ala$_5$-LysH$^+$ (black, bottom), Ac-Ala$_{10}$-LysH$^+$ (red, center), Ac-Ala$_{15}$-LysH$^+$ (green, center), and Ac-Ala$_{19}$-LysH$^+$ (light blue, top) in the amide-I band region. All spectra are normalized with respect to the highest peak intensity and the spectra are shifted upwards for better readability. The color of the dotted lines represents the assignment of the corresponding molecule to the peak position."

of the other molecules are calculated by the appearance of fragment ions. All spectra are normalized for the intensity of the highest peak. The data presented is an average of at least two scans for each molecule. These molecules are expected to absorb at least 10 photons before dissociating.

Although the spectra look fairly similar in the wave-number range presented, there are differences in the not-so-intense amide-III region (1000-1400cm^{-1}) and peak-shifts, in particular for the amide-I ($\approx$ 1700 cm^{-1}) band. In fact, a zoom of only the amide-I region shown in Figure 6.7 shows that with increasing length of the peptide chain, the position of the peak of this band is shifted to lower wave numbers in the infrared spectra, from 1690cm^{-1} for n=5, to 1665cm^{-1} for n=19. The amide-II peak (around 1500cm^{-1}) shows a slight shift to the blue with increasing length of the polypeptide, but not as pronounced as the amide-I peak (in the opposite direction). The peak clearly seen at 1790 cm^{-1} in the infrared spectrum of Ac-Ala$_5$-LysH$^+$ (and less clearly seen for the others, due to lack of laser power in the region) can be assigned to a free C=O stretching vibration of the C-terminal carbonyl group.

Mechanisms that broaden the measured IRMPD spectra

The measured spectra show that the amide-I/II/III region can get very crowded for such large and floppy molecules, and the peaks are considerably broad. In this section, a few reasons for mechanisms that may cause the broadening of the peaks in IRMPD spectra are briefly discussed.

One mechanism comes from the experimental apparatus itself, since the excitation laser used in these IRMPD experiments (FELIX), has a width of around 1% of the corresponding wavenumber [38]. Therefore, the measured peaks cannot be narrower than 10-18 cm^{-1} in the region that was measured.

An extra broadening mechanism comes from the fact that the measured IRMPD spectra correspond to a situation where the molecule absorbs tens of photons before dissociating [341]. If the multiple photon absorption, in experiment, would be coherent, there would be the following problem: The first photon would be absorbed in the first vibrational state, making the molecule go to the second vibrational state; the second photon would be absorbed in the second vibrational state, making the molecule jump to the third, and etc., as in a “ladder climbing” picture. Due to the anharmonic nature of the PES, where the spacings between the vibrational levels are not constant (see Figure 4.1), if this were true the molecule would rapidly be out of resonance with the IR laser, so that the dissociation at the resonant wavelength would fail. However, this is not the case, since spectra can be obtained by this technique and they yield a good agreement with the linear single-photon absorption picture [341, 369]. The good agreement comes from the fact that the photon-absorption process is not coherent, in the sense that the photons are not absorbed one after another in subsequent vibrational levels. In fact, the energy of each absorbed photon redistributes itself in the bath of vibrational states of the molecule, due to anharmonic coupling between vibrational modes. This process is often called intramolecular vibrational redistribution [370]. If the molecule is large enough so that the density of states is large enough and there is enough anharmonicity, this effect efficiently removes the population from the excited state into the bath of accessible vibrational states, so that the molecule can receive the next photon in the same vibrational level. At the same time that this process justifies, to a certain extent, the observed good agreement between IRMPD measured spectra and linear-absorption calculated ones, it can also lead to further broadening from the peaks and to distortion in their relative intensities.

Another source of possible broadening of the peaks, which is not particular to this technique, is the co-existence of different conformers in the beam.

Overall, the complexity of these spectra is such that theoretical predictions are necessary to provide a detailed interpretation of the vibrational peaks. In Section 9.2, analysis coming from the harmonic approximation and including anharmonic and temperature effects, through AIMD simulations (dipole autocorrelation function) will be presented. Reliability factors will be used to give quantitative information on the match between theory and experiment. The inclusion of anharmonicities improve considerably the fine agreement between theory and experiment.

6.4 Summary

The work summarized in this chapter provides experimental evidence for the relevant questions tackled from a first-principles point of view in this thesis, regarding alanine-based polypeptides in the gas phase, namely: How does the intrinsic secondary structure stability of the Ac-Ala$_n$-LysH$^+$ develop as a function of peptide length? Which intramolecular forces (vdW, H-bonds, electrostatic) are important for the stabilization of these molecules, and how accurately should they be described by first-principles methods in order to get reliable results? What is the detailed geometry of each molecule from this family of peptides, up to n=15? How do thermodynamic effects at finite temperature (free energies, entropy) affect these structures? How do the dynamics of these structures develop at different temperatures, and again, which intramolecular forces govern these dynamics? These questions will be addressed in the next chapters of this thesis.

Chapter 7

Search strategy to explore the conformational space of Ac-Ala$_n$-LysH$^+$

The polypeptides studied in this work have a broad conformational space. To illustrate this statement, we take the case of the smallest molecule studied in this work, namely Ac-Ala$_4$-LysH$^+$ (70 atoms), which is shown in its extended form in Figure 7.1(a). This molecule contains 5 amino acids, therefore 10 backbone torsional angles (ϕ, ψ). Forgetting about the degrees of freedom of the side chains (also the LysH$^+$ one) for a moment, and supposing that each of these torsional angles can adopt only 6 different values that can be considered "non-equivalent", there would be already 8000 possible conformations. The matter is much more complicated, since these angles can adopt much more than 6 non-equivalent values, and the Lys side chain is long and can also adopt several different configurations. Not all configurations are possible, since some would lead to steric clashes, but still, even for this 70-atoms molecule, there can be an *enormous* amount of possible configurations. In Figure 7.1(b) just a few of the possible configurations Ac-Ala$_4$-LysH$^+$ can adopt are shown, to give an idea of the possible variety. A few important points can be seen from these conformations, too. Consider the conformations shown in the center and at the far right (numbers 1 and 3). These two structures present helical loops, where number 1 is more compact and the number 3 more elongated (in fact, that is a 3_{10} helix). In the whole molecule, for n=4 there are 5 CO groups (excluding the COOH group) available to form H-bonds. When helical loops are formed, 3 or more (assuming that it is not a 2_7 helix) of these groups do not participate in backbone H-bonds. The lysine side-chain can saturate these dangling CO groups, as can be seen in the picture, and these will be found, further ahead, to be energetically stable conformations. Other possibilities are conformations like the number 4 and number 7, where the Lys side chain saturates CO groups lying in the Ac termination or in the neighboring Ala residues. These conformers tend to be more compact ("globular") conformations. For larger molecules more possibilities can be found, but similar geometrical considerations hold.

A broad search of the conformational space is required in order to determine, with a good degree of certainty, which are the most stable conformers of a given molecule. Using solely *ab initio* methods to do so is practically impossible, even with state of the art computational power. To solve this problem, in this work a two-step search is adopted, where the first step is based on a plain basin-hopping search using an empirical force field, and the second on full first-principles DFT calculations.

Although there are other, more refined (than plain basin-hopping), algorithms to probe the conforma-

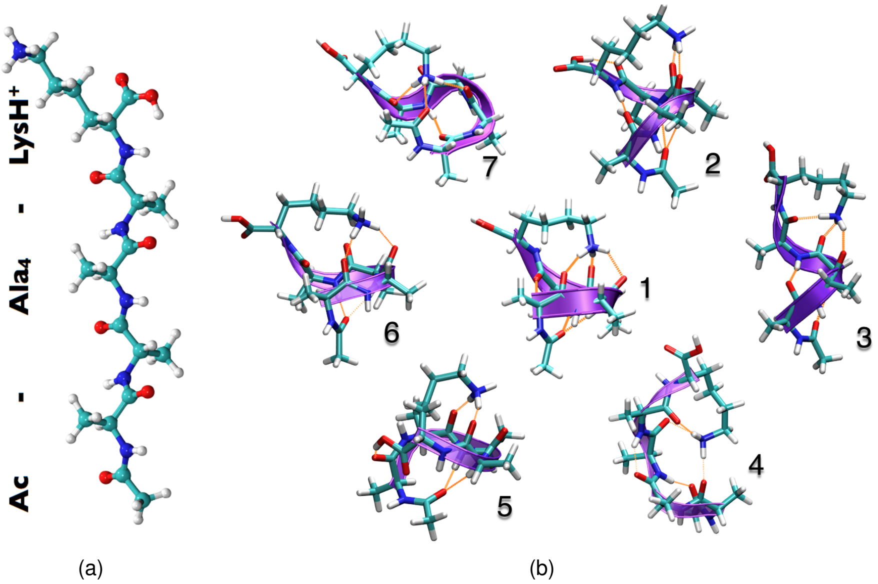


Figure 7.1: Ac-Ala$_4$-LysH$^+$: (a) extended conformation; (b) a few of the myriad of possible conformations this molecule can adopt.

tional space [25, 26, 371–374], which allow to find probable global minima more efficiently, the choice here was to be simply as broad as possible in the first step. The reason for that is to avoid any bias induced by inaccuracies in the force field, and use it just to get input structures to relax in DFT. It is well known [32] that conformational energy differences between different types of secondary structure may vary strongly between different force fields and/or DFT. Additionally, the interest here, especially for the smaller conformers, is to probe not only the global minimum but also other conformations that lie close in energy.

7.1 The strategy in detail

The first step of the conformational search consists of a comprehensive exploration of the conformational space using an empirical force field, used here just as a structure generator. This force field has been benchmarked against MP2 for small alanine based polypeptides in the gas-phase, producing reasonable results [30, 95].

Using this force field, a basin hopping search, as implemented in the TINKER [375] code package, was carried out. Starting from a local minimum, a generic basin-hopping algorithm consists of perturbing the coordinates of the atoms in a particular way and then applying again a local minimization for the new point in the potential energy surface (see Figure 7.2). If after the local minimization, one finds a new minimum, this is counted as a new structure. If one falls in the same minimum, that can in principle be defined by energetic and/or structural comparisons, the structure is disregarded. This technique, thus, effectively ignores transition state regions on the PES.

In the particular implementation found in TINKER's *scan* utility, the coordinates of a particular minimum are displaced in trial steps along the direction of the torsional normal modes of the molecule (effectively distorting the molecule). These modes are obtained by taking the vibrational modes (calculated in the harmonic approximation, such that for each mode there is an associated frequency) that correspond to torsional motions of groups. All possible torsions of all groups of the molecule are considered. The trial

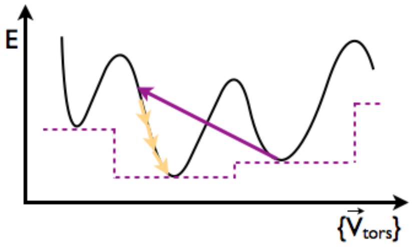


Figure 7.2: Schematic picture of a basin hopping procedure following the directions of the torsional normal modes $\vec{V}_{tors}$: a move is attempted represented by the purple arrow, followed by minimization, represented by the yellow arrows. The PES that is effectively seen by this scheme has no direct knowledge about the barriers between two minima, and is represented by the dashed lines.

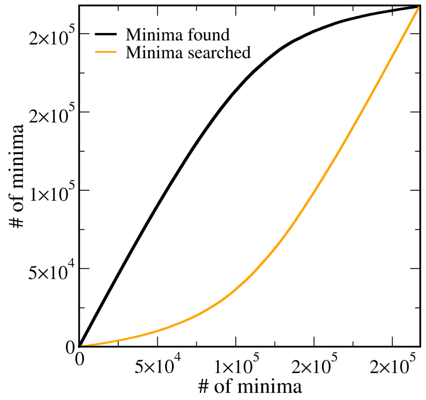


Figure 7.3: Convergence behavior of the search performed by the TINKER *scan* utility for Ac-Ala$_4$-LysH$^+$. The axis show the number of conformers explored in a force field search, using 15 torsional modes, 50kcal/mol for the energy window, and 10^{-4} kcal/mol for the energy comparison threshold.

steps are followed until the energy does not increase anymore, at which point a relaxation of the structure is performed. Only the energy of the minima are kept. After relaxation, the only check that is performed to define if the structure that was found is a new minimum or not is an energy comparison. There are three parameters that can be adjusted: (i) the number of torsional modes to be followed N_{modes}, starting from the one with lowest eigenvalue; (ii) the upper energy limit for keeping or discarding local minima e_{max}; and (iii) the accuracy for the energy convergence of local minimizations of the structures e_{thresh}, so that these can be compared to determine if there is a new minimum or not. The algorithm stops when all minima have been searched and no new minimum is observed (displayed in Figure 7.3).

The convergence of the number of conformers found with respect to N_{modes} was investigated, maintaining $e_{thresh} = 10^{-4}$ kcal/mol and $e_{max} = 50$ kcal/mol. The amount of conformers found in total and in the lowest 10 kcal/mol for an unconstrained search of Ac-Ala$_4$-LysH$^+$ are shown in Table 7.1. Although the total number of conformers found converges slowly, the extra conformers found are always high in energy. The number of conformers found in the lowest 10 kcal/mol in the unconstrained search, shown in Table 7.1, is already converged for 10 modes. In order to ensure convergence, $N_{modes} = 15$ was used in all "production" searches performed in this work. Searches with particular H-bonds constraints, for n=5, that forced to have an specific type of helicity (e.g. α, 3$_{10}$, etc.), were also tested, yielding the same convergence behavior.

The number of conformers found in the force field search was also investigated with respect to e_{thresh}, for the same unconstrained search of Ac-Ala$_4$-LysH$^+$, using 15 torsional normal modes. Decreasing e_{thresh} from 10^{-4} to 10^{-5} kcal/mol produces 1 order of magnitude more conformers, but again, all new conformers appear very high in energy. To show that, in Figure 7.4 the conformers found with both e_{thresh} are ranked according to their relative energy, i.e. lowest energy is index 1 and taken to be the zero-energy, next lowest is index 2, etc. The shape of the curve has a particular funnel-with-tail shape, where one can see that the tail has exactly the slope of the energy comparison threshold criterion, meaning that, at that part, there is one "different" [1] conformer at each 10^{-4} or 10^{-5} kcal/mol interval. When comparing

[1] Actually it is not clear (and here not investigated) if these conformers are really substantially different or if they have only very slight differences but belong to the same minima. Since the only criterion is an energy comparison, even if the geometry is

N_{modes}	total nr. of confs.	Nr. of confs. in lower 10 kcal/mol
2	131744	459
5	192794	479
10	204178	480
12	212432	480
15	214971	480
20	220130	480

Table 7.1: Number of conformers found with respect to the number of torsional modes used by the TINKER *scan* utility. The molecule of choice is Ac-Ala$_4$-LysH$^+$, there were no constraints in the search, the energy comparison threshold was 10^{-4} kcal/mol, and it was performed on 64 CPUs.

the "funnel" part of both searches in the lowest 10 kcal/mol, shown in the zoom of Figure 7.4 (which will be the energetic range of interest in this work), one can see that both searches produce the same conformers.

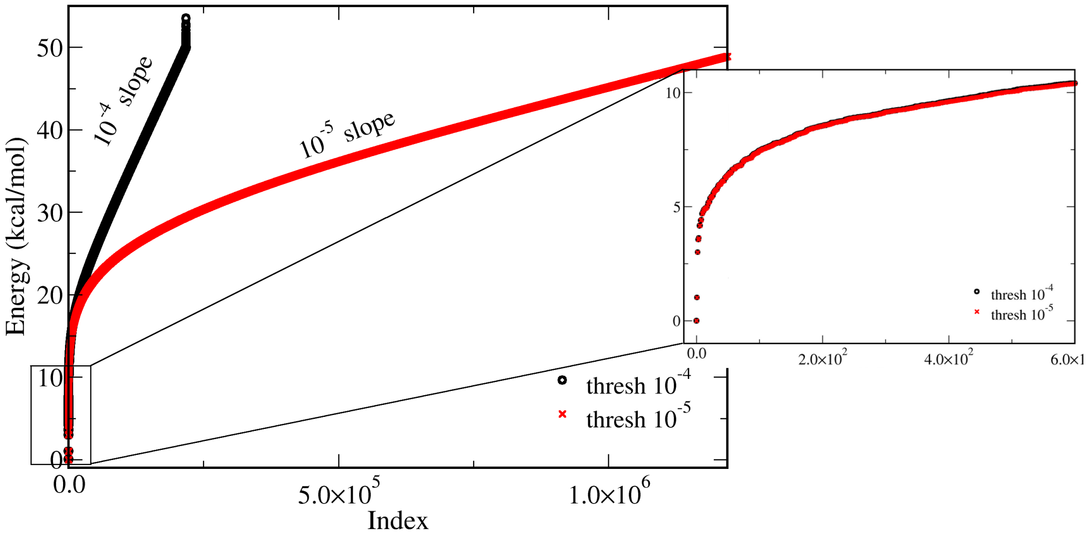


Figure 7.4: Number of conformers found by the TINKER *scan* utility with $e_{thresh} = 10^{-4}$ and $e_{thresh} = 10^{-5}$ kcal/mol (15 modes, e_{max}=50 kcal/mol). Conformers found in the searches are ranked (x-axis) according to their relative energy (y-axis) and given an index, i.e. lowest energy is index 1, next lowest is index 2, etc.

Changing e_{max} from 50 kcal/mol (2.2 eV) to 100 kcal/mol (4.3 eV) only produced more conformers above 50kcal from the lowest energy conformer. All searches reported in this thesis have been considerably sped up by using a parallel version of the TINKER basing hopping algorithm, implemented within our group. More details about this parallel implementation can be found in Ref. [368].

Based on the tests and considerations above, the parameters used for the basin-hopping searches in this work were: $N_{modes} = 15$, $e_{thresh} = 10^{-4}$kcal/mol, and $e_{max} = 50$kcal/mol.

The dependence on the energy hierarchy provided by the force field is reduced by performing the second step of the search. This second step consists of full relaxations of hundreds of conformers taken from the low-energy range of the force field using DFT. The relaxations are performed using the FHI-aims [1] program with the PBE generalized gradient approximation [131]. In order to consider van der Waals interactions, a C_6/R^6 term as proposed in the TS-vdW [2] scheme, (already discussed in Chapter 3) was

essentially identical, they might be assigned as different depending on the accuracy of the energy.

added to the DFT energy expression, and included in the relaxations.

The relaxations themselves are performed in two parts: first a full relaxation of all chosen conformers from the force field is done using the "*light*" settings for integration grids and basis sets [1]; then, only some of the lower energy conformers, now predicted from DFT, are taken and post-relaxed using the "*tight*" settings, discussed in Section 5. The forces in these relaxations were converged down to 5×10^{-3}eV/Å. How much of the lower energy range of the force field needs to be considered for the DFT relaxations and what are the limitations of this method are fundamental questions that will be studied for n=4 and 5 in the next Section.

7.2 How well does the search strategy work and which are the limitations?

7.2.1 Test cases: Ac-Ala$_4$-LysH$^+$ and Ac-Ala$_5$-LysH$^+$

It is essential to show that the search strategy presented here reliably ensures finding the energetically relevant conformers of a given molecule. The outcome of the unconstrained force field search with the standard set of parameters for Ac-Ala$_4$-LysH$^+$ is shown in Figure 7.5. Each black dot corresponds to a different minimum found in the force field, in the order that they were found. The fact that the dots overlap and, ultimately, form one big black area, is intentional. The number of possible candidate conformers is positively huge, which is a central constraint on any naive approaches to structure search and enumeration of peptide conformations.

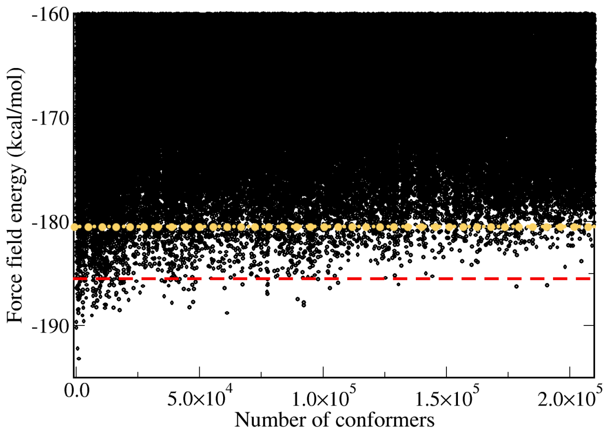


Figure 7.5: Outcome of an unconstrained force field basin-hopping search for Ac-Ala$_4$-LysH$^+$. Each black dot corresponds to a different force field minimum. Conformers below the red-line (~7 kcal/mol, or 0.3 eV) were considered for relaxation in DFT-PBE+vdW.

Three batches of relaxations with PBE+vdW, using the *light* settings of FHI-aims, were performed. One contained all conformers below the red line in Figure 7.5 (100 conformers and up to ~ 7 kcal/mol from the minimum), Another contained all other conformers up to the orange line in Figure 7.5 (900 more conformers and up to ~ 12 kcal/mol, or 0.52 eV, from the minimum). Finally, the last batch contained 68 conformers, chosen in intervals of 30, spanning up to 14 kcal/mol, or 0.61 eV, from the minimum. The idea was to compare all these force field minima with the DFT-PBE+vdW prediction (including the relaxations from the force field structure to the PBE+vdW structure) and check how much information could be gained already from the first set of relaxation, how much new information came from the second, and roughly how correlated the energy the FF and PBE+vdW hierarchies are. There are two kinds of possible problems that can be assessed by this exercise:

- The force field could fundamentally disagree with DFT for a given conformation.
- The force field minima could sometimes be unstable and relax into a totally different, lower minimum in DFT.

As we will see below, both problems are observed and need to be contained, by performing enough DFT relaxations. By analyzing the structural motifs lying low in energy from all these relaxations, it is possible to infer if other relevant structures will be missed by not relaxing in DFT-PBE+vdW all other conformers that lay even higher in energy in the force field.

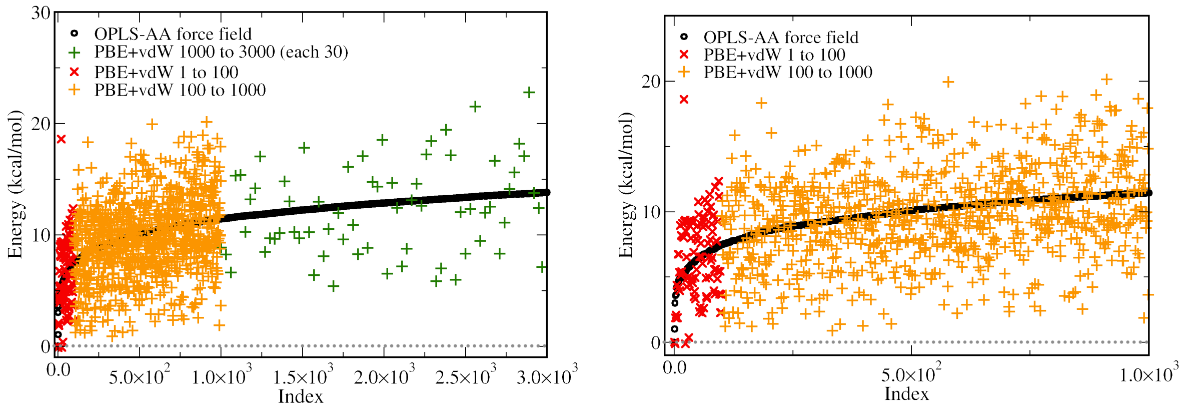
The conformers found in the force field were ranked according to their energy, as was done in Figure 7.4. This produced the black line seen in Figure 7.6(a). Then, the relative DFT-PBE+vdW energies of all the conformers relaxed that were below the red line in Figure 7.5) (rank 1-100, red symbols in Figure 7.6), and the DFT-PBE+vdW relaxed energies of all the conformers below the orange line (rank 100-1000, orange symbols in Figure 7.6) were plotted. In addition, the relative DFT-PBE+vdW energies for 68 chosen and relaxed conformers lying above that threshold are also plotted (rank 1000-3000, green symbols in Figure 7.6). By analyzing Figure 7.6, the first good news is that the lowest energy conformer is found among the first set of 100 relaxations. The second good news is that looking at the overall shape of the curves, including the green points, there is clearly some degree of correlation between the force field and DFT-PBE+vdW energy ordering. Although far from perfect, it means that at least weakly relying on the force field energy hierarchy is justified in this case [2]. However, low-lying energy conformers appear among the orange points. These conformers can be important if they are not already sampled among the first batch of relaxations (red crosses).

In order to further analyze the nature of these low-lying minima and compare them, the PBE+vdW relaxed conformers were sorted into families according to their H-bond pattern.[3] In total, from the 1000 lowest rank conformers that were relaxed with PBE+vdW, 200 families were identified. In the lower 3kcal/mol (0.13eV) range in the PBE+vdW energy hierarchy, there were 7 different families. Each one of them is labeled by a different color in Figure 7.7.

From the seven different families that are found in the low energy range of Figure 7.7 only 4 have representatives among the lowest 100 force field ranked structures, after relaxation in PBE+vdW (red crosses), the rest appearing only later in the force field rank. A detailed analysis of the geometric character (H-bond pattern) of these families is given in Table 7.2. The H-bonds are labeled from the Ace termination (N-terminus) to the LysH$^+$ termination (C-terminus) as is schematically represented for n=4 in Figure 7.8. Their respective relative energy in PBE+vdW (*tight* settings) are also reported. The families are labeled with numbers, where Family 1 corresponds to the green circles in Figure 7.7, Family 2 corresponds to the dark-gray circles, Family 1a corresponds to the light-gray circles, Family 3 corresponds to the purple circles, Family 1b corresponds to the light-pink circles, and Family 4 corresponds to the yellow circles. The 3D geometry of all these conformers can be seen in Figure 7.9.

[2] Work from our group, performed by Carsten Baldauf, has shown that when there are cations binding to the amino acids, the force field energy hierarchy is clearly less reliable and a much weaker correlation is found between the force field and the PBE+vdW energy hierarchy[376].

[3] In keeping with the definition in Chapter 2, we denote a H-bond when an (C)O atom is closer than 2.5Å to a (N)H atom and all possible H-bonds were considered for this classification, so that different conformers belonging to the same family only differ by slight bends of the backbone atoms.



(a)

(b)

Figure 7.6: Correlation between the force field and DFT-PBE+vdW (relaxed) energy hierarchies for Ac-Ala$_4$-LysH$^+$. In black, the force field conformers ordered by their energy. Red crosses (multiplication symbol): the outcome of PBE+vdW relaxations of all conformers lying within 7kcal/mol of the lowest force field conformer. Orange crosses (plus symbol): the outcome of PBE+vdW relaxations of all conformers lying within 12kcal/mol of the lowest force field conformer Green crosses (plus symbol): the outcome of PBE+vdW relaxations for conformers up to 14kcal/mol from the lowest force field conformer, taken in intervals of 30. The dotted grey lines in the plots mark the reference energy (zero), taken as the lowest-energy force field conformer.

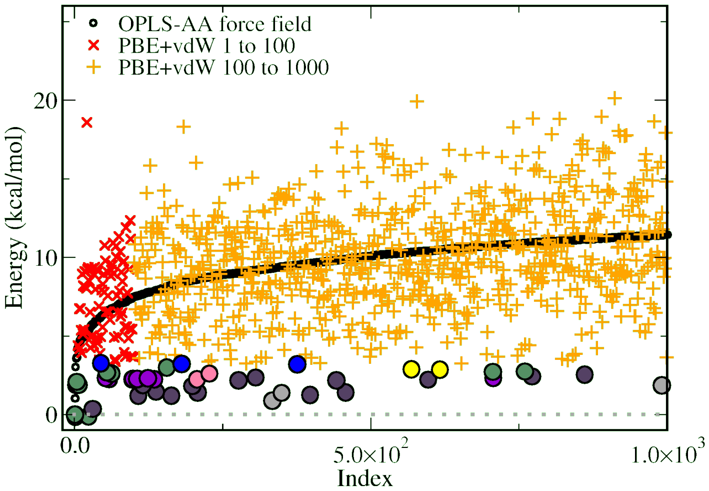


Figure 7.7: Same as Figure 7.6 but with each different H-bond family (see text) lying in the lower 3kcal/mol in PBE+vdW colored differently. In total there are 7 families.

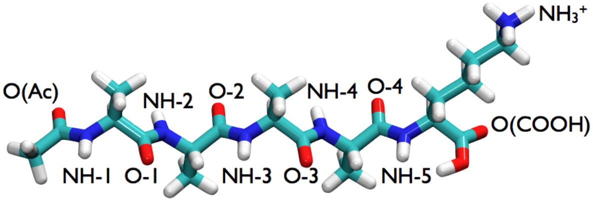
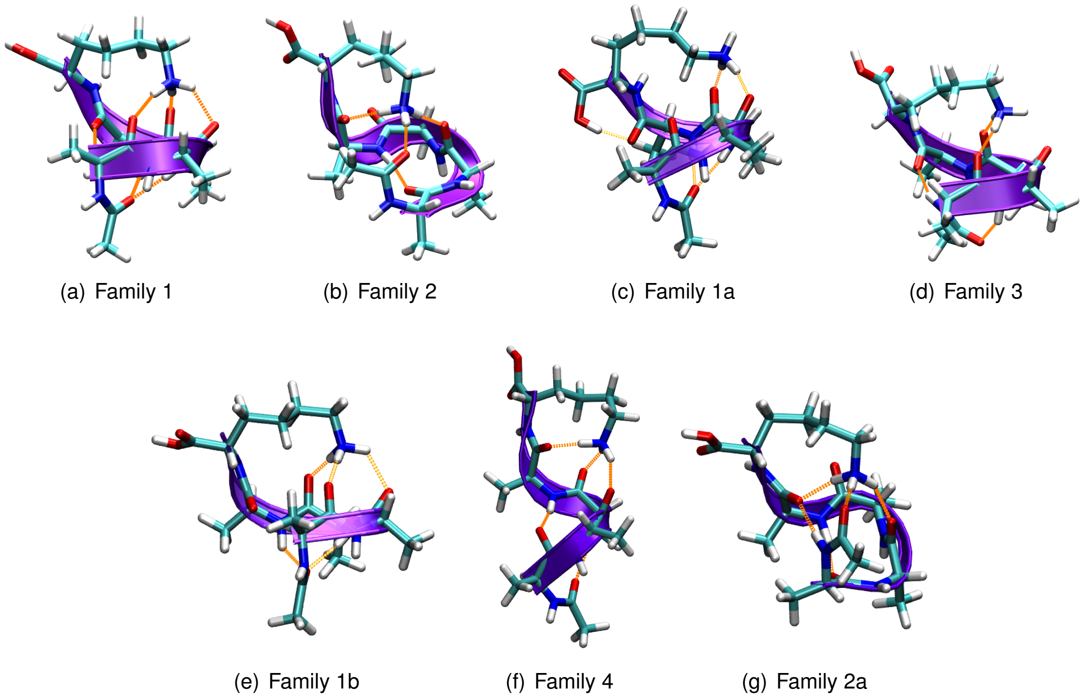


Figure 7.8: Labels used for the CO and NH groups in Table 8.1, represented schematically for Ac-Ala$_4$-LysH$^+$.


(a) Family 1 (b) Family 2 (c) Family 1a (d) Family 3

(e) Family 1b (f) Family 4 (g) Family 2a

Figure 7.9: Lowest six families of n=4, for which detailed H-bonds and energies are reported in Table 7.2.

From Table 7.2 and Figure 7.9, it is observed that the lowest energy conformer (Family 1, Fig. 7.9(a)) has a characteristic bifurcated H-bond of an α/3$_{10}$-helical character formed by the CO group of the Ac termination. This gives this conformer, though small, a helix-like loop. Family 2, the second lowest in energy, is very different, with its Ac termination connected directly to the NH$_3^+$ from the Lys side-chain, making it much more compact. The next family sampled among the first set of relaxations is Family 3. This family is particular for exhibiting a 2$_7$ loop and an "inverted" hydrogen bond (CO pointing to the N-terminus and NH pointing to the C-terminus), where the CO group points to the N-terminus and the NH group to the C-terminus. [4] The highest energy family that is sampled among the first batch of conformers is Family 2a (blue circles), which has exactly the same H-bond pattern as Family 2 but the NH group from the Ac terminus is turned, making a bond to O4 (which is also connected to the lysine NH$_3^+$). From the

[4] This motif will be encountered again for the longer molecules, and will prove to be a very important one.

Table 7.2: H-bond networks of the families within ~3kcal/mol (~0.13eV) of the lowest energy conformer for n=4 . The correspondence of the colored circles of Figure 7.7 with the labels given here are: Family 1 - green ; Family 2 - dark-gray; Family 1a - light-gray; Family 3 - purple; Family 1b - light-pink; Family 4 - yellow. The relative energies obtained with PBE+vdW (*tight* settings) are also reported, in eV.

n	Oxy.	Family 1	Family 1a	Family 1b	Family 2	Family 3	Family 4	Family 2a
4	O(Ac)	NH3 (3$_{10}$) NH4 (α)	NH3 (3$_{10}$) NH4 (α)	NH3 (3$_{10}$) NH4 (α)	NH$_3^+$	NH2 (2$_7$)	NH3 (3$_{10}$)	NH$_3^+$
	O-1	NH$_3^+$	free	NH$_3^+$	NH4 (3$_{10}$)	NH$_3^+$	NH4 (3$_{10}$)	NH4 (3$_{10}$)
	O-2	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$
	O-3	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH5 (2$_7$)	NH$_3^+$	NH$_3^+$	NH5 (2$_7$)
	O-4	free	OH	free	NH$_3^+$	NH1 (inverted)	NH$_3^+$	NH$_3^+$/NH(O-4)
	O(COOH)	NH5 weak	free	free	free	free	free	free
	E_{rel}(eV)	0.00	0.04	0.10	0.01	0.10	0.12	0.13

families appearing only among the relaxations of conformers with rank 100-1000, Families 1a (light-gray) and 1b (light-pink) are very similar to Family 1, having the exact same characteristic bifurcated H-bond coming from the Ac termination. The changes lie in the orientation and connection of the side-chain of the LysH$^+$ residue. These families are therefore classified as variants of Family 1.

In summary, the conformers discussed so far are thus either found already in the first 100 force field structures, or (like 1a, 1b, and 2a) are only small structural variations thereof. The only completely new low-energy (DFT) family encountered in the second batch of relaxations (rank 101-1000) is Family 4 (yellow). This family is actually a pure 3_{10} helix. By sorting all relaxed conformers into families and looking for the lowest rank from each family, it was found that the 3_{10}-helices only start being sampled at ≈10kcal/mol (0.4eV) from the lowest energy conformer in the force field. In PBE+vdW, though, this family is predicted to be much lower in energy, and this is precisely a case where the force field seems to be making a systematic error with a very well known helical motif.

In Figure 7.10 the relative energies from the force field and PBE+vdW for the relaxed conformers are compared, also including the relaxations with *tight* settings. The differences between the *light* and *tight* energy hierarchies are minimal and it is thus enough to relax only the PBE+vdW low energy conformers with *tight* settings, in order to get converged relative energies for the relevant conformers. Therefore, relying on this analysis, the conclusions are: (i) it looks very likely that the lowest energy structure motifs are sampled, although it is impossible to rule out grotesque systematic errors in the FF; (ii) many of the low-lying conformers do in fact correspond to the same overall structure classification (H-bond pattern), so even if small side group rotations or such might still differ, there is a high chance of actually having found the relevant low-lying motifs, and (iii) care must be taken to include conformers that are systematically overestimated in the force field (at least the ones that are known, e.g. 3_{10}-helices).

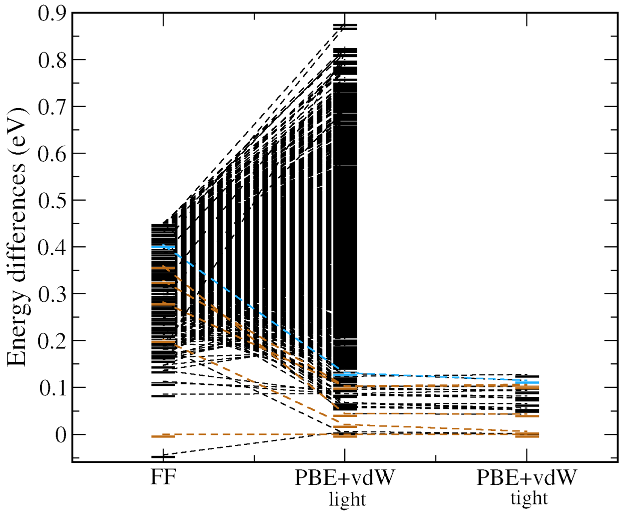


Figure 7.10: Energy hierarchies for conformers of Ac-Ala$_4$-LysH$^+$ in the *opls-aa* force field and fully relaxed with DFT-PBE+vdW, with *light* and *tight* settings of FHI-aims. Conformers coming from an unconstrained search in the force field (black) and the lowest energy "pure" 3_{10} (blue) conformers are shown. The low-energy families detailed in Table 7.2 are highlighted in brown. The zero in energy was taken to be the lowest energy conformer in DFT-PBE+vdW.

In order to search more carefully for systematic overestimations, the slightly larger Ac-Ala$_5$-LysH$^+$ (80 atoms) molecule was taken as a test-case. The reason is that in this molecule it is possible to define helices of various types more clearly, since it has one more H-bond available (for n=4 it is impossible to fold a pure α-helix, for example). By constraining specific H-bonds of the molecule, constrained force field searches for n=5 were performed to test all the helical motifs mentioned in Chapter 2, namely, α, 3_{10}, 2_7, and π. The constrained H-bonds were those in the backbone of the helix, avoiding, where possible, H-bonds that were connected to the terminations. The precise H-bonds constrained are highlighted in orange in Figure 7.11, amounting to 1 constrained H-bond for π-helices, 1 for α-helices, 2 for 3_{10}-helices, and 3 for 2_7-helices.

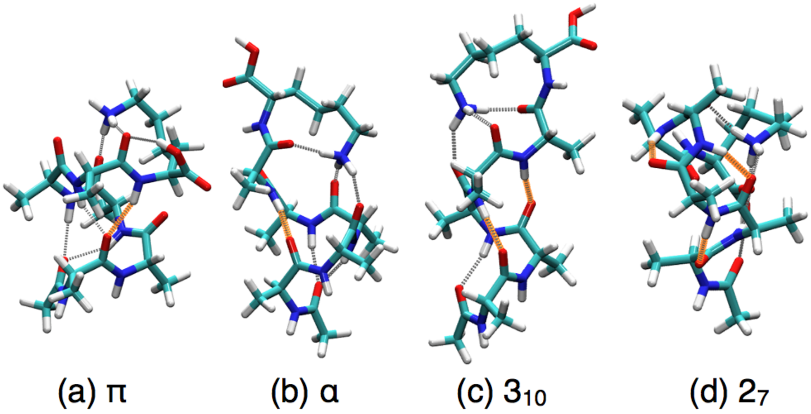


Figure 7.11: Examples of Ac-Ala$_5$-LysH$^+$ conformers coming out from the constrained searches. The constrained H-bonds are highlighted in orange.

The energy hierarchy predicted by the force field, plotted in Figure 7.12, places the α and 3$_{10}$ helices (red and blue points, respectively) as the "pure" helical motifs lying lower in energy, but both are higher in energy than conformers coming from the unconstrained search (black circles). The lowest energy conformer from the α-helical and 3$_{10}$-helical searches are 0.1 and 0.54 eV higher in energy, respectively, than the lowest "unconstrained" conformer. The 2_7 and π helices lie even higher in energy. Relaxations using DFT-PBE+vdW of the 10 lowest energy conformers from each search showed that the 3_{10} conformers are systematically overestimated in the force field also in this case, being predicted by PBE+vdW to occupy a much lower and relevant energy range, as seen in Figure 7.13. The relaxation with PBE+vdW does not induce any significant structural changes and the helices maintain the 3_{10} character. The same does not happen for the 2_7 and π helices, where either the conformers relax to α or mixed helices that are sampled already in the unconstrained search, or they stay substantially removed in energy ($>$0.26eV). The α-helical conformers stay roughly in the same energy window when relaxed with PBE+vdW (also seen in Figure 7.13).

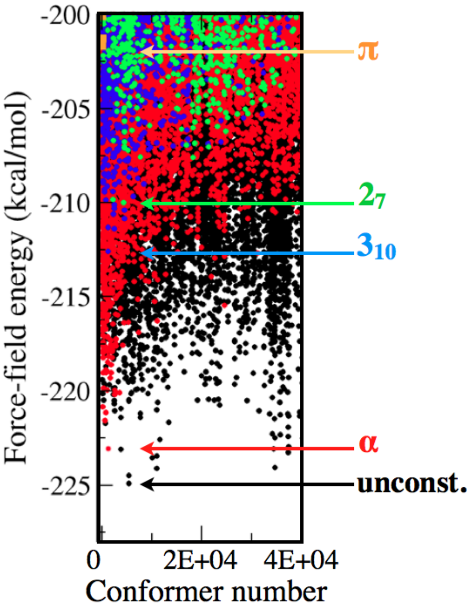


Figure 7.12: Energies of the Ac-Ala$_5$-LysH$^+$ force field conformers sorted by the order they were found in the basin-hopping search. Different constrained searches for the following helical motifs are shown: α-helices in red, 3_{10}-helices in blue, 2_7-helices in green, and π-helices in yellow. The lowest energy conformer for each search is pointed by an arrow.

In Figure 7.13, the energy hierarchies from OPLS-AA and PBE+vdW of all the relaxed conformers from the α (100), 3_{10} (20), and unconstrained (300) searches for n=5 are plotted. The relative energies of the first 300 conformers coming from the unconstrained search cover approximately the equivalent of the conformers ranked 0-100 in the relaxations of n=4 ($\approx$7kcal/mol or 0.3eV). Although this energy range is enough to sample the α-helices, it is surely not enough to sample the 3_{10}-helices. [5] Based on these tests, it becomes clear that beyond an unconstrained search, at least a 3_{10}-helical constrained search for the larger n is needed, followed by DFT relaxations.

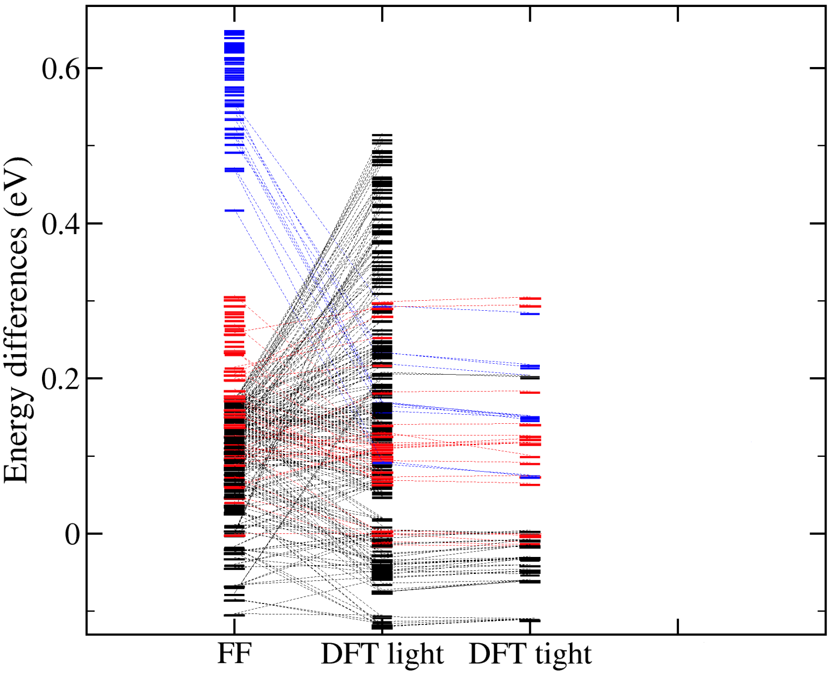


Figure 7.13: Energy hierarchies for conformers of Ac-Ala$_5$-LysH$^+$ in the *opls-aa* force field and fully relaxed with DFT-PBE+vdW. In black, conformers that came from an unconstrained search in the force field, in red conformers that came from an α-helical constrained search, and in blue conformers that came from a 3$_{10}$-helical constrained search. The zero in energy is the lowest energy α-helical conformer.

The unconstrained and constrained searches, as well as the large number of PBE+vdW relaxations aim to maximize the reliance on this search methodology. However, if the force field is blind to particular conformations, it is not obvious that the DFT-PBE+vdW relaxations will find them. Such a situation has been studied by Sucismita Chutia in our group, for a similar molecule (Ac-Phe-Ala$_5$-LysH$^+$). She illustrated the problem for the case of a bifurcated H-bond, which is possible in DFT-PBE+vdW, but not usually within standard force fields. In her case, the DFT-PBE+vdW minimizations find these “bifurcated H-bonds” minima, starting from different force field conformations. In order to obtain better results from the conformational search strategy, possible alternatives would be re-parametrizing force fields to describe better gas-phase properties of polypeptides (e.g. based on electronic-structure calculations), or performing searches directly with DFT (e.g. large-scale replica-exchange molecular dynamics), only starting from some FF-generated structures.

It should be noted that the larger the molecules become, the more difficult it becomes to explore the conformational space. Not only the DFT relaxations become more expensive but, since the conformational space is larger, more and more relaxations are required to span e.g. the lowest 7kcal/mol (0.3eV) energy

[5] The details about the geometries of the low-energy conformers for this case will be given in the next chapter, together with the searches and analysis performed for all other n up to 8.

window of the unconstrained force field searches. It will be seen, though, that computing relative free energies, destabilize considerably non-helical/compact conformers with respect to helical ones. Since the helical conformers are of particular importance in the context of this work, they are being explicitly considered here.

The next chapter will deal with characterizing the low energy-conformers for n=4-8 and assessing the quality of the PBE+vdW functional to describe relative energies for these molecules.

7.3 Summary

In this Chapter, a two-step procedure for searching the conformational space of Ac-Ala$_n$-LysH$^+$ was presented. The first step consists of a force field basin-hopping search, used only as a structure generator, while the second step consists of hundreds of relaxations using DFT, which gives reliable energy hierarchy. The capabilities and limitations of this procedure were discussed for n=4 and 5, finding that the OPLS-AA force field has systematic energy overestimations of certain conformers, in particular 3_{10} helices. The strategy employed here makes it very likely that the lowest energy structure motifs are sampled, although it is impossible to rule out grotesque systematic errors in the force field. In any case, care must be taken to account for systematic force field errors (at least the ones that are known, e.g. 3_{10}-helices).

Chapter 8

Onset of helical preference in the Ac-Ala$_n$-LysH$^+$ series

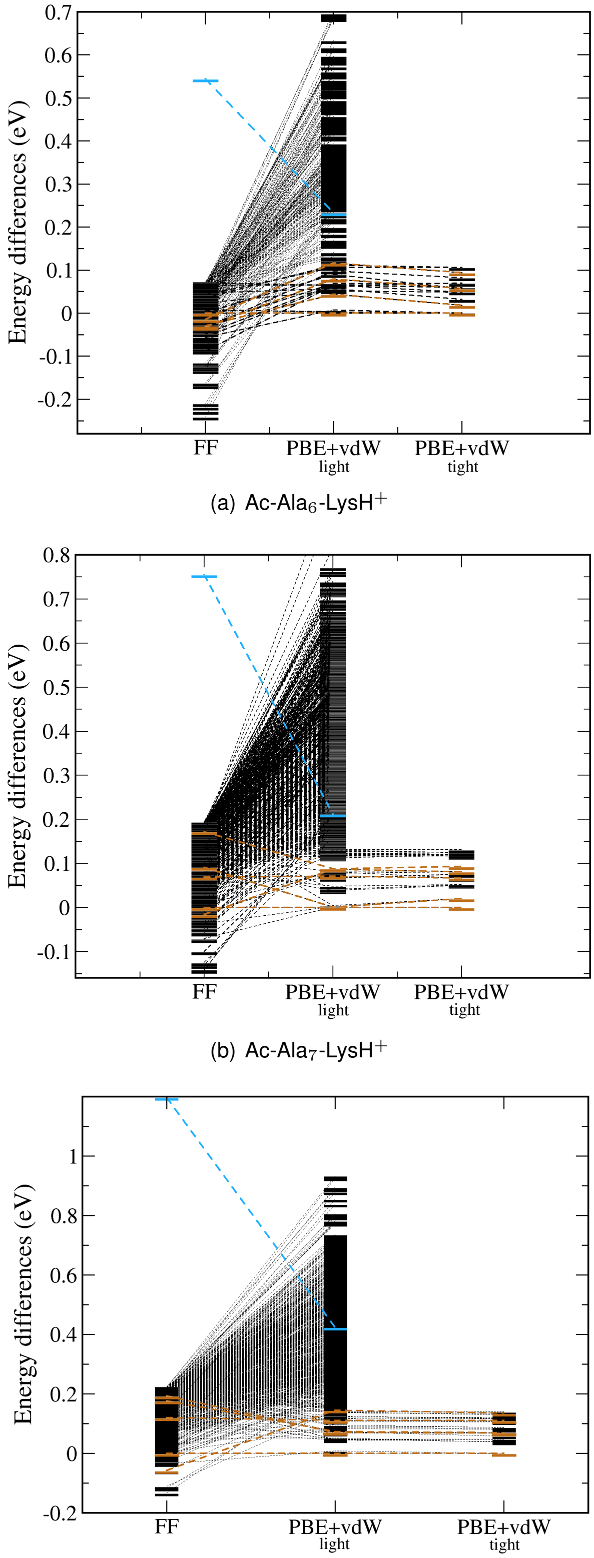
First-principles structure predictions for Ac-Ala$_n$-LysH$^+$, n=4-8

The goal in this Chapter is to use the search strategy described in the previous chapter to determine energetically-preferred stable conformational minima in DFT-PBE+vdW for Ac-Ala$_n$-LysH$^+$, n=4–8. Here it is found that the preference for α-helical conformers becomes dominant at n ≈7–8, which can be connected to the experimental work of Jarrold *et al.* [3], discussed in Section 6.3.2. The accuracy of the PBE+vdW functional for the energy hierarchies presented here is also studied and the addition of van der Waals interactions prove to be essential for an accurate understanding of this problem. The helices are further stabilized by computation of vibrational free energies. A short version of this search and conformer characterization for n=5 was published in Ref. [4].

8.1 Predicting stable conformations with DFT-PBE+vdW

The search strategy discussed and shown for n=4 and 5 in Chapter 7 is used here on the Ac-Ala$_n$-LysH$^+$ molecules with n= 4, 5, 6, 7, and 8. The largest molecule (n=8) contains 110 atoms. Two types of basin hopping searches were performed for each n, namely, an unconstrained search and another where a 3_{10} helical constraint on the geometries was imposed. All searches were run with the optimal parameters ($N_{modes} = 15$, $e_{max} = 50$ kcal/mol, $e_{thresh} = 10^{-4}$ kcal/mol) for the TINKER-*scan* program discussed in Chapter 7. For all unconstrained searches, $O(10^5)$ conformers are found, with the number increasing with increasing size of the molecule, as expected. From these searches, besides the n=4 and 5 relaxations already discussed in the previous chapter, 300 conformers of n=6, 800 of n=7, and 820 of n=8 were fully relaxed with DFT-PBE+vdW. These numbers correspond to *at least* the lowest 7 kcal/mol (0.3 eV) energy window of the force-field search for each n (the energy range explored can be explicitly seen in Figure 8.1). Plots similar to the ones shown in the previous chapter, comparing the force-field, PBE+vdW *light* and PBE+vdW *tight* energy hierarchies for n=6, 7, and 8 searches are shown in Figure 8.1. The conformers relaxed from the unconstrained searches are sorted into H-bond families, in the same way that was done for n=4 in the previous chapter. Among them, 71 different families were identified for n=5, 90 for n=6, 380 for n=7, and 392 for n=8.

A detailed account of the H-bond networks of the families within ∼3 kcal/mol (∼0.13eV) of the lowest



(a) Ac-Ala$_6$-LysH$^+$

(b) Ac-Ala$_7$-LysH$^+$

(c) Ac-Ala$_8$-LysH$^+$

Figure 8.1: Energy hierarchies for conformers of Ac-Ala$_n$-LysH$^+$ (n = 6, 7, 8) in the *opls-aa* force field and fully relaxed with DFT-PBE+vdW. Conformers coming from an unconstrained search in the force field (black) and the lowest energy "pure" 3_{10} (blue) conformers are shown. The lowest energy conformers from the families discussed in Table 8.1 are highlighted in brown. The zero in energy was taken to be the lowest energy conformer in DFT-PBE+vdW.

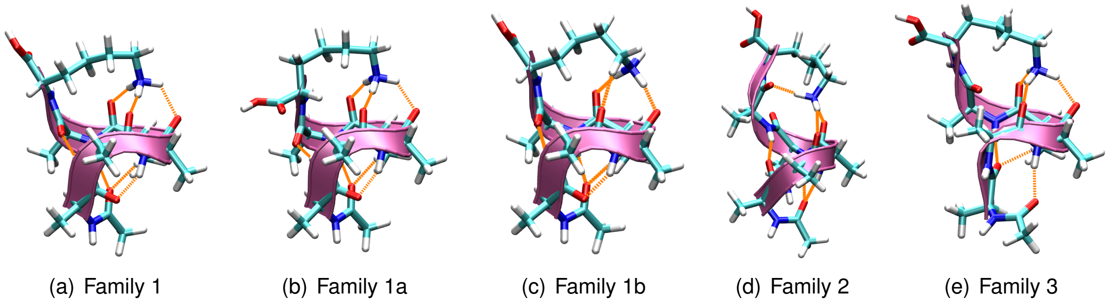

(a) Family 1 (b) Family 1a (c) Family 1b (d) Family 2 (e) Family 3

Figure 8.2: Lowest five families of n=5.

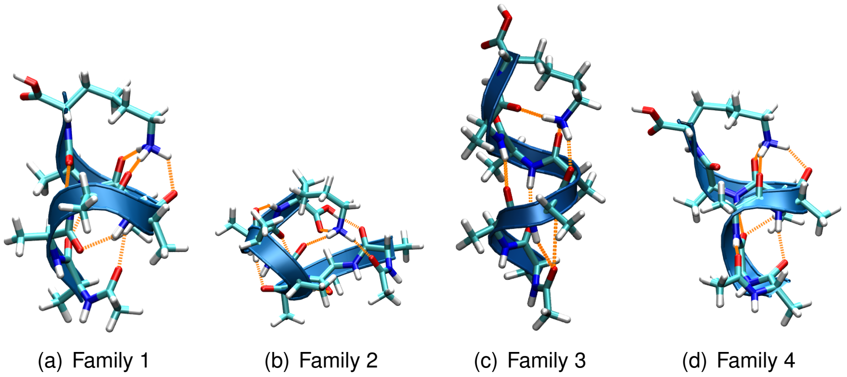
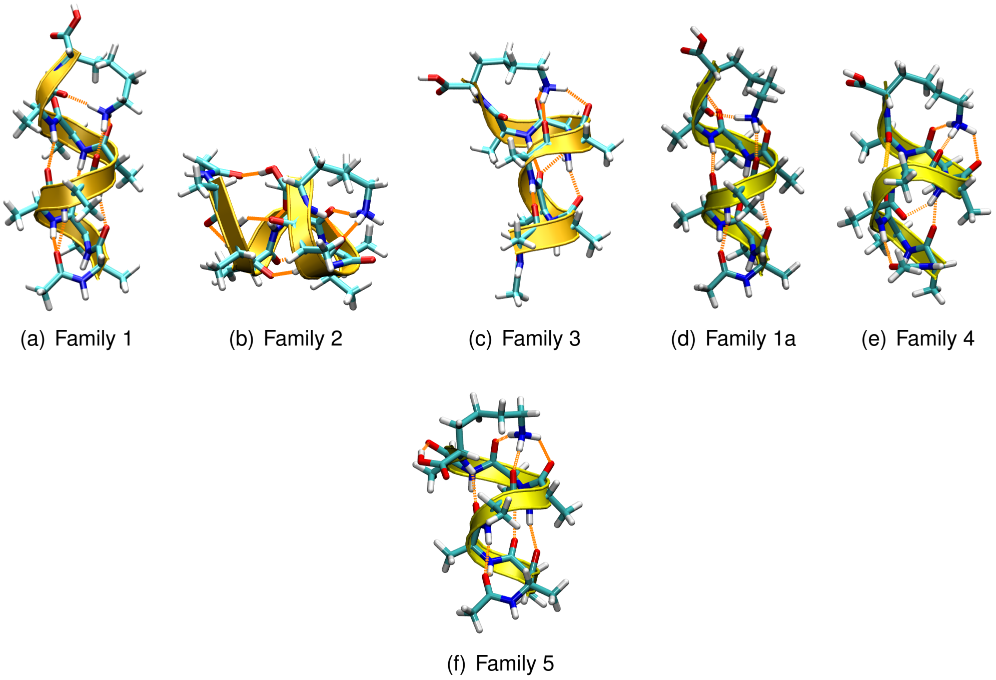

(a) Family 1 (b) Family 2 (c) Family 3 (d) Family 4

Figure 8.3: Lowest four families of n=6.


(a) Family 1 (b) Family 2 (c) Family 3 (d) Family 1a (e) Family 4

(f) Family 5

Figure 8.4: Lower energy families of n=7.

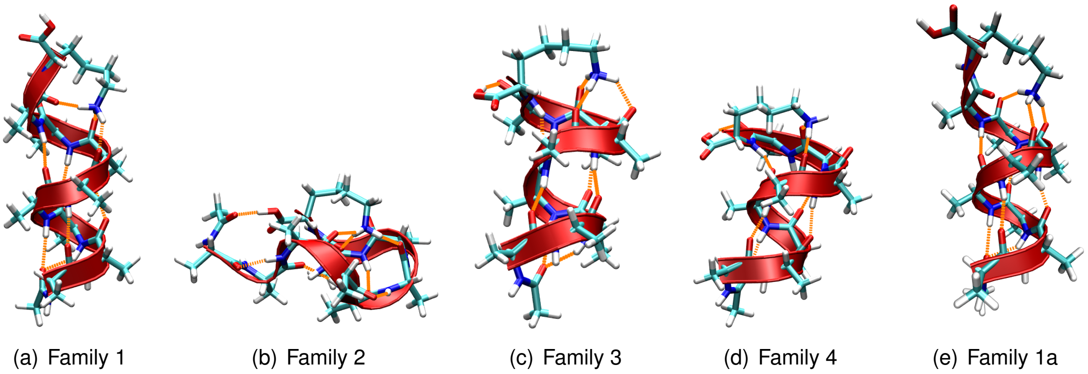


Figure 8.5: Lowest five families of n=8.

energy conformers for n=5-8 is given in Table 8.1. In Table 8.1, the H-bonds are labeled from the Ac termination (N-terminus) to the LysH$^+$ termination (C-terminus) as was schematically represented for n=4 in Figure 7.8. Family 1 is always the lowest energy family for each n. The detailed energetic ordering of these families for n=4–8, calculated with *tight* settings is also presented in Table 8.1. The 3D structures of all the respective conformers are shown in Figures 8.2–8.5 The low energy conformers for n=4 have been characterized in the previous chapter (see Table 7.2 and Figure 7.9), and the discussion will not be repeated here.

A discussion about the different H-bond families presented in Table 8.1, and shown in Figs. 8.2 – 8.5, for n=4-8, follows:

Table 8.1: H-bond network of the families within ~3kcal/mol (~0.13eV) of the lowest energy conformer for n=4-8. Energy differences (E_{rel}) obtained with DFT-PBE+vdW (PES only, *tight* settings). All energies in eV, lowest energy α-helical family in bold and red.

n	Oxy.	Family 1	Family 1a	Family 1b	Family 2	Family 3	Family 4
5	O(Ac)	NH4 (α) NH5 (π)	NH4 (α) NH5 (π)	NH4 (α) NH5 (π)	NH3 (3_{10}) NH4 (α)	NH3 (3_{10})	
	O1	NH3 (2_7)	NH3 (2_7)	NH3 (2_7)	NH5 (α)	NH4 (3_{10}) NH5 (α)	
	O2	NH$_3^+$	NH$_3^+$	CH (Lys)	NH$_3^+$	NH$_3^+$	
	O3	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$	NH$_3^+$	
	O4	NH$_3^+$	NH$_3^+$	NH$_3^+$	free	NH$_3^+$	
	O5	NH2 (inverted)	NH2 (inverted)	NH2 (inverted)	NH$_3^+$	NH$_3^+$ weak	
	O(COOH)	free/NH6	free	free/NH6	free	free	
	E_{rel}(eV)	0.00	0.05	0.06	**0.10**	0.11	
6	O(Ac)	NH5 (π)			NH$_3^+$	NH3 (3_{10}) NH4 (α)	NH3 (3_{10})
	O1	NH6 (π)			OH	NH5 (α)	NH4 (3_{10})
	O2	NH4 (2_7)			NH4 (2_7)	NH6 (α)	NH5 (3_{10}) NH6 (α)
	O3	NH$_3^+$			NH6 (3_{10})	NH$_3^+$	NH$_3^+$
	O4	NH$_3^+$			NH$_3^+$	NH$_3^+$	NH$_3^+$
	O5	NH$_3^+$			NH7 (2_7)	free	NH$_3^+$
	O6	NH3 (inverted)			NH2 (inverted)	NH$_3^+$	NH$_3^+$ weak
	O(COOH)	free			NH$_3^+$	free	free
	E_{rel}(eV)	0.00			0.02	**0.06**	0.09
7	O(Ac)	NH3 (3_{10}) NH4 (α)	NH4 (α)		OH	NH3 (3_{10})	NH3 (3_{10})
	O1	NH5 (α)	NH5 (α)		NH3 (2_7)	NH4 (3_{10})	NH6 (π)
	O2	NH6 (α)	NH6 (α)		NH7 (π)	NH5 (3_{10})	NH7 (π)
	O3	NH7 (α)	NH7 (α)		NH5 (2_7)	NH6 (3_{10}) NH7 (α)	NH5 (2_7)
	O4	NH$_3^+$	NH$_3^+$		NH$_3^+$	NH$_3^+$	NH$_3^+$
	O5	NH$_3^+$	NH$_3^+$		NH$_3^+$	NH$_3^+$	NH$_3^+$
	O6	free	NH8 weak		NH$_3^+$	NH$_3^+$	NH$_3^+$
	O7	NH$_3^+$	NH$_3^+$		NH2 (inverted)	free	NH4 (inverted)
	O(COOH)	free	free		free	free	free
	E_{rel}(eV)	**0.00**	0.08		0.02	0.07	0.09
8	O(Ac)	NH3 (3_{10}) NH4 (α)	NH3 (3_{10}) NH4 (α)		NH4 (α)	OH	NH3 (3_{10}) NH4 (α)
	O1	NH5 (α)	NH5 (α)		NH5 (α)	NH3 (2_7)	NH5 (α)
	O2	NH6 (α)	NH6 (α)		NH6 (α)	NH5 (3_{10})	NH6 (α)
	O3	NH7 (α)	NH7 (α)		NH7 (α)	NH$_3^+$	NH7 (α)
	O4	NH8 (α)	NH8 (α)		NH8 (α)	NH6 (2_7) NH9 (π)	NH8 (π)
	O5	NH$_3^+$	NH$_3^+$		NH$_3^+$	NH8 (3_{10})	NH$_3^+$
	O6	NH$_3^+$	NH$_3^+$		NH$_3^+$	NH$_3^+$	NH$_3^+$
	O7	free	NH$_3^+$		NH$_3^+$	free	NH$_3^+$
	O8	NH$_3^+$	NH$_3^+$ (weak)		OH	NH$_3^+$	OH
	O(COOH)	free	free		free	free	free
	E_{rel}(eV)	**0.00**	0.13		0.06	0.07	0.11

- n=5: The lowest energy conformer (Family 1, called “g-1” in Ref. [4]) of n=5 is not a simple helix. It exhibits a particular feature, already seen in Family 3 of n=4: there is one H-bond that goes against the “helical” dipole, highlighted in yellow in Figure 8.7(b), that is labeled “inverted” in Table 8.1. This conformer has a π-helical and a 2$_7$-helical loop starting from the Ac termination, but due to this inverted H-bond joining the C and N termini, it is much more compact than a pure helix. The next two families appearing for this n are extremely similar to Family 1 having the exact same H-bond connection for O(Ac), O1, O3, O4, and the inverted H-bond to O5. The only difference lies in the connection of O2 and O(COOH), but they do not cause an appreciable overall structural change in the molecule (see Figure 8.2 for details of the 3D structures). These changes are very similar to what is observed for Families 1a and 1b of n=4. Thus, for n=5, the lowest energy family is labeled Family 1, and these variations are also labeled Families 1a and 1b. Family 2 of n=5 is the α-helix (called “α-1” in Ref. [4]), presenting the characteristic α-helical H-bond between O1 and NH5. Family 3 is a mixed helix (called “α-2” in Ref. [4]) that has one 3$_{10}$ H-bond in the Ac termination, then one bifurcated 3$_{10}$-α H-bond and then the CO groups connecting to the Lys NH$_3^+$. The names “g-1”, “α-1”, and “α-2” will be used again throughout this thesis to refer to these conformers.

- n=6: The lowest energy conformer of n=6 is again not a simple helix, exhibiting the same 2$_7$ helical loop followed by an inverted H-bond, as was seen for the lowest energy of n=5 (g-1 motif). In this case, since the molecule is larger, these two H-bonds are preceded by one more π-helical loop, involving the Ac termination (see Fig. 8.3 and 8.7(c)). Family 2 of n=6 is a globular/compact conformer, with the CO coming from the Ac termination connecting directly to the NH$_3^+$ of the lysine. It also presents the inverted H-bond seen in Family 1, although the conformer is so compact that a “macro-dipole” of the molecule is hardly definable. Family 3 here is the α-helical conformer, the same motif of Family 2 of n=5, with one more α-helical bond. Family 4 is the mixed 3$_{10}$-α helix, characterized by 3_{10} helical network starting in the N-terminus and going up Ala-residue just before the one that connects to the LysH$^+$ side-chain. The bond then bifurcates to α and the Lys-termination connecting H-bonds assume the same pattern as for the α-helices.

- n=7: Here the α-helical conformer, the same motif as Family 3 of n=6 and Family 2 of n=5, is the lowest energy one (Family 1). At 0.08eV there is a very similar motif to Family 1, differing only in the connection of one CO group close to the Lys termination. This family is labeled Family 1a. Family 2 of n=7, very close in energy (0.02eV) to the lowest energy conformer, is again a very compact/globular conformer, with the CO from the Ac termination connecting directly to the OH group from the Lys COOH. In fact, upon close inspection, this conformer forms a turn, best seen in the 3D representation shown in Figure 8.4. Family 3 is the mixed α-3$_{10}$ helix, already found for the lower n, with mainly 3$_{10}$ loops and one bifurcated 3$_{10}$-α bond just before the CO groups connecting to the Lys termination. The inverted bond motif (g-1 motif) found for the lowest energy conformers of n=5 and 6 is also seen here, but it is now Family 4, which is 0.09 eV higher than the α-helix. Due to geometric constraints, this motif, for n=7, can only be stabilized by the appearance of a 3$_{10}$ helical loop on the Ac termination, which seems to be energetic unfavorable. Larger n would allow (geometrically) the appearance of more π-helical loops to stabilize this motif, but π helices have been shown to be the least energetically favored in the limit of infinite polyalanine helices [325]. The “g-1” motif is, thus, unlikely to appear in the low energy range for higher n. Finally, a last family, not explicitly shown in Table 8.1 but pictured in Figure 8.4, appears at 0.11eV, consisting of a ($\alpha \rightarrow \alpha \rightarrow \alpha \rightarrow \alpha/\pi \rightarrow$ Lys) sequence of H-bonds, being thus mainly an α-helix with just one bifurcated α-π H-bond.

- n=8: Family 1 is again the α-helical conformer in this case. At 0.13 eV there is a family very similar to Family 1, that has the same H-bond connections in the backbone but has the lysine side-chain connecting to the CO groups in the same way that it does for the α-3_{10} mixed helices of the other n (see Fig. 8.5(e)). This family is labeled 1a, and it comes from the third lowest energy structure in the force-field hierarchy. The mixed α-3_{10} helices are also among the relaxations for n=8, but they stay at least 0.14 eV higher in energy than the lowest energy conformer. Family 2 is a very compact/globular conformer, like in n=7, but is a bit more removed in energy (0.06eV), although not much. Family 3 is a mixed α-π helix, in the same motif described for Family 5 of n=7, and is essentially the same as Family 4 here, differing just by a bifurcated H-bond close to the Ac termination.

The trend for α-helical preference becomes stronger in PBE+vdW as n increases, based on the data presented. This is better visualized in Figure 8.6 which shows the relative energies (PBE+vdW, FHI-aims *tight* settings), with the respective geometries and the α-helical conformers highlighted in red. The structures of the lowest energy conformers, for each n, are shown in more detail in Figure 8.7, where the characteristic inverted H-bonds are highlighted in yellow for n=5 and 6.

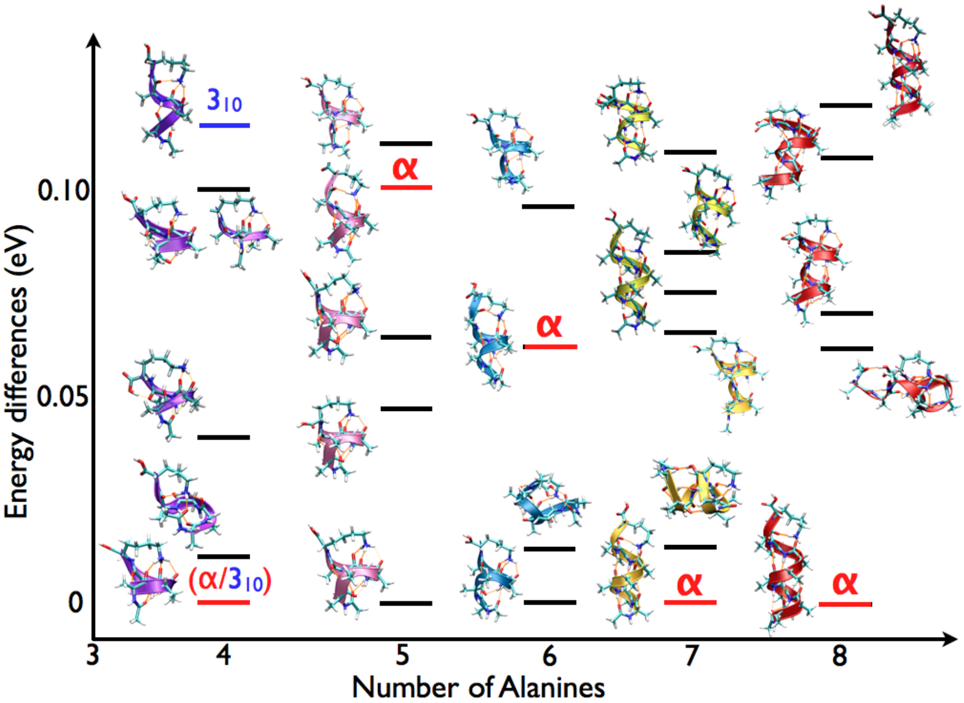


Figure 8.6: Energy differences for lowest energy conformers corresponding to different families of n=4-8. Highlighted in red are the α-helical conformers for each n.

The OPLS-AA force field predicts quite a different trend for the low energy conformers of n=4–8, and does not predict the α-helical stability seen for n=8 in PBE+vdW. For n=4 and 5 the lowest energy conformers in OPLS-AA belong to the same family as the lowest energy conformer predicted by PBE+vdW, although they do not relax to the global lowest energy in PBE+vdW. For n=6 and 7 the lowest energy conformers predicted by the force-field are very similar to each other (not to DFT). Starting from the N-terminus, there is one α-helical H-bond (two for n=7) then one π-helical H-bond, then a bond to the OH group from the COOH and the rest of the CO groups interacting with the NH_3^+ termination. These conformers turn out to be at least 0.16eV higher than the lowest energy conformer in PBE+vdW.

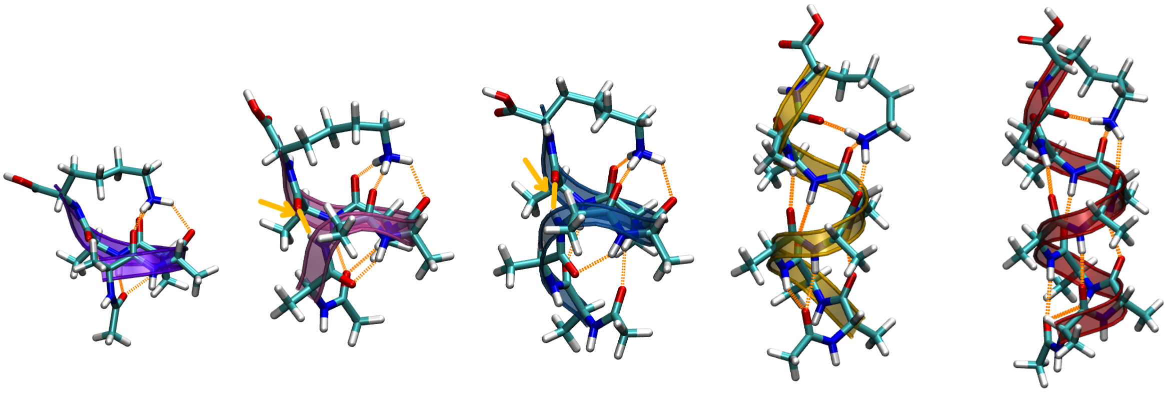

(a) n=4 (b) n=5 (c) n=6 (d) n=7 (e) n=8

Figure 8.7: Lowest energy conformer in PBE+vdW for Ac-Ala$_n$-LysH$^+$, n=4-8. Inverted H-bonds for n=5 and 6 are highlighted in yellow and pointed by an arrow.

Finally, for n=8 the force-field predicts the two lowest energy structures to be of the type (starting from the N-terminus): $\alpha \rightarrow \alpha/\pi \rightarrow \pi \rightarrow \pi \rightarrow$ Lys-termination. Only the third structure is of α helical type, and is in fact Family 1a, detailed in Table 8.1, which stays 0.13eV removed from the lowest energy conformer. The mixed α/π helices found as the lowest structures in the force field are predicted to be much higher in energy in PBE+vdW (at least 0.22eV higher than the lowest energy conformer).

The results obtained here seem to agree with the experimental observation (Ref. [3] and Section 6.3.2) of the helical preference onset at n=8, though they arise from a completely different analysis (no water adsorption here). The preference, however, is not that strong even for n=8, if only PES total energies are taken into consideration, as was done here. Since it was already discussed in the previous chapter that it is not impossible to be missing a few low energy conformer motifs, making a strong statement based solely on these numbers would be dangerous. Especially for n=8, where the conformational space is the largest among this set of molecules, and thus the most unlikely to be converged, the energy hierarchy could hypothetically change, even having performed the large amount of DFT relaxations that were performed here. Perhaps more important, though, as was already discussed in Chapters 2 and 4, the actual quantities that should be analyzed at finite temperatures, and that are also accessible in finite-temperature experiments, are *free energy* differences, instead of just the potential energies shown up to now. In the next sections, the role of vdW interactions and free energies in stabilizing the structures discussed here will be investigated.

8.2 Impact of different functionals and van der Waals interactions on the stabilization of conformers

When dealing with DFT, a common question is how different parametrizations of exchange-correlation functionals would affect the results. In this work, we have extensively tested this influence for the specific case of n=5. Since it was already discussed that Families 1, 1a, and 1b for this n are essentially the same, Families 1, 2 and 3 were chosen as the test-cases. In addition, now the pure 3_{10} helical conformer, shown in Figure 8.8 with its H-bond network, is also included. This conformer is 0.19 eV above the lowest energy conformer of n=5, using PBE+vdW, *tight* settings.

Oxy.	Bound-to
O(Ac)	NH3 (3_{10})
O1	NH4 (3_{10})
O2	NH5 (3_{10})
O3	NH$_3^+$
O4	NH$_3^+$
O5	NH$_3^+$
O(COOH)	free

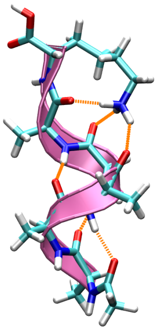
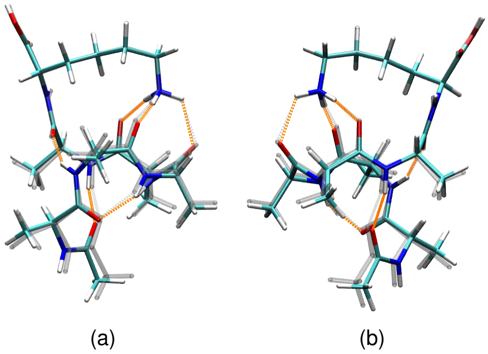


Figure 8.8: H-bond network and the 3D structure of the 3_{10} conformer for n=5.

These molecules are now labeled for their H-bond network character character (as was done in [4]): g-1 for Family 1, α-1 for Family 2, α-2 for (the mixed) Family 3, and 3_{10}-1 for the 3_{10} conformer.

As discussed in Chapter 3, there are many different possible approximations for the exchange-correlation potential. We chose here to consider only some of the most important functionals mentioned there that represent different rungs in the Perdew-ladder and also different "schools" (Perdew, Becke, and Truhlar). The following functionals are tested: (i) GGA's, in the PBE[131], PBE0 [144], and B3LYP [147] approximations ; and (ii) meta-GGA's, in the M06L[377] and M06 [378] approximations [1]. Of these functionals, PBE0, B3LYP, and M06 have a portion of Hartree-Fock exchange, which makes them the so-called "hybrid"-functionals. For each functional the relative energies are calculated with the addition of the van der Waals C_6/R^6 term from the TS-vdW scheme [2], and without.



(a)

(b)

Figure 8.9: Superimposed structures of the g-1 conformer relaxed with PBE (grey-transparent) and B3LYP (solid color). (a) and (b) are just different views.

For consistency, the different energies in each functional were calculated as single point energies at the PBE(+vdW) relaxed structures. Relaxations of the PBE structures using the B3LYP functional[2] for the g-1 and α-1 conformer yielded a Root Mean Square Deviation (RMSD) of only 0.02Å when considering either all or only backbone atoms. For a better visualization the structures of the g-1 conformer relaxed with PBE (grey-transparent) and B3LYP (solid color) are superimposed in Figure 8.9. Thus, the neglect of relaxation for each functional is not expected to change appreciably the results. The relative energetics for all functionals tested can be seen in Figure 8.10. The BLYP values are not highlighted here due to the problems of the TS-vdW scheme to couple to this functional. Nevertheless, the relative energetics would be (for fixed geometries): (i) g-1: 0.0 eV, α-1: 0.026 eV, α-2: 0.092 eV, 3_{10}-1: 0.015 eV for BLYP;

[1] The meta-GGA's were evaluated after PBE self-consistency.
[2] These relaxations were done using the Gaussian03[148] code.

(ii) g-1: 0.0eV, α-1: 0.128 eV, α-2: 0.075 eV, 3$_{10}$-1: 0.100 eV for BLYP+vdW. Notice that even including vdW interactions, the 3_{10} helices are more stable than α-helices in this functional.

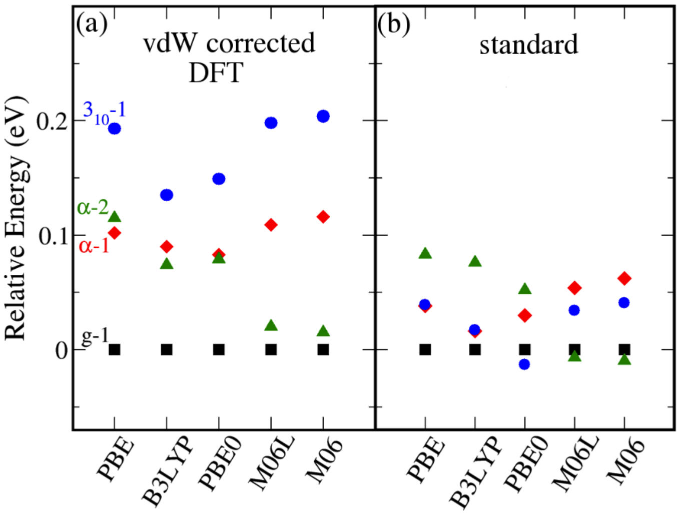


Figure 8.10: Energy hierarchies of conformers g-1 (Family 1), α-1 (Family 2), α-2 (Family 3) and 3$_{10}$-1 (see text) of Ac-Ala$_5$-LysH$^+$ in various functionals: (a) vdW corrected functionals; (b) standard funcionals.

Figure 8.10 shows that as long as vdW effects are considered, the g-1 conformer is the lowest energy one, followed by one of the α-1 or α-2 conformers. The 3$_{10}$-1 conformer is always the highest in energy and stays removed for all vdW corrected functionals. In contrast, when not considering vdW interactions, 3$_{10}$-1 can even have the lowest energy of the set (PBE0, for example) or at least be always energetically competitive, pointing to a completely different PES for secondary structure. Overall, the situation when not including the vdW correction is much less clear. All four conformers lie closer in energy and each functional predicts a different one to have the lowest energy. The M06 functionals[378], although built in order to mimic vdW interactions in the short range, here also exhibit a substantial energetic rearrangement upon inclusion of the long range C$_6$/R^6 correction. This has also been observed in Ref. [199].

Since vdW interactions are known to be of great importance for polypeptides, it is reassuring to see that the relative energetic trends do not depend strongly on the functional used as long as vdW interactions are properly considered.

Another important effect of including the vdW corrections is that in the absence of it, the α-helix crossover shown in Figure 8.6 would not be correctly predicted. To prove this, the first 600 conformers taken from the force-field for n=8 were relaxed using the standard PBE functional. The relative energy hierarchies for this case are shown in Figure 8.11 with the 3$_{10}$- and α-helical conformers highlighted, as well as the lowest energy conformer for PBE.

The α-helical conformer appears among the lowest energy conformers also in PBE. However, it is neither the global minimum, nor is it separated from other conformational motifs. The lowest energy conformer is shown in detail in Figure 8.12, being a mostly 3_{10} helical conformation, just like the mixed helices that were discussed for n=5, 6, and 7 in the PBE+vdW case in Section 8.1. This is in agreement with the fact already discussed in the literature [379, 380] that, in the absence of vdW interactions, 3$_{10}$ helices tend to be over-stabilized. Furthermore, comparing the number of families found in the lowest 0.1eV for PBE and PBE+vdW with n=8, taking exactly the same 600 conformers, there are only 3 different families appearing for the PBE+vdW case, while there are 8 for PBE, including very compact conformers

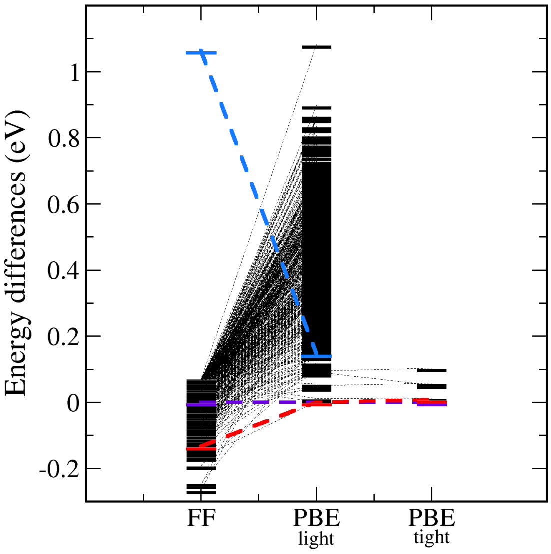


Figure 8.11: Energy hierarchies for n=8 relaxed with the standard PBE funcional, without including van der Waals corrections. The pure 3_{10} (blue), pure α (red) and the lowest energy conformer (purple) in DFT-PBE are highlighted.

Oxy.	Bound-to
O(Ac)	NH3 (3_{10})
O1	NH4 (3_{10})
O2	NH5 (3_{10})
O3	NH5 (3_{10})
O4	NH6 (3_{10}) NH7 (α)
O5	NH_3^+
O6	NH_3^+
O7	NH_3^+
O8	free
O(COOH)	free

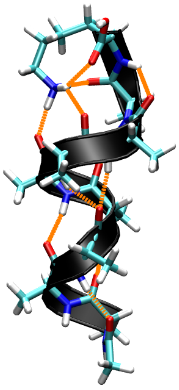


Figure 8.12: H-bond network and the 3D structure of the lowest energy conformer of n=8 using the standard PBE functional.

In order to access the accuracy of the PBE+vdW functional specifically for these molecules, it would be ideal to obtain benchmark data from methods that include outright non-local correlations. The gold standard method, briefly discussed in Chapter 3, would be CCSD(T) [86]. Unfortunately, due to its $O(N^7)$ scaling and slow convergence with basis-set size, this is unfeasible for molecules this size (110 atoms). We are left with the perturbative methods discussed in Chapter 3: RPA and MP2. However, converging relative energies within these methods is not a trivial task, and new basis sets specifically for this purpose had to be developed. A description of how and why these basis sets were developed will be shown towards the end of the thesis, in Chapter 11, with benchmark data for n=5 will be presented and discussed in Chapter 12.

8.3 Impact of free energies on α-helix stabilization.

In this section, the goal is to analyze how the vibrational free energy of Ac-Ala$_n$-LysH$^+$ (n=4-8) affects the energetic ordering of the lower energy conformers. Experiments, like the ones that were discussed in Chapter 6, are conducted at finite temperatures. The molecules are, thus, exploring the free energy surface at that temperature, and the experiments can access free energy differences, so that consideration of these effects is important. Here only the vibrational harmonic free energies of the molecules will be considered. The "configurational" free energy will not be explicitly included in the calculations below. Evaluation of this quantity would involve a very broad sampling of the possible configurations at a given temperature, something that can presently only be achieved with force fields. Since we have seen that the force fields would give wrong Boltzmann weights (e.g. because of energy over-estimations) for the molecules studied here, this has not been attempted.

The harmonic approximation normal mode analysis described Section 4.1 was used to estimate vibrational free energy differences of the lowest energy conformers of the families discussed in Section 8.1 for Ac-Ala$_n$-LysH$^+$, n=4-8. The equation used is Eq. 4.20, repeated here for clarity:

$$F_{vib}^{total}(T) = V_{BO} + U_{vib} - TS_{vib} = V_{BO} + \sum_{i=1}^{3N-6}\left[\frac{\hbar\omega_i}{2} + k_B T \ln\left(1 - \exp^{-\frac{\hbar\omega_i}{k_B T}}\right)\right], \qquad (8.1)$$

where ω_i are the harmonic normal modes of vibration, U_{vib} the vibrational internal energy (given by Eq. 4.21), S_{vib} the vibrational entropy, and V_{BO} is here taken to be the DFT-PBE+vdW total energy. The free energy contains the contributions from the internal thermal energy of the molecule and contributions from the entropy and temperature.

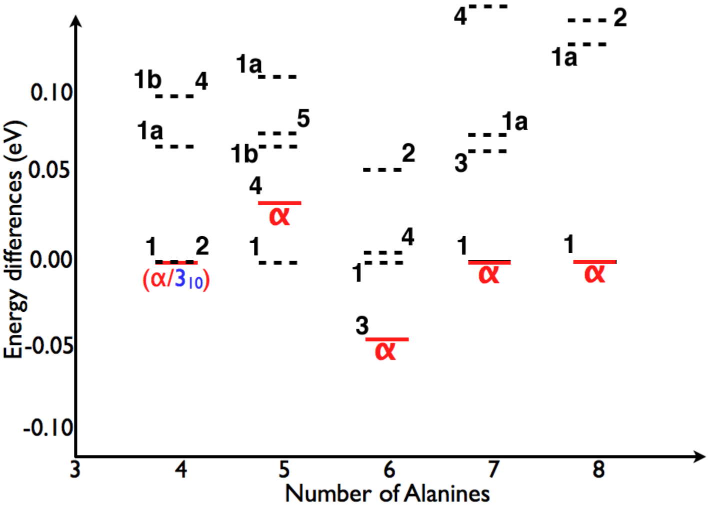


Figure 8.13: Relative harmonic free energies at T=300 K of the lowest-energy conformers of the families discussed for each n in . Only the ones lying within 0.15 eV from the PES minima for each n are shown, labeled by their respective family number. α-helical conformers are highlighted in red.

The relative free energies at T=300K, as well as the relative energies of PBE+vdW (*tight* settings), and the zero point energy contribution are reported in Table 8.3. In Figure 8.13, the free energy differences at 300K, with the α-helices highlighted and the families labeled by their number, are shown. The reference, in all cases, is taken to be the lowest PES total energy conformer (same as in Figure 8.6).

As can be seen in Table 8.3 and Figure 8.13, the stability of α-helices is enhanced by adding vibrational free energies, for all n, indicating that these helices are favored by vibrational entropy. For n=8 the energetic interval between the α-helical conformer and the next globular one at T=300 K is of 0.14 eV (the one at 0.13 eV is also an α-helix), making it extremely likely that in experiment this would be the first conformer of the series where mostly only α-helices would be present, now solidifying our predictions.

Here it is important to stress that all helical conformers are observed to be stabilized over compact ones, which connects well with the experimental observation of Ref. [352], where helices in Ala-Gly

Conformer	Family	DFT-PBE+vdW	+ZPE	ΔF (300 K)
Ac-Ala$_4$-LysH$^+$	1	0.00	0.00	0.00
	2	0.01	0.01	0.00
	1a	0.04	0.05	0.06
	3	0.10	0.12	0.16
	1b	0.10	0.10	0.10
	4	0.12	0.12	0.10
Ac-Ala$_5$-LysH$^+$	1	0.00	0.00	0.00
	1a	0.05	0.06	0.11
	1b	0.06	0.06	0.06
	4	**0.10**	**0.07**	**0.04**
	5	0.11	0.08	0.07
Ac-Ala$_6$-LysH$^+$	1	0.00	0.00	0.00
	2	0.02	0.02	0.05
	3	**0.06**	**0.03**	**-0.05**
	4	0.09	0.06	0.01
Ac-Ala$_7$-LysH$^+$	**1**	**0.00**	**0.00**	**0.00**
	2	0.02	0.07	0.22
	3	0.07	0.06	0.06
	1a	0.08	0.08	0.07
	4	0.09	0.11	0.15
Ac-Ala$_8$-LysH$^+$	**1**	**0.00**	**0.00**	**0.00**
	2	0.06	0.06	0.14
	3	0.07	0.09	0.16
	4	0.11	0.13	0.19
	1a	0.13	0.13	0.13

Table 8.3: Energy differences of lowest energy families for Ac-Ala$_n$-LysH$^+$, n=4-8: DFT-PBE+vdW (PES only), zero-point corrected total energy, and DFT-PBE+vdW harmonic free energy ΔF at 300 K. All energies in eV, lowest energy α-helical family in bold.

polypeptides are stabilized as the temperature is raised. Therefore, even if there could be a few more non-helical families in the low energy range for n=8, they would be considerably de-stabilized with respect to the α-helices that are seen there. The 3_{10} helix, which should be the next most stable pure helix, remains around 0.38 eV removed in energy for n=8, upon calculation of free energies at 300K.

Additionally, it might strike as odd that the conformers of n=8 belonging to Families 3 and 4, which are both essentially α-helical (see Table 8.1), are destabilized over Family 1 (the lowest energy α-helix). Families 3 and 4 contain a bond of the CO group from the 8^{th} Ala to the OH group from COOH. As can be seen in Figure 8.5, this bond creates strain, and has a direct impact on the harmonic vibrational frequencies and free energies, as will be seen below.

8.3.1 Reasons for the stabilization of helices

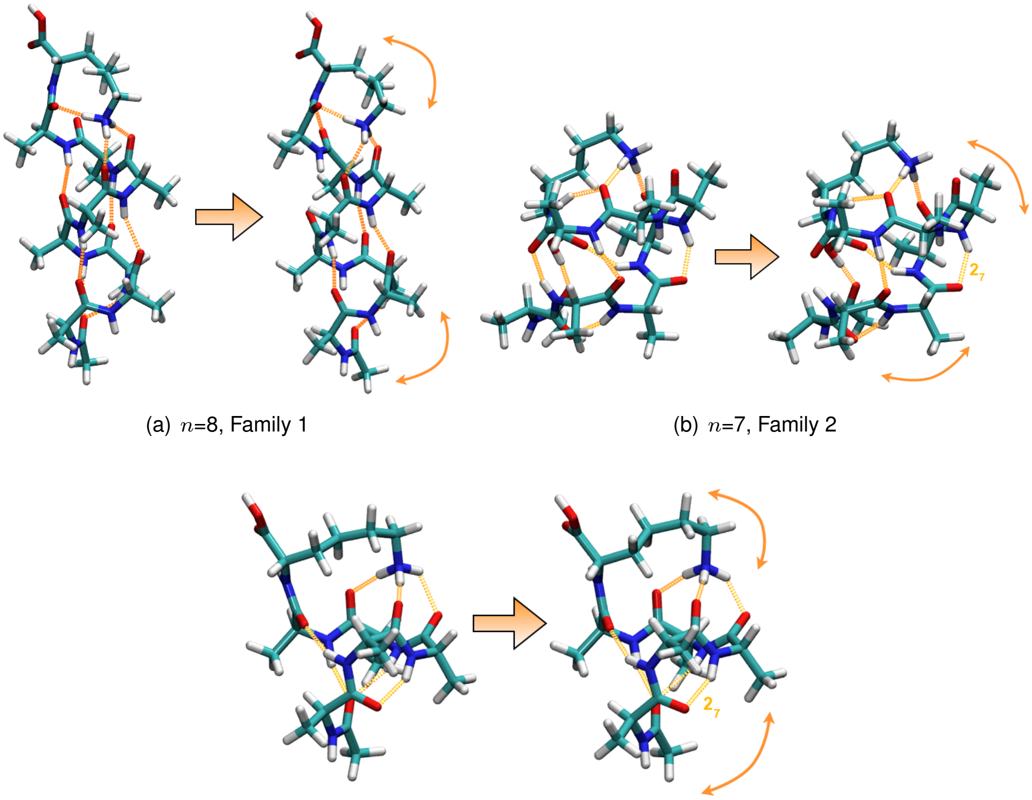
The observed stabilization of the α-helices can be understood by analyzing the first vibrational modes of different conformations. These modes are dominant on the evaluation of the harmonic free energies, as can be seen from Equation 8.1. The logarithmic term is negative and the higher in frequency the first vibrational mode is, the more it induces a destabilization in the computation of the free energy. The first mode, in particular, is of a delocalized character through the whole molecule, and if much enhanced, would correspond to a global bending of the molecule around its middle residues, dominated by the movement of the terminations. The deformation coming from this first vibrational mode is shown as an example in Figure 8.14 for the α-helix of n=8 (Family 1), the compact/turn conformer of n=7 (Family 2), and the conformer with the inverted H-bond defining a “non-simple helix” of n=5 (Family 1).[3]

For structures that exhibit a periodicity (e.g. α and 3_{10}-helices, or fully extended structure) the vibrational mode related to the bending of the terminations, shown in Figure 8.14, is softer. In fact, from calculations of the fully extended structure (FES) of n=15, which will be discussed in Chapter 10, it is seen that the FES has the lowest vibrational mode around only 2 cm^{-1}, while the α and 3_{10} helices of the same size have the first vibrational mode at 11 and 10 cm^{-1} respectively [4]. The “compact” or not strictly helical conformers discussed for n=5-8 have H-bond patterns that makes this vibration harder. In the cases shown in Figure 8.14 for n=5 (Family 1) and n=7 (Family 2), the 2_7 loop that these conformers exhibit (marked in the figure), oppose this vibration, making it effectively harder to happen. In Table 8.4, the position of the first mode of all conformers belonging to the families discussed in the previous section, for each n are shown. The α-helices presented here have a very low first vibrational mode between 8 and 13 wave numbers, depending on the size of the molecule (marked in red in Table 8.4). The other compact/not strictly helical conformers have a first vibrational mode at least at 20 wave numbers, for example: 28cm^{-1} for Family 2 of n=7, and 22cm^{-1} for Family 1 (g-1 motif) of n=5. The g-1 motif has the first vibrational mode lying around 20cm^{-1} for n=5, 6, and 7 (marked with * in Table 8.4).

As an example of the differences in distribution of the low energy modes for n=5, in Figure 8.15, a zoom into the lower wavenumber range of the normal modes of Family 1 (“g-1”) and of Family 2 (“α-1”) is shown, with an arrow pointing to the first vibrational mode. The accuracy of these modes can be inferred by the scatter of the first 6 rotational-translational modes, also shown in the figure, which should all be zero for perfect accuracy. They present, at the most, a 1cm^{-1} deviation. For these conformers, the

[3] Ismer *et al.* [324, 325] reported that the lowest three vibrational modes for an infinite helix are stretching, twisting and bending, respectively (i.e. bending is not the lowest vibrational branch). The relative stabilization between different helical motifs, in their case, could also be rationalized in terms of the first few vibrational modes. Here, instead, the movement of the terminations, non-existent in the infinite helix, composes the first vibrational mode. The second vibrational mode also corresponds to a bending of the terminations towards each other, but in another direction.

[4] While it is arguable that the accuracy of the calculation of the harmonic vibrational frequencies is smaller than 2 cm^{-1}, it is safe to say that this mode, for the FES of n=15, lies much lower in frequency than for the α and 3_{10} helices.



(a) n=8, Family 1

(b) n=7, Family 2

(c) n=5, Family 1

Figure 8.14: Deformation caused by the first vibrational mode (if enhanced 10 times) for: (a) Ac-Ala$_8$-LysH$^+$, Family 1 (α-helix); (b) Ac-Ala$_7$-LysH$^+$ Family 2 (compact/turn); (c) Ac-Ala$_5$-LysH$^+$ Family 1 (inverted H-bond, not simple helix).

stabilization contribution coming from the first mode at 300K is -0.06eV for g-1 and -0.07eV for α-1.

Clearly, although this mode is the one contributing most to the vibrational free energy, the free-energy differences between conformers depends on the whole distribution of modes, up to the mode that can be populated at that temperature. The population of each mode can be calculated as the internal thermal energy U_{therm} for each vibrational mode, given by:

$$U^i_{therm} = \frac{\hbar\omega_i}{e^{\frac{\hbar\omega_i}{k_B T}} - 1} \tag{8.2}$$

where ω_i is the frequency of vibration of the ith mode[5]. The normalization of this quantity by $k_B T$ also gives a measure of to which extent classical equipartition (each mode receives $k_B T/2$ of energy in the classical limit) is being fulfilled (or not). As can be seen in Figure 8.16, there is a considerable deviation from equipartition for all but the very first few vibrational modes of conformers g-1 and $\alpha - 1$ of n=5. Both molecules have vibrational modes up to ≈3600cm^{-1}, but modes higher than ≈1400cm^{-1} are not energetically accessible. This fact may have consequences for the *ab initio* molecular dynamics performed in Chapters 9.2 and 10, where the nuclei are assumed to be classical particles. The reader should be aware, though, that also anharmonicities can affect (and alleviate) this picture, since the "anharmonic" partition function would be different from the one used here (as well as the vibrations). In an anharmonic picture, modes of vibration can couple, also redistributing the available energy more

[5] This expression was obtained from Eq. 4.21, subtracting the ZPE contribution.

Table 8.4: Position of the lowest vibrational mode, in cm^{-1}, for the lowest energy conformers of each family discussed in the previous section (or previous chapter, for n=4), for n=4-8. . The lowest-energy α-helical conformers are in red, other α-helical conformers are in orange, the g-1-like conformers are marked with *, and the 3_{10}-helical conformer in blue.

	4	5	6	7	8
Family 1	23	22*	20*	12	11
Family 1a	27	23		10	11
Family 1b	25	26			
Family 2	17	13	20	28	22
Family 3	27*	20	8	15	16
Family 4	17		18	20*	15

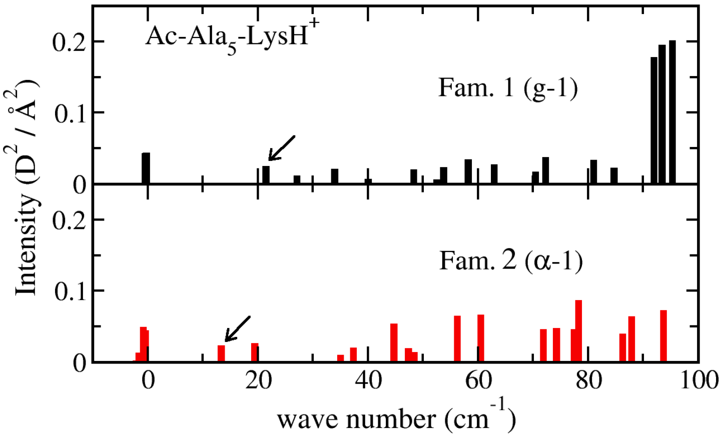


Figure 8.15: Normal modes of vibration of Family 1 and Family 2 (α-helical) for n=5; the arrow points to the first vibrational mode, and the 6 rotational-translational modes close to zero are shown for completeness only and can be taken as an indication of the (very low) level of numerical noise in the modes.

efficiently, and thus reaching equipartition sooner. Finally, α-1 has 7 modes below 50cm^{-1}, while g-1 has 5. Since these modes are the most populated, this suggests that α-1 has more vibrational states available than g-1.

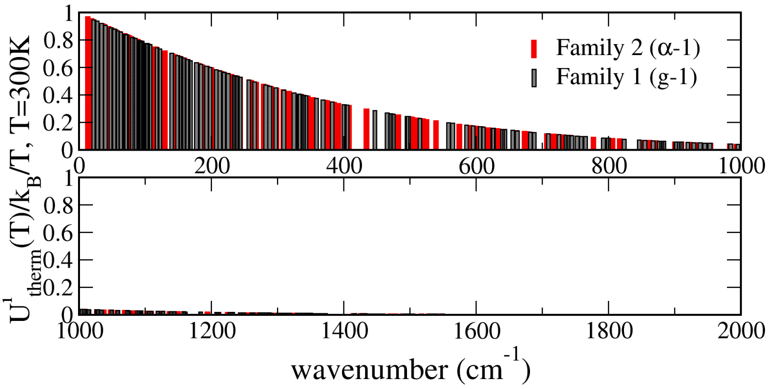


Figure 8.16: Average energy for each vibrational mode, normalized by k_BT, for T=300K, for the g-1 and $\alpha - 1$ (see text) conformers of n=5. Deviations from 1 denotes that equipartition is not fulfilled.

It is worth noting that all 3_{10} helices also have lower first vibrational modes than the compact conformers, being much less de-stabilized (or even not destabilized at all) in comparison to the α-helices, when including vibrational free energies. For example, for n=5, the lowest energy 3_{10}-helical conformer is 0.19eV higher in energy than g-1 at the PES, and gains a 0.02eV stabilization at 300K. The first vibrational mode of this particular conformer lies at 14cm^{-1} [6]. Since this mode is quite similar to the one

[6] Similar observations for the 3_{10}-helical conformers' vibrational modes have been made for n=10 and 15, that will be discussed in Chapter 10.

found in the α-helices (deviating only a couple of cm^{-1}), the slope of the free-energy curves is basically the same for α and 3_{10}, so that the energy difference at the PES for these two helical types is decisive for their relative ordering. This observation connects well with the observation in Ref. [325], where Ismer *et al.* report that in infinite, pure, polyalanine helices, the α and 3_{10} structures are similarly favored by vibrational entropy (while the more compact π-helices are the ones less entropically favored).

In order to prove that it is really the entropy term of the free energy that causes the stabilization of the helices, we take as an example the Ac-Ala$_6$-LysH$^+$ molecule. At the PES, a "g-1"-like conformer (Family 1) is the lowest energy structure, 60 meV lower than the α-helix, but at 300 K, the α-helical conformer (Family 3) is 50 meV more stable than Family 1, being the only molecule studied in the series where this change in stability from the PES to the FES at 300 K occurs. For these two conformers, the individual quantities composing the vibrational free energy in Eq. 8.1 (U from Eq. 4.21, TS, and E_{DFT}) are plotted in Figure 8.17 as energy differences, taking g-1 as the reference. The entropy term TS is the one with the largest contribution to the free energy difference between these two families at 300 K. However, also the internal energy U, more specifically the zero-point-energy term (see T=0 K), also favors helical structures over the compact ones, by a sizable amount. For n=6, for example, the internal energy contributes to 17 % of the α-helical stabilization at 300 K.

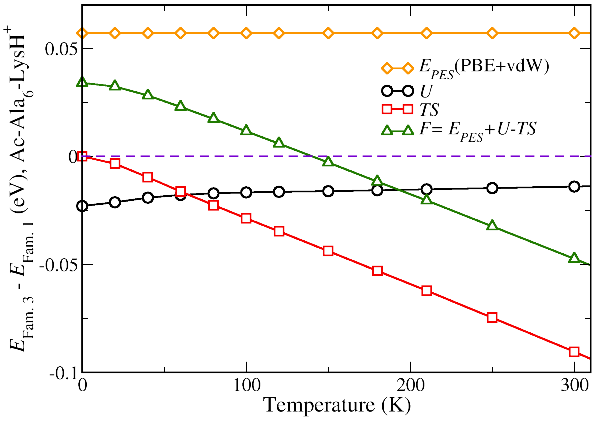


Figure 8.17: Energy differences for each term of the free energies for Family 3 (α) and Family 1 ("g-1" like) of Ac-Ala$_6$-LysH$^+$, taking Family 1 as the reference, as a function of temperature. The DFT-PBE+vdW energy differences are plotted as yellow diamonds, the internal energy U as black circles, the entropy term TS as red squares, and the full vibrational free energy F as green triangles.

Lastly, it is worth stressing that the stabilization of helices by vibrational entropic effects has been observed experimentally[352], even if indirectly, as mentioned in Chapter 6. Kinnear and coworkers observed, by performing ion-mobility experiments, that the Ac-Ala$_4$-Gly$_7$-Ala$_4$ polypeptide is globular at room temperature but becomes helical as the temperature is raised to 400K. The helix formation is expected to cause a loss of *conformational* entropy with respect to conformations found in the random-coil, so that the stabilization of helices with temperature increase could only be rationalized by a gain in *vibrational* entropy. This interpretation was also backed-up by force-field simulations of native and random-coil structures of more than 60 small peptides, including a 21-residues alanine-based peptide, by Ma *et al.* [353]. The free energies in the harmonic approximation were also calculated in that study and the authors conclude that both α-helices as well as β-hairpins are vibrationally more flexible than random-coil structures. The potential energy landscape should be, thus, one where the minima lying at higher values of energy should have narrow funnels while the native structure should have a lower energy and broader funnel. The study presented here provides the first first-principles evidence of this stabilization mechanism in the formation of helices.

8.4 Summary

In this chapter, a broad conformational search and characterization of low-energy DFT-PBE+vdW conformers was presented, for the series of alanine-based polypeptides Ac-Ala$_n$-LysH$^+$, n=4-8. The α-helix is seen to be the lowest energy structure for $n \approx$7-8, but with a very small energy gap on the PES to the next non-helical conformer. n=5 and 6 present a more compact structure, with an inverted H-bond, as the lowest energy PBE+vdW (PES) structure. For n=5 several DFT functionals were tested (GGAs, mGGAs and hybrids), with the energy hierarchy exhibiting a similar trend, as long as vdW interactions are added to all functionals. The situation for plain functionals (without vdW corrections) is inconclusive, with different functionals predicting different lowest energy structures. In fact, the preference for an α-helical structure at n=8 would not be predicted by plain PBE.

Computation of harmonic relative free energies makes the cross-over to α-helical preference safely predicted to happen at n=8, where a large free-energy gap at 300K (more than 0.1eV) is observed between the most stable α-helical conformer and the next non-helical one, in good agreement with experimental observations [3]. The stabilization of helices over the globular/compact conformers can be understood by the existence of a low frequency first vibrational mode involving the bending of the terminations, that makes conformations that are elongated and exhibit periodicity, (vibrationally) entropically favored.

Chapter 9

Connecting to experiment: ion mobility and IR spectroscopy

In the present chapter, a direct comparison of the first-principles predictions of the structure of the molecules composing the Ac-Ala$_n$-LysH$^+$ series with experimental data is presented. The results discussed in the previous Chapter for the α-helical preference onset seem to agree with observations made in water adsorption experiments [3, 348] (discussed in Section 6.3.2), but this agreement is indirect. Also, the search of the conformational space is quite a challenge for larger molecules, which renders a brute-force conformational search prohibitive as n grows. Having a theoretically calculated quantity that can be directly compared to structure-sensitive experimental data (like IR spectra, discussed in Chapter 4) is a great advantage, because it can, not only validate structure predictions, but also provide the means to determine the quality of the theory employed for this problem.

First, the ion mobility cross sections of the conformers discussed in the previous Chapter will be calculated, in order to connect to the experiment described in Section 6.3.1. Then, IR spectra of three members of the Ac-Ala$_n$-LysH$^+$ series, with n=5, 10, and 15 will be shown in the harmonic approximation (Section 4.1) and including anharmonic effects through the dipole auto-correlation function, as was explained in Section 4.3. The longer molecules are expected to be helical, but the situation for n=5 is unclear already from our conformational search and inclusion of harmonic free energies. IRMPD (infrared multiple photon dissociation) spectra of these molecules were measured in the FELIX free electron laser (Netherlands) by P. Kupser, working in the group of Gert von Helden, in the Molecular Physics department of the Fritz Haber Institute. Details of this experiment were given in Section 6.3.4. The key goal here is, thus, to verify the structure of these molecules, by comparing to the available experimental data.

The work presented in this chapter has been partly published in Ref. [4] [1].

9.1 Ion-mobility cross sections

The ion-mobility cross sections are calculated with the algorithm proposed in Ref. [340], using the Fortran program provided by the authors of Ref. [340]. The theoretical calculation involves computing the collision cross section Ω of equation 6.3. The average collision cross section is similar to the average

[1] As stated in Ref. [4], some values presented there for Ac-Ala$_5$-LysH$^+$ were calculated with an old (internal only) version of the TS-vdW scheme. The differences are not large between the two versions (less than 10meV in energy differences for all cases tested). Here all values presented are calculated with the final TS-vdW scheme.

area (2D) that the peptide spans when moving through the buffer gas. It is important, though, to consider the "long-range" interaction between the peptide and the buffer (helium) gas, since the hard-sphere approximation would give poor results [2].

In the scheme used here, the interaction between the peptide and the helium gas was approximated by a potential of a Lennard-Jones (LJ) type (with empirical parameters) for each pair of atoms (one in the molecule and one in the buffer gas), as described in Ref. [340]. Parameters for this potential, fitted to experimental data, are provided with the code. Knowing the form of the potential (with the respective parameters) and the temperature, an atom-atom collision integral $\Omega_{a,a}$ can be evaluated from tabulated data , as, for example, in Ref. [381]. The radius R^a of each atom in the peptide is then calculated as $R^a = \sqrt{\Omega_{a,a}/\pi}$. [3] The full cross section of the molecule is calculated by considering its geometry (as calculated here from the DFT-PBE+vdW relaxations), and projecting the shadow of each atom with radius given by R^a onto randomly chosen planes in space. A Monte Carlo integration is performed to determine the area of the projection. Many different randomly selected planes are considered and the area of the projections are calculated until the average converges to a value within specified error limits. More details and example of applications can be found in ref. [340].

For each n, this cross section was computed for all conformers shown in Table 8.1, using the PBE+vdW relaxed geometry for each of them (*tight* settings). Additionally, for n=10 and n=15, pure α-helical conformers were computed (also relaxed with PBE+vdW), since we expect the helix to remain α-helical for $n > 8$ [4]. The α-helical geometries used for n=10 and 15 are shown in Figure 9.1.

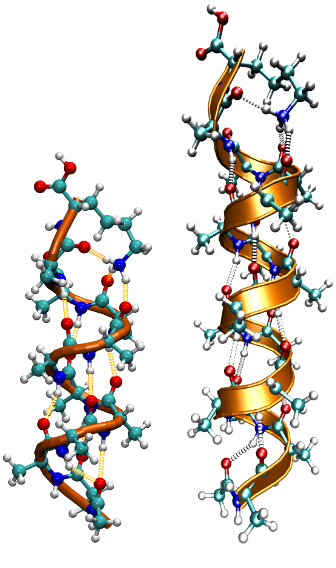

(a) n=10 (b) n=15

Figure 9.1: Representation of: (a) the Ac-Ala$_{10}$-LysH$^+$ molecule (130 atoms) in the α-helix (PBE+vdW ground state) geometry; (b) the Ac-Ala$_{15}$-LysH$^+$ molecule (180 atoms) in the α-helix (PBE+vdW ground state) geometry.

In order to compute the relative cross sections to a perfect α-helix for each n, the same formula used by Jarrold and coworkers (Equation 6.4) was applied to the calculated values. All relative cross sections are plotted in Figure 9.2(a). The scatter of the values is substantial for the different conformers of n=4-8. This is not surprising, because the conformers have very different geometries, some being much more

[2] The vdW interaction between the helium atoms and the peptide is non-negligible, and the scattering actually happens due to the vdW tails.

[3] The radius of each atom is calculated as if they were hard spheres, but the long-range interaction is indirectly accounted for in $\Omega_{a,a}$.

[4] The 3_{10}-helical motif was also tested for n=10, 15. Not only are their relative energies too high with respect to the α-helical conformer, as will be seen in the next chapter, but due to their more elongated geometry, the relative cross sections are way too high.

compact than others (e.g. Family 2 of n=7 vs. its α-helical Family 1). However, if one considers only the lowest free energy conformer (at 300 K) for each n, a similar trend for the cross section as the one observed in experiment can be seen, as plotted in Figure 9.2(b). In order to account for the relative contributions of all conformers of each n to the measured cross section, a possible approximation is to compute weighted averages, with the weights corresponding to the Boltzmann weights (populations) of each conformation at 300 K. This average cross section can be written as

$$\overline{\Omega}(T) = \frac{\sum_\lambda e^{\frac{-\Delta F_{vib,\lambda}}{k_B T}} \Omega_\lambda(T)}{\sum_\lambda e^{\frac{-\Delta F_{vib,\lambda}}{k_B T}}}, \tag{9.1}$$

where λ runs over the conformations lying in the lower 0.13 eV for each n (exactly the ones that were discussed in the previous chapter), $\Delta F_{vib,\lambda}$ is the difference of vibrational harmonic free energies between conformer λ, taking as the reference the lowest energy PBE+vdW PES conformer, and T is the temperature which is considered to be 300 K for comparison with experiment. Conformations with $|\Delta F_{vib,\lambda}| \geq 0.08$ eV have negligible weights. These average cross sections are plotted in Figure 9.2(c), where a very good agreement with the experimental data can be seen. The more compact conformers ("g-1"-like) appearing at low free-energies for n=5 and 6 pull down the relative cross sections of these n, with respect to larger n where the α-helical conformer really starts to dominate. This yields precisely the drop for short n observed in the experimental data. The discrepancy of around 6 Å^2 between the calculated and the experimental data is possibly due to inaccuracies in the parametrization of the LJ potential used, or due to a systematic effect related to the terminations. Nevertheless, it is only 6-7% of the calculated value, which renders these results quite reliable.

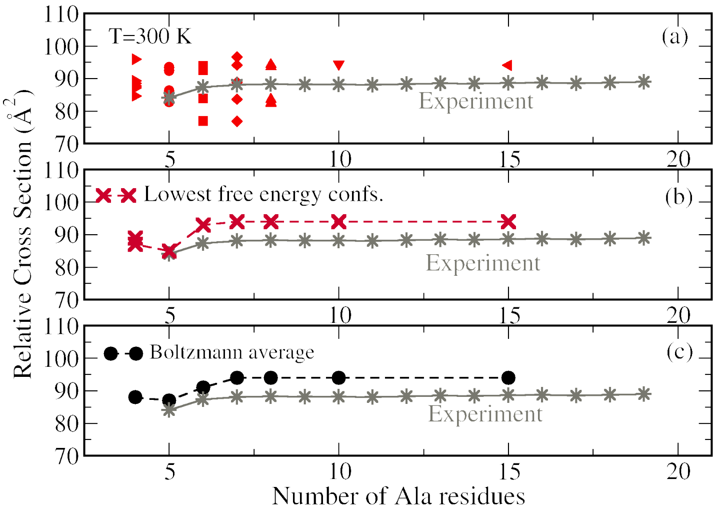


Figure 9.2: Estimated relative ion mobility cross sections for Ac-Ala$_n$-LysH$^+$, n=4-8, 10, and 15. (a) Calculated cross sections for all low-energy conformers discussed in Chapter 8 for n=4-8, plus α-helical conformations of n=10 and 15, at T=300 K. (b) Ion mobility cross sections only of the lowest free energy conformers for each n at T=300 K. (c) Boltzmann average of the cross sections presented in (a), calculated from Eq. 9.1 at T=300 K. The experimental data taken from Ref. [37] is shown in grey (stars), in all panels.

9.2 IR Spectroscopy of alanine-based polypeptides in the gas-phase

Calculations of both harmonic and anharmonic IR spectra for Ac-Ala$_n$-LysH$^+$, n=5, 10, and 15 (as well as n=8, as a test-case), were performed. The harmonic spectra were calculated with the double-harmonic approximation, as explained in Section 4.1. The frequencies of vibration were obtained through Eq. 4.9 and the intensities through Eq. 4.17. The anharmonic spectra were calculated through the Fourier transform of the dipole autocorrelation function, obtained from an *ab initio* molecular dynamics (AIMD) trajectory, as explained in Section 4.3 (Eq. 4.47).

9.2.1 Obtaining clean and converged spectra from AIMD dipole autocorrelation functions

For the calculation of the dipole autocorrelation function, AIMD runs of more than 20 ps in the microcanonical ensemble were performed for all molecules, using the PBE+vdW functional, and a time step (Δt) of 1 fs. The molecules were always initially equilibrated, by performing 4 ps of thermostated runs at 300 K. For the molecules studied here, $\sim$20 ps of simulation was enough to obtain converged spectra, both for the peak positions and the peak intensities. The convergence of the spectrum of Ac-Ala$_{15}$-LysH$^+$ (the largest molecule studied here) with time is shown in Figure 9.3. The peak positions and peak intensities do not change (notice particularly the peak just below 1400 cm^{-1}) when comparing the spectra obtained from a 20 ps trajectory and a 24 ps one.

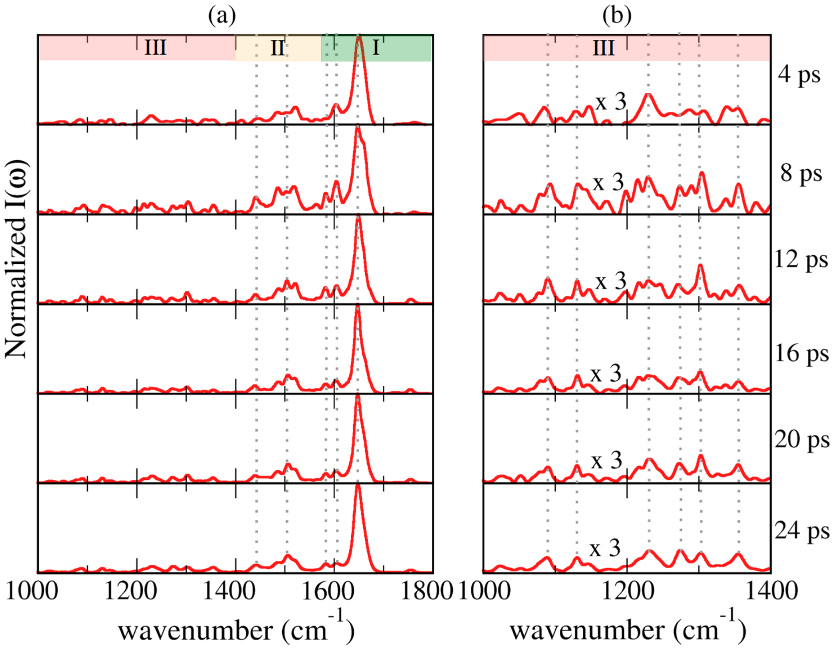


Figure 9.3: Convergence of the vibrational spectrum of Ac-Ala$_{15}$-LysH$^+$, calculated from AIMD, with the time of the trajectory. Grey lines serve as guides to the eye and are in the position of the converged peaks. All spectra are normalized to 1 for the highest peak. (a) Amide-I, amide-II, and amide-III regions, marked marked as I (red background), II (yellow background), and III (green background) in the figure. (b) Zoom into the amide-III region, with the intensities multiplied by 3.

Care had to be taken with the autocorrelation function in order to produce spectra with minimal noise, like the one shown in Figure 9.3. First, there is a problem with the number of points that is averaged for each term of the autocorrelation function. For example, if there are t_{max} steps in the AIMD simulation, and one considers a new $t = 0$ at each time step, the $\langle\mu(0)\mu(0)\rangle$ term will be the average of t_{max} times $\mu(0)\mu(0)$; $\langle\mu(1)\mu(0)\rangle$ will only have $t_{max} - 1$ possible combinations; $\langle\mu(2)\mu(0)\rangle$ will have only $t_{max} - 2$ possible combinations, and so on. The last term ($\langle\mu(t_{max})\mu(0)\rangle$) will not be an average at all, but will only be a single value. There is not enough statistics to compute the ensemble averages of the points of the autocorrelation function close to $t = t_{max}$ (see Figure 9.4(a)). Therefore, it is necessary to disregard part of this function in order to minimize noise. To maximize statistics, it was found that cutting $\sim 30\%$ of the tail is a good accuracy/price compromise. [5] A typical full dipole autocorrelation function and the one

[5] An alternative would be to compute the average of all terms considering only half of the trajectory. All terms $\mu(t)\mu(0)$ would have $t_{max}/2$ points available. The drawback is that it requires longer trajectories to produce spectra with good resolution. This procedure was tried here, but with 20 ps of trajectory produced spectra with a very low resolution, denoting that indeed, an even longer trajectory would be needed in this case.

where the tail was cut can be seen in Figure 9.4 (a) and (b) respectively.

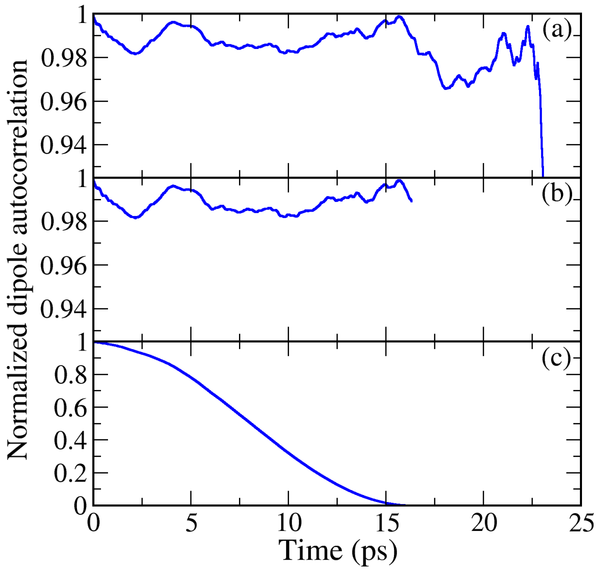


Figure 9.4: (a) Dipole autocorrelation function taken from a 22ps AIMD simulation of Ac-Ala$_{15}$-LysH$^+$. Notice the loss of statistics towards the end of the curve. (b) Dipole autocorrelation function with the tail ($\sim$ 30%) cuted out. (c) Dipole autocorrelation function after being multiplied by the Hann window ($w_t \langle \hat{\mu}(t)\hat{\mu}(0)\rangle_{t_0}$). The oscillations of the function cannot be seen anymore due to the change in scale in panel (c).

The autocorrelation functions here do not decay to zero because the molecule has a permanent dipole moment and it is not allowed to rotate in the MD trajectories. When taking the Fourier transform of this signal, the fact that it does not decay to zero produces noise, as will be explained below. This is a common problem in signal processing and a solution is well known, which is to multiply the signal by a window function. The purpose of the windowing functions is to improve the quality of the Fourier transform of signals (See Ref. [382], Chapter 13) by reducing leakage of one peak in its neighboring bins. In order to take the Fourier transform of a signal of finite time, and not necessarily decaying to zero, it is necessary to multiply it by a function that sets the signal to zero for all times greater than t_{max}. This is what would be called a "rectangular" (or square) window function, and which induces leakage. The decay of the rectangular function to zero is too harsh, inducing noise in the Fourier transform, which is reflected in the appearance of side-peaks (leakage) in neighboring bins. A trivial example of the Fourier transform of a (shifted) cosine function with a square function is shown in Figure 9.5, where small side-peaks can be seen (Figure 9.5(c)).

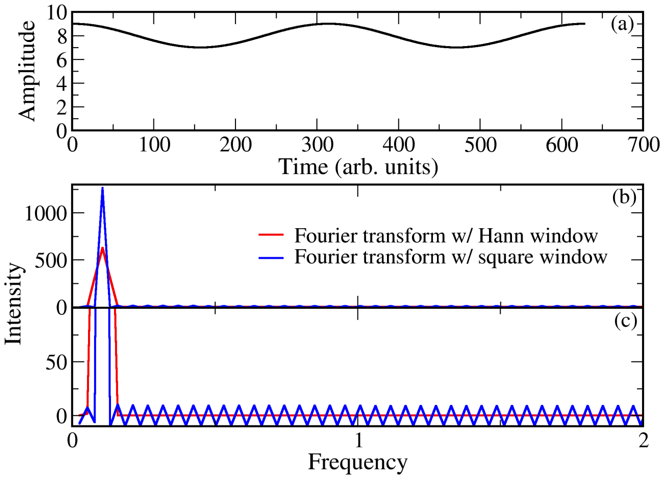


Figure 9.5: (a) A shifted cosine function. (b) Fourier transform of the function plotted in (a), using the rectangular window (blue) and the Hann window (red). (c) Zoom into the low intensity part of (b). Leakage is seen to be substantially reduced for the transformation using the Hann window.

In practice, any other type of function that decays smoothly to zero improves the quality of Fourier-transformed signal [382]. The function chosen to be used here is the Hann-type window, that has the following analytic form:

$$w_t(t) = \frac{1}{2}\left[1 + \cos\left(\frac{\pi t}{N}\right)\right] \tag{9.2}$$

where N is the total number of data points and t ranges from 0 to N. The Fourier-transform of the cosine function multiplied by this window is shown in Figure 9.5(b) and 9.5(c), where the leakage is seen to be much reduced with respect to the transformed signal multiplied by the rectangular function. In Figure 9.4(c) the dipole autocorrelation multiplied by the Hann window ($w_t\langle\hat{\mu}(t)\hat{\mu}(0)\rangle_{t_0}$) for an AIMD run of Ac-Ala$_{15}$-LysH$^+$ is shown. [6]

Finally, as discussed in Section 6.3.4, the excitation laser used to measure the IRMPD spectra has a width around 1% of the corresponding wavenumber. Therefore, all calculated spectra shown here (harmonic and anharmonic) were convoluted with a Gaussian broadening function with a variable width of 1% of the corresponding wave number, in order to account for this effect.

9.2.2 Pendry reliability factor

A central piece for the structure characterization of molecules based on IR spectra is the comparison between the theoretical spectra and the experimental ones. For large and complex molecules like the ones studied here, this comparison may be quite a challenge. As an example, a spectrum obtained for Ac-Ala$_{10}$-LysH$^+$ (AIMD, DFT-PBE+vdW, α-helix), that will be detailed further on, is shown in Figure 9.6, in comparison to the IRMPD experimental data from P. Kupser *et al.*, for the same molecule (Section 6.3.4). The question is: How can one quantify the degree of similarity (or deviation) of the spectra from one another?

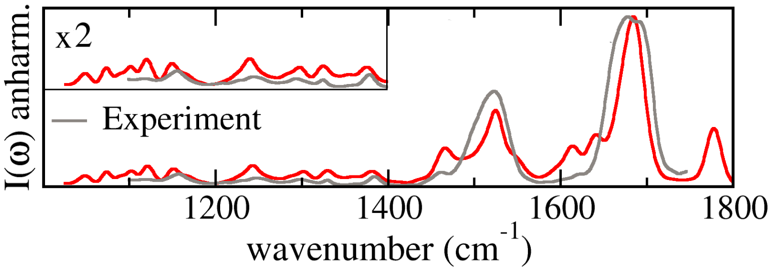


Figure 9.6: Comparison between experimental [gray] and theoretical [red] vibrational spectra for Ac-Ala$_{10}$-LysH$^+$, all normalized to 1 for the highest peak. Calculated spectrum from AIMD (including anharmonic effects) with the PBE+vdW functional, starting from an α-helix.

A visual comparison can always be made, but it is often found that the eyes of different observers do not always agree. A quantitative comparison, for example through a number that can define how correlated the curves are, is desirable, since it gives an unambiguous measure of the agreement between the spectra. A simple overall least squares fit for the intensities is not enough for these curves, since intensities may disagree to a certain degree, but the real important information is the position of the peaks. As was already discussed, the IRMPD spectra could have peak intensities that are distorted due to the absorption of many photons, however the peak positions should match those of the fundamental

[6] The oscillations of the autocorrelation function are not evident in Fig. 9.4(c) because the y-axis has been scaled so that the decay of the window function is visible.

modes of the molecule. Furthermore, especially for the case of the molecules studied here, there is important information about the structure of the molecules in the “not-so-intense” amide-III region of the spectrum [249, 250], discussed in Section 4.4 and zoomed in Figure 9.6 (1000-1400cm^{-1}). If only the peak intensities are taken into account, these small wiggles would be discarded, and only the very intense amide-I and amide-II peaks would matter.

This is a common problem in other areas of physics, for example X-ray diffraction and low energy electron diffraction (LEED), where sophisticated theories are employed to obtain information from measured signals by fitting parameters. In these areas, this problem has been solved by the use of so-called reliability factors (R-factors) [383–385]. In this work, we choose a particularly successful R-factor (in the above mentioned areas), proposed by Pendry [385]. The Pendry R-factor addresses the need to match mainly peak positions, rather than the intensities. Given two continuous curves with intensities $I_{exp}(\omega)$ and $I_{th}(\omega)$, this R-factor compares the renormalized logarithmic derivatives of the intensities, given by:

$$Y(\omega) = L^{-1}(\omega)/[L^{-2}(\omega) + W^2] \tag{9.3}$$

with $L(\omega) = I'(\omega)/I(\omega)$, and W approximately the half width of peaks in the spectra. The advantage is that the L functions have a sign inversion exactly where the maximum of the peak is, and if peaks are far enough apart, relative intensities are completely ignored, while if they are close together, $L(\omega)$ is moderately sensitive. However, the L functions would be too sensitive to zeroes in the intensity, since the logarithmic derivatives would have singularities in this case. The Y function is a simple transformation of L, which avoids such singularities, by giving similar weights to maxima and zeroes in the intensities. These functions are plotted as an illustration for a model peak in Figure 9.7. Several values of W are tested for the Y-function. The average half-width of the peaks would correspond approximately to the value of $W = 35$. In this work, when comparing experimental and theoretical IR spectra, the value used was of 10cm^{-1}. Changes of $\pm 50\%$ in this value induced a change of no more than 0.03 in the value of the calculated reliability factors, as shown in Appendix G.

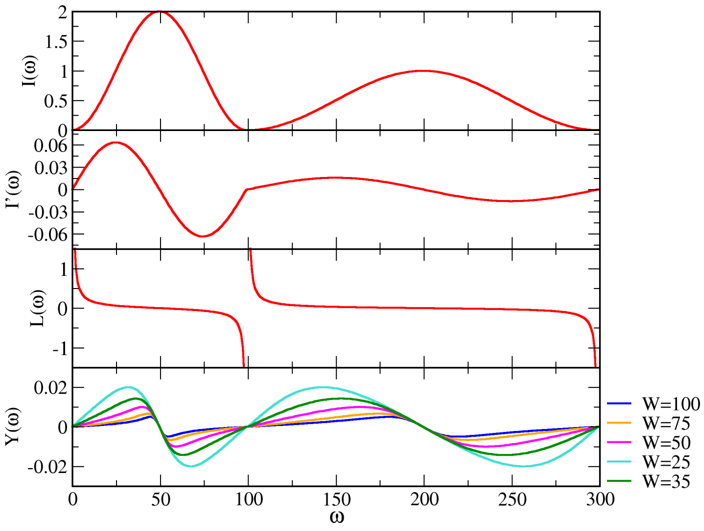


Figure 9.7: Pendry R-factor quantities. (a) Two model peaks; (b) the derivative of the intensity of the peaks; (c) the logarithmic derivative $L(\omega)$; (d) the corresponding Pendry $Y(\omega)$ functions (Eq. 9.3)

The Pendry R-factor (R_P) is then defined as:

$$R_P = \int d\omega (Y_{th} - Y_{exp})^2/(Y_{th}^2 + Y_{exp}^2) \tag{9.4}$$

In practice, this convention leads to R_P=0 for perfect agreement, 1 for uncorrelated spectra, and 2 for complete anticorrelation.

Since R_P is sensitive to small wiggles, experimental noise in low-intensity regions must be removed, which is achieved by splining and smoothing the raw data twice, using a 3-point formula, with the data interpolated through splines in a fine 0.5 cm^{-1} grid. The smoothing formula is given by:

$$s_j = \frac{y_{j-1} + 2y_j + y_{j+1}}{4} \tag{9.5}$$

An example of the raw experimental data compared to the smoothed experimental data can be seen in Figure 9.8 for Ac-Ala$_{15}$-LysH$^+$. Care was taken not to smooth the curves too much, so that even small peaks remain in the smoothed spectra, like the shoulder right above 1600cm^{-1}, or the peaks in the amide-III region (between 1000 and 1400 cm^{-1}). The comparison between raw and smoothed experimental spectra for n=5, 10, and 15 is shown in Appendix G. From now on, all the experimental data shown in the figures of this chapter will correspond to the smoothed data.

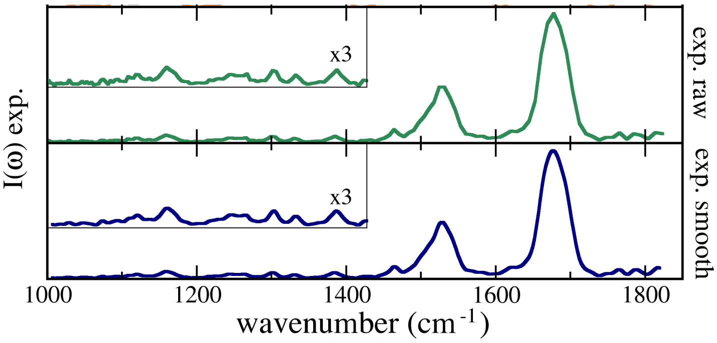


Figure 9.8: Raw experimental data (top) compared to the smoothed (Eq. 9.5)experimental data (bottom) for Ac-Ala$_{15}$-LysH$^+$.

The R_P factor between theoretical and experimental data is always calculated only for the range where the experimental data was measured and excluding the first and last 30 wave numbers, due to uncertainties in the experiment about how much beam power was available in the tails.

Rigid shifts between experimental and theoretical data.

If comparing calculated harmonic spectra to experimental data, variable shifts between theory and experiments are expected, due to the negligence of anharmonicities [7]. However, in this work, also when including the anharmonic effects, via the dipole autocorrelation function, *rigid* (not variable) shifts are observed between the theoretical data and the experimental one. Rigid shifts of calculated IR spectra including anharmonicity, compared to experiment, have been observed before in the literature [240]. This is most likely caused by a softening of the modes when using GGA functionals. This fact is exemplified by comparing, in Table 9.1 the force constants (k) for the frequencies of vibration of formamide, already presented in Section 4.4, calculated with PBE and B3LYP in the harmonic approximation. B3LYP is a hybrid GGA functional that usually gives frequencies of vibrations for molecules closer to the experimental value than PBE. The PBE force-constants are always smaller than the B3LYP ones, which makes the modes softer (i.e. the parabola is wider), and the frequencies underestimated. When including anharmonic effects to the PBE+vdW spectra and comparing to experiment, in the range from 1000 to 1800 cm^{-1} these underestimations seem to be systematic, as will be seen in the next section.

[7] In these cases, scaling factors, that are different for different functionals and different regions of the spectra, are commonly applied to the calculated spectra.

frequency nr.	ω_{PBE} (cm^{-1})	ω_{B3LYP} (cm^{-1})	k_{PBE}/k_{B3LYP}
1	239	263	0.83
2	545	570	0.90
3	633	640	0.98
4	991	1043	0.89
6	1233	1266	0.96
7	1363	1421	0.92
8	1559	1619	0.94
9	1737	1787	0.97
10	2850	2940	0.94
11	3487	3576	0.95
12	3626	3712	0.95

Table 9.1: Ratio PBE/B3LYP of the harmonic mass-weighted force constants (k) of the formamide molecule.

The R_P factors in this work are, thus, always determined including a rigid shift Δ of all calculated frequencies, but no scaling factors. The variation of R_P with respect to Δ for a few examples calculated in this work are shown in Appendix G.

Sensitivity of the R_P factor when comparing IR spectra

To illustrate an application and test the sensitivity of this factor to differences in the spectra, two different conformations of Ac-Ala$_8$-LysH$^+$ are taken as an example. These conformers are representative of Family 1 (α-helix) and Family 2 (compact, non-helical). For more details on these structures, see Chapter 8. Approximately 20ps of AIMD for each conformer is simulated and the spectra are calculated according to Equation 4.47. The idea is then simply to compare them both in the region between 1000 and 1800 cm^{-1}, which is where experimental data is available for n=5, 10, 15, and check two things: (i) Is this spectral region sensitive for structural changes?; (ii) Can Equation 9.4 (Pendry R-factor) give a good measure of how much these spectra (do not) agree?

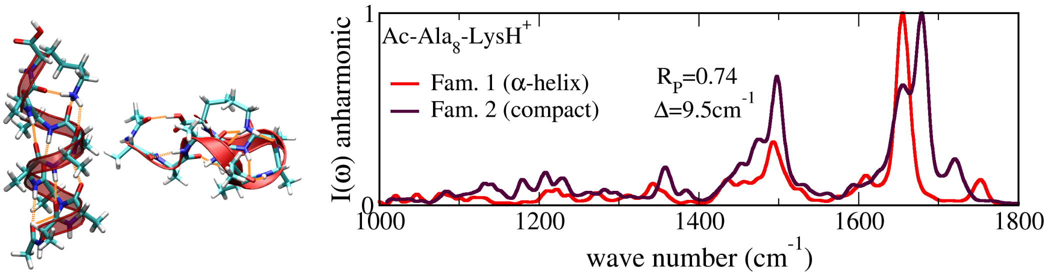


Figure 9.9: Comparison between Ac-Ala$_8$-LysH$^+$ vibrational spectra calculated from AIMD starting from the α-helical conformation [red line] and compact conformation [green line]. All spectra are normalized to 1 for the highest peak. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in the plot.

The two spectra for n=8 calculated from AIMD with the respective R_P-factor value (Eq. 9.4) and shift (Δ) are reported in Figure 9.9. They do look substantially different in this wavenumber region, meaning that if there were experimental data to compare, it should be possible to differentiate between these structures. The R_P factor associated with these two spectra is of 0.74, which is quite high, denoting very poor agreement, as expected. This factor is obtained upon a shift of 9.5cm^{-1} of the spectrum of the compact conformer (green line in Fig. 9.9) with respect to the one associated with the α-helical conformer (red line in Fig. 9.9). The small wiggles in the region of 1000 to 1400 cm^{-1} are not just random peaks. They belong to the amide-III region described in Section 4.4, which has been reported in the literature to give very valuable information even upon small structural changes in the molecules[34, 249, 250].

9.2.3 Effects of different functionals in the computation of harmonic and anharmonic spectra.

The Ac-Ala$_{10}$LysH$^+$ molecule is taken as a test case to investigate effects of using different functionals in the calculation of the harmonic and anharmonic spectra. The model is taken to be an α-helix, as pictured in Figure 9.1(a). The point here is not yet to assign a specific conformation for this molecule but just to access the quality of the functionals.

First, the quality of the harmonic approximation is addressed. The *harmonic* spectra for the α-helical model calculated with PBE+vdW and BLYP+vdW are shown in Figure 9.10, compared to the experimental spectrum described in Section 6.3.4 and Ref. [38] for this molecule. R_P factors comparing theory and experiment are reported, including the rigid shifts (Δ), explained in Section 9.2.2. Although there are many similarities between the PBE+vdW and the BLYP+vdW spectra, they differ substantially in the spacings between the amide-I and the amide-II peaks. The spectrum obtained with BLYP+vdW presents a much too small separation between these peaks. The very high R$_P$ factor of 0.79 reflects this fact. The shift Δ in this case is even found to be negative, because the best fit is actually found by aligning the peaks around 1400cm^{-1} and not the two other high-intensity ones.

The peak appearing in both PBE+vdW and BLYP+vdW spectra, between the amide-I and amide-II bands is the "scissor" motion (asymmetric bend of the three hydrogens with respect to the nitrogen)[8] of the NH$_3^+$ group in the Lys residue. The peak just below 1600cm^{-1}, which appears as a shoulder of the amide II band for PBE+vdW spectrum, but is quite pronounced in the BLYP+vdW spectrum is the umbrella vibration (symmetric bend of the three hydrogens with respect to the nitrogen) of the NH$_3^+$ group. The bad description of this peak, by the BLYP functional, has been reported in the literature, by Cimas and Gaigeot [244]. For the PBE+vdW functional the situation is bad as well, but looks slightly better in comparison to experiment, which is reflected by the lower (but still bad) R$_P$ factor of 0.54 upon a shift of +15cm^{-1}.

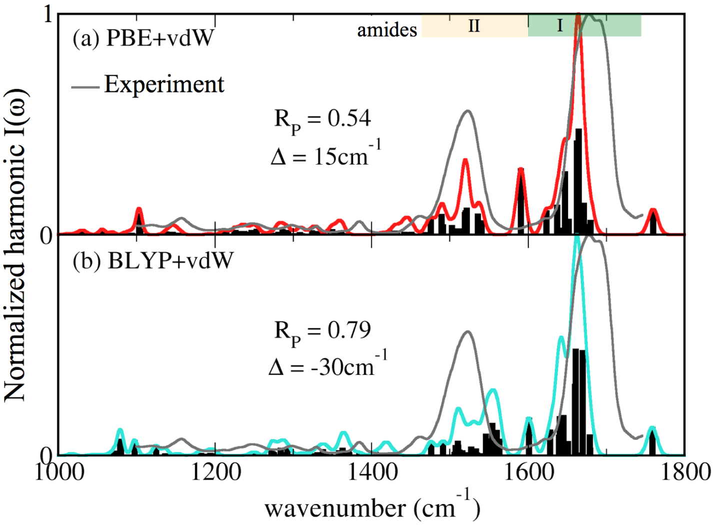
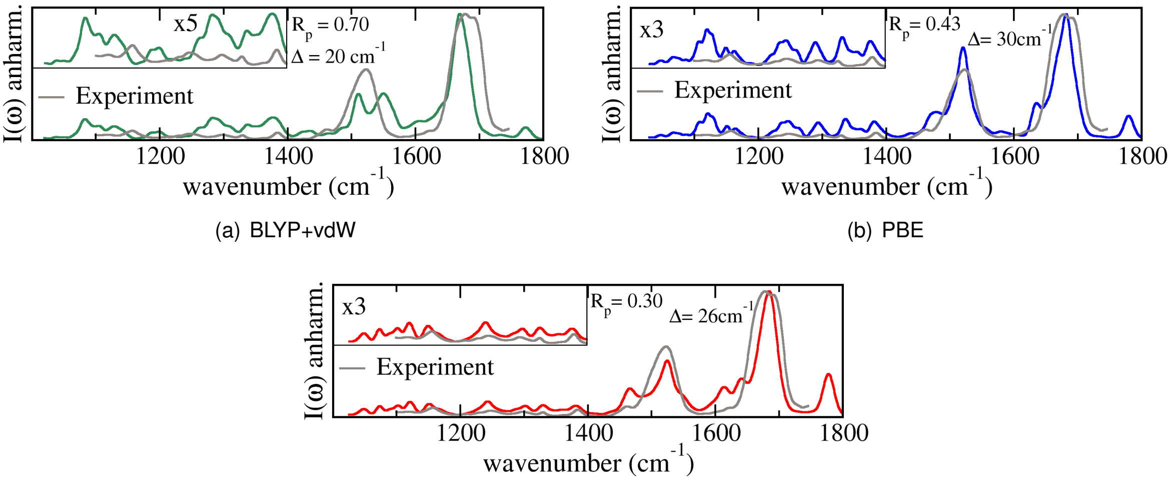


Figure 9.10: Comparison between harmonic spectra of α-helical motifs for n=10 for: (a) the PBE+vdW functional and (b) the BLYP+vdW functional. In grey, the experimental IRMPD (room-temperature) spectrum. The spectra have been artificially broadened by convolution with a gaussian function (the original "sticks" spectra are plotted in black in the Figure). Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph, but the plotted spectra are **not** shifted.

[8] A detailed account of the vibration assignment of the harmonic normal modes is given in Appendix F.

Next is the question of anharmonicities: By how much do they change this harmonic picture and what is the effect of different functionals in this case? Here it is important to stress that by “anharmonicities” here we mean not only the intrinsic anharmonic nature of the vibrational modes, but also the fact that the molecule dynamically explores the local conformational space. In order to study the effect of different functionals on the calculation of the AIMD derived spectra, the n=10 conformer was simulated with three different GGA functionals: PBE+vdW, BLYP+vdW and plain PBE. For the AIMD-derived IR spectra, the starting geometry and velocities for all three simulations were the same, and all of them started from a perfect α-helix. These spectra are shown in Figure 9.11 and are also compared to the experimental data.



(a) BLYP+vdW

(b) PBE

(c) PBE+vdW

Figure 9.11: Comparison between experimental [gray lines] and theoretical [colored lines] vibrational spectra for Ac-Ala$_{10}$-LysH^{+}, all normalized to 1 for the highest peak. (a)Calculated spectrum from AIMD (including anharmonic effects) with the BLYP+vdW functional, starting from an α-helix. (b) Same as (a), but with the PBE functional. (c) Same as (a), but with the PBE+vdW functional. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison).

First, we focus on the comparison of the AIMD-derived spectra obtained with PBE+vdW (Figure 9.11(c)) and BLYP+vdW (Figure 9.11(a)), as in the harmonic case. The BLYP+vdW spectrum gives also a very unsatisfactory R_P factor of (0.70), but now with a positive shift. This factor is quite high, denoting poor agreement (also clear by visual comparison).

The anharmonic PBE+vdW spectrum, on the other hand, now gives an excellent agreement, both visually and by a low R_P factor of 0.30. Based on this observation, a value of 0.30 for R_P is considered very good (or even the best) agreement for this problem. The reasons for not expecting a much lower R_P factor here are the following: DFT, with its exchange-correlation functional approximations is not an exact theory; the peak positions are not exact in an approximate functional; AIMD treats nuclei in terms of classical mechanics, but the “real” nuclei (and their corresponding wave functions) are probing the PES with a quantum distribution (which is sizable, as was seen in Section 8.3.1); and experimental data is obtained by the absorption of many photons while the theory is based in a linear, single-photon absorption picture.

Turning now to the PBE-derived spectrum, visually it looks very similar to the PBE+vdW one. However, the R_P factor is higher (0.43), denoting poorer agreement, and this must have a reason. In Figure 9.12 the PBE+vdW and PBE AIMD-derived spectra are superimposed. The relative distances between the peaks in the PBE functional do not match experiment as well as the PBE+vdW distances. The discrepancies are more pronounced in the shoulder of the amide-II peak and in the range of the amide-III vibrations.

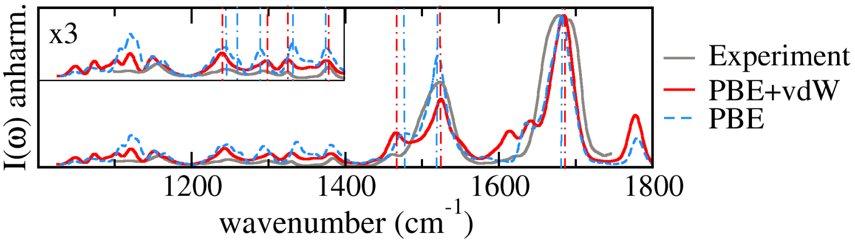
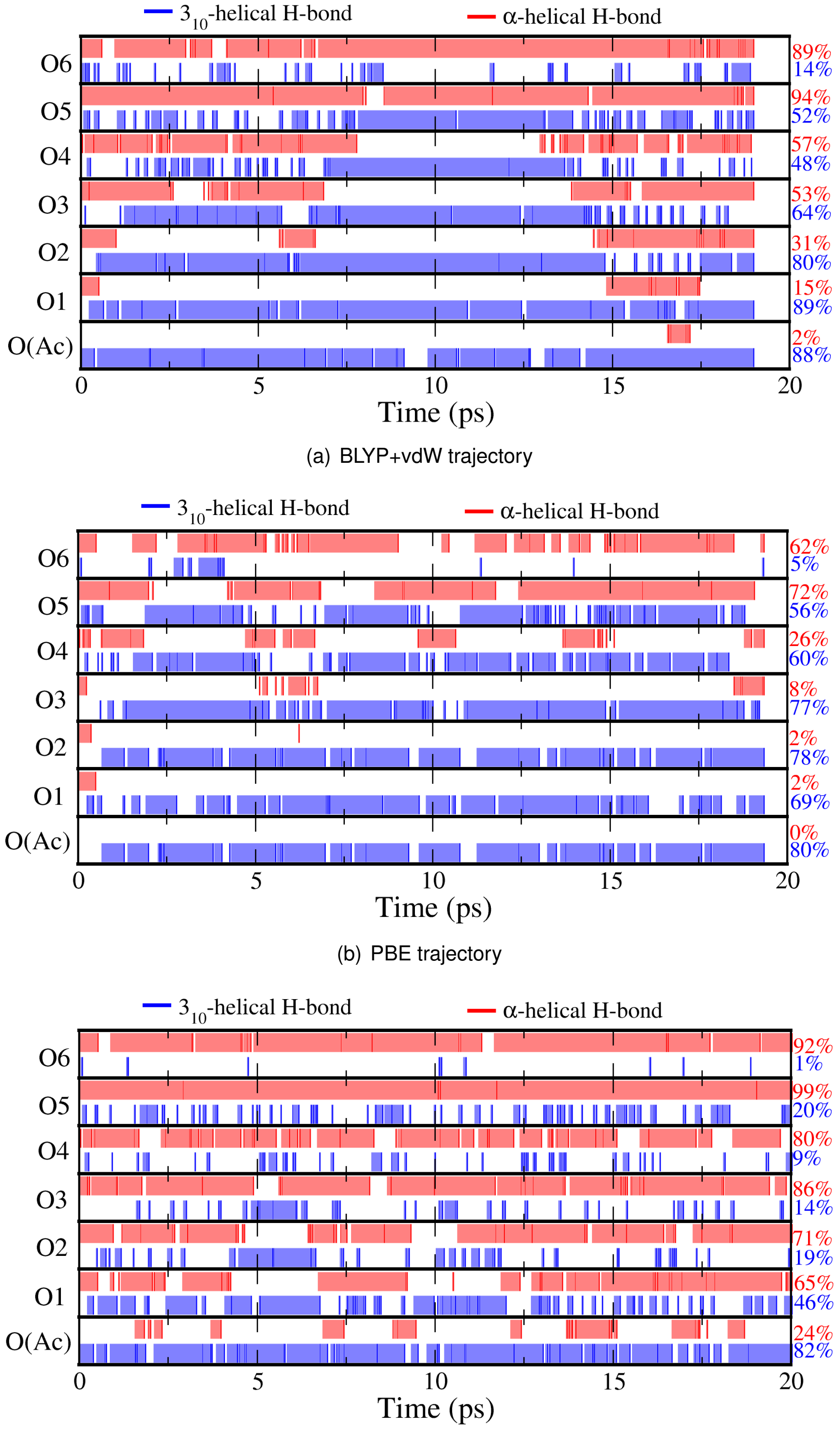


Figure 9.12: Comparison between experimental [gray lines] and theoretical AIMD-derived PBE [blue] and PBE+vdW [red] vibrational spectra for Ac-Ala$_{10}$-LysH$^+$, all normalized to 1 for the highest peak. These spectra are the same as shown in Figure 9.11(b) and (c), and the shifts Δ are the same as reported in that figure (30 cm^{-1} for PBE and 26 cm^{-1} for PBE+vdW) . The dot-dashed lines serve as a guide to the eye, in order to see the differences between the spectra.

In order to understand where these discrepancies come from, it is useful to look at the detailed conformations the molecule assumes in each of the trajectories leading to the spectra in Figure 9.11. The H-bond patterns for them are plotted in Figure 9.13. What is shown in that picture is the H-bond connection for each oxygen of the molecule starting from the Ac termination (O(Ac)), except for the CO groups that interact with the lysine termination, against the time of simulation. A H-bond is counted for every (C-)O - NH pair that is closer than 2.5Å. This is a conservative definition, in the sense that e.g. a 3_{10} bond might be counted even though not really there. Here it is more important that no possible bond is missed. Each color represents a different kind of H-bond, labeled in the figure, and the respective ratios of each type computed for the whole trajectory are shown on the right side of the plot. These ratios can exceed 100% because when a bond is bifurcated it is counted twice, once for each kind of H-bond.

The reason for the apparent disagreement with experiment for the spectra of n=10 calculated with PBE becomes apparent from Figure 9.13. While the trajectory for PBE+vdW is such that the molecule maintains mainly its α-helical character throughout the simulation, the PBE and BLYP trajectories favor a mostly 3_{10} helix. It is known that the absence of vdW favors 3_{10} helices, so the PBE data comes as no surprise, and the corresponding higher R_P factor is due precisely to this “conformational” change. For the BLYP+vdW trajectory, though, even though vdW interactions are included, the molecule also assumes a 3_{10} helical character. Indeed, the BLYP+vdW energy hierarchies for n=5, mentioned in Chapter 8, render 3_{10}-1 more stable than α-1. Here only one trajectory is presented and therefore there are not enough statistics to make a strong statement. The observed behavior might be solely due to the parametrization of the BLYP functional, but it is also known that the vdW correction as proposed in Ref. [165] does not couple as well to BLYP as it does to PBE. The value for the empirical parameter s_R of the scheme for BLYP is 0.62, which is rather small, meaning that vdW interactions affect rather unusually short bond distances by default. PBE has s_R=0.94, which is very close to ideal (if ideal can be considered s_R=1.0 in this scheme).

It is here assumed that the experimental spectrum corresponds to that of an α-helical conformation of n=10, which is based on the theoretical predictions presented in Chapter 8. Further comparisons between different geometry motifs will be given in the next section, but it will be seen that this is indeed the case for this molecule. For now, given that: (i) PBE(+vdW) is a non-empirical GGA functional; (ii) it has been shown to work very well for benchmark systems (Section 3.8.5) and for the energy hierarchy of the n=4-8 conformers (Chapters 8); (iii) and it shows better agreement with experiment than BLYP, in the



(a) BLYP+vdW trajectory

(b) PBE trajectory

(c) PBE+vdW trajectory

Figure 9.13: Evolution of the H-bond pattern of Ac-Ala$_{10}$-LysH$^+$ with time, withing a NVE *ab initio* molecular dynamics simulation: red corresponds to an α-helical H-bond, blue to 3_{10}. (a) trajectory with the PBE+vdW functional and (b) trajectory with the BLYP+vdW functional. O(Ac), O1, etc., correspond to each oxygen from the Ac and Ala residues in the molecule, starting from the N-terminus. Oxygens interacting with the Lys termination are not plotted.

harmonic and anharmonic approximation, this is the functional that we choose to use. The point to make here is that the use of the "right" functional, meaning one that contains vdW effects and that describes well the PES, is essential to obtain AIMD trajectories that lead to spectra that reproduce experimental data in detail.

9.2.4 Structure assignment for the longer helices: Ac-Ala$_n$-LysH$^+$, n=10, 15

For a careful structure assignment, we first focus on the larger molecules (n=10 and 15) where all available evidence so far (experimental, from Refs. [3, 37], and theoretical, from Chapter 8) points to a helical structure. The goal here is to obtain additional independent evidence, by comparison with experiment, that the α-helices are consistent in great detail with what is measured. In the following, two helical motifs for these molecules are tested: the α-helical and the 3_{10}-helical one. The geometries of these motifs for n=15 (very similar for n=10) are shown in Figure 9.14.

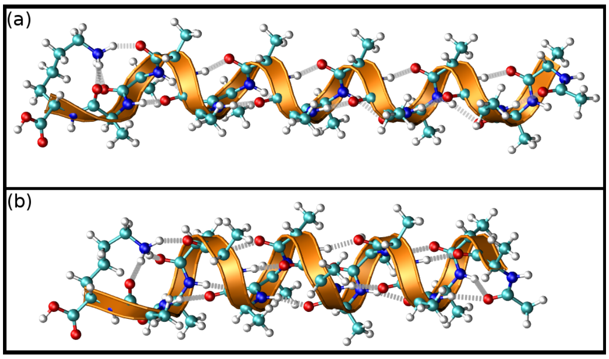


Figure 9.14: Ac-Ala$_{15}$-LysH$^+$ in: (a) 3_{10}-helical conformation; (b) α-helical conformation.

Harmonic spectra were calculated for these two helical motifs and compared to experiment. In Figure 9.15 and Figure 9.16, panels (a) and (b) show the harmonic vibrational spectra of n=15 and n=10 compared to experiment, respectively. Panel (a) in both figures shows the comparison of the calculated harmonic spectrum corresponding to the 3_{10}-helical structure motif with the experimental data, whereas panel (b) shows the comparison of the harmonic spectrum corresponding to an α-helical motif.

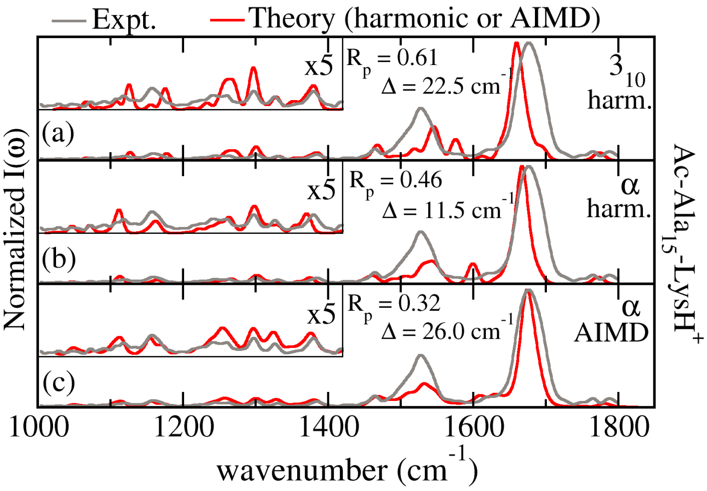


Figure 9.15: Comparison between experimental [gray lines] and theoretical [red lines] vibrational spectra, all normalized to 1 for the highest peak, for Ac-Ala$_{15}$-LysH$^+$: (a) - calculated spectra based on the harmonic approximation, for a 3_{10}-helical local minimum of the potential energy surface.; (b) same as (a) for α-helical minimum.; (c) - calculated spectrum from AIMD (including anharmonic effects), starting from an α-helix and α-helical in character throughout the simulation. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison).

It turns out that the spectra are indeed sensitive to different types of helical structure, already in the harmonic calculation, with the α-helical motif yielding a better theory-experiment match. This observation is based first on eye inspection, in which one can see the incorrect relative peak positions for the 3_{10} motif and other differences in the amide-III region (between 1000 and 1400 cm^{-1}). Then it is also based on the R_P factor values (reported in Figures 9.15 and 9.16), that clearly favors the α-helical motif.

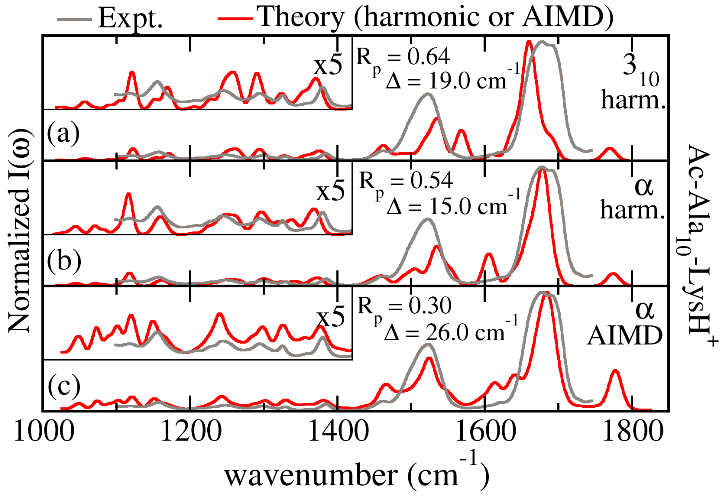


Figure 9.16: Comparison between experimental [gray lines] and theoretical [red lines] vibrational spectra, all normalized to 1 for the highest peak for Ac-Ala$_{10}$-LysH$^+$: (a) - calculated spectra based on the harmonic approximation, for a 3_{10}-helical local minimum of the potential energy surface.; (b) same as (a) for α-helical minimum.; (c) - calculated spectrum from AIMD (including anharmonic effects), starting from an α-helix and α-helical in character throughout the simulation. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison).

In the harmonic approximation, the agreement is not unambiguous even for the α-helical motifs. One can see that the R_P values of 0.46 for n=15 and 0.54 for n=10 are still to high, but it is known already from the previous section that the use of AIMD-derived spectra can improve the quality of theory-experiment comparison greatly. This is shown in panel (c) of Figures 9.15 and 9.16. These spectra now include all dynamical, temperature and anharmonic effects (but no nuclear quantum effects). Not only are R_P=0.32 for n=15 and R_P=0.30 for n=10 noticeably improved, but also the visual comparison of all relative peak positions is remarkably better. For example, the peak appearing around 1600 cm^{-1} in the harmonic spectra, lying in the middle of the gap observed in the experimental spectra, is considerably shifted upon calculation of the anharmonic spectra. This peak, as already mentioned earlier, comes from bends of the NH$_3^+$ group of the lysine. As discussed in Appendix F, the shape of this vibrational mode is not strongly anharmonic, so a strong shift in frequency is not expected. This mode probably loses intensity through anharmonic coupling with other modes. The separation between the highest peak of the amide-I, the highest peak of the amide-II and the beginning of the amide-III region at 1400cm^{-1} is very well reproduced for the anharmonic spectra, while this is not the case for the harmonic ones. The peaks in the not so intense amide-III region are also very well reproduced by the simulated anharmonic spectra, as can be seen in the zoom of panel (c) of Figures 9.15 and 9.16. Only some residual quantitative uncertainties remain, for example in the region of 1450-1600 cm^{-1}. A certain degree of quantitative discrepancies are expected due to residual errors of any approximate density functional, as well as the other approximations employed here (e.g., simulation of a linear absorption spectrum, absence of nuclear quantum effects, short trajectories). As has already been discussed differences in relative intensities of the peaks are possibly due to the non-linear nature of the measured spectra. Overall, the agreement between theory and experiment when including anharmonicities is remarkable.

Here, no anharmonic spectra for the 3_{10}-helical motif of n=10 and 15 are presented, and for a good reason: initial, locally stable 3_{10} arrangements tend to transform into α after only a few ps of AIMD, in at least three different (with different initial conditions) AIMD simulation attempts, making it impossible to obtain spectra. The reason is that for n=10 and 15, 3_{10} helical local energy minima are higher in energy than α by at least 0.41 and 0.82 eV respectively, in PBE+vdW. Computing harmonic free energies at

T=300K for these helical motifs does not change these energy differences noticeably. Similarly, it will be seen in the next chapter that simply running a thermostated molecular dynamics simulation at T=300K starting from a 3_{10} helical structure (Figure 10.8) transforms into a mainly α-helical one in less than 10ps of simulation time.

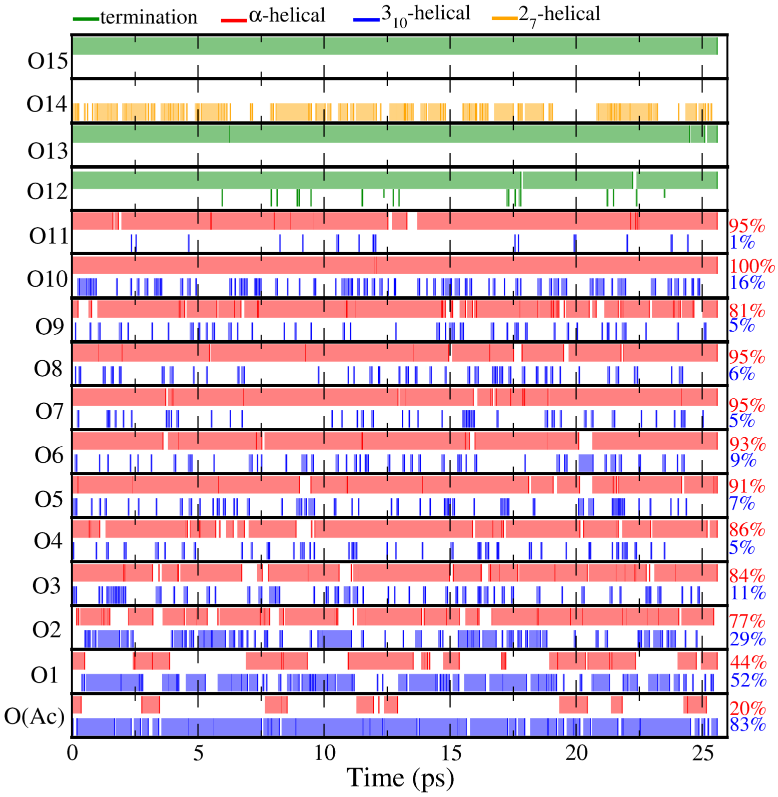


Figure 9.17: Evolution of the H-bond pattern of Ac-Ala$_{15}$-LysH$^+$ with time, within a microcanonical *ab initio* PBE+vdW molecular dynamics simulation: red corresponds to an α-helical H-bond, blue to 3_{10}, yellow to 2_7, and green corresponds to H-bonds to the NH_3^+ group of the Lysine termination.

Even though the molecules in question have been so far denoted α-helical, in reality (dynamics) a single pure, exact conformation is not expected. In Figure 9.17, the detailed H-bond network evolution of Ac-Ala$_{15}$-LysH$^+$ during the NVE molecular dynamics simulation that led to panel (c) Fig. 9.15. Examining Figure 9.17 closely, it is apparent that already at t=0 the molecule is not a perfect α-helix. This is because these runs are started after a few picoseconds of temperature equilibration with a thermostat. The H-bond associated with the Ac termination is predominantly 3_{10} helical all through the trajectory, the next one is predominantly bifurcated, and a bifurcation can happen with less probability for one H-bond further up. Yet, the next nine H-bonds in the structure (all those remaining up to the LysH+ termination) are α-helical more than 84% of the time. For n=10 and PBE+vdW the situation is quite similar, as was seen in Figure 9.13(a) of the previous section. It is, thus, possible to state that for n=10 and 15, the molecule is firmly α-helical in character.

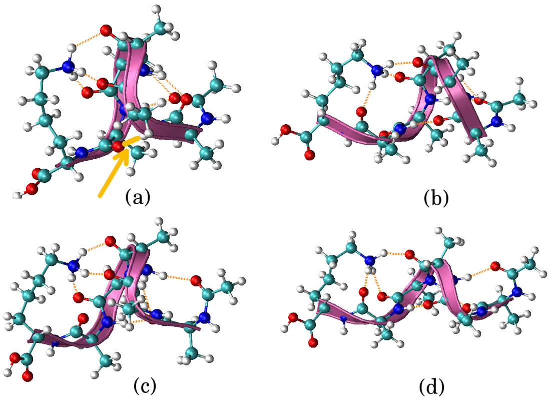
9.2.5 Short polyalanine: Ac-Ala$_5$-LysH$^+$.

For Ac-Ala$_5$-LysH$^+$, there is a competition between various different low energy conformations, as already discussed in great detail in Chapter 8. In the following, we shall address the question of whether it is possible to interpret the experimental spectrum also in this case. As low energy candidates, the same conformers as discussed in Chapter 8, Section 8.2 are chosen, plus the lowest 3_{10} helical conformer. For clarity, these conformers, with their respective H-bond patterns are depicted in Figure 9.18. Three of

Table 9.2: Total and free energy differences of the four chosen Ac-Ala$_5$-LysH$^+$ conformers wrt. g-1: DFT-PBE+vdW (PES only), zero-point corrected total energy, and DFT-PBE+vdW harmonic free energy ΔF at 300 K. All energies in eV.

	g-1	α-1	α-2	3_{10}-1
DFT-PBE+vdW	0.0	0.10	0.11	0.19
+ ZPE	0.0	0.07	0.08	0.17
ΔF(300 K)	0.0	0.04	0.07	0.17

them (labeled α-1, α-2, 3_{10}-1) are “helical” in the sense that they contain two well-separated terminations with the appropriate α- or 3_{10}-like H-bond loops in their Ala$_5$ section. The fourth conformer, however, labeled g-1, is the overall lowest-energy conformer, and is *not* “helical” in the same sense. In particular, it contains one H-bond (O-5 to NH-2) that runs *against* the normal helix dipole, effectively short-circuiting the terminations.



(e)	g-1	α-1	α-2	3_{10}-1

O(Ac)	NH-4(α)
	NH-5(π)
		NH-3(3_{10})
		NH-4(α)
			NH-3(3_{10})	NH-3(3_{10})

O-1	NH-3(2_7)	NH-5(α)	NH-4(3_{10})
			NH-5(α)
				NH-4(3_{10})

O-2	NH_3^+	NH_3^+	NH_3^+	NH-5(3_{10})
O-3	NH_3^+	NH_3^+	NH_3^+	NH_3^+
O-4	NH_3^+	NH-6	NH_3^+	NH_3^+
O-5	**NH-2**	NH_3^+	NH_3^+ weak	NH_3^+

Figure 9.18: Visualization of low-energy Ac-Ala$_5$-LysH$^+$ conformers: (a) g-1; (b) α-1; (c) α-2; (d) 3_{10}-1. (e) H-bond networks associated with each conformer. (C-)O and N-H groups are numbered starting from the N terminus and ending at the C terminus.

In Table 9.2, the relative energies of these four conformers are repeated for the PBE+vdW PES, including the zero point energy, as well as the free energy differences at T=300K (Eq. 4.20). These numbers, except for the 3_{10}-1 conformer, are also reported in Table 8.3: g-1 corresponds to Family 1 of n=5, α-1 to Family 2, and α-2 to Family 3.

In DFT-PBE+vdW, the g-1 conformer is more stable than its closest competitors by 0.1-0.2 eV. On the other hand, finite temperature effects reduce the relative stability of g-1, as has already been seen and discussed in Chapter 8.3. With the inclusion of the harmonic free energies at 300K, both the α-helical conformer α-1 and the mixed helix conformer α-2 are only some tens of meV removed from g-1; only 3_{10}-1 stays noticeably higher in energy, although it exhibits some stabilization with respect to g-1, as was expected from the discussion in Section 8.3.1.

The expected stability of at least three out of the four conformers is thus similar, and at room temperature, one would expect all of them to be present or even interconvert in experiment. Indeed, it was already seen in Section 9.1, when comparing ion mobility cross sections, that the consideration of all conformers, weighted by their relative Boltzmann population at 300K was able to reproduce the experimental curve. From the discussion in Chapter 8 - Section 8.2, it is also known that different functionals (corrected for vdW) can change the relative position of α-1 and α-2. The 3_{10}-1 conformer is always the highest in energy for the functionals tested.

Figure 9.19 (a)-(d) shows the computed *harmonic* spectra of all four conformers compared to experiment, while Figure 9.20 (a)-(d) shows the anharmonic ones obtained from AIMD trajectories. Since in this case $I(\omega)$ (anharmonic) are derived from dynamical trajectories, different conformers could interconvert over time, regardless of the starting structure. In fact, an interconversion happens for short periods between α-1 and α-2. The g-1 conformers stays firmly in its starting conformation through all the 20ps of simulation, while the 3_{10} conformer turns into α-2 in the final picosecond of the simulation only [9]. The detailed evolution of the H-bond pattern for these four conformers during the simulation can be seen in Appendix G. The spectra are still labeled for the initial (and predominant) conformation.

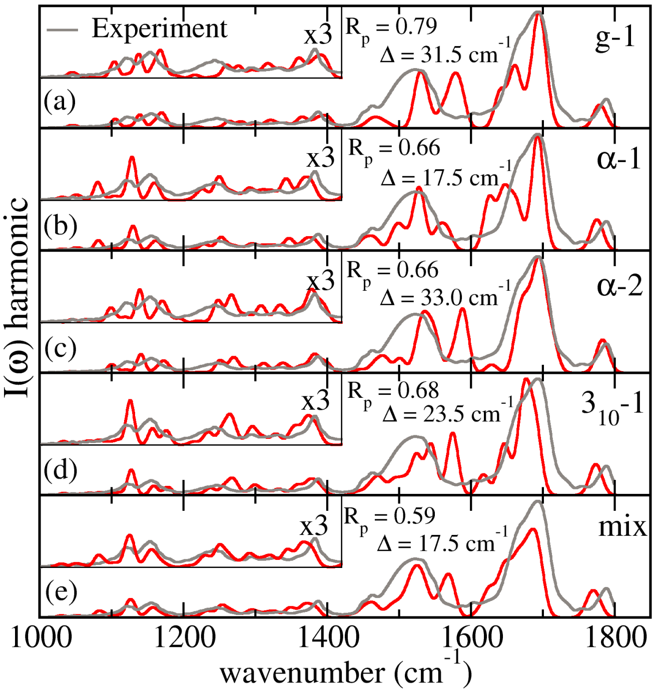


Figure 9.19: Ac-Ala$_5$-LysH$^+$: (a)-(d) Theoretical harmonic vibrational spectra (red lines) for the four chosen conformers of 9.18, compared with experiment (gray line); (e) optimum calculated spectrum when assuming a coexistence of more than one conformer in experiment. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison). All spectra have been normalized to 1 for the highest peak.

The anharmonic spectra presented in Figure 9.20(a)-(d) agree reasonably well with experiment, but each individual R_P factor somewhat high - between 0.44 and 0.60 - and only the 3_{10} conformer can be more readily set apart upon eye inspection by non-matching peak positions. Even so, this situation is much better than the harmonic one, shown in Figure 9.19(a)-(d). The visual comparison is not good for any of the individual spectra and the R_P factors are much higher - between 0.66 and 0.79. It is known from the cases of n=10,15 that a good match produces R_P around 0.3, and the free energies estimates presented above already give a hint about why the same good agreement is not seen here for n=5 even in the anharmonic case: there is probably a mix of different conformers being measured at the same time in the beam.

To test this possibility both for the harmonic and anharmonic case, Figures 9.19(e) and 9.20(e) show the result of simply linearly averaging theoretical spectra, with mixing factors derived by minimizing R_P. The result for the harmonic case still turns out to be unsatisfactory, with a minimum R_P of 0.59 (still too high) obtained for fractions of 15% g-1, 40% α-1, 25% α-2 and 20% 3_{10}-1. The result of this fit for the anharmonic case, on the other hand, is quite remarkable. The "optimum" theoretical spectrum achieves

[9] In particular, α-2 also visits the 3_{10}-1 conformation for a short period of time, as can be seen in Appendix G.

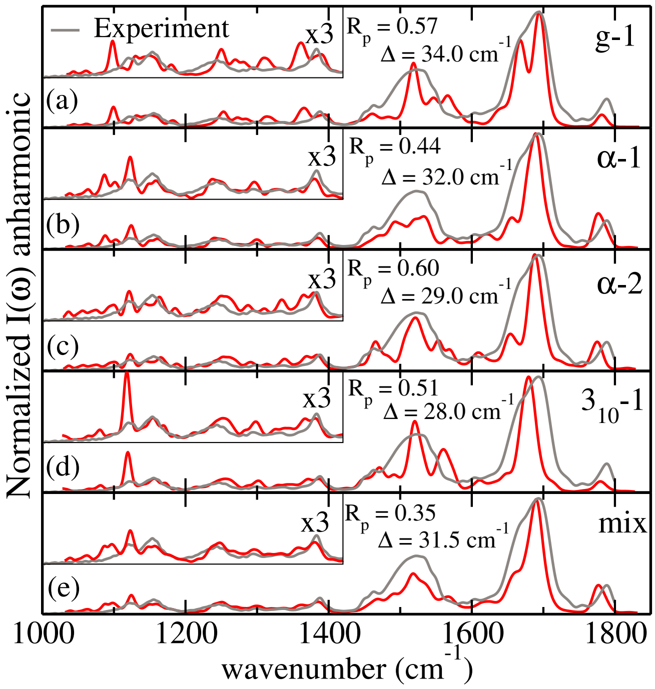


Figure 9.20: Ac-Ala$_5$-LysH$^+$: (a)-(d) Theoretical anharmonic vibrational spectra from AIMD trajectories (red lines) for the four chosen conformers of 9.18, compared with experiment (gray line); (e) optimum calculated spectrum when assuming a coexistence of more than one conformer in experiment. Pendry R-factors and rigid shifts Δ (see text) between measured and calculated spectra are included in each graph (calculated spectra are shifted by Δ for visual comparison). All spectra have been normalized to 1 for the highest peak.

R_P=0.35, with convincing visual agreement to experiment in the details. This could be interpreted as a simple outcome of a fit, but in addition, the computed optimum fractions are 25% g-1, 60% α-1, 15% α-2, and *no* contribution from 3_{10}-1 at all. This coincidence with our free-energy based conclusions strengthens our confidence on the results of our structural search significantly.

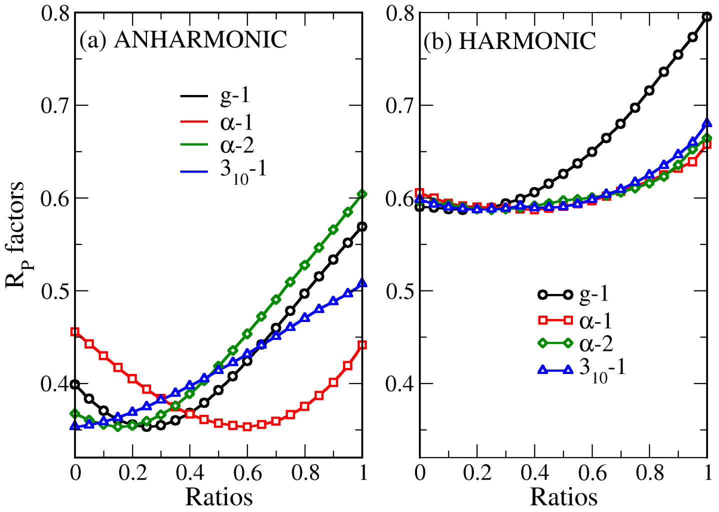


Figure 9.21: Variation of R_P factors with respect to the ratios of Ac-Ala$_5$-LysH$^+$ spectra of the four chosen low energy conformers: (a) anharmonic spectra; (b) harmonic spectra.

The averaging procedure produces only one minimum when mixing the anharmonic spectra of the four conformers, corresponding to the ratios described above. The variation of R_P with respect to the ratio of the anharmonic spectra is plotted in Figure 9.21(a). On the other hand, the variation of R_P with respect to the ratios of harmonic spectra, shown in Figure 9.21(b), has a much wider and less well-defined minimum. The good agreement with experiment is, therefore, tightly connected with the inclusion of all relevant physical effects: electronic structure through DFT, van der Waals interactions, temperature, conformational freedom, and anharmonicities.

High wavenumber range (amide-A and -B)

No experimental data was available for the amide-A and -B range (hydrogen stretches, $>$ 3000cm^{-1}) for the conformers discussed here. However, this range is much more sensitive to different H-bond patterns with respect to the amide-I, -II, and -III discussed above. Experimental data in this range could give much more precise information about the exact conformations that are being measured in experiment. As an example, the calculated harmonic (PBE+vdW, not broadened or shifted) spectra for the 4 conformers of n=5 is shown in this wave-number range in Figure 9.22. The differences between the spectra of the 4 conformers are much more pronounced in this vibrational range.

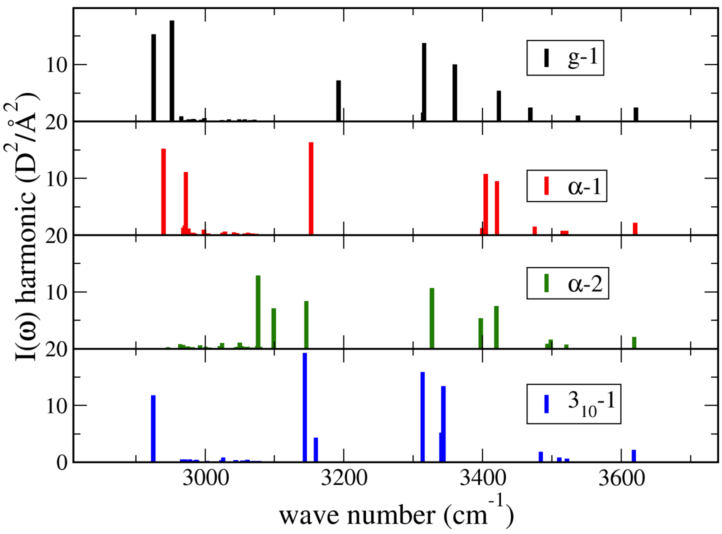


Figure 9.22: High wavenumber (amide-A and -B) range of the harmonic IR spectra (PBE+vdW, not broadened or shifted) of the four conformers of n=5 discussed in this chapter.

9.2.6 Amide-I and amide-II peak shifts

As has been shown in Section 6.3.4 a shift to the red is observed experimentally for the amide-I peak as n increases, while a shift to the blue is observed for the amide-II peak (Figures 6.6 and 6.7). These shifts are reproduced by the above calculations already in the harmonic approximation, assuming that the n=5 spectrum corresponds to a mix of different conformations and n=10 and 15 are α-helices, as was discussed above. The shift can be seen in Figure 9.23, where the harmonic spectra for the α-helices of n=10 and 15 were plotted with the mixed harmonic spectrum of Figure 9.19(e) for n=5, without applying any shifts to them.

As discussed in Section 4.4, these peaks are related mainly to collective C=O stretches (amide-I) and collective NH bends (amide-II). These modes can therefore be related to phonon modes that appear in infinite (pure) polyalanine helices [324]. The values for these peaks in the “phonon” limit can be taken from Ref. [324], where these phonon modes have been calculated for the infinite α-helices with DFT-PBE. Indeed, the amide-I peak in the phonon limit has frequency of $\sim$1650cm^{-1} for PBE, and the amide-II has a frequency of $\sim$1550cm^{-1}. These peaks for n=5, 10, and 15 have a higher value for the amide-I band and a lower value for the amide-II band. It can be thus interpreted that increasing the helix length and H-bond strength due to the cooperativity effect discussed in Section 2.2.1, these modes gradually approach the phonon limit.

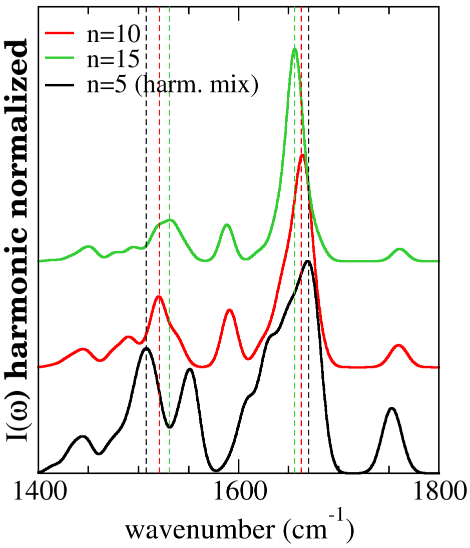


Figure 9.23: Calculated garmonic infrared spectra of Ac-Ala$_5$-LysH$^+$ (black, bottom), Ac-Ala$_{10}$-LysH$^+$ (red, center), and Ac-Ala$_{15}$-LysH$^+$ (green, top) in the amide-I and amide-II band region. All spectra are normalized to 1 with respect to the highest peak intensity and the spectra have an offset in the y-direction, for better readability. The color of the dashed lines represents the assignment of the corresponding molecule to the peak position.

9.3 Summary

In this chapter, a direct connection between the first-principles theoretical predictions and experiment has been presented.

Evaluation of ion mobility cross sections for the conformers that were characterized for n=4–8, 10 (α), and 15 (α), calculated here from an empirical potential, but with the geometries taken from DFT-PBE+vdW for each n was shown. They reproduce well the experimental data of Ref. [37], when considering a Boltzmann average (300K) with the weights calculated from the DFT-PBE+vdW harmonic free-energy differences.

First-principles IR spectra were compared with experimental (room temperature) IRMPD spectra, for Ac-Ala$_n$-LysH$^+$, n=5, 10, 15. Since the spectra and the simulations are structure-sensitive, it was possible to confirm that for n=10 and 15 the structure of this molecule is firmly α-helical, as was expected from the predicted α-helical onset at n=8. The experimental spectrum for the shorter conformer, n=5, could be explained by a mixture of low energy conformers co-existing in the beam at 300K. Both helical (α-1 and α-2) structures, as well as more compact (g-1) structures should be present, in agreement with the previous (harmonic) free-energy hierarchy analysis, made in Chapter 8. In all cases, the evaluation of spectra including anharmonicities, temperature, and configurational freedom effects via the dipole autocorrelation function, derived from AIMD runs, were seen to improve substantially theory-experiment comparison. This comparison was quantified via the use of a reliability factor (Pendry R-factor).

Chapter 10

Investigating the high temperature helical stability of Ac-Ala$_{15}$-LysH^{+}

In this part of the work, the goal is to find out how far first-principles DFT can go in order to understand the dynamics of folding and unfolding for a medium-sized polypeptide, namely Ac-Ala$_{15}$-LysH^{+} (180 atoms). This system was chosen because of its many interesting characteristics: (i) the extensive work of Jarrold *et al.* plus the calculations discussed in Chapter 9.2 have shown unambiguously that these systems form helical secondary structures in the gas phase [37]; and (ii) the ion-mobility experiment discussed in Section 6.3.3 showed that the helical structure is stable until ~700K *in vacuo* [5]. It is thus meaningful not only to simulate this molecule in the gas phase, but also to run the simulations at high temperatures, where the molecule can overcome barriers faster and the available time scales for first-principles methods can be used efficiently.

To give a better idea of the time scales involved in conformational changes, small and "fast-folding" proteins like the Trp-cage have been experimentally shown to fold (in solution) in 1 to 4 μs [386], while small polyalanine helices have been shown to unfold in between 200ns and 1μs (at $T \approx$ 340 K) [387, 388]. Initial secondary structure formation in small peptides, (both in the process of folding and in that of unfolding) happens much faster, though, in the scale of picoseconds to very few nanoseconds [389–391]. Experimental studies [392] of a small alanine-based polypeptide (trialanine) have shown that the changes in backbone angle that affect the structure of this molecule (in water) occur at a very short time-scale, of less than 1ps. Therefore, the time scale at least for local structural changes is reachable for *ab initio* methods. In practice, in this study, attempts are made to unfold (and later to fold) the α-helical motif of Ac-Ala$_{15}$-LysH^{+} at high temperatures, using AIMD. The work presented on this chapter has been partly published in Ref. [6].

10.1 Harmonic free energies

Before going to the AIMD simulations, in order to get an estimate of the expected temperature stability of Ac-Ala$_{15}$-LysH^{+} in PBE+vdW (TS-vdW scheme), harmonic free energy calculations for three different motifs of the structure were performed. The motifs chosen were the fully extended structure (FES), shown in Figure 10.1, the α-helix, and the 3_{10}-helix (previously shown in Figure 9.14, but repeated here in Figure 10.2). All of them are local minima of the PES in the PBE+vdW functional.

The temperature dependence between 0K and 500K of the relative free energies for these three

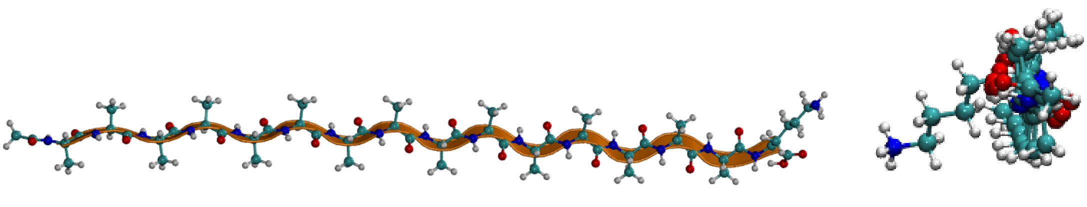


Figure 10.1: Fully extended structure (FES) of Ac-Ala$_{15}$-LysH^{+}, relaxed with PBE+vdW: (a) side-view; (b) end-view, exhibiting slight tilt from the pure β-strand conformation. The α helical and 3_{10} helical conformations considered in this part of the work are the same that have been shown in Figure 9.14

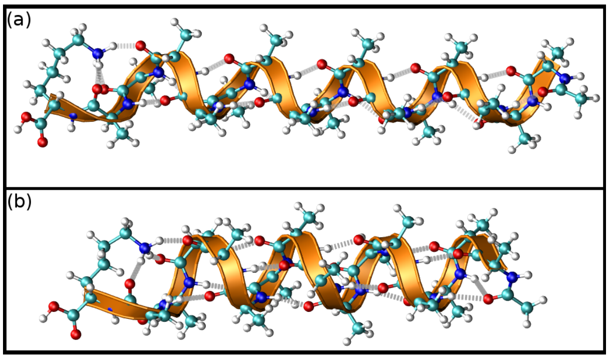


Figure 10.2: Schematic picture of Ac-Ala$_{15}$-LysH^{+} in: (a) 3_{10}-helical conformation; (b) α-helical conformation.

structural motifs is plotted in Figure 10.3. The potential-energy difference (at 0 K) between the α-helix and the fully extended structure amounts to -0.38 eV per residue. Inclusion of the zero-point energy increases this value to -0.36 eV and the free energy difference at 500 K further increases to -0.25 eV per residue, illustrating the role that vibrational free energy plays in destabilizing the α-helix with respect to the fully extended structure.

As discussed in Chapter 8, the vibrational free energy *stabilizes* α helices with respect to *globular*, *compact* structures. In an extended structure like the fully extended geometry, that exhibits a periodically repeating pattern, the lowest vibrational modes [1] start already at 2 cm^{-1}, corresponding to the same (delocalized) bending-of-terminations vibrations discussed in Chapter 8. These are the vibrational modes responsible for stabilizing the fully extended structure over the α-helix, since the α- and 3_{10}-helices studied here have these first "phonon" modes starting at higher frequency ($\approx$ 11 cm^{-1} and $\approx$ 10 cm^{-1}, respectively). Indeed, if one goes to the periodic, infinite limit with a purely alanine-based α helix and a fully extended structure, it has been previously shown that the optical phonon modes of the fully extended structure are always at a lower frequency, if compared to the α-helix (see Refs. [324, 325]). From Chapter 8, we have seen that the more compact/globular conformations have much harder low vibrational modes. Although the fully extended structure is taken here as a reference point, against which to compare the stability of the helical conformers, it is certainly not a general representative of the random-coil ("unfolded") ensemble [71, 82]. The FES would actually be lying on the tail of the end-to-end distance distribution of the random-coil. It is very hard to estimate quantitatively what is the vibrational free energy effect on the stability of the α-helix with respect to this ensemble, based solely on these "static" calculations. It is possible, though, to consider the destabilization with respect to the fully extended

[1] The accuracy of this calculation is also around 2 cm^{-1} for the vibrational frequencies.

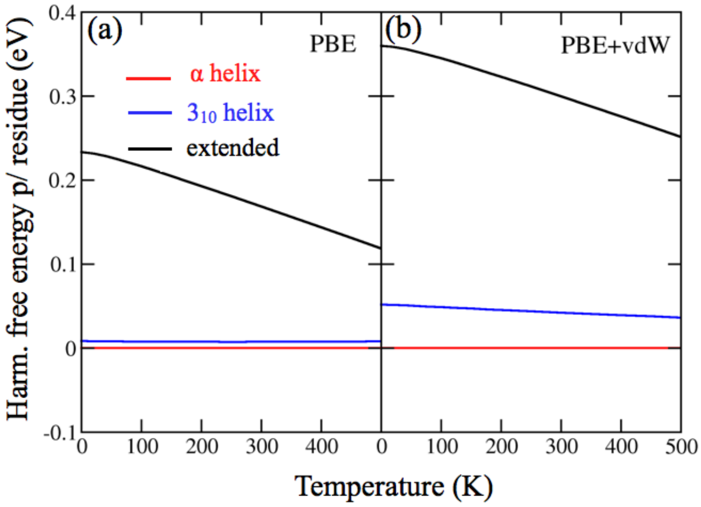


Figure 10.3: Vibrational free energies in the harmonic approximation of Ac-Ala$_{15}$-LysH$^+$, for the different structural motifs considered here: α (red), 3_{10} (blue), and fully extended structure (black). In (a) the PBE results are shown and in (b) the PBE+vdW results.

structure discussed above (-0.25eV per residue at 500K) as a lower limit to the possible destabilization of the α helix.

That said, there is still a missing contribution to the free energy difference between the unfolded random coil and the folded α-helix, which is the backbone conformational entropy, $\Delta S_{\rm conf}$ [50]. $\Delta S_{\rm conf}$ has been estimated for polyalanine helices from classical force field simulations in solution. The most recent estimates (weakly dependent on the employed force field) arrive at $\approx$ 0.09 eV per residue at 500 K, as reported by Baldwin [50]. The addition of $\Delta S_{\rm conf}$ would reduce the free-energy difference between Ac-Ala$_{15}$-LysH$^+$ α-helix and the limiting case of the fully extended structure, to -0.16 eV per residue. However, the α helix remains the most stable free-energy structure.

In contrast to PBE+vdW, in plain PBE (also plotted in Figure 10.3) the α and 3_{10} structures are similarly stable over the entire range of investigated temperatures, in disagreement with gas-phase experiments. Additionally, the fully extended structure at 500K, including all entropic contributions, would be almost as stable as the helices.

10.2 Dynamics and temperature stability of the helix

In this section, the dynamics and stability of Ac-Ala$_{15}$-LysH$^+$ will be studied by trying to unfold this polypeptide, performing AIMD simulations at various temperatures. All AIMD runs presented here used a Nosé-Hoover thermostat, with a 1 femtosecond time step, a SCF force convergence threshold of 5x10^{-4} eV/A, and a thermostat mass corresponding to 1700cm^{-1}. These settings have been shown to produce reliable results in Chapter 4.

10.2.1 Importance of van der Waals interactions for the gas phase high-temperature helix stability

To analyze the actual dynamics of polyalanine unfolding beyond the harmonic approximation, *ab initio* MD simulations of Ac-Ala$_{15}$-LysH$^+$ were performed with the PBE and the PBE+vdW functionals at different temperatures. Starting from a perfect α-helix, it is possible to monitor what the hydrogen bonds of the molecule do during an AIMD run. For PBE and PBE+vdW simulations at room temperature (300K), the total number of existing H-bonds is plotted against the time of simulation (taking a 100fs running average) in Figure 10.4. For a perfect α-helical structure, ignoring the H-bonds to the LysH^{+} termination, there is a total of 12 hydrogen bonds, which are here defined by a maximum CO–NH distance of 2.6Å [2]. PBE+vdW yields a stable α-helix in the short AIMD run shown in Figure 10.4. Plain PBE, on the other hand, shows a tendency to go to a predominantly 3_{10}-helical structure. This observations is again consistent with the previously observed fact that in the absence of vdW interactions, 3$_{10}$ helices are stabilized over α-helices (Ref. [379] and Chapter 8).

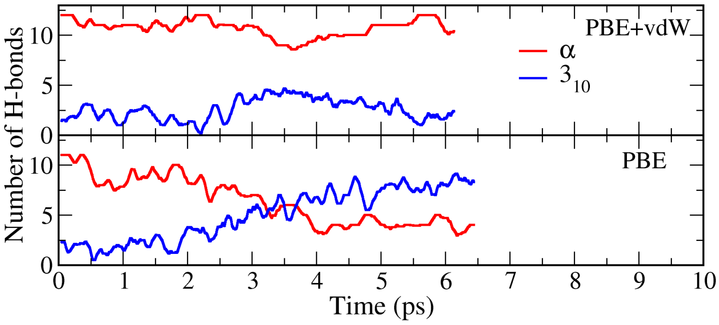


Figure 10.4: MD simulations of Ac-Ala$_{15}$-LysH+ peptide at 300 K. Number of hydrogen bonds throughout the MD trajectories for a PBE+vdW simulation (top) and a PBE simulation (bottom) are shown. Only α and 3$_{10}$-helical hydrogen bonds are presented, excluding π, 2$_{7}$, and non-bonded residues.

The helix is not expected to unfold at 300K, as was estimated by the free energy differences in the last section. Therefore, a more interesting “computer experiment” is to run the AIMD simulations at higher temperatures. An initial test (not shown) at T=1000K unfolds the molecule fully in just a few picoseconds (less than 5 ps) for both PBE and PBE+vdW [3]. This temperature is quite extreme, demonstrating that it is possible to unfold the molecule in a reasonable time scale.

The intermediary temperatures of 500, 700, and 800 K were also simulated. In Figure 10.5 the number of hydrogen bonds characteristic of a particular helix type (α and 3$_{10}$) is shown, as a function of MD simulation time up to 65 picoseconds (ps) for 500, 700, and 800 K, starting from a perfect α-helical geometry (the detailed H-bond pattern of these molecules during the 500 and 700K simulation can be seen in Figure 10.6). Not all simulations were of the same length, that is why some curves are shorter than 65ps. Snapshots of the conformations adopted by the molecules at particular times of each simulation are also presented. The PBE+vdW functional preserves a helical structure throughout the MD simulation at 500 K, with $\approx$ 80% of the H-bonds being of α-helical type at 500 K. [4] A substantially different behavior is predicted by plain PBE, in which after 10 ps at 500 K, the molecule preserves less than two “α-helical” H-bonds on average and its helical part ($\approx$ 50%) is mainly of 3$_{10}$ nature.

At 700 K, PBE+vdW yields a helix of a mixed α-helical and 3_{10}-helical nature, but the overall helical structure is preserved even in a MD simulation as long as 65 ps [5]. In contrast, at 800 K, the α-helical structure has essentially disappeared after 10 ps, denoting unfolding. Once again, PBE yields a very different picture. At 700K and 800K, the α-helical H-bonds largely disappear after 5-7 ps. At 800 K the polypeptide is almost fully unfolded after 7 ps and remains unfolded up to 30 ps. The conclusion is that

[2] This is a bit larger than the 2.5 Å used previously because at “high” temperatures the bonds tend to stretch more, although they do not break.

[3] In the PBE simulation, some hydrogens of the molecule, in the Ac termination, dissociate.

[4] For this temperature, in addition, MD simulations of up to 30 ns (a time scale that cannot be reached with DFT-PBE+vdW here) with the empirical OPLS-AA [30] force field were performed. The trend observed there is essentially a continuation of the trend seen over 30 ps in Fig. 10.5 with PBE+vdW, showing no unfolding.

[5] A movie showing the evolution of the helix at 700K, with the PBE and PBE+vdW functional can be found in http://www.youtube.com/watch?v=Y_7G8s26zzw.

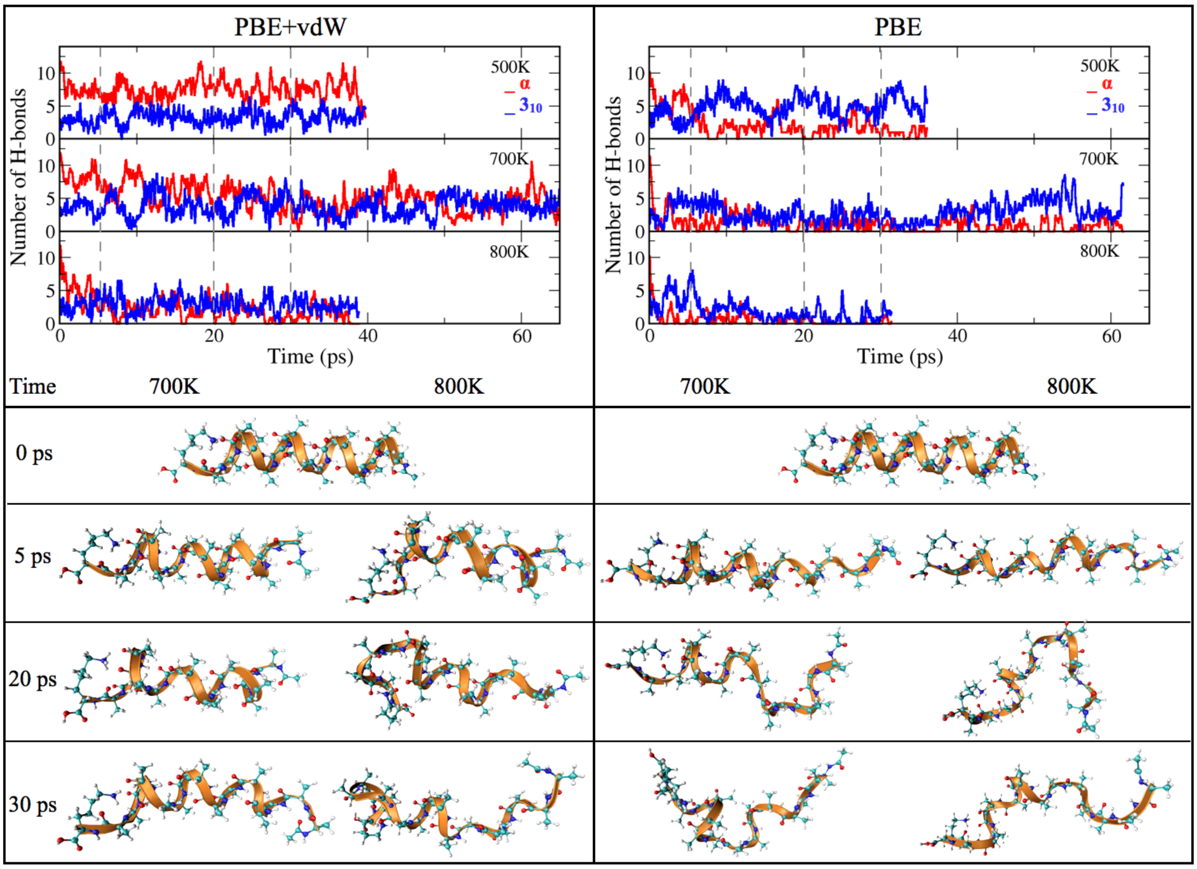


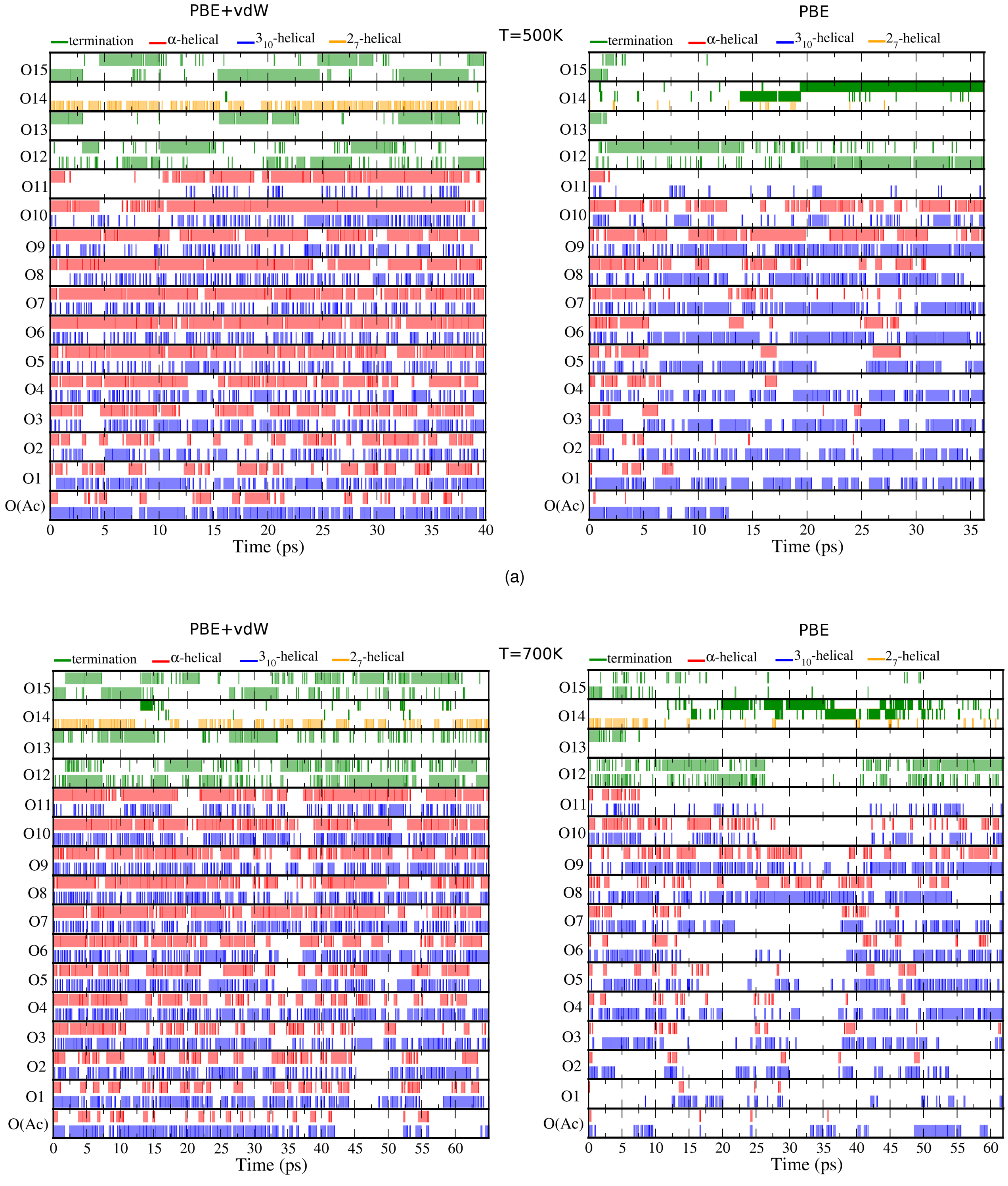
Figure 10.5: MD simulations of Ac-Ala$_{15}$-LysH+ peptide at 500, 700, and 800 K. Upper panels show the hydrogen bonding pattern throughout the MD trajectories, while lower panels present snapshots from MD simulations at 5, 20 and 30 ps. Only α (red) and 3$_{10}$-helical (blue) hydrogen bonds are shown, excluding π, 2$_7$, and non-bonded residues.

PBE yields a much less stable helix, which is not consistent with experiments. Nevertheless, longer AIMD simulations would be required to precisely determine the unfolding temperature of this molecule.

These differences between PBE+vdW and PBE on the H-bond network for the temperatures of 500K and 700K can be seen in even more detail in Figure 10.6(a) and (b). There, the H-bond connection of each oxygen of the molecule is plotted versus the time of simulation, in the same way that was explained in Figures 9.13 and 9.17 of the previous chapter. One can see clearly that at 700K in the PBE simulation, after only a few picoseconds almost all oxygens of the molecule show no connections, which is indicative of unfolding.

Clearly, PBE and PBE+vdW seem to be exploring different regions of the conformational space. This is best seen on the pitch/twist map (Ramachandran-type plot in cylindrical coordinates, see Chapter 2). The regions, in this kind of plot, corresponding to different secondary structure motifs was discussed in Section 2.2. Figure 10.7 shows these plots for the simulations at all temperatures, with and without vdW interactions. The quantity plotted there is the average, through all the simulation time, of all pitch-twist coordinates corresponding to the 12 alanine amino-acids lying in the center of the molecule (disregarding both terminations). The PBE+vdW simulation at 500K is the only one to explore the region corresponding to π-helices (pitch < 1.33Å), with small but not negligible probability. With the PBE functional, the molecule explores areas corresponding to more extended structures in the conformational space (2_7, PPII regions) at a much lower temperature than PBE+vdW. The highest probability for PBE is also always shifted to higher pitch (length of the helix) values if compared to PBE+vdW, confirming that the PBE conformers are always more elongated.

Another interesting observation came from starting again the simulations at 500K, with PBE and PBE+vdW, but from a 3_{10}-helical structure. The result, shown in Figure 10.8, is that the PBE+vdW simulation tries to take the molecule to an α-helical structure. The PBE simulation tries also to rearrange the molecule, but remains with very few α-helical connections, and many more 3_{10}-helical H-bonds. Similar behaviors have been reported in the literature, but from force-field studies. For example, a MD simulation starting form a 3_{10}-helical conformation of a 11 residues polyalanine [31] with OPLS in water, ended up in an α-helical conformation in about 20ps. Additionally, in Ref. [393] the stabilities of these two helical motifs for a deca-alanine polypeptide were compared in the gas-phase and in water (with the Cedar program), concluding that the α-helix is even more stabilized over 3_{10} structures in solution than in the gas-phase. Although these observations are important, we know from Chapter 7 that at least the OPLS-AA force-field overestimates the relative energies of 3_{10} helices, which may bias the results.



(a)

(b)

Figure 10.6: Detailed H-bond network of the unfolding simulations of Ac-Ala$_{15}$-LysH$^+$ at 500K and 700K using the PBE and PBE+vdW functionals. The oxygens are counted starting from the Ace termination and going up to the Lys termination. Red bars correspond to α-helical H-bonds, blue to 3_{10}, yellow to 2_7, and green to H-bonds connecting to the LysH$^+$ termination.

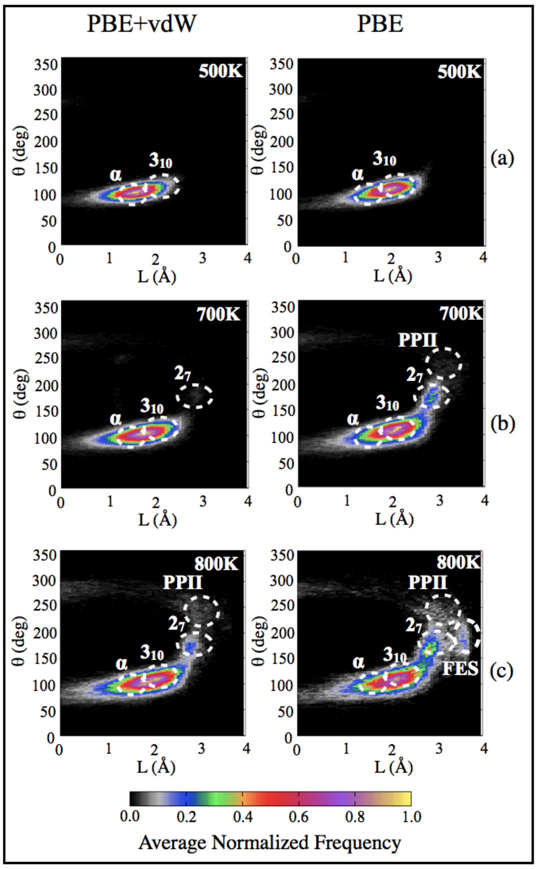


Figure 10.7: “Ramachandran map” in cylindrical pitch/twist coordinates [32, 35] for MD simulations with PBE and PBE+vdW at T=500K, 700K, and 800K. The color code corresponds to the probability (from 0 to 1) of visiting a certain conformation (see also text).

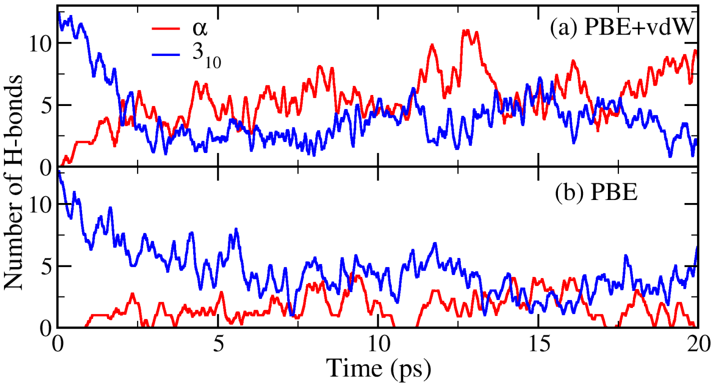


Figure 10.8: MD simulations of Ac-Ala$_{15}$-LysH^{+} peptide at 500 K, starting from a 3_{10} helical structure. Panels show the number of α-helical H-bonds (red) or 3_{10}-helical H-bonds (blue) versus time of simulation: (a) PBE+vdW and (b) plain PBE.

10.2.2 Importance of the termination (charge and connecting H-bonds)

The stabilizing effect of the H-bond network and the charge close to the C-termini of polyalanine helices has been studied using DFT(PBE+vdW) in Ref. [6], which is also (partly) the work of the author of this thesis. In order to explain here what was found, Figure 1 of Ref. [6] is reproduced in Figure 10.9. The quantity plotted is the energy to add one amino acid residue to a finite polyalanine chain is computed, as a function of chain length n:

$$E_{Ala}(n) = E_{tot}(Ala_n) - E_{tot}(Ala_{n-1}) \tag{10.1}$$

where E$_{tot}$(Ala$_n$) is the total energy of the n residues alanine helix, computed at the PBE and PBE+vdW level. The Ala$_n$ chain is frozen at the geometry of a hypothetical, infinite periodic α-helix, such that the periodic limit is systematically approached by the addition of extra residues. The terminating groups COOH and NH$_2$ as well as the Ala residues closest to the C terminus are relaxed for n =5, and then kept at that structure for larger n. Neutral helices, as well as one containing a Li$^+$ atom (depicted in Figure 10.9) are studied.

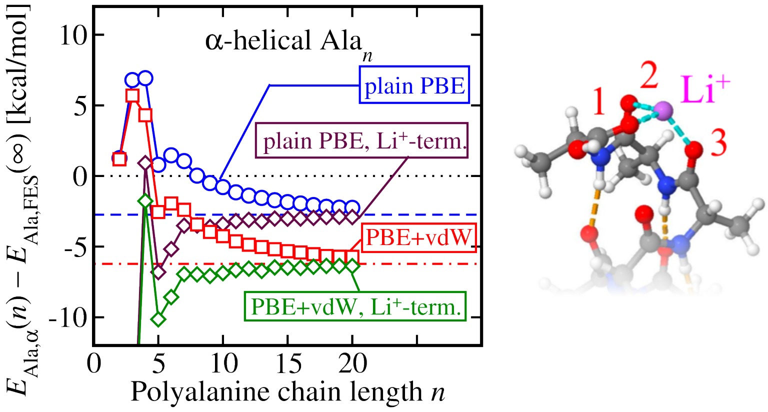


Figure 10.9: Reproduced from Ref. [6]: "Energy per added alanine peptide unit for idealized polyalanine helices as a function of peptide length, referenced to as infinite fully extended polyalanine structure. Circles and squares: neutral helices. Diamonds: Li$^+$-capped helices. A cartoon of the Li$^+$-capped C-terminus structure is shown on the right. The labels 1, 2, 3 indicate the dangling hydrogen bonds saturated by the ionic termination."

A significant cooperative effect between hydrogen bonds can be observed, for example, in the blue curve (circles, neutral helix) of Figure 10.9, for PBE. The energy is seen to go up until the point where the first H-bond is formed (n=5), where it undergoes a remarkable energy drop. From then on, the cooperative effect increases rather slowly with chain length n, towards the limit of an infinite periodic chain (dashed line). It approaches the limit only for n ≈20. Including the vdW contribution (red squares) essentially shifts down the PBE curve (for large n), more than doubling the α-helical stability. The fact that the PBE and PBE+vdW curves are parallel (for larger n), but start essentially at the same energy (n=2-3) points to the fact that the effect of vdW interactions should be of much shorter range than the cooperative effect of H-bonds. In fact, coooperativity requires chains of hydrogen bonds, whose dipolar interactions (including a possible density polarization) strengthen one another, which can only happen for longer polypeptide chains. The effect of a charge near the C-terminus, here represented by the Li$^+$ ion (green and bordeaux curves, for PBE+vdW and PBE respectively), but verified to be the same for the addition of a positive point charge, is that the helix is significantly stabilized, i.e. it reaches the periodic limit for much shorter n (n ≈11).

From the static picture presented above, it is clear that both the charge and the H-bond network (besides vdW, of course) have a stabilizing effect on the helical structure. However, it is still unclear how both the charge and the connecting H-bonds of the LysH^{+} termination of the Ac-Ala$_{15}$-LysH^{+} affect its *dynamic* helical stability. To investigate this further, simulations with PBE+vdW at 500K were performed for two modified versions of the molecule: one substituting the LysH^{+} amino acid for an Ala and another maintaining the Lys but taking out one proton (with the charge) from the NH$_3$ group.

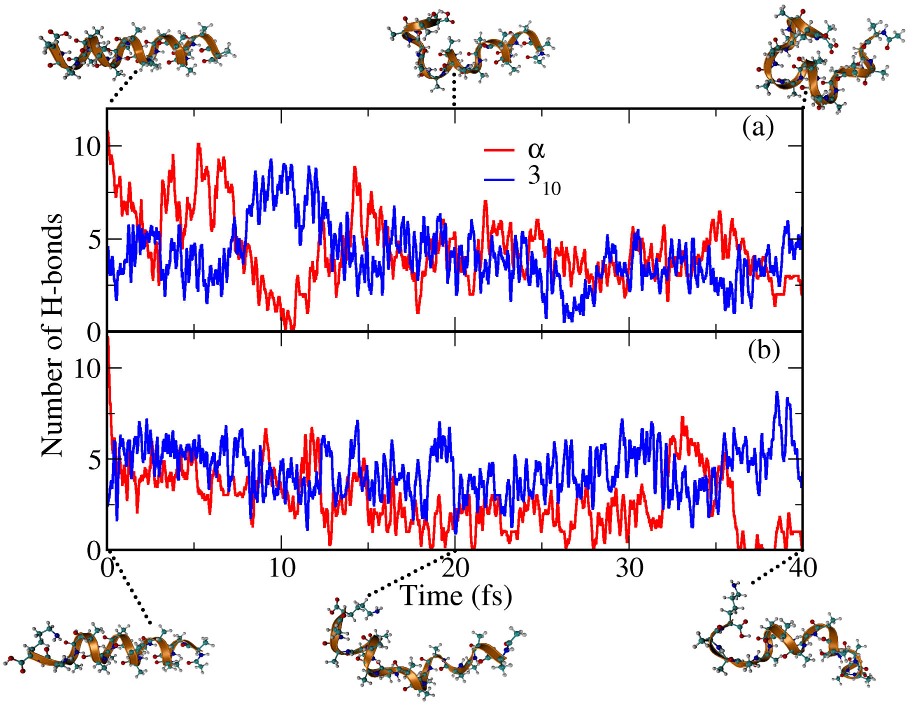


Figure 10.10: MD simulations of: (a) Ac-Ala$_{16}$ at 500K, PBE+vdW; (b) Ac-Ala$_{15}$-Lys at 500K, PBE+vdW. Panels show the number of α-helical H-bonds (red) or 3_{10}-helical H-bonds (blue) versus time of simulation.

The results from such simulations, in the form of number of H-bonds versus time and snapshots of the conformers at various times is shown in Figure 10.10. In *both* simulations, the helical H-bond pattern of the molecule becomes unstable after a few tens of picoseconds. This points clearly to the fact that a synergy between termination H-bonds, the charge, and vdW interactions rule the dynamical stability of the helix, making it an intricate phenomenon. In this case, only the H-bonds formed by the alanine residues are not enough to stabilize the structure at 500K.

10.3 Folding attempts

So far, only unfolding simulations were computed, starting from the α-helical structure of Ac-Ala$_{15}$-LysH^{+}. The inverse path, i.e. *folding*, is a much greater challenge and is shown here only as an outlook. Here only straight AIMD will be presented, but efforts are being made to connect such simulations to kinetic theories, from which great enhancements on the conformational space sampling efficiency are expected

The challenging character of these simulations is twofold (at least): (i) it is unclear what should be the starting structure for the folding attempt; and (ii) in folding, even at high temperatures, the molecule might become trapped in free energy minima for a “long” simulation time. A crude estimate of how long this process takes and how it happens can be obtained from force field simulations starting from a fully extended structure. Although the starting point is not ideal, it is an uniquely defined completely unfolded conformation, and therefore a reliable model. At 300K in the OPLS-AA force field [6], this process takes about 80ns. At 500K, this time is drastically diminished, but still takes between 2 and 3ns. For fully *ab initio* calculations this is still a time scale that is not reachable. However, current force fields are mainly parametrized for the folded structures of proteins and polypeptides, so that the folding path predicted by them might be biased towards the folded state. Ideally, though, in order to study the kinetics of folding, an unbiased theory would be desirable. In this respect, it is interesting to investigate how far DFT alone is able to go, and which are the relevant conformations appearing at short time scales in this folding process.

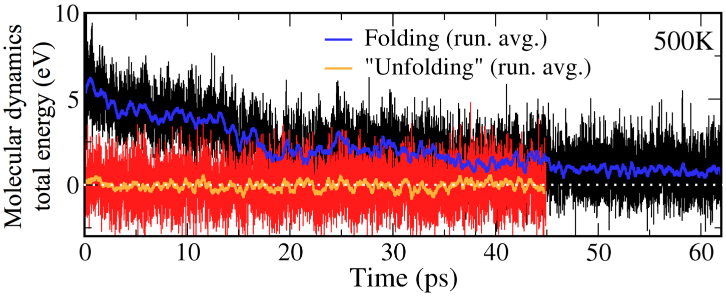
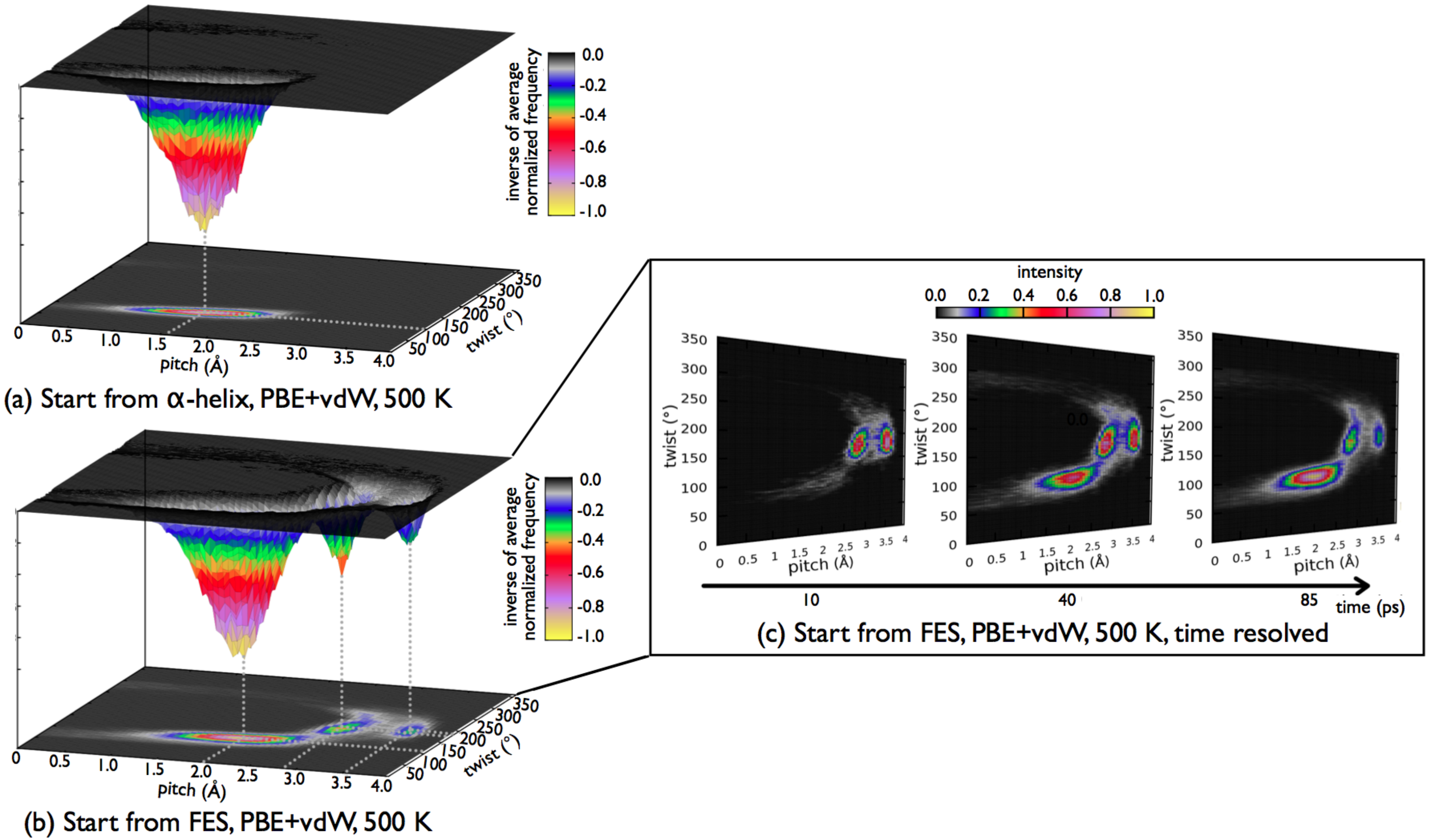


Figure 10.11: Total energies from thermostated MD runs at 500K, with PBE+vdW. Red curve corresponds to a run starting from the α-helical structure and remaining in it. Black curve corresponds to a run starting from the fully extended structure and folding. Orange and blue curves are the respective running averages.

The longest simulation presented here consists of almost 85 ps of AIMD at 500K (maintained with a Nosé-Hoover thermostat), starting from the fully extended structure. The molecular dynamics total energy of one of these simulations with respect to time is plotted in Figure 10.11 and compared to the total energy of the AIMD simulation starting from (and remaining in) the folded, α-helical molecule at 500K. One can see that the "folding" simulation is an out-of-equilibrium situation, where the energy is decaying until it eventually finds a (meta-)stable state. Still at $\sim$70ps the total energy of the "unfolded structure" is around 1eV higher than the folded one, which attests the impossibility of obtaining a fully folded structure within this time of simulation. Other two "folding" runs were attempted (that will be further discussed shortly), starting from the same geometry but with different initial velocities. All exhibit a similar behavior of the total energy, although following slightly different paths (as expected).

[6] The version of the TINKER program used for this simulation had, as a default thermostat, the Berendsen (velocity rescaling) thermostat.



(a) Start from α-helix, PBE+vdW, 500 K

(b) Start from FES, PBE+vdW, 500 K

(c) Start from FES, PBE+vdW, 500 K, time resolved

Figure 10.12: Probability “landscapes” in the pitch-twist cylindrical coordinates: (a) Folded molecule at 500K, PBE+vdW, averaged over 40ps; (b) AIMD starting from the fully extended structure, averaged over 85ps; (c) time resolved data shown in (b).

By looking at the pitch-twist coordinates of the folding simulation, it is possible to obtain a rough idea of the energy landscape explored. In Figure 10.12(a), the conformational space explored by the folded α-helical structure at 500K is shown as a reference. The 2D projection seen in Figure 10.12(a) is exactly the same shown in Figure 10.7 for PBE+vdW at 500K, i.e. the average, through all the simulation time, of all pitch-twist coordinates corresponding to the 12 alanine amino-acids lying in the center of the molecule (disregarding both terminations). The maximum intensity in this case is centered at 1.6Å and 100^o. The 3D rendering offers a better visualization, in which the "z-axis" correspond to the negative of the (normalized) frequency with which the alanine residues adopt the respective pitch and twist values, through the whole trajectory. The same plot is shown in Figure 10.12(b) for 85 ps of a folding run starting from the fully extended structure at 500K, with PBE+vdW. Two regions of high probability at larger lengths can be seen. The one at 3.5 Å and 180^o corresponds to the fully extended structures range (as well as β-sheets), while the one at 2.8 Å and 175^o corresponds to 2_7 and turn conformations. The highest probability, in the 85ps average, is centered at 2.0 Å and 105^o, which is still too extended (if compared to the α-helical) minimum, but is already in a region where some helical content is present in the polypeptide. This region actually corresponds to a meta-stable state, at least in this short simulation time, that exhibits some (even if small) helical content at the terminations, but has a turn effectively folding the molecule in two, as will be detailed further down. This region is not explored in the first picoseconds of the simulation, though, and this becomes clear in Figure 10.12(c) where cuts of the folding run showing the time evolution of the 2D projection of the pitch-twist plots are shown.

It is interesting to note that in Ref. [394], Levy and *et al.* performed a study on solvent effects on folding kinetics of the Ala$_{12}$ molecule, using a force-field. Their gas-phase simulation finds a pronounced and narrow free-energy minimum in the α-helical region and a broad "multi"-minimum, higher in energy, around the turns and β-sheets regions in coordinate space, which corresponds to the random-coil ensemble. While here the statistics are not enough to recover a free energy landscape, the situation looks similar. In their case, the inclusion of the solvent stabilized the random-coil minima with respect to the α-helical one.

Based on the first-principles folding simulations performed in this work, it is possible to make a statement about the initial conformations adopted during the first steps of the kinetics of the molecule starting from the fully extended structure. In the three PBE+vdW folding runs at 500K (with different starting velocities for the atoms), a characteristic structure is formed, involving a turn in the region encompassing the fifth to the eighth alanine (starting to count from the Ace termination). An example of such a "turn" configuration is shown in Figure 10.13 with the residues usually taking part on this particular structure formation in solid colors. These conformations form very fast in the simulations. Representative snapshots of the three folding simulations are shown in Figure 10.14.

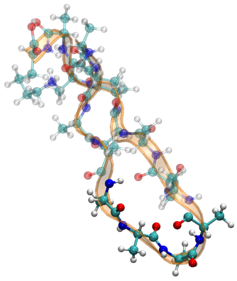


Figure 10.13: Example of a geometry with the characteristic turn (in solid colors) appearing in the AIMD folding simulations.

While it is not possible to claim that this "turn" state is an intermediate of a folding process based on so few statistics, even very short proteins have been shown to fold via such states [77]. Remarkably,

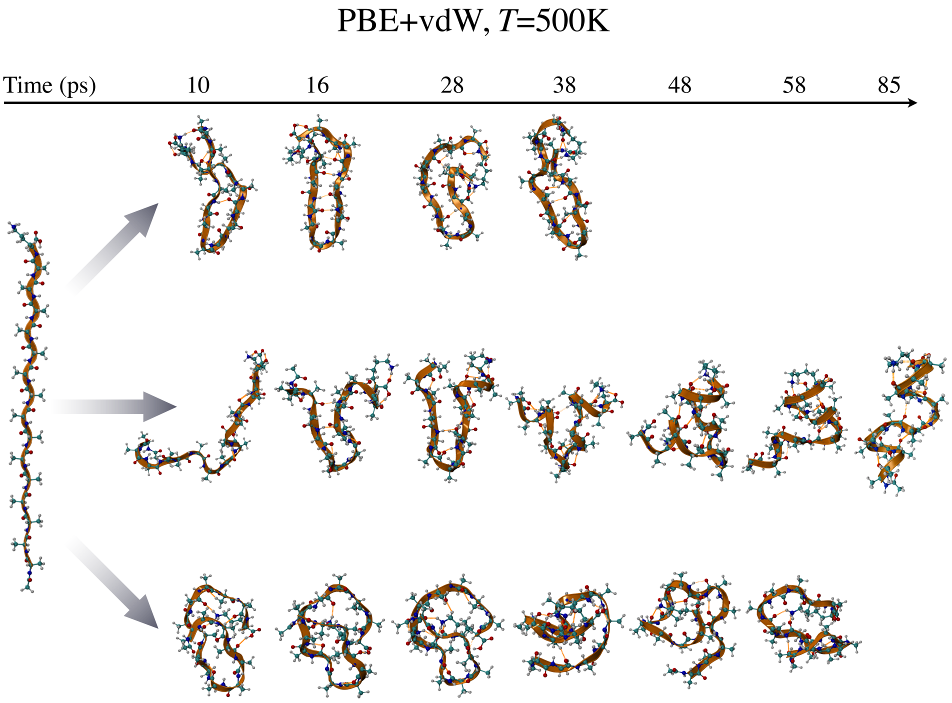


Figure 10.14: Snapshots of representative conformations appearing in three different folding runs at 500K, with PBE+vdW.

a PBE simulation without inclusion of vdW interactions did not form this “turn” conformation, indicating again that the free energy landscape is strongly altered.

10.4 Summary

In conclusion, direct *ab initio* molecular dynamics simulations reveal that vdW interactions explain the remarkable stability of Ac-Ala$_{15}$-LysH^{+} *in vacuo* up to high temperatures. DFT-PBE simulations were seen to render the molecule too unstable, rapidly unfolding already at 700K, in disagreement with experiments. The inclusion of vdW interactions dramatically changes the conformational landscape explored, favoring the exploration of more compact helices, exhibiting a mostly α-helical character at finite temperatures. PBE favors a mostly 3_{10} helix. A synergy between the charge, the connecting H-bonds of the Lys termination, and vdW was essential to explain the dynamical high-temperature stability of the helix. Although first-principles *folding* simulations for even this small(ish) peptide at high temperatures are still unreachable, it is possible to learn something about the probable beginning of the folding path and the landscape explored.

Part III

Steps toward benchmarks from high-level electronic-structure methods

Chapter 11

Development of NAO basis sets for explicitly correlated calculations

Benchmarking the DFT results against methods that take non-local correlation explicitly into account should give a measure for the reliability of the DFT-PBE+vdW data, since these methods should be, in principle, more accurate for the long-range correlation. Although it would be ideal to use the "gold-standard" of these methods (coupled cluster plus singles, doubles and perturbative triples (CCSD(T)) [86]), its formal $O(N^7)$ scaling (where N is, for example, the number of basis set functions of the system) makes this method prohibitively expensive for large molecules. Feasible explicitly correlated methods for molecules of ≈100 atoms, as studied in this work, are for example EX+cRPA and MP2, which are perturbative methods.

The goal here is to obtain converged relative energies between different polypeptide conformations. Even though MP2 and EX+cRPA are not the ideal benchmark, obtaining converged energy hierarchies for them (preferably with a small basis set size) is treated here as a first step towards obtaining accurate benchmark data. As discussed in Chapter 3, one of the drawbacks is that these methods suffer from a large basis set superposition error, even for very large basis sets. In order to deal with this problem, efficient and *small* numeric atom-centered basis sets for each species, to be used in addition to the standard *tiers*, explained in Section 5 and Appendix B, will be developed in this chapter. These new basis sets reduce substantially the BSSE, and the non-local correlation effects converge systematically with basis set size. The development of these basis sets, the reasons for this development, and benchmarks for their performance are treated in this chapter.

11.1 Statement of the problem

As has been seen in Section 5.2, for the water dimer, converging energy differences within explicitly correlated methods, i.e. methods that treat non-local correlation explicitly, is not a trivial task. Special basis functions that systematically converge the non-local correlation have to be used, and this was realized early on, when Almlof and Taylor [395] proposed special atomic natural orbitals (ANO) for this purpose. Since then, much work has been done to develop especially gaussian based [279, 396–399] localized basis sets that allow to converge energies within such methods. On the other hand, concerning numeric atom-centered orbitals (NAO) not much (if anything) can be found in the literature in the lines of basis sets devised for these explicitly correlated methods.

Even with the gaussian-based basis sets the energy convergence for these methods is worse when compared to DFT or HF. The reason, already discussed in Section 5.2, is related to the fact that the correlation energy evaluation typically depends on sums over the unoccupied orbitals. These states are much more difficult to converge than the occupied ones, because ideally they would have to be summed up to the continuum. In practice, however, it would be enough to describe (converge) the part of the Hilbert space corresponding to the relevant unoccupied orbitals. The problem is that with a finite basis set it is impossible to sum these states up to infinity, so that if these basis sets are not well devised for the correlation, BSSE will be present even with large basis sets. Moreover, the explicit consideration of the unoccupied orbital space of each atom implies that the fragments with a smaller basis set will span a volume of the unoccupied space that is forcefully different from that spanned by the full complex with more basis sets.

A very successful and widely employed remedy for BSSE, already discussed in Section 5.2 is the counterpoise correction of Boys and Bernardi [268]. As written in Eq. 5.18, in this scheme each fragment of a molecular complex is calculated in the presence of the basis sets of the other fragments, in the places where their atoms would be (“ghost” atoms), and energy differences are then calculated. It does, therefore, correct mostly the *inter*molecular BSSE, relying on the fact that the *intra*molecular one will cancel out when calculating energy differences.

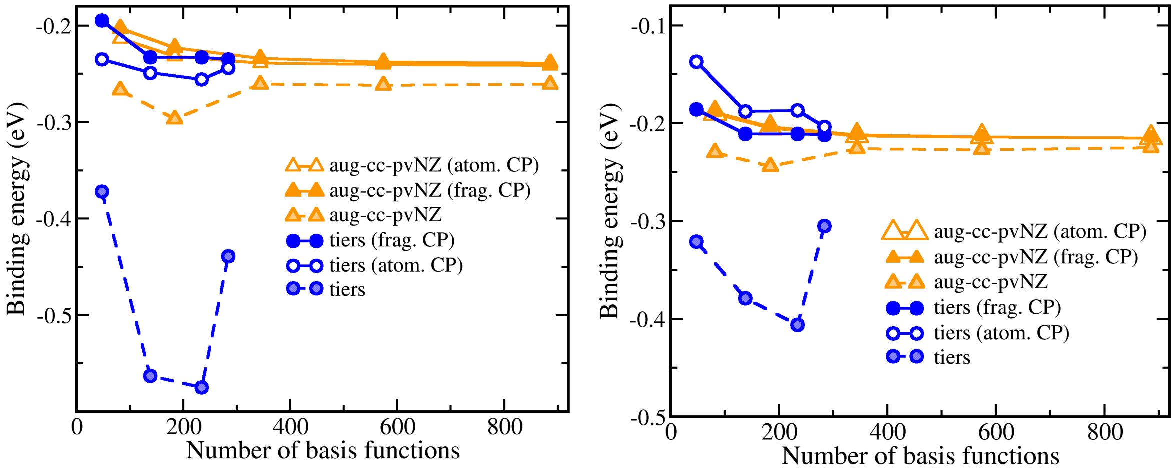
This procedure is sufficient in most cases but presents two problems for the goals of this work: (i) there is an ambiguity related to the definition of fragments when one wants to study different conformations of the same molecule; (ii) it does not correct the *intra*molecular BSSE, which might be different for different molecules, and has been shown to be very relevant for conformations of alanine-based polypeptides [270] (as well as other systems [274, 400, 401]). An obvious way out is to take this scheme to the limit and define each atom as a fragment. This is commonly called atomization counterpoise correction (Refs. [271, 282] and references therein). The correction that is used in the following is the same as the correction reported in Ref. [282] and a special case of the one reported in Ref. [271]. The correction is obtained by the following expression:

$$\Delta_{ac} = \sum_{a}^{N_{atoms}} [E^a(a) - E^a(sys)] \tag{11.1}$$

where $E^a(a)$ is the total energy of atom a calculated only with its own basis functions, $E^a(sys)$ the total energy of that atom calculated with the basis sets of the whole system via “ghost” atoms, with the sum running over all atoms in the system. In cases where one is interested in binding energies of complexes, this correction is applied to the total energy of the full complex and that of its fragments. After that, the binding energy is obtained according to Eq. 5.16. This correction is the only rigorous way of calculating *a posteriori* corrections for the energy differences between different conformations of the same molecule, since there is no ambiguity in choosing different fragments. The drawbacks (and the reasons why this correction is not very popular) are that the calculation is much more expensive than the fragmentation scheme (it needs one full calculation for each atom of the molecule in the presence of all basis functions of the full system) and possible inaccuracies in describing the core electrons of the atoms (in an all-electron picture) are not necessarily canceled out, as will be illustrated in the next Section, for the water dimer.

11.1.1 Water dimer

The atomization counterpoise correction has already been applied to the water dimer in Ref. [282], using gaussian basis sets [1]. The only explicitly correlated method reported in that reference is MP2. [2] Here the objective is first to test how the NAO basis sets that are standard in FHI-aims, perform for explicitly correlated methods, specifically the EX+cRPA+SE method. These NAO basis sets are the ones detailed in Table B.1, Appendix B.



(a) EX+cRPA+SE@PBE

(b) MP2

Figure 11.1: Convergence of the binding energy of the water dimer (frozen at the relaxed MP2 geometry) with different types of basis functions. Full lines are corrected for BSSE and dashed lines are not. Orange triangles correspond to the aug-cc-pVNZ, $N = D, T, Q, 5, 6$; Blue circles correspond to tier$N$, $N = 1, 2, 3, 4$. Values corrected with the fragmentation BSSE correction (full symbols) and with the atomization BSSE correction (empty symbols) are shown. (a) EX+cRPA+SE@PBE values; (b) MP2 values. For MP2, the aug-cc-pVNZ fragmentation and atomization correction curves are on top of one another in this scale, so that one is not visible.

In Figure 11.1(a) the binding energy of the water dimer, calculated with EX+cRPA+SE at PBE self-consistency (EX+cRPA+SE@PBE), is plotted. The geometries of the molecules (including the fragments) are frozen at the relaxed MP2 geometry, the same as can be found in the S22 data base [178] (see Appendix E). The aug-cc-pVNZ basis sets curves, including the fragment counterpoise correction, the atomization counterpoise correction, and no correction, are shown in orange. The remaining curves in the plot, in blue, are calculated with the standard basis sets distributed with FHI-aims (tiers$N$, $N = 1, 2, 3, 4$, see Appendix B). The first thing to notice is that the uncorrected values of the binding energy for the NAO basis sets (blue filled circles connected by a dashed line in Figure 11.1) are very bad and do not show any convergence pattern with basis set size. When applying a standard fragmentation counterpoise correction (Eq. 5.17) the convergence behavior is regained to a remarkable accurate extent (blue filled circles connected by a full line in Figure 11.1 (a)), and the value agrees with the one obtained when applying the same correction to gaussian basis sets. It is worth noting that the the fragmentation BSSE behavior with the $tiers$ basis sets has been verified over an extensive series of test cases, including the S22 set, Au dimer [1], etc. Finally, when applying the atomization counterpoise correction for the NAO basis sets (Eq. 11.1) although the values get much closer to the expected result (blue empty circles connected by a full line in Figure 11.1), they don't show a smooth convergence. For the gaussian basis

[1] In Ref. [282], it is not explicitly said if the calculations were performed including or not spin polarization.

[2] The MP2 results for gaussian basis sets and fragmentation BSSE correction in this work (see Fig. 5.4) agree with the literature.

sets, on the other hand, both the atomization and fragmentation BSSE correction converge smoothly and agree at convergence. In Figure 11.1(b) the same curves discussed above are shown for the binding energy of the water dimer using MP2. Essentially the same conclusions as for EX+cRPA+SE@PBE can be drawn. Therefore, with the standard NAO basis sets from FHI-aims, problems seem to remain: it is not possible to converge outrightly the binding energies, and the atomization BSSE correction that would be needed to calculate relative energies between different conformations does not converge smoothly.

It is important to note that all calculations presented in this work that include the atomization counterpoise correction are done without considering spin polarization for the single atoms. Atoms with degenerate valence states were allowed fractional occupation numbers and care was taken to check that all atoms of the same type had the same fractional occupation in a calculation. While fractional occupations are not a problem for DFT, it is not usual to use it with quantum-chemistry (HF, MP2, CC) methods. The situation here, though, is not one where physical values for the total energy of each atom are sought, but one where these atoms' energies should mimic the situation they encounter inside the molecule. Let us take the example of the water dimer calculated with MP2, in order to put numbers to these arguments. If one takes the hydrogen atom, for example, and performs a spin polarized calculation with MP2, the correlation energy will be (physically correct) zero. If the HF calculation (where MP2 is based on) is converged, this would lead to absolutely no contribution from this atom for the BSSE correction of Eq. 11.1. This correction is needed though, to approach the correct limit of the binding energy. By calculating the binding energy of the water dimer using MP2 and both the fragmentation and the atomization scheme with Gaussian basis sets, it is observed that when performing spin *un*polarized calculations the atomization scheme with the aug-cc-pV6Z [3] basis set gives $E_b = -0.215$eV which agrees, to the sub-meV level, with the value obtained by the fragmentation scheme for the same basis set (see Figure 11.1(b), where the atomization corrected and fragmented corrected curves are on top of one another, in that scale). It is important to stress, though, that while in DFT (and also, to some extent, in HF) the fractional occupation numbers can be regarded as an average of different possible ground state determinants, for a physically meaningful perturbation theory with fractional occupation numbers, the formulation would have to be completely rewritten (see Ref. [402]). The MP2 correlation energy and RPA polarizability (χ^0) with fractional occupation numbers used in this work are the following:

$$E^{MP2}_{corr,frac} = \frac{1}{4}\sum_{ijab} f_i f_j \frac{|\langle ij||ab\rangle|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b} \tag{11.2}$$

$$\chi^0_{frac}(\vec{r},\vec{r}',i\omega) = \sum_j^{\text{occ}} \sum_a^{\text{unocc}} f_j \frac{\phi_j^*(\vec{r})\phi_a(\vec{r})\phi_a^*(\vec{r}')\phi_j(\vec{r}')}{i\omega - \epsilon_a + \epsilon_j} + \text{c.c.}, \tag{11.3}$$

where i, j denote occupied orbitals, a, b unnocupied, and f_i, f_j the respective occupation numbers.

Also, all calculations presented here are performed with the *tight* settings of the FHI-aims code [1] for accuracy and grids, with only one modification to the onset of the cutoff potential, that will be discussed in detail in Section 11.3.2.

11.1.2 S22 data set

The water dimer is only a single case. To broaden the observations made for that single system, the two different BSSE correction schemes were compared for the S22 set of molecular complexes proposed in

[3] Calculated with FHI-aims.

Ref. [178] and shown explicitly in Appendix E. In this set, non-covalent interactions like H-bonds and van der Waals are represented. The reference values for this set of molecules using the EX+cRPA+SE@PBE method are taken from Ref. [190]. These values were obtained with the $tier4$ basis plus diffuse gaussian functions, including fragmentation counterpoise correction. The quantities used to compare the data are the Mean Absolute Error (MAE) and the Mean Absolute Relative Error (MARE) defined by:

$$\text{MAE} = \frac{1}{M}\sum_{i}^{M} |E_{\text{calc},i} - E_{\text{ref},i}| \tag{11.4}$$

$$\text{MARE} = \frac{1}{M}\sum_{i}^{M} \left|\frac{E_{\text{calc},i} - E_{\text{ref},i}}{E_{\text{ref},i}}\right| \tag{11.5}$$

where E_{calc} is the calculated value, E_{ref} the reference value and the sum runs over all components of the set. The MARE is important because it tells which percentage of the total value the error represents. For example, a MAE of 10meV might not be so important if one is talking about binding energies of ≈1eV, but very important if the binding energies are of the order of 20meV.

In Table 11.1 the behavior of $tier2$ and $tier4$ standard basis sets are summarized by computing the MAE and the MARE for the binding energies of S22 data set. Three values are shown for each basis set: the one obtained when no counterpoise correction (CP) is performed, the one obtained when applying the *atomization* CP correction, and the one obtained when applying the fragmentation CP correction.

basis set	No CP	Atom. CP	Frag. CP
$tier2$	420 (171%)	15 (10%)	27 (15%)
$tier4$	440 (249%)	13 (9%)	5 (2%)

Table 11.1: Mean Absolute Error (MAE) in meV for the S22 data set, using the RPA+SE method on top of PBE. No CP, Atom. CP, and Frag. CP stand for no correction, atomization counterpoise correction, and fragmentation counterpoise correction respectively. In parenthesis the Mean Absolute Relative Error (MARE).

From this table, it becomes again clear that applying no correction produces meaningless values for the binding energies. Moreover although the MAE and the MARE at the $tier2$ level are smaller for the *atomization* CP correction than for the fragmentation CP correction, they do not decrease substantially when increasing the basis set to $tier4$, and become larger than the fragmentation CP correction at this level. Furthermore, performing the atomization CP correction for $tier2$ and $tier4$ still produces a maximum absolute relative error of 29% (conformer 10, benzene-CH_4) and 40% (conformer 11, stacked benzene dimer) respectively. Meanwhile, the fragmentation scheme produces a maximum absolute relative error of 44% for $tier2$ (conformer 11, stacked benzene dimer), but only of 6% for $tier4$ (conformer 14, stacked indole-benzene). The errors in binding energy for each conformer, given by:

$$\Delta E_i = E_{\text{calc},i} - E_{\text{ref},i}, \tag{11.6}$$

where in this case E_i refers to the binding energies of the complexes, are plotted in Figure 11.2.

Based on these test cases, the conclusions drawn from the H_2O-dimer remains: the standard $tiers$ basis sets are insufficient for straight, non CP corrected calculations, and are limited for atomization CP corrections, since the convergence behavior is not optimal (although the improvement is substantial over the non-corrected values).

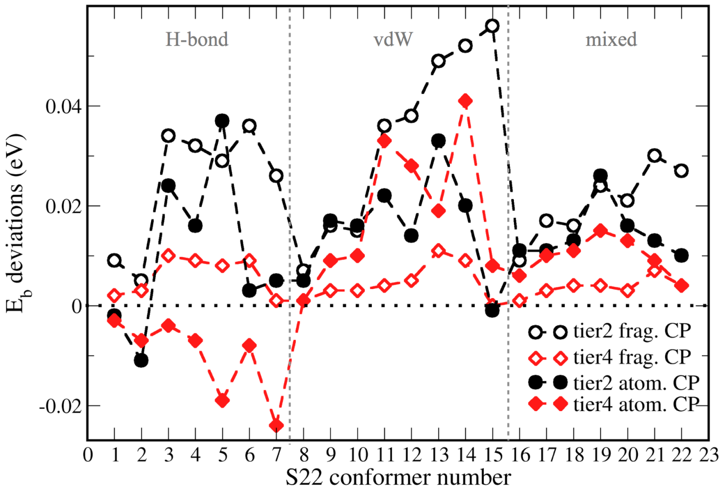


Figure 11.2: Errors with respect to EX+cRPA+SE@PBE converged values from Ref. [190]. Filled symbols correspond to values including the fragmentation BSSE correction (Eq. 5.17), and open symbols correspond to values including atomization BSSE correction (Eq. 11.1).

11.2 Core-correlation basis functions

The pronounced errors seen in the previous section for the standard NAO basis set of FHI-aims can be understood by an analysis of the non-local correlation. Taking, for example, the MP2 expression for the correlation (last term of Eq. 3.33) it is possible to take apart the sum and calculate the contribution of each single occupied molecular orbital interacting with all others, by summing up all but one of the occupied orbitals indexes (e.g., perform the sum on j, a, and b and get the MP2 correlation for each occupied orbital i).

Considering the monomer of the water dimer, calculate the contribution to the MP2 correlation for each occupied state when including the basis set of the other conformer (counterpoise correction) and when including only the basis set of the monomer itself (standard). If one does this exercise with $tier4$, one sees that the main difference lays in the term coming from the first (core) orbital interacting with all others. The difference for this term, between the standard calculation and the counterpoise corrected one, is of 65 meV (for only one monomer). The second orbital already presents a difference of only 6 meV, the third 2 meV and the others 1 meV or less. On the other hand, when using the correlation consistent Dunning basis sets, at the aug-cc-pVQZ level, *all* terms present differences of 1meV or less, when comparing the counterpoise corrected calculation with the standard one (again for a single monomer). Thus, the errors for the NAO $tiers$ seem to be coming mainly from the correlated core. Indeed, the minimal basis obtained from the DFT(LDA) free atoms describe the core of the DFT atoms almost exactly, even in a bonded situation, but according to the data presented so far, this is not the case when explicit non-local correlation is considered. The aug-cc-pVNZ basis sets, on the other hand, have been optimized for CI total energies of atoms [279].

When performing the fragmentation correction for BSSE, any remaining errors due to these core orbitals mostly cancel when calculating energy differences. The atomization BSSE correction is more prone to errors, since $N + 1$ calculations (N =number of atoms of the complex), are needed to compute the total energy correction. Small errors can add up substantially in this procedure. This may explain the lack of convergence observed when performing the atomization correction with increasing NAO basis set size.

The idea of this work is to find a specific and finite set of basis functions that can be added to the standard NAO $tiers$, so that the correlated core of the atoms is correctly described. The development of the correlation consistent gaussian basis sets follows a known prescription in quantum chemistry: given a primitive set of s and p functions, for example (for atoms of the first row of the periodic table), the so-called polarization functions are added systematically, i.e. first d functions, then f functions, then g functions, etc., with their exponents optimized at each step. This procedure generates a systematic but slow convergence of the correlation energy with basis set size. Here the goal is to obtain a fast convergence with basis set size, in order to work with small basis sets. Moreover, in the optimization procedure, as little as possible human bias is desired.

11.2.1 Optimization procedure

The procedure to optimize the basis sets used in this work is similar to the standard optimization procedure for the basis sets distributed with the FHI-aims code package, that was briefly described in Section 5, and can be found in detail in Ref. [1]. The optimization here, instead of starting from a minimal basis for the atoms and optimizing LDA total energies for dimers of each species, starts from the $tier2$ default basis set of each element (see Appendix B on these basis sets) and optimizes MP2 total energies for the dimers. A pool of possible basis functions is defined, including (i) radial functions of doubly positive charged free ions and (ii) hydrogen-like radial functions for a potential Z/r with Z in range 0.1 $\leq Z \leq$ 60, where Z is allowed to be non-integer [4]. These radial functions are obtained via Eq. 5.4. The angular momenta were considered from l=0 (s) to l=5 (h). As the optimization target, the MP2 total energy of non-spin polarized symmetric dimers of each element is minimized, taking the average at $N_d = 5$ different interatomic distances (detailed below) for the elements studied here. MP2 is not a variational method, in the sense that the expectation value of the energy in this method does not obey a variational principle. This could, in principle, lead to problems when trying to minimize this energy. Nevertheless, from all calculations performed in this work, it was never observed that the MP2 energy was not "variational" with respect to the addition of basis functions. The energy is always lowered by addition of extra functions and this improvement gets smaller and smaller (i.e. there is convergence) with the basis set size. At any rate, we are here interested in finding radial functions which capture the source of potential energy fluctuations as effectively as possible, which justifies our procedure.

At each step, the quantity Δ^{MP2}_{basis}, defined by:

$$\Delta^{MP2}_{basis} = \frac{1}{N_d}\sum_{i=1}^{N_d}[E^{MP2}_{basis}(d_i) - E^{MP2}_{basis-1}(d_i)] \tag{11.7}$$

is calculated for each function of the pool. In Eq.11.7 $E^{MP2}_{basis-1}(d_i)$ is the MP2 total energy for the dimer calculated with all basis chosen up to the previous step and $E^{MP2}_{basis}(d_i)$ is the MP2 total energy calculated with all previously chosen basis plus an additional candidate function to be evaluated in the current step. From the pool of possible basis functions, the procedure automatically chooses the one that causes the largest energy improvement for the system. In this work we are interested in the light elements C, N, H, and O. The distances used were, respectively, $d_i = 0.5, 0.7, 1.0, 1.5, 2.5$ Å for H, $d_i = 1.0, 1.25, 1.5, 2.0, 3.0$ Å for C, $d_i = 1.0, 1.1, 1.5, 2.0, 3.0$ Å for N, and $d_i = 1.0, 1.208, 1.5, 2.0, 3.0$ Å for O.

The goal, with this procedure, is not to achieve absolute total energy convergence or even a semblance of convergence. We aim to find a limited set of basis functions that can capture the main error, so that the

[4] In order to optimize the shape of the radial functions, Z can be treated as a continuous parameter, which allows more flexibility on the functions.

rest can be eliminated more effectively by atomization BSSE corrections. In Figure **??** the MP2 energy changes per atom and averaged over all distances are plotted, for each new basis function (detailed in the next paragraph), and taking as the reference the MP2 total energy calculated with the standard $tier2$ basis set.

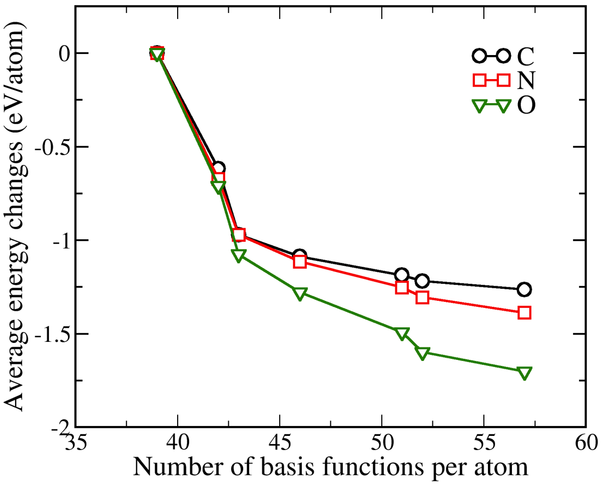
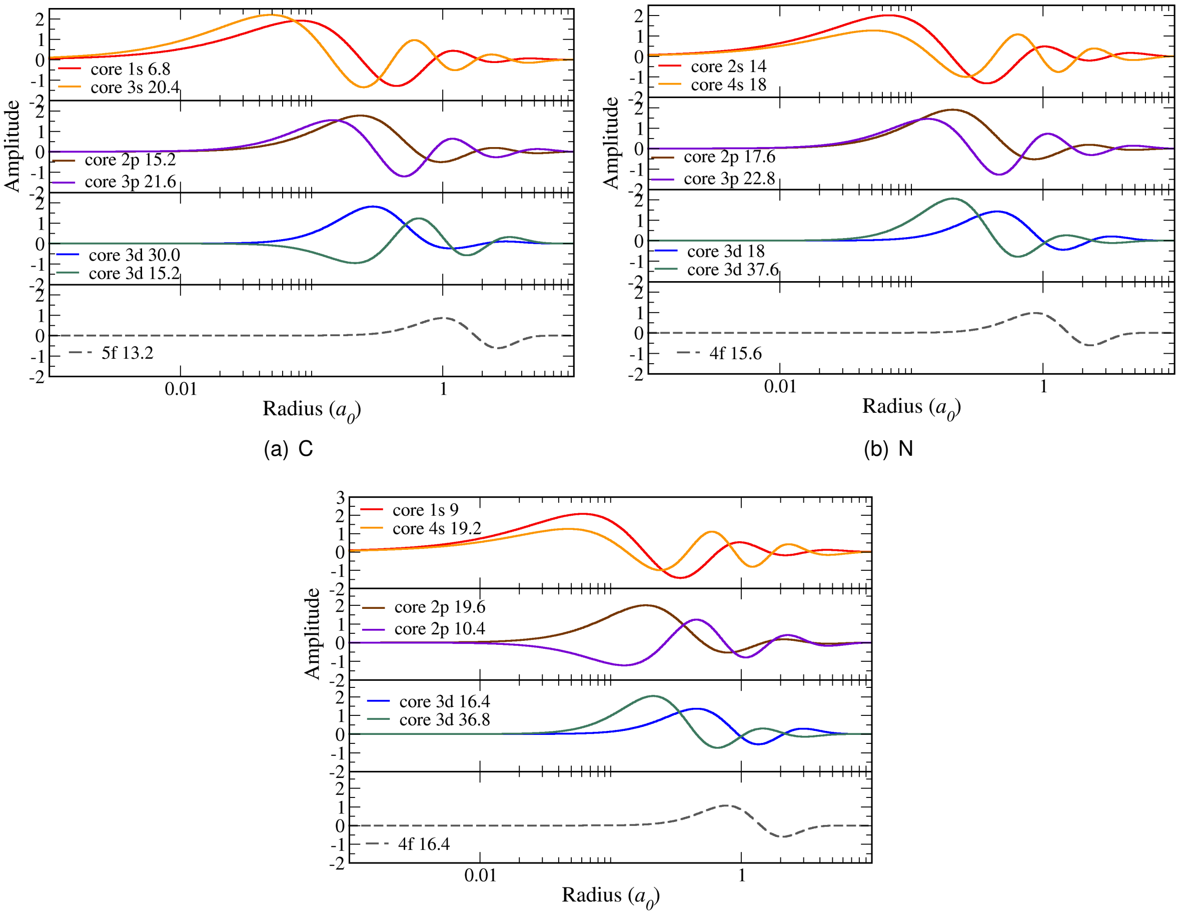


Figure 11.3: Energy changes, averaged over all interatomic distances, for adding each chosen basis function, with the reference taken as the MP2 energy per atom calculated with the standard $tier2$ basis set.

For hydrogen, when starting this procedure from $tier2$, only basis functions very similar to the ones that were already present in the standard $tier3$ for this species were produced. When starting from $tier3$ the highest improvement for an extra basis-function was of 4 meV. Therefore, no additional basis functions are proposed for this element. This is not surprising, since H has only one electron and in principle should not have core-correlation functions.

For C, N, and O the largest improvements were obtained by hydrogen-like radial functions with l=0 (s), and l=1 (p), listed in Table 11.2. These were followed by another set of four hydrogen-like radial functions, one with p, one d, and one with f character, automatically from the script. In Figure 11.4 the respective radial functions after on-site orthonormalization (in order to remove linear dependences with all other functions, including the standard $tier2$) for the basis sets are plotted for each element. The f functions present the highest peak at a distance of ≈1 a_0 (Bohr radius) for all elements, and the first node at ≈1.8 a_0, while all the other functions have the highest peak and first node at shorter radii. The f functions are, thus, spatially more delocalized than the others. This observation prompted a reassessment of the functions found in the script. It was observed that when the f function was chosen, the next function inducing almost the same energy improvement was an s function of core character. This s function was then "promoted", in the sense that it was forced to be chosen instead of the f function and the optimization procedure was re-started from that point, yielding another d function for all elements. The f function was kept for *a posteriori* tests, but as will be seen later on, it indeed does not bring any appreciable improvements, if used in conjunction with the other "core" basis functions, when calculating energy differences.

The chosen basis functions, are detailed in Table 11.2. The name and number of the functions mean the same as in the standard $tiers$, explained in detail in Appendix B. From now on "$+sp$" will be used to name the addition of only the first s and p functions of each element (shown in Table 11.2) to the standard basis sets. In the same fashion, "$+spspdd$" will be used to name the addition of all basis sets shown in Table 11.2 for each element.



(a) C

(b) N

(c) O

Figure 11.4: Radial functions of the "core" optimized basis functions after on-site orthonormalization for (a) C, (b) N, and (c) O.

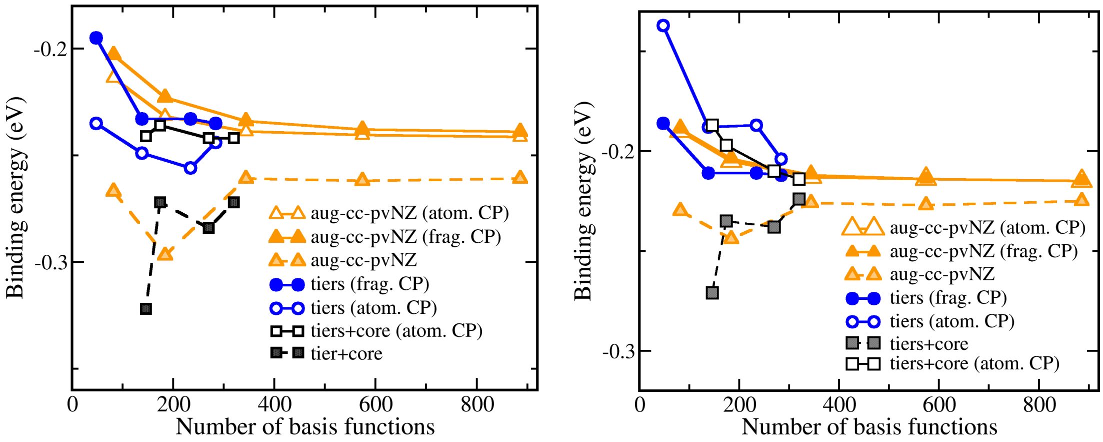
C core functions	N core functions	O core functions
hydro 2 p 15.2	hydro 2 p 17.6	hydro 2 p 19.6
hydro 1 s 6.8	hydro 2 s 14	hydro 1 s 9
hydro 3 p 21.6	hydro 3 p 22.8	hydro 2 p 10.4
hydro 3 d 30	hydro 3 d 18	hydro 3 d 16.4
hydro 3 s 20.4	hydro 4 s 18	hydro 4 s 19.2
hydro 3 d 15.2	hydro 3 d 37.6	hydro 3 d 36.8

Table 11.2: Extra functions of core character chosen for C, N, and O, ordered by the energy improvements given by Eq. 11.7, and seen in Figure 11.3. The name "hydro 2 p 15.2 " denotes a hydrogen-like function (see Eq. 5.4) of the 2p type with an effective Z of 15.2.

11.3 Performance

11.3.1 Water dimer

Returning to the water dimer, already studied in Section 11.1.1, the idea is now to test the newly developed basis sets.



(a) EX+cRPA+SE@PBE

(b) MP2

Figure 11.5: (a) Convergence of the EX+cRPA+SE@PBE binding energy of the water dimer (frozen at the relaxed MP2 geometry) with different types of basis functions. Full lines are corrected for BSSE and dashed lines are not. Orange stars correspond to the aug-cc-pVNZ, $N = D, T, Q, 5, 6$; blue circles correspond to tier$N$, $N = 1, 2, 3, 4$; and black squares correspond to $tier2 + sp \rightarrow tier2 + spspdd \rightarrow tier3 + spspdd \rightarrow tier4 + spspdd$. (b) Same as (a), but for the MP2 binding energy of the water dimer.

In Figure 11.5(a), the convergence of the EX+cRPA+SE binding energy of the water dimer with the new basis sets (black curves) is plotted, together with the other curves already shown in Figure 11.1. In Figure 11.5(b), the same is shown for MP2. Notice the dramatically changed scale, with respect to Figure 11.1, though. The sequence of points for the NAO basis, with increasing basis set size is: $tier2 + sp \rightarrow tier2 + spspdd \rightarrow tier3 + spspdd \rightarrow tier4 + spspdd$. As compared to the uncorrected values for the standard $tiers$, (shown only in Figure 11.1) the uncorrected values for the new basis sets (filled black squares) present a dramatically reduced BSSE, which is now of the known order, or better, than the one obtained with the gaussian aug-cc-pVNZ functions. In this case, adding the discarded f function to the $tier2 + spspdd$ changed 0.5meV the non-corrected binding energy and 1meV the atomization-corrected binding energy.

When applying the atomization CP correction scheme for the EX+cRPA+SE method, the curve including the core functions now looks smooth and converges to the expected value. For the MP2 method, the atomization CP corrected curve [open black squares in Figure 11.5(b)] is similarly improved with respect to the standard NAO ($tiers$), atomization CP corrected, curve [open blue circles in Figure 11.5(b)] . Furthermore, the value at $tier4 + spspdd$ agrees with the aug-cc-pV6Z value. The convergence for smaller basis sets, however, is not quite as rapid if compared to the gaussian basis sets of similar size - unlike in the EX+cRPA+SE case [see same curves in Figure 11.5(a)]. Still, it is of the same order of magnitude as the gaussian basis sets [5]. From now on only atomization CP corrections for the

[5] The author suspects that the reason for the slightly less rapid convergence of $tier2$+core basis sets in MP2, compared to EX+cRPA+SE, is due to the treatment of the oxygen atom reference in the atomization scheme. In RPA, the reference wave function is DFT-PBE. In contrast, the reference in MP2 is straight HF. It is well known that electronic configurational symmetry

EX+cRPA+SE method will be shown.

11.3.2 S22 data set

The natural step now is to apply the new basis sets also to the S22 data set. In Figure 11.6 the errors in the binding energies (Eq. 11.6) of each complex obtained adding the new basis sets on top of the standard $tiers$ ($tier2 + sp, tier2 + spspdd, tier4 + spspdd$) is plotted with respect to EX+cRPA+SE@PBE converged values (Ref. [190]). The full symbols correspond to binding energy values including the atomization CP correction of Eq. 11.1 and the open symbols correspond to binding energy values without this correction. In Table 11.3 the MAE and the MARE for the whole S22 set and the different basis sets are summarized.

basis set	No CP	Atom. CP
$tier2 + sp$	140(57%)	9(7%)
$tier2 + spspdd$	71(32%)	11(8%)
$tier4 + spspdd$	79(34%)	12(6%)

Table 11.3: Mean absolute error (MAE) for the S22 data set, using the EX+cRPA+SE method on top of PBE with values in meV. In parenthesis the mean absolute relative error (MARE).

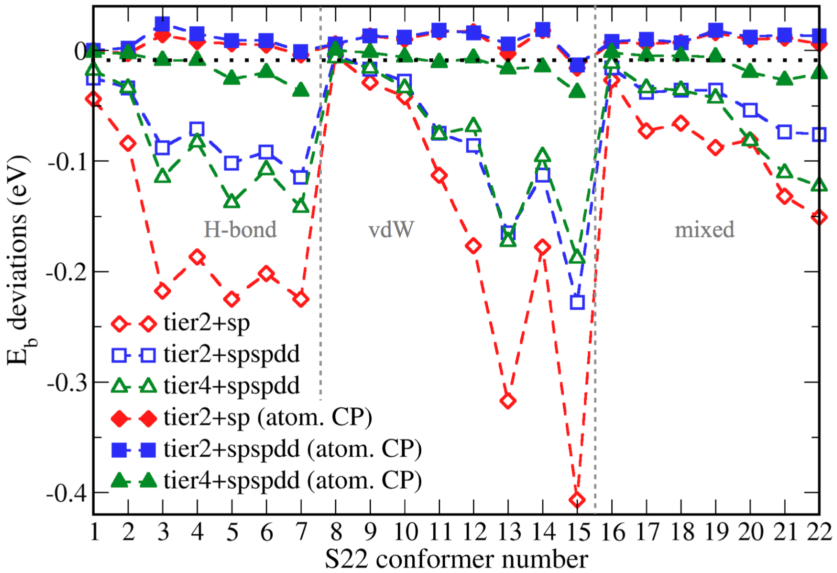


Figure 11.6: Errors with respect to EX+cRPA+SE@PBE converged values from Ref. [190]. Filled symbols correspond to values including the atomization BSSE correction (Eq. 11.1), and open symbols correspond to uncorrected values.

First, focusing on the values *without* CP correction in Table 11.3 and comparing them to the ones reported in Table 11.1, already at the $tier2 + sp$ level the MAE is reduced by a factor of ~ 3. The subsequent set of $spdd$ functions for each element reduces again this error by 2, and going to $tier4$ plus the full extra set of core basis functions has a similar performance. This fact might indicate that these

breaking (splitting the $2p$ orbital) occurs in such methods, but the effect is much larger (1-2 orders of magnitude) for HF than DFT. Most likely, this large change is captured in a more systematic way by the correlation-consistent gaussian basis sets.

core functions work optimally for $tier2$, which is the basis set with which they were optimized [6]. In any case, even at the biggest basis set, the performance is not satisfactory, if no further corrections are applied.

When including the atomization CP correction all errors are dramatically reduced. Interestingly, all basis sets tested show a similar performance. Now, as opposed to what was seen in Section 11.1.2, the maximum absolute relative error when including the atomization CP correction is of 23% (conformer 8, CH_4 dimer) and 13% (conformer 11, stacked benzene dimer) for the $tier2 + spspdd$ and $tier4 + spspdd$ basis sets, respectively. The absolute values for the errors of these conformers are 6 meV ($tier2+spspdd$, conformer 8) and 11meV ($tier4 + spspdd$, conformer 11). This is a much better behavior, very similar to the one obtained for the standard $tiers$ with the fragmentation CP correction. Moreover, the reduction of the MARE to 6% at $tier4 + spspdd$ with atomization CP correction means that now the biggest errors are found for the conformers with the larger binding energies [7].

The agreement with the reference data is still not optimal though. As the MARE for $tier2 + spspdd$ and $tier4 + spspdd$ points out, binding energies calculated in this way (with atomization BSSE correction) can be around 10% off. This is an error that has to be kept in mind. Here the goal is to calculate relative energies between different polypeptide conformations. Therefore, these basis sets and BSSE corrections need to be tested for precisely this situation. This will be done for the alanine dipeptide in the next section, but before a short discussion about the impact of the radial cutoff potential will be given.

Impact of the cutoff potential

The impact of changing the value of r_{onset}, from Eq. 5.5, on the binding energies obtained with MP2 and EX+cRPA was tested. Its value was found to strongly affect the binding energies mainly of dispersion bonded conformers. This effect is depicted in Figure 11.7, only for the EX+cRPA+SE method. In that Figure, the MARE of the dispersion bonded complexes of the S22 data set (numbers $8 - 15$) is shown for three different values of r_{onset}: 4 Å, 5 Å, and 6 Å. Only values calculated with the $tier4 + spspdd$ basis sets (including atomization BSSE correction) are shown. The reference values are taken from Ref. [190].

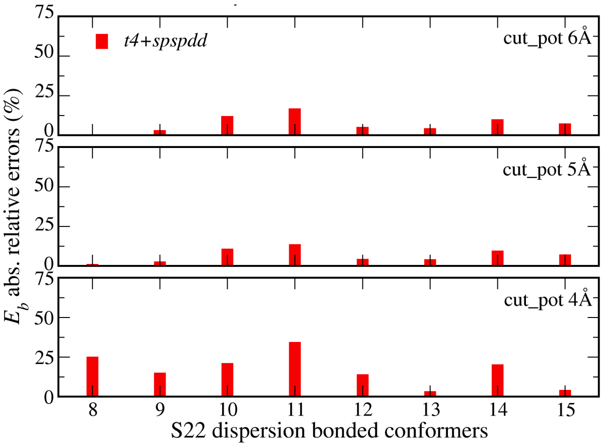


Figure 11.7: MARE of the dispersion bonded complexes for three different onsets of the cutoff potential: 4Å, 5Å, and 6Å, calculated using EX+cRPA+SE@PBE, with the $tier4 + spspdd$ basis sets, and including atomization BSSE correction.

While for LDA/GGA calculations, $r_{onset} = 4$ Å produced converged values, for the dispersion bonded complexes of the S22 basis set, using EX+cRPA+SE, it induces errors as large as 40% on the binding

[6] A similar observation was made by Woon and Dunning [403] when optimizing correlation-consistent gaussian function to include core correlation. In that case, the optimized exponents of the gaussians strongly depended on the "valence" basis set (e.g. cc-pVDZ, cc-pVTZ, etc.) with which they were optimized.

[7] This is the reason why in Figure 11.6 the atomization CP corrected curve for $tier4 + spspdd$ looks worse than $tier2 + spspdd$. In reality, it is not worse: it has small errors where the binding energies are smaller and larger errors where binding energies are larger, having an overall better performance than $tier2 + spspdd$.

energy, as seen for conformer 11 for example. Increasing r_{onset} to 5 Å dramatically reduced all the errors. Going from 5 Å to 6 Å the improvements are not so dramatic. The cutoff potential of 5 Å, was chosen for being the best compromise between cost and accuracy.

11.3.3 Alanine dipeptide

The alanine dipeptide (Ac-Ala-NHMe), has been the subject of a number of first-principles studies, e.g. [27, 45, 67, 95, 192, 243, 272, 404–410] and many more, becoming a paradigm for first-principles calculations and force-field benchmarking. Several low-energy conformations have been characterized by theory. In Figure 11.8 the geometries of five different low energy MP2 conformation minima [8], discussed in Ref. [192], called C$_{7eq}$, C$_5$, C$_{7ax}$, β_2, and α_1 are shown schematically with the estimated CCSD(T) relative energies from Ref. [192]. The lowest-energy conformer predicted by most first-principles methods, called C$_{7eq}$ due to the existence of one H-bond of 2_7 type, has been observed experimentally in the gas-phase [411], in the only gas-phase experiment known involving this molecule. Since the main goal here is calculating relative energies of different conformers of the same biomolecule, these structures provide a good model system.

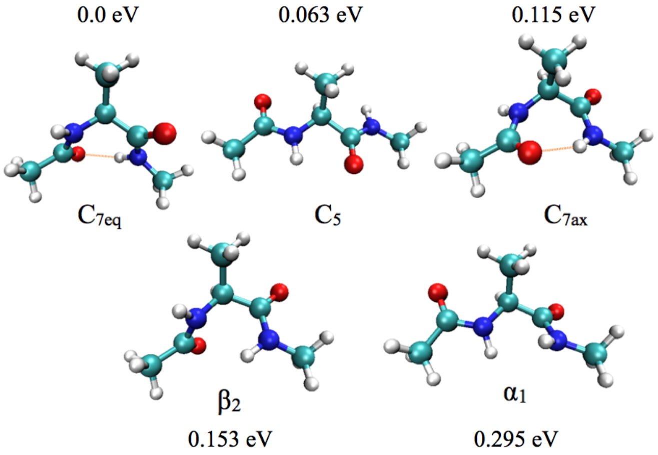


Figure 11.8: Geometries of the five studied alanine dipeptide conformers. The relative energy values were taken from Ref. [192] and are calculated with CCSD(T), extrapolated to the complete basis set limit.

The energy hierarchies obtained with EX+cRPA+SE@PBE and different basis sets are shown in Figure 11.9, with the C$_{7eq}$ conformer taken as the reference point. The orange curves represent the relative energy obtained with the aug-cc-pVNZ basis sets (N = D, T, Q, 5, 6), without any BSSE correction. The aug-cc-pV6Z value is considered the reference value here. The black curves were obtained with the NAO $tiers$ of FHI-aims with the addition of the core functions of Table 11.2, but without any BSSE correction. The purple symbols correspond to the black symbols, but including the atomization BSSE correction. For these relative energies, the addition of the core basis to the standard NAO $tiers$ give a good accuracy already for the uncorrected curve. The BSSE corrected values (purple) and the uncorrected ones differ by a maximum of 20meV as long as the extra core functions are added, and this error occurs only for the highest energy conformer, representing only 6% of the energy difference. The others present differences of maximum 10meV. This error is, thus, smaller than what was observed for the binding energies of the molecules in the S22 set. Moreover, convergence is reached for the $tier2 + spspdd$ basis sets, which are

[8] These particular five conformations are present in all levels of theory (DFT-B3LYP, MP2, HF, CCSD(T))[192].

of similar size to aug-cc-pVTZ.

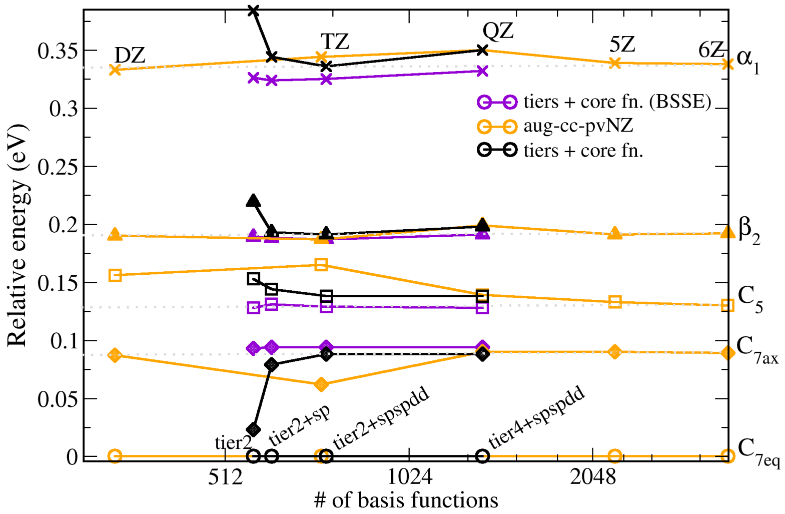


Figure 11.9: Relative energies for different conformations of the alanine dipeptide in EX+cRPA+SE with respect to basis set size. Comparison between the NAO basis sets (black symbols) and the Dunning gaussian basis sets aug-cc-pv$N$Z (orange symbols). In purple, the relative energies including the atomization BSSE correction (Eq. 11.1) for the NAO basis sets. Each different symbol shape corresponds to a different conformer, labeled on the right side of the plot.

It should be mentioned that the energetic ordering of the C_5 and C_{7ax} conformers is switched with respect to CCSD(T) data in the literature [192], reported in Figure 11.8. This is, in fact, a deficiency of EX+cRPA+SE@PBE, which is not related to the performance of the basis sets. Interestingly, results obtained with EX+cRPA+SE@PBE0 using the $tier2 + sp$ basis set, and including atomization BSSE correction, yield the correct energy hierarchy, with the C_5 conformer 80meV above C_{7eq} and the C_{7ax} conformer 102meV above C_{7eq}. Since the difference between these two conformers is the formation of a H-bond, this points to the fact, already discussed in Section 3.10, that PBE0 often describes the H-bond better than PBE.

The relative energies of these alanine dipeptide conformers were also calculated with MP2. Again, two different sets of basis functions were tested: the Dunning correlation consistent gaussian aug-cc-pv$N$Z (N = D, T, Q, 5, 6) basis set and the newly developed NAO basis sets. The results are shown in Figure 11.10, where both orange and black curves are without any BSSE correction. MP2 also predicts the correct ordering of the conformers, as can be seen by the label of the curves at the right side of the plot.

For the case of the water dimer shown in Section 5.2, MP2 shows less BSSE than EX+cRPA@PBE. Additionally, from the results for EX+cRPA@PBE in Figure 11.9, it was seen that these relative energies are even less affected by BSSE than the binding energies discussed in the previous section. Thus, the very good agreement shown in Figure 11.10 between the values obtained with $tier4 + spspdd$ and aug-cc-pv6Z are not expected to change appreciably by inclusion of an atomization BSSE correction.

Performing EX+cRPA or MP2 calculations with the additional NAO basis sets presented in this chapter, in order to obtain energy hierarchies of the Ac-Ala$_n$-LysH$^+$ conformers, should be possible and reliable. In the following, the $tier2 + spspdd$ basis sets will be used. The remaining errors, when not including the BSSE correction for the alanine dipeptide, are expected to be ≈1 meV per atom in the worst case (seen for the EX+cRPA@PBE calculations), and well below that for the best cases. The errors obtained when not including BSSE corrections for the S22 set were larger, though, although not drastically so. Furthermore the $tier2 + spspdd$ results in all cases show essentially the same accuracy as the

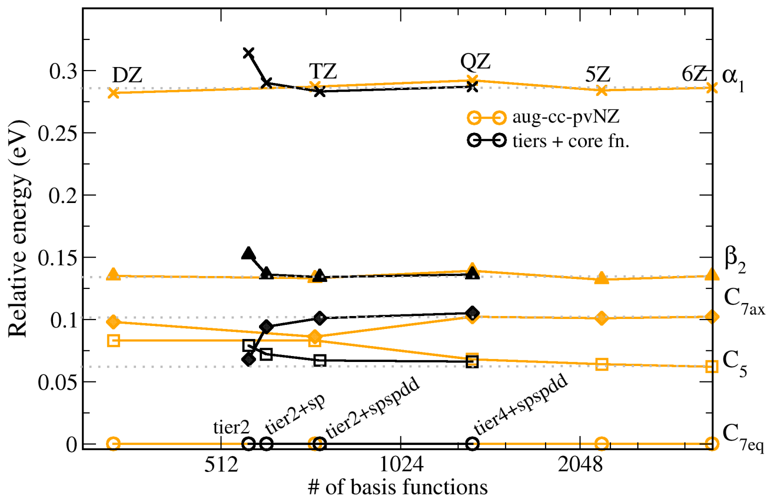


Figure 11.10: Relative energies for different conformations of the alanine dipeptide in MP2 with respect to basis set size. Comparison between the NAO basis sets (black symbols) and the Dunning gaussian basis sets aug-cc-pv$N$Z (orange symbols), without BSSE corrections. The gray dotted lines are a guide to the eye, aligned at the aug-cc-pV6Z value. Each different symbol shape corresponds to a different conformer, labeled on the right side of the plot.

$tier4 + spspdd$. When including atomization BSSE corrections the results are expected to be converged with negligible errors for the $tier2 + spspdd$ basis.

11.4 Summary

In summary, the advantage of the additional basis sets presented here are: (i) they substantially reduce the BSSE when used with methods that take non-local correlation explicitly into account, even without any BSSE correction (ii) for relative energies, $tier2 + spspdd$ achieves a very good accuracy for a small basis set size (comparable to the size of aug-cc-pVTZ). Going to a finer accuracy for energy differences still requires the use of a BSSE correction, but with these basis sets it is possible to converge atomization BSSE corrected energies.

Chapter 12

Explicitly correlated methods for Ac-Ala$_5$-LysH$^+$ conformational energies

In this chapter, the basis sets discussed in Chapter 11 are used, in connection with MP2 and EX+cRPA, to obtain converged relative energies for the low-energy conformers of Ac-Ala$_5$-LysH$^+$, discussed in Chapter 8.

12.1 Ac-Ala$_5$-LysH$^+$: benchmarking against explicitly correlated methods

For the case of the Ac-Ala$_5$-LysH$^+$ molecule, the considered test cases are again the four conformers studied in Section 8.2, namely, the ones labeled g-1 (Family 1), α-1 (Family 2), α-2 (Family 3), and 3_{10}-1 (lowest energy 3_{10}-helical conformer). For the MP2 calculations, the starting point are the self consistent converged Hartree-Fock (HF) orbitals as is conventional. For the EX+cRPA(+SE) calculations, three different starting points are used: PBE, PBE0, and HF.

In Figure 12.1(b) the relative energies of the g-1, α-1, α-2, and 3_{10}-1 conformers are plotted, for the above mentioned methods. For EX+cRPA@HF the singles correction does not contribute, since, in this case, the orbitals are eigenfunctions of the HF hamiltonian and all matrix elements corresponding to singly excited orbitals interacting with the ground-state are zero. The basis set used was $tier2 + spspdd$ and the atomization BSSE correction was performed for EX+cRPA(+SE)@PBE and EX+cRPA(+SE)@PBE0, where consistent atomic ground states can be guaranteed with relative ease. The largest correction in the relative energies was of 30meV, observed between the g-1 and 3_{10}-1 conformers, inducing a stabilization of the 3_{10}-1 over the g-1. The correction for EX+cRPA@HF and MP2 is expected to be at the most of the same magnitude.

EX+cRPA@PBE and EX+cRPA+SE@PBE agree with the trend seen from the vdW corrected functionals. For the values calculated at PBE0 self-consistency, although g-1 continues to be the lowest energy conformer (if only barely for the EX+cRPA data point), the situation is already slightly modified, with the 3_{10} conformer getting stabilized, being almost as stable as α-1. EX+cRPA@HF and MP2 predict the 3_{10} conformer to be even more stabilized over g-1. In particular, for EX+cRPA@HF, g-1 becomes the

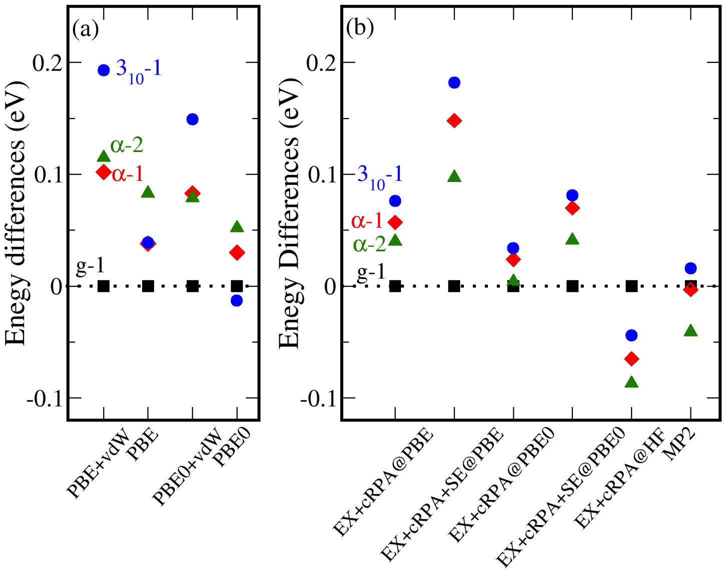


Figure 12.1: In panel (a) the energy hierarchies for the for the the chosen conformers of Ac-Ala$_5$-LysH$^+$: g-1 (black), α-1 (red), α-2 (green), and 3_{10}-1 (blue), are shown again (Fig. 8.10) for PBE, PBE+vdW, PBE0, and PBE0+vdW. In panel (b), energy hierarchies calculated with EX+cRPA@PBE, EX+cRPA+SE@PBE, EX+cRPA@PBE0, EX+cRPA+SE@PBE0, EX+cRPA@HF, EX+cRPA+SE@HF, and MP2 are shown for the same conformers. All points are corrected for atomization BSSE, except for EX+cRPA@HF and MP2.

highest energy conformer between the four. This situation is quite striking, and contradicts not only what was seen for all other vdW-corrected functionals (Section 8.2) and the EX+cRPA(+SE)@PBE data, but also the direct comparison with experimental IR spectra (Chapter 9.2), where the 3_{10}-1 conformer is not compatible with the measured spectrum.

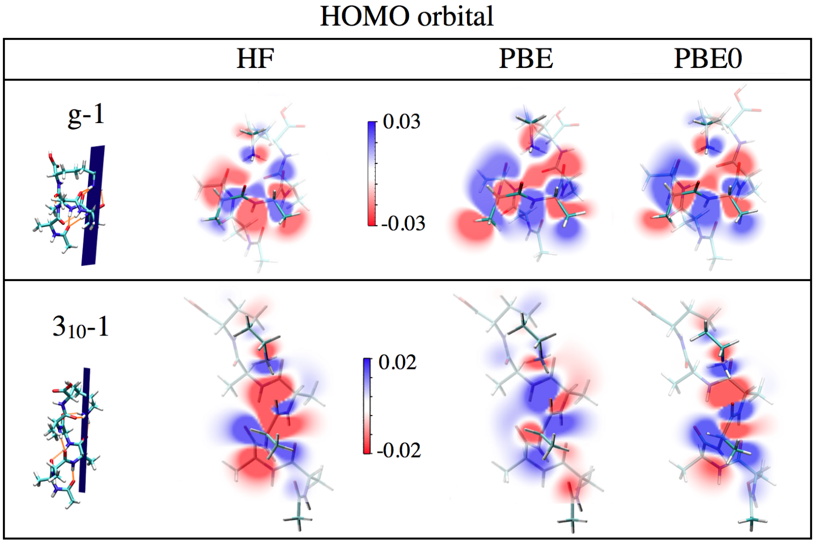


Figure 12.2: Top panel: HOMO level of the g-1 conformer of n=5 calculated with PBE, HF, and PBE0. Bottom panel: same as top one for the 3_{10}-1 conformer of n=5. The values corresponding to the colors are shown in the color-bar, with units of e/Å^3.

In order to understand this discrepancy, the starting point of each calculation (PBE, PBE0, and HF) was analyzed more carefully. While the PBE (standard) functional predicts g-1 to be the most stable conformer, PBE0 already predicts the 3_{10}-1 conformer to be slightly preferred energetically, as can be seen in Figure 12.1(a). Calculating the HF energy hierarchy for these conformers, it was found that the 3_{10}-1 conformer is 0.24 eV *more* stable than g-1, and α-1 and α-2, 0.20 eV and 0.19 eV more stable, respectively. This points to the fact that, upon inclusion of exact exchange in the functionals, the g-1 conformer gets destabilized while the 3_{10}-1 conformer gets stabilized. The correlation tries to correct this trend, but apparently is not enough in the case of MP2 and EX+cRPA based on HF orbitals.

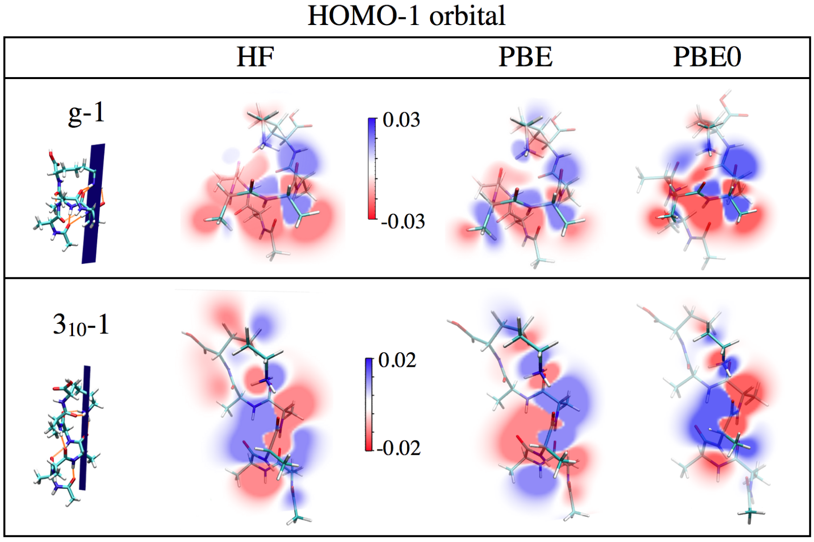


Figure 12.3: Top panel: HOMO-1 level of the g-1 conformer of n=5 calculated with PBE, HF and PBE0. Bottom panel: same as top one for the 3$_{10}$-1 conformer of n=5. The values corresponding to the colors are shown in the color-bar, with units of e/Å^3.

Inspecting the orbitals themselves, very different electronic state descriptions are found when comparing HF, PBE0 and PBE. While the lowest unoccupied orbital (LUMO) appear quite similar in all methods for the g-1 and the 3$_{10}$-1 conformers, the highest occupied molecular orbital (HOMO) and the orbital just below it (HOMO-1) present differences. A cut of these orbitals, coming from all 3 functionals and passing through the NH$_3^+$ group of the molecules is shown in Figures 12.2 and 12.3. On the backbone of the molecules, there is mainly just a phase-shift of of the orbitals (positive to negative), but close to the NH$_3^+$ group the changes are clearly more drastic, with the orbitals having different characters.

Taking into account all the results presented so far in this thesis, especially the good agreement with experiment found for PBE+vdW energy hierarchies, IR spectra and dynamics for Ac-Ala$_n$-LysH$^+$, it is tempting to conclude that HF and the perturbative methods based on its orbitals are giving a wrong description of these systems. Moreover, the EX+cRPA+SE@PBE method has been shown [190] to give results closer (than MP2 or RPA) to CCSD(T) data for other systems. On the other hand, the over-localization of the electronic density in HF [402, 412, 413], may produce discrepancies, when dealing with charged systems. Nevertheless, the data presented here proves that this problem is a really complicated one, and ideally, one would need to go to CCSD(T), with converged basis sets, to have a definitive benchmark. This is at present unfeasible, but there is, clearly, a dire need for a benchmark quality method that describes systems of this size reliably. Steps towards this goal are being taken and this will be the subject of future work performed in our group.

Chapter 13

Conclusions and outlook

In this thesis several standing challenges for the accurate description of the secondary structure formation in peptides are addressed, using first-principles electronic structure methods. The model system used here was the alanine-based polypeptide series Ac-Ala$_n$-LysH$^+$ in the gas-phase, which is a prototype for the formation of helical secondary structure.

The challenge of performing an extensive and efficient exploration of the large conformational space was approached by a two-step procedure. The first step consists of a broad conformational screening, using a basin-hopping search with an empirical force-field (OPLS-AA) as a structure generator. The second step consists of hundreds of relaxations using density-functional theory (PBE exchange-correlation functional), with the inclusion of van der Waals corrections [2] (PBE+vdW) [1]. The capabilities and limitations of this search procedure were discussed for n=4 and 5. It was found that there is a weak correlation between the OPLS-AA and PBE+vdW energy hierarchies, and that the OPLS-AA force-field induces systematic energy overestimations of certain conformers, in particular 3_{10} helices, when compared to PBE+vdW. Since the PBE+vdW functional was benchmarked against coupled cluster data for small alanine peptides (see Section 3.10), and it is an *ab initio* electronic structure method (as opposed to force fields), it is taken as the “trusted” method in this work. In order not to miss relevant conformers, a large amount of DFT relaxations must be performed and care must be taken to include known force-field limitations, as, e.g., 3_{10}-helices.

The conformational search was applied for n=4-8, and characterization of low-energy DFT-PBE+vdW conformers was presented. The α-helix is found to be the lowest energy structure for n≈7-8 in the DFT-PBE+vdW potential energy surface (PES), but with a very small energy difference to the next non-helical conformer. n=5 and 6 present a more compact structure, with an inverted H-bond, as the lowest energy PBE+vdW (PES) structure. For molecules larger than n=8 (110 atoms) the conformational freedom becomes prohibitive for the search strategy presented here, due to the combinatorial explosion of possible conformations and the bad force field-PBE+vdW correlation for energy hierarchies. This limitation motivates future work on the enhancement of this search strategy, either by performing it fully within *ab initio* methods or developing better force-fields (possibly also parametrized for gas-phase molecules), in order to generate more reliable input structures for the first-principles calculations.

The quality of the DFT-PBE functional and the description of non-covalent interactions (especially vdW) was also studied. For n=5 several DFT functionals were tested (GGAs, mGGAs and hybrids), with the energy hierarchy exhibiting a similar trend, as long as vdW interactions are added to all functionals.

[1] For the DFT calculations, the all-electron, localized-basis code FHI-aims [1] was used. Development contributions to this code were also necessary.

The situation for the “plain” functionals (no vdW) is inconclusive, with different functionals predicting different lowest energy structures. In fact, the “isolated” preference for an α-helical structure at n=8 would not be predicted by plain PBE.

From the predicted low energy conformations of n=4-8, it was possible to make a connection to the real world, by computing (harmonic) vibrational free energies and comparing to several experiments. Computation of harmonic relative free energies predicts the cross-over to α-helical preference to happen, safely, at n=8, with PBE+vdW. At this size, a large free-energy difference at 300K (more than 0.1eV) is observed between the most stable α-helical conformer and the next non-helical one. This observation is in good agreement with the experiment of Ref. [3], where it was suggested that the helix would be stabilized at n=8 for these molecules (see Section 6.3.2). The stabilization of helices over the globular/compact conformers could be understood by the existence of a low frequency first vibrational mode involving the bending of the terminations, that makes conformations that are elongated and exhibit periodicity (vibrationally) entropically favored.

The first-principles results were compared to a few experiments. The ion mobility cross-sections for the conformers that were characterized for all n were calculated from an empirical potential for the ion-buffer gas interaction, but with the geometries taken from DFT-PBE+vdW. The cross-sections reproduce well the experimental data of Ref. [37] when considering a Boltzmann average (300K) with the weights calculated from the DFT-PBE+vdW harmonic free-energy differences. For small conformers, several structures contribute to the measured cross-section. The IR spectra of Ac-Ala$_n$-LysH$^+$, n=5, 10, 15 was calculated and compared to room-temperature IRMPD data obtained for the exact same molecules [38]. It was possible to provide strong evidence that for n=10 and 15 the structure of this molecule is firmly α-helical, consistent with the predicted α-helical onset at n=8. The experimental spectrum for the shorter conformer, n=5, could be explained by a mixture of low energy conformers co-existing in the beam at 300K. Both helical (α-1 and α-2) structures, as well as more compact (g-1) structures should be present, in agreement with the above mentioned (harmonic) free-energy hierarchy analysis and ion-mobility cross sections assessment. For all spectra computed, the inclusion of anharmonicities (including local configurational freedom) and temperature effects, via the dipole autocorrelation function derived from *ab initio* molecular dynamics runs, were seen to improve substantially the theory-experiment comparison. The match between theory and experiment was quantified by means of a reliability factor (Pendry R-factor), which is a great improvement on qualitative comparisons based only on visual inspections.

The unfolding mechanism of Ac-Ala$_{15}$-LysH$^+$ was studied by means of several picoseconds of *ab initio* molecular dynamics simulations, in order to probe if DFT could reproduce the experimentally observed [5] remarkable stability of Ac-Ala$_{15}$-LysH$^+$ *in vacuo* up to high temperatures. At high temperatures (500K-800K) several tens of picoseconds of simulation ($30-80$ps) are enough to describe the unfolding mechanism. VdW interactions are seen to be essential to predict the correct stability of the helix up to ≈700K. DFT-PBE simulations render the molecule too unstable, being already unfolded at 700K, in disagreement with experiments. The inclusion of vdW interactions dramatically changes the conformational landscape explored, favoring the exploration of more compact helices, with the helices exhibiting a mostly α-helical character at finite temperatures. PBE favors a mostly 3$_{10}$ helix. A synergy between the charge, the connecting H-bonds of the Lys termination, and vdW is crucial to explain the dynamical high-temperature stability of the helix. Attempts to *fold* the molecule were shown. A full first-principles folding simulation for even this “not so large” peptide at high temperatures is still unreachable, but it is possible to learn something about the probable beginning of the folding path and the landscape explored. A connection to accurate statistical methods that allow for an efficient first-principles exploration of the folding landscape of the molecules needs to be developed, so that these simulations can span longer times and explore the full folding mechanism.

Finally, benchmarks to methods that explicitly describe the non-local, long range electronic correlation (MP2, EX+cRPA) were computed. Additional NAO basis sets that substantially reduce the BSSE within these methods and that allow the convergence of the relative energies between these conformers were developed. The benchmark data presented for n=5 (80 atoms) proved that the problem is very intricate, with a strong dependence on the starting point (orbitals) used for perturbative methods like MP2 and EX+cRPA. Given the good agreement of PBE+vdW data with the several experiments mentioned above, it seems that the perturbative methods based on HF orbitals (MP2 and EX+cRPA@HF) do not describe correctly these molecules. In order to settle this issue, it would be necessary to obtain a definitive benchmark data from even more accurate quantum-chemistry methods [ideally, CCSD(T)], which is at present unfeasible due to the high computational cost. Steps in this direction must be pursued, though.

The work shown here underlines the importance of the non-covalent vdW interactions in the description of polypeptides. The PBE functional, allied to the TS-vdW correction, yields a good description of the polypeptides addressed in this work, with an excellent match to several experiments. The computational cost of these calculations are essentially the same as a DFT calculation, so that they are very appealing for use in a wide range of problems. Moreover, PBE+vdW may be used as an affordable, benchmark-quality method, for treating these systems. The results obtained in this thesis represent a step leading to a full, quantum-mechanical based, *in silico* understanding of the properties of proteins.

Appendices

Appendix A

Exchange enhancement factors of various GGAs

The reduced gradient:

$$s = \frac{|\nabla n(\vec{r})|}{2(3\pi^2)^{1/3}n^{4/3}(\vec{r})} \tag{A.1}$$

The enhancement factors:

$$F_x^{PBE} = 1+\kappa-\tfrac{\kappa}{1+\mu s^2/\kappa}, \text{with: } \kappa = 0.804 \text{ and } \mu = 0.2195 \tag{A.2}$$

$$F_x^{PBEsol} = 1+\kappa-\tfrac{\kappa}{1+\mu s^2/\kappa}, \text{with: } \kappa = 0.804 \text{ and } \mu = 0.1235 \tag{A.3}$$

$$F_x^{revPBE} = 1+\kappa-\tfrac{\kappa}{1+\mu s^2/\kappa}, \text{with: } \kappa = 1.245 \text{ and } \mu = 0.2195 \tag{A.4}$$

$$F_x^{RPBE} = 1+\kappa(1-e^{-\mu s^2/\kappa}), \text{with: } \kappa = 0.804 \text{ and } \mu = 0.2195 \tag{A.5}$$

$$F_x^{B88} = 1+\tfrac{\mu s^2}{1+s\beta \textit{arcsinh}(cs)}, \text{with: } c = 2^{4/3}(3\pi^2)^{1/3} \;\; \mu \approx 0.2743 \;\; \beta = \tfrac{9\mu(6/\pi)^{1/3}}{2c} \tag{A.6}$$

$$H_x^{AM05} = X(s)+[1-X(s)]F_x^{LAA}(s) \tag{A.7}$$

$$X(s) = \frac{1}{1+\alpha s^2}, \;\alpha = 2.804 \tag{A.8}$$

$$F_x^{LAA}(s) = \frac{cs^2+1}{cs^2/F_x^b+1}, \;c = 0.7168 \tag{A.9}$$

$$F_x^b = \frac{1}{\epsilon_x^{LDA}\tilde{n}_0(s)2\xi(s)} \tag{A.10}$$

$$\xi(s) = \{[(4/3)^{1/3}2\pi/3]^4\zeta(s)^2+\zeta(s)^4\}^{1/4} \tag{A.11}$$

$$\zeta(s) = \left[\frac{3}{2}W\left(\frac{s^{3/2}}{2\sqrt{6}}\right)\right]^{2/3} \quad , \quad \tilde{n}_0(s) = \frac{\zeta(s)^{3/2}}{3\pi^2 s^3} \tag{A.12}$$

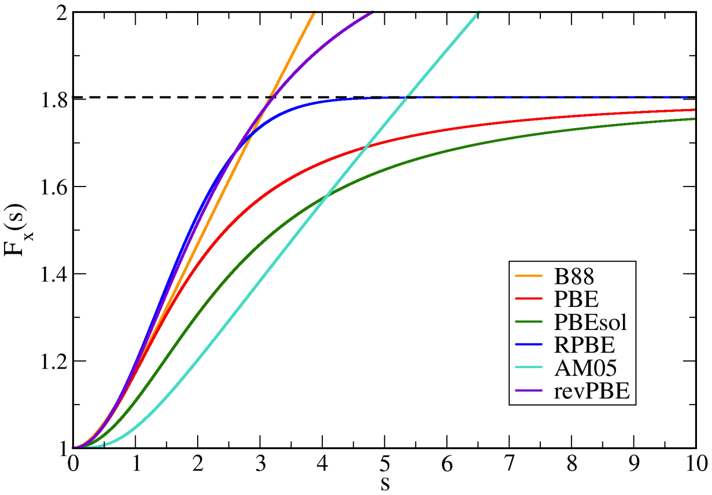


Figure A.1: Exchange enhancement factors F_x as a function of s for different GGA functionals. The dashed line at 1.804 corresponds to the local Lieb-Oxford upper bound.

Appendix B

FHI-aims standard NAO basis sets for H, C, N, and O.

The standard NAO basis sets of FHI-aims are developed by optimizing the binding energies of symmetric dimers of each element. The strategy (similar to the one discussed in Chapter 11) includes defining a large pool of possible radial function shapes with a variable confinement potential and, starting from the minimal basis of the free atoms, run through the pool and choose the function that lowers the most the LDA total energy of the dimer (averaged for various separations). After adding the function to the initial basis set, this process is repeated until there are no more significant improvements in the total energy. The basis sets obtained in this way are shown in Table B.1 for hydrogen, carbon, nitrogen and oxygen. Each radial function labeled has its corresponding $(2l + 1)$ angular functions. There are two types of function: hydrogen-like with effective ionic charge Z, and ionic functions of doubly positively charged ions of the corresponding elements. In this way, a function called “hydro 2 p 1.7” means that it is a hydrogen-like function of the 2p type with an effective Z of 1.7. A function called “ionic 2 p auto” for carbon, for example, means that it corresponds to the 2p function of the C^{2+} atom and the onset radius of confinement potential is automatically chosen so that it is the same as the one specified for the radial function equation.

The functions chosen for each element are sorted into $tiers$, in the order that they were chosen automatically from the script, and therefore meaning that the first $tier$ is the one that brought most energetic improvement, the second $tier$ less, etc. For the *light* elements discussed here, the functions in $tier1$ are the ones that represented total energy improvements for the dimers of ≈1eV to ≈100meV and $tier4$ of ≈20meV. The separation between each of them is not unique, but, for example, for *light* elements of the second row of the periodic table the first $tier$ always includes s, p, d functions.

	H	C	N	O
$minimal$	1s	[He] + 2s2p	[He] + 2s2p	[He] + 2s2p
$tier1$	hydro 2 s 2.1 hydro 2 p 3.5	hydro 2 p 1.7 hydro 3 d 6 hydro 2 s 4.9	hydro 2 p 1.8 hydro 3 d 6.8 hydro 3 s 5.8	hydro 2 p 1.8 hydro 3 d 7.6 hydro 3 s 6.4
$tier2$	hydro 1 s 0.85 hydro 2 p 3.7 hydro 2 s 1.2 hydro 3 d 7	hydro 4 f 9.8 hydro 3 p 5.2 hydro 3 s 4.3 hydro 5 g 14.4 hydro 3 d 6.2	hydro 4 f 10.8 hydro 3 p 5.8 hydro 1 s 0.8 hydro 5 g 16 hydro 3 d 4.9	hydro 4 f 11.6 hydro 3 p 6.2 hydro 3 d 5.6 hydro 5 g 17.6 hydro 1 s 0.75
$tier3$	hydro 4 f 11.2 hydro 3 p 4.8 hydro 4 d 9 hydro 3 s 3.2	hydro 2 p 5.6 hydro 2 s 1.4 hydro 3 d 4.9 hydro 4 f 11.2	hydro 3 s 16 ionic 2 p auto hydro 3 d 6.6 hydro 4 f 11.6	ionic 2 p auto hydro 4 f 10.8 hydro 4 d 4.7 hydro 2 s 6.8
$tier4$		hydro 2 p 4.5 hydro 5 g 16.4 hydro 4 d 13.2 hydro 3 s 13.6 hydro 4 f 17.6	hydro 2 p 4.5 hydro 2 s 2.4 hydro 5 g 14.4 hydro 4 d 14.4 hydro 4 f 16.8	hydro 3 p 5 hydro 3 s 3.3 hydro 5 g 15.6 hydro 4 f 17.6 hydro 4 d 14

Table B.1: Standard NAO basis for hydrogen, carbon, nitrogen and oxygen, distributed with FHI-aims[1].

Appendix C

Time correlation functions

A (classical) time correlation function is defined as:

$$C(t) = \langle A(0)A(t) \rangle, \tag{C.1}$$

where A is the observable of interest and the angle brackets represent an ensemble average.

Since the starting time is arbitrary, in a molecular dynamics trajectory one can choose as many time for $t = 0$ as one wants, and C.4 would be rewritten as:

$$C(t) = \langle A(t_0)A(t_0 + t) \rangle \tag{C.2}$$

Therefore, in only one (long enough) trajectory it is already possible to obtain an ensemble average. Very short time intervals between two different t_0 will not help in this average since they will still be very correlated and will produce essentially the same curve. Only time origins that are far enough apart so that the correlation between them is low will contribute.

If the system is ergodic (which should be true for molecular dynamics)[201], then (weighted) ensemble averages of the auto-correlation functions are equal to their time averages:

$$\langle A(0)A(t) \rangle = \overline{A(0)A(t)} = \lim_{T \to \infty} \frac{1}{T} \int A_i(\tau)A_i(t + \tau)d\tau \tag{C.3}$$

The auto-correlation function presents an initial decay that is proportional to the correlation time and represents the loss of correlation between time t and t_0.

A quantum time auto-correlation function is generally defined as

$$C_{AA} = \frac{\text{Tr}[A(t)A(0)\exp(-\beta \hat{H})]}{\text{Tr}[\exp(-\beta \hat{H})]}, \tag{C.4}$$

where A(t) is defined as the Heisenberg time-dependent operator $\exp(i\hat{H}t/\hbar)\hat{A}\exp(-i\hat{H}t/\hbar)$.

Appendix D

Benchmarks for Ammonia (NH_3)

This appendix shows the tests made for ammonia (Figure D.1) with respect to the accuracy parameters of the AIMD microcanonical simulations and the harmonic vibrations.

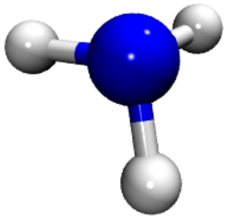


Figure D.1: Schematic drawing of a NH_3 molecule.

D.1 Ab initio Molecular Dynamics

Figure D.2 shows the total energy of an AIMD microcanonical simulation with respect to simulation time. The time step (Δt) is 0.5fs and the Verlet integration algorithm was used. The black curve was obtained with the tight convergence settings for the self-consistent cycle used through this thesis. In terms of the FHI-aims code flags they are:

```
sc_accuracy_rho 1E-5
sc_accuracy_eev 1E-4
sc_accuracy_etot 1E-6
sc_accuracy_forces 5E-4
```

The red curve in Figure D.2 was obtained with the less accurate settings of:

```
sc_accuracy_rho 1E-2
sc_accuracy_eev 1E-1
sc_accuracy_etot 1E-3
sc_accuracy_forces 1E-2
```

The energy drift of the red curve is completely unphysical, since the total energy should be conserved in a microcanonical simulation.

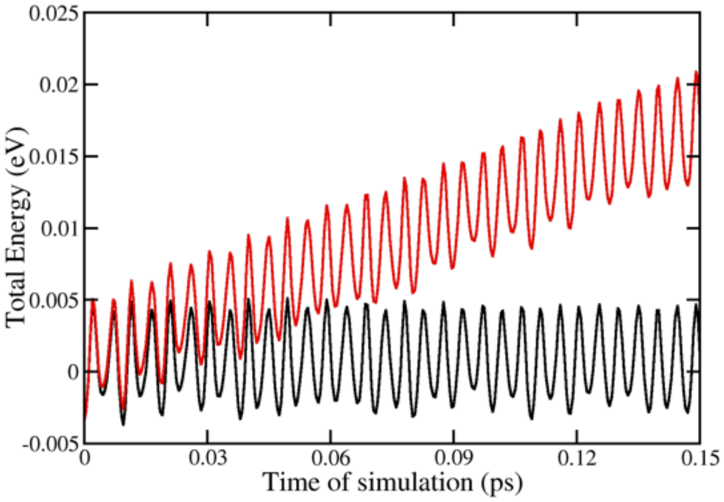


Figure D.2: Molecular dynamics total energy as a function of AIMD time step. Black: tight convergence settings (see text). Red: light convergence settings (see text), exhibiting unphysical energy drift. The zero value of the y-axis was assigned to the average total energy of the tightly converged simulation.

Figure D.3 shows a comparison of the Verlet integration algorithm using Δt=0.5fs (black) and Δt=1fs (red) , plus a 4th order integrator (goes up to fourth order expansion in the Taylor series of the position expression) studied in Ref. [215], with Δt=2fs (blue). Comparing the Verlet curves, one can see that the one obtained with the smaller time-step (Δt=0.5fs, black) presents oscillation amplitudes smaller than the one of the larger time-step (Δt=1fs, red). Although for a system of the size of NH_3 the 1fs time-step is not ideal because a 0.02-0.03eV energy fluctuation represents a relevant fraction of these systems, it is not so much for the larger systems studied in this thesis. Moreover, for the evaluation of the vibrational spectra of the systems this time-step is accurate enough, since the vibrations involved have a much lower frequency than the accuracy involved (1fs $\approx$ 33000cm^{-1}).

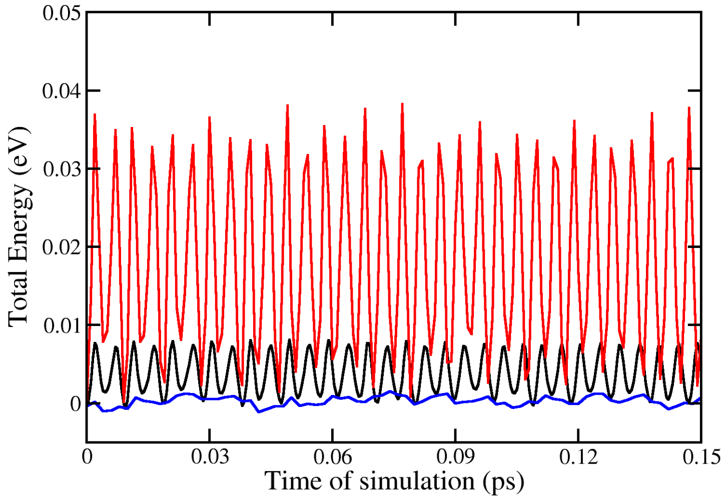


Figure D.3: Molecular dynamics total energy as a function of AIMD time step, always using tight convergence settings. Blue: 4th order integrator[215], with Δt=2 fs. Black: velocity Verlet, with Δt=0.5 fs. Red: velocity Verlet, with Δt=1.0 fs. The zero value of the y-axis was assigned to the average total energy of the 4th order integrator (blue) curve.

The 4th-order integrator shows the best accuracy for the larger time step in Figure D.3. However, as can be seen in Figure D.4, it also involves more force-evaluations. Since force-evaluations are very expensive for *ab initio* simulations, one needs to use the largest time step possible with the fewer number of force-evaluations for a fixed amount of simulation time. These requirements lie in the left region of Figure D.4, where the Verlet-algorithm presents a better accuracy. Therefore, in this thesis, the Verlet algorithm with Δt=1fs was used for all simulations.

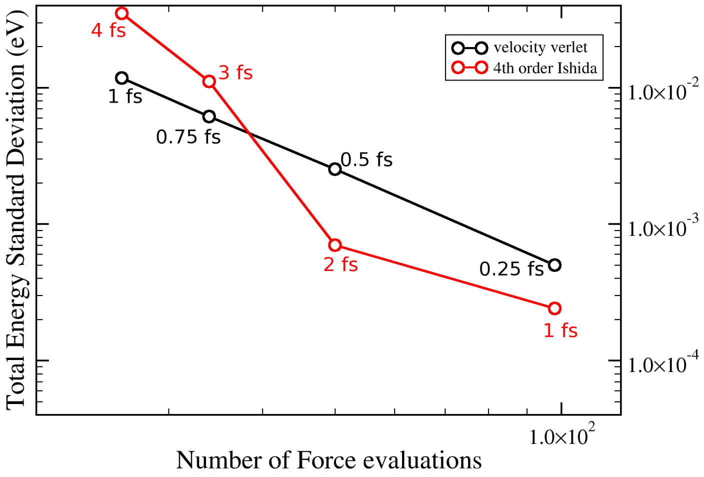


Figure D.4: Comparison of the standard deviation of the total energy for 0.15ps AIMD runs for the Verlet and 4th order integrator[215], as a function of Δt and number of force evaluations. In the region of interest (left side of the plot, larger time-steps and less force evaluations), velocity Verlet performs better for the same number of force evaluations.

D.2 Harmonic vibrations

For the harmonic vibrations, there is one parameter possible to be adjusted, namely, the length of the displacements Δ used to calculate the forces in all the cartesian directions. This was varied from 0.001Åto 0.010Å, using the FHI-aims $tight$ settings for basis-set and grids for ammonia. Monitoring the frequencies obtained (reported in Table D.1), especially numbers 2 and 3, and 5 and 6 which should be degenerate, one concludes that up to 0.007Åthe accuracy is still acceptable, but not optimal. 0.001Åis possibly too small, being subject to eventual wiggles in the PES due to grid inacuracies. 0.003Åor 0.005Åshould be the better choices. In both cases, the accuracy is less than 1cm^{-1}

Table D.1: Harmonic frequencies of vibration (in cm^{-1}) of NH$_3$ calculated with the PBE functional and *tight* settings of FHI-aims. Various displacements for the finite differences method of calculating second derivatives are tested.

Displacement (Å)	0.001	0.003	0.005	0.007	0.010
1	1016.4	1016.6	1016.5	1016.4	1016.1
2	1018.2	1618.2	1618.2	1618.0	1017.8
3	1018.3	1618.3	1618.3	1618.2	1018.2
4	3387.9	3388.0	3388.0	3388.0	3388.2
5	3512.3	3512.3	3512.3	3512.3	3512.4
6	3512.3	3512.3	3512.3	3512.4	3512.5

D.3 Anharmonic vibrational spectrum

In Figure D.5 the IR spectrum of NH_3 was calculated in the harmonic approximation (Section 4.1) and from the dipole autocorrelation function (Section 4.3) coming from a 4ps long AIMD (DFT-LDA) run thermalized at 600K. Both calculations were done with the *light* settings of FHI-aims and $tier1$ basis sets, hence the difference in the frequencies of vibrations if compared to Table D.1. The purpose here was just to compare the two approximations for a simple molecule, so that these settings are sufficient. In the anharmonic case, overtones and frequencies combinations (e.g. the one at 623 cm^{-1}) can be observed, which are forbidden in the harmonic approximation.

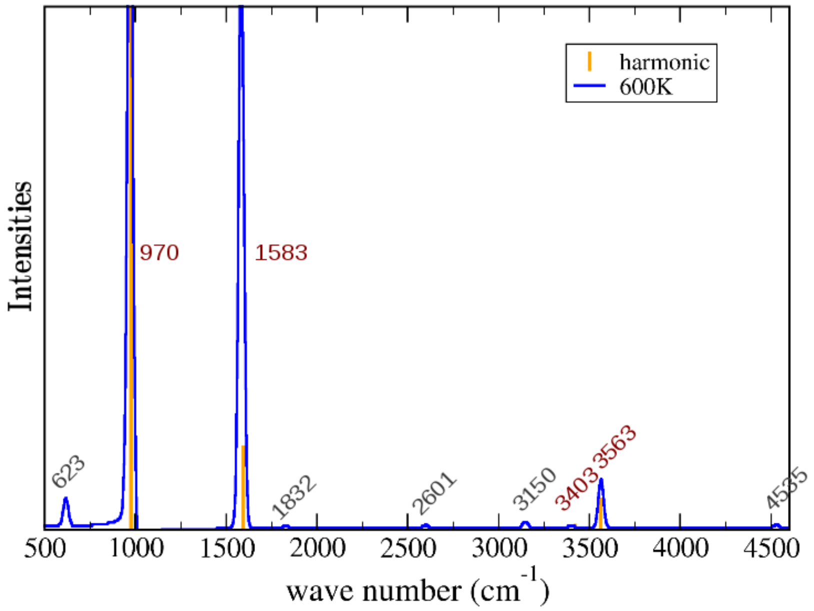
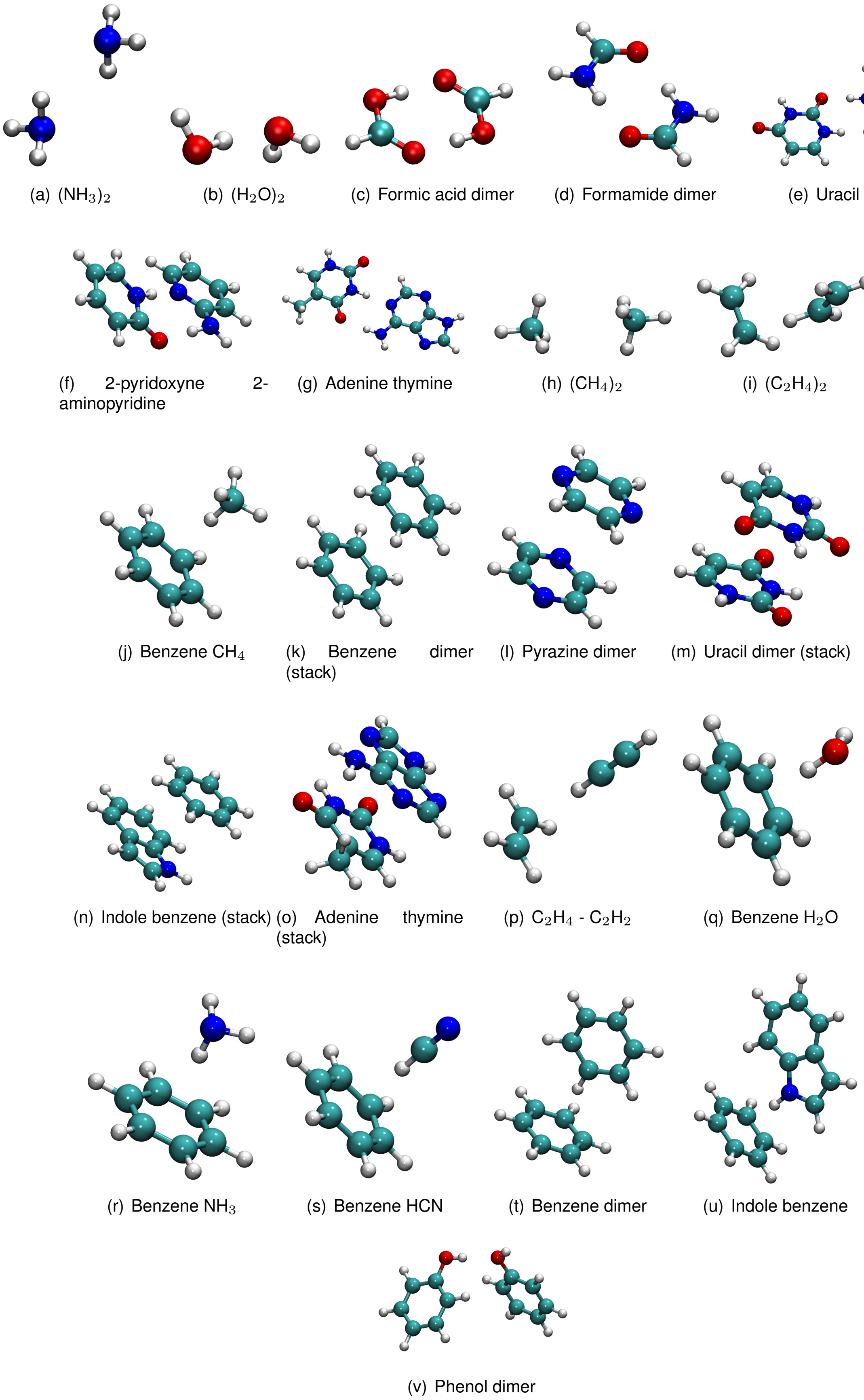


Figure D.5: Comparison of the harmonic spectrum of vibration calculated with *light* settings and the one obtained from 4ps of AIMD at $< T >= 600K$, also with light settings. Black numbers in the figure indicate peaks that do not correspond to fundamental vibrations, like overtones or combinations of the fundamental peaks.

Appendix E

Geometries of the molecules in the S22 database

The following figure shows the geometries and names of the molecules belonging to the S22 database, as proposed in ref. [178]. This set of molecules is proposed as a training set of non-covalent interaction. In the original publication, they are divided in three groups, according to the predominant character of the non-covalent bond present: numbers 1–7 are hydrogen bonded; numbers 8–15 are dispersion bonded; and numbers 16–22 are "mixed" complexes. CCSD(T) binding energy values extrapolated to the complete basis set limit are also reported in ref. [178] for all the 22 complexes.



(a) $(NH_3)_2$

(b) $(H_2O)_2$

(c) Formic acid dimer

(d) Formamide dimer

(e) Uracil dimer

(f) 2-pyridoxyne 2-aminopyridine

(g) Adenine thymine

(h) $(CH_4)_2$

(i) $(C_2H_4)_2$

(j) Benzene CH_4

(k) Benzene dimer (stack)

(l) Pyrazine dimer

(m) Uracil dimer (stack)

(n) Indole benzene (stack)

(o) Adenine thymine (stack)

(p) C_2H_4 - C_2H_2

(q) Benzene H_2O

(r) Benzene NH_3

(s) Benzene HCN

(t) Benzene dimer

(u) Indole benzene

(v) Phenol dimer

Figure E.1: Geometries and names of the complexes belonging to the S22 database.

Appendix F

Vibrations assignment to normal modes

The case of α-helical Ac-Ala$_{10}$-LysH$^+$

Upon a normal mode analysis in the harmonic approximation, it is possible to assign vibrations of certain groups in the molecule for each peak in the IR spectrum. In Figure F.1 a detailed account of the character of the vibration corresponding to each peak of the IR spectrum of Ac-Ala$_{10}$-LysH$^+$ (α-helical conformation), between 1000 and 1800 cm^{-1} is shown.

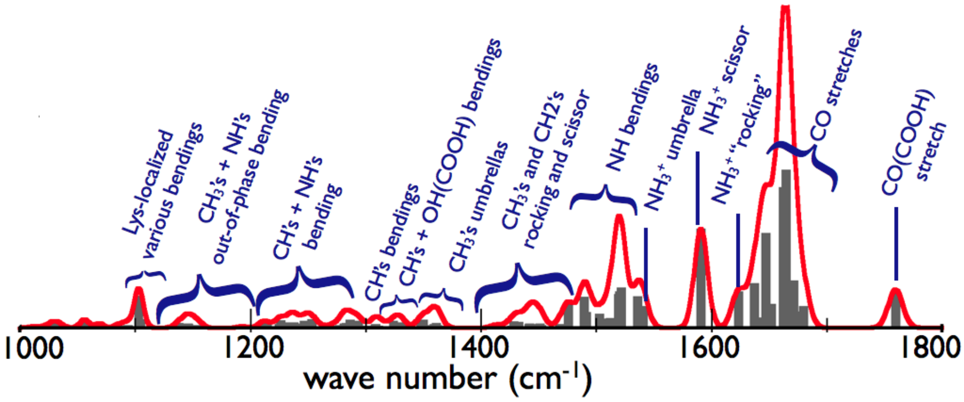
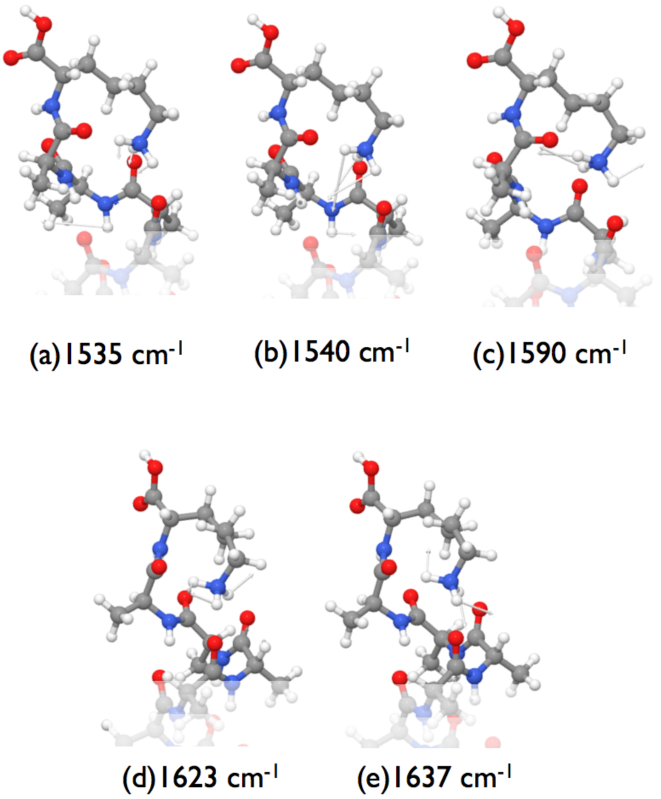


Figure F.1: Assignment of specific vibrations to each IR active frequency of the α-helical geometry of Ac-Ala$_{10}$-LysH$^+$.

The normal modes corresponding to the NH bends involving the NH$_3^+$ group, with their corresponding frequencies of vibration, are shown in Figure F.2.

The peak at ≈ 1590 cm^{-1} ("scissor" vibration), does not show such a high intensity in the experimental spectrum, as well as in the anharmonic spectra computed from AIMD (see Chapter 9.2). Two things can happen: either the mode is too anharmonic and undergoes a substantial shift, or it loses intensity through anharmonic coupling with other modes. We have investigated the anharmonic character of the normal mode by calculating single-point energies upon displacements on the direction of this mode. The energies obtained for each displacement, compared to a purely harmonic potential, is shown in Figure F.3. Since the harmonic approximation seems to be quite fulfilled for this mode, it probably loses intensity



(a) 1535 cm^{-1} (b) 1540 cm^{-1} (c) 1590 cm^{-1}

(d) 1623 cm^{-1} (e) 1637 cm^{-1}

Figure F.2: Vibrations localized at the NH_3^+ group. Arrows point to the directions of displacement.

through anharmonic coupling with other modes.

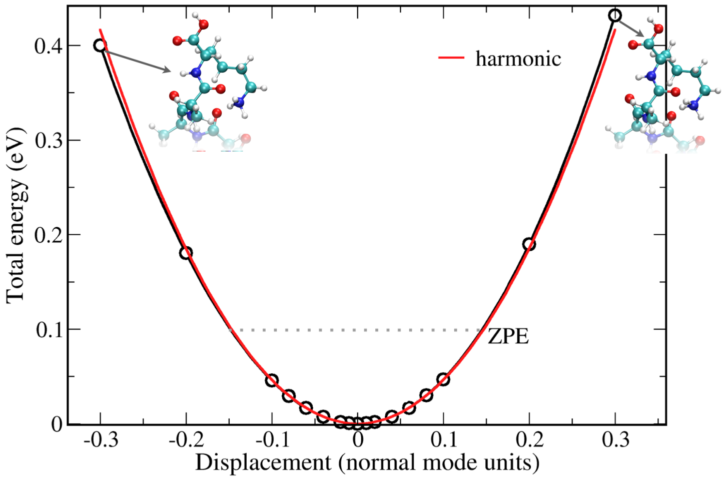


Figure F.3: Total energies with respect to mode displacement for the scissor vibrational mode of the NH_3^+ group. The bottom of the potential has been shifted to zero. The ZPE line corresponds to the zero-point vibrational energy for this mode. In red, a parabola fitted to the [-0.01, 0.01] region. The geometries of the molecule at maximum displacements are also shown.

Appendix G

Extra details on calculated and experimental IR spectra

Comparison between raw and smoothed experimental data

Figures G.1 - G.3 show a comparison between the raw experimental data and after smoothing via the 3-point formula of Eq. 9.5.

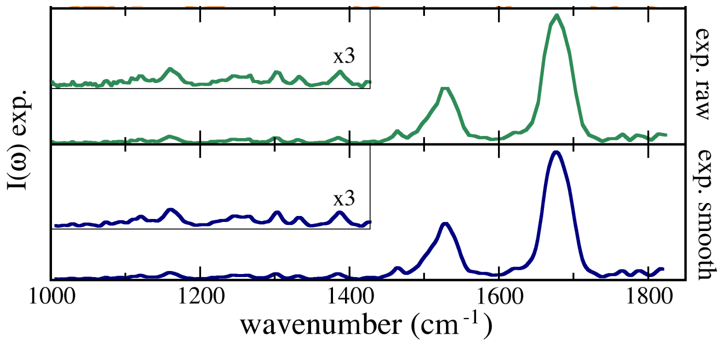


Figure G.1: Raw experimental data (top) compared to the smoothed (Eq. 9.5)experimental data (bottom) for Ac-Ala$_{15}$-LysH^{+}.

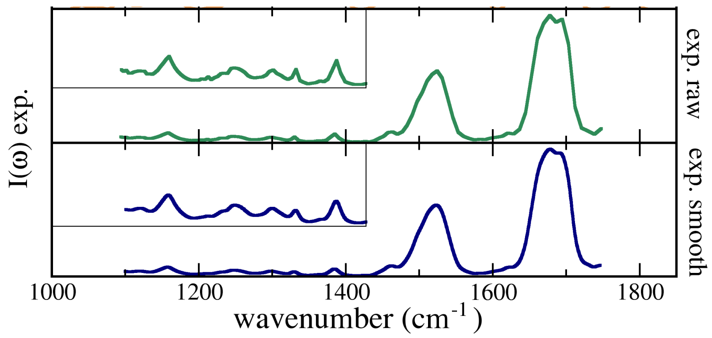


Figure G.2: Raw experimental data (top) compared to the smoothed (Eq. 9.5)experimental data (bottom) for Ac-Ala$_{10}$-LysH^{+}.

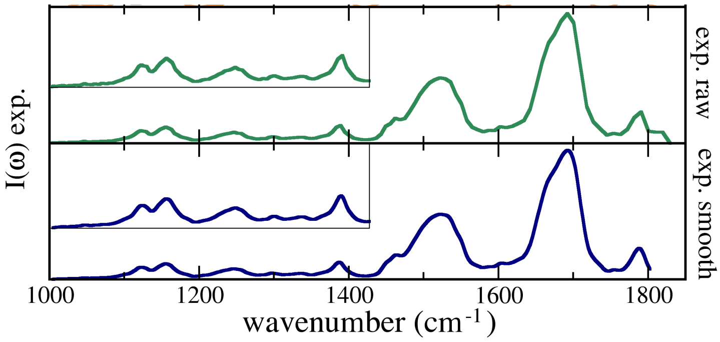


Figure G.3: Raw experimental data (top) compared to the smoothed (Eq. 9.5)experimental data (bottom) for Ac-Ala$_5$-LysH$^+$.

Detailed H-bond connection for the AIMD runs of Ac-Ala$_5$-LysH$^+$

Figures G.4 - G.7 show the detailed H-bond evolution in the AIMD run that led to the IR spectra shown in Figure 9.20. The H-bond connection of all backbone oxygens of the four chosen conformers of Ac-Ala$_5$-LysH$^+$ (g-1, α-1, α-2, and 3_{10}-1) is shown.

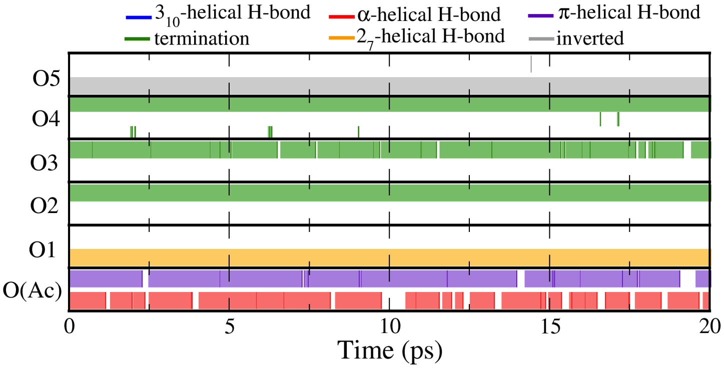


Figure G.4: Detailed H-bond network of the PBE+vdW AIMD simulation leading to the anharmonic spectrum of the g-1 conformer of Ac-Ala$_5$-LysH$^+$

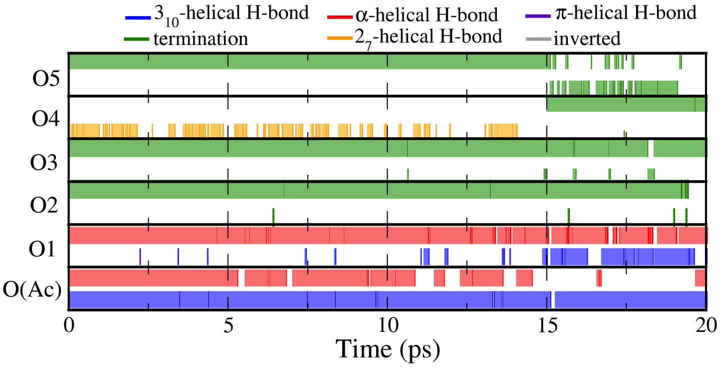


Figure G.5: Detailed H-bond network of the PBE+vdW AIMD simulation leading to the anharmonic spectrum of the α-1 conformer of Ac-Ala$_5$-LysH$^+$

Variation of R_P as a function of parameters

As an example of the variation of the R_P factor as a function of shift Δ of the theoretical spectrum with respect to experiment, in Figure G.8 this variation is plotted for Ac-Ala$_{10}$-LysH$^+$ and Ac-Ala$_{15}$-LysH$^+$. The data corresponds to the comparison between the anharmonic (AIMD derived) spectra of these molecules (α-helical) and the smoothed experimental data.

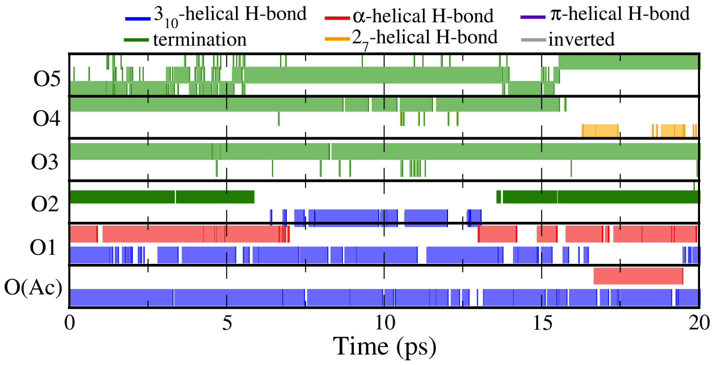


Figure G.6: Detailed H-bond network of the PBE+vdW AIMD simulation leading to the anharmonic spectrum of the α-2 conformer of Ac-Ala$_5$-LysH$^+$

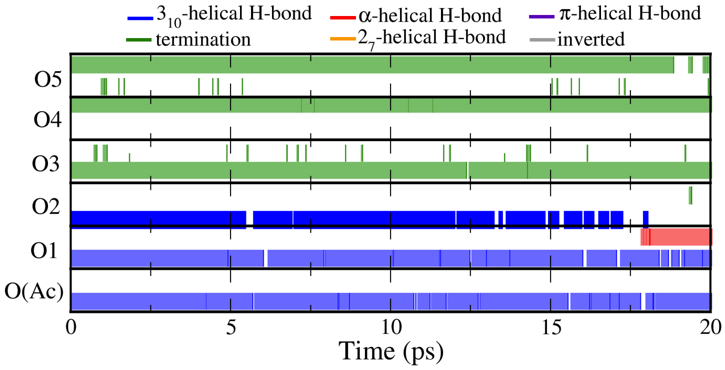
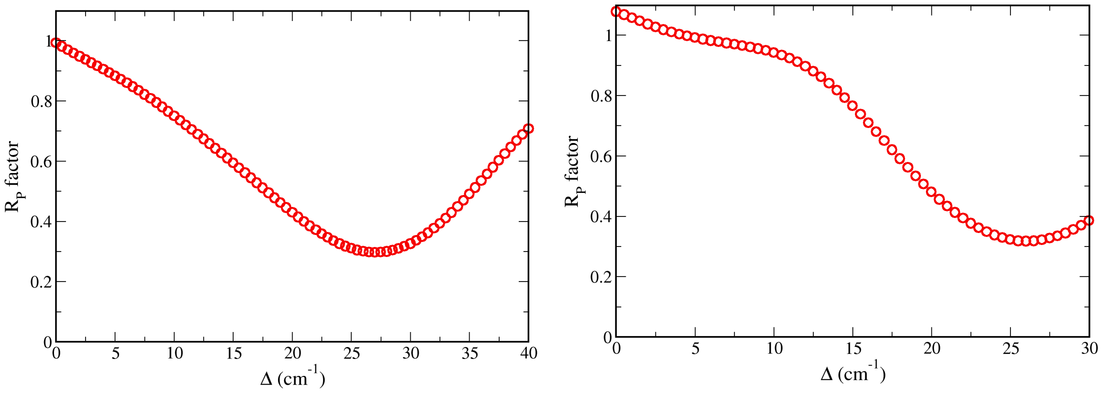


Figure G.7: Detailed H-bond network of the PBE+vdW AIMD simulation leading to the anharmonic spectrum of the 3_{10}-1 conformer of Ac-Ala$_5$-LysH$^+$

The variation of R_P as a function of the parameter W of Eq. 9.3 is shown in Figure G.9, for the harmonic and anharmonic IR spectra of Ac-Ala$_{15}$-LysH$^+$. As discussed in the main text, the reliability factor is not extremely sensitive to this parameter.



(a) Ac-Ala$_{10}$-LysH$^+$

(b) Ac-Ala$_{15}$-LysH$^+$

Figure G.8: Variation of R_P as a function of the shift Δ, for experiment versus anharmonic spectra of (a) Ac-Ala$_{10}$-LysH$^+$ and (b)Ac-Ala$_{15}$-LysH$^+$.

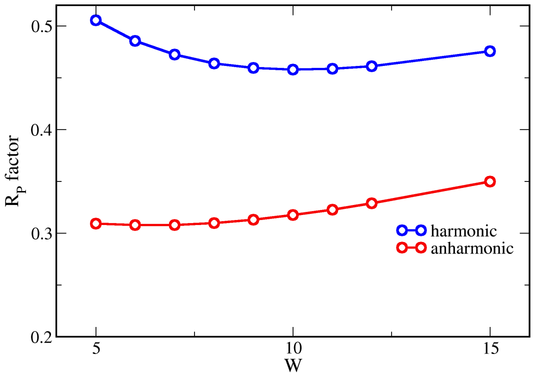


Figure G.9: Variation of R_P as a function of W, for experiment versus harmonic and anharmonic spectra of Ac-Ala$_{15}$-LysH$^+$.

CV

Personal data

Name:	Mariana Rossi Carvalho
Date of birth:	18.04.1983
Place of birth:	Campinas, SP, Brazil
Sex:	Female
Nationality:	Brazilian

Education

2007-2011	Ph.D. in Physics, Fritz Haber Institute of the Max Planck Society, Berlin, Germany
2005-2007	M. Sc. in Physics, Universidade de São Paulo (USP), São Paulo, Brazil
2001-2004	B. Sc. in Physics, Universidade de São Paulo (USP), São Paulo, Brazil

Published Papers

1. *Dispersion interactions with density-functional theory: Benchmarking semi-empirical and interatomic pair-wise corrected density functionals.* N. Marom, A. Tkatchenko, M. Rossi, V.V. Gobre, O. Hod, M. Scheffler, and L. Kronik, submitted to J. Chem. Theory Comput. (August 11, 2011)

2. *Unraveling the Stability of Polypeptide Helices: Critical Role of van der Waals Interactions*, A. Tkatchenko, M. Rossi, V. Blum, J. Ireta, and M. Scheffler, Phys. Rev. Lett. **106**, 118102 (2011)

3. *Secondary Structure of Ac-Ala$_n$-LysH$^+$ Polyalanine Peptides (n=5, 10, 15) in Vacuo: Helical or Not?*, M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, and M. Scheffler, J. Phys. Chem. Lett. **1**, 3465 (2010)

4. *Realistic calculations of carbon-based disordered systems*, A. Rocha, M. Rossi, A. J. R. da Silva, and A. Fazzio, J. Phys. D: Appl. Phys. **43**, 374002 (2010)

5. *Designing Real Nanotube-Based Gas Sensors*, A. Rocha, M. Rossi, A. Fazzio, and A. J. R. da Silva, Phys. Rev. Lett. **100**, 176803 (2008)

Acknowledgement

Any research work is not (or at least should not be) the work of a single person. So many other people contribute to it, sometimes quite directly, but also in many other indirect ways.

In this work, I must say that the most essential person was Volker Blum. Vielen Dank, Volker, wirklich! I have learned so much with you over all these years. People say I have even almost learned to give talks as fast as you! Jokes apart, having you as my direct supervisor in this work was really an extremely fruitful experience. If I would thank you for every single thing you helped me with for all these years, one page would not be enough. So I will just say thanks for all the support. I honestly hope (and I am also quite sure) that we stay in contact for still many years.

Then, I must thank Matthias Scheffler. The work in the theory department of the FHI for this last 4 (almost 5...) years has been anything but boring. The atmosphere is extremely challenging and the possibilities that I had to talk to so many experts in so many subjects has been certainly unique (not to talk about the easy access to amazingly good and powerful computers). I fear that when I get out of here I will be hopelessly spoiled by the quality and “possibilities” standards. Matthias, thank you for giving me the opportunity to write my Ph.D. here, and thank you also for so many good and wise advices on my research, my papers, and my presentations.

I also want to thank my “personal experts” with whom I have had contact here in the FHI and lots of fruitful discussions in all these years. Alex, muchas gracias! Your ideas have certainly contributed a whole lot to the work that I have written up here. I hope that our collaborations continue to give as many nice results as these ones. Joel, a ti tambien, muchas gracias! Without you it would have been so much harder to understand all this bio world. I find the discussions we have always quite enlightening. My colleagues from the (bio part of the) bio-group, Suchi, Matti, Carsten, Franziska, thank you all! Carsten and Franziska, a special thanks to you for correcting my Deutsch and helping with the translation of the Zusammenfassung, and also for reading many parts of the thesis. Luca, grazie for so many clarifications about the the “thermodynamical” and “statistical mechanical” world. Felix, Jörg, Michael, Erik, John, Viktor, Xinguo, Patrick, Eli, Hakim, Matteo, Marco, a big thank you for being supportive of my work! You know that many of you have also helped me a lot outside of the working environment, by sharing so many beers, cocktails, and parties!

Gert, Peter, and Frauke, our collaboration has been so great! It was probably the part of this work where I had more fun. Thanks for all the nice spectra and the great discussions.

To my spanish (and one polish!) friends Ania, Daggi, Eli, Isa, Pili, Núria, Rocio, gracias por todo. Chicas, without you it would have been hard not to go crazy over all this Ph.D.

I also cannot forget my brazilian friends, that like me, have been doing their Ph.D. in Europe in these past years. Maintaining contact with all of you has certainly been a great source of happiness. Bruno, Diogo, Gabriel, Sabrina, Mauricio, Priscila, Lucas, conseguimos! Terminamos, ou quase terminamos, o doutorado. A próxima fase dá um pouquinho mais de medo, não é? Bu, obrigada por todas as ajudinhas em python e tantas outras coisas.

My family, that is far far away, in Brazil. Their help was, of course, fundamental. I would never have gotten here in the first place (literally), if it were not for them. Mãe, vovó, Ju, pai, vocês foram tão lindos e especiais, e eu tenho tantas saudades. Ester, obrigada por mesmo longe continuar sendo essa prima amiga que eu gosto tanto e que me entende tão bem.

Fa, cosa ti posso dire? Ogni giorno, è tuo il mio sorriso. Adesso e per tanti altri anni, ovunque sia.

Bibliography

[1] V. Blum, R. Gehrke, F. Hanke, P. Havu, X. Ren, K. Reuter, and M. Scheffler, “Ab initio molecular simulations with numeric atom-centered orbitals,” *Comp. Phys. Comm.*, vol. 180, pp. 2175–2196, Jan 2009.

[2] A. Tkatchenko and M. Scheffler, “Accurate molecular van der waals interactions from ground-state electron density and free-atom reference data,” *Phys. Rev. Lett.*, vol. 102, p. 073005, 2009.

[3] M. Kohtani and M. F. Jarrold, “Water molecule adsorption on short alanine peptides: How short is the shortest gas-phase alanine-based helix?,” *Journal of the American Chemical Society*, vol. 126, no. 27, pp. 8454–8458, 2004.

[4] M. Rossi, V. Blum, P. Kupser, G. von Helden, F. Bierau, K. Pagel, G. Meijer, and M. Scheffler, “Secondary structure of ac-alan-lysh+ polyalanine peptides (n = 5,10,15) in vacuo: Helical or not?,” *The Journal of Physical Chemistry Letters*, vol. 1, no. 24, pp. 3465–3470, 2010.

[5] M. Kohtani, T. Jones, J. Schneider, and M. Jarrold, “Extreme stability of an unsolvated alpha-helix,” *J. Am. Chem. Soc.*, vol. 126, no. 24, pp. 7420–7421, 2004.

[6] A. Tkatchenko, M. Rossi, V. Blum, J. Ireta, and M. Scheffler, “Unraveling the stability of polypeptide helices: Critical role of van der waals interactions,” *Phys. Rev. Lett.*, vol. 106, p. 118102, 2011.

[7] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. Phillips, “A three-dimensional model of the myoglobin molecule obtained by x-ray analysis,” *Nature*, vol. 181, pp. 662–666, 1958.

[8] C. Anfinsen, “Principles that govern the folding of protein chains,” *Science*, vol. 181, pp. 223 – 230, 1973.

[9] C. Levinthal, “Are there pathways for protein folding?,” *Journal de Chimie Physique*, vol. 65, p. 44, 1968.

[10] M. Karplus, “The levinthal paradox: Yesterday and today,” *Folding and Design*, vol. 2, pp. S69 – S75, 1997.

[11] J. Bryngelson, H. N. Onuchic, N. D. Socci, and P. G. Wolynes, “Funnels, pathways, and the energy landscape of protein folding: A synthesis,” *Proteins: Structure, Function, and Genetics*, vol. 21, pp. 167 – 195, 1995.

[12] A. Bartlett and S. Radford, “An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms,” *Nature Structural and Molecular Biology*, vol. 16, pp. 582 – 588, 2009.

[13] C. L. Brooks, J. N. Onuchic, and D. J. Wales, “Taking a walk on a landscape,” *Science*, vol. 293, no. 5530, pp. 612–613, 2001.

[14] K. E. Riley, M. Pitonack, P. Jurecka, and P. Hobza, “Stabilization and structure calculations for noncovalent interactions in extended molecular systems based on wave function and density functional theories,” *Chemical Reviews*, vol. 110, no. 9, pp. 5023–5063, 2010.

[15] S. Grimme, “Density functional theory with london dispersion corrections,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 2, pp. 211–228, 2011.

[16] H. Taniuchi and C. B. Anfinsen, “An experimental approach to the study of the folding of staphylococcal nuclease,” *Journal of Biological Chemistry*, vol. 244, no. 14, pp. 3864–3875, 1969.

[17] J. E. Brown and W. A. Klee, “Helix-coil transition of the isolated amino terminus of ribonuclease,” *Biochemistry*, vol. 10, no. 3, pp. 470–476, 1971.

[18] W. SHEARER, R. BROWN, G. BRYCE, and F. GURD, “Reversible disruption by cupric ions of a helical conformation of a polypeptide derived from ribonuclease,” *J Biol Chem*, vol. 241, pp. 2665–&, Jan 1966.

[19] S. Marqusee and R. L. Baldwin, “Helix stabilization by glu-...lys+ salt bridges in short peptides of de novo design,” *Proceedings of the National Academy of Sciences*, vol. 84, no. 24, pp. 8898–8902, 1987.

[20] J. M. Scholtz, E. J. York, J. M. Stewart, and R. L. Baldwin, "A neutral, water-soluble, .alpha.-helical peptide: the effect of ionic strength on the helix-coil equilibrium," *Journal of the American Chemical Society*, vol. 113, no. 13, pp. 5102–5104, 1991.

[21] J. Vila, R. Williams, J. Grant, J. Wójcik, and H. Scheraga, "The intrinsic helix-forming tendency of l-alanine," *Proc. Natl. Am. Sc. U.S.A.*, vol. 89, p. 7821, 1992.

[22] K. Bolin and G. Millhauser, "α and 3(10): The split personality of polypeptide helices," *Accounts Chem Res*, vol. 32, no. 12, pp. 1027–1033, 1999. and references therein.

[23] M. Jarrold, "Peptides and proteins in the vapor phase," *Annual Review of Physical Chemistry*, vol. 51, pp. 179–207, 2000.

[24] T. R. Rizzo, J. A. Stearns, and O. V. Boyarkin, "Spectroscopic studies of cold, gas-phase biomolecular ions," *Int Rev Phys Chem*, vol. 28, no. 3, pp. 481–515, 2009.

[25] P. N. Mortenson, D. A. Evans, and D. J. Wales, "Energy landscapes of model polyalanines," *The Journal of Chemical Physics*, vol. 117, no. 3, pp. 1363–1376, 2002.

[26] D. Wales, "Energy landscapes and properties of biomolecules," 2005.

[27] D. Wales and B. Strodel, "Free energy surfaces from an extended harmonic superposition approach and kinetics for alanine dipeptide," *Chem Phys Lett*, vol. 466, pp. 105–115, 2008.

[28] S. Gnanakaran and A. E. Garcia, "Helix-coil transition of alanine peptides in water: Force field dependence on the folded and unfolded structures," *Proteins*, vol. 59, no. 4, pp. 773–782, 2005.

[29] D. Paschek, M. Puehse, A. Perez-Goicochea, S. Gnanakaran, A. E. Garcia, R. Winter, and A. Geiger, "The solvent-dependent shift of the amide i band of a fully solvated peptide as a local probe for the solvent composition in the peptide/solvent interface," *Chem Phys Chem*, vol. 9, no. 18, pp. 2742–2750, 2008.

[30] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, "Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides," *J. Phys. Chem. B*, vol. 105, pp. 6474–6487, 2001.

[31] J. Tirado-Rives, D. S. Maxwell, and W. L. Jorgensen, "Molecular dynamics and monte carlo simulations favor the α-helical form for alanine-based peptides in water," *J. Am. Chem. Soc.*, vol. 115, pp. 11590–11593, 1993.

[32] E. Penev, J. Ireta, and J.-E. Shea, "Energetics of infinite homopolypeptide chains: A new look at commonly used force fields," *The Journal of Physical Chemistry B*, vol. 112, no. 22, pp. 6872–6877, 2008.

[33] P. Salvador, A. Asensio, and J. J. Dannenberg, "The effect of aqueous solvation upon alpha-helix formation for polyalanines," *J Phys Chem B*, vol. 111, no. 25, pp. 7462–7466, 2007.

[34] J. Kubelka, R. Huang, and T. A. Keiderling, "Solvent effects on ir and vcd spectra of helical peptides: Dft-based static spectral simulations with explicit water," *The Journal of Physical Chemistry B*, vol. 109, no. 16, pp. 8231–8243, 2005.

[35] J. Ireta, J. Neugebauer, M. Scheffler, A. Rojo, and M. Galvan, "Structural transitions in the polyalanine alpha-helix under uniaxial strain," *J. Am. Chem. Phys.*, vol. 127, p. 17241, 2005.

[36] A. Cimas, T. D. Vaden, T. S. J. A. de Boer, L. C. Snoek, and M.-P. Gaigeot, "Vibrational spectra of small protonated peptides from finite temperature md simulations and irmpd spectroscopy," *Journal of Chemical Theory and Computation*, vol. 5, no. 4, pp. 1068–1078, 2009.

[37] R. R. Hudgins, M. A. Ratner, and M. F. Jarrold, "Design of helices that are stable in vacuo," *Journal of the American Chemical Society*, vol. 120, no. 49, pp. 12974–12975, 1998.

[38] P. Kupser, *Infrared spectroscopic characterization of secondary structure elements of gas-phase biomolecules*. PhD thesis, Fritz-Haber-Institut der Max-Planck-Gesellschaft and Freie Universität Berlin, 2011.

[39] J. A. Stearns, C. Seaiby, O. V. Boyarkin, and T. R. Rizzo, "Spectroscopy and Conformational Preferences of Gas-Phase Helices," *Physical Chemistry Chemical Physics*, vol. 11, p. 125, 2009.

[40] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000. http://www.pdb.org.

[41] C. A. Opitz, M. Kulke, M. C. Leake, C. Neagoe, H. Hinssen, R. J. Hajjar, and W. A. Linke, "Damped elastic recoil of the titin spring in myofibrils of human myocardium," *Proceedings of the National Academy of Sciences*, vol. 100, no. 22, pp. 12688–12693, 2003.

[42] R. B. Corey and L. Pauling, "Molecular models of amino acids, peptides, and proteins," *Review of Scientific Instruments*, vol. 24, no. 8, pp. 621–627, 1953.

[43] S. Nanita and R. Cooks, "Serine octamers: Cluster formation, reactions, and implications for biomolecule homochirality," *Angew Chem Int Edit*, vol. 45, no. 4, pp. 554–569, 2006.

[44] D. Voet and J. G. Voet, *Biochemistry*. Wiley, 4th ed., 2011.

[45] T. Head-Gordon, M. Head-Gordon, M. J. Frisch, C. L. Brooks, and J. A. Pople, "Theoretical study of blocked glycine and alanine peptide analogs," *Journal of the American Chemical Society*, vol. 113, no. 16, pp. 5989–5997, 1991.

[46] L. Pauling, R. Corey, and H. Branson, "The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain.," *Proc. Natl. Acad. Sci. USA*, vol. 37, pp. 205–211, 1951.

[47] L. Pauling and R. Corey, "The pleated sheet, a new layer configuration of polypeptide chains," *Proc. Natl. Acad. Sci. USA*, vol. 37, p. 251256, 1951.

[48] D. Eisenberg, "The discovery of the α-helix and β-sheet, the principal structural features of proteins," *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 11207–11210, 2003.

[49] K. U. L. m Lang, "Proteins and enzymes," *Lane Medical Lectures, Stanford University Publications*, vol. 6, 1952.

[50] R. L. Baldwin, "Energetics of protein folding," *J. Mol. Biol.*, vol. 371, p. 283, 2007.

[51] L. Pauling, *The Nature of the Chemical Bond*. Cornell University Press, 3rd. ed., 1960.

[52] F. London, "Zur theorie und systematik der molekularkräfte," *Zeitschrift für Physik A Hadrons and Nuclei*, vol. 63, pp. 245–279, 1930. 10.1007/BF01421741.

[53] S. J. Grabowski, "What is the covalency of hydrogen bonding?," *Chemical Reviews*, vol. 111, no. 4, pp. 2597–2625, 2011.

[54] P. Salvador, N. Kobko, R. Wieczorek, and J. J. Dannenberg, "Calculation of trans-hydrogen-bond 13c-15n three-bond and other scalar j-couplings in cooperative peptide models. a density functional theory study," *Journal of the American Chemical Society*, vol. 126, no. 43, pp. 14190–14197, 2004. PMID: 15506785.

[55] J. Ireta, J. Neugebauer, M. Scheffler, A. Rojo, and M. Galván, "Density functional theory study of the cooperativity of hydrogen bonds in finite and infinite α-helices," *The Journal of Physical Chemistry B*, vol. 107, no. 6, pp. 1432–1437, 2003.

[56] P. T. Van Duijnen and B. T. Thole, "Cooperative effects in α-helices: An *ab initio* molecular-orbital study," *Biopolymers*, vol. 21, no. 9, pp. 1749–1761, 1982.

[57] J. Cruzan, L. Braly, K. Liu, M. Brown, J. Loeser, and R. Saykally, “Quantifying hydrogen bond cooperativity in water: Vrt spectroscopy of the water tetramer,” *Science*, vol. 271, no. 5245, pp. 59–62, 1996.

[58] M. Elrod and R. Saykally, “Many-body effects in intermolecular forces,” 1994.

[59] S. Xantheas, “Cooperativity and hydrogen bonding network in water clusters,” *Chemical Physics*, vol. 258, no. 2-3, pp. 225–231, 2000.

[60] S. Suhai, “Cooperativity and electron correlation effects on hydrogen bonding in infinite systems,” *Int. J. Quantum Chem.*, vol. 52, pp. 395–412, 1994.

[61] J. J. Dannenberg, “Cooperativity in hydrogen bonded aggregates. models for crystals and peptides,” *Journal of Molecular Structure*, vol. 615, no. 1-3, pp. 219 – 226, 2002.

[62] P. Hobza, R. Zahradník, and K. Müller-Dethlefs, “The world of non-covalent interactions,” *Collect. Czech. Chem. Commun.*, vol. 71, no. 4, pp. 443–531, 2006.

[63] B. L. Sibanda and J. M. Thornton, “β-hairpin families in globular proteins,” *Nature*, vol. 316, pp. 170–174, July 1985.

[64] K. Möhle, M. Gußmann, and H. Hofmann, “Structural and energetic relations between β-turns,” *Journal of Computational Chemistry*, vol. 18, no. 11, pp. 1415–1430, 1997.

[65] C. Venkatachalam, “Stereochemical criteria for polypeptides and proteins .v. conformation of a system of 3 linked peptide units,” *Biopolymers*, vol. 6, no. 10, p. 1425, 1968.

[66] E. G. Hutchinson and J. M. Thornton, “A revised set of potentials for β-turn formation in proteins,” *Protein Science*, vol. 3, pp. 2207–2216, Dec. 1994.

[67] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” *Journal of Molecular Biology*, vol. 7, no. 1, pp. 95 – 99, 1963.

[68] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. D. Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, “Structure validation by cα geometry: ϕ, ψ and cβ deviation,” *Proteins*, vol. 50, no. 3, pp. 437–450, 2003.

[69] B. K. Ho, A. Thomas, and R. Brasseur, “Revisiting the ramachandran plot: Hard-sphere repulsion, electrostatics, and h-bonding in the -helix,” *Protein Science*, vol. 12, pp. 2508–2522, Jan 2003.

[70] J. Ireta and M. Scheffler, “Density functional theory study of the conformational space of an infinitely long polypeptide chain,” *J. Chem. Phys.*, vol. 131, no. 8, p. 085104, 2009.

[71] N. Fitzkee and G. Rose, “Reassessing random-coil statistics in unfolded proteins,” *P Natl Acad Sci Usa*, vol. 101, no. 34, pp. 12497–12502, 2004.

[72] P. Palen**c**ár and T. Bleha, “Molecular dynamics simulations of the folding of poly(alanine) peptides,” *J Mol Model*, Mar 2011.

[73] J. A. Stearns, O. V. Boyarkin, and T. R. Rizzo, “Spectroscopic signatures of gas-phase helices: Ac-phe-(ala)5-lys-h+ and ac-phe-(ala)10-lys-h+,” *Journal of the American Chemical Society*, vol. 129, no. 45, pp. 13820–13821, 2007.

[74] J. P. M. Lommerse, S. L. Price, and R. Taylor, “Hydrogen bonding of carbonyl, ether, and ester oxygen atoms with alkanol hydroxyl groups,” *Journal of Computational Chemistry*, vol. 18, no. 6, pp. 757–774, 1997.

[75] V. Pande, “Folding@home distributed computing,” 2011.

[76] R. Zwanzig, A. Szabo, and B. Bagchi, "Levinthal's paradox," *Proc. Natl. Acad. Sci. USA*, vol. 89, pp. 20 – 22, 1992.

[77] D. Brockwell and S. Radford, "Intermediates: ubiquitous species on folding energy landscapes?," *Current Opinion in Structure Biology*, vol. 17, pp. 30 – 37, 2007.

[78] C. M. Dobson, "Protein-misfolding diseases: Getting out of shape," *Nature*, vol. 418, pp. 729–730, August 2002.

[79] F. Chiti and C. M. Dobson, "Amyloid formation by globular proteins under native conditions," *Nature Chemical Biology*, vol. 5, pp. 15–22, January 2009.

[80] S. B. Prusiner, "Prions," *Proceedings of the National Academy of Sciences*, vol. 95, no. 23, pp. 13363–13383, 1998.

[81] M. Grabenauer, T. Wyttenbach, N. Sanghera, S. E. Slade, T. J. T. Pinheiro, J. H. Scrivens, and M. T. Bowers, "Conformational stability of syrian hamster prion protein prp(90231)," *Journal of the American Chemical Society*, vol. 132, no. 26, pp. 8816–8818, 2010.

[82] A. Möglich, K. Joder, and T. Kiefhaber, "End-to-end distance distributions and intrachain diffusion constants in unfolded polypeptide chains indicate intramolecular hydrogen bond formation," *Proceedings of the National Academy of Sciences*, vol. 103, no. 33, pp. 12394–12399, 2006. Correction: PNAS **105**, pp. 6787 (2008).

[83] P. Privalov, "Cold denaturation of protein," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 25, no. 5, pp. 281–306, 1990.

[84] A. Pastore, S. R. Martin, A. Politou, K. C. Kondapalli, T. Stemmler, and P. A. Temussi, "Unbiased cold denaturation: Low- and high-temperature unfolding of yeast frataxin under physiological conditions," *Journal of the American Chemical Society*, vol. 129, no. 17, pp. 5374–5375, 2007. PMID: 17411056.

[85] D. J. Wales, *Energy Landscapes*. Cambridge University Press, 1st ed., 2003.

[86] A. Szabo and N. Ostlund, *Modern Quantum Chemistry*. Dover, 1st, revised. ed., 1996.

[87] E. K. U. Gross and R. Dreizler, *Density Functional Theory*. Springer Verlag, Berlin, 1st ed., 1995.

[88] R. Mattuck, *A Guide to Feynman Diagrams in the Many-Body Problem*. Dover, 1st edition ed., 1992.

[89] I. G. Kaplan, *Intermolecular Interactions: Physical Picture, Computational Methods and Model Potentials*. John Wiley & Sons Ltd., 1st ed., 2006.

[90] C. Fiolhais, F. Nogueira, and M. Marques, *A Primer in Density Functional Theory*. Springer Verlag, Berlin, 1st ed., 2003.

[91] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society*, vol. 118, no. 45, pp. 11225–11236, 1996.

[92] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *Journal of the American Chemical Society*, vol. 117, no. 19, pp. 5179–5197, 1995.

[93] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[94] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, “Current status of the amoeba polarizable force field,” *The Journal of Physical Chemistry B*, vol. 114, no. 8, pp. 2549–2564, 2010. PMID: 20136072.

[95] M. Beachy, D. Chasman, R. Murphy, T. Halgren, and R. Friesner, “Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields,” *J Am Chem Soc*, vol. 119, no. 25, pp. 5908–5920, 1997.

[96] M. Kolar, K. Berka, P. Jurecka, and P. Hobza, “On the reliability of the amber force field and its empirical dispersion contribution for the description of noncovalent complexes,” *ChemPhysChem*, vol. 11, no. 11, pp. 2399–2408, 2010.

[97] M. Born and J. Oppenheimer, “Zur quantentheorie der moleküle,” *Ann. Phys. Leipzig*, vol. 84, pp. 457–484, 1927.

[98] D. R. Hartree, “The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, pp. 89–110, 1928.

[99] V. Fock, “Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems,” *Zeitschrift für Physik*, vol. 61, pp. 126–148, 1930.

[100] T. Koopmans, “Ueber die zuordnung von wellenfunktionen und eigenwerten zu den einzelnen elektronen eines atoms,” *Physica*, vol. 1, no. 1-6, pp. 104 – 113, 1934.

[101] E. Schwegler and M. Challacombe, “Linear scaling computation of the hartree–fock exchange matrix,” *The Journal of Chemical Physics*, vol. 105, no. 7, pp. 2726–2734, 1996.

[102] C. Møller and M. S. Plesset, “Note on an approximation treatment for many-electron systems,” *Phys. Rev.*, vol. 46, pp. 618–622, Oct 1934.

[103] R. J. Bartlett and M. Musial, “Coupled-cluster theory in quantum chemistry,” *Rev Mod Phys*, vol. 79, no. 1, pp. 291–352, 2007.

[104] F. Coester and H. Kummel, “Short-range correlations in nuclear wave functions,” *Nucl Phys*, vol. 17, no. 3, pp. 477–485, 1960.

[105] J. CIZEK, “On correlation problem in atomic and molecular systems . calculation of wavefunction components in ursell-type expansion using quantum-field theoretical methods,” *J Chem Phys*, vol. 45, no. 11, pp. 4256–&, 1966.

[106] L. H. Thomas, “The calculation of atomic fields,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 23, no. 05, pp. 542–548, 1927.

[107] E. Fermi, “Un metodo statistico per la determinazione di alcune priopriettà dell’atomo,” *Rend. Accad. Naz. Lincei*, vol. 6, pp. 602–607, 1927.

[108] J. C. Slater, “A simplification of the hartree-fock method,” *Phys. Rev.*, vol. 81, pp. 385–390, 1951.

[109] P. Hohenberg and W. Kohn, “Inhomogeneous electron gas,” *Phys. Rev.*, vol. 136, no. 3B, pp. B864–B871, 1964.

[110] M. Levy, “Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem,” *Proceedings of the National Academy of Sciences*, vol. 76, no. 12, pp. 6062–6065, 1979.

[111] W. Kohn, “v-representability and density functional theory,” *Phys. Rev. Lett.*, vol. 51, pp. 1596–1598, Oct 1983.

[112] J. Chayes, L. Chayes, and M. Ruskai, “Density functional approach to quantum lattice systems,” *Journal of Statistical Physics*, vol. 38, pp. 497–518, 1985.

[113] W. Kohn and L. J. Sham, “Self-consistent equations including exchange and correlation effects,” *Phys. Rev.*, vol. 140, no. 4A, pp. A1133–A1138, 1965.

[114] L. Curtiss, P. Redfern, K. Raghavachari, and J. Pople, “Assessment of gaussian-2 and density functional theories for the computation of ionization potentials and electron affinities,” *J Chem Phys*, vol. 109, no. 1, pp. 42–55, 1998.

[115] U. von Barth, “Basic density-functional theory-an overview,” *Physica Scripta*, vol. 2004, no. T109, p. 9, 2004.

[116] K. Burke, “The abc of dft,” 2010.

[117] J. P. Perdew and K. Schmidt, “Jacob’s ladder of density functional approximations for the exchange-correlation energy,” *AIP Conference Proceedings*, vol. 577, no. 1, pp. 1–20, 2001.

[118] C. Dykstra, G. Frenking, K. Kim, and G. Scuseria, *Theory and Applications of Computational Chemistry: The First 40 Years*. Elsevier, Amsterdam, 1st ed., 2005.

[119] M. Gell-Mann and K. A. Brueckner, “Correlation energy of an electron gas at high density,” *Phys. Rev.*, vol. 106, p. 364, 1957.

[120] D. M. Ceperley and B. J. Alder, “Ground state of the electron gas by a stochastic method,” *Phys. Rev. Lett.*, vol. 45, pp. 566–569, Aug 1980.

[121] J. P. Perdew, E. R. McMullen, and A. Zunger, “Density-functional theory of the correlation energy in atoms and ions: A simple analytic model and a challenge,” *Phys. Rev. A*, vol. 23, pp. 2785–2789, Jun 1981.

[122] 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*, vol. 58, no. 8, pp. 1200–1211, 1980.

[123] J. P. Perdew and Y. Wang, “Accurate and simple analytic representation of the electron-gas correlation energy,” *Phys. Rev. B*, vol. 45, pp. 13244–13249, Jun 1992.

[124] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebegue, J. Paier, O. A. Vydrov, and J. G. Angyan, “Assessing the performance of recent density functionals for bulk solids,” *Phys Rev B*, vol. 79, no. 15, p. 155107, 2009.

[125] J. Harl, L. Schimka, and G. Kresse, “Assessing the quality of the random phase approximation for lattice constants and atomization energies of solids,” *Phys Rev B*, vol. 81, no. 11, p. 115126, 2010.

[126] J. Perdew, R. Parr, M. Levy, and J. Balduz, “Density-functional theory for fractional particle number - derivative discontinuities of the energy,” *Physical Review Letters*, vol. 49, no. 23, pp. 1691–1694, 1982.

[127] S.-K. Ma and K. A. Brueckner, “Correlation energy of an electron gas with a slowly varying high density,” *Phys. Rev.*, vol. 165, pp. 18–31, Jan 1968.

[128] P. Haas, F. Tran, and P. Blaha, “Calculation of the lattice constant of solids with semilocal functionals,” *Phys. Rev. B*, vol. 79, p. 085104, Feb 2009.

[129] P. Haas, F. Tran, P. Blaha, K. Schwarz, and R. Laskowski, “Insight into the performance of gga functionals for solid-state calculations,” *Phys. Rev. B*, vol. 80, p. 195109, Nov 2009.

[130] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin, and J. Sun, “Erratum: Workhorse semilocal density functional for condensed matter physics and quantum chemistry [phys. rev. lett. 103, 026403 (2009)],” *Phys. Rev. Lett.*, vol. 106, p. 179902, Apr 2011.

[131] J. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters*, vol. 77, no. 18, pp. 3865–3868, 1996.

[132] J. P. Perdew and Y. Wang, "Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation," *Phys. Rev. B*, vol. 33, pp. 8800–8802, Jun 1986.

[133] J. P. Perdew, *Electronic Structure of Solids.* Akademie Verlag, Berlin, 1991.

[134] Y. Zhang and W. Yang, "Comment on "generalized gradient approximation made simple"," *Phys. Rev. Lett.*, vol. 80, p. 890, Jan 1998.

[135] B. Hammer, L. B. Hansen, and J. K. Nørskov, "Improved adsorption energetics within density-functional theory using revised perdew-burke-ernzerhof functionals," *Phys. Rev. B*, vol. 59, pp. 7413–7421, Mar 1999.

[136] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, "Restoring the density-gradient expansion for exchange in solids and surfaces," *Phys. Rev. Lett.*, vol. 100, p. 136406, Apr 2008.

[137] A. D. Becke *Journal of Chemical Physics*, vol. 88, p. 1053, 1988.

[138] A. D. Becke, "Density functional calculations of molecular bond energies," *The Journal of Chemical Physics*, vol. 84, no. 8, pp. 4524–4529, 1986.

[139] C. Lee, W. Yang, and R. Parr *Physical Review B*, vol. 37, p. 785, 1988.

[140] S. Kurth, J. Perdew, and P. Blaha, "Molecular and solid-state tests of density functional approximations: Lsd, ggas, and meta-ggas," *Int J Quantum Chem*, vol. 75, no. 4-5, pp. 889–909, 1999.

[141] R. Armiento and A. E. Mattsson, "Functional designed to include surface effects in self-consistent density functional theory," *Phys. Rev. B*, vol. 72, p. 085108, Aug 2005.

[142] A. E. Mattsson and R. Armiento, "Implementing and testing the am05 spin density functional," *Phys. Rev. B*, vol. 79, p. 155101, Apr 2009.

[143] W. Kohn and A. E. Mattsson, "Edge electron gas," *Phys. Rev. Lett.*, vol. 81, pp. 3487–3490, Oct 1998.

[144] M. Ernzerhof and G. E. Scuseria *Journal of Chemical Physics*, vol. 110, p. 5029, 1999.

[145] C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The pbe0 model," *The Journal of Chemical Physics*, vol. 110, no. 13, pp. 6158–6170, 1999.

[146] J. P. Perdew, M. Ernzerhof, and K. Burke, "Rationale for mixing exact exchange with density functional approximations," *The Journal of Chemical Physics*, vol. 105, no. 22, pp. 9982–9985, 1996.

[147] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields," *The Journal of Physical Chemistry*, vol. 98, no. 45, pp. 11623–11627, 1994.

[148] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko,

P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, "Gaussian 03, Revision C.02." Gaussian, Inc., Wallingford, CT, 2004.

[149] R. H. Hertwig and W. Koch, "On the parameterization of the local correlation functional. what is becke-3-lyp?," *Chemical Physics Letters*, vol. 268, no. 5-6, pp. 345 – 351, 1997.

[150] A. V. Arbuznikov and M. Kaupp, "The self-consistent implementation of exchange-correlation functionals depending on the local kinetic energy density," *Chemical Physics Letters*, vol. 381, no. 3-4, pp. 495 – 504, 2003.

[151] J. P. Perdew and L. A. Constantin, "Laplacian-level density functionals for the kinetic energy density and exchange-correlation energy," *Phys Rev B*, vol. 75, no. 15, p. 155109, 2007.

[152] C. Adamo, M. Ernzerhof, and G. E. Scuseria, "The meta-gga functional: Thermochemistry with a kinetic energy density dependent exchange-correlation functional," *The Journal of Chemical Physics*, vol. 112, no. 6, pp. 2643–2649, 2000.

[153] J. P. Perdew, S. Kurth, A. c. v. Zupan, and P. Blaha, "Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation," *Phys. Rev. Lett.*, vol. 82, pp. 2544–2547, Mar 1999.

[154] J. P. Perdew, S. Kurth, A. c. v. Zupan, and P. Blaha, "Accurate density functional with correct formal properties: A step beyond the generalized gradient approximation," *Phys. Rev. Lett.*, vol. 82, pp. 2544–2547, Mar 1999.

[155] J. Tao, J. P. Perdew, V. N. Staroverov, and G. E. Scuseria, "Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids," *Phys. Rev. Lett.*, vol. 91, p. 146401, Sep 2003.

[156] Y. Zhao, N. E. Schultz, and D. G. Truhlar, "Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions," *The Journal of Chemical Physics*, vol. 123, no. 16, p. 161103, 2005.

[157] Y. Zhao and D. G. Truhlar, "A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions," *The Journal of Chemical Physics*, vol. 125, no. 19, p. 194101, 2006.

[158] Y. Zhao and D. Truhlar, "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 (Theoretica Chimica Acta)*, vol. 120, pp. 215–241, 2008. 10.1007/s00214-007-0310-x.

[159] Y. Zhao and D. G. Truhlar, "Exploring the limit of accuracy of the global hybrid meta density functional for main-group thermochemistry, kinetics, and noncovalent interactions," *Journal of Chemical Theory and Computation*, vol. 4, no. 11, pp. 1849–1868, 2008.

[160] J. Vondrá**s**ek, L. Bendová, , V. Klusák, and P. Hobza, "Unexpectedly strong energy stabilization inside the hydrophobic core of small protein rubredoxin mediated by aromatic residues: correlated ab initio quantum chemical calculations," *J Am Chem Soc*, vol. 127, pp. 2615–2619, 2005.

[161] H. B. G. Casimir and D. Polder, "The influence of retardation on the london-van der waals forces," *Phys. Rev.*, vol. 73, pp. 360–372, Feb 1948.

[162] A. W. Rodriguez, F. Capasso, and S. G. Johnson, "The casimir effect in microstructured geometries," *Nature Photonics*, vol. 5, pp. 211–221, Apr 2011.

[163] T. Janowski and P. Pulay, “High accuracy benchmark calculations on the benzene dimer potential energy surface,” *Chemical Physics Letters*, vol. 447, pp. 27–32, Jan 2007.

[164] T. van Mourik and R. J. Gdanitz, “A critical note on density functional theory studies on rare-gas dimers,” *00219606*, vol. 116, no. 22, p. 9620, 2002.

[165] A. Tkatchenko, J. Robert A. DiStasio, M. Head-Gordon, and M. Scheffler, “Dispersion-corrected møller–plesset second-order perturbation theory,” *The Journal of Chemical Physics*, vol. 131, no. 9, p. 094106, 2009.

[166] K. T. Tang, “Dynamic polarizabilities and van der waals coefficients,” *Phys. Rev.*, vol. 177, pp. 108–114, Jan 1969.

[167] J. Hepburn and G. Scoles *Chemical Physics Letters*, vol. 36, p. 451, 1975.

[168] Q. Wu and W. Yang, “Empirical correction to density functional theory for van der waals interactions,” *Journal of Chemical Physics*, vol. 116, p. 515, 2002.

[169] M. Elstner, P. Hobza, T. Frauenheim, S. Suhai, and E. Kaxiras *Journal of Chemical Physics*, vol. 114, p. 5149, 2001.

[170] S. Grimme *J. Comp. Chem.*, vol. 25, p. 1463, 2004.

[171] S. Grimme *J. Comp. Chem.*, vol. 27, p. 1787, 2006.

[172] P. Jurecka, J. Cerny, P. Hobza, and D. R. Salahub *J. Comp. Chem.*, vol. 28, p. 555, 2007.

[173] F. Ortmann, F. Bechstedt, and W. G. Schmidt, “Semiempirical van der waals correction to the density functional description of solids and molecular structures,” *Phys Rev B*, vol. 73, p. 205101, Jan 2006.

[174] A. BECKE and E. Johnson, “A density-functional model of the dispersion interaction,” *J Chem Phys*, vol. 123, no. 15, p. 154101, 2005.

[175] F. O. Kannemann and A. D. Becke, “van der waals interactions in density-functional theory: Intermolecular complexes,” *Journal of Chemical Theory and Computation*, vol. 6, no. 4, pp. 1081–1088, 2010.

[176] F. L. Hirshfeld, “Bonded-atom fragments for describing molecular charge densities,” *Theoretica Chimica Acta*, vol. 44, pp. 129–138, 1977.

[177] X. Chu and A. Dalgarno, “Linear response time-dependent density functional theory for van der waals coefficients,” *The Journal of Chemical Physics*, vol. 121, no. 9, pp. 4083–4088, 2004.

[178] P. Jurecka, J. Sponer, J. Cerný, and P. Hobza, “Benchmark database of accurate (mp2 and ccsd(t) complete basis set limit) interaction energies of small model complexes, dna base pairs, and amino acid pairs,” *Phys. Chem. Chem. Phys.*, vol. 8, p. 1985, Jan 2006.

[179] Y. Andersson, D. C. Langreth, and B. I. Lundqvist, “van der waals interactions in density-functional theory,” *Phys. Rev. Lett.*, vol. 76, pp. 102–105, Jan 1996.

[180] M. Dion, H. Rydberg, E. Schroder, D. Langreth, and B. Lundqvist, “Van der waals density functional for general geometries,” *Physical review letters*, vol. 92, no. 24, p. 246401, 2004.

[181] O. A. Vydrov, Q. Wu, and T. V. Voorhis *J Chem Phys*, vol. 129, no. 1, p. 014106, 2008.

[182] K. Lee, E. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, “Higher-accuracy van der waals density functional,” *Phys Rev B*, vol. 82, no. 8, p. 081101, 2010.

[183] D. Pines and D. Bohm, “A collective description of electron interactions: Ii. collective vs individual particle aspects of the interactions,” *Phys. Rev.*, vol. 85, pp. 338–353, 1952.

[184] D. Bohm and D. Pines, “A collective description of electron interactins: Iii. coulomb interaction in a degenerate electron gas,” *Phys. Rev.*, vol. 92, p. 609, 1953.

[185] J. Harris and A. Griffin, “Correlation energy and van der waals interaction of coupled metal films,” *Phys. Rev. B*, vol. 10, p. 3669, 1975.

[186] D. C. Langreth and J. P. Perdew, “Exchange-correlation energy of a metal surface: Wave-vector analysis,” *Phys. Rev. B*, vol. 15, p. 2884, 1977.

[187] O. Gunnarsson and B. I. Lundqvist, “Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism,” *Phys. Rev. B*, vol. 13, p. 4274, 1976.

[188] S. L. Adler, “Quantum theory of the dielectric constant in real solids,” *Phys. Rev.*, vol. 126, p. 413, 1962.

[189] N. Wiser, “Dielectric constant with local field effect included,” *Phys. Rev.*, vol. 129, p. 62, 1963.

[190] X. Ren, A. Tkatchenko, P. Rinke, and M. Scheffler, “Beyond the random-phase approximation for the electron correlation energy: The importance of single excitations,” *Phys. Rev. Lett.*, vol. 106, p. 153003, Apr 2011.

[191] B. Santra, *Density-Functional Theory Exchange-Correlation Functionals for Hydrogen Bonds in Water*. PhD thesis, Fritz-Haber-Institut der Max-Planck-Gesellschaft and Technische Universität Berlin, 2010.

[192] R. Vargas, J. Garza, B. P. Hay, and D. A. Dixon, “Conformational study of the alanine dipeptide at the mp2 and dft levels,” *The Journal of Physical Chemistry A*, vol. 106, no. 13, pp. 3213–3218, 2002.

[193] A. Gruneis, M. Marsman, and G. Kresse, “Second-order m[o-slash]ller–plesset perturbation theory applied to extended systems. ii. structural and energetic properties,” *The Journal of Chemical Physics*, vol. 133, no. 7, 2010.

[194] B. Santra, A. Michaelides, and M. Scheffler, “On the accuracy of density-functional theory exchange-correlation functionals for h bonds in small water clusters: Benchmarks approaching the complete basis set limit,” *J Chem Phys*, vol. 127, no. 18, p. 184104, 2007.

[195] B. Santra, A. Michaelides, M. Fuchs, A. Tkatchenko, C. Filippi, and M. Scheffler, “On the accuracy of density-functional theory exchange-correlation functionals for h bonds in small water clusters. ii. the water hexamer and van der waals interactions,” 2008.

[196] B. Santra, A. Michaelides, and M. Scheffler, “Coupled cluster benchmarks of water monomers and dimers extracted from density-functional theory liquid water: The importance of monomer deformations,” *J Chem Phys*, vol. 131, no. 12, p. 124509, 2009.

[197] V. Staroverov, G. Scuseria, J. Tao, and J. Perdew, “Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes,” *J Chem Phys*, vol. 119, no. 23, pp. 12129–12137, 2003.

[198] J. Ireta, J. Neugebauer, and M. Scheffler, “On the accuracy of dft for describing hydrogen bonds: Dependence on the bond directionality,” *The Journal of Physical Chemistry A*, vol. 108, no. 26, pp. 5692–5698, 2004.

[199] N. Marom, A. Tkatchenko, M. Rossi, V. V. Gobre, O. Hod, M. Scheffler, and L. Kronik, “Dispersive interactions within density functional theory: Benchmarking semi-empirical and pair-wise corrected density functionals,” *submitted*, 2011.

[200] D. McQuarrie, *Statistical Mechanics*. University Science Books, 1st ed., 2000.

[201] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press, 2nd ed., 2002.

[202] C. Chipot and A. Pohorille, *Free Energy Calculations: Theory and Applications in Chemistry and Biology*. Springer Verlag, Berlin-Heidelberg, 1st ed., 2007.

[203] E. Wilson, J. Decius, and P. Cross, *Molecular Vibrations: the theory of infrared and Raman vibrational spectra*. Dover, new ed., 2003.

[204] P. M. Morse, "Diatomic molecules according to the wave mechanics. ii. vibrational levels," *Phys. Rev.*, vol. 34, pp. 57–64, Jul 1929.

[205] J. Neugebauer, M. Reiher, C. Kind, and B. A. Hess, "Quantum chemical calculation of vibrational spectra of large moleculesraman and ir spectra for buckminsterfullerene," *Journal of Computational Chemistry*, vol. 23, no. 9, pp. 895–910, 2002.

[206] D. Porezag and M. R. Pederson, "Infrared intensities and raman-scattering activities within density-functional theory," *Phys. Rev. B*, vol. 54, pp. 7830–7836, Sep 1996.

[207] M. R. Pederson, T. Baruah, P. B. Allen, and C. Schmidt, "Density-functional-based determination of vibrational polarizabilities in molecules within the double-harmonic approximation: Derivation and application," *Journal of Chemical Theory and Computation*, vol. 1, no. 4, pp. 590–596, 2005.

[208] J. Neugebauer and B. A. Hess, "Fundamental vibrational frequencies of small polyatomic molecules from density-functional calculations and vibrational perturbation theory," *The Journal of Chemical Physics*, vol. 118, no. 16, pp. 7215–7225, 2003.

[209] J. Antony, G. von Helden, G. Meijer, and B. Schmidt, "Anharmonic midinfrared vibrational spectra of benzoic acid monomer and dimer," *J. Chem. Phys.*, vol. 123, p. 014305, 2005.

[210] D. Benoit, "Rationalizing the vibrational spectra of biomolecules using atomistic simulations," *Frontiers in Bioscience*, vol. 14, no. 1, pp. 4229–4241, 2009.

[211] B. R. Brooks, D. Janezic, and M. Karplus, "Harmonic analysis of large systems. i. methodology," *Journal of Computational Chemistry*, vol. 16, no. 12, pp. 1522–1542, 1995.

[212] C. Herrmann, J. Neugebauer, and M. Reiher, "Finding a needle in a haystack: direct determination of vibrational signatures in complex systems," *New Journal of Chemistry*, vol. 31, pp. 818–831, 2007.

[213] N. S. Bieler, M. P. Haag, C. R. Jacob, and M. Reiher, "Analysis of the cartesian tensor transfer method for calculating vibrational spectra of polypeptides," *Journal of Chemical Theory and Computation*, vol. 0, no. 0, 0.

[214] R. Car and M. Parrinello, "Unified approach for molecular dynamics and density-functional theory," *Phys. Rev. Lett.*, vol. 55, pp. 2471–2474, Nov 1985.

[215] H. Ishida, Y. Nagai, and A. Kidera, "Symplectic integrator for molecular dynamics of a protein in water," *Chem Phys Lett*, vol. 282, pp. 115–120, Jan 1998.

[216] A. Odell, A. Delin, B. Johansson, N. Bock, M. Challacombe, and A. M. N. Niklasson, "Higher-order symplectic integration in born-oppenheimer molecular dynamics," *J Chem Phys*, vol. 131, no. 24, p. 244106, 2009.

[217] G. Zheng, A. M. N. Niklasson, and M. Karplus, "Lagrangian formulation with dissipation of born-oppenheimer molecular dynamics using the density-functional tight-binding method," *J Chem Phys*, vol. 135, no. 4, p. 044122, 2011.

[218] T. D. Kühne, M. Krack, F. R. Mohamed, and M. Parrinello, "Efficient and accurate car-parrinello-like approach to born-oppenheimer molecular dynamics," *Phys. Rev. Lett.*, vol. 98, p. 066401, Feb 2007.

[219] A. M. N. Niklasson, C. J. Tymczak, and M. Challacombe, "Time-reversible born-oppenheimer molecular dynamics," *Phys. Rev. Lett.*, vol. 97, p. 123001, Sep 2006.

[220] M. E. Tuckerman, D. Marx, M. L. Klein, and M. Parrinello, “On the quantum nature of the shared proton in hydrogen bonds,” *Science*, vol. 275, no. 5301, pp. 817–820, 1997.

[221] J. A. Morrone and R. Car, “Nuclear quantum effects in water,” *Phys. Rev. Lett.*, vol. 101, p. 017801, 2008.

[222] C. Vega, M. M. Conde, C. McBride, J. L. F. Abascal, E. G. Noya, R. Ramirez, and L. M. Sesé, “Heat capacity of water: A signature of nuclear quantum effects,” *J. Chem. Phys.*, vol. 132, p. 046101, 2010.

[223] X.-Z. Li, B. Walker, and A. Michaelides, “Quantum nature of the hydrogen bond,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 16, pp. 6369–6373, 2011.

[224] H. C. Andersen, “Molecular dynamics simulations at constant pressure and/or temperature,” *J. Chem. Phys.*, vol. 72, p. 2384, 1980.

[225] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, “Molecular dynamics with coupling to an external bath,” *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[226] A. Lemak and N. Balabev, “On the berendsen thermostat,” *Molecular Simulation*, vol. 13, pp. 177–187, 1994.

[227] A. Mor, G. Ziv, and Y. Levy, “Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats,” *Journal of Computational Chemistry*, vol. 29, no. 12, pp. 1992–1998, 2008.

[228] S. Nosé, “A unified formulation of the constant temperature molecular dynamics methods.,” *J. Chem. Phys.*, vol. 81, p. 511, 1984.

[229] W. Hoover, “Canonical dynamics: Equilibrium phase-space distributions,” *Phys Rev. A*, vol. 31, p. 1695, 1985.

[230] G. Martyna, M. Kleun, and M. Tuckerman, “Nose-hoover chains - the canonical ensemble via continuous dynamics,” *J Chem Phys*, vol. 97, no. 4, pp. 2635–2643, 1992.

[231] G. Bussi, D. Donadio, and M. Parrinello, “Canonical sampling through velocity rescaling,” *The Journal of Chemical Physics*, vol. 126, no. 1, p. 014101, 2007.

[232] G. Bussi and M. Parrinello, “Stochastic thermostats: comparison of local and global schemes,” *Computer Physics Communications*, vol. 179, no. 1-3, pp. 26 – 29, 2008. Special issue based on the Conference on Computational Physics 2007 - CCP 2007.

[233] R. Ramírez, T. López-Ciudad, P. Kumar, and D. Marx, “Quantum corrections to classical time-correlation functions: Hydrogen bonding and anharmonic floppy modes,” *J. Chem. Phys.*, vol. 121, no. 9, pp. 3973–3983, 2004.

[234] A. Witt, S. D. Ivanov, M. Shiga, H. Forbert, and D. Marx, “On the applicability of centroid and ring polymer path integral molecular dynamics for vibrational spectroscopy,” *J Chem Phys*, vol. 130, no. 19, p. 194510, 2009.

[235] S. D. Ivanov, A. Witt, M. Shiga, and D. Marx, “Communications: On artificial frequency shifts in infrared spectra obtained from centroid molecular dynamics: Quantum liquid water,” *J. Chem. Phys.*, vol. 132, no. 3, p. 031101, 2010.

[236] R. Zwanzig, “Time-correlation functions and transport coefficients in statistical mechanics,” *Annu. Rev. Phys.*, vol. 16, p. 67, 1965.

[237] J. Borisow, M. Moraldi, and L. Frommhold, “The collision induced spectroscopies,” *Molecular Physics*, vol. 56, no. 4, pp. 913–922, 1985.

[238] R. Ramirez, T. Lopez-Ciudad, P. Kumar, and D. Marx, “Quantum corrections to classical time-correlation functions: Hydrogen bonding and anharmonic floppy modes,” *J Chem Phys*, vol. 121, no. 9, pp. 3973–3983, 2004.

[239] M. Schmitz and P. Tavan, “Vibrational spectra from atomic fluctuations in dynamics simulations. ii. solvent-induced frequency fluctuations at femtosecond time resolution,” *The Journal of Chemical Physics*, vol. 121, no. 24, pp. 12247–12258, 2004.

[240] M.-P. Gaigeot, M. Martinez, and R. Vuilleumier, “Infrared spectroscopy in the gas and liquid phase from first principle molecular dynamics simulations: application to small peptides,” *Mol Phys*, vol. 105, no. 19-22, pp. 2857–2878, 2007.

[241] M.-P. Gaigeot and M. Sprik, “Ab initio molecular dynamics computation of the infrared spectrum of aqueous uracil,” *J. Phys. Chem.*, vol. 107, pp. 10344–10358, 2003.

[242] M.-P. Gaigeot, “Theoretical spectroscopy of floppy peptides at room temperature. a dftmd perspective: gas and aqueous phase,” *Phys. Chem. Chem. Phys.*, vol. 12, pp. 3336–3359, 2010. See also references therein.

[243] M.-P. Gaigeot, “Infrared spectroscopy of the alanine dipeptide analog in liquid water with dft-md. direct evidence for pii/[small beta] conformations,” *Phys. Chem. Chem. Phys.*, vol. 12, pp. 10198–10209, 2010.

[244] A. Cimas and M.-P. Gaigeot, “Dft-md and vibrational anharmonicities of a phosphorylated amino acid. success and failure,” *Phys Chem Chem Phys*, vol. 12, no. 14, pp. 3501–3510, 2010.

[245] X. Li, D. T. Moore, and S. S. Iyengar, “Insights from first principles molecular dynamics studies toward infrared multiple-photon and single-photon action spectroscopy: Case study of the proton-bound dimethyl ether dimer,” *J Chem Phys*, vol. 128, no. 18, p. 184308, 2008.

[246] M. Schmitz and P. Tavan, “Vibrational spectra from atomic fluctuations in dynamics simulations. i. theory, limitations, and a sample application,” *The Journal of Chemical Physics*, vol. 121, no. 24, pp. 12233–12246, 2004.

[247] D. McNaughton, C. Evans, S. Lane, and C. Nielsen, “The high-resolution ftir far-infrared spectrum of formamide,” *Journal of Molecular Spectroscopy*, vol. 193, pp. 104–117(14), January 1999.

[248] A. Barth and C. Zscherp, “What vibrations tell about proteins,” *Quarterly Reviews of Biophysics*, vol. 35, no. 04, pp. 369–430, 2002.

[249] M. Oboodi, C. Alva, and M. Diem, “Solution-phase raman studies of alanyl dipeptides and various isotopomers - a reevaluation of the amide-iii vibrational assignment,” *J Phys Chem-Us*, vol. 88, no. 3, pp. 501–505, 1984.

[250] T. Weymuth, C. R. Jacob, and M. Reiher, “A local-mode model for understanding the dependence of the extended amide iii vibrations on protein secondary structure,” *J Phys Chem B*, vol. 114, no. 32, pp. 10649–10660, 2010.

[251] E. G. Robertson, M. R. Hockridge, P. D. Jelfs, and J. P. Simons, “Ir-uv ion-depletion and fluorescence spectroscopy of 2-phenylacetamide clusters: hydration of a primary amide,” *Phys. Chem. Chem. Phys.*, vol. 3, pp. 786–795, 2001.

[252] M. Valiev, E. Bylaska, N. Govind, K. Kowalski, T. Straatsma, H. V. Dam, D. Wang, J. Nieplocha, E. Apra, T. Windus, and W. de Jong, “Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations,” *Computer Physics Communications*, vol. 181, no. 9, pp. 1477 – 1489, 2010.

[253] G. Kresse and J. Furthmller, “Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set,” *Computational Materials Science*, vol. 6, no. 1, pp. 15 – 50, 1996.

[254] J. Soler, E. Artacho, J. Gale, A. Garcia, J. Junquera, P. Ordejón, and D. Sánchez-Portal, “The siesta method for ab initio order- n materials simulation,” *Journal of Physics: Condensed Matter*, vol. 14, no. 11, p. 2745, 2002.

[255] C. C. J. Roothaan, “New developments in molecular orbital theory,” *Rev. Mod. Phys.*, vol. 23, pp. 69–89, Apr 1951.

[256] B. Delley, “An all-electron numerical method for solving the local density functional for polyatomic molecules,” *The Journal of Chemical Physics*, vol. 92, no. 1, pp. 508–517, 1990.

[257] V. Lebedev, “A quadrature formula for the sphere of the 131st algebraic order of accuracy,” *Dokl. Math.*, vol. 3, p. 477, 1999.

[258] R. P. Feynman, “Forces in molecules,” *Phys. Rev.*, vol. 56, pp. 340–343, Aug 1939.

[259] H. Hellmann, “Zur rolle der kinetischen elektronenenergie für die zwischenatomaren kräfte,” *Zeitschrift fuer Physik*, vol. 85, pp. 180–190, 1933.

[260] P. Pulay, “Ab initio calculation of force constants and equilibrium geometries in polyatomic molecules,” *Molecular Physics*, vol. 17, pp. 197–204(8), 1969.

[261] R. Gehrke, *First-principles basin-hopping for the structure determination of atomic clusters*. PhD thesis, Fritz-Haber-Institut der Max-Planck-Gesellschaft and Freie Universität Berlin, 2009.

[262] T. Auckenthaler, V. Blum, H.-J. Bungartz, T. Huckle, R. Johanni, L. KrŁmer, B. Lang, H. Lederer, and P. R. Willems, “Parallel solution of partial symmetric eigenvalue problems from electronic structure calculations.,” *Parallel Computing*, 2011. submitted.

[263] J. L. Whitten, “Coulomb potential energy integrals and approximations,” *J. Chem. Phys.*, vol. 58, p. 4496, 1973.

[264] O. Vahtras, J. Almlöf, and M. W. Feyereisen, “Integral approximations for lcao-scf calculations,” *Chem. Phys. Lett.*, vol. 213, p. 514, 1993.

[265] F. Weigend, “A fully direct ri-hf algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency,” *Phys. Chem. Chem. Phys.*, vol. 4, p. 4285, 2002.

[266] F. Weigend, M. Häser, H. Patzelt, and R. Ahlrichs, “Ri-mp2: optimized auxiliary basis sets and demonstration of efficiency,” *Chem. Phys. Lett.*, vol. 294, p. 143, 1998.

[267] B. Liu and A. D. McLean, “Accurate calculation of the attractive interaction of two ground state helium atoms,” *J. Chem. Phys*, vol. 59, no. 8, pp. 4557–4558, 1973.

[268] S. Boys and F. Bernardi, “Calculation of small molecular interactions by differences of separate total energies - some procedures with reduced errors,” *Mol Phys*, vol. 19, p. 553, Jan 1970.

[269] E. Clementi, *Modern techniques in computational chemistry*. ESCOM Science Publishers, Leiden, 1st ed., 1991.

[270] R. M. Balabin, “Communications: Is quantum chemical treatment of biopolymers accurate? intramolecular basis set superposition error (bsse),” *J Chem Phys*, vol. 132, p. 231101, Jan 2010.

[271] F. Jensen, “An atomic counterpoise method for estimating inter- and intramolecular basis set superposition errors,” *J Chem Theory Comput*, vol. 6, pp. 100–106, Jan 2010.

[272] H. Fujitani, A. Matsuura, S. Sakai, H. Sato, and Y. Tanida, “High-level *ab initio* calculations to improve protein backbone dihedral parameters,” *Journal of Chemical Theory and Computation*, 2009.

[273] I. Mayer, “Towards a “chemical” hamiltonian,” *International Journal of Quantum Chemistry*, vol. 23, no. 2, pp. 341–363, 1983.

[274] I. Mayer, “The chemical hamiltonian approach for treating the bsse problem of intermolecular interactions,” *Int J Quantum Chem*, vol. 70, no. 1, pp. 41–63, 1998.

[275] H. B. Jansen and P. Ros, “Non-empirical molecular orbital calculations on the protonation of carbon monoxide,” *Chemical Physics Letters*, vol. 3, no. 3, pp. 140 – 143, 1969.

[276] I. Mayer and P. Valiron, “Second order møller–plesset perturbation theory without basis set superposition error,” *The Journal of Chemical Physics*, vol. 109, p. 3360, Jan 1998.

[277] A. Halkier, T. Helgaker, P. Jorgensen, W. Klopper, H. Koch, J. Olsen, and A. K. Wilson, “Basis-set convergence in correlated calculations on ne, n2, and h2o,” *Chemical Physics Letters*, vol. 286, no. 3-4, pp. 243 – 252, 1998.

[278] F. van Duijneveldt, J. van Duijneveldt-van de Rijdt, and J. van Lenthe, “State of the art in counterpoise theory,” *Chemical Reviews*, vol. 94, no. 7, pp. 1873–1885, 1994.

[279] T. Dunning, “Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen,” *J. Chem. Phys.*, vol. 90, p. 1007, Jan 1989.

[280] P. Fast, M. Sanchez, and D. Truhlar, “Infinite basis limits in electronic structure theory,” *J Chem Phys*, vol. 111, no. 7, pp. 2921–2926, 1999.

[281] D. G. and Truhlar, “Basis-set extrapolation,” *Chemical Physics Letters*, vol. 294, no. 1-3, pp. 45 – 48, 1998.

[282] A. Galano and J. R. Alvarez-Idaboy, “A new approach to counterpoise correction to bsse,” *J. Comp. Chem.*, vol. 27, p. 1203, 2006.

[283] R. M. Epand and H. A. Scheraga, “The influence of long-range interactions on the structure of myoglobin,” *Biochemistry*, vol. 7, no. 8, pp. 2864–2872, 1968.

[284] M. CRUMPTON and P. SMALL, “Conformation of immunologically-active fragments of sperm whale myoglobin in aqueous solution,” *J Mol Biol*, vol. 26, p. 143, Jan 1967.

[285] W. KLEE, “Studies on conformation of ribonuclease s-peptide,” *Biochemistry*, vol. 7, p. 2731, Jan 1968.

[286] T. Keiderling, “Protein and peptide secondary structure and conformational determination with vibrational circular dichroism,” *Current Opinion in Chemical Biology*, vol. 6, no. 5, pp. 682 – 688, 2002.

[287] P. C. Lyu, L. A. Marky, and N. R. Kallenbach, “The role of ion pairs in α-helix stability: two new designed helical peptides,” *Journal of the American Chemical Society*, vol. 111, no. 7, pp. 2733–2734, 1989.

[288] S. Marqusee, V. Robbins, and R. Baldwin, “Unusually stable helix formation in short alanine-based peptides,” *Proceedings of the National Academy of Sciences U.S.A.*, vol. 86, pp. 5286–5290, 1989.

[289] A. Chakrabartty and R. L. Baldwin, “Stability of α-helices,” vol. 46, pp. 141 – 176, 1995.

[290] J. M. Scholtz and R. L. Baldwin, “The mechanism of α-helix formation by peptides,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 21, no. 1, pp. 95–118, 1992. and references therein.

[291] A. Chakrabartty, T. Kortemme, and R. L. Baldwin, “Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions,” *Protein Science*, vol. 3, no. 5, pp. 843–852, 1994.

[292] S. PADMANABHAN, S. Marqusee, T. RIDGEWAY, T. LAUE, and R. BALDWIN, “Relative helix-forming tendencies of nonpolar amino-acids,” *Nature*, vol. 344, no. 6263, pp. 268–270, 1990.

[293] L. Williams, K. Kather, and D. S. Kemp, “High helicities of lys-containing, ala-rich peptides are primarily attributable to a large, context-dependent lys stabilization,” *Journal of the American Chemical Society*, vol. 120, no. 43, pp. 11033–11043, 1998.

[294] J. A. Vila, D. R. Ripoll, and H. A. Scheraga, “Influence of lysine content and ph on the stability of alanine-based copolypeptides,” *Biopolymers*, vol. 58, no. 3, pp. 235–246, 2001.

[295] Z. Shi, C. Olson, A. Bell, and N. Kallenbach, “Stabilization of $alpha$-helix structure by polar side-chain interactions: complex salt bridges, cation-pi interactions, and c-h em leader o h-bonds.,” *Biopolymers*, vol. 60, pp. 366–380, 2001.

[296] C. Rohl, A. Chakrabartty, and R. Baldwin, “Helix propagation and n-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol,” *Protein Sci*, vol. 5, pp. 2623–2637, Jan 1996.

[297] E. Spek, C. Olson, Z. Shi, and N. Kallenbach, “Alanine is an intrinsic α-helix stabilizing amino acid,” *Journal of the American Chemical Society*, vol. 121, no. 23, pp. 5571–5572, 1999.

[298] T. E. Creighton, “Stability of α-helices,” *Nature*, vol. 326, pp. 547–548, 1987.

[299] S. Ihara, T. Ooi, and S. Takahashi, “Effects of salts on the nonequivalent stability of the α-helices of isomeric block copolypeptides,” *Biopolymers*, vol. 21, no. 1, pp. 131–145, 1982.

[300] S. Takahashi, E.-H. Kim, T. Hibino, and T. Ooi, “Comparison of α-helix stability in peptides having a negatively or positively charged residue block attached either to the n- or c-terminus of an -helix: The electrostatic contribution and anisotropic stability of the α-helix,” *Biopolymers*, vol. 28, no. 5, pp. 995–1009, 1989.

[301] S. MIICK, G. MARTINEZ, W. FIORI, A. TODD, and G. MILLHAUSER, “Short alanine-based peptides may form 3(10)-helices and not α-helices in aqueous-solution,” *Nature*, vol. 359, no. 6396, pp. 653–655, 1992.

[302] Z. Shi, C. A. Olson, G. D. Rose, R. L. Baldwin, and N. R. Kallenbach, “Polyproline ii structure in a sequence of seven alanine residues,” *Proc. Natl. Am. Sc. U.S.A.*, vol. 99, pp. 9190–9195, 2002.

[303] G. Merutka and E. Stellwagen, “Analysis of peptides for helical prediction,” *Biochemistry*, vol. 28, no. 1, pp. 352–357, 1989.

[304] G. Merutka and E. Stellwagen, “Positional independence and additivity of amino acid replacements on helix stability in monomeric peptides,” *Biochemistry*, vol. 29, no. 4, pp. 894–898, 1990.

[305] T. P. Creamer and G. D. Rose, “Side-chain entropy opposes alpha-helix formation but rationalizes experimentally determined helix-forming propensities,” *Proceedings of the National Academy of Sciences*, vol. 89, no. 13, pp. 5937–5941, 1992.

[306] P. Lyu, M. Liff, L. Marky, and N. Kallenbach, “Side chain contributions to the stability of alpha-helical structure in peptides,” *Science*, vol. 250, no. 4981, pp. 669–673, 1990.

[307] N. Pace and M. Scholtz, “A helix propensity scale based on experimental studies of peptides and proteins,” *Biophysical Journal*, vol. 75, no. 1, pp. 422–427, 1998.

[308] B. H. Zimm and J. K. Bragg, “Theory of the phase transition between helix and random coil in polypeptide chains,” *The Journal of Chemical Physics*, vol. 31, no. 2, pp. 526–535, 1959.

[309] S. Lifson and A. Roig, “On the theory of helix—coil transition in polypeptides,” *The Journal of Chemical Physics*, vol. 34, no. 6, pp. 1963–1974, 1961.

[310] G. E. Job, R. J. Kennedy, B. Heitmann, J. S. Miller, S. M. Walker, and D. S. Kemp, “Temperature- and length-dependent energetics of formation for polyalanine helices in water: Assignment of w$_{Ala}$(n,t) and temperature-dependent cd ellipticity standards,” *Journal of the American Chemical Society*, vol. 128, no. 25, pp. 8227–8233, 2006.

[311] R. J. Kennedy, S. M. Walker, and D. S. Kemp, “Energetic characterization of short helical polyalanine peptides in water: Analysis of ^{13}co chemical shift data,” *Journal of the American Chemical Society*, vol. 127, no. 48, pp. 16961–16968, 2005.

[312] B. Heitmann, G. Job, R. Kennedy, S. Walker, and D. Kemp *Journal of the American Chemical Society*, vol. 127, no. 6, pp. 1690–1704, 2005.

[313] G. Job, B. Heitmann, R. Kennedy, S. Walker, and D. Kemp, “Calibrated calculation of polyalanine fractional helicities from circular dichroism ellipticities,” *Angew Chem Int Edit*, vol. 43, no. 42, pp. 5649–5651, 2004.

[314] R. Schweitzer-Stenner, F. Eker, K. Griebenow, X. Cao, and L. A. Nafie, “The conformation of tetraalanine in water determined by polarized raman, ft-ir, and vcd spectroscopy,” *Journal of the American Chemical Society*, vol. 126, no. 9, pp. 2768–2776, 2004.

[315] A. E. García and K. Y. Sanbonmatsu, “α-helical stabilization by side chain shielding of backbone hydrogen bonds,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 5, pp. 2782–2787, 2002.

[316] A. E. Garcia, “Characterization of non-alpha helical conformations in ala peptides,” *Polymer*, vol. 45, no. 2, pp. 669 – 676, 2004.

[317] D. Paschek, S. Gnanakaran, and A. E. Garcia, “Simulations of the pressure and temperature unfolding of an alpha-helical peptide,” *P Natl Acad Sci Usa*, vol. 102, no. 19, pp. 6765–6770, 2005.

[318] B. Roux and T. Simonson, “Implicit solvent models,” *Biophysical Chemistry*, vol. 78, no. 1-2, pp. 1 – 20, 1999.

[319] R. A. G. D. Silva, S. Yasui, J. Kubelka, F. Formaggio, M. Crisma, C. Toniolo, and T. A. Keiderling, “Discriminating 3(10)- from alpha helices: Vibrational and electronic cd and ir absorption study of related aib-containing oligopeptides,” 2002.

[320] R. A. G. D. Silva, J. Kubelka, P. Bour, S. M. Decatur, and T. A. Keiderling, “Site-specific conformational determination in thermal unfolding studies of helical peptides using vibrational circular dichroism with isotopic substitution,” *Proc. Natl. Am. Sc. U.S.A.*, vol. 97, pp. 8318–8323, 2000.

[321] P. Bour, J. Kubelka, and T. A. Keiderling, “Ab initio quantum mechanical models of peptide helices and their vibrational spectra,” *Biopolymers*, vol. 65, no. 1, pp. 45–59, 2002.

[322] P. Salvador, R. Wieczorek, and J. J. Dannenberg, “Direct calculation of trans-hydrogen-bond c-13-n-15 3-bond j-couplings in entire polyalanine alpha-helices. a density functional theory study,” *J Phys Chem B*, vol. 111, no. 9, pp. 2398–2403, 2007.

[323] M. I.-H. Tsai, Y. Xu, and J. J. Dannenberg *J Phys Chem B*, vol. 113, no. 1, pp. 309–318, 2009.

[324] L. Ismer, J. Ireta, S. Boeck, and J. Neugebauer, “Phonon spectra and thermodynamic properties of the infinite polyalanine α helix: A density-functional-theory-based harmonic vibrational analysis,” *Phys. Rev. E*, vol. 71, no. 3, p. 031911, 2005.

[325] L. Ismer, J. Ireta, and J. Neugebauer, “First-principles free-energy analysis of helix stability: The origin of the low entropy in π helices,” *The Journal of Physical Chemistry B*, vol. 112, no. 13, pp. 4109–4112, 2008.

[326] J. P. Simons, “Good vibrations: probing biomolecular structure and interactions through spectroscopy in the gas phase,” *Molecular Physics*, vol. 107, pp. 2435–2458(24), July 2009.

[327] M. Karas, D. Bachmann, U. Bahr, and F. Hillenkamp, “Matrix-assisted ultraviolet laser desorption of non-volatile compounds,” *International Journal of Mass Spectrometry and Ion Processes*, vol. 78, pp. 53 – 68, 1987.

[328] J. Fenn, M. Mann, C. Meng, S. Wong, and C. Whitehouse, “Electrospray ionization for mass spectrometry of large biomolecules,” *Science*, vol. 246, no. 4926, pp. 64–71, 1989.

[329] R. Aebersold and M. Mann, “Mass spectrometry-based proteomics,” *Nature*, vol. 422, pp. 198 – 207, 2003.

[330] N. Polfer and J. Oomens, “Vibrational spectroscopy of bare and solvated ionic complexes of biological relevance,” *Mass Spectrometry Reviews*, vol. 28, pp. 468–494, 2009.

[331] T. Wyttenbach and M. T. Bowers, “Hydration of biomolecules,” *Chemical Physics Letters*, vol. 480, no. 1-3, pp. 1 – 16, 2009.

[332] M. Kohtani, B. Kinnear, and M. Jarrold, “Metal-ion enhanced helicity in the gas phase,” *J. Am. Chem. Soc.*, vol. 122, pp. 12377–12378, 2000.

[333] M. Kohtani, M. Jarrold, S. Wee, and R. O’Hair, “Metal ion interactions with polyalanine peptides,” *J. Phys. Chem. B*, vol. 108, pp. 6093–6097, 2004.

[334] M. S. de Vries and P. Hobza, “Gas-phase spectroscopy of biomolecular building blocks,” *Annual Review of Physical Chemistry*, vol. 58, no. 1, pp. 585–612, 2007.

[335] M. F. Jarrold, “Helices and sheets in vacuo,” *Phys. Chem. Chem. Phys.*, vol. 9, pp. 1659–1671, 2007. See also references therein.

[336] J. L. P. Benesch and C. V. Robinson, “Dehydrated but unharmed,” *Nature*, vol. 462, no. 7273, pp. 576–577, 2009.

[337] *Phys. Chem. Chem. Phys.*, vol. 6, 2004.

[338] P. Dugourd, R. Hudgins, D. Clemmer, and M. Jarrold, “High-resolution ion mobility measurements,” *Rev Sci Instrum*, vol. 68, no. 2, pp. 1122–1129, 1997.

[339] T. Wyttenbach, G. vonHelden, and M. Bowers, “Gas-phase conformation of biological molecules: Bradykinin,” *J Am Chem Soc*, vol. 118, no. 35, pp. 8355–8364, 1996.

[340] T. Wyttenbach, G. von Helden, J. Batka, D. Carlat, and M. Bowers, “Effect of the long-range potential on ion mobility measurements,” *J. Am. Soc. Mass. Spectr.*, vol. 8, no. 3, pp. 275–282, 1997.

[341] J. Oomens, B. G. Sartakov, G. Meijer, and G. von Helden, “Gas-phase infrared multiple photon dissociation spectroscopy of mass-selected molecular ions,” *Int. J. Mass Spectrom.*, vol. 254, pp. 1–19, 2006.

[342] J. A. Stearns, O. V. Boyarkin, and T. R. Rizzo, “Spectroscopic signatures of gas-phase helices:Ac-Phe-(Ala)5-Lys-H+ and Ac-Phe-(Ala)10-Lys-H+,” *Journal of the American Chemical Society*, vol. 129, p. 13820, 2007.

[343] M. Kohtani, J. Schneider, T. Jones, and M. F. Jarrold, “The mobile proton in polyalanine peptides,” *J Am Chem Soc*, vol. 126, pp. 16981–16987, 2004.

[344] R. Hudgins and M. Jarrold, “Helix formation in unsolvated alanine-based peptides: Helical monomers and helical dimers,” *J. Am. Chem. Soc*, vol. 121, pp. 3494–3501, 1999.

[345] M. Kohtani and M. F. Jarrold, “The initial steps in the hydration of unsolvated peptides: Water molecule adsorption on alanine-based helices and globules,” *Journal of the American Chemical Society*, vol. 124, no. 37, pp. 11148–11158, 2002.

[346] M. Kohtani, B. Kinnear, and M. Jarrold, “Metal-ion enhanced helicity in the gas phase,” *J Am Chem Soc*, vol. 122, no. 49, pp. 12377–12378, 2000.

[347] J. R. McLean, J. A. McLean, Z. Wu, C. Becker, L. M. Perez, C. N. Pace, J. M. Scholtz, and D. H. Russell, "Factors that influence helical preferences for singly charged gas-phase peptide ions: The effects of multiple potential charge-carrying sites," *The Journal of Physical Chemistry B*, vol. 114, no. 2, pp. 809–816, 2010.

[348] D. Liu, T. Wyttenbach, and M. T. Bowers, "Hydration of protonated primary amines: effects of intermolecular and intramolecular hydrogen bonds," *International Journal of Mass Spectrometry*, vol. 236, no. 1-3, pp. 81 – 90, 2004. Special issue: In honour of Dudley H. Williams.

[349] P. E. Barran, N. C. Polfer, D. J. Campopiano, D. J. Clarke, P. R. Langridge-Smith, R. J. Langley, J. R. Govan, A. Maxwell, J. R. Dorin, R. P. Millar, and M. T. Bowers, "Is it biologically relevant to measure the structures of small peptides in the gas-phase?," *International Journal of Mass Spectrometry*, vol. 240, no. 3, pp. 273 – 284, 2005.

[350] J. Gidden, E. S. Baker, A. Ferzoco, and M. T. Bowers, "Structural motifs of dna complexes in the gas phase," *International Journal of Mass Spectrometry*, vol. 240, no. 3, pp. 183 – 193, 2005.

[351] B. Kinnear and M. Jarrold, "Helix formation in unsolvated peptides: Side chain entropy is not the determining factor," *J Am Chem Soc*, vol. 123, no. 32, pp. 7907–7908, 2001.

[352] B. Kinnear, M. Hartings, and M. Jarrold, "The energy landscape of unsolvated peptides: Helix formation and cold denaturation in ac-a(4)g(7)a4+h+," *J Am Chem Soc*, vol. 124, no. 16, pp. 4422–4431, 2002.

[353] B. Ma, C.-J. Tsai, and R. Nussinov, "A systematic study of the vibrational free energies of polypeptides in folded and random states," *Biophys J*, vol. 79, no. 5, pp. 2739–2753, 2000.

[354] W. Chin, F. Piuzzi, I. Dimicoli, and M. Mons, "Probing the competition between secondary structures and local preferences in gas phase isolated peptide backbones," *Phys Chem Chem Phys*, vol. 8, no. 9, pp. 1033–1048, 2006.

[355] E. Gloaguen, R. Pollet, F. Piuzzi, B. Tardivel, and M. Mons, "Gas phase folding of an (ala)4 neutral peptide chain: spectroscopic evidence for the formation of a -hairpin h-bonding pattern," *Phys. Chem. Chem. Phys.*, vol. 11, no. 48, pp. 11385–11388, 2009.

[356] W. Chin, F. Piuzzi, J. Dognon, L. Dimicoli, B. Tardivel, and M. Mons, "Gas phase formation of a 3(10)-helix in a three-residue peptide chain: Role of side chain-backbone interactions as evidenced by ir-uv double resonance experiments," *J Am Chem Soc*, vol. 127, no. 34, pp. 11900–11901, 2005.

[357] V. Brenner, F. Piuzzi, I. Dimicoli, B. Tardivel, and M. Mons, "Chirality-controlled formation of beta-turn secondary structures in short peptide chains: Gas-phase experiment versus quantum chemistry," *Angew Chem Int Edit*, vol. 46, no. 14, pp. 2463–2466, 2007.

[358] T. D. Vaden, T. S. J. A. de Boer, J. P. Simons, L. C. Snoek, S. Suhai, and B. Paizs, "Vibrational spectroscopy and conformational structure of protonated polyalanine peptides isolated in the gas phase," *J Phys Chem A*, vol. 112, no. 20, pp. 4608–4616, 2008.

[359] K. Pagel, P. Kupser, F. Bierau, N. C. Polfer, J. D. Steill, J. Oomens, G. Meijer, B. Koksch, and G. von Helden, "Gas-phase ir spectra of intact [alpha]-helical coiled coil protein complexes," *International Journal of Mass Spectrometry*, vol. 283, no. 1-3, pp. 161 – 168, 2009.

[360] H. Zhu, M. Blom, I. Compagnon, A. M. Rijs, S. Roy, G. von Helden, and B. Schmidt, "Conformations and vibrational spectra of a model tripeptide: change of secondary structure upon micro-solvation," *Phys. Chem. Chem. Phys.*, vol. 12, pp. 3415–3425, 2010.

[361] P. Kupser, K. Pagel, J. Oomens, N. Polfer, B. Koksch, G. Meijer, and G. v. Helden, "Amide-i and -ii vibrations of the cyclic -sheet model peptide gramicidin s in the gas phase," *Journal of the American Chemical Society*, vol. 132, no. 6, pp. 2085–2093, 2010.

[362] J. Valle, J. Eyler, J. Oomens, D. Moore, A. van der Meer, G. von Helden, G. Meijer, C. Hendrickson, A. Marshall, and G. Blakney, “Free electron laser-fourier transform ion cyclotron resonance mass spectrometry facility for obtaining infrared multiphoton dissociation spectra of gaseous ions,” *Rev Sci Instrum*, vol. 76, no. 2, p. 023103, 2005.

[363] D. Reha, H. Valdes, J. Vondrasek, P. Hobza, A. Abu-Riziq, B. Crews, and M. de Vries, “Structure and ir spectrum of phenylalanyl-glycyl-glycine tripetide in the gas-phase: Ir/uv experiments, ab initio quantum chemical calculations, and molecular dynamic simulations,” *Chem-Eur J*, vol. 11, no. 23, pp. 6803–6817, 2005.

[364] S. Grimme, J. Antony, T. Schwabe, and C. Mueck-Lichtenfeld, “Density functional theory with dispersion corrections for supramolecular structures, aggregates, and complexes of (bio)organic molecules,” 2007.

[365] T. Haeber, K. Seefeld, G. Engler, S. Grimme, and K. Kleinermanns, “Ir/uv spectra and quantum chemical calculations of trp-ser: Stacking interactions between backbone and indole side-chain,” *Phys Chem Chem Phys*, vol. 10, no. 19, pp. 2844–2851, 2008.

[366] P. Dugourd, R. R. Hudgins, D. E. Clemmer, and M. F. Jarrold, “High-resolution ion mobility measurements,” *Review of Scientific Instruments*, vol. 68, no. 2, pp. 1122–1129, 1997.

[367] A. Shvartsburg and M. Jarrold, “An exact hard-spheres scattering model for the mobilities of polyatomic ions,” *Chem Phys Lett*, vol. 261, no. 1-2, pp. 86–91, 1996.

[368] S. Chutia, M. Rossi, V. Blum, and M. Scheffler, “Microsolvation of two benchmark gas-phase peptides from first-principles methods,” 2011. in preparation.

[369] J. Oomens, G. Meijer, and G. von Helden, “Gas phase infrared spectroscopy of cationic indane, acenaphthene, fluorene, and fluoranthene,” *The Journal of Physical Chemistry A*, vol. 105, no. 36, pp. 8302–8309, 2001.

[370] K. Lehmann, B. Pate, and G. Scoles, “Intramolecular dynamics from eigenstate-resolved infrared spectra,” *Ann. Rev. Phys. Chem.*, vol. 45, pp. 241–74, 1994.

[371] D. Wales, “Exploring the energy landscape,” *Int J Mod Phys B*, vol. 19, no. 15-17, pp. 2877–2885, 2005.

[372] A. Laio and M. Parrinello, “Escaping free-energy minima,” *P. Natl. Acad. Sci. USA*, vol. 99, no. 20, pp. 12562–12566, 2002.

[373] G. Bussi, A. Laio, and M. Parrinello, “Equilibrium free energies from nonequilibrium metadynamics,” *Physical review letters*, vol. 96, no. 9, p. 090601, 2006.

[374] G. Torrie and J. Valleau, “Non-physical sampling distributions in monte-carlo free-energy estimation - umbrella sampling,” *J Comput Phys*, vol. 23, no. 2, pp. 187–199, 1977.

[375] J. Ponder, “Tinker - software tools for molecular design.” In this work we used versions 4.2 and 5.1 of the program and the force-fields’ versions distributed within the package.

[376] C. Baldauf, V. Blum, and M. Scheffler, 2011. in preparation.

[377] Y. Zhao and D. Truhlar, “A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions,” *The Journal of chemical physics*, 2006.

[378] Y. Zhao and D. Truhlar, “The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements,” *Theor Chem Acc*, 2008.

[379] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, and W. Yang, “Quantum mechanics simulation of protein dynamics on long timescale,” *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 4, pp. 484–489, 2001.

[380] M. Elstner, K. J. Jalkanen, M. Knapp-Mohammady, T. Frauenheim, and S. Suhai, “Dft studies on helix formation in n-acetyl-(-alanyl)n-n -methylamide for n=1-20,” *Chemical Physics*, vol. 256, no. 1, pp. 15 – 27, 2000.

[381] L. Viehland and E. Mason, “Tables of transport collision integrals for (n, 6, 4) ion-neutral potentials,” *Atomic Data and Nuclear Data Tables*, vol. 16, pp. 495–514, 1975.

[382] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in Fortran 90*. Cambridge University Press, 2nd ed., 1996.

[383] M. V. Hove, W. Weinberg, and C. Chan, *Low-energy electron diffraction: experiment, theory and surface structure determination*. Springer-Verlag, 1 ed., 1986.

[384] V. Blum and K. Heinz, “Fast leed intensity calculations for surface crystallography using tensor leed,” *Comput Phys Commun*, vol. 134, no. 3, pp. 392–425, 2001.

[385] J. Pendry, “Reliability factors for leed calculations,” *J Phys C Solid State*, vol. 13, no. 5, pp. 937–944, 1980.

[386] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, “Smaller and faster: The 20-residue trp-cage protein folds in 4 ?s,” *Journal of the American Chemical Society*, vol. 124, no. 44, pp. 12952–12953, 2002.

[387] I. K. Lednev, A. S. Karnoup, M. C. Sparrow, and S. A. Asher, “α-helix peptide folding and unfolding activation barriers: A nanosecond uv resonance raman study,” *Journal of the American Chemical Society*, vol. 121, no. 35, pp. 8074–8086, 1999.

[388] E. K. Asciutto, A. V. Mikhonin, S. A. Asher, and J. D. Madura, “Computational and experimental determination of the α-helix unfolding reaction coordinate,” *Biochemistry*, vol. 47, no. 7, pp. 2046–2050, 2008.

[389] S. Woutersen and P. Hamm, “Time-resolved two-dimensional vibrational spectroscopy of a short α-helix in water,” *J. Chem. Phys.*, vol. 115, no. 16, p. 7737, 2001.

[390] S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer, “Fast events in protein folding: Helix melting and formation in a small peptide,” *Biochemistry*, vol. 35, no. 3, pp. 691–697, 1996.

[391] W. A. Hegefeld, S.-E. Chen, K. Y. DeLeon, K. Kuczera, and G. S. Jas, “Helix formation in a pentapeptide: Experiment and force-field dependent dynamics,” *J. Phys. Chem. A*, vol. 114, pp. 12391–12402, 2010.

[392] S. Woutersen, Y. Mu, G. Stock, and P. Hamm, “Subpicosecond conformational dynamics of small peptides probed by two-dimensional vibrational spectroscopy,” *Proceedings of the National Academy of Sciences*, vol. 98, no. 20, pp. 11254–11258, 2001.

[393] L. ZHANG and J. HERMANS, “3(10)-helix versus alpha-helix - a molecular-dynamics study of conformational preferences of aib and alanine,” *Journal of the American Chemical Society*, vol. 116, no. 26, pp. 11915–11921, 1994.

[394] Y. Levy, J. Jortner, and O. Becker, “Solvent effects on the energy landscapes and folding kinetics of polyalanine,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 5, pp. 2188–2193, 2000.

[395] J. Almlof and P. R. Taylor, “General contraction of gaussian basis sets. i. atomic natural orbitals for first- and second-row atoms,” *The Journal of Chemical Physics*, vol. 86, no. 7, pp. 4070–4077, 1987.

[396] R. Kendall, T. Dunning, and R. Harrison, “Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions,” *J. Chem. Phys.*, vol. 96, pp. 6796–6806, Jan 1992.

[397] P. C. Hariharan and J. A. Pople, “The influence of polarization functions on molecular orbital hydrogenation energies,” *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, vol. 28, pp. 213–222, 1973.

[398] A. J. Sadlej, “Medium-size polarized basis sets for high-level correlated calculations of molecular electric properties,” *Collect. Czech. Chem. Commun.*, vol. 53, pp. 1995–2016, 1988.

[399] A. Schafer, H. Horn, and R. Ahlrichs, “Fully optimized contracted gaussian basis sets for atoms li to kr,” *The Journal of Chemical Physics*, vol. 97, no. 4, pp. 2571–2577, 1992.

[400] D. Asturiol, M. Duran, and P. Salvador *J Chem Phys*, vol. 128, p. 144108, Jan 2008.

[401] H. Valdes, V. Klusak, M. Pitonak, O. Exner, I. Stary, P. Hobza, and L. Rulisek *J Comput Chem*, vol. 29, pp. 861–870, Jan 2008.

[402] A. J. Cohen, P. Mori-Sanchez, and W. Yang, “Second-order perturbation theory with fractional charges and fractional spins,” *J Chem Theory Comput*, vol. 5, no. 4, pp. 786–792, 2009.

[403] D. Woon and T. Dunning, “Gaussian basis sets for use in correlated molecular calculations. v. core-valence basis sets for boron through neon,” *The Journal of chemical physics*, vol. 103, p. 4572, Jan 1995.

[404] D. M. Philipp and R. A. Friesner, “Mixed ab initio qm/mm modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide,” *Journal of Computational Chemistry*, vol. 20, no. 14, pp. 1468–1494, 1999.

[405] R. A. DiStasio, Y. Jung, and M. Head-Gordon, “A resolution-of-the-identity implementation of the local triatomics-in-molecules model for second-order møller-plesset perturbation theory with application to alanine tetrapeptide conformational energies,” *Journal of Chemical Theory and Computation*, vol. 1, no. 5, pp. 862–876, 2005.

[406] J. SCARSDALE, C. VANALSENOY, V. KLIMKOWSKI, L. SCHAFER, and F. MOMANY, “Abinitio studies of molecular geometries .27. optimized molecular-structures and conformational-analysis of n-alpha-acetyl-n-methylalaninamide and comparison with peptide crystal data and empirical calculations,” *J Am Chem Soc*, vol. 105, no. 11, pp. 3438–3445, 1983.

[407] H. BOHM and S. BRODE, “Abinitio scf calculations on low-energy conformers of n-acetyl-n’-methylalaninamide and n-acetyl-n’-methylglycinamide,” *J Am Chem Soc*, vol. 113, no. 19, pp. 7129–7135, 1991.

[408] D. Philipp and R. Friesner, “Mixed ab initio qm/mm modeling using frozen orbitals and tests with alanine dipeptide and tetrapeptide,” *J Comput Chem*, vol. 20, no. 14, pp. 1468–1494, 1999.

[409] K. Jalkanen and S. Suhai, “N-acetyl-l-alanine n’-methylamide: A density functional analysis of the vibrational absorption and vibrational circular dichroism spectra,” *Chemical Physics*, vol. 208, no. 1, pp. 81–116, 1996.

[410] R. Kaschner and D. Hohl, “Density functional theory and biomolecules: A study of glycine, alanine, and their oligopeptides,” *J Phys Chem A*, vol. 102, no. 26, pp. 5111–5116, 1998.

[411] R. Lavrich, D. Plusquellic, R. Suenram, G. Fraser, A. Walker, and M. Tubergen, “Experimental studies of peptide bonds: Identification of the c-7(eq) conformation of the alanine dipeptide analog n-acetyl-alanine n-methylamide from torsion-rotation interactions,” *J Chem Phys*, vol. 118, no. 3, pp. 1253–1265, 2003.

[412] A. J. Cohen, P. Mori-Sanchez, and W. Yang, “Insights into current limitations of density functional theory,” *Science*, vol. 321, no. 5890, pp. 792–794, 2008.

[413] A. J. Cohen, P. Mori-Sanchez, and W. Yang, “Fractional spins and static correlation error in density functional theory,” *J Chem Phys*, vol. 129, no. 12, p. 121104, 2008.