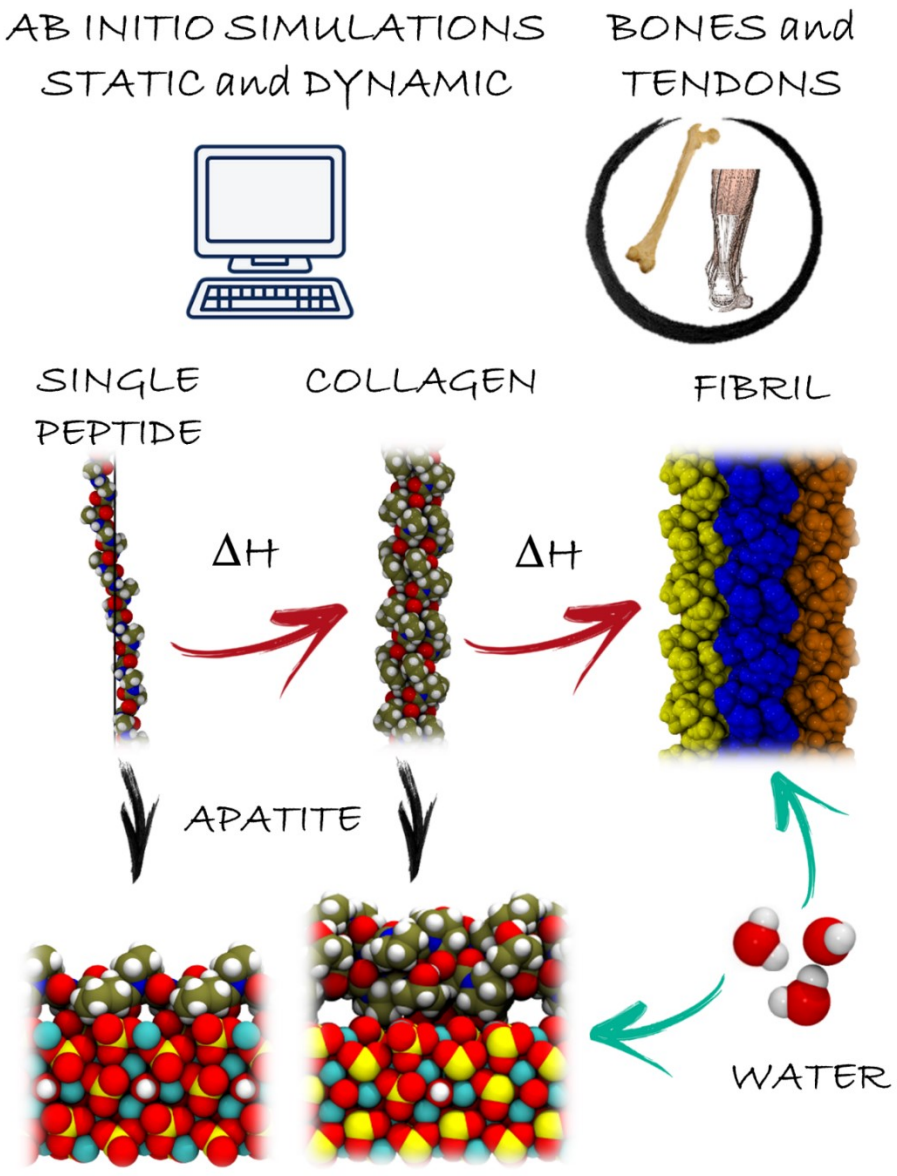
University of Turin
Doctoral School of Sciences and Innovative Technologies
PhD Programme in Chemical and Material Sciences XXX Cycle

"Towards the *Ab Initio* Simulations of Tendon and Bone Materials"



Michele Cutini

Supervisor
Prof. Piero Ugliengo

University of Turin
Doctoral School of Sciences and Innovative Technologies
PhD Programme in Chemical and Material Sciences

XXX Cycle (2014-2017)

Towards the *Ab Initio* Simulations of Tendon and Bone Materials

Candidate: Michele Cutini

Supervisor: Prof. Piero Ugliengo

Jury Members: **Prof. Maria Cristina Menziani**

Università di Università degli Studi di Modena e Reggio Emilia

Dipartimento di Dipartimento di Scienze Chimiche e Geologiche

Prof. Marco Milanesio

Università di Università degli Studi del Piemonte Orientale "Amedeo Avogadro"

Dipartimento di Dipartimento di Scienze ed Innovazione Tecnologica

Prof. Mariona Sodupe Roure

Universitat Autònoma Barcelona

Department of Chemistry

Head of the Doctoral School: Prof. Massimo Maffei

PhD Programme Coordinator: Prof. Mario Chiesa

La mente intuitiva è un dono sacro e la mente razionale è un fedele servo.
Noi abbiamo creato una società che onora il servo e ha dimenticato il dono.
Einstein

Contents

Abbreviations

COL – Collagen
TC – Tropocollagen
Hap – Hydroxyl-apatite
PPII – poly-L-proline type II helix
D – DOWN pucker of pyrrolidine ring
U – UP pucker of pyrrolidine ring
Gly – Glycine
Pro – Proline
Lys – Lysine
Arg – Arginine
Asp – Aspartate
Glu – Glutamic acid
Hyp – (2S,4R)-4-hydroxyl-proline
GPP – Collagen-like peptide with Gly-Pro-Pro composition
GPO – Collagen-like peptide with Gly-Pro-Hyp composition
HEL – Collagen-like triple helix
HPC – High performance computing

Outlook

This PhD thesis resumes the work I have done in the last three years. The project main goal was to elucidate the structural and energetic features of **bone** and **tendon** materials by the use of ***ab initio* simulations**. This approach has been chosen as it complies with the **line of research on bio-materials** of the group in which my PhD thesis has been carried out. The works in the bio-materials field started in 1990, and since then, several aspects on inorganic bio-materials were addressed. My supervisor and former group members investigated silica and hydroxyl-apatite crystals and surfaces, as well as their interactions with small organic molecules of biological interest (water, carbon monoxide, glycine and a small α-helix). This PhD work grows on this background, focusing, however, on a poorly understood field, i.e. the **quantum mechanical simulation of collagen protein**, both free and adsorbed on hydroxyl-apatite. This choice derives from the simple consideration that the collagen protein is the principal component of bones and tendons. Therefore, this work is a natural extension of its background through more realistic and complex simulations of one of the most important biomaterials in nature.

Collagen protein is the most abundant protein in all vertebrate, it is found in plenty of different tissues with several functions. To appreciate how large is the scientific background accumulated on that topic, we quoted Kadler et al.[1]: "newcomers to the field", referring to collagen, "face the daunting task of sifting through 100,000 research papers that span 40 years". Therefore, one of the major effort of this PhD had been dealing with the sorting of the massive amount of research on the topic. Trying to understand what has been already done is fundamental to simulate, in a broader sense, this complex system. A further complication arose from the chosen method of investigation. The complexity of tendon and bone materials is not naturally suited for *ab initio* simulation. Therefore, even after three years of "work on the field", my research project is still at a prelaminar stage. The main slowing factors were defining proper collagen models suitable for *ab initio* calculations, and, finding enough high performance computing to tackle systems of very large size. The outcome of this effort is that several interesting side-questions have emerged during the study of these systems. Anyways, we did not "throw in the sponge". Currently realistic bone and tendons models are under deep investigation in our laboratory.

The thesis consists of three Chapters. **Chapter 1** is a state-of-the-art review of the structure of collagen and its aggregates (tendons and bones). **Chapter 2** includes a brief description of the computational modelling and techniques employed in this thesis. Moreover, I have included a summary on the current scenario of simulations on tendons and bones. **Chapter 3** summarizes the scientific goal achieved during my PhD and it consists of three sub-section. In the first, I tested a promising "fast and accurate" *ab initio* method that can be employed for treating very large systems. In the second, we simulate the interface between a simple collagen model and the hydroxyl-apatite surface, for mimicking the inorganic/organic interface at bones. In the third, I report some preliminary results on the energetic of the collagen protein folding.

Each Chapter is constituted by three sections: *abstract*, *core*, and *perspectives*. The *abstract* is a very brief summary of the content of the *core* section, where the topic of the chapter is discussed verbosely. The *perspective* section draws the research path inspired by the results discussed in the chapter. At the end of the thesis, I have included a schematic representation of the work done as a **graphical summary**, to facilitate a bird-eye view of the whole PhD study.

Bibliography

1. Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **120,** 1955–1958 (2007).

CHAPTER I. Introduction

Abstract

In this chapter, which is almost entirely based on literature data, I have summarized a detailed description of collagen protein and its simplest assembly (tendon). Moreover, I also provided a brief description of hydroxyapatite mineral (Hap) and its inclusion in collagen matrix (to create bone material). Most of the effort was spent in the description of the rules guiding the collagen triple helix stability.

1.1 Collagen

Collagen (COL) is a protein family, which includes 28 members to date. COLs are made by three identical (homo-trimeric) or different (hetero-trimeric) polypeptide chains.[1,2] These chains are wrapped together giving rise to the well-known collagen triple helix, see Figure 1.

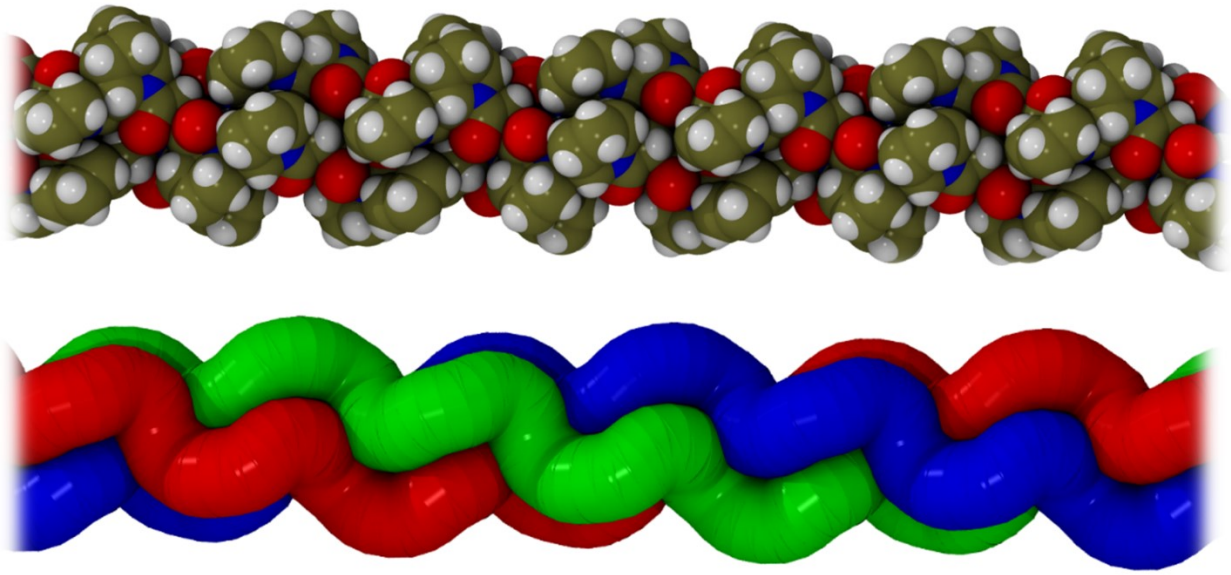


Figure 1. Collagen triple helix as van der Waals radii (top) and with single strand represented in different colours (bottom). Colour code: Carbon/cyan, Oxygen/red, Nitrogen/blue and Hydrogen/light-grey.

The various proteins of the family are labelled with roman numerals (I, II, etc.) in the chronological order of their discovery and grouped in categories mainly based on: (i) their propensity of associating in well-organized supra-molecular assemblies, e.g. collagen fibrils; (ii) the number of triple helical domain interruptions.[2]

The fibril-forming collagens present mainly the triple helical domain with only two shorts non-helical regions at the ends, called telo-peptides. The ~90% of human collagen is fibril-forming collagen (collagen type I, II, III, V and XI). These are found abundantly in vertebrate tendons, bones, skin, cartilage, cornea, blood vessels, etc. Regarding humans, "collagens comprise one-third of the total protein, accounting for three-quarters of the dry weight of skin".[2] Remarkably, collagen was found in a 68 Myr old *Tyrannosaurus rex* fossil and from a 78 Myr old *Brachylophosaurus canadensis* fossil. Disregarding the discovery uncertainties due to potentials cross-contaminations,[3–6] collagen is the oldest protein detected to date.

As mentioned before, the **collagen triple helix** motif defines the entire family of collagens. A seminal work on COL structure determination was first proposed by Ramachandran and Kartha in 1954,[7] and then refined in the later years[8] by the contribution of Rich and Crick[9] and Cowan et al.[10] In each individual collagen protein, usually called tropocollagen (TC), the helical domain is composed by three parallel polypeptides strands wrapped together in right handed helix. Each strand has a poly-proline or poly–glycine type II (**PPII**) geometry, which is a left-handed helix with 3 aminoacids per turn, see Figure 2.[11–13]

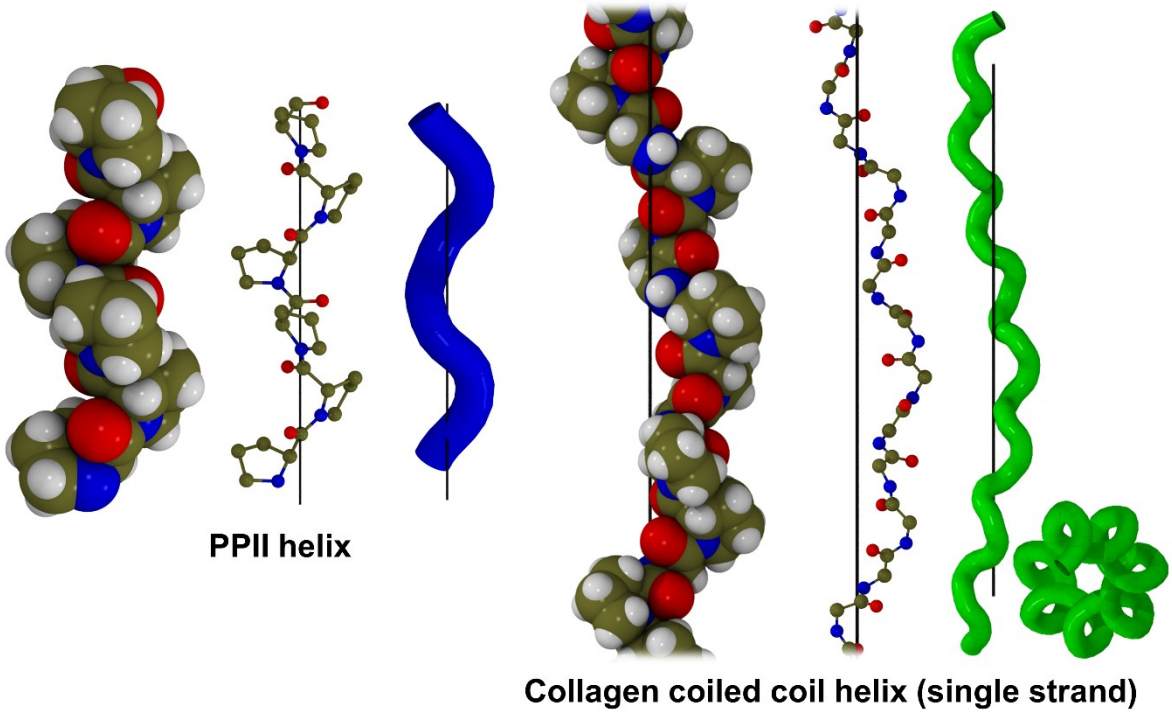


Figure 2. Left: single polypeptide chain (poly-Pros) in PPII conformation in a vdW, ball and stick, and tube representation. Right: single collagen chain (coiled coil) in vdW, ball and stick (backbone atoms only), and tube (side and top views) representation.

For each strand, the primary structure follows a strict repetitive pattern of triplets where Glycine (Gly) occurs always in the same position (**Gly-X-Y**). The aminoacids in X and Y positions are exposed to solvent and Gly is buried inside the helix core, Figure 3 A-B. The three strands wrap together with one residue staggering for having a close packed structure, Figure 3 C-D.

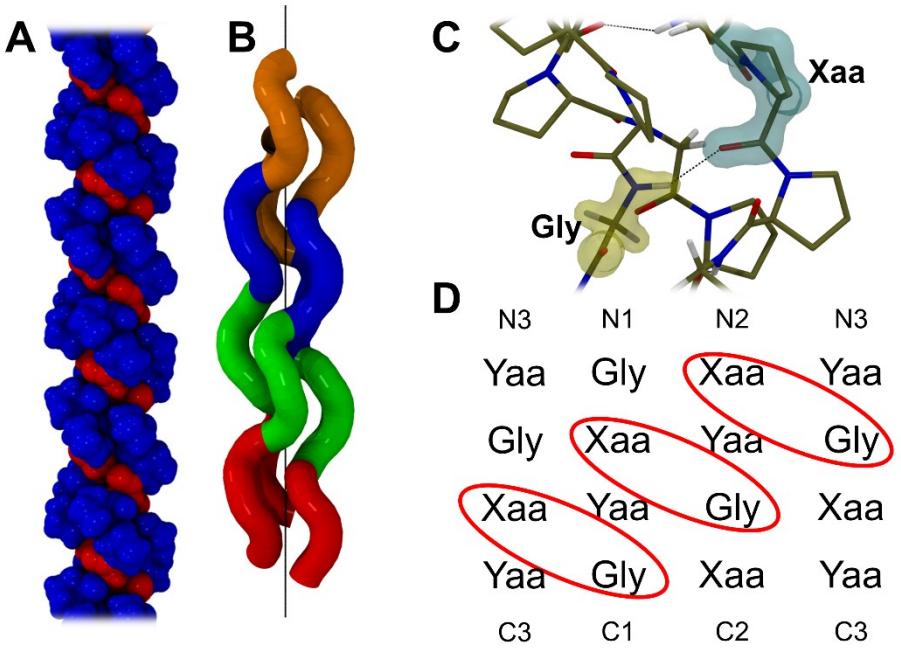


Figure 3. A: collagen side view with Glys in red and X and Y amino acids in blue. B: collagen side view (tube) with each Gly-X-Y triplet coloured differently to highlight the staggering between the chains. C: detail on the inter-chain H-bond. D: H bond pattern between poly peptides chains, which

are numbered. C and N indicate the direction of the chain (C- N- termini) and the red circle the H bonds.

For its small size Gly is the only L-aminoacid that would fit in the core of the helix. Indeed, the side chain for a L-aminoacid points towards the center of the triple helix, and any side chain bigger than H (Gly case) would cause a weakening of the structure for steric hindrance. The **prerequisite of Gly** in each triplet is not related to its small size only.[14] Indeed, if D aminoacid replaces Gly, the side chain would point away from the core of the protein and it should result in a stable triple helix.[14] In fact, D aminoacid substitution on Gly position leads unstable triple helices. This happens because the main chain dihedral preorganization of Gly is fundamental to give stable triple helices, along with the small steric hindrance at the helix core.[15] The three peptide strands are kept together by H-bonds between the carboxylic group of the aminoacids in the second position and the N-H of Gly, Figure 3 C-D.[2] Proline (Pro) and Hydroxy-4R-Proline (Hyp) are the most common aminoacids in the second and the third position (28% and 38%, respectively). In all collagens, the Gly-Pro-Hyp (GPO) triplet is the most common triplet (10.5%),[16] and Pro and Hyp represent the ~22% of all residues. The **high content of Pro** and derivatives is thought to give stability to the triple helix. This type of stabilization is manly entropic, pre-organizing the individual strands to a PPII geometry in the unfolded state, hence to the geometry more similar to the folded triple helix than a random conformation.[1,2]

The **side-chain pucker of the proline and derivatives** (pyrrolidine ring) plays an important role in the stability of COL.[16] Pro is known to have two stable side chain conformations, hereafter indicated as DOWN (D) and UP (U), Figure 4 A.

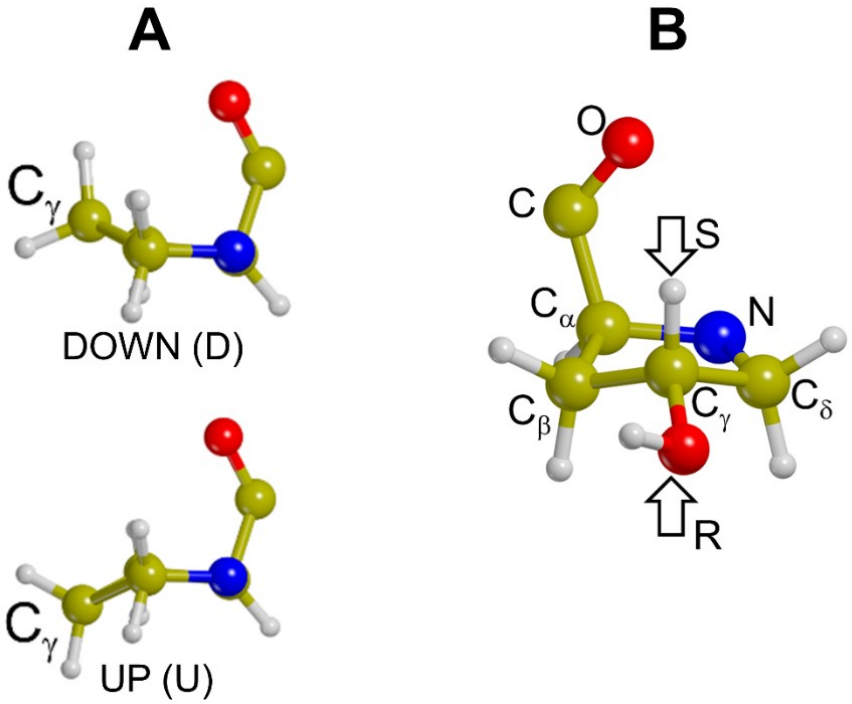


Figure 4. A: UP (U) and DOWN (D) Prolines. B: L-Hydroxy-4R-Proline. All the atoms are labelled and S and R positions are shown.

For *trans* prolines in gas phase,[17,18] and in proteins,[19] flipping from one pucker to another costs ~1 kcal·mol^{-1}, in favour of the D state, thus both state are almost equally populated at room temperature (RT). **Interestingly, electron-withdraw substituents on the Cγ of the pyrrolidine ring can alter the Pro puckering propensity**. Depending on the position of the electron-withdraw substitution it is possible to enhance the propensity for a U or D pucker. In details, electron-withdraw substituents on R/S stabilizes U/D puckering in Pros, see Figure 4 B. Moreover, the more the substituent is electronegative the more the stabilization is pronounced.[20] Hyp is a natural example of Cγ substitution (with OH group) which is found in all collagens. In this case, the OH substitution

happens on R position leading to a stabilization of the U pucker. The enzymatic hydroxylation of the Pros in Y position takes place after collagen biosynthesis. This step is fundamental for collagen stability, indeed the hydroxylation process enhance the temperature of unfolding of collagen type I just above the human average body temperature.[1] A similar effect is found for Pros in X position. In this very case, a Cγ substitution stabilizing the D puckering (electron-withdraw substituents in S position) increases the triple helix stability. A possible explanation for these phenomena derives from **the strong correlation between the main dihedral angle of the Pros and its side chain conformation** (peculiar of Pros having a five-member-ring as side chain). Pros in D state have values of $\phi \sim -69.8°$ while Pros in U state exhibit $\phi \sim -58.7°$.[19] The expected ϕ values of Pros match closely the observed ones in the collagen triple helix: $-72° \pm 6°$ and $-58° \pm 4°$ in X and Y position, respectively (average values from collagen peptide crystal structures at high resolution).[1] **Therefore, Pro and derivatives pre-organized the main chain dihedral angles to fit into a collagen triple helix geometry without deformation and entropic cost.**[1,18,21] Interestingly, the preorganization of Pros stabilizes more the collagen helix when takes place in Y rather than X position, for several reasons: (i) Pro is slightly more favored in the D pucker rather than the U pucker. So the substitution would just reinforce the innate propensity of Pro to D pucker; (ii) forcing the D pucker in Pro stabilizes the Pro isomer with *cis* peptide bond thus destabilizing the triple helix.[19,22,23] Quoting reference:[2] "In collagen, all peptide bonds are *trans*, thus, before a (Gly-X-Y) strand can fold into a COL triple helix, all the *cis* peptide bonds must isomerize to *trans*".

In summary, Pro derivatives on Y/X positions with enhanced tendency to U/D pucker stabilize the triple helix. The opposite is true as well, Pro derivatives on Y/X position with enhanced tendency to D/U pucker destabilizes the triple helix. This is known as **propensity based hypothesis**. Following this principle, they were able to synthetize several hyperstable collagen-like peptides.[1,2] One example is a homo-trimeric peptide made by three chains of Gly-Pro-Flp (GPF) triplets, where Flp stands for Fluoro-4R-Proline.[24] The Fluoro-substitution in 4R forces Pro to have U puckering and it is place in Y position, therefore, following the propensity hypothesis, GPF has a stable triple helix.

Unfortunately, this hypothesis does not work in all cases.[1,25,26] The propensity-based hypothesis take into consideration only **entropy**: **pre-organizing the aminoacids to the collagen helix conformation simplified the polypeptide coiled-coiling**. What about **enthalpy**? Several experiments have been performed to rationalize the stability of collagen models triple helix. For instance, the homo-trimeric GPO and GPF helices are over-stabilized (with respect to the reference homo-trimeric Gly-Pro-Pro (GPP) helix) by enthalpic phenomena in the first case and entropic ones in the second case.[27] The enthalpy stabilization arises from an increased solvent hydration due to the OH group of Hyp in GPO. The entropic stabilization of GPF arises from the pre-organization of the dihedral angles thanks to Flp.[27] **In general the collagen stability can be explained only with a clear understating of the entropic and enthalpic character of both triple helix and single coil systems to obtain a comprehensive explanation for the stabilizing mechanics of the triple helix**.[28] The propensity model provide a very simply way to assess the stability of a collagen like peptide, but sometimes it fails. Even if for all the breaking-rule cases a rational explanation has been provided,[1,29] we are far from a real comprehension of the phenomena guiding the folding/unfolding of collagen triple helices.

Despite the rate of published works on the origin of collagen stability slowed down considerably in the last decade, the general interest of scientific community in the synthesis of collagen like peptides has not fell off. Indeed, they demonstrated that it is possible to create stable triple helices by:

(i) changing a CH_2 group with a NH group (Gly → Aza-Gly) in the core of the triple helix increases the numbers of H-bonds between strands-,[30–32]

(ii) crosslinking (covalent bonding between collagen strands).[33]

Interestingly, introducing an amino-Pro (Pro with amino group as ring substituent) in collagen peptides creates helical structures that can fold and unfold reversibly in function of the solution pH.[34,35]

The last aspect of the collagen protein treated in this paragraph regards the triple helix geometry. Uncertainties on the geometry of the collagen triple helix appeared in 1977 when Okuyama et al. proposed a new structural model for the collagen protein.[36] The new model was coherent with the previous one[7–10] on most of the aspects, but it presented some differences in the helical pattern. In the first place collagen was thought as a 10_3 helix,[37] while this new model refined it to a 7_2 helix.[36,38] These two different typology of helix are associated to two different radial projections which are reported in Figure 5.

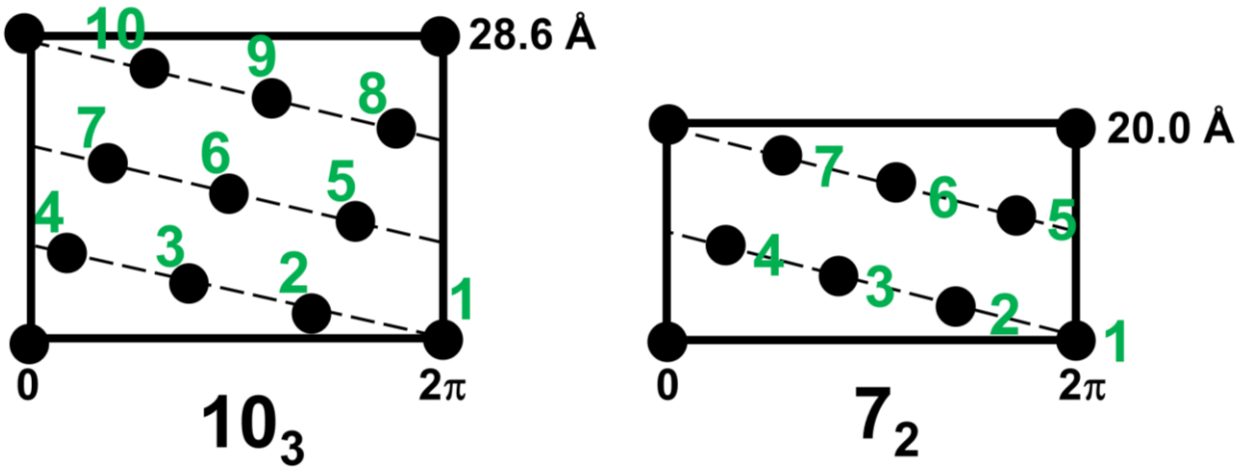


Figure 5. Radial projections for the 10_3 and 7_2 helices. Each black point represent a scattering unit, e.g. an aminoacid triplet (Gly-X-Y). The axis of abscissas is in angular unit, the axis of ordinates is in spatial unit.

In the 10_3 helix, the number of scattering units within a translational repeat is 10 while in the 7_2 case it is 7. Following the dashed line in Figure 5, the helix makes three turns within a translational repeat in the former case and just two turns in the latter. The careful reader has already understood the meaning of the 10_3 and 7_2 notation. The uppercase number indicate the number of units in the radial projections and the lowercase one defines the number of helix turns. The translational repeat length, that is the length in which the helix is translationally repeated, is 28.6 Å and 20.0 Å for the 10_3 and 7_2 helices, respectively. The helices in Figure 5 are both left handed, but the resulting collagen triple helix is right handed. This arises from the contradiction between the way the diffraction units are linked in the radial projections and the way the aminoacids are covalently bonded in the protein. To deepen our understanding on this point, I reported (in Figure 6) the radial projection for the 10_3 helix, highlighting the peptide bond connection. Now, to make a COL triple helix, we need three peptide strands that are reported with different colors. Each single peptide strand has a 10_1 helicity with translational repeat of 85.8 Å. As it should be, the overall structure is now right handed.

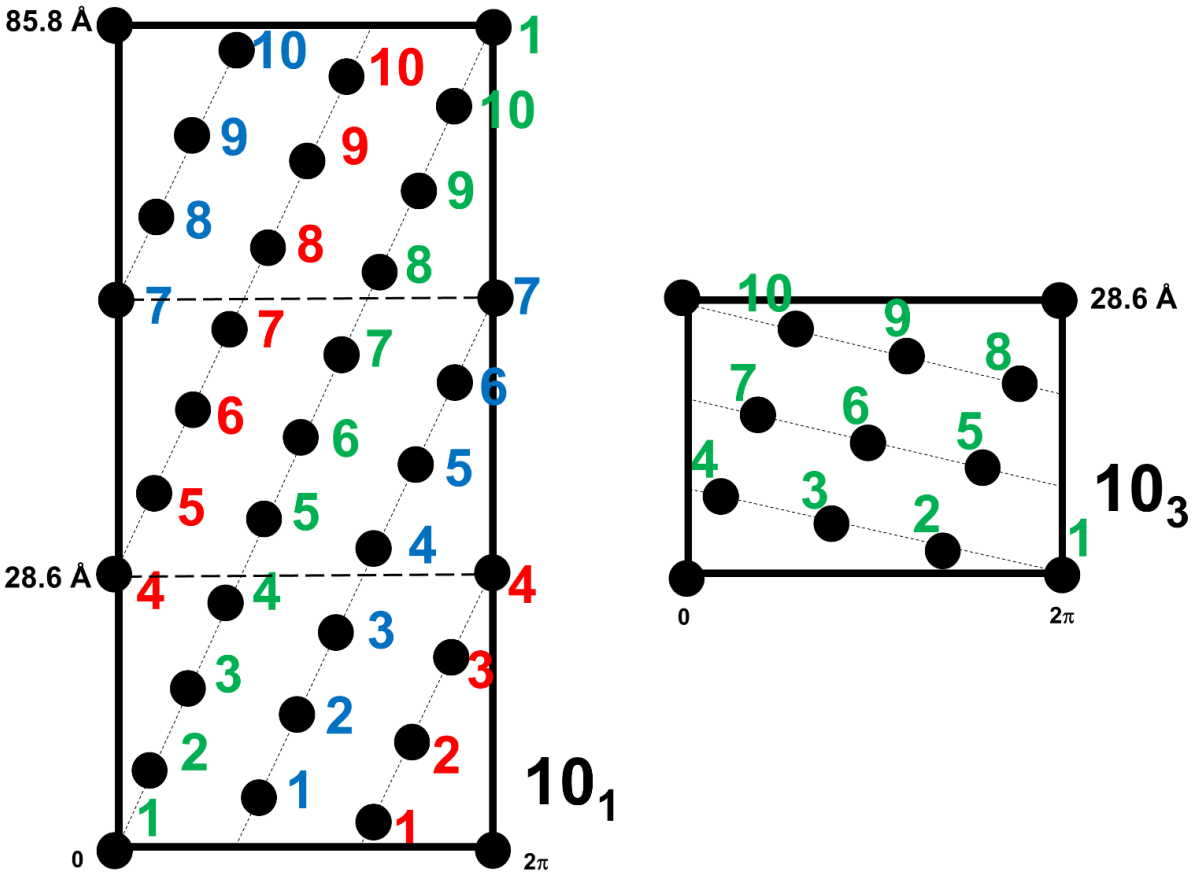


Figure 6. Radial projections for the 10_1 (left) and 10_3 (right) helices. The former is three times longer than the latter.

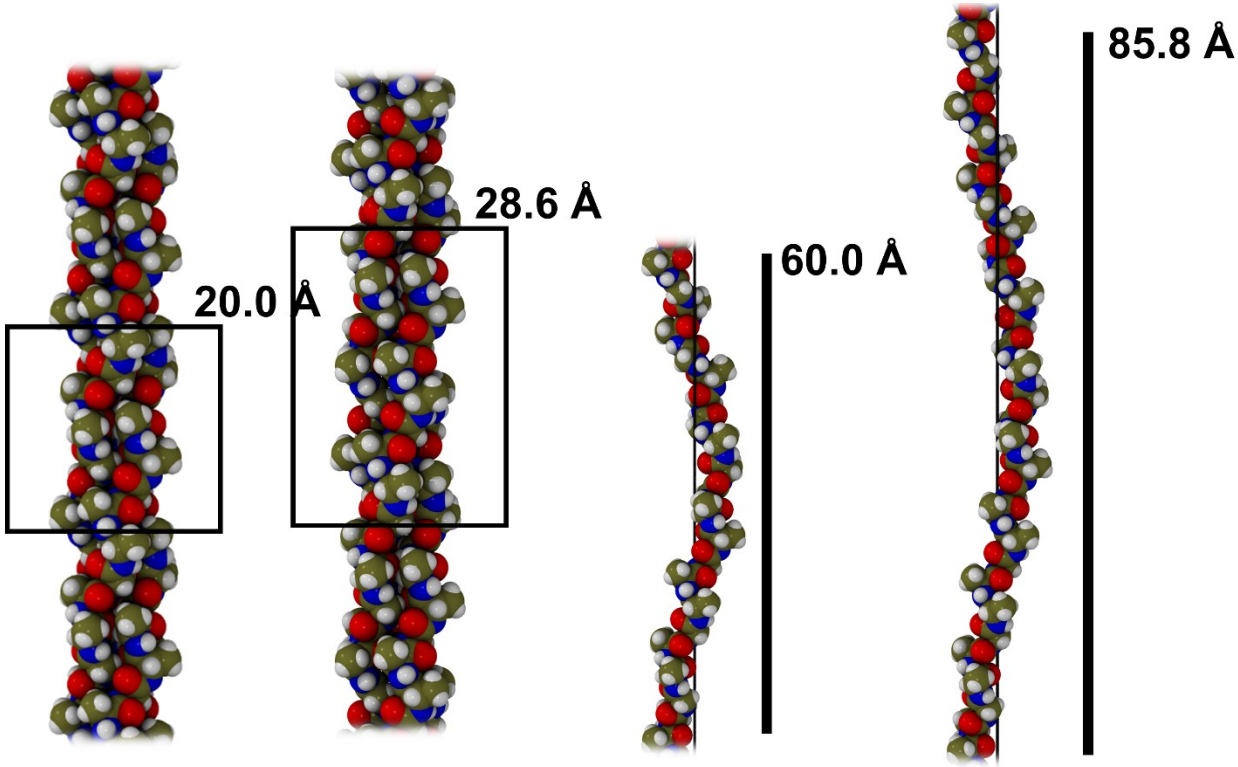


Figure 7. 10_3 and 7_2 triple strand helices (on the left) and the corresponding single strands (on the right).

A first glance at Figure 7, 10_3 and 7_2 triple helices look very similar. The differences become clear when a single COL strand is isolated from the triple helix geometry (Figure 7 right). In the 10_3 geometry, each individual chain repeats after 10 triplets with a translational length of 85.6 Å. Each triplet is separated from the other by a rotation of 36° (360°/10), and a translation of 8.56 Å (85.6 Å/10) along the helix axes. In the 7_2 geometry, each individual chain repeats after 7 triplets with a translational length of 60.0 Å. Each triplet are separated by the faster rotation of 51.4° (360°/7), and a translation of 8.57 Å (85.6 Å/10) (Figure 7). The important point is that **there is agreement in the scientific community on the collagen geometry: in Pro-rich regions the collagen protein is a 7_2 helix, in Pro-free regions it is a 10_3 helix**.[1,2,39,40]

1.2 Tendon

Tendons are tough bands of fibrous tissues whose purpose is to transmit muscular force to bones.[11] Tendons have a hierarchical structure with **collagen** as basic building block. Collagen single protein, called tropo-collagen, is 300 nm long and it assemblies into larger aggregates called **fibrils**, ~1 mm long. In fibrils the tropo-collagen proteins are covalently bonded (cross-linked) by glycosaminoglycan, which are long unbranched polysaccharides, to form **fibres** (ranging from 1 to 300 mm long). In the upper-level, fibres gather together into **fascicles**. Finally, more fascicles (on a centimeter scale), surrounded by a connective layer, form the whole tendon tissue.[11,41] Figure 8 graphically summarize the hierarchical organization of tendon.

Collagens contribute up to 80% of tendon dry weight. The second main component of tendons is water (up to ~70% in weight). Among the different type of collagen, type I is the most abundant (95%) and only small amounts of types III, V, XII and XIV are found.[11,42,43]

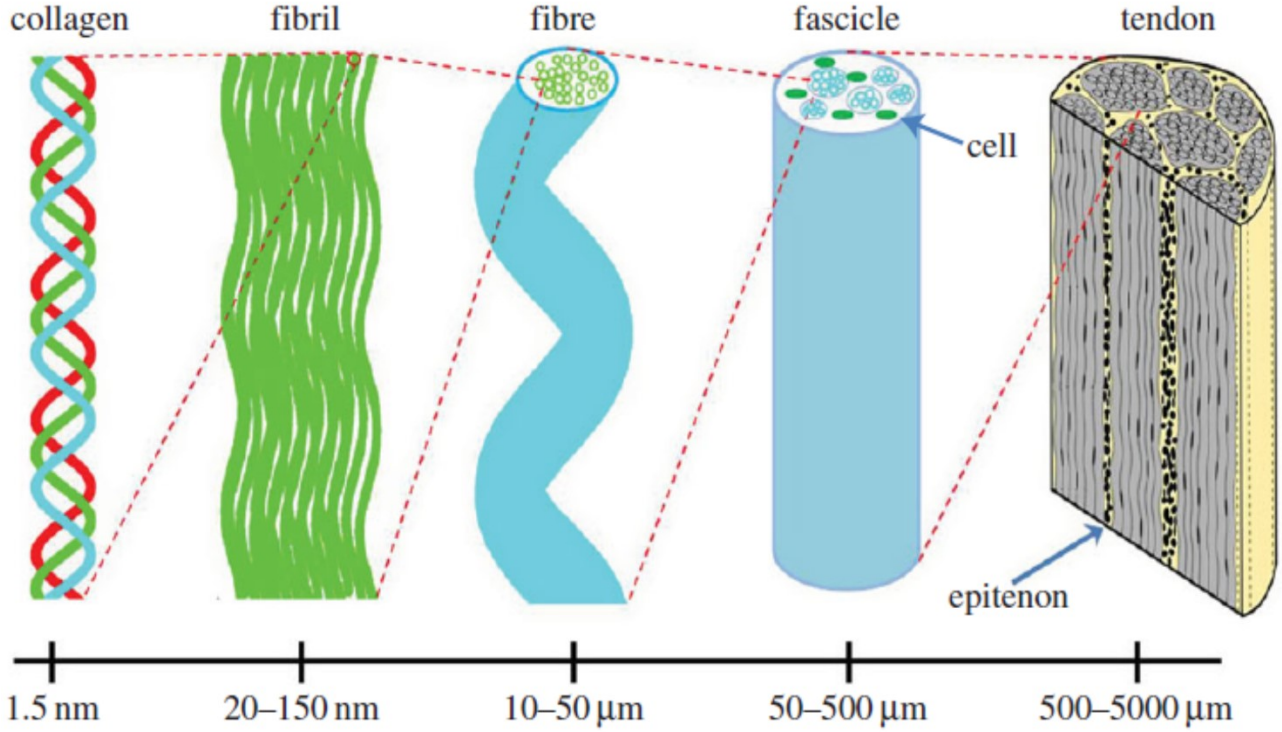


Figure 8. Hierarchical structure of tendon. The lateral dimension of the collagen aggregates is on the abscissa axis. Image taken from Ref. [44].

Here, I briefly resume some of the models that has been proposed for describing the most simple collagen aggregate, e.g. the tendon fibril. A peculiar feature of tendon fibril is a periodic pattern observable at the transmission electronic microscope (TEM). This structure is known since '40 and it is called **D periodicity**.[45,46] The periodic pattern consists in a repetition of matter-poor (0.54·D) and matter-rich (0.46·D) areas, where D stands for ~67 nm. This two different zones of the fibril, called gap and overlap, correspond to the lighter and darker areas of Figure 9, respectively.

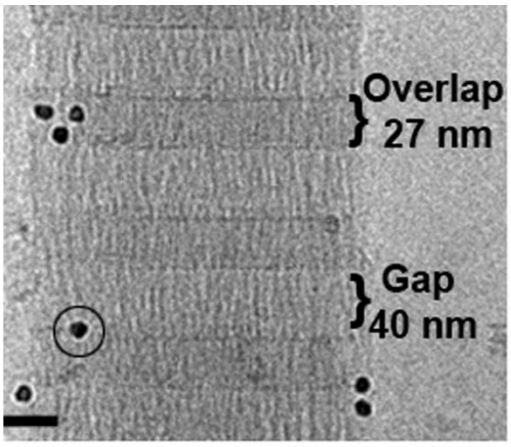


Figure 9. Transmission electron microscopy image of a collagen fibril taken from Ref. [47].

How to explain this periodic pattern, considering that the tropocollagen length is not multiple of D? (It is indeed 4.46 times D). In 1964, **Hodge and Petruska** proposed a model where five tropocollagen molecules were staggered side-by-side, with neighbors offset by 67 nm (D), as reported in Figure 10 A-B.[48] It is evident, by looking to Figure 10 A, that the model reproduces the fibril D pattern revealed by TEM. This bi-dimensional proteins arrangement is replicated in the perpendicular direction, ending up in a tri-dimensional layered structure, Figure 10 C. Few years later, **Smith**, who was dubious on the stacked layers arrangement, proposed an alternative model for the collagen fibril. In this new model, five D staggered tropocollagens were disposed concentrically into a filament with empty space inside. This supramolecular assembly took the name of **micro-fibril**, which is reported in Figure 11.[49] This model left some uncertainties on the later packing between micro-fibrils and so that, in the following years, the micro-fibril existence was questioned.[50] Ten years ago there was a breakthrough on the determination of the collagen fibril structure by **Orgel** et al. who resolved the electron density map from X-ray diffraction pattern of collagen unit cell within a tendon fibril. They came out with a model confirming the micro-fibril organization of tropocollagens. Each micro-fibril is made by five tropocollagen molecules, which are (see Figure 12 A):

- axially D staggered (which gives the observed macroscopic D periodicity),
- super-twisted with right handedness,
- covalently bonded (cross-linked) with collagens of neighboring micro-fibrils

Indeed, micro-fibrils are inter-connected by covalent bonds and this provides a rationale to the difficulty of isolate a single tendon micro-fibril. This model is the first one to consider the collagen N- and C-terminal (telopeptides) segments contained within neighboring micro-fibrils, see Figure 12 B.

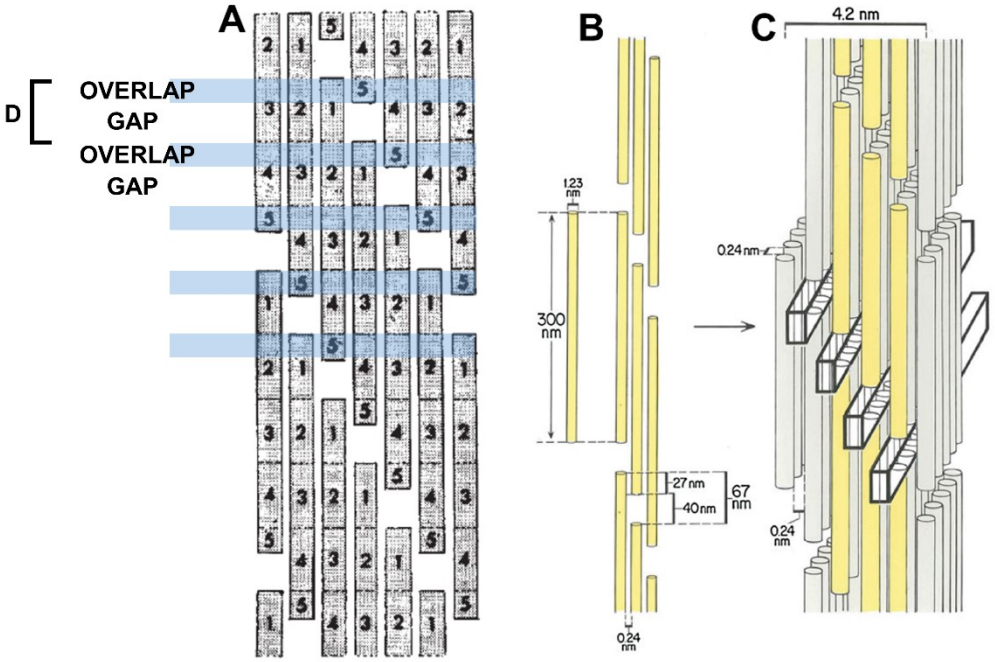


Figure 10. Graphical representation of the Hodge and Petruska model. Images taken from Ref. [49,51].

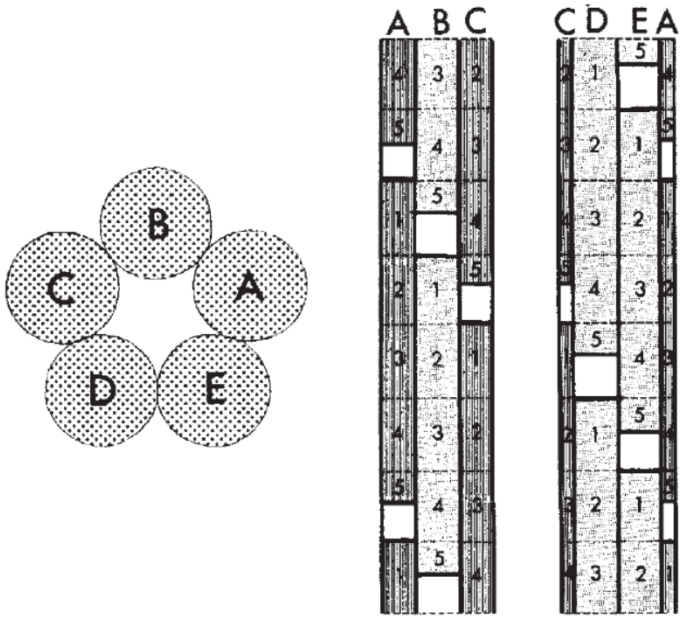


Figure 11. Graphical representation of the Smith model. Top view in the overlap region on the left, and side view on the right. A-D are the labeling of the five collagen proteins defining the micro-fibril structure. Images taken from Ref. [49].

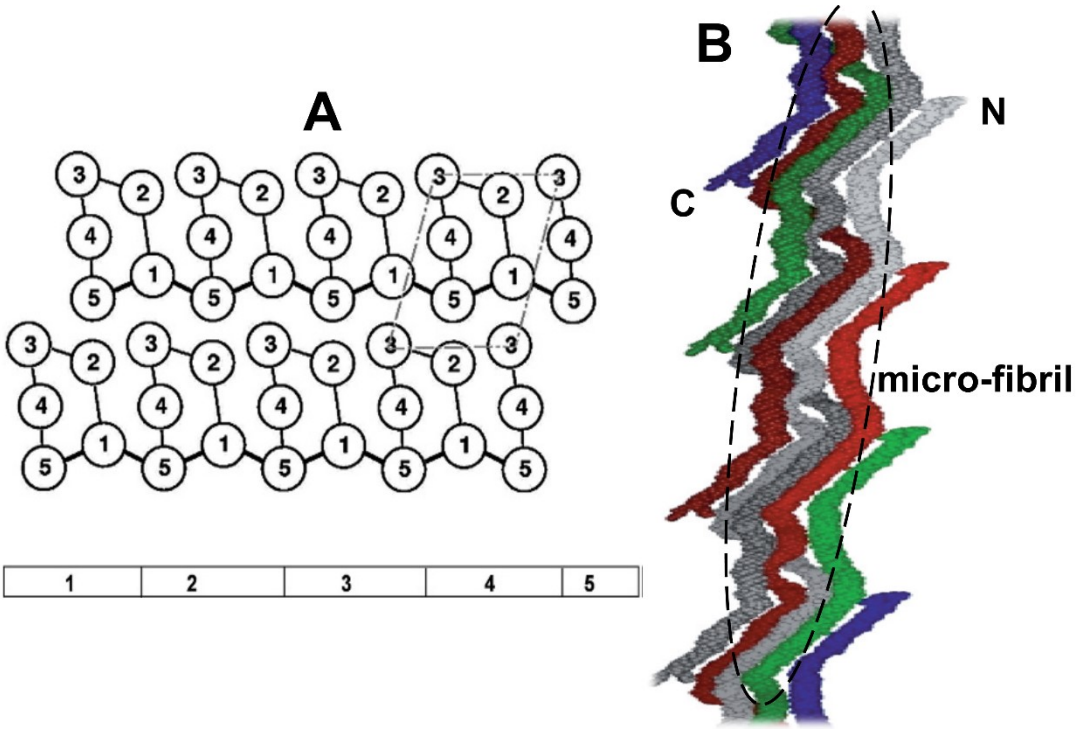


Figure 12. A: On the bottom, a tropocollagen molecule divided by five segments. Each segment is D long but the fifth one which is assumed half the length of other segments (0.46·D). On the top, the cross-section of a collagen fibril in the overlap region highlighting the in-plane inter-protein interactions. The cross-linkage are reported as thicker line. The thin in-plane linkage suggests the helical climb of contiguous protein segments. For instance, displacing of D the collagen segment 1 along the fiber axis (normal to the paper plane), it gets in position 2, due to the right hand supertwist of the fiber. Image taken from Ref. [11]. B: Isolated micro-fibril structure from Orgel's model. It looks clear that the C and N terminal section are outside the micro-fibril structure. Image taken from Ref. [52].

Interestingly, collagen type I is not stable at room temperature,[53] therefore, its self-assembly into fibrils prevents the inevitable protein denaturation. According to Ref. [54]: "The assembly of collagen molecules into fibrils is an entropy driven process, similar to that occurring in other protein self-assembly systems, such as microtubules, actin filaments and flagella. These processes are driven by the loss of solvent molecules from the surface of protein molecules and result in assemblies with a circular cross-section, which minimizes the surface area/volume ratio of the final assembly. Although the broad principles of collagen fibril self-assembly are generally accepted, **less is known about the molecular mechanisms of the assembly process**". The process of self-assembly is spontaneous in vitro,[55] and, following the speculation of someone, it is enthalpically driven by London dispersion forces.[56] The covalent cross-linking in the collagen fibril, *vide supra*, is a post-fibrillogenesis process of reinforcement, where the non-helical domains bind the helical domain of adjacent collagen molecules. This process is key in the determination of the mechanical property of the fibril.[11] In general fibril collagens contain higher percentage of charged residues (Lys, Arg, Asp Glu) (~20%) than of hydrophobic residues (~6%).[57] Analyzing the primary structure of collagen, it was demonstrated that the D staggering maximize the interaction between the hydrophobic residues of COLs, which gives credit to the hypothesis of a hydrophobic driven fibrillogenesis.[58] Interestingly, a similar pattern for charged residues is not found. Although charged residues in water environment can strongly bond with the solvent, most of the times the charged residues in collagen type I sit close in the aminoacidic sequence. This vicinity may lead to intra-protein charge interactions between chains, which is a standard phenomenon in viruses' collagens. Indeed, in these *alien* collagens the most common triplet is not Gly-Pro-Hyp but Glycine, Aspartic acid (negative charged) and Lysine (positive charged).[1]

1.3 Bone

Bone is composite material made by an organic matrix combined with a mineral. The main organic component is **collagen** protein and water molecules, and the mineral component is **hydroxyapatite** (Hap). In fresh bones, water constitutes around 25% of the weight. In dry bones, the mineral accounts for ~75% of bone weight, and one-quarter of weight corresponds to collagen protein along with lipids and other proteins. **In general, a composite is a material where two or more dissimilar phases combine to produce a materials with properties superior to both the starting components**. In the bone case, the properties that are optimized in the composite formation are the **mechanical properties**.

Since now, in this PhD thesis, I have never mentioned the concept of mechanical properties of a material. I have always looked to the things from the chemical perspective rather than from the engineering one for a limited background on the topic. I have started studying the mechanical property of tendons and bones models during my last part of the PhD work, but it is still at an embryonic stage. Considering the importance of the mechanical property for both the bone features and as a research field, as written also in the next Chapter, I will now briefly describe the mechanical feature of bone without any presumption of being comprehensive. Most of the following information are taken from Ref. [59,60]

Three important properties for describing the mechanics of a material are: **stiffness, strength and toughness**. The exact definition of these term is not clear from literature,[59] and so I will provide a qualitative explanation for each of them with the help of the **stress–strain curve** (Figure 13) which reports the deformation occurring in a material under the action of a mechanical force:

- **Stiffness** is the rigidity of an object. When a force is applied to a material the more it resist to deformation the more it is stiff. It is common to associate the Young's modulus to the stiffness of a material, see Figure 13 A. The **Young's modulus** is the slope of the stress-strain curve in the elastic regime (the stress/strain curve has a linear trend).
- **Strength** is the ability of an object to withstand to stress without failure, see Figure 13 B.
- **Toughness** is the ability of an object to absorb energy without failure. In general the measurement of toughness may be taken graphically from the stress–strain curve as the area under the curve, see Figure 13 C. The larger the area under the curve, the higher the toughness of the material.

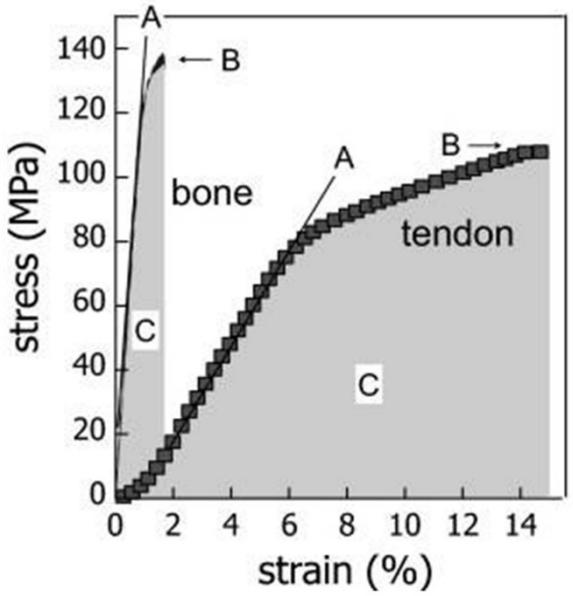


Figure 13. On the left: General stress-strain curve of bone and tendon materials. A: Slope of the stress-strain curve in the elastic regime which defines the material Young modulus. B: A material strength is the maximum stress it can sustain before failure. C: The shaded area under the curve can be a measure for toughness (the energy needed to induce the failure of a material). The stress/strain curve of the Hap mineral is not reported, but it would be much steeper than the bone one. Image taken from Ref. [59].

Hap crystal is a stiff and strong material, but not very tough, with a Young's modulus around 114 GPa.[61] Conversely, collagen is tougher but less stiff, with a Young's modulus of 0.4-12 GPa.[62] Moreover, when organized in fibrils, collagen stiffness can even reduce of several order of magnitude.[63] **The composite remarkably combines the properties of the two components: the stiffness of the mineral and the toughness of the organic matrix**.[64] The composite stiffness is always between the lowest and the highest components values. For instance, the cortical bone has a Young's modulus in the range of 10-20 GPa, which in the between of the Hap and COL ones. Water is also an important factor which has to be considered. Indeed, reducing the water content in bones increases the stiffness of the material. Unfortunately, the situation is much more complex when it comes to strength and toughness. Indeed, these quantities depend on the much more entangled interfacial phenomena.

The reasons of the mechanical property of bone are hidden within the hierarchical multicomponent organization of its structure. *For the purpose of this thesis I am interested at the atomic scale phenomena occurring in bone only. Therefore, I will not describe its entire hierarchical organization, as done for tendon and the description will be restricted only to the basic building block of bone, e.g. the mineralized fibril, which micro-structure is shared by all different types of bones*. Although with a major difference, bones share the basic structural building block with tendons, e.g. the collagen fibril. Indeed, in the case of bone fibril, the collagen matrix is reinforced with the inclusion of Hap. Hap is a calcium phosphate mineral belonging to the crystal family of calcium apatite with chemical formula of $Ca_{10}(PO_4)_6(OH_2)$. In biological environment, it has a hexagonal crystal packing as shown in Figure 14.

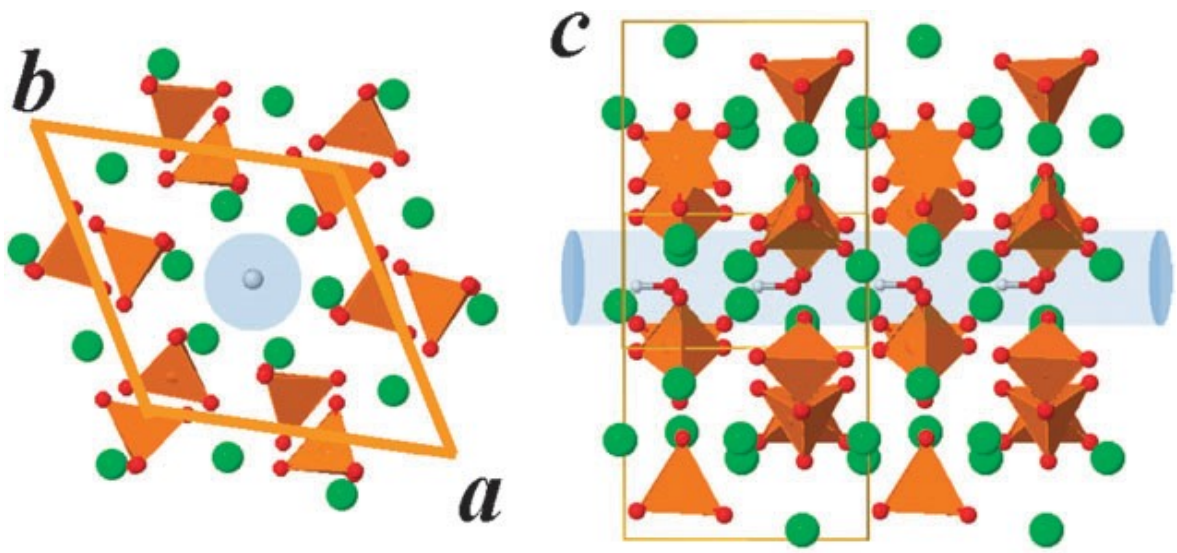


Figure 14. Top and lateral view of the Hap channel of hexagonal bulk Hap. The *a*, *b* and *c* labels refer to the crystallographic cell vector. Color code: Calcium/green, phosphate group/orange tetrahedron, Oxygen/red, Hydrogen/light gray. Image taken from Ref. [65].

Hap mineralizes in the form of flat nanoplatelets (3-4 nm thick),[51,66] both inside and outside the fibril,[67] with no clear preference. Within the fibril, the gaps between tropocollagens are sites for the calcium apatite nucleation, see Figure 15 A-B. The mineral grows mainly along the *c* crystallographic axis exposing the {010} surface.[68] The *c* axis (to which the OH channel is aligned to, see Figure 14) of all Hap nanoplatelets, lies parallel to each other and also to the collagen protein axes.[67,69] In human bone, the unit cell of the mineral contains substantial carbonate CO_3, nominally about 5% on weight. Recently, it was shown that physicochemical effects of carbonate integration within the apatite lattice may be the cause for the plate-like morphology of bone mineral in bones.[70]

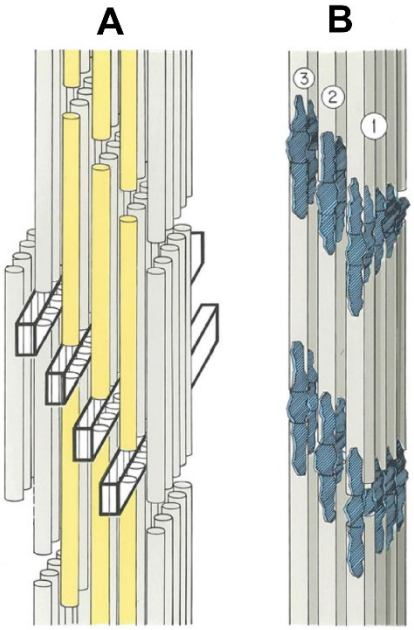


Figure 15. A: 3D organization of collagen within the fibril (Hodge and Petruska model). B: Intra-fibrillar crystallization of Hap in the gap region of the fibril. The obsolete Hodge and Petruska model of collagen fibril is used here because of its graphical clearance. Image taken from Ref. [67].

Perspectives

In the last decade, the scientific community interest on the **stability of collagen** like peptides fade away. In most of the cases, the propensity hypothesis predicts the stability of collagen models successfully, and, in the out-of-trend cases, studies were performed for revealing the general reasons for the unexpected behavior. Although much work has been done, an understating of the phenomena at the atomic scale in most of the cases is missing. For instance, the role of water solvation is still uncertain, also with respect to the biologically fundamental Hyp aminoacid.

To an upper-level, the molecular mechanisms of the **assembly** process of the **collagen aggregates** in fibrils, and the **interface phenomena between collagen and apatite crystals** in bones are not well known.

Ab initio simulation could give a rational to these phenomena, but **the extreme complexity of the system under question has been since now an insurmountable step**. For instance, the experimental works done to substitute Gly with a D aminoacid (which ended up to be impossible) were suggested by *ab initio* simulations (which came out to be wrong).[14] The next paragraph guides the reader though the state of the art of computer simulations on these scientific topics, so the scenario will be clearer.

Bibliography

1. Bella, J. Collagen structure: new tricks from a very old dog. *Biochem. J.* **473,** 1001–25 (2016).
2. Shoulders, M. D. & Raines, R. T. Collagen structure and stability. *Annu. Rev. Biochem.* **78,** 929–58 (2009).
3. Buckley, M., Warwood, S., van Dongen, B., Kitchener, A. C. & Manning, P. L. A fossil protein chimera; difficulties in discriminating dinosaur peptide sequences from modern cross-contamination. *Proc. R. Soc. B Biol. Sci.* **284,** 20170544 (2017).
4. Schroeter, E. R. *et al.* Expansion for the Brachylophosaurus Canadensis Collagen I Sequence and Additional Evidence of the Preservation of Cretaceous Protein. *J. Proteome Res.* **16,** 920–932 (2017).
5. Schweitzer, M. H. *et al.* Analyses of Soft Tissue from Tyrannosaurus rex Suggest the Presence of Protein. *Science* **316,** 277–281 (2007).
6. Schweitzer, M. H. *et al.* Biomolecular Characterization and Protein Sequences of the Campanian Hadrosaur B. canadensis. *Science* **324,** 626–631 (2009).
7. Ramachandran, G. N. & Gopinath, K. Structure of Collagen. *Nature* **174,** 269–270 (1954).
8. Ramachandran, G. N. & Gopinath, K. Structure of Collagen. *Nature* **176,** 593 (1955).
9. Rich, A. & Crick, F. H. C. The molecular structure of collagen. *J. Mol. Biol.* **3,** 483–506 (1961).
10. Cowan, P. M., Mcgavin, S. & North, a C. The polypeptide chain configuration of collagen. *Nature* **176,** 1062–1064 (1955).
11. Fratzl, P. *Collagen. Structure and Mechanics*. (Springer Berlin / Heidelberg, 2008).
12. Crick, F. H. C. & Rick, A. Structure of Polyglycine II. *Nature* **176,** 780–781 (1955).
13. Arnott, S. & Dower, S. D. The structure of poly-l-proline II. *Acta Cryst. B* **24,** 599–601 (1968).
14. Tsai, M., Xu, Y. & Dannenberg, J. J. Completely geometrically optimized DFT/ONIOM triple-helical collagen-like structures containing the ProProGly, ProProAla, ProProDAla, and ProProDSer triads. *J. Am. Chem. Soc.* **127,** 14130–14131 (2005).
15. Horng, J.-C., Kotch, F. W. & Raines, R. T. Is glycine a surrogate for a D-amino acid in the collagen triple helix? *Protein Sci.* **16,** 208–215 (2006).
16. Ramshaw, J. A. M., Shah, N. K. & Brodsky, B. Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides. *J. Struct. Biol.* **122,** 86–91 (1998).
17. Kapitán, J., Baumruk, V., Kopecký, V., Pohl, R. & Bouř, P. Proline zwitterion dynamics in solution, glass, and crystalline state. *J. Am. Chem. Soc.* **128,** 13451–13462 (2006).
18. Cutini, M., Corno, M. & Ugliengo, P. Method Dependence of Proline Ring Flexibility in the Poly-L-Proline Type II Polymer. *J. Chem. Theory Comput.* **13,** 370−379 (2017).
19. Vitagliano, L., Berisio, R., Mastrangelo, A., Mazzarella, L. & Zagari, A. Preferred proline puckerings in *cis* and *trans* peptide groups: implications for collagen stability. *Protein Sci.* **10,** 2627–32 (2001).
20. Improta, R., Benzi, C. & Barone, V. Understanding the role of stereoelectronic effects in determining collagen stability. 1. A quantum mechanical study of proline, hydroxyproline, and fluoroproline dipeptide analogues in aqueous solution. *J. Am. Chem. Soc.* **123,** 12568–12577 (2001).
21. Vitagliano, L., Berisio, R., Mazzarella, L. & Zagari, A. Structural bases of collagen stabilization induced by proline hydroxylation. *Biopolymers* **58,** 459–464 (2001).
22. Sarkar, S. K., Young, P. E., Sullivan, C. E. & Torchia, D. a. Detection of *cis* and *trans* X-Pro peptide bonds in proteins by 13C NMR: application to collagen. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 4800–4803 (1984).
23. DeRider, M. L. *et al.* Collagen stability: Insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations. *J. Am. Chem. Soc.* **124,** 2497–2505 (2002).
24. Holmgren, S. K., Taylor, K. M., Bretscher, L. E. & Raines, R. T. Code for collagen's stability

deciphered. *Nature* **392,** 666–667 (1998).
25. Motooka, D. *et al.* The triple helical structure and stability of collagen model peptide with 4(S)-hydroxyprolyl-Pro-Gly units. *Biopolymers* **98,** 111–121 (2012).
26. Hodges, J. A. & Raines, R. T. Stereoelectronic and steric effects in the collagen triple helix: Toward a code for strand association. *J. Am. Chem. Soc.* **127,** 15923–15932 (2005).
27. Nishi, Y. *et al.* Different effects of 4-hydroxyproline and 4-fluoroproline on the stability of collagen triple helix. *Biochemistry* **44,** 6034–6042 (2005).
28. Kawahara, K. *et al.* Effect of hydration on the stability of the collagen-like triple-helical structure of [4(R)-hydroxyprolyl-4(R)-hydroxyprolylglycine]10. *Biochemistry* **44,** 15812–15822 (2005).
29. Shoulders, M. D., Kotch, F. W., Choudhary, A., Ilia A. Guzei & Raines, R. T. The Aberrance of the 4S Diastereomer of 4-Hydroxyproline. *J. Am. Chem. Soc.* **132,** 10857–10865 (2010).
30. Zhang, Y., Malamakal, R. M. & Chenoweth, D. M. Aza-Glycine Induces Collagen Hyperstability. *J. Am. Chem. Soc.* **137,** 12422–12425 (2015).
31. Kasznel, A. J., Zhang, Y., Hai, Y. & Chenoweth, D. M. Structural Basis for Aza-Glycine Stabilization of Collagen. *J. Am. Chem. Soc.* **139,** 9427–9430 (2017).
32. Zhang, Y., Herling, M. & Chenoweth, D. M. General Solution for Stabilizing Triple Helical Collagen. *J. Am. Chem. Soc.* **138,** 9751–9754 (2016).
33. Hentzen, N. B., Smeenk, L. E. J., Witek, J., Riniker, S. & Wennemers, H. Cross-Linked Collagen Triple Helices by Oxime Ligation. *J. Am. Chem. Soc.* **139,** 12815–12820 (2017).
34. Siebler, C., Erdmann, R. S. & Wennemers, H. Switchable proline derivatives: tuning the conformational stability of the collagen triple helix by pH changes. *Angew. Chem. Int. Ed. Engl.* **53,** (2014).
35. Egli, J. *et al.* pH-Responsive Aminoproline-Containing Collagen Triple Helices. *Chem. - A Eur. J.* **23,** 7938–7944 (2017).
36. Okuyama, K., Takayanagi, M., Ashida, T. & Kakudo, M. A New Structural Model for Collagen. *Polym. J.* **3,** 341–343 (1977).
37. Fraser, R. D. B., Macrae, T. P. & Suzuki, E. Chain Conformation in the Collagen Molecule. *J. Mol. Biol.* **129,** 463–481 (1979).
38. Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M. & Kakudo, M. Crystal and molecular structure of a collagen-like polypeptide (Pro-Pro-Gly)10. *J. Mol. Biol.* **152,** 427–443 (1981).
39. Okuyama, K., Xu, X., Iguchi, M. & Noguchi, K. Revision of Collagen Molecular Structure. *Biopolymers* **84,** 181 (2006).
40. Bella, J. A new method for describing the helical conformation of collagen: Dependence of the triple helical twist on amino acid sequence. *J. Struct. Biol.* **170,** 377–391 (2010).
41. Gautieri, A., Vesentini, S., Redaelli, A. & Buehler, M. J. Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. *Nano Lett.* **11,** 757–766 (2011).
42. Benjamin, M., Kaiser, E. & Milz, S. Structure-function relationships in tendons: A review. *J. Anat.* **212,** 211–228 (2008).
43. Birch, H. L. Tendon matrix composition and turnover in relation to functional requirements. *Int. J. Exp. Pathol.* **88,** 241–248 (2007).
44. Fang, F. & Lake, S. P. Modelling approaches for evaluating multiscale tendon mechanics. *Interface Focus* **6,** 20150044 (2016).
45. Gross, J. & Schmitt, F. O. The structure of human skin collagen as studied with the electron microscope. *J. Exp. Med.* **88,** 555–68 (1948).
46. Tomlin, S. G. & Worthington, C. R. Low angle X-ray diffraction patterns of collagen. *Nature* **175,** 811 (1955).
47. Nudelman, F., Lausch, A. J., Sommerdijk, N. A. J. M. & Sone, E. D. In vitro models of collagen biomineralization. *J. Struct. Biol.* **183,** 258–269 (2013).
48. Petruska, J. A. & Hodge, A. J. A Subunit Model for the Tropocollagen Macromolecule. *Proc.*

Natl. Acad. Sci. **51,** 871–876 (1964).
49. Smith, J. W. Molecular Pattern in Native Collagen. *Nature* **219,** 157–158 (1968).
50. Hulmes, D. J. & Miller, A. Quasi-hexagonal molecular packing in collagen fibrils. *Nature* **282,** 878–880 (1979).
51. Landis, W. J., Song, M. J., Leith, A. & McEwen, L. Mineral and organic matrix interaction in normally calcifying tendon visualized in three dimensions by high-voltage electron microscopic tomography and graphic image reconstruction. *J. Struct. Biol.* **110,** 39–54 (1993).
52. Orgel, J. P. R. O., Irving, T. C., Miller, A. & Wess, T. J. Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci.* **103,** 9001–9005 (2006).
53. Leikina, E., Mertts, M. V., Kuznetsova, N. & Leikin, S. Type I collagen is thermally unstable at body temperature. *Proc. Natl. Acad. Sci.* **99,** 1314–1318 (2002).
54. Kadler, K. E., Holmes, D. F., Trotter, J. A. & Chapman, J. A. Collagen fibril formation. *J. Biochem.* **316,** 1–11 (1996).
55. Kadler, K. E., Hill, A. & Canty-Laird, E. G. Collagen fibrillogenesis: fibronectin, integrins, and minor collagens as organizers and nucleators. *Curr. Opin. Cell Biol.* **20,** 495–501 (2008).
56. Streeter, I. & De Leeuw, N. H. Atomistic modeling of collagen proteins in their fibrillar environment. *J. Phys. Chem. B* **114,** 13263–13270 (2010).
57. An, B., Lin, Y. S. & Brodsky, B. Collagen interactions: Drug design and delivery. *Adv. Drug Deliv. Rev.* **97,** 69–84 (2016).
58. Hulmes, D. J. S., Miller, A., Parry, D. A. D., Piez, K. A. & Woodhead-Galloway, J. Analysis of the primary structure of collagen for the origins of molecular packing. *J. Mol. Biol.* **79,** 137–148 (1973).
59. Wagermaier, W., Klaushofer, K. & Fratzl, P. Fragility of Bone Material Controlled by Internal Interfaces. *Calcif. Tissue Int.* **97,** 201–212 (2015).
60. Vaz, M. F., Canhão, H. & Fonseca, J. E. in *Advances in Composite Materials - Analysis of Natural and Man-Made Materials* (ed. Těšinova, P.) 195–228 (InTech, 2011).
61. Gilmore, R. S. & Katz, J. L. Elastic properties of apatites. *J. Mater. Sci.* **17,** 1131–1141 (1982).
62. Sun, Y.-L., Luo, Z.-P., Fertala, A. & An, K.-N. Stretching type II collagen with optical tweezers. *J. Biomech.* **37,** 1665–1669 (2004).
63. Grant, C. A., Brockwell, D. J., Radford, S. E. & Thomson, N. H. Tuning the elastic modulus of hydrated collagen fibrils. *Biophys. J.* **97,** 2985–2992 (2009).
64. Zimmermann, E. A. & Ritchie, R. O. Bone as a Structural Material. *Adv. Healthc. Mater.* **4,** 1287–1304 (2015).
65. Corno, M., Rimola, A., Bolis, V. & Ugliengo, P. Hydroxyapatite as a key biomaterial: quantum-mechanical simulation of its surfaces in interaction with biomolecules. *Phys. Chem. Chem. Phys.* **12,** 6309–6329 (2010).
66. Hassenkam, T. *et al.* High-resolution AFM imaging of intact and fractured trabecular bone. *Bone* **35,** 4–10 (2004).
67. Stock, S. R. The Mineral–Collagen Interface in Bone. *Calcif. Tissue Int.* **97,** 262 (2015).
68. Fratzl, P. *et al.* Mineral crystals in calcified tissues: A comparative study by SAXS. *J. Bone Miner. Res.* **7,** 329–334 (1992).
69. Handschin, R. G. & Stern, W. B. X-ray diffraction studies on the lattice perfection of human bone apatite (Crista Iliaca). *Bone* **16,** 355S–363S (1995).
70. Deymier, A. C. *et al.* Protein-free formation of bone-like apatite: New insights into the key role of carbonation. *Biomaterials* **127,** 75–88 (2017).

CHAPTER II. Computer Simulation on Tendon and Bone Materials

Abstract

This chapter includes:

- a brief description of the computational modelling as a complementary technique to study the matter at atomistic level.
- an overview of the theoretical frameworks used in this thesis, with emphasis to the London dispersion correction schemes and the "semi-empirical" HF-3c method.
- the description of the binding energy calculation for a molecule within a crystal and adsorbed on a surface.
- a state of the art bibliographic review of the computational modelling on tendon and bone materials to contextualize this PhD thesis work.

2.1 Computational Approach

As many of the concept reported in this chapter are now part of the background of any master student in chemistry and physics we chose not to give too many formal details to the discussion, limiting the focus on concepts. Because of the well-established matter, nothing in this chapter is original as it has been taken from standard books and publications from References.[1–4]

Computer modelling

"Computational chemistry/molecular modelling is the science (or art) of representing molecular structures numerically and simulating their behaviour with the equations of quantum and classical physic".[5]

The delicate step of any computational work is the choice of the model, e.g. the numerical representation of the chemical system of interest. For our purposes, we use compositions of atoms as models (molecular model). In this scenario, the main difficulty becomes providing the right atomic details for the correct representation of the real problem. We must also take into account that picking a molecular model conditions the level of theory, which is the physical framework used for the properties prediction. Generally, increasing the details in the model or the accuracy of the methodology increases the computational burden of the simulation. Therefore, it is important to find a balance between the complexity of the model and the accuracy of the level of theory, see Figure 1.

The theoretical frameworks are usually divided in three groups: *ab initio* methods, *semi-empirical* methods and *molecular mechanics*. *Ab initio* and *molecular mechanics* methods describe the atom behaviour by applying the law of quantum and classical physics, respectively. *Semi-empirical* methods are based on a rather drastic approximation of *ab initio* methods, which therefore require referring to experimental data through a proper parametrization to recover the errors due to the adopted simplifications. Many computational techniques can be placed in the between of the above-mention general definitions. The accuracy trend is *ab initio* > *semi-empirical* > *molecular mechanics*, which follows the gradual methodology separation from the true quantum mechanical nature of molecules. Indeed, in *ab initio* simulations electrons are explicitly treated by solving the Schrodinger equation. Vice versa, in *molecular mechanics*, electrons disappear and atoms and bonds simply become "balls and springs" with electrons characterizing the natural structural parameters (bond length, angles, torsions, etc.) and force constants. To account for their quantum nature, the classical potentials (springs) describing the interactions between atoms (balls) are derived from *ab initio* calculations. The collection of all these potentials is called force-field (FF).

As previously mentioned, the higher the accuracy, the lower the model size, see Figure 1. The largest classical MD simulation ever done envisages about one billion atoms and the simulation was run for 2 ns (nucleation of iron).[6] Another notable work regards the study of the HIV virus where the simulation was run on a model of about 64 million atoms for over 1 μs.[7] Conversely, one of the largest static *ab initio* simulation (using the hybrid B3LYP functional) was performed on the water solvated crambin protein crystal which includes ~1000 atoms in the unit cell.[8] From this observation classical MD simulations seems to be the only choice for investigating complex macroscopic materials like bone and tendons. This might be true, but there are several tricks that can be applied to enlarge the range of applicability of more accurate theoretical frameworks, e.g. use of symmetry and the coupling with the multiscale modelling.

In the next section, I will go into a brief excursus of the most important *ab initio* approaches currently available to the scientific community. More emphasis is put on those employed during this PhD work.

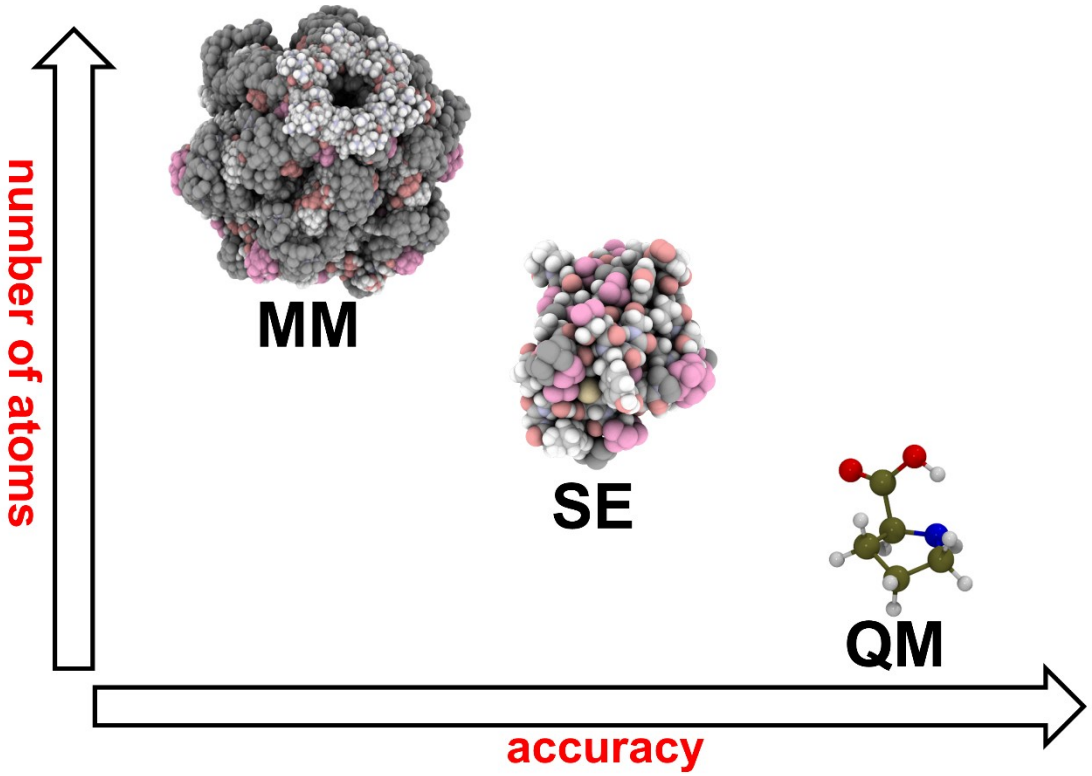


Figure 1. How do molecular mechanics (MM), semi-empirical (SE) and *ab initio* methods (QM) perform?

Hartree-Fock (HF) method and beyond

In quantum physics, all the information of a system is gathered in the **wave function (Ψ)**, which depends on all particles coordinates, spin and time. The Hamiltonian operator $\hat{H}$ when applied to the Ψ (of a certain state) gives back the energy of the system (in that state), through the **time-independent Schrödinger equation** (differential eigenvector equation)

$$\hat{H}\Psi = E\Psi \qquad (1)$$

For a non-relativistic system of nuclei and electrons, $\hat{H}$ is the sum of kinetic energy of nuclei ($\hat{T}_n$) and electrons ($\hat{T}_e$) and the potential energy of the system. We can make a good approximation of the latter term considering only Columbian force arising from the nucleus-nucleus, $\hat{U}_n(R)$, electron-electron, $\hat{U}_e(r)$, and nucleus-electron interaction, $\hat{U}_{n\text{-}e}(R, r)$,

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{U}_n(R) + \hat{U}_e(R, r) + \hat{U}_{n\text{-}e}(r) \qquad (2)$$

Separating the motion of electrons (e) and nuclei (n) (Born–Oppenheimer approximation), (1) becomes:

$$\hat{H} = \hat{H}_e + \hat{H}_n \qquad (3)$$

$$\Psi = \Psi_e \cdot \Psi_n \qquad (4)$$

$$E = E_e + E_n \qquad (5)$$

Let us set aside the nuclear part of the Schrödinger equation and focus on the Ψ_e where now the nuclei coordinates are only parameters. If we assume **electrons as non-interacting** (and we neglect the spin that can be re-introduced later), we can describe each of them with a single function in the spatial coordinates, called orbital (ϕ). In a complete base of ortho-normal orbital functions $\{\phi_i\}$, Ψ_e takes the form of a sum (this is true if the $\{\phi_i\}$ basis set is complete):

$$\Psi_e = \sum_i c_i \phi_i \qquad (6)$$

Where c_i are coefficients. The same relation works for any product of orbitals, let us call it $\{\Phi_i\}$,

$$\Psi_e = \sum_i g_i \Phi_i \qquad (7)$$

A property of Ψ_e arising from its quantum nature is **antisymmetry**, which imposes that Ψ_e must change sign to a pair permutation of electrons. This is true only if Ψ_e is a precise type of sum of product of (re-introducing the electron spin) spin-orbitals called Slater determinant, Ξ (electronic configuration). So Ψ_e is a linear combination of electronic configuration with coefficients p:

$$\Psi_e = \sum_i p_i \Xi_i \qquad (8)$$

The equation (8) is true until $\{\phi_i\}$ is complete and the sum in (8) is not truncated.

The Hartree-Fock method (HF) is a computational technique for computing the best mono-determinant ground-state Ψ_e. Within the base of the spin-orbital functions, $\{\phi_i\}$, the HF equations state as follows:

$$\hat{F}\,\phi_i = \varepsilon_i \phi_i \tag{9}$$

$$\hat{F} = \hat{h} + \hat{J} + \hat{K} \tag{10}$$

$$E = \frac{1}{2}\,\mathrm{Tr}\,(\hat{h} + \hat{F}) \tag{11}$$

$\hat{F}$ is the Fock operator. $\hat{h}$, $\hat{J}$ and $\hat{K}$, that are the mono-electronic Hamiltonian (kinetic energy + electrons-nuclei interaction), Columbian and Exchange operators. Both the kinetic and Columbian operators have a correspondence in classical mechanics; the exchange operator appears for the antisymmetric nature of the wave function. In these equations, the electrons are treated like independent, but feeling the motion of all the other electrons. The $\hat{J}$ and $\hat{K}$ operators include and act on the spin-orbital functions, therefore the HF Ψ_e is obtained iteratively (SCF), minimizing E (c_i) in function of the coefficients c_i. This ensure to have the Ψ_e in the system ground state (variational theorem). The difference between the real energy of a given system and HF one defines the correlation energy. There are several ways to account for the correlation effects (e.g. the electrons motion is not independent). For instance, it is possible to include more than one determinant in (8), e.g. configuration interaction methods (CI). For example, including all determinants within a chosen basis set gives the exact wave function of a system for that given basis set (FCI). Even if accurate, the CI methods are computationally very demanding. Conversely, it is a common procedure to couple the Perturbation Theory to the HF method to correct the HF energy. This is the Møller-Plesset approach, which provides a correction for the HF energy including a part of the correlation energy (E_{MP2}).

Density functional theory (DFT)

In the HF method, the energy of the system is a functional of the electronic orbitals and so of the wave function. A completely different point of view emerges from the first theorem of Hohenberg and Kohn. It states that **the energy of the ground state must be a functional E of the corresponding electron density ρ(x,y,z)** where ρ depends on three spatial variables only, irrespectively on the system size, at variance with Ψ_e which is a function of all electron cartesian and spin coordinates. The problem is that we do not have the exact expression for the functional E capable to extract the energy from ρ. We can separate the E[ρ] in two terms $E_{known}[\rho] + E_{xc}[\rho]$, where the first term includes the electron kinetic energy, the nuclei-electron potential and the classical electron-electron repulsion energy. The second term includes all quantum electron-electron interaction (exchange and correlation energy) not considered in the first term.

The second theorem of Hohenberg and Kohn states that **the density that minimizes the energy of the overall functional is the true electron density corresponding to the full solution of the Schrödinger equation**. With this, we can tackle the problem variationally. But how to calculate ρ? The Kohn Sham approach (KS-DFT) suggest using an auxiliary system of independent electrons that share the same ρ with the real system. In this case, the variational problem in ρ is identical to a case of independent electrons with $v_{eff}(r)$ as external potential (Coulomb nuclei-electrons and electron-electron plus the exchange correlation potential).

$$[-\frac{1}{2}\nabla + v_{eff}]\,\phi_i = \varepsilon_i\phi_i \qquad (12)$$

The problem now is finding an expression for the exchange correlation potential, which is strictly related to the $E_{xc}[\rho]$. The simpler expression of the exchange correlation potential is the sum of the exchange of a uniform electron gas plus the correlation coming from Monte Carlo simulation on a uniform electron gas. At this level, the potential depends linearly from the density (LDA), and the accuracy of DFT is not enough for chemical applications. Introducing the density gradient in the exchange functional (GGA) increases the accuracy of the method and it can be applied for problems of chemical nature (LYP[9] and PBE[10] functionals). Another clear advance in the theory is the inclusion of some HF exchange into the functional. This new family of exchange-correlation potential (hybrid functionals) had large success for their improved accuracy (B3LYP).[11]

The low cost and wide applicability of DFT is based on the fact that it does not treat the two electrons interaction explicitly like HF based methods.[4] Unfortunately this leads to the well-known large errors of DFT for one electron systems. The unphysical interaction of a single electron with it-self (self-interaction error, SIE) makes DFT very inaccurate for very simple systems (e.g. H atom). Anyways, the large accuracy/cost ratio of DFT made it the most used *ab initio* procedure of the last decades.

London dispersion correction

The London dispersion is a long-range attractive interaction between un-bonded atoms and belongs to the wider van der Waals group of interactions. It arises from instantaneous dipole–induced dipole interaction and its deep quantum mechanical nature is due to the instantaneous electron correlation. In the calculation of ground state Ψ or ρ by HF or DFT methods, some orbital functions, called the virtual orbitals, are set aside. These functions contain information on the electron excitations and response of the system to external perturbation (Koopmans' Theorem); therefore, the London interaction description is missing in both methods. These interactions are ubiquitous and their description is fundamental in many phenomena, e.g. complexes formation, adsorption process, molecular crystal stability, etc. For the reasons described above, even modern DFT implementations were lacking this important contribution to the total energy, biasing the results dramatically. At variance with DFT, since their very beginning this type of interactions were included in the classical force fields, where the usual interaction potential is a 12-6 Lennard-Jones. Historically, London and Eisenschits were the first to derive the equation to describe the dispersion energy at long range:[12]

$$E_{disp}^{AB} \approx \frac{3}{2}\frac{I_A \cdot I_B}{I_A + I_B} \cdot \alpha_A \alpha_B R^{-6} = -C_{6\,eff}^{AB} \cdot R^{-6} \qquad (13)$$

In their formulation the energy between atom A and B, E_{disp}^{AB}, depends from the atomic ionization potential I_A and I_B and the atomic static polarizability α_A and α_B of both species. It is interesting to notice that $\mathbf{E_{disp}^{AB}}$ **is always attractive and tends to stabilize dense over dilute structures**. A more general formulation for the dispersion interactions derives from the second order perturbation theory, PT2 (at the limit of long distances), with a general expression similar to the London one:

$$E_{disp}^{(6)AB} = -C_6^{AB} \cdot R^{-6} \qquad (14)$$

$$C_6^{AB} = 3/2\pi \int_0^{\infty} \alpha_A(i\omega)\alpha_B(i\omega)d\omega \qquad (15)$$

Here the $\alpha(i\omega)$ is the isotropic dynamical polarizability of the fragments A and B. It useful to recall that imposing the unperturbed Hamiltonian as the HF Hamiltonian in PT2 we directly obtain the MP2 energy, which therefore account for London dispersion contribution. If we consider higher orders of the perturbation theory, we obtain similar expression for the $E_{disp}^{(8)AB}$ and $E_{disp}^{(10)AB}$ terms. The C_6^{AB}, C_8^{AB} and C_{10}^{AB} terms describe the dipole-dipole, the dipole-quadrupole interaction and the dipole-octupole and the quadrupole-quadrupole interactions between the A and B fragments, respectively. The first term dominates largely on the others due to its slower decay with the system separation. Nonetheless, this pair-wise approach cannot take into account the many-body nature of the problem. Indeed, the dispersion energy of a three-body system ABC is not simply the sum of E_{AB}, E_{BC} and E_{AC} pairwise terms. Nevertheless, the pairwise approach is employed in which the usual fragments are atoms. The oversimplifications arises from the fact that:

- the coefficient of the free atoms are different from those of the atoms in the molecule.
- interference effects (if the atom A is in the middle between B and C, it will screen the interaction between B and C) are not taken into account.

The simplest workaround the HF and DFT deficiencies in that respect is to add a pairwise approach to atomic fragments to supplement the pure electronic energy with a London term:

$$E_{disp}^{(n)} = -\sum_{AB}\sum_{n=6,8,10}\frac{C_n^{AB}}{R^n} \qquad (16)$$

The result is called "dispersion corrected approach", the most popular one being the DFT-D methods proposed by Grimme and coworkers[13,14]. In particular, the D2 correction and its recent improved D3 scheme, received a large consensus from the scientific community of molecular modeller for their simple implementation and cheap computational cost. In all cases, the quantum and dispersion energies are un-coupled:

$$E_{TOT} = E_{QM} + E_{DISP} \qquad (17)$$

For the D2 dispersion scheme,[13] the expression of the dispersion energy is equivalent to (16) with n=6 multiplied by a proper damping function which usually converges to 0 at small distances to avoid the singularity of the dispersion energy at small R. Furthermore, this function avoids double-counting the contribution coming from the pure DFT part and the empirical dispersive one at separations within the classical bond lengths. In other words, it is assumed that pure DFT handles correctly the short-range interactions (covalent bond) while the D term handles the long range, dispersive contribution. Usually, the C_6^{AB} terms are evaluated as geometric mean of the respective homo-atomic values. The main disadvantage of this technique is the lack of any dependences on the molecular environment. To improve its performances for molecular crystals in combination with the B3LYP functional, Civalleri et al.[15] rescaled the whole E_{DISP} term in eq. (17) and the van der Walls radii (D* scheme). To address this point affecting the D2 scheme, Grimme recently proposed the D3 dispersion scheme.[14] Here the chemical environment of each atoms is taken into account explicitly. For different reference geometries, the C_6^{AB} coefficients are computed for each atoms by the *ab initio* calculations following eq. (15). Furthermore, the D3 scheme includes also the $E_{disp}^{(8)AB}$ term.[16,17] An auxiliary contribution to the London forces arises from the three body interaction, the so-called Axilrod–Teller–Muto (ATM)-three-body-term,[18,19] accounting for the dipole-dipole-dipole interaction. This $E_{disp}^{(9)}$ term depends directly on the $\sqrt{C_6^{AB}C_6^{BC}C_6^{AC}}$ and inversely on the $(R_{AB}R_{BC}R_{AC})^3$ and, usually, contributes less than 5% to the total E_{DISP}. This correction scheme has been implemented in combination with the D3 scheme ($D3^{ABC}$).

Basis set

The previously presented formalism works for molecules but also for crystals (periodic case), if considering that electrons in periodically repeating environment are described by Bloch waves:

$$\Psi_e(\mathbf{r},\mathbf{k}) = e^{i\mathbf{kr}}u(\mathbf{r}) \qquad (18)$$

Where **k** is the crystal wave vector, $e^{i\mathbf{kr}}$ is a plane-wave function, **r** is the position vector and u(**r**) is a periodic function specific for the treated crystal. The two types of basis sets concur to define the Bloch waves $\Psi_e(\mathbf{r},\mathbf{k})$. These can be represented through atom centered *Gaussian* functions or *plane-wave* functions. One of the disadvantage of employing a Gaussians based basis set is the basis set superposition error (BSSE). It is a well-known effect which takes place mainly in the simulation of weakly bonded complexes and adsorption processes. The BSSE arises from the basis orbitals sharing between the complexes components which induces adduct energy lowering. A plane-wave basis set fills the space homogenously so it is inherently BSSE free. The BSSE can be avoided by using very large set of Gaussians functions, which increases notably the computational burden of the simulations. The standard method of correcting for BSSE is the counterpoise (CP) method of Boys and Bernardi (BB-CP).[20] In the CP approach the energy of the complexes parts are calculated by using the whole complex basis set. This is more expensive but include the effect of the other basis set functions present in the adduct calculation. By mapping the BB-CP onto a semi-empirical atom-pairwise potential, it is possible to obtain a repulsive potential which depends only on the system geometry (gCP),[21] with energy expression as follows

$$E_{BSSE}^{gCP} = \sum_{\substack{A,B \\ (A\neq B)}}^{atoms} e_A^{miss} \frac{exp\left(-\alpha(R_{AB})^{\beta}\right)}{\sqrt{S_{AB}N_B^{virt}}} \qquad (19)$$

e_A^{miss} measures the incompleteness of the atomic target basis set, R_{AB} is the interatomic distance, S_{AB} and N_B^{virt} are the Slater-type overlap and the number of virtual orbitals, respectively. α and β are simply empirical parameters. This type of "classical" treatment for the BSSE and dispersion forces (see previous paragraph) can be coupled effectively within the *ab initio* framework. For instance, Grimme recently proposed the HF-3c method, which is dispersion corrected and BSSE free thanks to the D3 and the gCP correction schemes (*vide infra*).

The computer codes

Most of the theoretical frameworks presented in this chapter are implemented in the CRYSTAL, CP2K and CRYSCOR codes.[22–24] The **CRYSTAL** program uses atom centered Gaussian functions to describe the crystal or molecular orbitals and computes **energy and gradient at the HF and DFT level** (inclusive of *Dispersion* correction). The adoption of localized basis functions allows an efficient implementation of the exact exchange even for crystal materials, at variance with the much more problematic implementation using plane-waves basis functions. The **CP2K** suit uses a mixture of plane-wave and Gaussian basis set to generate the Bloch waves, and is capable of running the **dynamical evolution of the systems**. In this type of simulations called *ab initio* molecular dynamic (**AIMD**), the system time evolution is described classically by the Newton laws of motion while the forces are computed on the fly through the solution of the Schrödinger equation. Finally, the **CRYSCOR** code computes **the energy correction at the second order of the Møller-Plesset perturbation theory (MP2)[25] to the HF energy for periodic systems**.

The semi-empirical HF-3c method

In this section, we briefly described the "semi-empirical" HF-3c method, which has been extensively used in this PhD thesis.[26] The HF-3c method consists in a HF calculation with a small previously published basis set (MINIX) and three semi-empirical corrections (3c) to the HF energy to recover the intrinsic errors due to the adoption of the HF Hamiltonian coupled with a minimal basis set. The quality of the basis set representing the atomic electron shells varies with the atom type. For the most common biological atomic species, e.g. H C O and N, a STO-3G like quality basis set is employed (MINIX set, details in Ref. [26]).[27] The HF method with a basis set quality lower than 6-31G* shows strong deficiencies,[4] hence a key role is played by the empirical corrections. Essentially, the 3c terms are added in order to:

(i) include long-range London dispersion interactions by the D3 scheme
(ii) correct for the BSSE through the gCP method
(iii) correct for the covalent bond length that are systematically overestimated for electronegative elements in the HF/MINIX method (SRB)

We have described (i) and (ii) in details in the previous paragraphs. (iii) is defined as follows:

$$E_{SRB} = -s \sum_{A,B(A\neq B)} (Z_A Z_B)^{3/2}\, e^{-\gamma (R_{AB}^0)^{\frac{3}{4}} R_{AB}}$$

Where Z is atomic charge and $R_{AB}^{0,D3}$ is the cut off radii. s and γ are fitting parameter determined to produce vanishing HF-3c total atomic forces on around 100 small organic molecules optimized at the B3LYP-D3/def2-TZVPP level.

The corrected total energy, HF-3c, is then calculated as

$$E_{tot}^{HF-3c} = E_{tot}^{HF/MINIX} + E_{disp}^{D3(BJ)} + E_{BSSE}^{gCP} + E_{SRB} \qquad (20)$$

The method presents several advantages:

- it is very fast due to minimal basis
- it provides on average correct bond lengths
- it is purely analytical (grid free) thus leading to noise free derivatives
- it is corrected for weak dispersion forces and it is BSSE free.

However, HF-3c cannot simulate negative charged as well as strongly correlated systems properly, due to the electron correlation omission and the limited size of the basis set.

Computational details

In this paragraph, I brief describe the methodology details employed in the simulations reported in this PhD thesis. A more detailed description can be found in the published papers.[28–30]
For the **DFT** calculations run with CRYSTAL code, we used both the B3LYP hybrid functional,[31] and the PBE one.[32,33] To account for **dispersion forces** we employed the D2 dispersion scheme,[25] and the D3 scheme.[34] In some cases, the D3 scheme is also coupled with the Axilrod–Teller–Muto three-body-term.[35]

The convergence in energy for the **SCF** cycle is set to 10^{-7} Ha. Often, the recently introduced DIIS extrapolator technique, which reduces the number of SCF cycles by almost a factor two, has been employed to speed up the SCF convergence.[36] Tolerance values controlling the Coulomb and exchange series (the latter for B3LYP and HF calculations) in periodic systems[37,38] vary in function of the system investigated. In most of the cases reported in this thesis the tolerances are set to the default values.[25] The k-point grid is generated via the Monkhorst−Pack scheme[39–42] with a system dependent number of k points. The choice of k point is made to ensure good convergence of the energy and geometry on the number of k points.

Most of the times valence triple-zeta polarized (VTZP) **basis sets** were adopted for carbon, nitrogen and oxygen atoms,[43] while hydrogen, calcium and phosphorus atoms are described with a 3-1G(p), 865-1G(p,d) and 85-2G(p,d) basis set, respectively.[44]

For the **geometry optimization**, the internal coordinates were relaxed using the analytical gradient method. The Hessian is upgraded with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.[45–47] I kept the tolerances for the maximum allowed gradient and the maximum atomic displacement for convergence at the default values (0.00045 Ha·bohr^{-1} and 0.0018 Bohr, respectively).

For the elastic property calculations, the default values are tighten to 0.00045 Ha·bohr^{-1} and 0.00093 Bohr, respectively. Moreover, in the latter case the SCF convergence on energy is set to 10^{-8} Ha.

For the **vibrational frequency calculations** run with the CP2k code, the mass-weighted force-constant matrix was computed at the Γ point by numerical derivative of the analytic nuclear gradients. A value of 0.003 Å was chosen as the displacement of each atomic coordinate and the tolerance for the SCF cycle convergence was tightened from 10^{-7} to 10^{-11} Ha.

For the **dynamic simulations** with CP2K the Quickstep technique with a geometry mixed plane wave and Gaussian basis set methodology (Gaussian and Plane Wave method, GPW) was employed to calculate the electronic structure.[33] We used the PBE functional, with the Goedecker-Teter-Hutter pseudopotentials,[33] and a Gaussian triple-ζ basis set with polarization functions (TZVP) for all atoms, but P and Ca for which a double-ζ basis set with polarization functions (DZVP) is employed. The cut-off for the plane wave basis was set to 400 Ry. Dispersion forces are taken into account with the Grimme D2 method. A time step of 0.5 fs was chosen and the Nosé–Hoover thermostat was employed.[33]

The **graphical visualization** and **structural manipulation** were performed with MOLDRAW version 2.0.[37,38] Dynamical data analysis was performed employing the VMD program.[48] Images were rendered with VDM, VESTA and POV-Ray softwares.[49,50]

Cohesive energy and sublimation enthalpy calculation for molecular crystals

One of the type of systems investigated in this thesis are molecular crystals, which have been used to test and refine the simplified HF-3c method. This was done considering that the kind of interactions in molecular crystals are very close to those present in the COL models. Comparing the computed cohesive energy and sublimation enthalpy with the experiment can give insights on the quality of the adopted method.

Here, we briefly detail how we compute those quantities and how we compare them with the experiment. The cohesive energy is defined as:

$$\Delta E_c = E_{cry}/Z - E_{mol} \quad (21)$$

where E_{cry} is the total energy of the crystal unit cell with optimized cell parameters and internal coordinates, Z the number of molecules in the unit cell, and E_{mol} is the total energy of the fully optimized isolated molecule in the gas phase. To have correct results the ΔE_c must be correct for the BSSE.

The computed cohesive energies can be compared with thermodynamically back-corrected experimental sublimation enthalpies. The back-correction procedure can take into account the thermal and zero point vibrational (ZPE) energy explicitly (see Ref. [51,52]). Conversely, they can be estimated by subtracting a constant factor 2RT (3/2RT for linear molecules) to the experimental sublimation enthalpy

$$\Delta E_{exp} = \Delta H_{exp} - 2RT \quad (22)$$

The sublimation enthalpy can be directly computed by the following expression:

$$\Delta H_{sub} = -\Delta E_c + \Delta E_{vib} + 4RT \quad (23)$$

where the constant term is 7/2 RT for linear molecules and ΔE_{vib} includes both the ZPE and the thermal vibrational energy contributions. This expression is true only when the isolated molecule has a well-defined conformation at a given temperature, behaves as an ideal gas and no phase transformation occurs between 0 K and T. We calculated the vibrational frequencies in the harmonic approximation. To compute the thermal vibrational contribution to enthalpy the Einstein model has been applied. In this model, each phonon branch is approximated with a single frequency obtained at the Γ point, the acoustic ZPE is neglected and the acoustic thermal energy is evaluated at the high temperature limit as 3RT.[53,54]

Binding Energy (BE) of the collagen model on the Hap surface

In a periodic treatment, the binding energy (BE) *per* unit cell *per* adsorbate is defined as:

$$BE = E(S//S) + E_M(P//P) - E(SP//SP) \quad (24)$$

where the symbol following the double slash identifies the optimized geometry at which the energy has been computed. Therefore, E(S//S), E_M(P//P) and E(SP//SP) are the energies of a bare slab S, the free adsorbate P and the interacting system SP in their fully optimized geometry. BE is a positive quantity for a bounded system and can be recast in terms of the geometrical deformation cost of the surface ($\delta E_S>0$), the adsorbate ($\delta E_P>0$), the lateral adsorbate−adsorbate interactions (ΔE_L) between polymer images, and the binding energy between the pre-deformed constituents (BE*):

$$\delta E_S = E(S//SP) - E(S//S) \quad (25)$$

$$\delta E_P = E_M(P//SP) – E_M(P//P) \quad (26)$$

$$\Delta E_L = E(P//SP) – E_M(P//SP) \quad (27)$$

$$BE^* = E(S//SP) + E(P//SP) - E(SP//SP) \quad (28)$$

$$BE = BE^*- \delta E_S - \delta E_P - \Delta E_L \quad (29)$$

ΔE_L is the difference between the energy of P organized as a periodic 2D system in the SP geometry and the energy of a single deformed P. This term can then be either positive/negative as a function of the favorable/un-favorable lateral interaction. Obviously, BE* > BE, as the deformation costs are usually larger than the lateral interactions ΔE_L, even when the latter are repulsive. As we are dealing with localized Gaussian-type basis set, we should correct the BE for the BSSE using a well consolidated procedure already adopted in previous works.[35,55] Only ΔE_L, BE* and BE terms are affected by BSSE (all deformation terms are BSSE free) and their corrected values are indicated with the notation ΔE_L^C, BE^{*C} and BE^C, respectively. To avoid a relevant role of the lateral interactions between adsorbate replicas, which would have obscured the importance of the surface, we adopted unit cell sizes large enough to ensure to have ΔE_L negligible.

2.2 Current Scenario of Simulations on Tendon and Bone Materials and Components

Reviewing collagen and bone bibliography is a challenging task. Ten years ago, Kadler et al. state referring to collagen protein "newcomers to the field face the daunting task of sifting through 100,000 research papers that span 40 years".[56] Furthermore, the rate of published papers even increased in the last decade.[57] A glance to the SCIFINDER website[58] reinforces the idea on the large amount of information on the topic to date: 400,000 and 1,3000,000 papers for the "collagen" and "bone" entries. Even if only a small fraction of these works employed simulations as method of investigation the amount of information is still large. To review them in a suitable way for a PhD thesis, I had only to full report the works based on quantum mechanical (QM) simulations, referring to review papers for most of all others approaches. Only in few cases, I made exceptions to the above mention standpoint to prove the potentialities of QM simulations on the field.

As for the *ab initio* works on collagen models (comprising of triple helical structure), they can be counted on the fingers of one hand. All simulations refer to finite molecular models, in which the 300 nm long collagen protein is modelled with much shorter molecules, usually around one thousand times shorter in length. Clearly, a polymeric model (1D periodic) for the COL protein would be much more representative of the real protein than any finite and short model (molecule), Figure 2. Moreover, a periodic model solve the need of the capping substituents at the peptide termini. These are included to avoid the inter-chains interaction of the free COOH and NH_2 groups at the molecule edges (Figure 2).

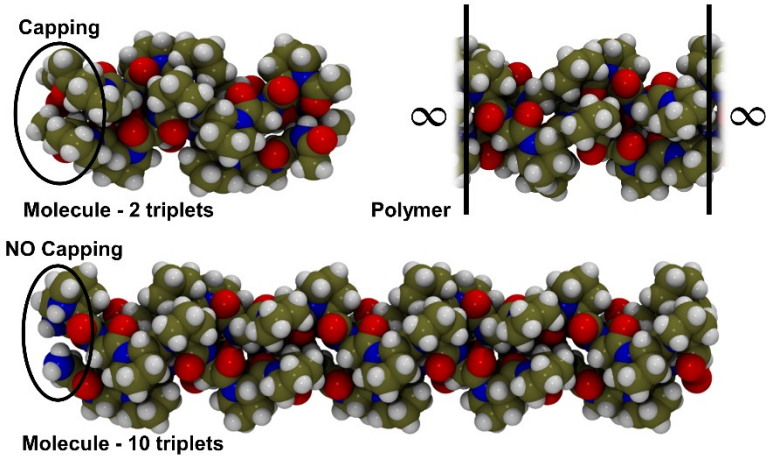


Figure 2. Comparision of different collagen (COL) models. The number of triplets refers to each COL single strand. The 10 triplets model is reported without capping (NO capping) to highlight the difference between capped and un-capped models only. All the models reported in literature include the ends capping. Color code: Nitrogen/blue, Oxygen/red, Carbon/cyan and Hydrogen/light gray.

QM simulations have been often used to study the inter-strand interaction in COL. In Ref. [59], they computed the folding energy of a handful of COL-like peptides (2 triplets – Figure 2) at the DFT/ONIOM level of theory. Moreover, they investigated the effect of Gly substitution, finding that that using D-aminoacids in place of Glycine aminoacids generates stable triple helices. **It is worthwhile mentioning the difficulties in the folding energy calculation: the chosen single strand starting geometry is completely arbitrary, thus, the final reaction energy value is questionable**. A reasonable starting conformation for a single COL strand is the PPII, which has been related to the geometry of un-folded proteins.[60] We will explore the PPII conformation as possible starting geometry for a COL single strand later in Chapter 3.

The problem is faced in a different way in Ref. [61] where they compare the collagen conformation (COL models, 2 triplets – Figure 2) with the β-sheet conformation. They investigated several primary structures, finding that the imino-acid containing peptides prefer triple helical structure due to the missing H-bond donor (N-H) in the β-sheet conformation.

Semi-empirical methods coupled with DFT simulations were used on simple triple helical models (similar to those mentioned in Ref. [59,61]) to compute the circular dichroism spectra.[62] Two works based on DFT-LDA were performed on a large triple helical COL models (10 triplets – Figure 2). The computed properties were (i) the density of state and atomic charges of the protein.[63] (ii) the binding energy between protein chains and with respect to the solvent (water).

Regarding the COL interaction with hydroxyapatite (Hap), de Leeuw et al. computed the gas-phase interaction energy of aminoacids triplets made of Gly, Pro, Hyp and Lys (single strand) with the Hap (001) and (010) surfaces represented as slab models.[64]

The most common computational framework used to study bone and tendon materials is molecular mechanics (MM). Due to the relatively low cost of classical FFs which usually scales almost linearly with the system size, realistic models of COL become feasible allowing several interesting questions to be addressed. The first MM simulations on models of COL started in the '70s (for a review see Ref. [65]). At that stage, it was only feasible to relax the atomic position on short COL models. The COL model length was only few nm long and imposes severe restriction on the primary structure of the protein, which is drastically simplified in these works. These studies shed some light on the role of the inter-chains interaction on the stability of the triple helix,[66] and the lateral inter-proteins interaction.

In the last two decades, thanks to the increase of the available computational resources and the MD codes improvements, the use of FF was extended to MD simulations, allowing investigating several different features of bone and tendons components, e.g. mechanical properties. Below, some peculiar examples of the use of classical molecular dynamic (CMD) are listed:

a) **structure**/**stability** correlation in collagens (e.g. the role of cross-chain ionic interactions, aminoacidic pre-organization and protein model chain length).[67–70]
b) **mechanic** of COL protein itself,[71] or on high stress regimes.[72]
c) cross linking a single **water** bridge between COL chains[73] also in the case of Gly mutation.[74]
d) **COL/Hap interface**. One of the first work on this topic employed a simple molecular tripeptides as collagen models, see Figure 3 A. It is interesting to notice that the *classical adsorption energies are comparable (~15% dev) with the those obtained at the GGA DFT level*.[75] In this case, a polarizable FF was used for describing the Hap mineral. This type of FF is to be considered more accurate than non-polarizable FFs because it models the changes of atomic charge distribution as a function of the environment. In a more recent work they computed the experimental binding energy of the COL protein on the (100) and (110) Hap surfaces. Adsorption energies were performed with classical FF envisaging a short triple helical collagen model, envisaging 10 triplets, Figure 2. Both method agrees on the preferential adsorption of COL on Hap (100) face. Unfortunately, comparing the experimental adsorption energies with the computed ones was not carried out as the computed ones were not reported in the paper. Therefore only relative energies between different conformations was addressed.[76]
e) **mechanics** of the **COL-Hap** adduct employing simple bone models, see Figure 3 B (for a review see Ref. [77]).[78–80]
f) early stage **nucleation of Hap** on the COL protein.[81,82]

It was only the x-ray structure determination of tendon fibril by Orgel and coworkers,[83] that the quality of the modelling increased.

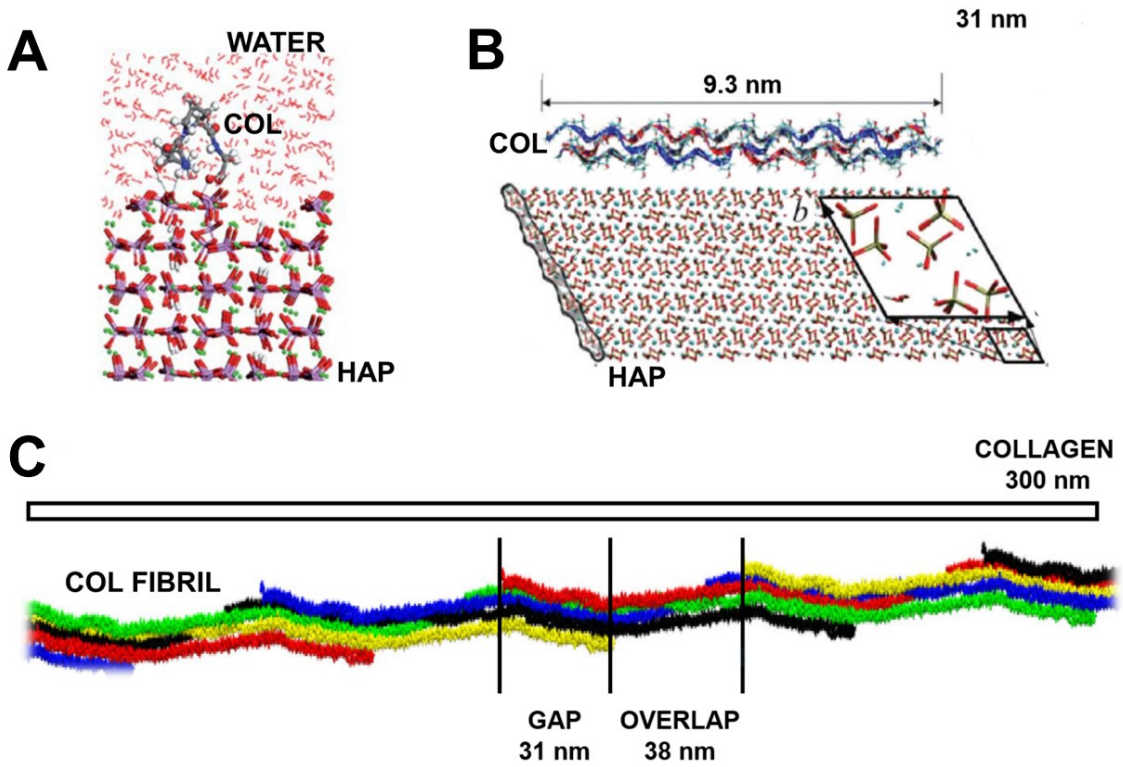


Figure 3. Different modelling for COL and COL/Hap interface. Taken from Ref. [75,79,84].

It became possible to full atomistic model the fibril environment of bone and tendons, see Figure 3 C. This more accurate modelling, which includes all the COL fibril features, e.g. the gap-overlap regions, 3D periodicity, etc., (see Chapter 1 for more details), allowed a better understanding on:

g) the nature of the collagen fibril interactions (protein-protein and protein-water).[85,86]
h) the effect of water and Hap inclusion in the COL matrix on the mechanical properties of tendon and bone.[84,87–89]
i) the effect of mutations, such as those happening in human diseases, e.g. osteogenesis imperfecta, on tendons mechanics.[90]
j) the mechanical damaged of tendons due to stress.[91]
k) the cross-linking glycation phenomena, occurring with age and diabete.[92,93]
l) the mechanical property of collagen protein hetero-structure. The gap and overlap regions of the tendon fiber have different mechanical properties.[94,95]
m) the clustering of Hap on a collagen fibril under stress.[96]

Unfortunately, these works based on CMD often couple an accurate description of the atomistic complexity of the fibril and mineralized fibril, with an inaccurate description of the interatomic forces. Sometimes, the FF misses the explicit treatment of H-bonds, fundamental for describing the protein-water interaction, (g). Moreover, when the COL protein is studied in interaction with Hap, the parameters in charge of treating the interaction between protein atoms with the Hap ions were missing. Therefore, combination rules were adopted with loss of accuracy in region which is crucial for the features of the adsorbed COL.[97] This problem becomes smaller when a polarized FF is employed for the Hap mineral, as confirmed by the agreement between the FF and DFT interaction energies, (c). Unfortunately, due to the higher cost of polarized FF over the non-polarized ones, only small model systems can be investigated. **Therefore, the quality of the forces acting between COL and Hap are questionable and, consequently, the *ab initio* treatment may be a fruitful approach to provide unbiased results and provide clues to improve the FF parametrization.** The *ab initio* treatment is also the only solution when bond breaking/making is occurring at the COL/Hap interface,

due to the strong interaction between ions and the COL aminoacids. The only MM alternative relies upon the adoption of reactive force fields like ReaxFF,[98] which is capable to handle chemical reactions at the price of specific and heavy parametrization.

A different and valuable approach is the multiscale technique. The basic idea of these works is to use full atomistic CMD simulations on small COL models to derive parameters for a mesoscopic bead-spring model, therefore dramatically reducing the system size and the simulation time. Several interesting studies adopting this approach appeared on tendons simulation already ten year ago.[99–102]The mesoscopic model is made to simulate the energetic and force response of the real protein and assemblies (fibril), as they were obtained from fully atomistic simulations. In this way they have simulated COL protein elasticity both as a single molecule[99] and as a fiber.[102] Moreover, they elucidated the role of the Hap,[100] and the mutations (osteogenesis imperfecta)[101] onto the fibril mechanics. A limitation of this approach arises from the oversimplification of the collagen models adopted as reference. In several cases,[99–102] the starting COL model is made by 10 triplets, see Figure 2, which neglects the highly heterogeneous primary structure of COL. Indeed, no double-check studies were carried out on the effect of the primary structure on the reference COL model at the mesoscopic level of treatment. **Nevertheless, the agreement of this approach with the experiment is generally good, pushing the interest on this multiscale approaches for simulating bone and tendon materials at scales approaching the real system**.[103,104] Therefore, there is no surprise that the adoption of the coarse-grained modelling is indeed rising.[77,105–107]

Perspectives

Despite the large number of *ab initio* works on Hap, there are only few studies devoted to model the interaction of COL with Hap at the *ab initio* level due to the large complexity of the hybrid system. Among these studies, no one was focusing on reliable tendon models, e.g. triple helical collagens, nor bone models, e.g. triple helix/ Hap interfaces. Nevertheless, a very recent multiscale modelling review[108] indicates the potentiality of QM simulations in bone research as a method parameters-free. "The *ab initio* simulation on tendon and bone material" seems a promising un-explored research direction from two point of views:

- The parameter free approach provided by QM can be a solid base for a multi-technique approach.
- Moreover, several atomic-scale questions of interest, can be addressed by QM simulations thanks to their widely accepted high accuracy.

Along this research path, the main difficulty is producing reliable models for the computational expensive QM simulations. As we will see later in the thesis, by exploiting both the symmetry as encoded in computer codes like CRYSTAL17 and the high performance resource together with simplified, but still reliable, quantum mechanical methods we will be able to simulate periodic models of COL.

Bibliography

1. Grimme, S., Hansen, A., Brandenburg, J. G. & Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chem. Rev.* **116,** 5105−5154 (2016).
2. Tarantelli, F. *'Appunti di Chimica Teorica'*. (Università di Perugia, 2013).
3. Kohn, W. Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Rev. Mod. Phys.* **71,** 1253–1266 (1999).
4. Pople, J. A. in *In Nobel Lecture, Chemistry 1996-2000* (ed. Grenthe, I.) (World Scientific Publishing, 2003).
5. Richon, A. B. An Introduction to Molecular Modeling. *Mathematech* **1,** 83 (1994).
6. Shibuta, Y. *et al.* Heterogeneity in homogeneous nucleation from billion-atom molecular dynamics simulation of solidification of pure metal. *Nat. Commun.* **8,** 10 (2017).
7. Perilla, J. R. & Schulten, K. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nat. Commun.* **8,** 1–10 (2017).
8. Delle Piane, M., Corno, M., Orlando, R., Dovesi, R. & Ugliengo, P. Elucidating the fundamental forces in protein crystal formation: The case of crambin. *Chem. Sci.* **7,** 1496–1507 (2016).
9. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37,** 785–789 (1988).
10. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **77,** 3865–3868 (1996).
11. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98,** 5648–5652 (1993).
12. Eisenschitz, R. & London, F. Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften. *Eur. Phys. J. A* **60,** 491−527 (1930).
13. Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **27,** 1787–1799 (2006).
14. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132,** 154104–19 (2010).
15. Civalleri, B., Zicovich-Wilson, C. M., Valenzano, L. & Ugliengo, P. B3LYP augmented with an empirical dispersion term (B3LYP-D*) as applied to molecular crystals. *Cryst. Eng. Comm.* **10,** 405–410 (2008).
16. Becke, A. D. & Johnson, E. R. A density-functional model of the dispersion interaction. *J. Chem. Phys.* **123,** 154101 (2005).
17. Johnson, E. R. & Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *J. Chem. Phys.* **123,** 24101 (2005).
18. Axilrod, B. M. & Teller, E. Interaction of the van der Waals Type Between Three Atoms. *J. Chem. Phys.* **11,** 299–300 (1943).
19. Muto, Y. The Force Between Nonpolar Molecules. *Proc. Phys. Math. Soc. Japan* **17,** 629–31 (1944).
20. Boys, S. F. & Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **19,** 553–566 (1970).
21. Kruse, H. & Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems. *J. Chem. Phys.* **136,** 154101 (2012).
22. Dovesi, R. *et al.* CRYSTAL14 : A program for the ab initio investigation of crystalline solids. *Int. J. Quantum Chem.* **114,** 1287–1317 (2014).
23. Hutter, J., Iannuzzi, M., Schiffmann, F. & Vandevondele, J. CP2K: Atomistic simulations of

condensed matter systems. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4,** 15–25 (2014).
24. Pisani, C. *et al.* Cryscor: a program for the post-Hartree–Fock treatment of periodic systems. *Phys. Chem. Chem. Phys.* **14,** 7615 (2012).
25. Dovesi, R. *et al. CRYSTAL14, User's Manual*. (Università di Torino, 2014).
26. Sure, R. & Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **34,** 1672–1685 (2013).
27. Tatewaki, H. & Huzinaga, S. A systematic preparation of new contracted Gaussian-type orbital sets. III. Second-row atoms from Li through Ne. *J. Comput. Chem.* **1,** 205–228 (1980).
28. Cutini, M., Corno, M. & Ugliengo, P. Method dependence of proline ring flexibility in the poly-L-proline type II polymer. *J. Chem. Theory Comput.* **13,** 370−379 (2017).
29. Cutini, M. *et al.* Assessment of Different Quantum Mechanical Methods for the Prediction of Structure and Cohesive Energy of Molecular Crystals. *J. Chem. Theory Comput.* **12,** 3340–3352 (2016).
30. Cutini, M., Corno, M., Costa, D. & Ugliengo, P. How Does Collagen Adsorb on Hydroxyapatite? Insights From Ab Initio Simulations on a Poly-Proline Type II Model. *J. Phys. Chem. C* (2017) DOI: 10.1021/acs.jpcc.7b10013.
31. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).
32. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44,** 1272–1276 (2011).
33. Czinki, E. & Császár, A. G. Conformers of gaseous proline. *Chem. - A Eur. J.* **9,** 1008–1019 (2003).
34. Corno, M., Rimola, A., Bolis, V. & Ugliengo, P. Hydroxyapatite as a key biomaterial: quantum-mechanical simulation of its surfaces in interaction with biomolecules. *Phys. Chem. Chem. Phys.* **12,** 6309–6329 (2010).
35. Jimenez-Izal, E., Chiatti, F., Corno, M., Rimola, A. & Ugliengo, P. Glycine Adsorption at Nonstoichiometric (010) Hydroxyapatite Surfaces: A B3LYP Study. *J. Phys. Chem. C* **116,** 14561–14567 (2012).
36. Pulay, P. Improved SCF convergence acceleration. *J. Comput. Chem.* **3,** 556–560 (1982).
37. Detar, D. F. & Luthra, N. P. Conformations of Proline. *J. Am. Chem. Soc.* **76,** 1232–1244 (1977).
38. Ho, B. K., Coutsias, E. A., Seok, C. & Dill, K. A. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* **14,** 1011–8 (2005).
39. Grimme, S., Brandenburg, J. G., Bannwarth, C. & Hansen, A. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *J. Chem. Phys.* **143,** 54107 (2015).
40. Brandenburg, J. G., Maas, T. & Grimme, S. Benchmarking DFT and semiempirical methods on structures and lattice energies for ten ice polymorphs. *J. Chem. Phys.* **142,** 124104 (2015).
41. Erba, A., Maul, J. & Civalleri, B. Thermal properties of molecular crystals through dispersion-corrected quasi-harmonic ab initio calculations: the case of urea. *Chem. Commun.* **52,** 1820–1823 (2016).
42. Heit, Y. N., Nanda, K. D. & Beran, G. J. O. Predicting finite-temperature properties of crystalline carbon dioxide from first principles with quantitative accuracy. *Chem. Sci.* **7,** 246–255 (2016).
43. Kilpatrick, J. E., Pitzer, K. S. & Spitzer, R. The Thermodynamics and Molecular Structure of Cyclopentane. *J. Am. Chem. Soc.* **69,** 2483–2488 (1947).
44. Altona, C. & Sundaralingam, M. Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Am. Chem. Soc.* **94,** 8205–8212 (1972).
45. Broyden, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.* **6,** 76–90 (1970).

46. Fletcher, R. A. New Approach to Variable Metric Algorithms. *Comput. J.* **13,** 317–322 (1970).
47. Shanno, D. F. & Kettler, P. C. Optimal Conditioning of Quasi-Newton Methods. *Math. Comput.* **24,** 657–664 (1970).
48. London, R. E. On the interpretation of carbon-13 spin-lattice relaxation resulting from ring puckering in proline. *J. Am. Chem. Soc.* **100,** 2678–2685 (1978).
49. Zhong, H. & Carlson, H. A. Conformational Studies of Polyprolines. *J. Chem. Theory Comput.* **2,** 342–353 (2006).
50. Ramshaw, J. A. M., Shah, N. K. & Brodsky, B. Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides. *J. Struct. Biol.* **122,** 86–91 (1998).
51. Brandenburg, J. G. *et al.* Geometrical correction for the inter- and intramolecular basis set superposition error in periodic density functional theory calculations. *J. Phys. Chem. A* **117,** 9282–9292 (2013).
52. Reilly, A. M. & Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **139,** 24705 (2013).
53. Riley, K. E., Platts, J. a., Řezáč, J., Hobza, P. & Hill, J. G. Assessment of the performance of MP2 and MP2 variants for the treatment of noncovalent interactions. *J. Phys. Chem. A* **116,** 4159–4169 (2012).
54. Cybulski, S. M. & Lytle, M. L. The origin of deficiency of the supermolecule second-order Møller-Plesset approach for evaluating interaction energies. *J. Chem. Phys.* **127,** 141102 (2007).
55. Corno, M., Busco, C., Bolis, V., Tosoni, S. & Ugliengo, P. Water adsorption on the stoichiometric (001) and (010) surfaces of hydroxyapatite: A periodic B3LYP study. *Langmuir* **25,** 2188–2198 (2009).
56. Kadler, K. E., Baldock, C., Bella, J. & Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **120,** 1955–1958 (2007).
57. Pacchioni, G. *Scienza, quo vadis? Tra passione intellettuale e mercato*. (2017).
58. SCIFINDER. Available at: https://scifinder.cas.org.
59. Tsai, M., Xu, Y. & Dannenberg, J. J. Completely geometrically optimized DFT/ONIOM triple-helical collagen-like structures containing the ProProGly, ProProAla, ProProDAla, and ProProDSer triads. *J. Am. Chem. Soc.* **127,** 14130–14131 (2005).
60. Adzhubei, A. A., Sternberg, M. J. E. & Makarov, A. A. Polyproline-II Helix in Proteins: Structure and Function. *J. Mol. Biol.* **425,** 2100–2132 (2013).
61. Palfi, V. K. & Perczel, A. How Stable is a Collagen Triple Helix? An Ab Initio Study on Various Collagen and b-Sheet Forming Sequences. *J. Comput. Chem.* **29,** 1374–1386 (2008).
62. Jalkanen, K. J., Olsen, K., Knapp-Mohammady, M. & Bohr, J. First-principles structures for the close-packed and the 7/2 motif of collagen. *EPL (Europhysics Lett.* **100,** 28003 (2012).
63. Eifler, J., Rulis, P., Tai, R. & Ching, W. Y. Computational study of a heterostructural model of type I collagen and implementation of an amino acid potential method applicable to large proteins. *Polymers (Basel).* **6,** 491–514 (2014).
64. Almora-Barrios, N. & De Leeuw, N. H. A density functional theory study of the interaction of collagen peptides with hydroxyapatite surfaces. *Langmuir* **26,** 14535–14542 (2010).
65. Domene, C., Jorgensen, C. & Abbasi, S. A perspective on structural and computational work on collagen. *Phys. Chem. Chem. Phys.* **18,** 24802–24811 (2016).
66. Miller, M., Némethy, G. & Scheraga, H. Calculation of the structures of collagen models. Role of interchain interactions in determining the triple-helical coiled-coil conformation. 2. Poly (glycyl-prolyl-hydroxyprolyl). *Macromolecules* **478,** 470–478 (1980).
67. Keshwani, N., Banerjee, S., Brodsky, B. & Makhatadze, G. I. The role of cross-chain ionic interactions for the stability of collagen model peptides. *Biophys. J.* **105,** 1681–1688 (2013).
68. Gurry, T., Nerenberg, P. S. & Stultz, C. M. The contribution of interchain salt bridges to triple-helical stability in collagen. *Biophys. J.* **98,** 2634–2643 (2010).

69. S. Sundar, R., Gopalakrishnan, R., Wade, R. C. & Subramanian, V. Structural basis for the varying propensities of different amino acids to adopt the collagen conformation. *J. Phys. Chem. B* **115,** 2593–2607 (2011).
70. Raman, S. S., Parthasarathi, R., Subramanian, Y. & Ramasami, T. Role of length-dependent stability of collagen-like peptides. *J. Phys. Chem. B* **112,** 1533–1539 (2008).
71. Vesentini, S., Fitié, C. F. C., Montevecchi, F. M. & Redaelli, A. Molecular assessment of the elastic properties of collagen-like homotrimer sequences. *Biomech. Model. Mechanobiol.* **3,** 224–234 (2005).
72. Chang, S. W. & Buehler, M. J. Molecular biomechanics of collagen molecules. *Mater. Today* **17,** 70–76 (2014).
73. Tourell, M. C. & Momot, K. I. Molecular dynamics of a hydrated collagen peptide: Insights into rotational motion and residence times of single-water bridges in collagen. *J. Phys. Chem. B* **120,** 12432–12443 (2016).
74. Fu, I., Case, D. A. & Baum, J. Dynamic Water-Mediated Hydrogen Bonding in a Collagen Model Peptide. *Biochemistry* **54,** 6029–6037 (2015).
75. Almora-Barrios, N. & De Leeuw, N. Modelling the interaction of a Hyp-Pro-Gly peptide with hydroxyapatite surfaces in aqueous environment. *Cryst. Eng. Comm.* **12,** 960–967 (2010).
76. Tao, J. *et al.* Energetic basis for the molecular-scale organization of bone. *Proc. Natl. Acad. Sci.* **112,** 326–331 (2015).
77. Nair, A. K., Libonati, F., Qin, Z., Dimas, L. S. & Buehler, M. J. in *Biomineralization Sourcebook. Characterization of Biominerals and Biomimetic Materials* (ed. Gower, E. D. and L. B.) 337–349 (CRC Press, Taylor & Francis Group, 2014). doi:10.1007/s13398-014-0173-7.2
78. Katti, D. R., Pradhan, S. M. & Katti, K. S. Directional dependence of hydroxyapatite-collagen interactions on mechanics of collagen. *J. Biomech.* **43,** 1723–1730 (2010).
79. Qin, Z., Gautieri, A., Nair, A. K., Inbar, H. & Buehler, M. J. Thickness of hydroxyapatite nanocrystal controls mechanical properties of the collagen-hydroxyapatite interface. *Langmuir* **28,** 1982–1992 (2012).
80. Libonati, F., Nair, A. K., Vergani, L. & Buehler, M. J. Mechanics of collagen-hydroxyapatite model nanocomposites. *Mech. Res. Commun.* **58,** 17–23 (2014).
81. Almora-Barrios, N. & De Leeuw, N. H. Molecular dynamics simulation of the early stages of nucleation of hydroxyapatite at a collagen template. *Cryst. Growth Des.* **12,** 756–763 (2012).
82. Ma, J. A molecular dynamics study on the nucleation of calcium phosphate regulated by collagen. *J. Mater. Sci.* **49,** 3099–3106 (2014).
83. Orgel, J. P. R. O., Irving, T. C., Miller, A. & Wess, T. J. Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci.* **103,** 9001–9005 (2006).
84. Nair, A. K., Gautieri, A., Chang, S.-W. & Buehler, M. J. Molecular mechanics of mineralized collagen fibrils in bone. *Nat. Commun.* **4,** 1724 (2013).
85. Streeter, I. & De Leeuw, N. H. A molecular dynamics study of the interprotein interactions in collagen fibrils. *Soft Matter* **7,** 3373–3382 (2011).
86. Streeter, I. & De Leeuw, N. H. Atomistic modeling of collagen proteins in their fibrillar environment. *J. Phys. Chem. B* **114,** 13263–13270 (2010).
87. Gautieri, A., Vesentini, S., Redaelli, A. & Buehler, M. J. Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. *Nano Lett.* **11,** 757–766 (2011).
88. Nair, A. K., Gautieri, A. & Buehler, M. J. Role of Intrafibrillar Collagen Mineralization in Defining the Compressive Properties of Nascent Bone. *Biomacromolecules* **15,** 2494–2500 (2014).
89. Masic, A. *et al.* Osmotic pressure induced tensile forces in tendon collagen. *Nat. Commun.* **6,** 5942 (2015).
90. Li, T. *et al.* Studies of chain substitution caused sub-fibril level differences in stiffness and

ultrastructure of wildtype and oim/oim collagen fibers using multifrequency-AFM and molecular modeling. *Biomaterials* **107,** 15–22 (2016).
91. Zitnay, J. L. *et al.* Molecular level detection and localization of mechanical damage in collagen enabled by collagen hybridizing peptides. *Nat. Commun.* **8,** 14913 (2017).
92. Collier, T. A., Nash, A., Birch, H. L. & De Leeuw, N. H. Preferential sites for intramolecular glucosepane cross-link formation in type I collagen: A thermodynamic study. *Matrix Biol.* **48,** 78–88 (2015).
93. Collier, T. A., Nash, A., Birch, H. L. & de Leeuw, N. H. Intra-molecular lysine-arginine derived advanced glycation end-product cross-linking in Type I collagen: A molecular dynamics simulation study. *Biophys. Chem.* **218,** 42–46 (2016).
94. Tang, M., Li, T., Gandhi, N. S., Burrage, K. & Gu, Y. T. Heterogeneous nanomechanical properties of type I collagen in longitudinal direction. *Biomech. Model. Mechanobiol.* **16,** 1023–1033 (2017).
95. Zhou, Z., Minary-Jolandan, M. & Qian, D. A simulation study on the significant nanomechanical heterogeneous properties of collagen. *Biomech. Model. Mechanobiol.* **14,** 445–457 (2015).
96. Zhou, Z., Qian, D. & Minary-Jolandan, M. Clustering of hydroxyapatite on a super-twisted collagen microfibril under mechanical tension. *J. Mater. Chem. B* **5,** 2235–2244 (2017).
97. Schröder, K. P. *et al.* Bridging hydrodyl groups in zeolitic catalysts: a computer simulation of their structure, vibrational properties and acidity in protonated faujasites (H-Y zeolites). *Chem. Phys. Lett.* **188,** 320–325 (1992).
98. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, A. W. ReaxFF: A Reactive Force Field for Hydrocarbonds. *J. Phys. Chem. A* **105,** 9396–9409 (2001).
99. Buehler, M. J. Atomistic and continuum modeling of mechanical properties of collagen: Elasticity, fracture, and self-assembly. *J. Mater. Res.* **21,** 1947–1961 (2006).
100. Buehler, M. J. Molecular nanomechanics of nascent bone: fibrillar toughening by mineralization. *Nanotechnology* **18,** 295102 (2007).
101. Gautieri, A., Uzel, S., Vesentini, S., Redaelli, A. & Buehler, M. J. Molecular and mesoscale mechanisms of osteogenesis imperfecta disease in collagen fibrils. *Biophys. J.* **97,** 857–865 (2009).
102. Buehler, M. J. Nature designs tough collagen: Explaining the nanostructure of collagen fibrils. *Proc. Natl. Acad. Sci.* **103,** 12285–12290 (2006).
103. Fang, F. & Lake, S. P. Modelling approaches for evaluating multiscale tendon mechanics. *Interface Focus* **6,** 20150044 (2016).
104. Buehler, M. J. & Genin, G. M. Integrated multiscale biomaterials experiment and modelling: a perspective. *Interface Focus* **6,** 20150098 (2016).
105. Gautieri, A., Russo, A., Vesentini, S., Redaelli, A. & Buehler, M. J. Coarse-grained model of collagen molecules using an extended MARTINI force field. *J. Chem. Theory Comput.* **6,** 1210–1218 (2010).
106. Depalle, B., Qin, Z., Shefelbine, S. J. & Buehler, M. J. Influence of cross-link structure, density and mechanical properties in the mesoscale deformation mechanisms of collagen fibrils. *J. Mech. Behav. Biomed. Mater.* **52,** 1–13 (2015).
107. Depalle, B., Qin, Z., Shefelbine, S. J. & Buehler, M. J. Large Deformation Mechanisms, Plasticity, and Failure of an Individual Collagen Fibril with Different Mineral Content. *J. Bone Miner. Res.* **31,** 380–390 (2016).
108. Jung, G. S. & Buehler, M. J. Multiscale Modeling of Muscular-Skeletal Systems. *Annu. Rev. Biomed. Eng.* **19,** 435–457 (2017).

CHAPTER III. Results and Discussion

Abstract

The open questions presented in the previous chapters (see perspectives sections of Chapter 1-2) can be addressed employing *ab initio* simulation as investigation technique. The main challenge in this approach is reducing the complexity of the biological systems (tendon and bones) to a simpler level treatable by the costly *ab initio* calculations. Its use is inevitable considering our ambitious goals:

- unrevealing the energetic reasons for the collagen folding and collagens aggregation (tendons),
- clear up the collagen-hydroxyapatite interface, which is a fundamental as much as un-known interface between the organic and inorganic components of bones.

A rational for these phenomena is not entirely feasible to date, but we hope to address several interesting side questions.

As explained in Chapter 2, in any computational study, the first step is the choice of the model-size/theoretical-framework. In this PhD thesis, I followed different roads:

- Using "**fast" *ab initio* methods** which enlarges dramatically the possibility in molecular modelling with respect to conventional *ab initio* methods. I am referring to the HF-3c method, described in Chapter 2. Due to severe classical correction in the HF-3c method, I had to test it extensively. All test cases were periodic systems such as: molecular crystals, inorganic materials and collagen models. The choice of the systems was dictated by the need to have all the representative interactions occurring in collagen models interacting with hydroxyapatite. The work done in this direction is described in the **Section 3.1 Seeking a Chimera. Cost-effective *Ab Initio* Method for Large Periodic Simulations**.
- Focusing on **specific phenomena** that can be flawlessly extracted from the biological complexity and so described with simple models with high accuracy. See **Section 3.2 Role of Pro Pucker in Bones**.
- **The smart modelling of collagen and aggregate**, to reach protein complexity in a step by step fashion taking advantage of the symmetric nature of collagen and aggregates. **Section 3.3 Roto-translational Symmetry of Collagen Protein and Aggregates. Few Applications**.

The results that lead to a publication are only reported briefly. Full details are in the original papers, which are referred in this thesis.

3.1 Seeking a Chimera. Cost-Effective *Ab Initio* Method For Large Periodic Simulations

Among all the "fast" ab initio methods, we chose the HF-3c method. According to Ref. [1] *the method is thought to "fill the gap between semi-empirical and DFT methods in terms of cost and accuracy". It is one the simplest flavour of an ab initio calculation but it can still compute interaction energies and geometries with reasonable accuracy for molecular systems with a tiny fraction of the computational cost of standard DFT methods.*[1] *The interest on the method is rising as it has been implemented in ORCA*[2]*, TURBOMOLE*[3] *and CRYSTAL*[4] *(for periodic calculations) suites, and very recently, it has been also employed for AIMD simulations.*[5] *Unfortunately, in the application of last year, some concern about the accuracy emerged.*[6,7] *Due to the un-clear accuracy and large amount of classical corrections, the method needs extensive testing. HF-3c has been tested on molecular systems, but we are interested in simulation with periodic boundary condition because tendon and bone are much simpler modelled by 3D structures rather than molecules (it is not easy to make a reasonable molecular cluster of hydroxyapatite). Therefore, we extensively tested it in the periodic systems such as: molecular crystals, inorganic layered materials and polymeric models of collagen.*

Molecular crystals

In this section of the thesis, I report the work done in testing the HF-3c and other methods in predicting **geometry**, **cohesive energy** and **vibrational spectra** of **molecular crystals**. Molecular crystals are a good choice for our purposes, because:

- the experimental data on geometry and energy of several molecular crystals are available in literature in the form of benchmark sets
- different theoretical approaches have been already tested on the above-mentioned sets, thus facilitating the comparative study
- a molecule in the 3D array of a crystal experiences all types of intermolecular interactions ranging from weak London dispersion forces to strong H-bonds. As already mentioned, the description of weak London dispersion forces is missed in standard DFT or HF methods, see Chapter 2. Therefore, molecular crystals are the perfect ground for testing the reliability of dispersion corrected methods.

In this work we compared several methods, such as:

- **HF-3c** (as fast *ab initio*),
- **B3LYP-D*/TZP** (as standard *ab initio*) and
- **Local MP2 with medium size basis set** (post-HF method). MP2 method, which is routinely used in non-periodic simulation, is rarely employed for 3D systems due to the complexity in extending the formalisms from molecules (0D) to crystals (3D). As a consequence of this, the MP2 forces calculation is not currently available for periodic calculation (no geometry relaxation is possible). Despite this, it is the only method which includes a fraction of dynamical correlation to naturally take London dispersion into account, without resorting to additional empirical or classical correction to the energy.

Anticipating the results which will follow later in the text, during the HF-3c testing, we understood that there was room for improvement. The HF-3c method systematic over-binds purely dispersion bonded systems, e.g. crystals of benzene, naphthalene etc. To increase its accuracy without altering too much the whole set of parameters, we simply scaled the dispersion energy term in HF-3c energy expression (Chapter 2). As trial case for the tuning procedure, we have used graphite, whose carbon graphene layers are mutually interacting through dispersive London forces. This led to three different HF-3c flavours:

- **HF-3c** (the standard version)
- **S-HF-3c** (the s_8 term on the D3 scheme is reduced to 70%)

- **HF-3c-027** (the s_8 term on the D3 scheme is reduced to 27%).

The scaling factors on the s_8 term of D3 scheme were chosen because they gave the best agreement with testing cases on the energy side (for the 70% case) and on structures (for the 27% case).

We considered **82 molecular crystal** belonging to three literature benchmark sets, e.g. the **X23** set, the **G60** set and the **K7** set, ranging from purely **dispersion** bonded to **H-bond** dominated systems. The results for **geometry** (**Figure 1)**, cohesive **energy (Figure 2)** and **vibrational frequencies (Figure 3)** predictions are reported only for the X23 set. This is done because the X23 set is the most known and well-tested molecular crystal benchmark set (even so we have found some relevant errors in it, see Ref. [8] for further details). A **statistic for the overall 82 systems (MC82 set)** is reported in **Figure 4**.

Let us start the discussion from the crystal cell volume prediction by HF-3c and B3LYP-D*/TZP reported in Figure 1, where we illustrate the relative deviations in percentage of the optimized crystal cell volumes from low temperature experimental data. As expected, HF-3c shows a systematic tendency to shrink the unit cell volume with respect to experiment. This contraction effect is larger for the systems in which dispersion forces are more important, e.g. dispersion dominated (crystal label 1-10) and mixed subgroup (H bond + dispersion, crystal label 20-23). The strongest differences can either arise due to an intrinsic error of HF-3c or due to the typically shallower potential energy surface of pure π-systems in contrast to hydrogen bonded ones (crystal label 11-19). In contrast, but not unexpectedly, the B3LYP-D*/TZP method performs better with the computed mean absolute relative error in percentage (MARE%) of 3.0% and a percentage standard deviation (SD%) of 1.9%. HF-3c method is less accurate, the error in the computes volumes being 8.0% of MARE% and SD% of 2.5%.

The volumes computed with the scaled dispersion term are in better agreement with the experiment than the pure HF-3c, see Figure 1. As expected, the scaling of the s_8 factor leads to a systematic expansion of the cell volume due to the reduced role of London interactions. The MARE% of 8.0% for plain HF-3c, now drops to 1.9% for HF-3c-0.27. It is worth noticing that the percentage standard deviation does not change when tuning the dispersion term. The scaling procedure leads to a systematic expansion of the unit cell volume, that induces a reduction of the mean relative error in percentage (MRE%), i.e. -8.0%, -5.2% and -0.8% for the HF-3c, S-HF-3c and HF-3c-0.27 methods, respectively. **Notably, the HF-3c-027 computes molecular crystal volumes closer to experiment than the B3LYP-D*/TZP method, see Figure 1.**

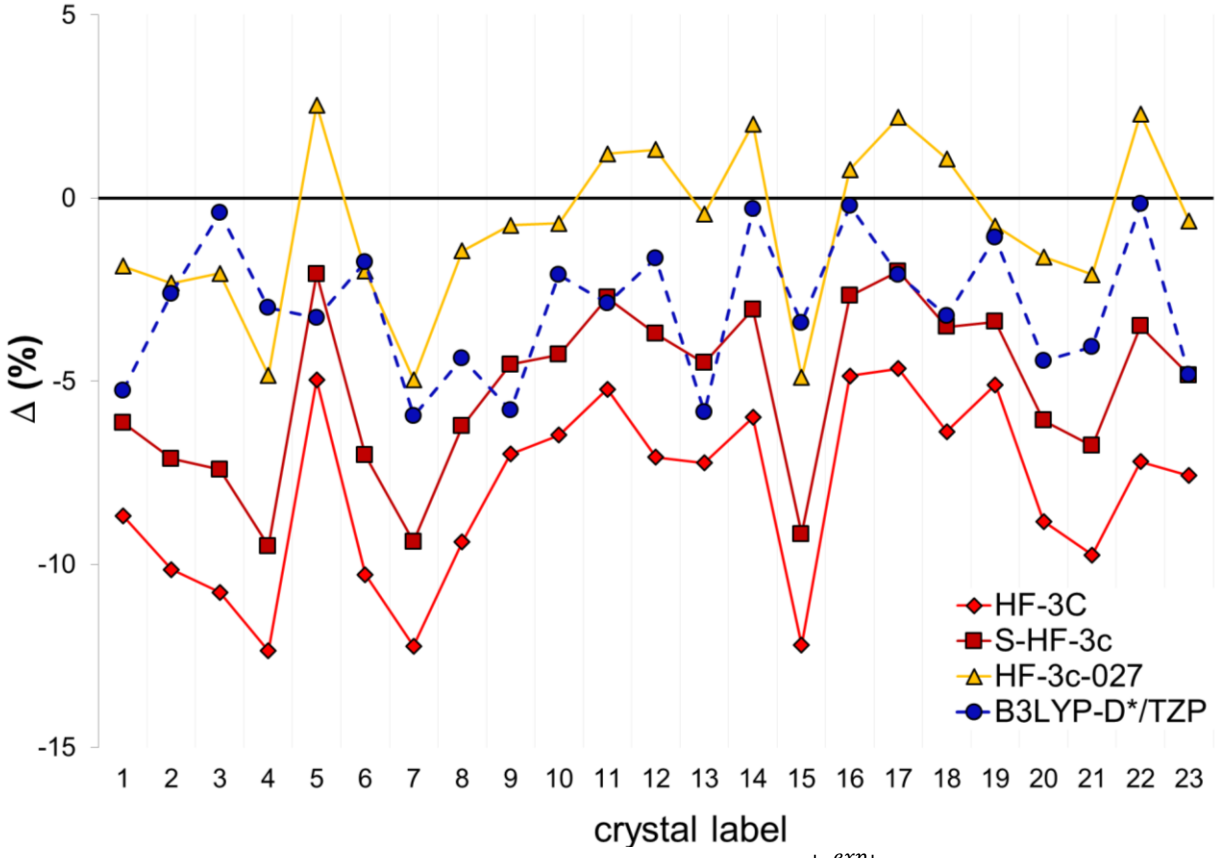


Figure 1. Relative deviation in percentage, $\Delta_i(\%) = 100 \cdot \frac{|x_i| - |x_i^{exp}|}{|x_i^{exp}|}$, of the computed cell volumes from the experimental reference values for the X23 set as obtained at the HF-3c and the B3LYP-D*/TZP level of theory as well as for the scaled S-HF-3c and HF-3c-0.27 methods. For details of crystal labelling see Ref. [8].

As observed for the unit cell volumes, HF-3c gives overestimated cohesive energies (excess of binding), with the largest deviation for systems dominated by dispersion interactions with a mean absolute error (MAE) of 10.3 kJ·mol^{-1} (see Figure 2, crystal label 1-10). For the overall X23 set, the computed ΔE_c are overestimated, as shown by a mean signed error (ME) of 9.4 kJ·mol^{-1}. Remarkably, the B3LYP-D*/TZP method performs very well: we attained a MAE of 4.6 kJ·mol^{-1} for the X23 set that is close to the "chemical accuracy" (i.e. 4.2 kJ·mol^{-1}).[9] As the volume analysis suggests, when decreasing the dispersive contribution to the HF-3c total energy leads to cohesive energies in closer agreement with the experiment. The S-HF-3c method reduces the MAE from 10.3 to 4.5 kJ·mol^{-1}, for the dispersion dominated systems. This scaling increases the overall accuracy thus yielding a MAE of the whole X23 set of 6.9 kJ·mol^{-1}. A further reduction of the dispersion term, i.e. by the 0.27 scaling factor, leads to a loss of binding energy within the crystals, with a general underestimation of the ΔE_c. Indeed, the computed MAE is 13.1 kJ·mol^{-1} for HF-3c-0.27.

In addition, since the HF-3c-0.27 resulted to give the best crystal structures among the HF-3c scaled variants, the cohesive energies for the X23 were also computed through single-point energy calculations at the B3LYP-D*/TZP level on the HF-3c-0.27 optimized structures. This approach has been denoted hereafter as **SP-B3LYP-D***. Encouraging, high accuracy is achieved by this approach, which shows a MAE of 5.2 kJ·mol^{-1} with a SD of 5.8 kJ·mol^{-1}. This is a very promising result that paves the way to its application in crystal structure prediction due to the very cheap cost of crystals geometry optimization at the HF-3c-0.27 level.

As another important benchmark in the present work, we computed the cohesive energies of the X23 set at the LMP2 level of theory. In this case, single-point energy calculations were carried out on the B3LYP-D*/TZP optimized structures by using a p-aug-6-31G(d,p) basis set (LMP2/p-aug-

6-31G(d,p)). Since the LMP2 energy includes electron correlation effects, we expected that the LMP2 method would properly account for the intermolecular dispersion forces. LMP2 performs moderately well in computing the cohesive energies. However, the accuracy is lower than most of the methods presented above with a MAE of 9 kJ·mol^{-1} and a SD of 11 kJ·mol^{-1}, see Figure 2.

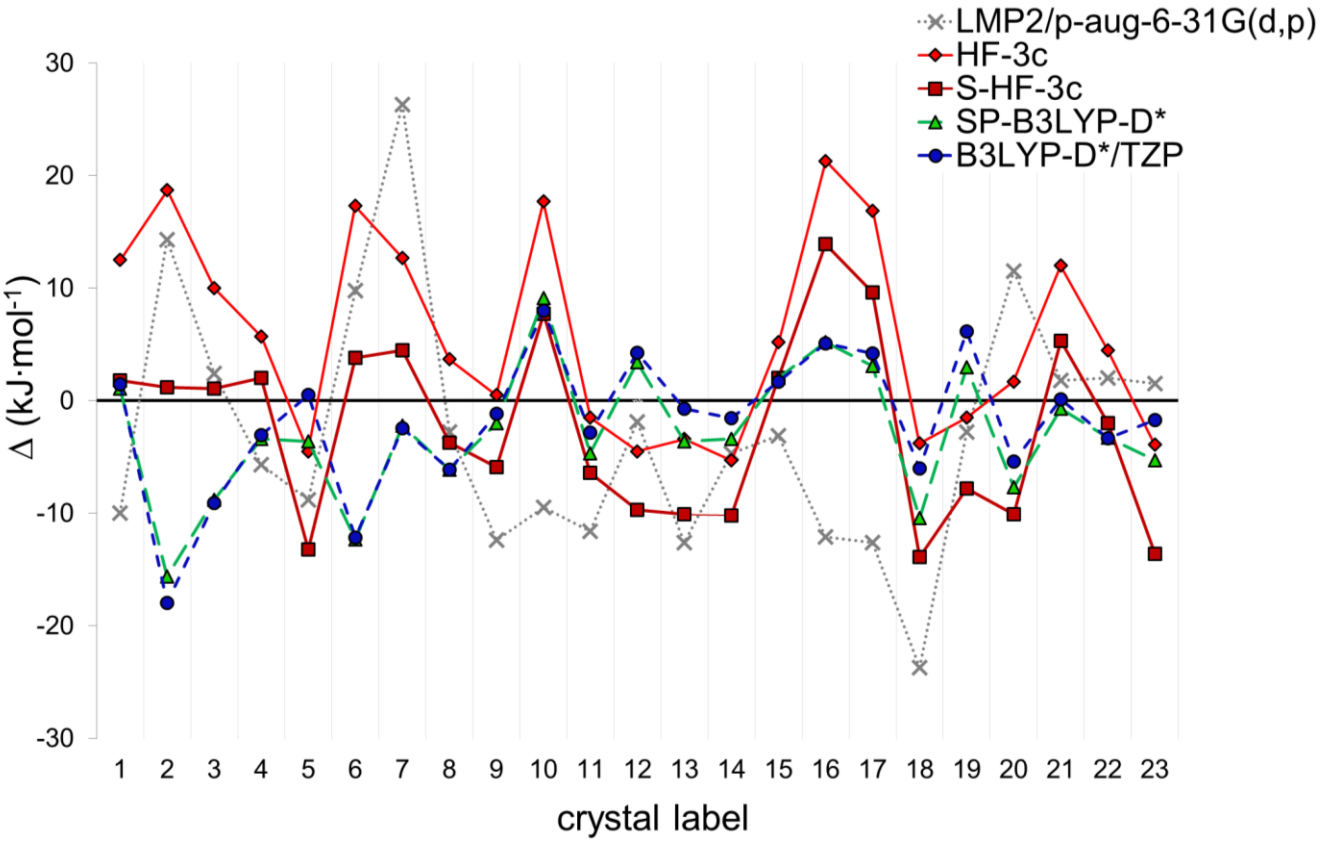


Figure 2. Deviation ($\Delta_i = |x_i| - |x_i^{exp}|$) of the ΔE_c from the thermodynamically corrected experimental cohesive energies for the X23 set as computed at the LMP2, HF-3c, S-HF-3c, SP-B3LYP-D* and B3LYP-D*/TZP levels of theory. For details on crystal labelling see Ref. [8].

To further analyse the accuracy of the HF-3c method in computing the properties of molecular crystals we have evaluated the sublimation enthalpies, ΔH_{sub}, for the X23 set. This is a more stringent test as it implies the calculation of the vibrational frequencies of the crystals. For the HF-3c and S-HF-3c methods computed cohesive energies were corrected by including the thermal contribution to enthalpy ($\mathbf{\Delta E_{vib} + 4RT}$). In Figure 3, we compare the ΔE_{vib} + 4RT obtained for the S-HF-3c with accurate DFT calculations,[10] and the widely adopted 2RT correction. We also considered a reduced ΔE_{vib} contribution by scaling the zero point and thermal energy to account for the systematic error of the HF method which systematically overestimates the vibrational frequencies.[11] Figure 3 shows that the S-HF-3c computed thermal corrections for the different molecular crystals are in good agreement with the DFT results. Interestingly, the best agreement is reached for the dispersive subgroup. **It is important to point out that, on average, the thermal correction to enthalpy is -7.1 kJ·mol^{-1} and -6.3 kJ·mol^{-1} for S-HF-3c and DFT, respectively. These values are not far from the 2RT constant correction (about -5.0 kJ·mol^{-1})**. This then suggests that such correction can be used as an easy and fast way to estimate experimental cohesive energies from the experimental sublimation enthalpies without the need of expensive vibrational frequencies calculations. Figure 3 also shows that the scaling of the vibrational quantities does not lead to any change in our results.

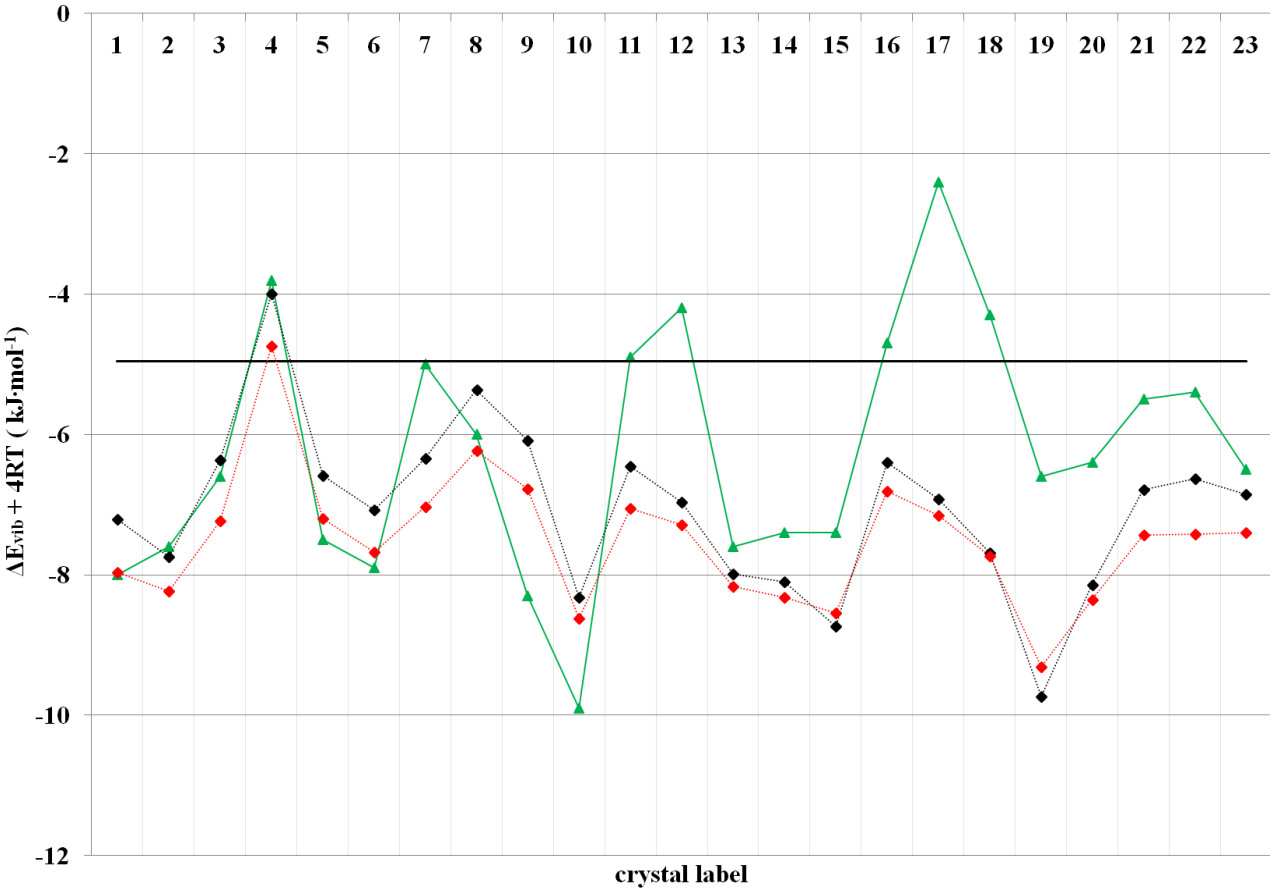


Figure 3. Comparison of the thermal correction to enthalpy ($\Delta E_{vib} + 4RT$) for the X23 set as evaluated at S-HF-3c (red squares, dotted red line) and at the DFT level of theory (green triangle, solid green line).[10] The usually adopted 2RT value is also shown as black solid line. In addition, the results for the S-HF-3c level of theory with scaled vibrational frequencies are also reported (black squares, black dotted line). For details see Ref. [8].

A summary of the performances of HF-3c, S-HF-3c, SP-B3LYP-D* (HF-3c-027 for volumes and B3LYP-D*/TZP for energies) and full B3LYP-D*/TZP on all the sets is reported in Figure 4. The reference cohesive energies have been estimated from the experimental sublimation energies by applying the 2RT correction factor, which we have shown is not far from the thermal contribution to enthalpy obtained through vibrational frequencies calculations. The scaled HF-3c-0.27 method shows extraordinary accuracy in the volume prediction with an overall MARE% of 3.4%. On the energy side, the SP-B3LYP-D* approach gives a good accuracy with a global MAE of 7.4 kJ/mol. The main source of error in the energy estimation arises from the G60 set. This may be due to the presence of some old experimental data with very large error bar.

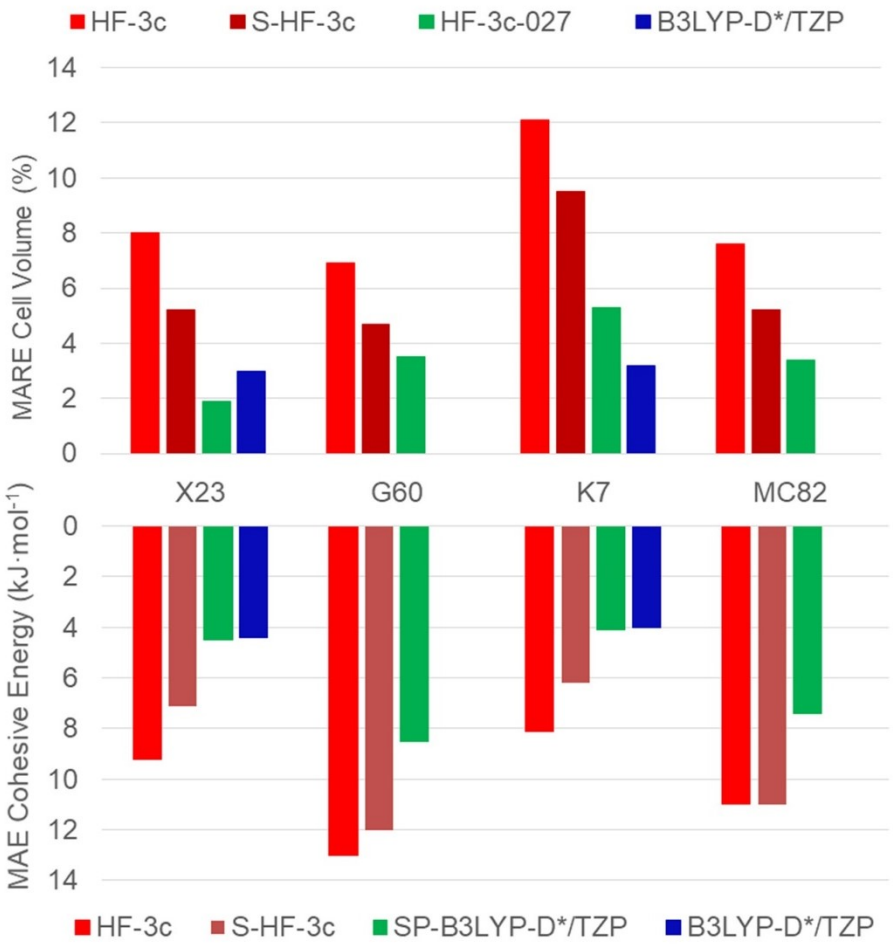


Figure 4. X23, G60 and K7 summary (82 molecular crystals, MC82 set). MAE of the ΔE_c and MARE% for the cell volumes evaluated at the HF-3c, S-HF-3c, HF-3c-0.27 (SP-B3LYP-D* for the energy only) and B3LYP-D*/TZP levels of theory.

Inorganic crystals

In this section of the thesis, I have testes the HF-3c method and their refinements (HF-3c-027 and SP-B3LYP-D) on a selected group of inorganic materials. This is a mandatory step for our purposes. Indeed, bone contains large amount of inorganic material and, at moment, studies dealing with the performance of HF-3c on this type of systems are entirely missing. As test cases, I focused on microporous material, e.g. **zeolites** as they have been heavily studied from both theory and experiments. I employed a set of **14 zeolites,** recently studied by Zicovich-Wilson et al.[12] The test is performed on the prediction of geometrical, energetical and vibrational features of the above mentioned set.

First of all we tested the HF-3c methods in computing the fully relaxed geometry for the zeolite set. Figure 5 reports the percentage deviation on the HF-3c and HF-3c-027 computed crystal volumes with respect to low temperature experimental data. I have included in Figure 5 also the results for the hybrid DFT functionals, PBE0-D2 and B3LYP-D2, taken from Ref.[12]. As for molecular crystals, the **HF-3c-027 optimized volumes agree remarkably well with the experimental data**. Interestingly, in this case the dispersion plays a small role. Indeed, the un-scaled HF-3c method performs just slightly worse than the scaled HF-3c-027, see Figure 5. Conversely, hybrid-DFT systematically overestimates the volume for all zeolite set, with PBE0-D2 slightly more accurate than B3LYP-D2 for all the set elements. Surprisingly, **HF-3c methods outperform hybrid-DFT**. The computed MARE% are 1.9%, 2.2%, 5.9% and 7.1% for the HF-3c-027, HF-3c, PBE0-D2 and B3LYP-D2 methods, respectively.

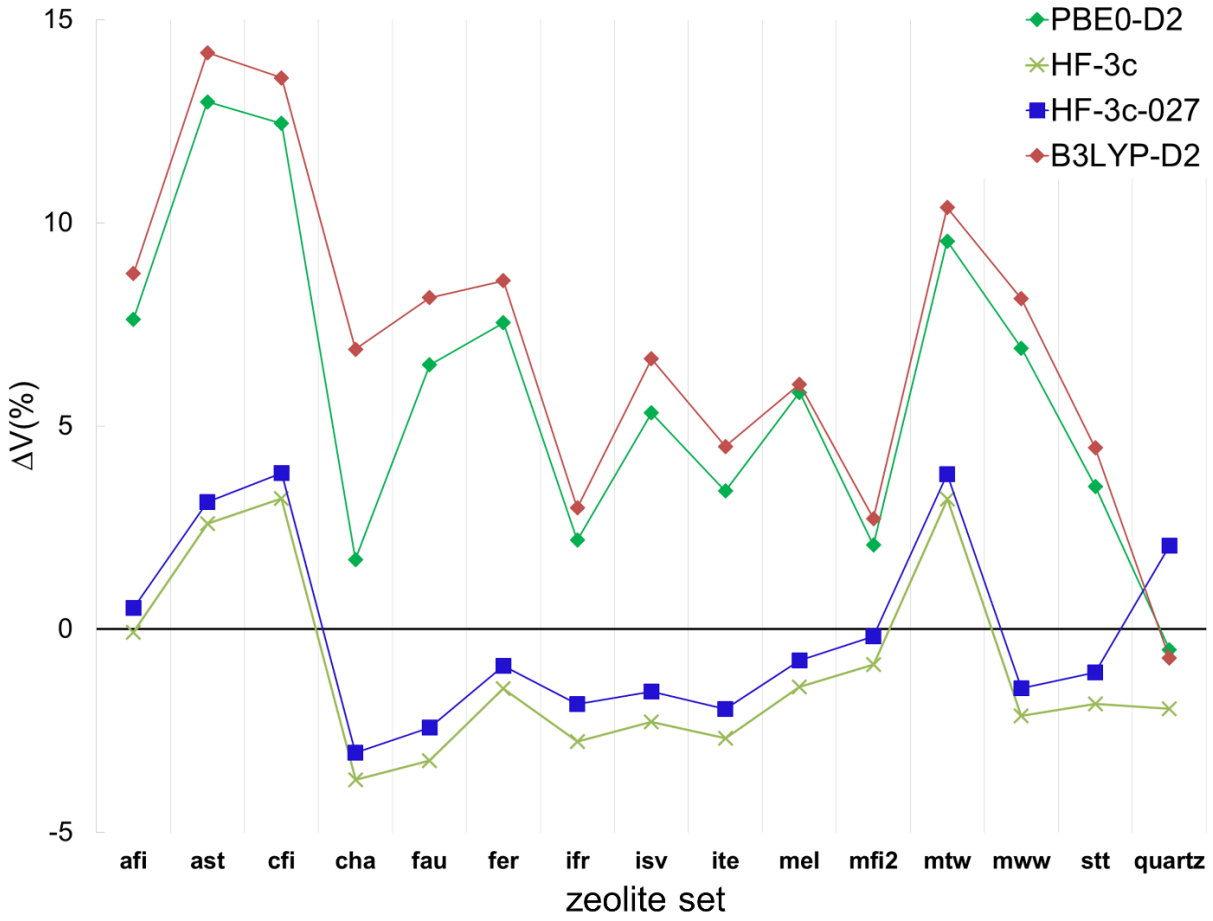


Figure 5. Optimized volume percentage error (ΔV) at the B3LYP-D2/TZP, PBE0-D2/TZP, HF-3c and HF-3c-027 levels with respect to the experiment for the zeolite set and α-quartz system. The data of for the hybrid-DFT are taken from Ref. [12].

To assess the performance of HF-3c methods on the energy, I have calculated the **energy of formation of all the zeolites with respect to that of α-quartz (the most stable all silica polymorph)** comparing with accurate experimental enthalpy of formation, see Figure 6. I have tested the HF-3c methods (HF-3c and HF-3c-027) along with the B3LYP-D//HF-3c-027 (D≡D* and $D3^{ABC}$) approach, which gave excellent results for molecular crystals. Figure 6 summarize the results, with the hybrid DFT ones (B3LYP-D2 and PBE0-D2), taken from Ref. [12]. The results in Figure 6A indicate that HF-3c overshoots zeolite energy of formation, but reducing the dispersion energy (HF-3c-027) leads to better results. In statistical terms, reducing the dispersion reduces the MAE from 7.4 $kJ \cdot mol^{-1}$ to 3.7 $kJ \cdot mol^{-1}$, with constant SD of 2.4 $kJ \cdot mol^{-1}$. The results are less scattered for hybrid DFT with a SD of 1.3 $kJ \cdot mol^{-1}$ for both B3LYP-D2 and PBE0-D2, with the latter more accurate than the former. Indeed, the computed MAE is 6.8 $kJ \cdot mol^{-1}$ and 2.6 $kJ \cdot mol^{-1}$for B3LYP-D2 and PBE0-D2, respectively.

Remarkably, using the **B3LYP-D//HF-3c-027 approach gives excellent results**. The bests are achieved employing the D* dispersion scheme (**SP-B3LYP-D***), which lead to a MAE ± SD of 1.5 $kJ \cdot mol^{-1}$ ± 1.3 $kJ \cdot mol^{-1}$, **outperforming the full hybrid DFT-D** (DFT-D//DFT-D) **approaches, at a tiny fraction of their computational cost.**

We extended the testing of the HF-3c methods on **the estimation of the vibrational zero point energy (ZPE)**. Including the ZPE in the energy of formation calculation leads to a quantity

which is in the between of a pure electronic **energy of formation** and an **enthalpy of formation**. I followed a procedure of this type:

- I have computed the frequencies of vibration at the HF-3c-027 level for all zeolites and quartz.
- I have scaled the frequency of vibrations by a scaling factor s, (s=0.8771). This value minimizes the RMSD of the HF-3c-027 vibrational frequencies from the B3LYP-D*/TZP ones for the α–quartz case.
- In this way, I have calculated the scaled ZPE corrections, and I have combined them with the SP-B3LYP-D* electronic energy of formation. The resulting method is named ΔH*-SP-B3LYP-D* (0 K).

The inclusion of the ZPE leads to a general up-shift of the formation energy, see Figure 6B, which decreases the agreement of the pure energetic SP-B3LYP-D* with the experiments.

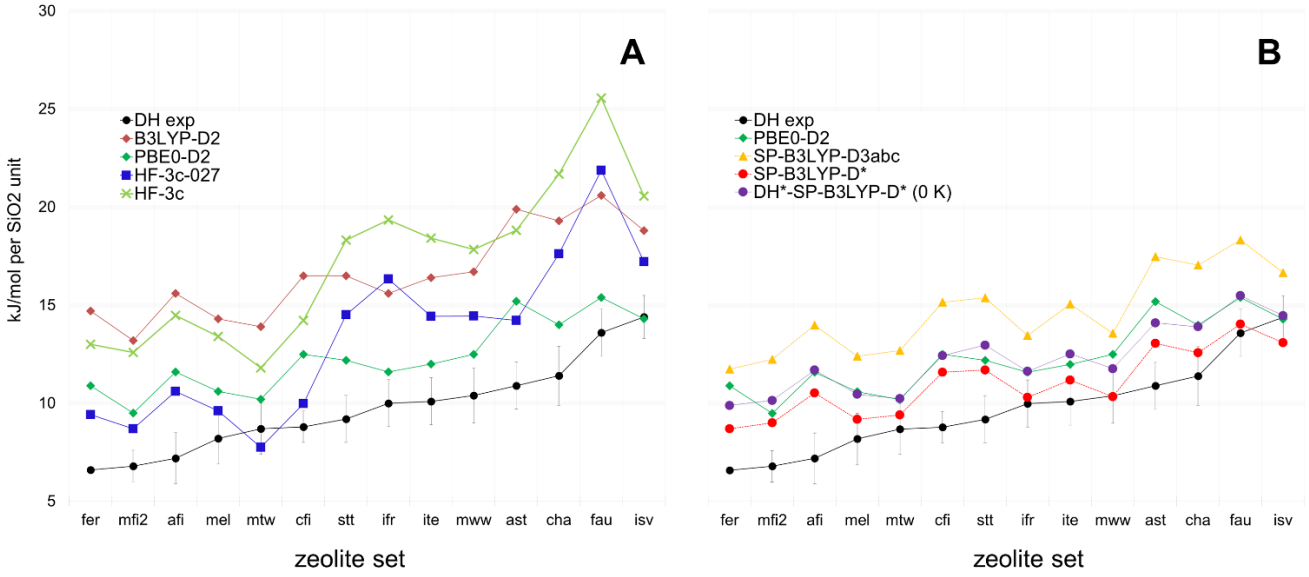


Figure 6. Comparison between calculated and experimental energy of formation of zeolites with respect to α-quartz. The experimental error bar is also included. All the methods employed are divided in two groups, A and B, for sake of clarity. The data of for the B3LYP-D2 and PBE-0 methods are taken from Ref. [12].

Pro ring potential energy surface within a protein

In this section of the thesis I wanted to assess the capacity of the HF-3c method in all its flavours (standard HF-3c, scaled version HF-3c-027, and coupled to hybrid DFT, SP-B3LYP-D*) on a feature which is of interest for the study of collagen protein, e.g. **the side chain conformation of Pro**. To do so, we employed different model systems which contain Pro residues. In this context, I will report only two, e.g. a single **Pro molecule (0D)** and a **Poly-L-Proline type II polymer (PPII) (1D)**. Several DFT functionals, e.g. B3LYP and PBE, with different dispersion corrections schemes, D2, D3 and $D3^{ABC}$, are compared with the HF-3c outcomes. To make this wide comparison we developed a strategy to simplify the intrinsically 2D conformational PES calculation to a 1D PES.

The matter is studied in depth by investigating the **electronic energy potential energy surface** (PES) of the **pyrrolidine ring on pseudo-rotational coordinates** (bi-dimensional), e.g. phase angle (P) and ring amplitude (Q). These two variables, P and Q, can define the low-energy conformations of a five-membered heterocycle. Physically, Q represents the maximum value assumed by χ^j, the j-th torsional angle of a pyrrolidine ring, see Figure 7 A, and P identifies the type of ring puckering. If P assumes values that are even multiples of 18° (0, 36°, 72°, …), the ring is puckered in the half chair conformation (H), while for odd multiples of 18° (18, 54°, …), it adopts an envelope conformation (E). To clarify graphically the matter, Figure 7 B reports the cyclopentane in the E and H conformations. In the E conformation, four atoms lie on the same plane and the remaining atom, which is out of plane, is labelled as C. In the H conformation, A and B are defined as the atoms not directly bonded to F, that is the atom crossed by the C_2 axis. The stable conformers of Pro side chain have been already presented in Figure 4 of Chapter 1. Usually, this ring puckers either in an UP (U) or DOWN (D) conformations, where four out of five ring atoms, namely N, C_α, C_β, and C_δ, lie on a common plane and C_γ carbon is relaxed out of the plane, see Figure 7 C. The reader can find the two puckers (green circles) in the pseudo-rotation pathway of the pyrrolidine ring reported in Figure 7 D. For further details on the concept of pseudo-rotational applied to pyrrolidine rings see Ref.[13].

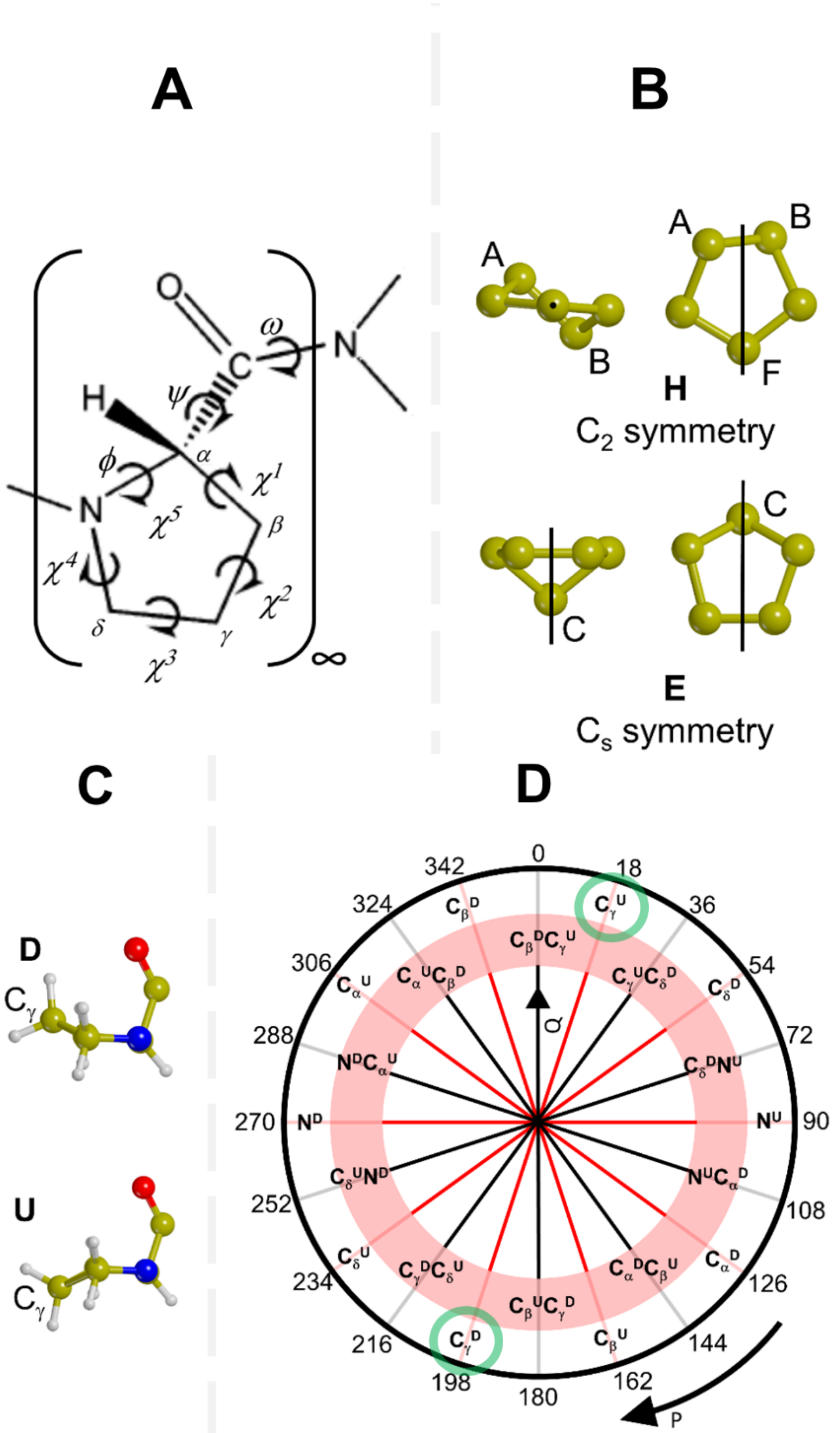


Figure 7. A: Definition of the atoms and torsional angles of a Pro within a polymeric structure. B: Five-membered homo-cycles in the H and E conformations with symmetry axis and mirror plane. C: Pro molecule extracted from a PPII polymer, in an E conformation with D and U C_γ atom, respectively. D: Pseudo-rotation pathway of a pyrrolidine ring (Q is radial and P is angular). For the odd (red line) and even (black line) multiples of 18° (e.g. E and H conformations) C, A and B atoms are reported, respectively.

Initially, we have tested the performance of the HF-3c and HF-3c-027 as well as of the hybrid B3LYP-D*/TZP//HF-3c-027 (SP-B3LYP-D*) approach in computing the 2D-PES for the side chain of PRO in vacuum, see Figure 8. The results for the HF-3c show the coexistence of three stable ring conformations, i.e. A, B and C, Figure 8.The adoption of HF-3c-027 method does not change the minima positions, only slightly affecting their relative stability. When SP-B3LYP-D* is computed,

the shape of the energy surface changes radically. A new large and shallower minimum (D) assimilates the B and C minima leading to a 2D-PES with two wells only, with the D state 2.3 kJ·mol^{-1} higher in energy than the A state, as also reported from CCSD(T)/CBS calculations.[14] The validity of the hybrid SP-B3LYP-D* approach is confirmed also by the full B3LYP-D3^{ABC} method, that is here considered as a reference method within the DFT approaches. The latter method combines both energy and geometry relaxation at the B3LYP level inclusive of the recent D3^{ABC} dispersion scheme. Interestingly, the reference method maintains the same shape of the SP-B3LYP-D* energy surface, only slightly affecting the conformer relative stability. The geometry of the ring in the A and D minima are equivalent to UP and DOWN puckers (with only slight deviation), see Figure 8.

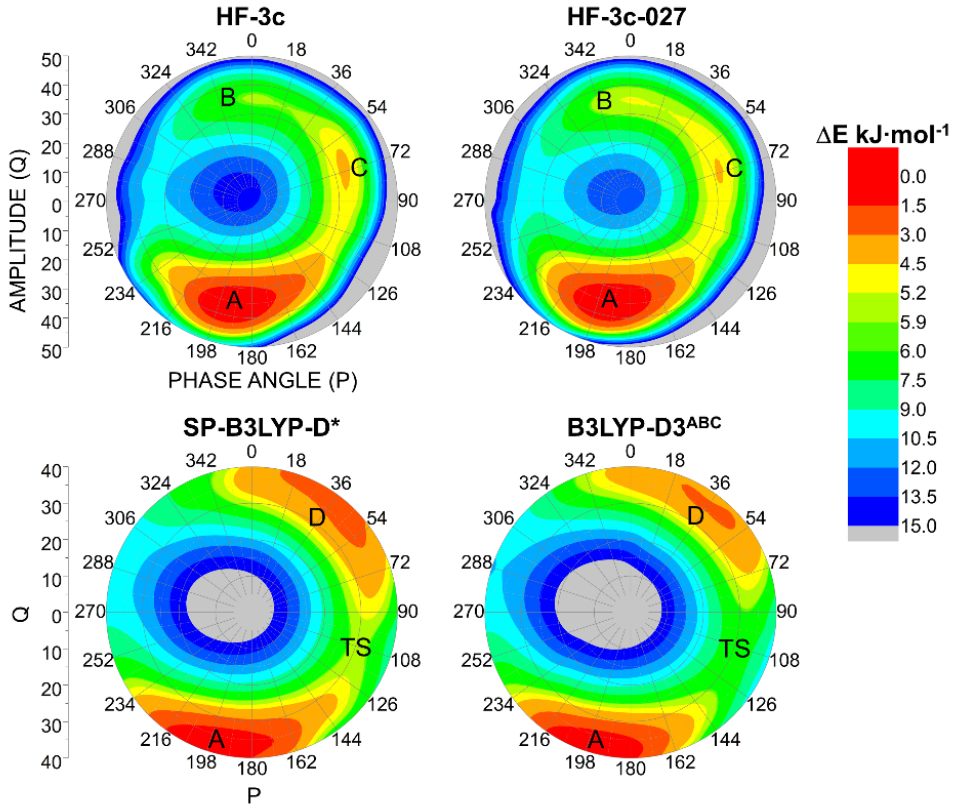


Figure 8. 2D-PES for the pyrrolidine ring of Pro (single Pro molecule) computed with different methods.

As anticipated, we have reduced the computational burden associated to the 2D-PES through the 1D energy scan only of the χ^2 dihedral angle. Scanning the χ^2 angle is equivalent of moving on the 2D PES minimum energy path that links the D and U minima, for further details see Ref.[13] With this simplified approach, we studied the Pro pucker for the PPII polymer.

The PPII polymer is reported in Figure 9 A. Due to the presence of three proline groups in the unit cell, different conformers are possible as shown in Figure 9 B. The DDD state has all Pros in D conformation, the UDD only one Pro in U, the UUD state has only one Pro in the D state while the UUU conformation has all the Pros in U state. It is worth mentioning that the periodic boundary conditions imposed on the PPII polymer, imposes UDD≡DUD≡DDU, as well as UUD≡UDU≡DUU. We focused mainly in the DDD→UDD transition whose reaction path (on the χ^2 dihedral angle only) is reported in Figure 10. The Pro pucker is defined D puckered for negative value of χ^2 and U puckered for positive values of χ^2 angle.

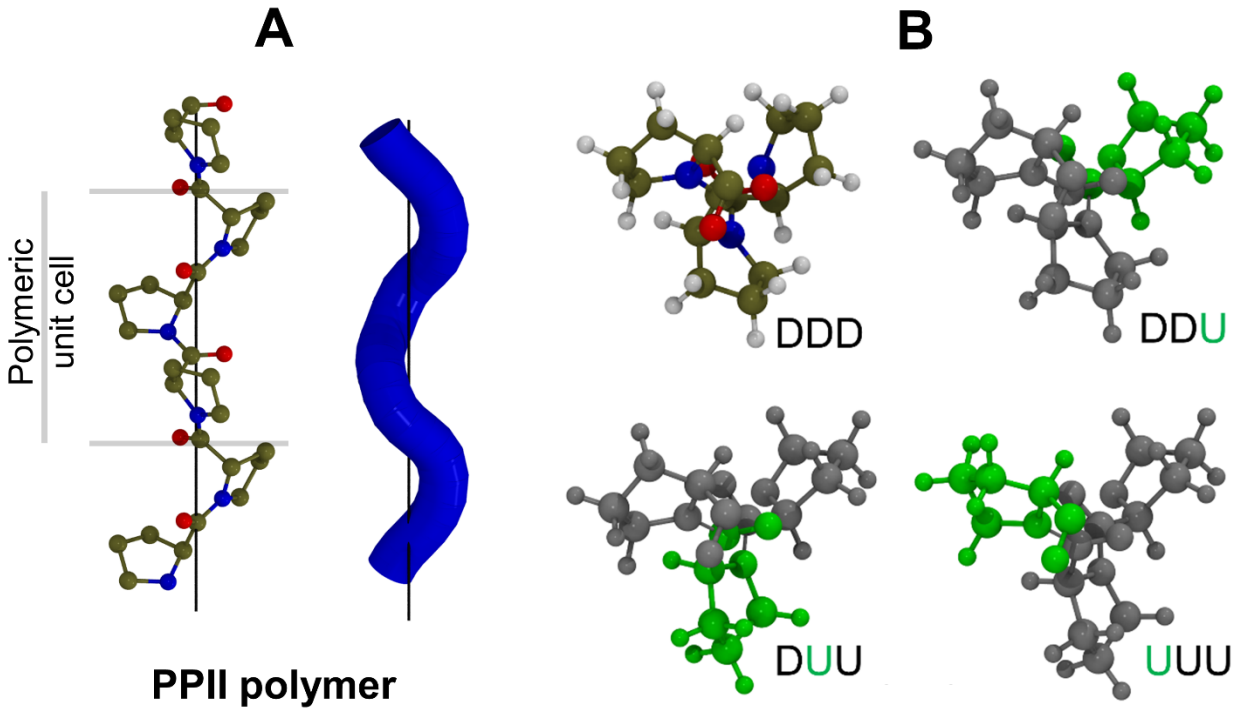
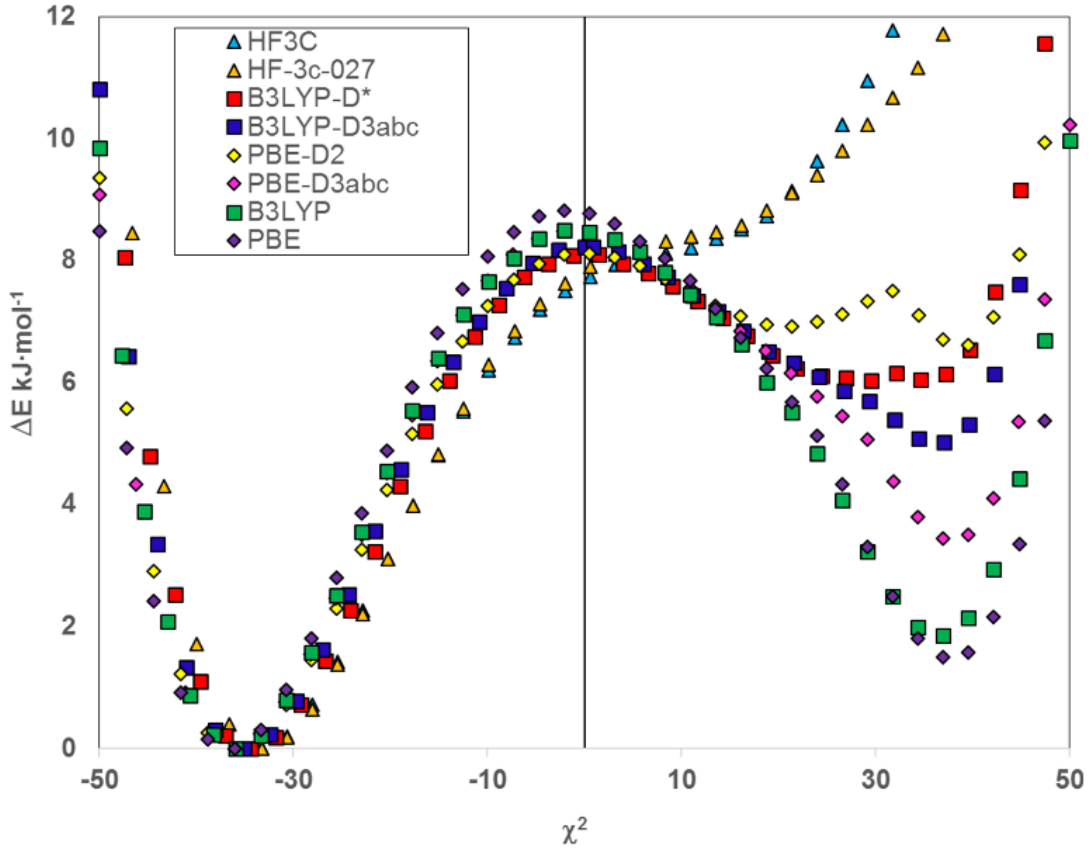


Figure 9. A: Perpendicular view of the PPII polymer. Each PPII unit cell contains three Pros. B: Views along the polymer main axis of the PPII conformers



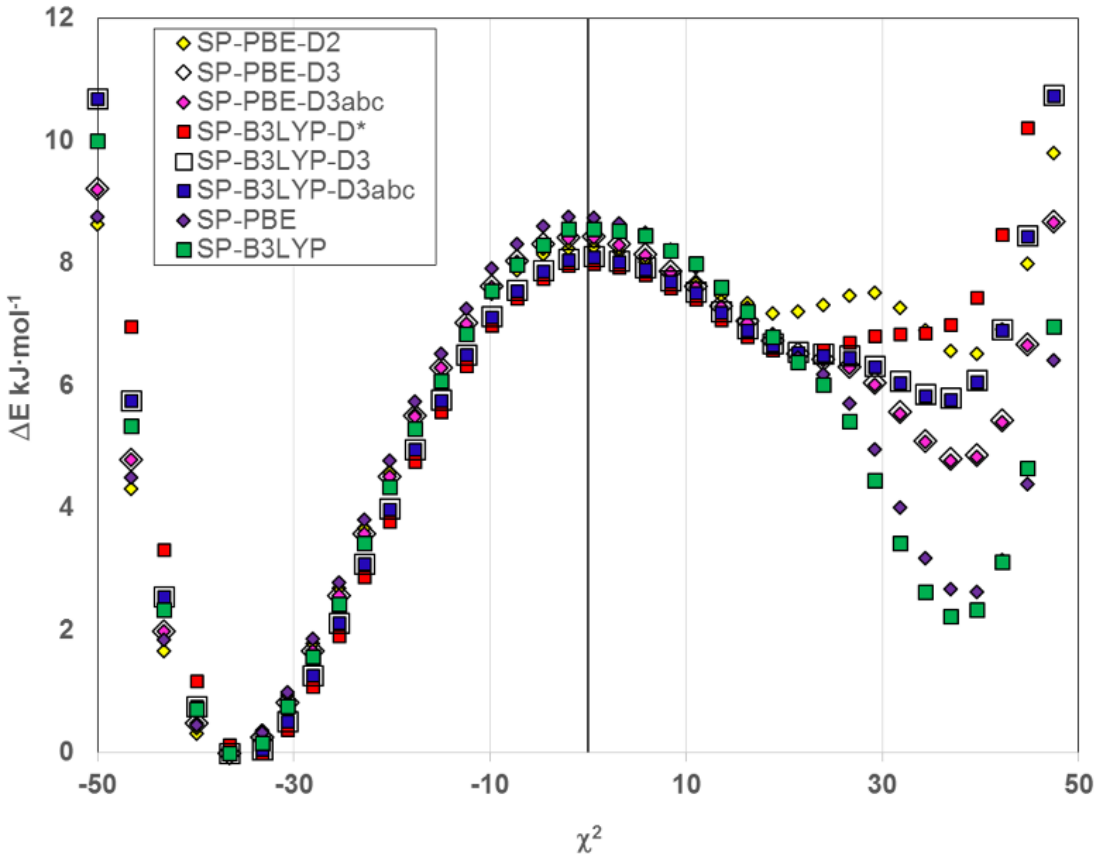


Figure 10. DDD→UDD reaction on the χ^2 variable at different level of theory (D pucker for $\chi^2<0$, U pucker for $\chi^2>0$). On the top: both geometry optimization and energy estimation are at same level

of theory. On the bottom: DFT-D//HF-3c-027 methods, with DFT≡PBE and B3LYP, and D≡D2 (D*), D3 and $D3^{ABC}$.

In all cases, **the results for the DFT-D//DFT-D approach is fully consistent with the DFT-D//HF-3c-027 one (SP-DFT-D method)**, see Figure 10. For this reason, in the discussion we refer only to one case, for instance the SP-DFT-D one. Focusing on the energy estimations at the B3LYP, B3LYP-D*, B3LYP-D3 and B3LYP-$D3^{ABC}$ levels, we can see that the dispersion free B3LYP energies exhibit a very deep minimum for the UDD conformer with a ΔE (E_{UDD}-E_{DDD}) of 2.2 kJ·mol^{-1}. **The inclusion of dispersion contribution to the energy, clearly destabilizes the UDD conformation**, filling the well of the pure electronic calculation and leading to a rather flat energy profile in the region of the UDD state. The ΔE is 6.5 kJ·mol^{-1} and 5.8 kJ·mol^{-1} for the SP-B3LYP-D* and SP-B3LYP-D3 methods, respectively. **The contribution of the ABC correction to the D3 energy is negligible**. The destabilization effect due to the dispersion energy contribution on the U pucker is unrelated to the periodicity of the model, also occurring in the molecular case.

The SP-B3YLP-D* predicts a minimum in χ^2 = ~21° with P = 50° while the SP-B3LYP-D3 methods gives χ^2 = ~37° with P = 13° which are quite different (P is the phase angle, see Figure 7). Interestingly, the B3LYP calculation confirms the position of the UDD SP-B3LYP-D3 minimum. From high resolution X-ray structure analysis on the Protein DataBank,[15,16] we expected two minima on the PES, with values of P for U Pro of 12° and a χ^2 of ~37°. To investigate on this point, also the PBE functional coupled with the D2 dispersion scheme is tested. It computes the UDD region of the PES either with a minimum at χ^2= ~20° or with two almost degenerate minima, see Figure 10 (yellow rhombus). Overall, **the D3 scheme excels in accounting for dispersion interactions in PPII polymers compared to D2 and D* schemes**.

Moreover, we have investigated the effect of the chosen DFT functional on the energy. **The SP-B3LYP-$D3^{ABC}$ and SP-PBE-$D3^{ABC}$ 1D-PES exhibit similar energy profiles**, see Figure 10. The TS and minima occur at the same position, and the energetic only differs by no more than 1 kJ·mol^{-1}. Interestingly, for the **HF-3c and HF-3c-027 method the UDD minimum do not exists**, see Figure 10.

Then I extended the analysis to all the puckering state of the PPII polymer, by computing the energy path, e.g. 1D PES on χ^2, **from DDD towards UUU conformer** with the SP-PBE-$D3^{ABC}$ and SP-B3LYP-$D3^{ABC}$ method. Moreover, full optimization at the all minima positions with the full DFT-D and HF-3c methods is performed and the results are gathered in Figure 11. Our results demonstrate that by **increasing the U Pro content in the polymer chain, the overall stability decreases**, Figure 11. The only exception is the UUU state, which is slightly more stable than the UUD state at the PBE-$D3^{ABC}$ level only. **The trend in the relative stabilities is in agreement between single point and full geometry optimization calculations, with an energy up-shift within 2 kJ·mol^{-1}**. In contrast with the good results for SP-DFT-D methods, pure HF-3c methods performs poorly. Indeed, the UDD conformers is predicted as unstable, and the UUD and UUU states are at least 10 kJ·mol^{-1} less stable with respect to the DFT-D approach. Even though HF-3c-027 is unreliable on the energies, we found a **very good agreement between SP-B3LYP-$D3^{ABC}$ and B3LYP-$D3^{ABC}$ geometries** for all considered PPII conformers. In particular, the Pro stable puckering states are fully consistent between the full DFT-D optimization and SP methods, and largely agree with the experimental results on Pro ring pucker in biological systems.[17]

As for the main chain torsional angles, the results show a progressive reduction of the ϕ from -77° to -56° and of the ψ angle from 160° to 132° moving from DDD to UUU state, respectively. The ψ trend parallels the results for the B3LYP/6-31G* relaxed Pro_6 geometries, as reported in Ref. [18]. In the molecule, the reported average ψ values are 153° and 126° for the DDD and UUU cases, respectively. Conversely, the ϕ angle for the B3LYP relaxed hexamer moves from -72° to -66° on average. This trend is slightly different with respect to our results and this may be due to the periodic nature of our model. During the DDD→UUU transition ω oscillates of ±5° around 174°, which is close to the ideal *trans* peptide bond of 180°.

Interestingly, the length of the unit cell shortens as the Pro ring puckers in the U fashion. The B3LYP-D3^{ABC} method computes a length of the unit cell of 9.12 Å, 8.96 Å, 8.77 Å and 8.61 Å for the DDD, UDD, UUD and UUU conformers, respectively. The PBE-D3^{ABC} gives similar trend with unit cell length slightly longer, e.g. 9.16 Å, 8.98 Å, 8.79 Å and 8.62 Å, respectively. At room temperature, the all DOWN state is the most populated one, for both B3LYP-D3^{ABC} and PBE-D3^{ABC} methods with a Boltzmann distribution of the 80% and 60%, respectively. If we keep a Pro ring fixed in the U puckering, the population of the UDD, UUD and UUU states at RT will be almost equal for PBE-D3^{ABC}. Moving from the UDD to the UUU states will result in almost no energy change and a reduction of the unit cell length of ~4%. Therefore, **the above data shows that the U pucker of a Pro in PPII system enhances the flexibility of the polymer chain at RT**.

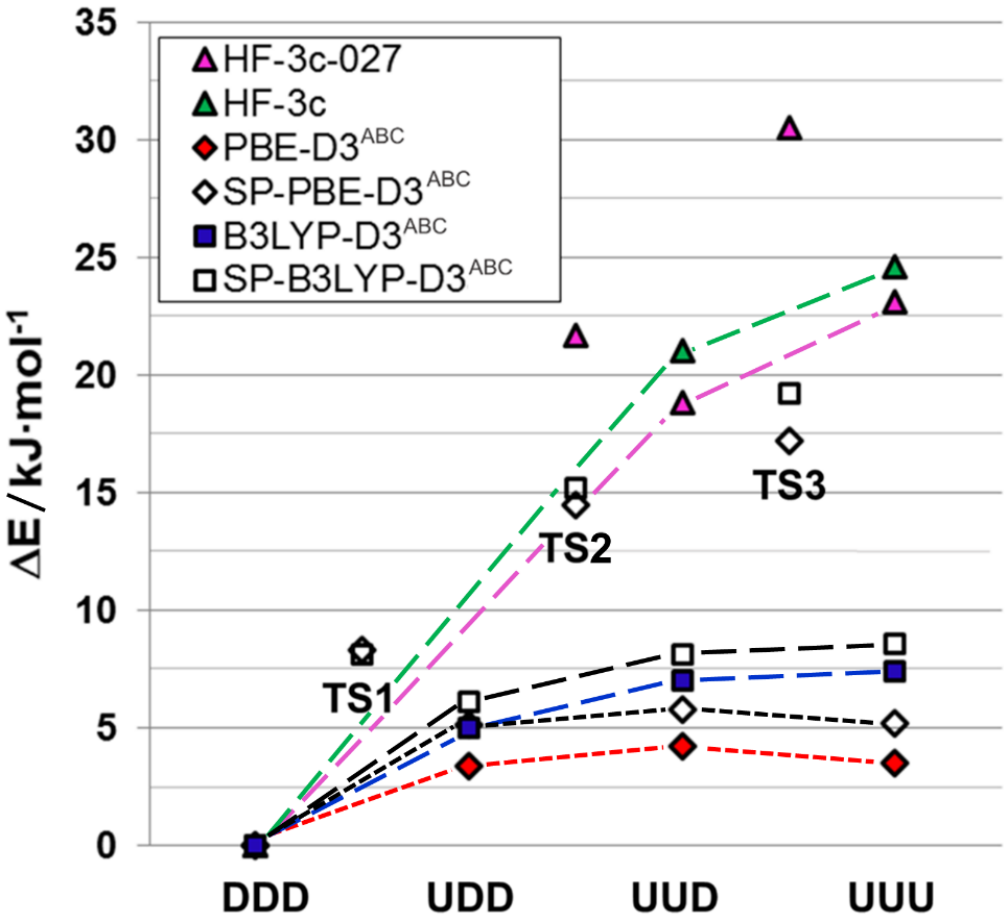


Figure 11. Energy ranking for the PPII polymer conformations. Lines are included to guide the reader eyes only. Transition state energies (TS) for all the process are reported. The UDD conformer is unstable at the HF-3c and HF-3c-027 levels (missing symbol).

3.2 Role of Pro Pucker in Bones

In this section of the thesis, we focus on a **specific feature** of the collagen-single-strand/Hap interface. Indeed, we are interested in evaluating **the effect of the side chain of Pro on the COL/Hap interface**. The relative small size of the adopted models allows us to employ a standard DFT approach, whose details are given later in the text.

As for the adopted models, we used the {**010} P-rich Hap slab (2D periodic) as Hap nano-platelet model** and **PPII polymer as model of a single strand of collagen**. The Hap model have been extensively studied in previous works in the group in which I am currently working, see Ref. [19,20] PPII is a good collagen model due to the high content of imino-acids in collagens and the geometry of each single strand, see Chapter 1. The use of a polymeric system as collagen model allows us to get rid of the terminus COOH-NH_2/surface interaction, which is usually the leading force in the adsorption process of small molecular peptides onto surfaces, blurring the real interaction between the core peptide and the mineral surface atoms.[21] This is even more so because of the very long collagen axial length: therefore, simulating the collagen with short peptide, i.e. molecular Pro tripeptide, will emphasize the role of terminal groups, with misleading results. We aligned the polymer along the *b* Hap vector. Even though in bones the experimental collagen alignment is along the *a* Hap vector (parallel to the O-H channel), recent AFM and MD simulations indicated that the preferred one has an inclination of 72.5° with respect to the *a* vector on the {010} face of Hap,[22] in agreement with our alignment.

To unravel the effect of Pro side chain on the adsorption process of PPII on Hap, we computed **static adsorption geometries** of all side chain conformers and two polymer alignments along *b* Hap vector. We calculated the **energy of the process** and to include thermal and entropic contributions in our simulations, we investigated the **dynamic evolution of the polymer when free and in interaction with the Hap** surface, by means of **AIMD** simulation. For all simulations, the PBE functional with a TZP basis set and the D2 dispersion scheme is employed. **Both static and dynamic calculations demonstrate that the surface has a fundamental role in guiding the formation of specific PPII conformers.**

The Hap (010) P-rich surface exposes Ca cations embedded in a matrix of PO_4 anions, while OH anions are hidden within the slab playing no role in adsorption. The electrostatic potential rendered over the surface of constant electron density (see Figure 12, right) shows well defined regions of relatively high positive/negative values which are active in the adsorption of biomolecules. The corresponding electrostatic potential of the PPII (Figure 12, left) reveals only important deviation from neutrality in the Proximity of the C=O groups, in which negative (red coloured) basins of potential occur. We have therefore docked the PPII polymer towards the Hap (010) P-rich surface following the electrostatic complementarity principle, *i.e.* by maximizing the match between the blue/red zones of the respective maps. Considering that the Ca1 is the most active Ca ion on the surface, we oriented the PPII C=O group toward it. The alignment of the PPII polymer with the Hap *b* vector can occur in two different orientations both related by a rotation of 180° around the z-axis perpendicular to the Hap surface. To simplify the discussion, we will focus only on the most stable orientation, as shown in Figure 13. With the polymer oriented in this way, the Pro1 carbonylic group points toward the Ca1 surface atom with the Pro1 pyrrolidine ring lying parallel to the Hap surface. In Pro2 and Pro3 residues, the C=O group points away from the surface with the pyrrolidine ring lying parallel and normal to the Hap surface, respectively.

To fully explore the role of the Pro side chain in the PPII/Hap adduct, we investigated the interaction with the surface of all PPII conformers, e.g. DDD, UDD, UUD and UUU. *The reader would find useful the information on the previous paragraph for a deeper understating of PPII conformers*. For UDD and UUD, there are three different ways to place the PPII polymer at the surface, as a function of the position of the triplet sequence. In the UDD/UUD cases, the U/D puckered Pro may occur in Pro1, Pro2 or Pro3 aminoacids, see Figure 13. We labeled all possible

cases as: *U*DD/Hap-*D*UU/Hap, *D*UD/Hap-*U*DU/Hap and *D*DU/Hap-*U*UD/Hap, respectively. The puckering of the Pro in Pro1 position is always reported in italic. To avoid confusion between the conformers of the free and adsorbed polymer, from now on, we will refer to the latter case following the format XXX/Hap, with X = D or U.

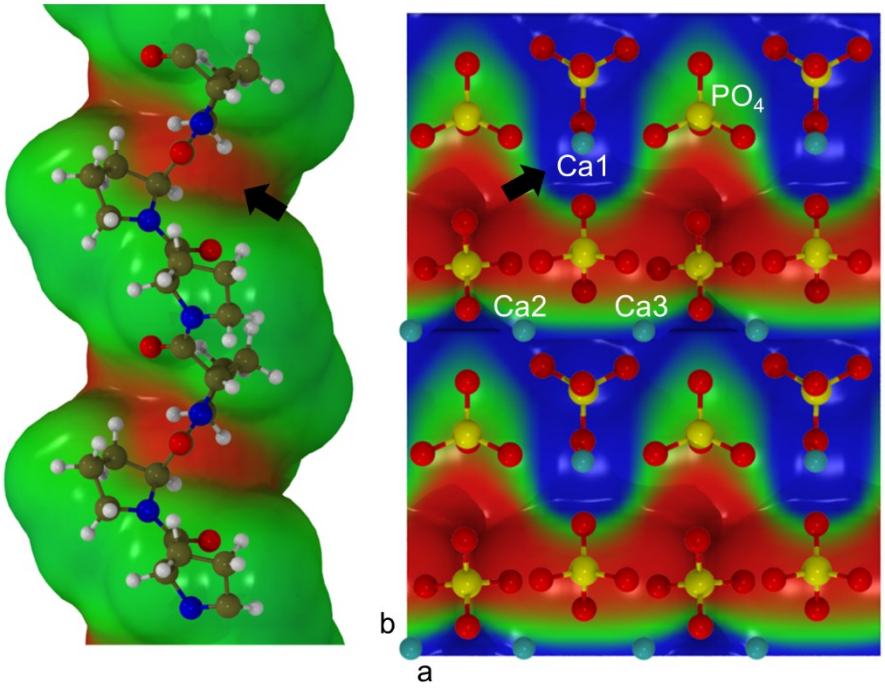


Figure 12. Left: PBE PPII polymer electrostatic potential surface viewed perpendicularly to the main PPII axis. Right: top view along the *z* axis of the PBE electrostatic potential surface of the P-rich (010) Hap surface. Color coding for atoms: carbon (green), nitrogen (blue), calcium (cyan), phosphorus (yellow), oxygen (red) and hydrogen (light grey).

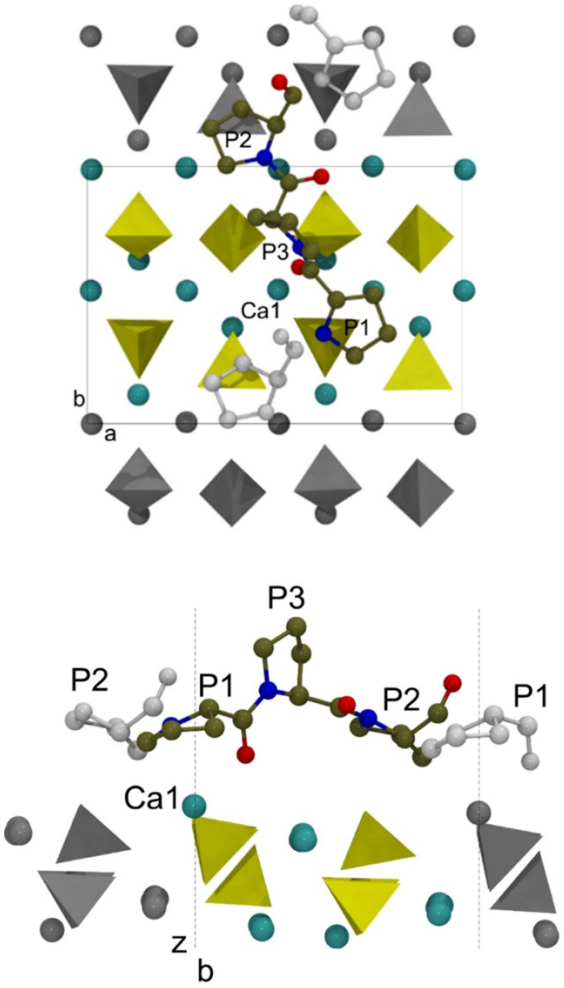


Figure 13. Top: top view along the *z*-axis of the adsorbed *U*DD PPII polymer. Bottom: lateral view of the *U*DD PPII polymer adsorbed on (010) P-rich Hap surface. Color coding for atoms inside the unit cell: carbon (green), nitrogen (blue), calcium (cyan), PO_4 (yellow tetrahedron), oxygen (red). The atoms outside the unit cell are in white (Pros) and silver (Hap). H atoms are hidden for sake of clarity.

The adsorption of the PPII polymer to the Hap surface by means of static PBE-D2 simulations showed a favourable process with binding energies (BE) values in the 63-126 kJ·mol^{-1} range. The adsorption features for all the PPII conformers are reported in Table 1. The geometry optimization procedure confirms a specific C=O···Ca electrostatic interaction, as suggested by the electrostatic potential maps (see Figure 12). Indeed, the C=O···Ca1 distance is in the 2.30-2.45 Å range depending on the PPII side chain conformation. Interestingly, the side chain of Pro3, which points perpendicular to the surface, does not influence the adduct features. This is understandable, considering that, at variance with Pro3, Pro1 and Pro2 are in contact with the surface. Pro3 puckering indirectly affects the BE^C value through the deformation energy, δE (see Chapter 2 for binding energy equations). For D puckering in Pro3, δE_P is smaller and so BE^C larger, see Table 1.

Table 1. Binding energy components (kJ·mol^{-1}) (see Chapter 2 for binding energy equations) and relevant structural data of the PPII/Hap system relaxed at PBE-D2/TZP level (Figure 13). R (Å) and α (degree) are the distance between the oxygen atom of the C=O group and the Ca1 ion and the C=O···Ca1 angle.

PPII	BE	BE*	δE_S	δE_P	BE*^{C}	BE^C	%BSSE	R	α
*D*DD	103.0	189.1	29.8	56.4	148.8	62.6	39.2	2.38	173
Disp	*39.0*	*46.1*	*-0.1*	*7.1*	*46.1*	*39.0*			
*D*DU (*U*DU)*	157.0	254.8	43.1	54.8	212.9	115.0	26.7	2.44	127
Disp	*58.6*	*61.9*	*-0.7*	*4.0*	*61.9*	*58.6*			
*D*UD	144.4	220.6	20.5	55.7	186.5	110.4	23.6	2.33	150
Disp	*48.7*	*50.4*	*-1.0*	*2.7*	*50.4*	*48.7*			
*U*DD	168.2	255.5	42.8	44.5	213.3	126.1	25.1	2.45	126
Disp	*60.4*	*61.9*	*-0.8*	*2.3*	*61.9*	*60.4*			
*D*UU	136.5	220.8	20.7	63.7	186.7	102.3	25.0	2.32	151
Disp	*48.3*	*50.4*	*-1.0*	*3.2*	*50.4*	*48.3*			
*U*DU	158.1	254.3	42.8	53.3	212.4	116.3	26.5	2.44	127
Disp	*59.6*	*61.8*	*-0.7*	*2.9*	*61.8*	*59.6*			
*U*UD	164.8	239.0	36.7	37.5	200.5	126.3	23.4	2.42	131
Disp	*56.5*	*57.6*	*-0.8*	*1.8*	*57.6*	*56.5*			
*U*UU	153.3	235.1	36.3	45.5	197.4	115.6	24.6	2.43	134
Disp	*55.5*	*56.6*	*-0.6*	*1.8*	*56.6*	*55.5*			

**During the geometry optimization, the PPII changes conformation from* D*DU/*Hap *to* U*DU/*Hap.

We computed the relative stabilities of the PPII, as free polymer, with the cell length fixed to that of the Hap. The free UDD, UUD and UUU conformers are 9.7 kJ·mol^{-1}, 17.8 kJ·mol^{-1} and 28.3 kJ·mol^{-1} less stable than the DDD. **The relative stability of the PPII conformers adsorbed at the Hap surface, reveals that the adhesion process at the Hap surface stabilizes conformers which are inaccessible to the free polymer.** On the Hap, the presence of U puckered Pros in Pro1 or Pro2 positions, increases the BE with the surface with respect to the *D*DD/Hap case. This effect is quantified to be ~48 kJ·mol^{-1} at least, as revealed by the comparison of BE*^{C} values for rows *D*DD and *D*UD/*D*UU of Table 1. This, in turn, shows that the *U*DD/Hap, *U*UD/Hap and *U*UU/Hap cases are more stable than *D*DD/Hap, at variance with respect to the free PPII polymer. Figure 14 compares the relative stability of the various conformers for three different situations: i) the PPII as a free polymer; ii) the PPII as a free polymer with the cell parameter fixed at that of the Hap unit cell; iii) the PPII/Hap system. Comparison between case i) and ii) revealed that stretching the unit cell size of the free PPII polymer, as imposed by the adsorption at the Hap surface (from 9.120 Å to 9.538 Å), causes an increase in the instability of the different conformations. Also, while the UDD, UUD and UUU relative stability with respect to DDD are very close to each other, they dramatically differ to each other upon stretching the PPII unit cell. Indeed, the UUU conformer is three times more unstable when stretched than as a free conformer. This is due to the intrinsic contraction of the polymer unit cell when switching from the DDD to the others conformers, as we pointed out in the previous paragraph. Figure 14 also highlights how the adsorption of PPII at the Hap surface brings about a

dramatic change of the relative stability compared to the free (stretched) polymer. **Indeed, the UDD conformation becomes the most stable one at the Hap surface by about 60 kJ mol^{-1}. Energy differences reveals that the population of all other conformations, but the *U*UD case, are almost negligible.** Unfortunately, there are no experimental data to compare with our simulations.

We studied possible correlations between energetic data of Table 1, responsible of the stabilities shown in Figure 14 and geometrical details of both the PPII polymer and the Hap surface. We only find a good correlation between the displacements of the *z* coordinate of the Ca1 ion from the equilibrium position of the free Hap with the deformation energy δE_S of the Hap surface (see Figure 15, left). This is in agreement with the localized character of the PPII/Hap interaction, specifically due to the C=O···Ca1 bond. Figure 15 right, reveals that the deformation energy of the polymer does not correlate with the variation of the PPII torsional angles, dictating the polymer conformation.

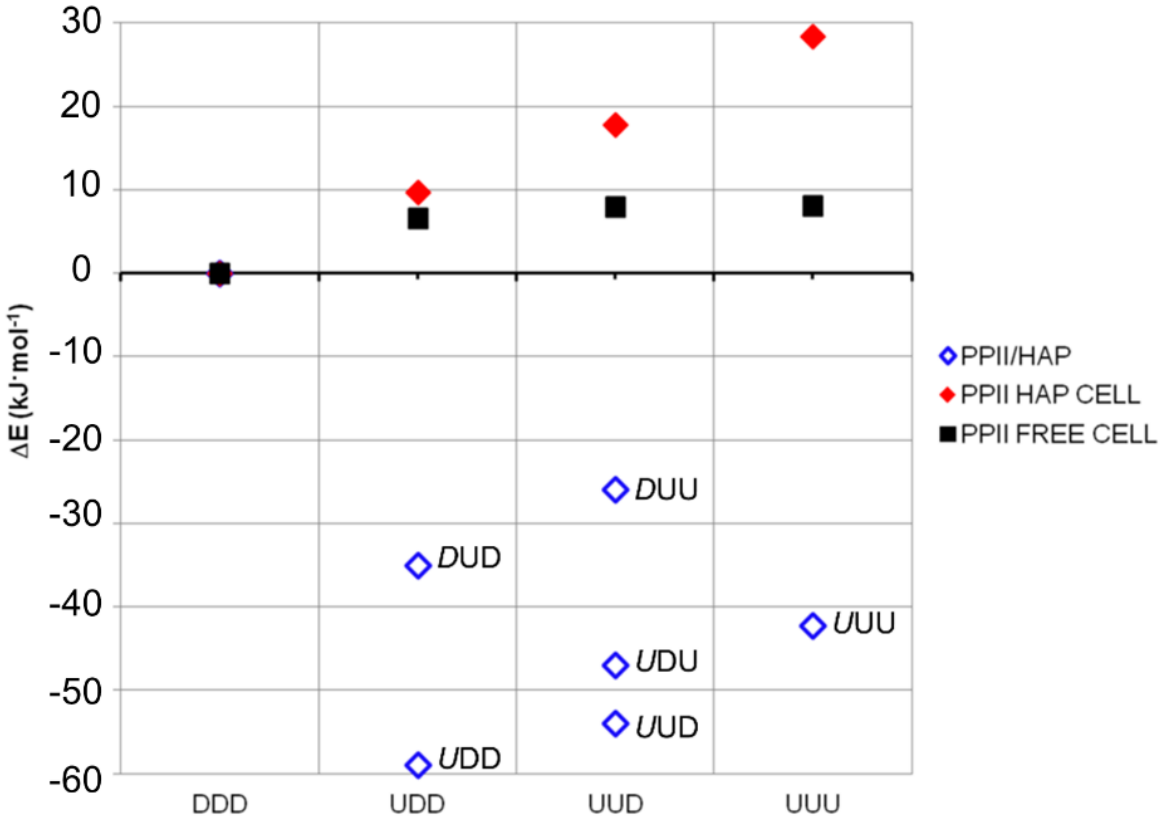


Figure 14. Relative stability with respect to the DDD conformer $\Delta E = E_{XXX} - E_{DDD}$; X is either U or D. Cases are: the free PPII polymer fully relaxed (filled square), the free PPII polymer relaxed with cell length fixed to the Hap *b* cell parameter (filled red diamond), and the PPII/Hap system (empty blue diamond).

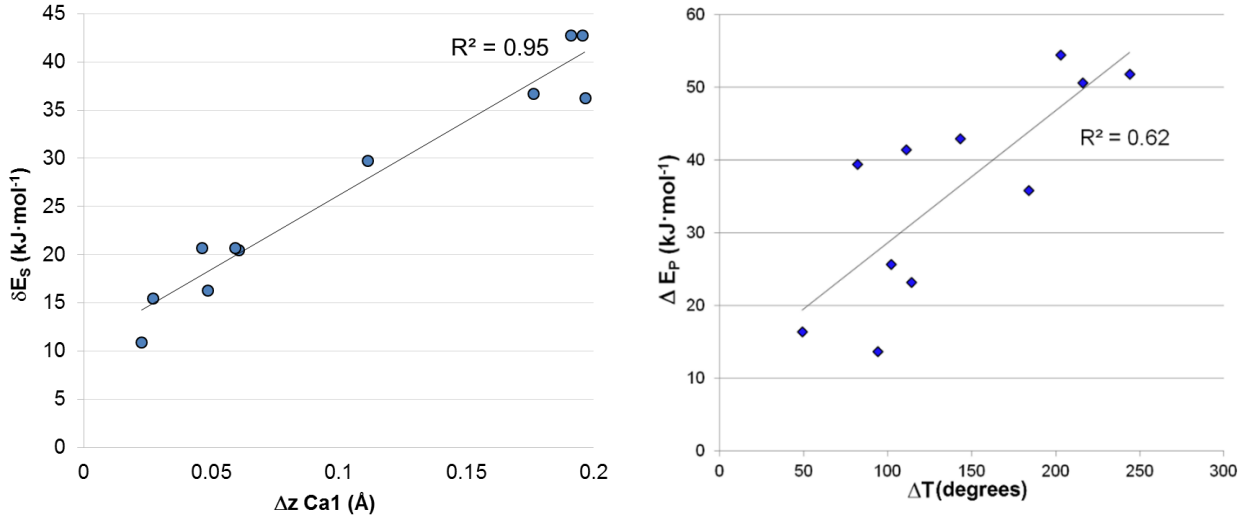


Figure 15. Left: surface deformation energy (δE_S) *vs* the displacement of the z coordinate of the most exposed Ca1 ion (Δz) with respect to the clean surface. Right: polymer deformation energy ΔE_P *vs* ΔT. $\Delta T = |\phi^{ads} - \phi^{r}| + |\omega^{ads} - \omega^{r}| + |\psi^{ads} - \psi^{r}|$, where ϕ^{ads}, ω^{ads} and ψ^{ads} are the back bone dihedral angles of the polymer adsorbed on to the Hap surface and ϕ^{r}, ω^{r} and ψ^{r} those of the relaxed polymer with the cell length fixed at the Hap unit cell *b* parameter, see Ref. [23] for further details.

To study the structural evolution in time of our models, we run AIMD calculations for PPII, both as a free polymer and when adsorbed on the Hap surface for a time evolution window of 22.6 ps. The advantage of dynamic simulations is to include thermal and entropic effects, which are neglected in the static approaches. To start with, we set the free PPII polymer structure in the DDD state and the adsorbed one in the *D*DD/Hap conformation. The DDD state is the most stable one from static calculations.

During the dynamic of the free PPII polymer, two of the four available conformational states, DDD and UDD are visited. A statistical analysis of the trajectories for the whole production time revealed that DDD and UDD minima are populated for 92% and 8% of the simulation time, respectively (see Figure 16). This result is in good agreement with the prediction based on static calculations, see Figure 14. Indeed, from a Boltzmann analysis of the static PBE-D2/TZP relative stabilities of PPII at T=300 K the occupancy of the DDD, UDD, UUD and UUU conformers is 87%, 6%, 4% and 3%, respectively.

When the PPII polymer interacts with the Hap surface, the conformation evolves from the *D*DD/Hap to the *U*DD/Hap state, see Figure 16. The D→U ring flipping occurs immediately during the equilibration time (first 2.4 ps). Interestingly, the dynamic of the event is in agreement with the estimated half-life time, $\tau_{1/2}$, for a DDD→UDD ring flip for the free polymer ($\tau_{1/2}$= 2.8 ps), as previously computed in Ref.[13] After the initial flip, the *U*DD/Hap state remains stable for the whole production time, e.g. 10 ps, confirming the result of the static simulations. As already pointed out, the static simulations indicate the *U*DD/Hap and *U*UD/Hap as the lowest energy states for the PPII/Hap system. Therefore, we may expect the dynamic simulation to evolve from *U*DD/Hap to *U*UD/Hap providing a long enough production time. During the dynamics, the system is anchored onto the surface stably in one state, as shown in Figure 17. Indeed, the C=O···Ca1 distances indicate an almost pure one-modal distribution. Moreover, the angle between the C=O···Ca1 oscillates around the average 120°, in agreement with the oxygen sp^2 hybridization, dictating the final geometry (Figure 17).

As expected, the surface adsorption induces a geometrical distortion of the polymer torsion bonds to maximize the contact with the Hap surface, which can be summarized as follows:

i) for the backbone dihedrals, we computed the Ramachandran plot of the free and adsorbed PPII polymer, as shown in Figure 18. It is worth noting that Pro3, which is the residue

least involved with the Hap surface, has the largest deviation on the (ϕ,ψ) values with respect to the free PPII system. Indeed, the Pro1 and Pro2 regions of existence largely overlap with those of the free PPII one (yellow area on the graph), including also the ideal value for PPII conformation (black cross on the graph).

ii) for the side chain puckering, Figure 16 clearly shows the evolution of the χ^2 torsional angle from a one-modal distribution type for the free PPII polymer to a bi-modal one for PPII at Hap.

It is important to point out that the agreement between *U*DD/Hap geometry computed by static and dynamical simulations is remarkable. The dynamical average values of the adsorbed PPII at Hap differ from the static result by less than 1° on average and the C=O···Ca1 distance and angle from static simulations fit well within the computed dynamic oscillations (see Ref. [23] for further details).

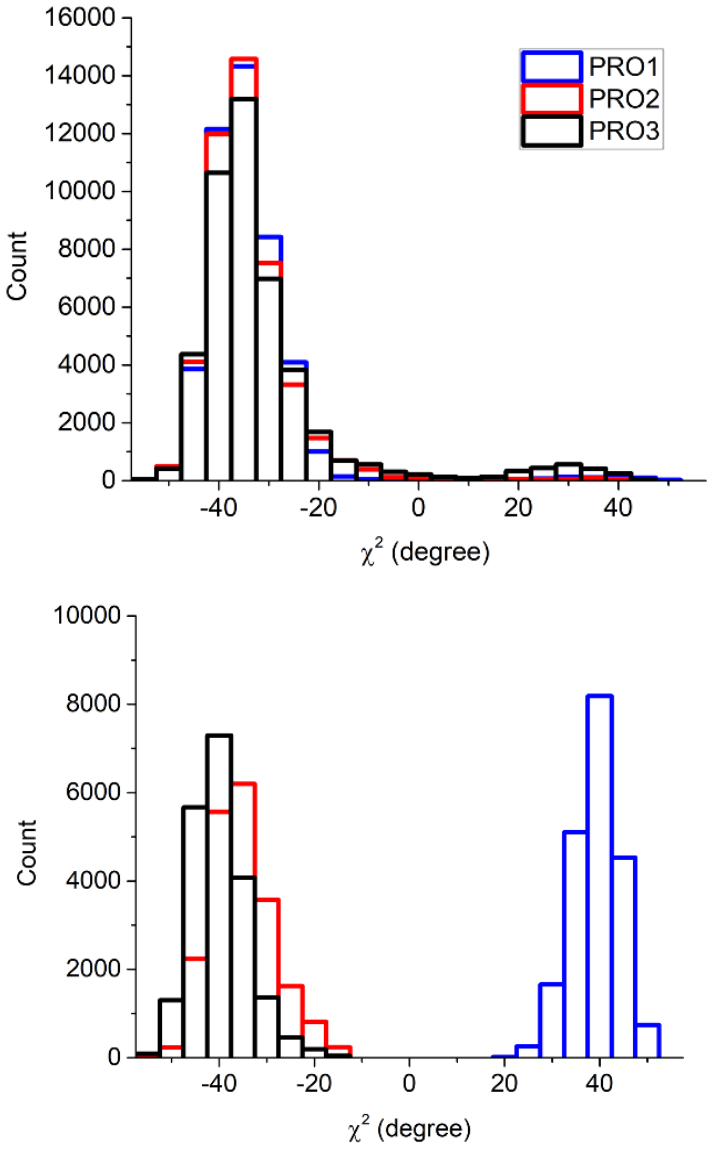


Figure 16. Frequency of occurrence of χ^2 (ring) torsional angle for Pro1, Pro2 and Pro3 during the dynamics for the free PPII polymer (top graph) and when adsorbed at the Hap surface (bottom graph).

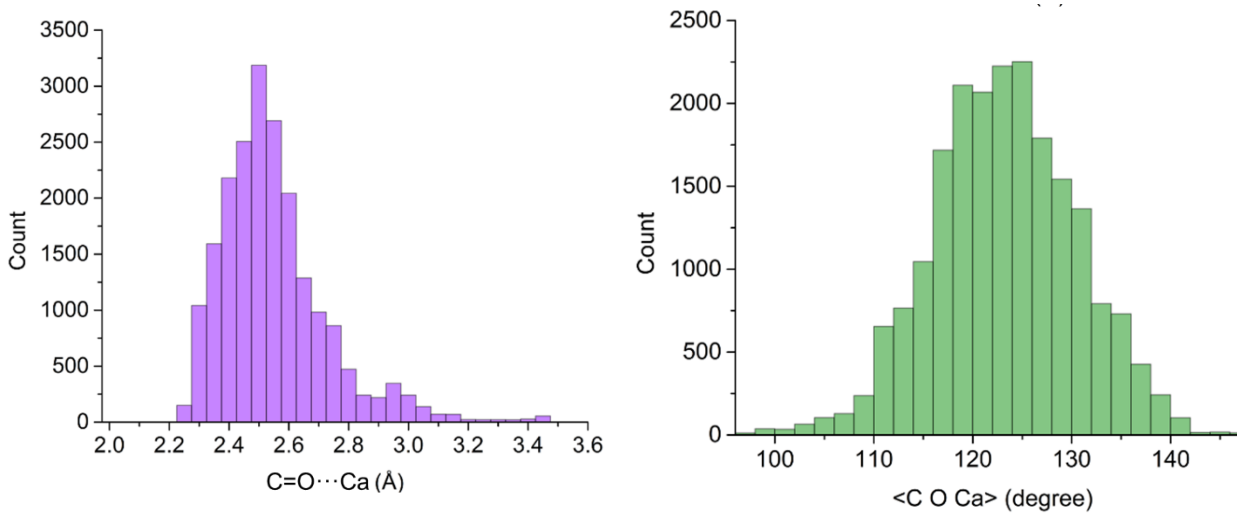


Figure 17. Pro1-Ca frequency count for the C=O···Ca1 distance (left) and angle (right) during the PPII/Hap dynamic calculation.

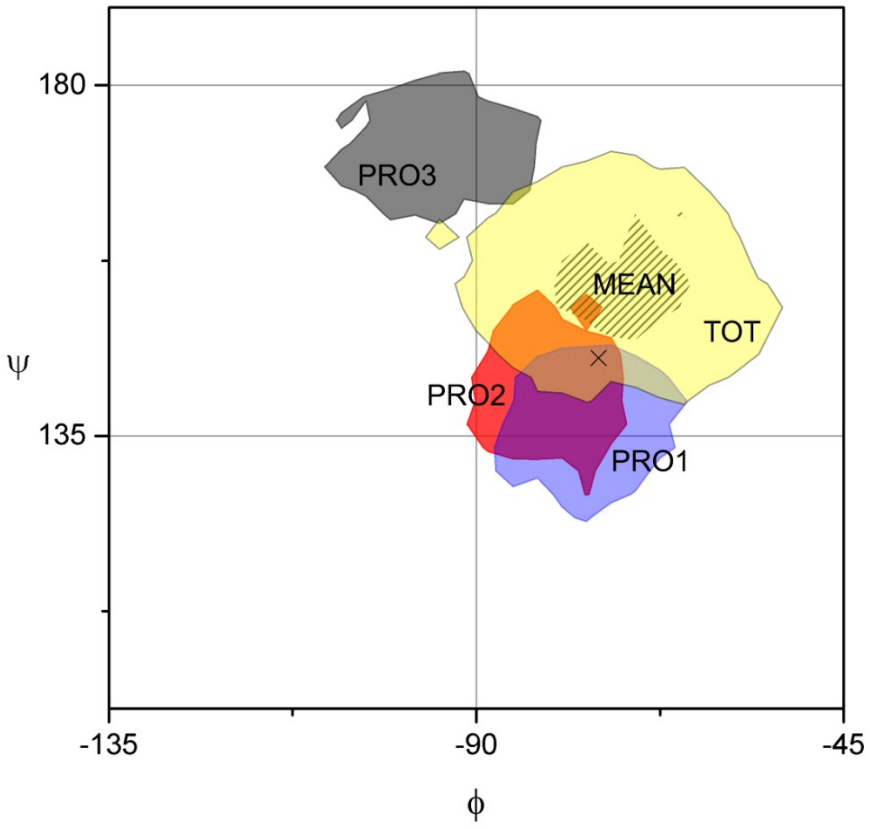


Figure 18. Ramachandran plot (explored ϕ and ψ dihedrals) for the PPII polymer dynamic both free and adsorbed on Hap. Color code for the PPII/Hap dynamic: Pro1 in blue, Pro2 in red, Pro3 in dark grey. For the dynamic of the free PPII, we report the (ϕ,ψ) region visited by all Pros (TOT, in yellow) as well as the mean values (as black lines). The black cross represents the most probable (ϕ,ψ) values for the free PPII, (-75°,145°).[24]

3.3 Roto-Translational Symmetry of Collagen Protein and Aggregates. Few Applications

A starting point for investigating the **driving forces in the collagen protein folding** is hereafter reported. The process involves the association of three independent strands into a superstructure and it takes place in water solvent, see Chapter 1 for further details. In the presented approach, we make three approximations:

i. The process happens in vacuum condition.
ii. The free single strand geometry is forced to be in PPII conformation.
iii. The collagen protein is simplified with a homo-trimeric 1D periodic system with only one type of triplet as aminoacidic composition.

The simulation of the collagen protein is possible thanks to its highly symmetrical nature, and the capability of the CRYSTAL code to exploit it efficiently. In Figure 5-6 of Chapter 1 I have shown the diffraction pattern of a collagen protein. Employing the roto-translational symmetry, as encoded in the CRYSTAL code, it is possible to reproduce those patterns, creating a one dimension infinite collagen-like polymer, starting from one aminoacidic triplet only. So, all the collagen models hereafter presented are triple helix, 1D infinite, and have seven triplets per polymeric unit cell, but only one belongs to the asymmetric unit cell, thanks to the symmetry. This allows us to run truly fast simulation with the only drawback of a limited number of the degrees of freedom, which are restricted to the atoms in the asymmetric unit cell, see Figure 19.

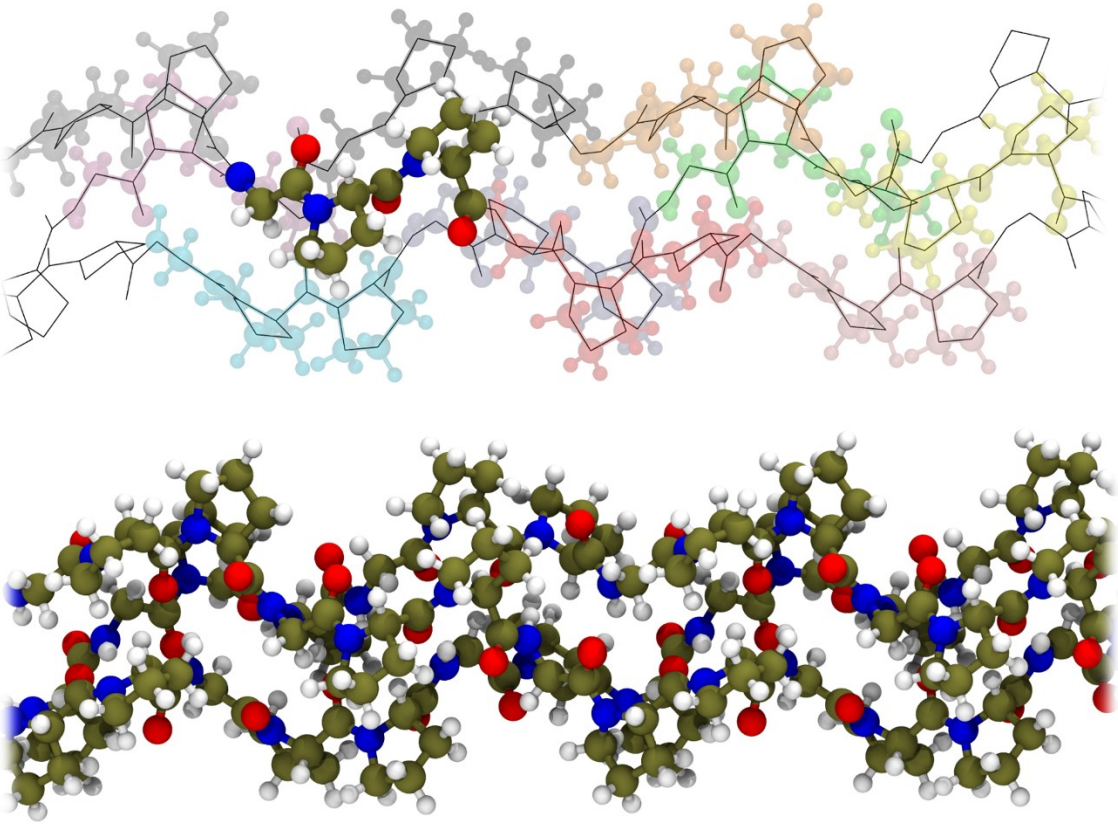


Figure 19. Polymeric collagen protein model. Top: The asymmetric unit (Gly-Pro-Pro triplet) is reported in full colours. The replicas due to the symmetry operators are coloured in transparency. Bottom: Resulting model with full atoms details.

I took in consideration two different compositions, e.g. GPP and GPO. In this way, we try to address **the role of the OH substitution**, in the wrapping reaction which makes three single strands into a triple collagen helix. From experiments, we known that the Pro hydroxylation stabilizes collagen protein. We remind to the reader, that the PPII geometry imposes to have 3 residues per unit cell (each residue rotates of 120° along the polymer axes), thus **both the PPII helix and collagen triple helix (HEL) have one residue triplet in the asymmetric only**. For both aminoacidic compositions, we took into consideration all the possible polymer conformers. In the GPP case, due to the imposition of the helicity to the main chain of the protein, the only possible conformers differ for the puckering of the Pro rings, Figure 20. The complexity arises in the GPO case due to the two possible orientations of the OH group, see Figure 21. For both PPII and HEL geometries, we took into account four and eight conformations for the GPP and GPO cases, respectively, see later in the text.

The case of Gly-Pro-Pro (GPP)

As first step, we focus on the simplest composition, e.g. the Gly-Pro-Pro (GPP), and the simplest geometry, e.g. **PPII**, of our models. The asymmetry arising from the introduction of the Gly residue induces the other two positions in the triplet to be not equivalent. Therefore, the position within the triplet are named after the current definition of collagen triplet, see Chapter 1, and Figure 20, e.g. X and Y. **Position X and Y** can be occupied by (see Figure 20 and Table 2):

- both Pros in the D state (**DD or 1**)
- both Pros in the U state (**UU or 4**),
- a U Pro and a D Pro, respectively (**UD or 3**)
- a D Pro and a U Pro, respectively (**DU or 2**)

The **four stable polymer conformers,** for both single and triple helical systems, undergo to geometry optimization of both cell length and internal positions with the **B3LYP/TZP-D3^{ABC}** method (Table 2 and Figure 20). We include also the results for the Pro-Pro-Pro compositions in a PPII geometry (PPP, see previous paragraphs) at the same level of theory as a comparison (see Table 3).

In agreement with the PPP case, in GPP, the state with all the Pros in a DOWN (D) puckering is the most stable one (conformer **1**, see Table 2). From **1**, the side chain flipping from D to UP (U) costs ~6.5 kJ·mol^{-1}. Conversely, when a Pro is already in the U state within the polymer the cost of the flipping of the other one is reduced to ~1.5 kJ·mol^{-1} (**2→4** and **3→4**). These results are in total agreement with the previous finding on the PPP polymer, therefore **the Gly introduction does not affect the ranking stability of the Pro side conformers within the polymer**.

Conversely, the Pro→Gly increases the length of the polymer for the all D case (9.12 Å for the DDD case and 9.33 Å for the GPP case, respectively, all in the D state). When both the Pros are U, conversely, both PPP and GPP have the same length, 8.75 Å and 8.77 Å, respectively.

Regarding the dihedrals, ϕ = -72.5° ψ = 157.2° for a PPP in the DDD state. Due to the Gly presence, in the DD state, the ϕ splits in -79.3° and -68.4° for Pro in X and Y respectively, both of them having the same ψ = 161.2°. In the UU state, ϕ converges to ~ -59° for Pro in X and Y, while ψ splits to ~131° and ~144°. It seems that the U puckering makes the ring stiffer as emerged by the PES analysis. Indeed, the U basin looked less shallow that then the D one. Within a protein database analysis, the standard deviation on the ϕ is 9° and 7°+ for D and U puckered prolines.

Moving from an isolated PPII geometry to triple helical geometry, the conformers relative stabilities change. **Interestingly, the most stable triplet state in HEL becomes DU, in agreement with the propensity model,** see Chapter 1.

From X-ray structures on triple helical GPP peptides (wet by water), Pros in Y are 100% of the cases in U state and Pros in (X) are most of the times in D state.[25,26] In our GPP HEL models, the D→U interconversion in Y costs only ~1.5 kJ·mol^{-1}. Conversely, the D→U interconversion in X costs

~6 kJ·mol^{-1}, see Table 4. So, **our modelling better reproduce the puckering preference of position X rather than for Y, in agreement with the experimental findings.**

Table 2. Geometry and energy of the **GPP models in a PPII helical fashion.** The angles are expressed in degrees and cell axis lengths (cell) in Å whilst the energies in kJ·mol^{-1} per triplet. The name (n) is associated to a symbol (syb) which summarizes the pucker of Pro in X and Pro in Y (XY). The ideal PPII (ϕ, ψ, ω) are (-75°, 145°, 180°).

		Gly			Pro (X)				Pro (Y)							
n	syb	ϕ	ψ	ω	ϕ	ψ	ω	χ^2	ϕ	ψ	ω	χ^2	cell	ΔE_{tot}	ΔE_{el}	ΔE_{disp}
1	DD	-83.3	163.1	173.7	-79.3	161.2	162.9	-35.5	-68.4	161.3	159.8	-35.9	9.33	0.00	0.00	0.00
2	DU	-84.1	162.9	175.1	-78.2	156.1	160.2	-35.4	-61.8	158.5	160.4	31.5	9.27	5.43	3.06	2.37
3	UD	-81.3	159.6	170.6	-75.4	162.2	163.2	21.3	-68.7	162.4	159.6	-35.8	9.24	7.60	6.56	1.04
4	UU	-63.9	131.2	173.4	-59.0	130.9	175.7	38.5	-58.8	144.0	175.7	38.2	8.75	8.48	4.55	3.93

Table 3. Geometry details of the PPP conformers at the B3LYP-D3^{ABC}. Angles in degrees and cell axis length in Å. Label of Pro pucker after previous paragraph and Ref.[13].

name	ϕ	ψ	ω	χ^2	cell axis
DDD	-72.5	157.2	166.6	-35.1	9.12
UDD	-60.2	146.7	176.0	36.9	8.96
U**D**D	-73.8	154.6	173.8	-36.7	
UD**D**	-70.0	138.1	167.1	-35.4	
UUD	-57.3	131.3	174.2	38.1	8.77
U**U**D	-61.5	148.3	179.0	37.8	
UU**D**	-69.6	137.9	170.3	-36.1	
UUU	-57.5	133.3	174.5	38.5	8.61

Table 4. Geometry and energy of the **GPP triple helical models**. The angles are expressed in degrees, the hydrogen bonds (Hb) and cell axis length (cell) in Å whilst the energies in kJ·mol^{-1} per triplet. The name is associated to a symbol (syb) which summarizes the pucker of Pro in X and Y (XY). The experimental SD is reported within brackets.

		Gly			Pro (X)				Pro (Y)									
n	syb	ϕ	ψ	ω	ϕ	ψ	ω	χ^2	ϕ	ψ	ω	χ^2	Hb	α	cell	ΔE_{tot}	ΔE_{el}	ΔE_{disp}
1	DU	-70.6	177.9	175.8	-69.9	166.2	165.4	-36.5	-56.1	156.7	167.6	33.2	2.074		19.89	0.00	0.00	0.00
2	UU	-74.0	180.1	175.3	-65.4	164.4	161.8	30.8	-56.7	157.3	170.3	32.9	2.069		19.87	6.88	2.17	4.71
3	UD	-76.3	177.7	174.8	-61.5	158.5	165.3	33.1	-61.7	157.9	179.2	-37.3	1.976		19.96	6.42	0.84	5.58
4	DD	-70.4	174.6	176.2	-69.0	163.6	168.6	-35.9	-57.7	155.2	170.8	-37.2	2.018		19.97	1.57	-1.95	3.52
GPP 7_2 helix[27]		-70.2	175.4	178.2	-75.5	152.0	181.8		-62.6	147.2	183.2							
GPP exp.[26]		-71.7 (3.7)	175.9 (3.1)	179.7 (2.0)	-74.5 (2.9)	164.3 (4.1)	176.0 (2.5)		-60.1 (3.6)	152.4 (2.6)	175.4 (3.4)							

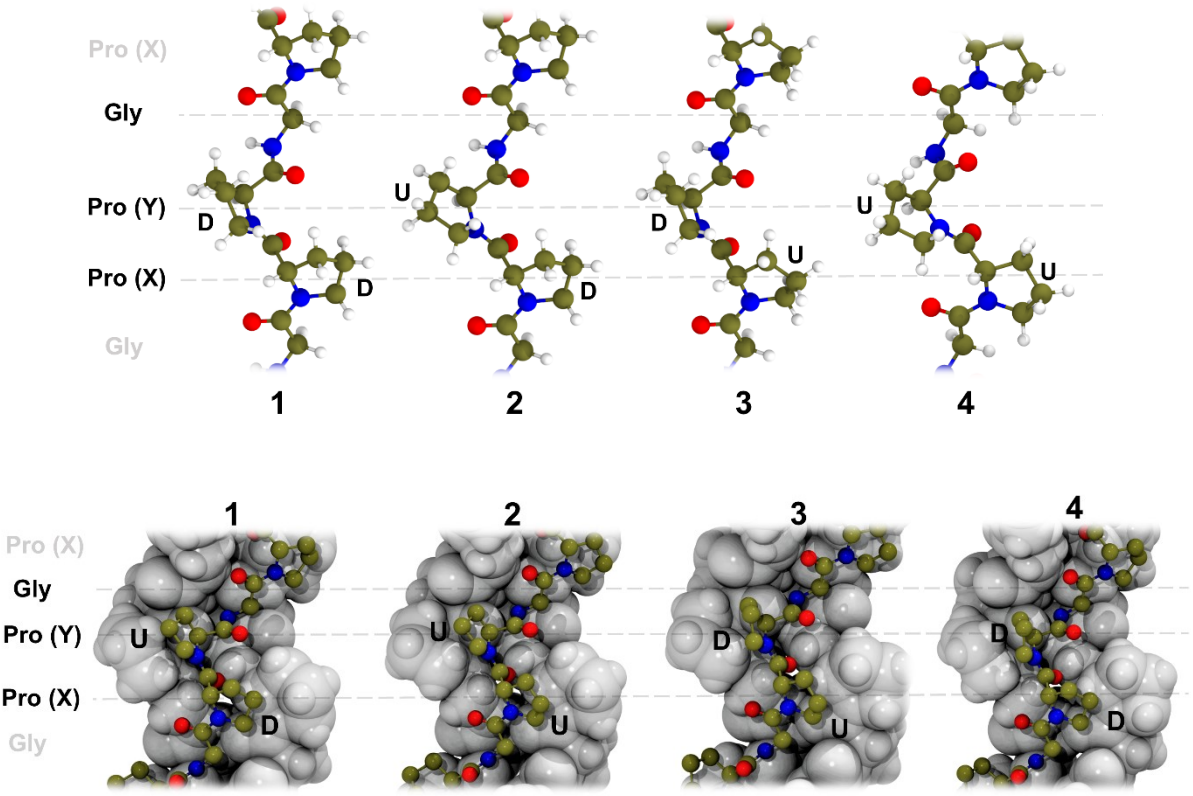


Figure 20. Top: single collagen strand with GPP composition in a PPII geometry. Bottom: triple helical collagen with GPP. The aminoacidic position in the collagen triplet are highlighted.

The case of Gly-Pro-Hyp (GPO)

The presence of the OH group increases the complexity of the systems, thus increasing the number of stable conformers from four to eight in both PPII and HEL structures. Indeed, for each D/U combination in X and Y, the OH group can have two different orientation, which are defined in Figure 21 with respect to the C-N peptide group direction.

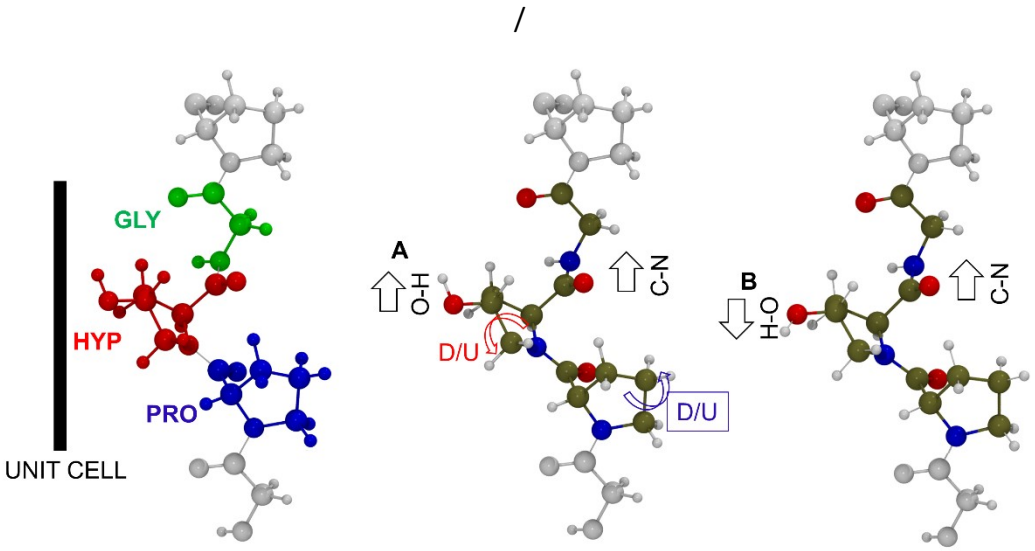


Figure 21. Description of the degrees of freedom of the Gly-Pro-Hyp PPII polymer. The orientation of the OH group is compared with the direction of the peptide bond (C-N bond).

As for the former case, all the systems were fully optimized employing the **B3LYP/TZP-D3^{ABC}** method. Starting the discussion with the simpler PPII geometry, all the conformers are graphically reported in Figure 22 and the geometrical features gathered in Table 5. **The stability ranking in this case is ruled by an intramolecular H-bond between the hydroxyl group of Hyp and the carboxyl group of Gly, see Figure 22**. When the Hyp in Y position is in the U state, the H-bond takes place for the B orientation of the OH group, involving the Gly not directly bonded to Hyp (conformer **2** and **4**, Figure 22). Conversely, when the Hyp in Y position is in the D state, the H-bond takes place for the A orientation of the OH group, and it involves the Gly directly bonded to Hyp (conformer **5** and **7**, Figure 22). In the former case, the H-bond is much shorter and it guides the stability ranking of the GPO PPII conformers. Indeed, **2** and **4** are the most stable conformers, differing only of 4.44 kJ·mol^{-1} in favour of conformer **2**. This small energy difference from **2** to **4** is due to the lengthening of the H-bond. This is directly correlated to the different pucker of Pro in position X. In state **2**, it is U while in **4** it is D. Even if the D state is usually more stable than the U one, this latter leads to a shorter unit cell causing a shortening of the intramolecular H-bond, see Table 5. An estimation of the BE for the intra-molecular H-bond in **2** and **4** comes from the re-arrangement of OH group which takes place in the **1→2** and **3→4** interconversions. Indeed, **1** and **2** conformers have the same puckering of the pyrrolidine ring as **3** and **4** ones. Therefore, if we assume the reorientation of OH as a free process, the H-bond formation must release ~20 kJ·mol^{-1} and ~11 kJ·mol^{-1} for the former and latter case, respectively.

The same rationalization can be applied to the conformers with Hyp in the D state (**5-8**). In this case, there is no intermolecular H-bond, but just a weak long-range electrostatic interaction. Assuming the A/B OH orientations as iso-energetic, we can speculate on the energetic of this weak interaction to be ~2 kJ·mol^{-1} (energy difference between the **5-6** and **7-8** cases). Focusing on the cases without any OH···C=O interaction, e.g. **1**, **3**, **6** and **8**, we can guess the cost of a D→U flipping for a Pro (in X), to be 5-7 kJ·mol^{-1}(**1→3** and **6→8**), and we can guess the cost of a D→U flipping for a

Hyp (in Y), to be 4-6 kJ·mol^{-1}(**1→6** and **3→8**), see Table 5. **Therefore, the puckering propensity seems respected for the Pro residue only. This may be due to the assumption of an OH group which can almost freely rotate around the C-O bond.**

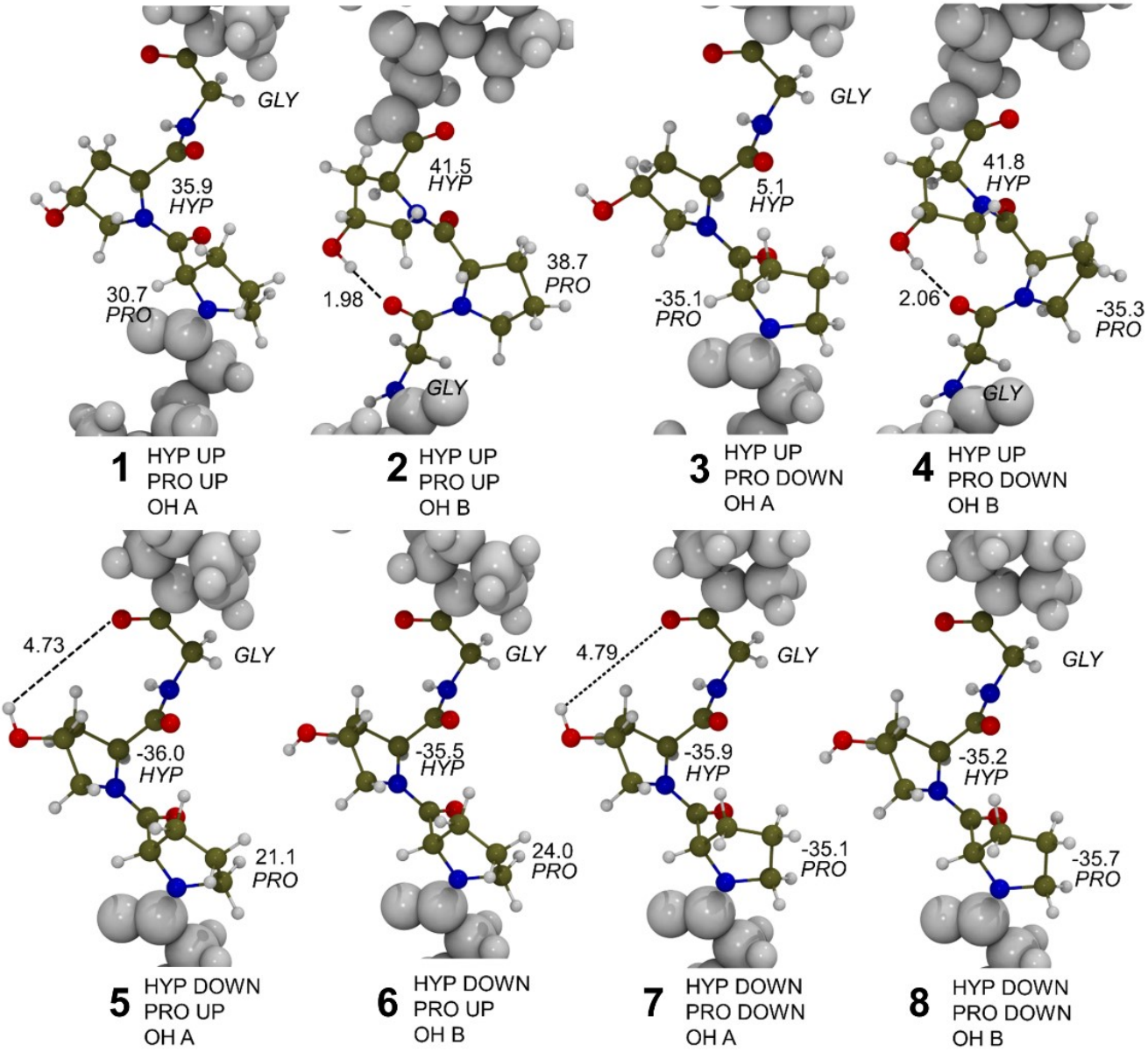


Figure 22. Different conformations for a single collagen strand (GPO) in a PPII geometry.

Table 5. Geometry and energy of the **GPO models in a PPII helical fashion.** The angles are expressed in degrees and cell lengths (cell) in Angstrom whilst the energies in kJ·mol^{-1} per triplet. The name (n) is associated to a symbol (syb) which summarizes the pucker of Pro in X and Pro in Y (XY).

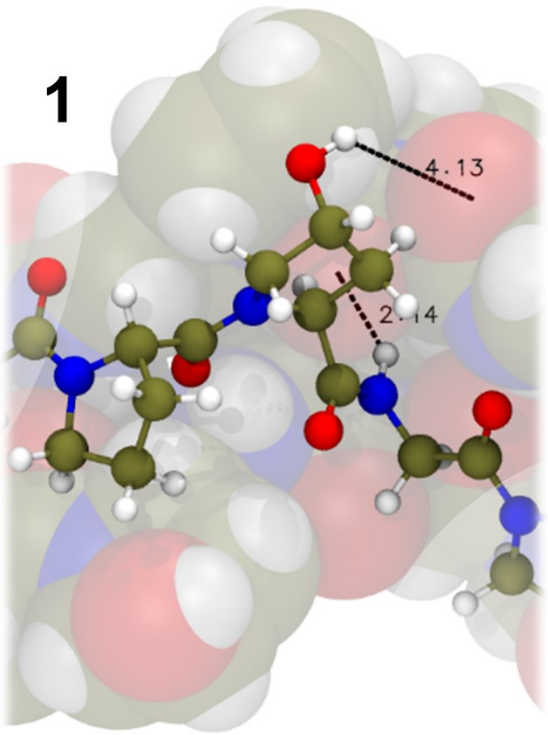
			Gly			Pro (X)				Hyp (Y)						
n	syb	OH	ϕ	ψ	ω	ϕ	ψ	ω	χ^2	ϕ	ψ	ω	χ^2	Hb	cell	ΔE
1	UU	A	-73.3	149.1	172.7	-70.8	151.9	164.2	30.7	-57.5	150.6	161.2	35.9	none	9.01	20.07
2	UU	B	-60.6	125.8	170.6	-57.3	122.1	-172.8	38.7	-70.2	152.9	-175.5	41.5	1.979	8.68	0.00
3	DU	A	-83.2	166.2	171.4	-78.4	162.9	159.8	-35.1	-69.4	164.4	158.7	5.1	none	9.30	15.56
4	DU	B	-65.8	141.2	178.0	-68.7	121.9	-173.6	-35.3	-68.4	148.0	-169.7	41.8	2.056	8.75	4.44
5	UD	A	-79.5	160.6	169.1	-75.3	162.3	162.3	21.1	-68.4	162.9	158.6	-36.0	4.726	9.20	14.31
6	UD	B	-78.8	156.6	171.0	-73.6	160.8	163.9	24.0	-67.3	160.0	159.7	-35.5	none	9.17	16.35
7	DD	A	-81.1	163.7	172.1	-79.3	161.4	162.6	-35.1	-68.3	161.6	158.8	-35.9	4.793	9.29	7.02
8	DD	B	-81.9	161.8	174.2	-78.6	160.4	163.1	-35.7	-67.6	159.6	159.8	-35.2	none	9.29	9.10

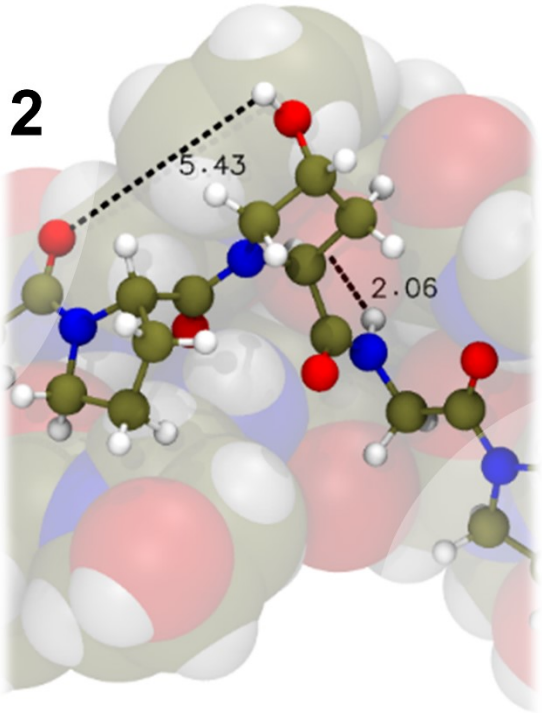
As for the GPP case, also for GPO the DU state is the most stable when the chains are wrapped in a triple helix (HEL), see **1** Table 6. **Again, this is in agreement with the propensity model**, see Chapter 1. Interestingly, in this case the OH group in the **A** orientation, gets involved in weak electrostatic interactions, which seems to induce an internal re-arrangement with consequent stabilization of the helix, see Table 6 and Figure 23. In **1→2** and **3→4** the energy gain for this OH induced reorganization is ~5 kJ·mol^{-1} and for **5→6** and **7→8** of ~2 kJ·mol^{-1}, see Table 6.

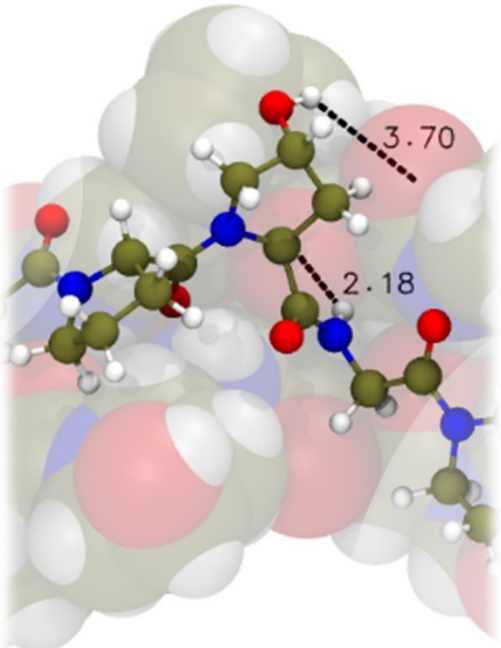
Interestingly, in the most stable cases (**1-4**), the D→U process in X costs less than 2 kJ·mol^{-1}. This is in complete agreement with experimental data on collagen-like peptides. Indeed, in the Gly-Pro-Pro (GPP) collagen-like peptide, the Pro puckering in the X position prefers (7:2[25] or 10:0[26]) a D conformation. Instead, in Gly-Pro-Hyp (GPO), the Pro in X has a more uniformly distributed pucker among the D/U states (4:3, 28:23).[28,29] Due to the interplay between the pucker and the main chain, the side chain mobility induces structural flexibility to the collagen protein. Recently, Chow et al.,[30] by means of solid state nuclear magnetic resonance (ss-NMR) spectroscopy and classical molecular dynamics (CMD) calculations related the free ring flipping of Pro in GPO peptides to the presence of Hyp in Y position. As the relative stabilities in all considered processes are small numbers, we checked our results by increasing the precision of the CRYSTAL calculations by setting the bi-electronic and exchange repulsion integrals accuracy to much higher values than the default ones. It is reassuring that these results are indistinguishable from the less accurate ones.

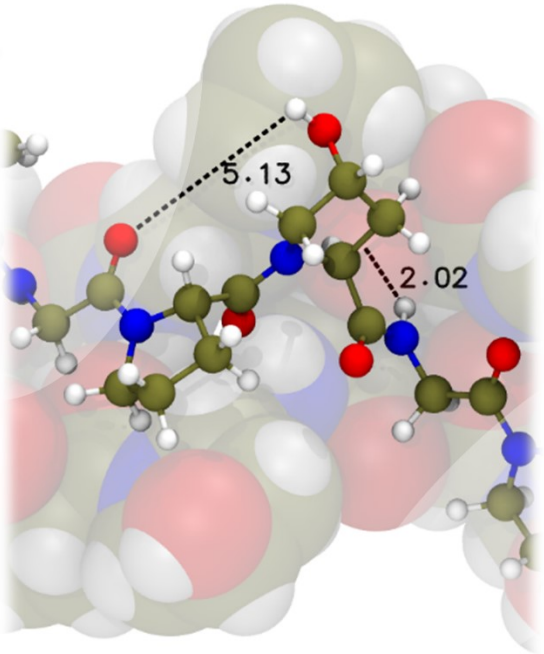
Table 6. Geometry and energy of the **GPO triple helical models.** The angles are expressed in degrees, the hydrogen bond (Hb) and cell lengths (cell) in Angstrom whilst the energies in kJ·mol^{-1} per triplet. The name (n) is associated to a symbol (syb) which summarizes the pucker of Pro in X and Hyp in Y (XY).

			Gly			Pro (X)				Hyp (Y)									
n	syb	OH	ϕ	ψ	ω	ϕ	ψ	ω	χ^2	ϕ	ψ	ω	χ^2	Hb	hb	cell	ΔE_{tot}	ΔE_{el}	ΔE_{disp}
1	DU	A	-68.3	177.1	175.4	-73.9	172.0	168.8	-37.8	-57.5	158.4	162.3	31.3	2.14	4.13	19.93	0.00	0.00	0.00
2		B	-69.8	176.6	176.1	-71.8	167.6	169.8	-36.2	-59.0	156.8	167.9	30.0	2.06	5.43	19.98	5.55	2.84	2.71
3	UU	A	-71.2	179.9	174.1	-67.7	170.3	161.3	29.2	-55.1	158.6	162.9	33.9	2.18	3.70	19.80	1.92	0.26	1.66
4		B	-74.2	179.5	174.7	-63.5	161.9	162.8	32.5	-58.0	156.9	173.2	31.4	2.02	5.13	19.85	6.34	2.25	4.09
5	UD	A	-76.6	179.0	175.6	-63.0	160.2	164.9	32.4	-61.9	158.6	177.6	-37.0	1.99	5.35	20.01	10.00	3.08	6.92
6		B	-75.9	179.1	175.6	-62.2	159.2	164.4	32.7	-61.0	157.0	178.0	-36.9	1.98	/	19.9	12.17	4.95	7.22
7	DD	A	-70.7	176.4	176.9	-70.1	164.7	167.4	-36.2	-57.2	155.6	169.1	-37.2	2.03	5.07	19.99	5.14	0.69	4.45
8		B	-70.2	176.0	176.8	-69.1	163.6	167.6	-35.9	-57.0	154.4	170.2	-36.9	2.02	/	19.9	7.59	2.49	5.10


1
4.13
2.14


2
5.43
2.06


3.70
2.18
3


5.13
2.02
4

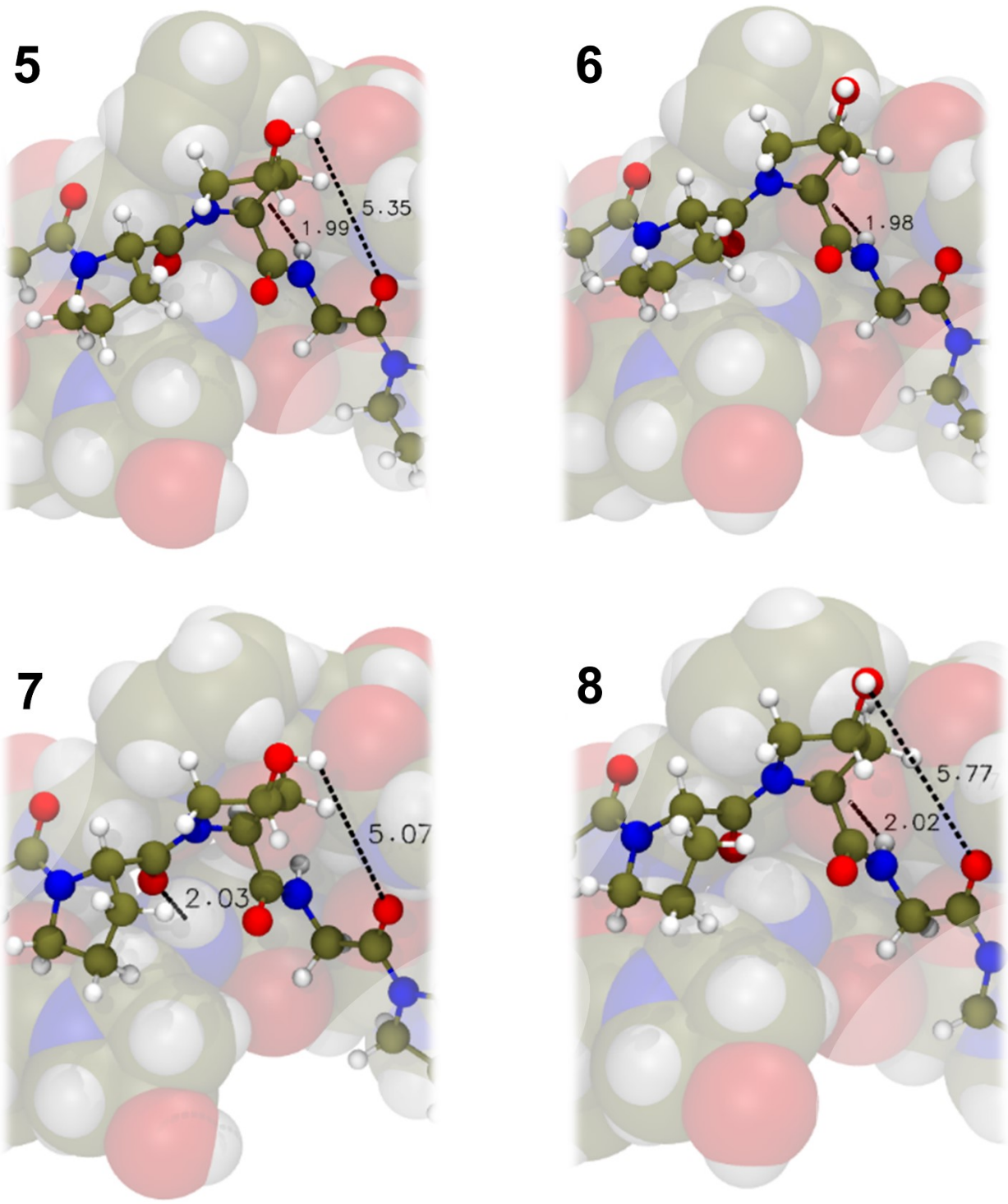


Figure 23. Triple helical collagen with GPO composition (**1**-**8**), see Table 6. The asymmetric unit cell within the protein structure is highlight as well as the H contacts.

Folding of the triple helix with GPP and GPO composition.

Finally, we compute the folding energy reaction for both the GPP and GPO compositions. The initial and final conformations are the most stable for each system, see Table 2 and 4-6. The results along with the energy decomposition (see Chapter 2) are gathered in Table 7. First of all, **the interaction energy between strands (BE*) is higher for the GPO than GPP, with 95.2 and 90.6 kJ·mol^{-1} per triplet, respectively. This agrees with the expected energy ranking for this type of systems**. Anyways, due to a larger deformation energy (δE_P, see Table 7) in the GPO case, the overall BE is slightly higher for the GPP case. **The choice of a proper description for the starting geometry in the folding energy calculation is a complex matter**. For instance, regarding the peptide bond only, the *cis* peptide bond in tertiary amides is frequent (Pro has a secondary amino group which forms tertiary amides in proteins). In a protein database survey the 5.2% of Pros has the preceding peptide bond in the *cis* state.[31] Therefore, **the all *trans* PPII geometry results un appropriate as a reference for the starting geometry of a collagen single strand.** Moreover A D puckered Pro is more likely to have a peptide bond in *cis* conformation than a U puckered one (in unfolded COLs, *cis*/*trans* rate of Gly-Pro (16%) and X-Hyp (8%) peptide bonds).[32,33] Therefore, the amount of *cis* bonds is also different in the GPO and GPP single strands, leading to different reference starting geometries.

Interestingly, **the contribution of the dispersion component to the final binding energy is around two times larger than the pure DFT contribution** (see BE* in Table 7).

Table 7. Energetic of the folding of the COL protein for the GPP and GPO compositions. The δE_P term refers to the deformation energy of transition from the PPII to the 7_2 helix for a single strand. Values in kJ·mol^{-1} per triplet.

	δE_P	BSSE	BE*	BE*^{C}	BE	BE^C
			GPO			
B3LYP/TZP	20.5	7.0	35.2	28.6	14.7	7.8
D3^{ABC}	2.1	0.0	60.0	60.0	57.9	57.9
Total	22.6	7.0	95.2	88.3	72.7	65.7
			GPP			
B3LYP/TZP	14.4	5.8	31.3	25.6	17.0	11.2
D3^{ABC}	1.6	0.0	59.3	59.3	57.7	57.7
Total	15.9	5.8	90.6	84.9	74.7	68.9

Perspectives

The "fast *ab initio*" HF-3c method was shown to be accurate enough in both organic and inorganic benchmark sets, and therefore it is suitable for our purposes. Due to the remarkable speed-up with respect to standard DFT, e.g. ~30 times with respect to B3LYP-D3^{ABC} and ~10 times with respect to pure GGA DFT functionals, more realistic models become at hands. For instance, the simple model of bone presented in Section 3.2 can be upgraded to triple helical collagen-like polymer in interaction with a Hap surface slab (Figure 24).

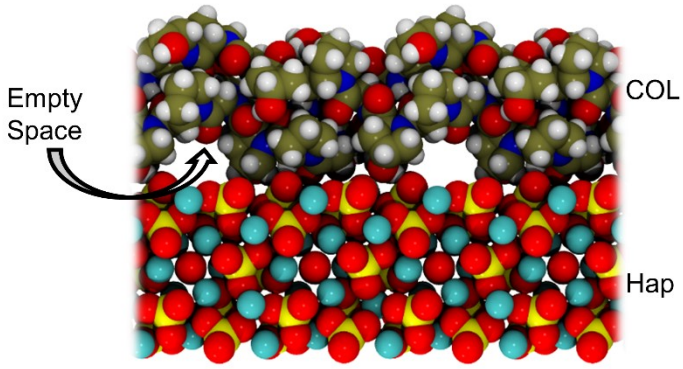


Figure 24. COL/Hap interface. The empty space in the image are due to the forcing of the collagen protein to have a certain periodic pattern which limits the adhesion of the protein onto the Hap.

Currently, we are also extending the work presented in Section 3.3 to:

- the folding of collagens into the upper level of the hierarchical structure of tendon, e.g. the collagen fibril (Figure 25), to gain some insights on the driving forces in collagen fibrillogenesis. Due to the large size of the models, we must either rely on a HF-3c-like method or on the high-performance computing (HPC) approach by exploiting the massive parallel features of CRYSTAL17 code.

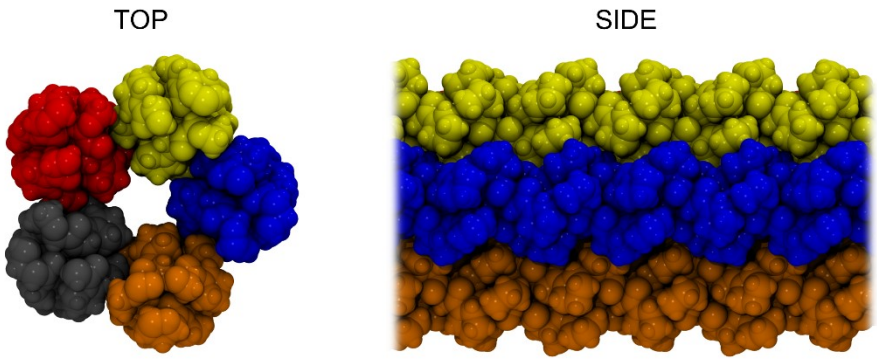


Figure 25. Top and side view of a tendon model following the structure of Smith's fibril,[34] see Chapter 1.

- the role of water (modelled through explicit micro-solvation) in the triple helix formation. In this way, we hopefully account for the enthalpy contribution of solvent in collagen stability, Figure 26.

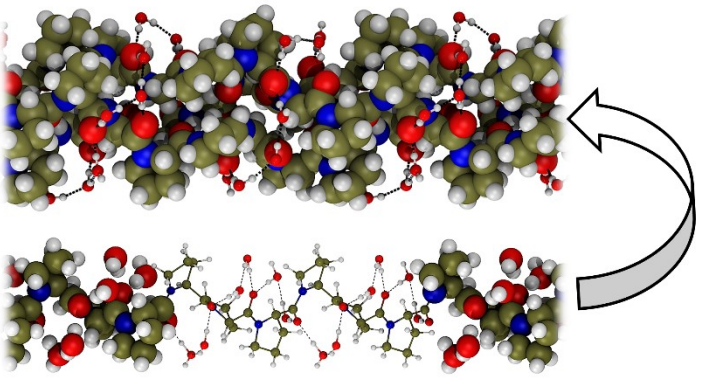


Figure 26. PPII→HEL reaction with explicit waters inclusion.

Bibliography

1. Sure, R. & Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **34,** 1672–1685 (2013).
2. Neese, F. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2,** 73–78 (2012).
3. Furche, F. *et al.* Turbomole. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4,** 91–100 (2014).
4. Dovesi, R. *et al.* CRYSTAL14 : A program for the *ab initio* investigation of crystalline solids. *Int. J. Quantum Chem.* **114,** 1287–1317 (2014).
5. Peters, L. D. M., Kussmann, J. & Ochsenfeld, C. Efficient and Accurate Born–Oppenheimer Molecular Dynamics for Large Molecular Systems. *J. Chem. Theory Comput.* **13,** 5479–5485 (2017).
6. Caldararu, O., Olsson, M. A., Riplinger, C., Neese, F. & Ryde, U. Binding free energies in the SAMPL5 octa-acid host–guest challenge calculated with DFT-D3 and CCSD(T). *J. Comput. Aided. Mol. Des.* **31,** 87–106 (2017).
7. Christensen, A. S., Kromann, J. C., Jensen, J. H. & Cui, Q. Intermolecular interactions in the condensed phase: Evaluation of semi-empirical quantum mechanical methods. *J. Chem. Phys.* **147,** 161704 (2017).
8. Cutini, M. *et al.* Assessment of Different Quantum Mechanical Methods for the Prediction of Structure and Cohesive Energy of Molecular Crystals. *J. Chem. Theory Comput.* **12,** 3340–3352 (2016).
9. Chickos, J. S. Enthalpies of Sublimation after a Century of Measurement : A View as Seen through the Eyes of a Collector. *Netsu Sokutei* **30,** 116–124 (2003).
10. Reilly, A. M. & Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. *J. Chem. Phys.* **139,** 24705 (2013).
11. Sinha, P., Boesch, S. E., Gu, C., Wheeler, R. A. & Wilson, A. K. Harmonic vibrational frequencies: Scaling factors for HF, B3LYP, and MP2 methods in combination with correlation consistent basis sets. *J. Phys. Chem. A* **108,** 9213–9217 (2004).
12. Román-Román, E. I. & Zicovich-Wilson, C. M. The role of long-range van der Waals forces in the relative stability of SiO2-zeolites. *Chem. Phys. Lett.* **619,** 109–114 (2015).
13. Cutini, M., Corno, M. & Ugliengo, P. Method Dependence of Proline Ring Flexibility in the Poly-L-Proline Type II Polymer. *J. Chem. Theory Comput.* **13,** 370−379 (2017).
14. Czinki, E. & Császár, A. G. Conformers of gaseous proline. *Chem. - A Eur. J.* **9,** 1008–1019 (2003).
15. Detar, D. F. & Luthra, N. P. Conformations of Proline. *J. Am. Chem. Soc.* **76,** 1232–1244 (1977).
16. Ho, B. K., Coutsias, E. A., Seok, C. & Dill, K. A. The flexibility in the proline ring couples to the protein backbone. *Protein Sci.* **14,** 1011–8 (2005).
17. London, R. E. On the interpretation of carbon-13 spin-lattice relaxation resulting from ring puckering in proline. *J. Am. Chem. Soc.* **100,** 2678–2685 (1978).
18. Zhong, H. & Carlson, H. A. Conformational Studies of Polyprolines. *J. Chem. Theory Comput.* **2,** 342–353 (2006).
19. Jimenez-Izal, E., Chiatti, F., Corno, M., Rimola, A. & Ugliengo, P. Glycine Adsorption at Nonstoichiometric (010) Hydroxyapatite Surfaces: A B3LYP Study. *J. Phys. Chem. C* **116,** 14561–14567 (2012).
20. Corno, M., Rimola, A., Bolis, V. & Ugliengo, P. Hydroxyapatite as a key biomaterial: quantum-mechanical simulation of its surfaces in interaction with biomolecules. *Phys. Chem. Chem. Phys.* **12,** 6309–6329 (2010).
21. Almora-Barrios, N. & De Leeuw, N. H. A density functional theory study of the interaction of

collagen peptides with hydroxyapatite surfaces. *Langmuir* **26,** 14535–14542 (2010).
22. Tao, J. *et al.* Energetic basis for the molecular-scale organization of bone. *Proc. Natl. Acad. Sci.* **112,** 326–331 (2015).
23. Cutini, M., Corno, M., Costa, D. & Ugliengo, P. How Does Collagen Adsorb on Hydroxyapatite? Insights From *Ab initio* Simulations on a Poly-Proline Type II Model. *J. Phys. Chem. C* (2017).
24. Adzhubei, A. A., Sternberg, M. J. E. & Makarov, A. A. Polyproline-II Helix in Proteins: Structure and Function. *J. Mol. Biol.* **425,** 2100–2132 (2013).
25. Okuyama, K., Nagarajan, V. & Kamitori, S. 7/2-Helical model for collagen - Evidence from model peptides. *Proc. Indian Acad. Sci. - Chem. Sci.* **111,** 19–34 (1999).
26. Berisio, R., Vitagliano, L., Mazzarella, L. & Zagari, A. Crystal structure of the collagen triple helix model [(Pro-Pro-Gly)10]3. *Protein Sci.* **11,** 262–270 (2002).
27. Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M. & Kakudo, M. Crystal and molecular structure of a collagen-like polypeptide (Pro-Pro-Gly)10. *J. Mol. Biol.* **152,** 427–443 (1981).
28. Okuyama, K., Miyama, K., Mizuno, K. & Bächinger, H. P. Crystal structure of (Gly-Pro-Hyp)9: Implications for the collagen molecular model. *Biopolymers* **97,** 607–616 (2012).
29. Okuyama, K. *et al.* Crystal structures of collagen model peptides with Pro-Hyp-Gly repeating sequence at 1.26 Å resolution: Implications for proline ring puckering. *Biopolym. - Pept. Sci. Sect.* **76,** 367–377 (2004).
30. Ying Chow, W. *et al.* Hydroxyproline Ring Pucker Causes Frustration of Helix Parameters in the Collagen Triple Helix. *Sci. Rep.* **5,** 12556 (2015).
31. Vitagliano, L., Berisio, R., Mastrangelo, A., Mazzarella, L. & Zagari, A. Preferred proline puckerings in *cis* and *trans* peptide groups: implications for collagen stability. *Protein Sci.* **10,** 2627–32 (2001).
32. Sarkar, S. K., Young, P. E., Sullivan, C. E. & Torchia, D. a. Detection of *cis* and *trans* X-Pro peptide bonds in proteins by 13C NMR: application to collagen. *Proc. Natl. Acad. Sci. U. S. A.* **81,** 4800–4803 (1984).
33. DeRider, M. L. *et al.* Collagen stability: Insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations. *J. Am. Chem. Soc.* **124,** 2497–2505 (2002).
34. Smith, J. W. Molecular Pattern in Native Collagen. *Nature* **219,** 157–158 (1968).

Graphical Summary