Balancing Cost and Accuracy in Quantum Mechanical Simulations on Collagen Protein Models

Balancing Cost and Accuracy in Quantum Mechanical Simulations on Collagen Protein Models

Michele Cutini,* Irene Bechis, Marta Corno, and Piero Ugliengo*

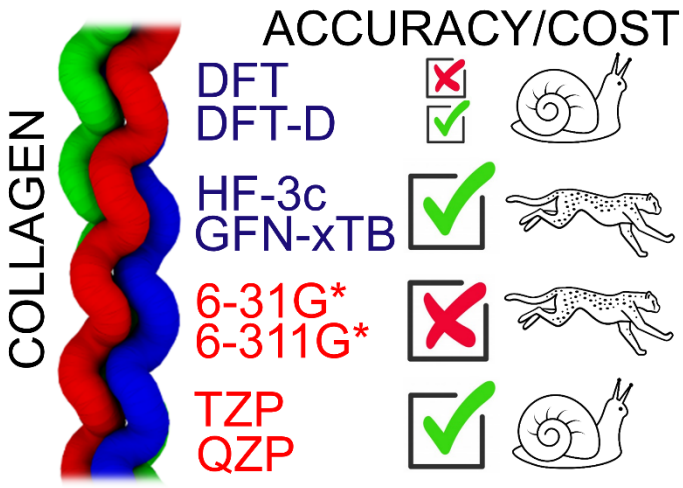
University of Turin, Department of Chemistry and NIS (Nanostructured Interfaces and Surfaces) Centre, Turin, 10125, IT

I.B. present address: Department of Chemistry, Imperial College London, Molecular Science Research Hub, White City Campus, Wood Lane, London, W12 0BZ, UK

e-mail: michele.cutini@unito.it, piero.ugliengo@unito.it

Abstract and TOC

Collagen proteins are spread in almost every vertebrate's tissue with mechanical function. The defining feature of this fundamental family of proteins is its well-known collagen triple helical domain. This helical domain can have different geometry, varying in helical elongation and inter-strands contact, as a function of the aminoacidic composition. The helical geometrical features play an important role in the interaction of the collagen protein with cell receptors, but for the vast majority of collagen compositions, these geometrical features are unknown. Quantum mechanical (QM) simulations based on density functional theory provide a robust approach to characterize the scenario on the collagen composition-structure relationships. In this work we analyze the role of the adopted computational method in predicting collagen structure for two purposes. Firstly, we look for a cost-effective computational approach to apply to a large scale composition-structure analysis. Secondly, we attempt to assess the robustness of the predictions by varying the QM methods. Therefore, we have run geometry optimization on periodic models of collagen protein using a variety of approaches, based on the most commonly used DFT functionals (PBE, HSE06 and B3LYP) with and without dispersion correction ($D3^{ABC}$). We have coupled these methods with several different basis sets, looking for the highest accuracy/cost ratio. Furthermore, we have studied the performance of the composite HF-3c method, and the semiempirical GFN1-xTB method. Our results identify a computational recipe that is potentially capable of predicting collagen structural features in line with DFT simulations, with orders of magnitude reduced computational cost, encouraging further investigations on the topic.



Introduction

Collagen proteins are key elements of most vertebrate's tissues with mechanical function. These proteins are known to give elasticity to tissues, leading to extraordinary light and strong materials (e.g. mammal's bones) when combined with bio-minerals. The peculiar triple helical motif (known as collagen triple helix) defines entirely this family of proteins.[1–3] This compact organization requires a strict aminoacid sequence. Indeed, in each of the three strands, a Glycine (Gly) is found every three aminoacids, leading to the peculiar Gly-X-Y pattern. The other aminoacids (X and Y) can vary, but Proline (Pro) and (2S,4R)-4-Hydroxyl-proline (Hyp) are the most commonly found, in X and Y, respectively.[4] The collagen triple helix can be more or less packed, depending on the amino acids in X and Y.[5,6] This topic has been debated for decades.[7] Nowadays, there is general consensus that Pro-rich collagens prefer a 7/2 helical packing, and Pro-poor collagens a 10/3 one. A 7/2 helix contains 7 aminoacidic triplets into two helix turns, while a 10/3 helix contains 10 aminoacidic triplets in three helix turns, making a 7/2 helix tighter than a 10/3 helix.[8,9]

The main experimental technique providing evidence on collagen triple helicity is X-ray diffraction on crystals of collagen-like peptides. Unfortunately, the amount of data on Pro-poor collagens is scarce, mainly coming from Pro-rich collagens with only few different residues in the core of the peptide sequence.[5,10] This lack of experimental data does not allow to have a clear understanding of the role of the composition on the structure of collagen, which is important on both basic and applicative levels. Indeed, predicting collagen structure is a key information in the interaction of collagen protein with cell receptors.[11]

In this scenario, molecular simulations are a promising tool for predicting the geometrical packing of collagen triple helices as a function of the aminoacidic composition. We have recently proposed a computational approach based on hybrid DFT-D simulation, which can

predict correctly the helical packing for Pro-rich collagens.[9] The computational procedure is fairly simple. It consists of simulating collagen triple helices with exactly the ideal 7/2 and 10/3 helical geometry and directly compare their energy. Its intrinsic simplicity allows a straight-forward extension of this type of analysis to all natural (and not) Gly-X-Y collagen triplets. Unfortunately, the large composition/conformational variability of collagen discourages the use of state-of-the-art hybrid DFT simulation, despite recent progress in speeding up DFT when dealing with the simulation of large molecules. Indeed, the adoption of DFT to study large systems requires high performance computing (HPC) facilities to account for the relatively large request of central memory and specific computer codes able to exploit massive parallelism, as shown recently for the case of Crambin protein.[12] However, when considering that collagen features depend upon its aminoacid composition/conformation and, in perspective, the most promising ones should also be studied when interacting with the hydroxyapatite (the natural partner in the definition of many collagen-based biomaterials), it turned out that DFT is a too costly approach. A possible alternative is to adopt semiempirical methods, as reviewed in Ref. [13] for noncovalent interactions for chemical and biochemical applications.

Therefore, in this work, we have investigated the role of the computational approach in the prediction of collagen protein helical features, looking for a cost-effective approach to substitute to DFT simulations. Furthermore, we have tested the consistency of collagen protein predictions as a function of the adopted Quantum Mechanical (QM) method. This would allow a future use of QM techniques for reliable structural (helical packing) and properties predictions on a large number of collagen protein models, thus filling the gap on the collagen structure-composition relationship.

The main “hyper-parameters” that can affect a QM investigation are the Hamiltonian type and the basis set employed. Regarding the Hamiltonian, we have chosen to compare the most common DFT functionals employed in both plain-wave and Gaussian-based simulations, e.g. the pure DFT PBE functional,[14] the hybrid HSE06 functional,[15] and the well-established hybrid B3LYP functional.[16–18] The role of dispersion interactions has been taken into account by comparing dispersion-corrected (DFT-D) with plain DFT simulations. Furthermore, we have also tested the HF-3c method,[19] which gave cost-effective results in predicting molecular crystals,[20] polymers,[21] microporous,[22] and layered materials properties.[23] Among the various available semiempirical Hamiltonians, we have tested the GFN1-xTB method, which provided excellent results for several applications.[24,25] To our

knowledge, this is the first application of the xTB method, in the GFN1-xTB flavor, to true periodic bio-polymeric systems.

As for the basis set (BS) employed for the DFT simulations, we have relied on several BSs made of atom-centered Gaussian-type functions as well as plane waves functions. Simulations run with Gaussian functions BSs suffer the well-known basis set superposition error (BSSE), which is relevant for weakly bonded systems, such as the collagen triple helices. BSSE can artificially shorten the inter-molecular distances and overestimate interaction energies. These spurious effects can be minimized by using large BSs, but the cost of the simulation rapidly increases with the BS size reducing its applicability. Therefore, we have compared different types of Gaussian based BSs, with a growing number of BS functions, analyzing the accuracy/cost of the resulting simulations. We compared two types of Pople type basis sets,[26] as well as two types of Ahlrichs type basis sets.[27–29] A plane wave basis set is also employed as reference method, being intrinsically BSSE-free.

Finally, we have also tested "hybrid" methodologies such as the DFT-D//HF-3c approach, which combines a fast geometry optimization run at the HF-3c level with an accurate energy prediction at the DFT level,[20–22] and, the even faster HF-3c//GFN1-xTB and DFT-D//GFN1-xTB approaches in which the geometry optimization is run at the very cheap semiempirical GFN1-xTB level. The comparison between these computational approaches is performed evaluating several features of the collagen triple helix, such as the helical packing, the geometry and the inter-strand interaction energy. The results of this work allow to clearly identify a cost-effective computational procedure, capable of predicting preferred collagen helical organization in line with state-of-the-art hybrid DFT simulations at a reduced computational cost.

Computational Details

We computed energies and relaxed geometries within the HF and DFT frameworks, by means of the CRYSTAL17 code.[30] For the HF framework, we employed the HF-3c method,[19] based on a HF calculation with the minimal MINIX Gaussian-type basis set. Despite the minimal quality basis set, the inclusion of the pair-wise semi-empirical corrections, i.e. i) BSSE correction,[31] ii) inclusion of dispersion interactions,[32] iii) correction of the systematic error in the inter-atomic distances,[19] greatly improves the results. We have recently refined the HF-3c method for periodic system geometries, by reducing the D3 dispersion term. Specifically, the dipole-quadrupole term (s_8) of the D3 scheme is scaled by a factor of 0.27. We named the resulting method as HF-3c-027.[20] The HF-3c-027 gave excellent results in

computing protein, molecular crystal and microporous crystal structures, see Ref.[20–22] Most DFT simulations were run using the B3LYP hybrid functional,[14] and corrected with the D3 scheme which is available in the CRYSTAL17 suite,[32] including the Axilrod–Teller–Muto (ATM)-three-body-term ($D3^{ABC}$).[33,34] The recent D4 dispersion scheme has not been implement in the CRYSTAL suite, but it will be take into consideration in the future.[35] We also employed the GGA PBE functional,[14] and the hybrid HSE06 functional,[15] which are the most common functionals employed in plane-wave simulations. DFT simulations were carried out using several basis sets made of Gaussian functions, such as Pople-type 6-31G* and 6-311G* basis sets, and Ahlrichs-type VTZP and QZVP basis sets.[26–29] The basis sets are fully reported in Table S1-S4.

Atomic positions and cell vectors optimization were performed adopting the analytical gradient method. The Hessian was upgraded with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.[36–38] We set tolerances for the convergence of the maximum allowed gradient and the maximum atomic displacement to default values. The recently introduced DIIS extrapolator technique has been employed to speed up the SCF convergence.[39] Details on the tolerance values controlling the Coulomb and exchange series in periodic systems,[40] and the shrink factor (k points sampling) used in the calculations are set to 6 6 6 6 14 and 4 4, respectively. Only when the QZVP basis set is employed the tolerances on the integral are tighten to 7 7 7 7 25 for ensuring SCF convergence. We tested the effect of tightening integral tolerance to have a formally correct comparison between all methodologies. The variations in the values obtained using different integral tolerances are negligible for all the cases.

To compare Gaussian and plane wave basis sets, we run plane wave periodic simulations with the Vienna *Ab-initio* Simulation Package (VASP),[41–44] using the PBE and PBE-D2 functionals.[45] The kinetic energy cut-off has been set to 500 eV and 1000 eV, and the SCF iterative procedure was converged to a tolerance in total energy of $\Delta E = 10^{-6}$ eV. The Monkhorst-Pack sampling of the Brillouin zone was used for the k-points mesh, with the same number of k-points as in CRYSTAL17 calculations.

We run the semiempirical GFN1-xTB Hamiltonian[24] simulations within the periodic boundary conditions as implemented in the CP2K code,[46,47] version 7.1. The graphical visualization and structural manipulation of structures was performed with MOLDRAW version 2.0.[48] Images were rendered with VMD.[49]

Results and Discussion

Molecular Models and Computed Quantities

In this work, we consider four simplified models of collagen as 1D biopolymers, by varying their aminoacidic composition and triple helix features. Regarding the composition, all collagen models are homo-trimeric collagens, e.g. the three strands composing the protein are identical. Furthermore, each strand is made by the repetition of only one type of Gly-X-Y aminoacidic triplet. The two different collagen compositions studied here are Gly-Gly-Hyp (GGO) and Gly-Ala-Hyp (GAO), which are reported graphically in Figure 1 a) and b). They differ for one aminoacid only, i.e. the aminoacid in X position of the triplet. As for the helicity, we have simulated collagen models that have 7/2 and 10/3 helicity, see Figure 2. We refer to Ref.[8,9] for a detailed description of these different helical models. Combining the two different helicities and compositions results in four collagen models, which are named hereafter as GXO-H with X = G and A as a function of the composition, and H = 7/2 and 10/3 in function of the helical symmetry imposed to the model, e.g. GGO-7/2, GGO-10/3, GAO-7/2 and GAO-10/3.

As for the side chain aminoacidic conformation, the Hyp puckering is kept fixed to the UP conformation with the most stable OH orientation found in collagens with GPO composition by DFT simulations.[8] The UP puckering of Hyp is the side conformation expected by theoretical and experimental works,[8,21,50,51] due to the OH substitution on position 4R of the pyrrolidine ring, which differs from the most common side chain conformation of plain Pro residue.[21,52] The other aminoacids within the models (Gly and Ala) have no relevant side chain conformation variability.

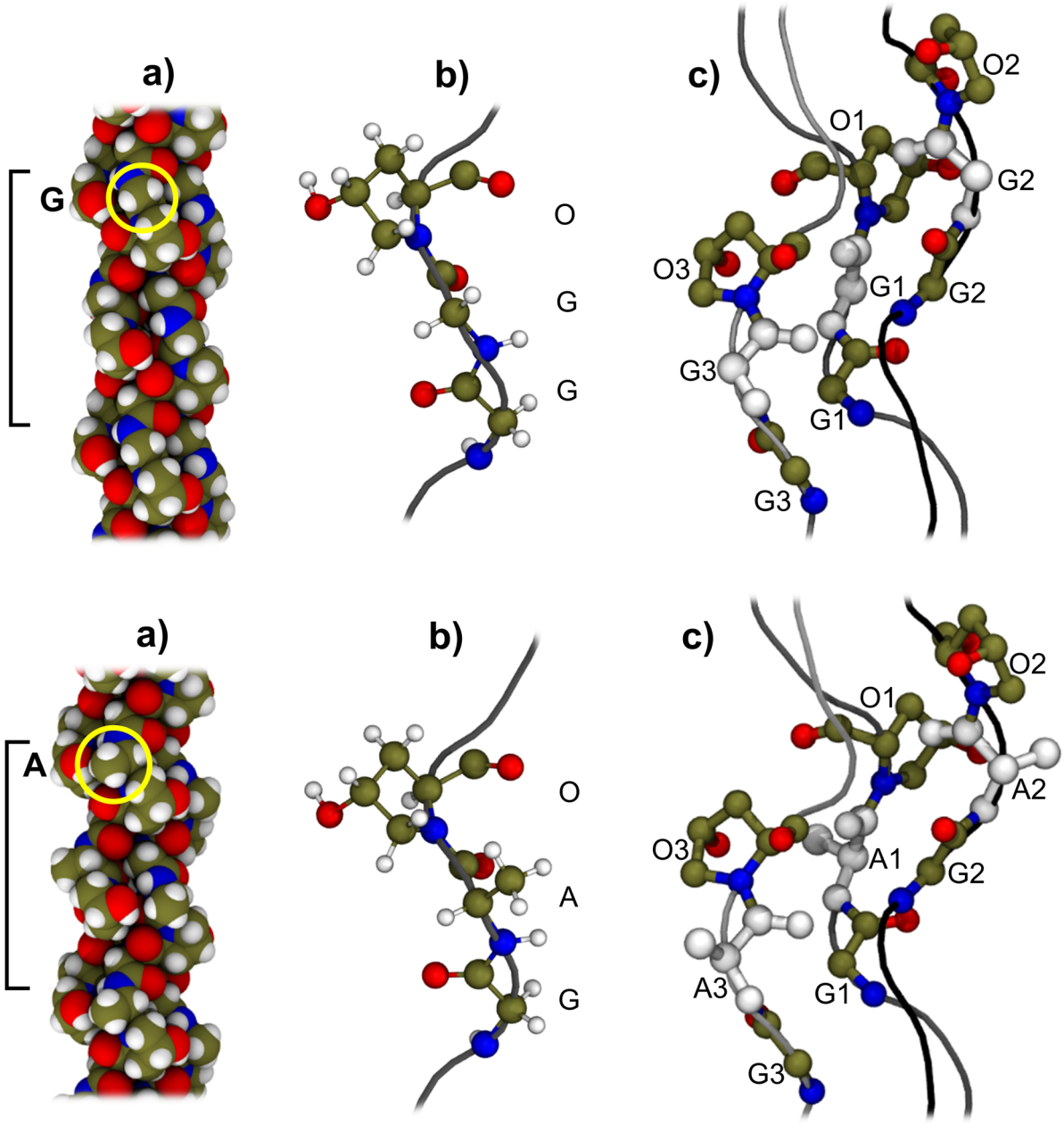


Figure 1 Graphical representation of the GGO-7/2 (**TOP**) and GAO-7/2 (**BOTTOM**) helices relaxed at the B3LYP-D/VTZP level. Colour code: Oxygen in red, Hydrogen in white, Nitrogen in blue, Carbon in brown-green. **a)**: van der Waals representation of collagen triple helix highlighting the unit cell length and the aminoacid in the X position. **b)**: Single aminoacidic triplet extracted from the collagen triple helix as balls and sticks. **c)**: Interaction between GXO triplets (X= G or A). All atoms are reported in balls and sticks with hydrogen atoms omitted from clarity of representation. We have reported the three collagen strands

as tubes in grey scale, and the aminoacid in X position in white colour. Each aminoacid is labelled with a number associated to the collagen strand number (named as 1, 2 and 3).

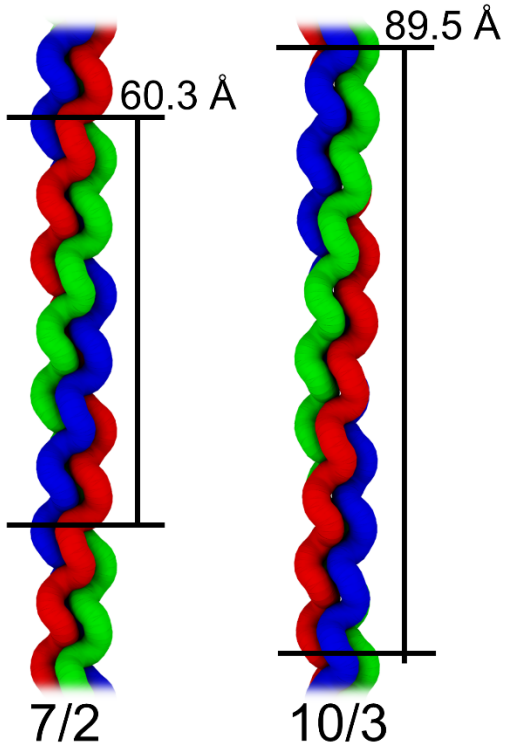


Figure 2 Graphical representation of collagen 7/2 and 10/3 helices represented as a 1D polymers. Each of the three independent collagen strands are represented as coloured tubes. The length of the helical repetition is also reported for clarity.

For all our triple helical models, we have computed the binding energy between collagen single strands (BE*), with the following expression:

$$BE^* = E(F//COL) - E(COL//COL) \quad (1)$$

The name following the double slash identifies the optimized geometry at which the energy has been computed. For instance, E(COL//COL) is the energy of a collagen triple helix, COL, in its fully optimized geometry, and E(F//COL) is the energy of the single collagen strand, F, in the triple helix optimized geometry. The use of Gaussian-type orbitals poses a severe problem when the interaction energy between various molecules is computed due to basis set superposition error (BSSE). Therefore, we corrected the BE* for BSSE (BE^{*C}) through the counter-poise method (CP).

$$BE^{*C} = BE^* - BSSE \quad (2)$$

We have also compared helical packing stability (ΔE-helix) within the same compositions. This is defined as (for X= G and A):

$$\Delta E\text{-helix} = E(GXO\text{-}10/3) - E(GXO\text{-}7/2) \quad (3)$$

where E(GXO-10/3) is the E(COL//COL) for the 10/3 helix and E(GXO-7/2) is the E(COL//COL) for the 7/2 helix.

The Role of Basis Set

To assess the role of the basis set in collagen DFT simulations we have run B3LYP-D simulations using four different basis sets. Two BSs are of the Pople split-valence type,[26] and two are of the Ahlrichs VZ type.[27] For the former case, we have employed 6-31G* and 6-311G* basis sets,[26] for the latter case, TZP and QZP basis sets. We have carried out the results and the discussion focusing on the computational cost of the calculation, along with the geometry and energy (BE^{*C}, BSSE and ΔE-helix) prediction. Regarding geometry, we have relegated the torsional angles values and the inter-strands electrostatic contact lengths in the SI, see Figure S2-S3. To have a more concise view on collagen geometry, we have only reported the geometrical rise per triplet compared with the experimental values in the main text.[5] This geometrical feature is computed as the length of the polymer unit cell divided by the number of triplets within the unit cell.

The first element of discussion is the computational cost of the selected BSs. For instance, for the GGO-7/2 case, the AO number in the 1D unit cell are 2023, 2471, 2695 and 3325 for the 6-31G*, 6-311G*, VTZP and QZVP BSs, respectively. The needed time for an energy and gradient calculation (SCF+G) increases by 2, 9 and 30 times for 6-311G*, VTZP and QZVP cases with respect to the fastest 6-31G* BS.

As for the geometry, in Figure 3 we have compared the rise per triplet of 10/3 and 7/2 helices in real collagens and in our models. We expect some discrepancy between the predicted values and the experimental ones, as the latter are averaged on different aminoacidic compositions and not specifically set for GAO and GGO homo-trimeric collagens. We also have run geometry optimization without relaxing the unit cell parameters and using the ideal geometrical rise per triplet values. Such helices are the ones usually considered in the X-ray diffraction works on collagen. We showed that this does not give different results on values like the ΔE-helix, with exception for the GFN1-xTB case, which is discussed later in the text. At the same time, relaxing the polymer unit cell gives an indication of the deviation of the real collagen helices from the ideal cases. In our opinion this is an interesting argument of discussion, thus we performed in all cases cell optimization.

Relaxed structures adopting the Pople type BSs are not close to the experimental ones.[5] The 6-31G* BS predicts the 10/3 helix as more compressed than the 7/2 helix, for both GGO and GAO cases. By incrementing the BS size, the agreement improved. Indeed, the 6-311G* BS correctly predicts the rise per triplet order, despite the relative internal difference being underestimated. This is computed as 0.08/0.06 Å for GGO/GAO, to be compared to an experimental value of 0.34 Å. Using the larger Ahlrichs type BSs, the computational prediction is closer to experiment. The relative difference between helices is 0.23/0.24 Å for GGO/GAO (at the B3LYP-D/QZP level) with negligible differences between TZP and QZP basis sets.

All methodologies give similar estimation of the binding energy between collagen strands (BE^{*C}), for both GGO-7/2 and GAO-10/3 helices (see Figure 4). This is not the case for the other helices, e.g. GGO-10/3 and GAO-7/2. Indeed, using the 6-31G* BS for the GGO-10/3 case, and the 6-31G* and the 6-311G* BSs for the GAO-7/2 case, we compute BE^{*C} fairly different from the ones obtained with Ahlrichs BSs. This is correlated with differences in the electrostatic contacts between the protein strands, see Figure S2 and S3, which are the main responsible of the electrostatic component of the interstrand interaction energy. Due to the lower BSSE, whose contribution to the BE^{*C} is commented hereafter, the geometries, and thus the BE^{*C}, computed with the Ahlrichs BSs are considered as internal reference. The BSSE reduces by increasing the quality of the basis set, see Figure 4. The high value for Pople type BSs (more than 23 $kJ \cdot mol^{-1} \cdot triplet^{-1}$) is at the origin of the variations in the electrostatic contact lengths. The Ahlrichs VZ BSs reduces the BSSE by a factor three (up to values lower than 7 $kJ \cdot mol^{-1} \cdot triplet^{-1}$), with similar results for TZP and QZP BSs.

The computed ΔE-helix values are reported in Figure 5. Interestingly, all BSs compute the 7/2 helix more stable than the 10/3 helix, but with different ΔE-helix values. The results at the 6-31G* level give a different order of ΔE-helix values for GGO and GAO composition, with respect to QZP. Conversely, the results at the 6-311G* level tend to over-stabilize the 7/2 helix for both GGO and GAO, with respect to QZP. As for the results on geometry and BE^{*C}, the TZP BS gives coherent results with the more expensive QZP BS.

Furthermore, we have checked the role of the residual BSSE of the QZP basis set, by comparing with the BSSE-free plane-wave (PW) basis set, using the PBE-D functional with both the QZP BS and the PW BS. The results are gathered in Table S5, which indicates that both BSs give very similar relaxed structures. This is an indication of the reliability of the QZP BS as our internal selected reference BS.

Considering the discussion carried out so far, the best cost-effective BS is the Ahlrichs TZP BS. Indeed, this BS ensures the quality of the results of the QZP BS at 1/3 of its computational cost. Therefore, in the next paper section, we will address the role of the Hamiltonian adopting the TZP BS as standard.

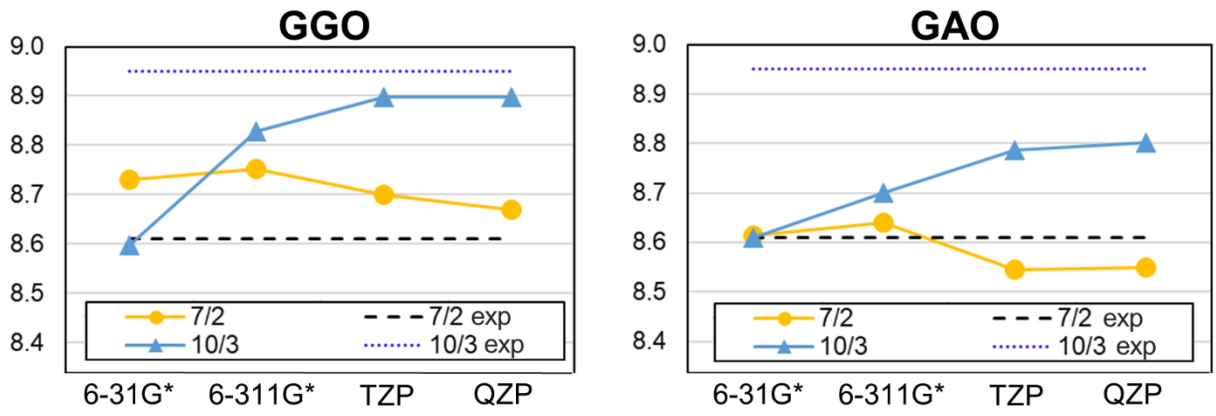


Figure 3. BS effect on the collagen rise per residue, using the B3LYP-D functional. Experimental rise per triplet reported as dashed line. Values are in Å.

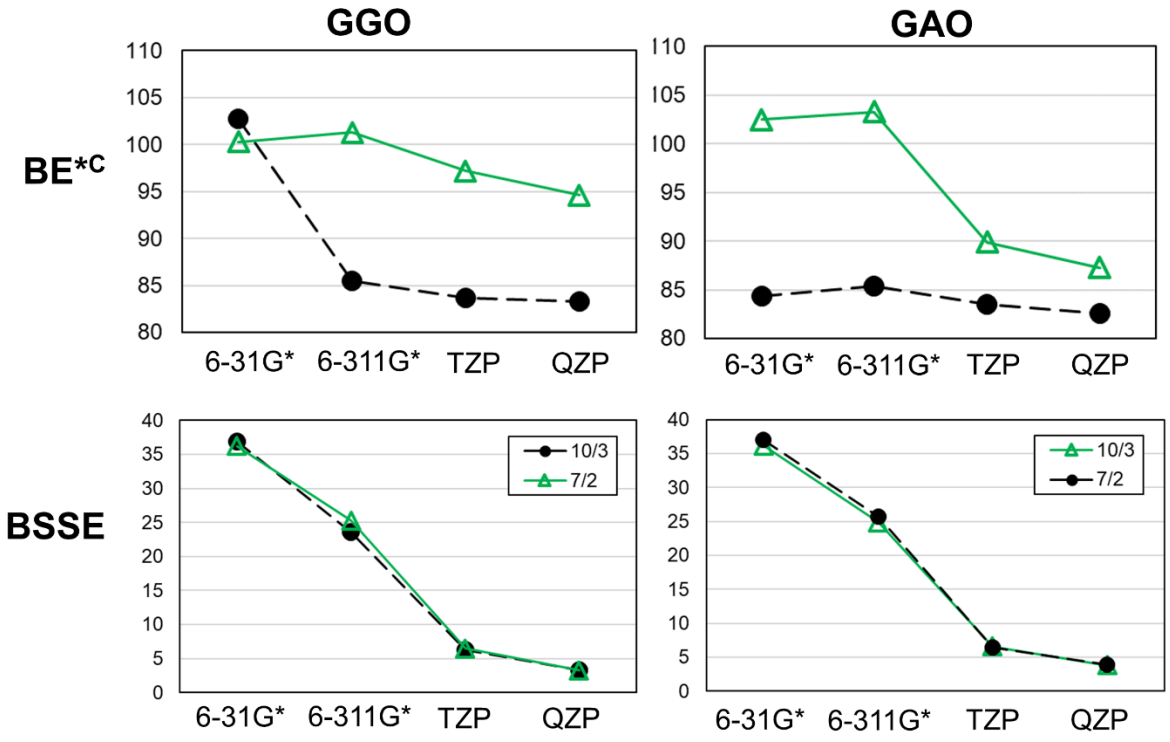


Figure 4. BE^{*C} (first row) and CP correction for BSSE (second row) in $kJ \cdot mol^{-1} \cdot triplet^{-1}$ by varying the BS quality at the B3LYP-D level. Results reported for GGO (first column) and GAO (second column) helices.

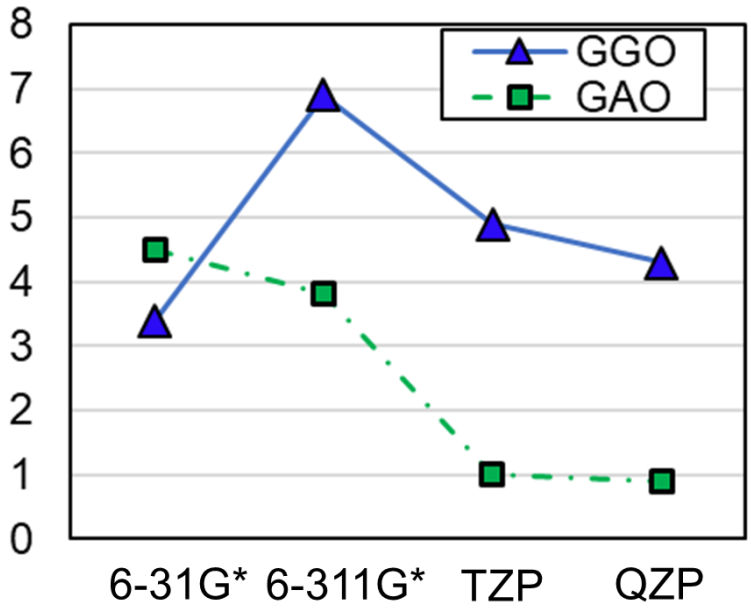


Figure 5. BS effect on ΔE-helix using the B3LYP-D functional. Positive values of ΔE-helix indicate that a 7/2 packing is preferred. Values in $kJ \cdot mol^{-1} \cdot triplet^{-1}$.

The Role of the Hamiltonian

To understand the role of the DFT functional in collagen simulations, we have tested some of the most commonly used DFT functionals for periodic and molecular simulations, e.g. the hybrid B3LYP-D and HSE06-D functionals, and the DFT PBE-D functional. We have also considered the role of dispersion, by the data computed with the B3LYP functional in the bunch of tested methods. Among the non-DFT Hamiltonians, we have included in the analysis the cheap HF-3c, HF-3c-027 and GFN1-xTB methods. The results analysis is carried out in line with the previous paper section, e.g. we will focus on the method's computational burden, computed rise per triplet value, and estimated energy values (BE^{*C} and ΔE-helix).

Regarding the computational burden for a SCF+G calculation, B3LYP/TZP, HSE06/TZP and PBE/TZP methods are respectively 32, 48 and 10 times slower than HF-3c. HF-3c is faster also than the B3LYP/6-31G* method by at least 4 times. Clearly, the semiempirical GFN1-xTB method is the fastest among the chosen Hamiltonians, being several orders of magnitude faster than HF-3c.

As for the geometry, the computed values of rise per triplet for the four collagen models as a function of the Hamiltonian type are reported in Figure 6. By analyzing the results, we can state the following:

- For the DFT-D methods, all functionals give quite similar results, with negligible differences from B3LYP-D and HSE06-D, and PBE-D predicting slightly more elongated helices than the hybrid functionals in agreement with previous work.[53–55]
- B3LYP functional computes over-elongated helices. This is in line with previous finding of Ref.[9] for a Gly-Pro-Pro collagen model, in which we demonstrated that dispersion interactions must be included in the simulation to improve the agreement with the experiments.
- HF-3c methods compute relative difference between 7/2 and 10/3 helices in good agreement with the experiments, as the computed mean value is 0.41 ± 0.02 Å which is fairly close to the experimental one of 0.34 Å. Furthermore, the dispersion scaled version of HF-3c (HF-3c-027) accurately matches the experiment, with a deviation lower than 0.1 Å.
- Regardless the adopted methodology, the GAO helices are more compressed than the GGO ones, e.g. with a smaller rise per triplet. To gain further insights on this, we have relaxed the geometry for straight polymers with a poly-proline type II geometry,[8] and GGO and GAO compositions. The computed unit cell lengths are 9.813 Å (GGO) and 9.522 Å (GAO), at the B3LYP-D level, which elongate to 9.966 Å (GGO) and 9.721 Å (GAO) at the B3LYP level. These results indicate that the compression due to the Ala presence does not depend on neither the triple helical organization of collagen nor the dispersion interactions. We believe that this indicates a structural propensity of Ala (bond lengths, bond angles and dihedral angles) to make shorter peptide chains than Gly.
- GFN1-xTB computes helical rise per triplet in good agreement with experimental values for the 7/2 helix. Unfortunately, for the 10/3 helices, the rise per triplet value almost collapses on to the 7/2 experimental value, making the packing of 10/3 and 7/2 helices very similar.

In Figure 7 we have compared the BE^{*C} computed for our models as a function of the Hamiltonian. The main findings are the followings:

- The B3LYP method underestimates the BE^{*C} compared to DFT-D methods. This underestimation depends on the lack of dispersion correction, which accounts for more than 50% of the inter-strands energy (for the B3LYP-D case), see Figure S11.

- The HF-3c and HF-3c-027 methods underestimate the BE^{*C} with respect to DFT-D methods. The reason is that the pure electrostatic interaction (HF/MINIX energy) is underestimated due to the adopted minimal basis set (Figure S10).
- All DFT-D methods predict the 7/2 helices to have higher inter-strands interactions than the 10/3 helices. Instead, pure DFT (no dispersion included), compute an inter-strands interaction similar for both helices. This is in line with the findings for the GPP case, see Ref.[9] On average, the computed BE^{*C} difference between 7/2 and 10/3 helices of 13 $kJ \cdot mol^{-1} \cdot triplet^{-1}$, comes from both dispersion interactions, (higher in a more compact 7/2 helix) and better inter-strand electrostatic contacts, see Figure S10.
- GFN1-xTB computes inter-strand energies that are generally in line with full hybrid and much more expensive DFT simulations. Interestingly, the GFN1-xTB results are in better agreement with DFT-D simulations than the more computationally demanding HF-3c method.

As for the preferred helical packing, for which the results are gathered in Figure 8, we can state that:

- The 7/2 helical packing is clearly favored for GGO collagens using any HF-D or DFT-D method, with small differences among methodologies. For GAO compositions, the 7/2 packing is only slightly more stable than the 10/3. This result indicates that increasing the size of the residue in the X position may induce collagen to pack into more elongated helices.
- The B3LYP method tends to over-stabilize elongated helices. This trend is in line with the results obtained for the GPP composition, see Ref. [9] As we have already pointed out, including the -D correction is crucial to have results in line with the experimental evidence.
- Interestingly, the B3LYP-D/TZP//HF-3c-027 (SP-B3LYP-D/TZP) method, which combines a geometry optimization carried out with the fast HF-3c-027 method and an energy estimation only at the more expensive B3LYP-D/TZP method, gives results quantitatively in line with those obtained by full DFT calculations.
- Plain GFN1-xTB slightly stabilizes the 10/3 helices more than the reference DFT-D. This tendency of GFN1-xTB leads to results in fair agreement with DFT-D for the GGO case. Conversely, for the GAO case, the 10/3 helix is found to be more stable than the 7/2 one. Using either a HF-3c or B3LYP-D single point energy estimation on the GFN1-xTB relaxed geometries, over stabilizes the 7/2 geometry, leading to

results in qualitative agreement with those obtained from full DFT-D simulations, with larger deviations. These deviations arise from the over-shrunk geometry of 10/3 helix, see Figure 6. To solve this problem, we relaxed the geometry of collagen keeping the unit cell length fixed to the ideal values of 7/2 and 10/3 helices, e.g. GFN1-xTB-cellfix method, see Figure 8. In this way, the results are in quantitative agreement with full DFT-D results.

- We have also explored the performance of the small basis set global hybrid functional PBEh-3c, its screened exchange variant HSE-3c and a generalized gradient approximated B97-3c functionals evaluated in a medium-sized basis set, all recently implemented in the CRYSTAL17 code. We have computed the relative stability of the 7/2 vs 10/3 helix for GAO and GGO polymers as a single point energy evaluation (SP) at the HF-3c-027 geometries. This class of methods extends the accuracy of the simplest HF-3c by increasing significantly the computer time, as reported in Ref.[56]. The time ratio for modeling large *molecular* complexes was: HF-3c(1):B97-3c(10):PBEh-3c(50) and HF-3c(1):HSE-3c(8):PBEh-3c(10) for a single *crystal* case. For the present case of collagen (*polymer*), the ratio is similar: HF-3c(1):PBEh-3c(6):HSE-3c(20):B97D-3c(38). The difference with the *molecular* case can arise from the different techniques adopted by molecular codes for the calculation of long-range effects in the evaluation of the bi-electronic integrals compared to the CRYSTAL17 implementation. Regarding the periodic case, the chosen integral tolerances and different periodicity can alter slightly the trend cost of the different methodologies. In the present case, the cost of B97D-3c is due to the long tails of the adopted triple zeta basis set which overcome the cost of handling the exact exchange in PBEh-3c. While it would be worth adopting in future work the PBEh-3c method, here we only focused on the relative stability of the 7/2 *vs* 10/3 helix for GAO and GGO polymers computed as a single point energy evaluation (SP) at the HF-3c geometries. The results are shown in the Figure S11 of the supplementary information and showed that the more sophisticated methods reduced the absolute stability of both biopolymers with respect to HF-3c, while maintaining the proper order of stability. The relative stability of about 2 $kJ \cdot mol^{-1} \cdot triplet^{-1}$ at HF-3c, becomes about 3 $kJ \cdot mol^{-1} \cdot triplet^{-1}$ for B97-3c, HSE-3c, PBEh-3c and increases up to about 4 $kJ \cdot mol^{-1} \cdot triplet^{-1}$ at B3LYP-D3.

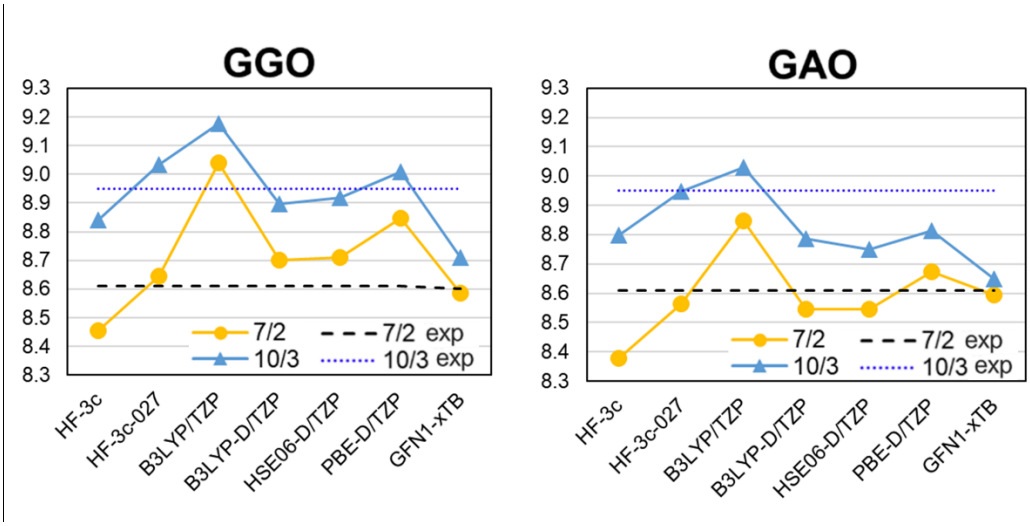


Figure 6. Effect of the methodology on the rise per triplet (Å) compared with experimental helices (exp). D = $D3^{ABC}$.

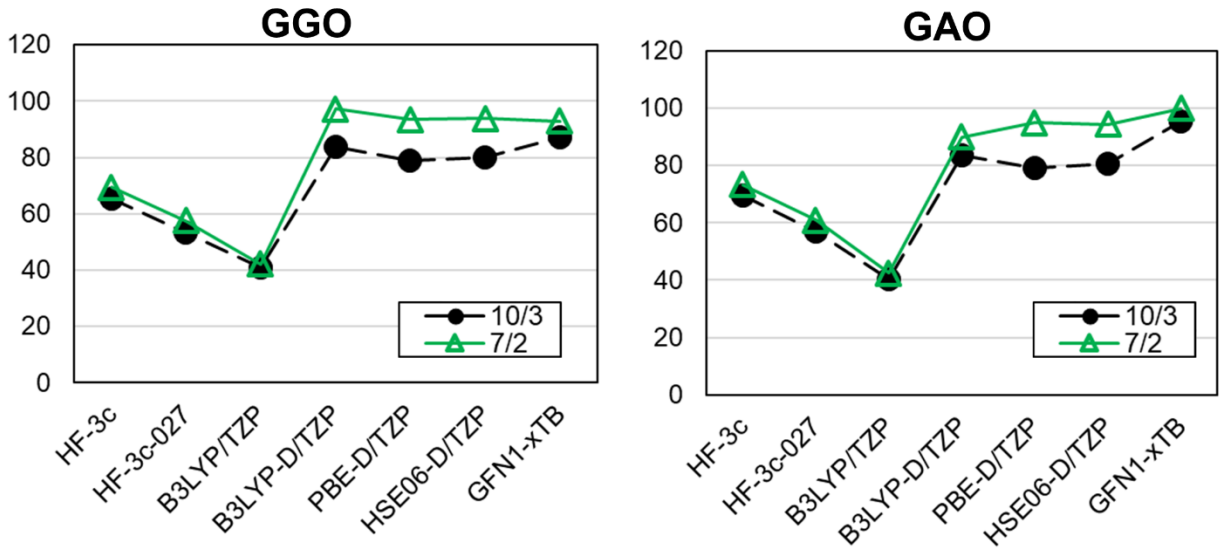


Figure 7. BE^{*C} (in $kJ \cdot mol^{-1} \cdot triplet^{-1}$) for GGO and GAO, with 7/2 and 10/3 helices and different methodology. D = $D3^{ABC}$.

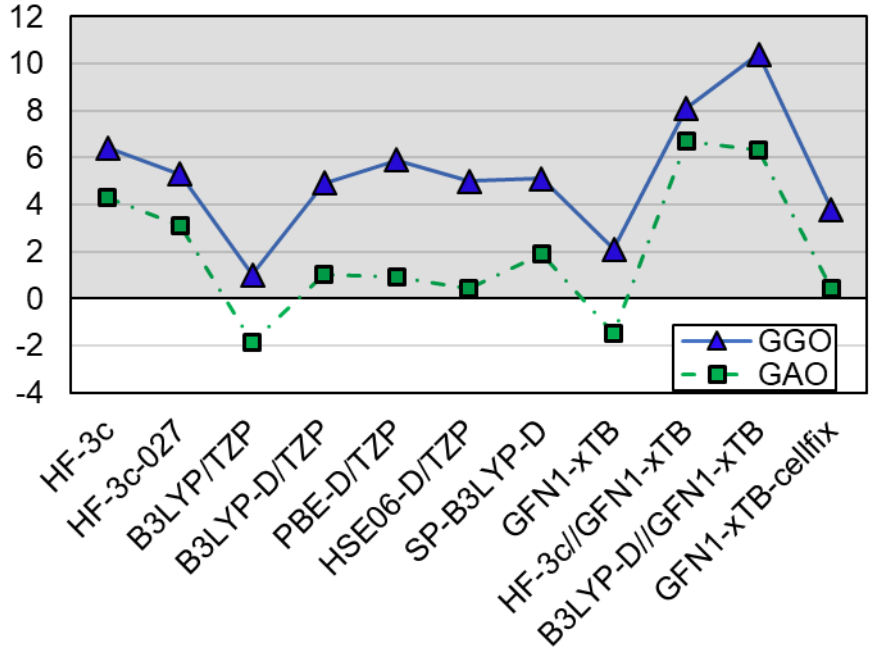


Figure 8. Effect of the methodology on the ΔE-helix (in kJ·mol^{-1}·triplet). D = D3^{ABC}. Positive value of ΔE-helix stabilizes the 7/2 helix over the 10/3 one (grey shaded area). The best reference is for B3LYP-D/TZP method.

Conclusions

In this work we have carefully checked the role of the quantum mechanical approach on computing structure, inter-strands binding energy and helical packing of Pro-rich collagen protein models. Our purpose is to establish a cost-effective approach to apply for a comprehensive study on the relationship between collagen structure and composition. This is of fundamental importance on both fundamental and applicative levels.

Within the DFT framework, we have chosen the most commonly employed DFT functionals in solid state and molecular simulations, i.e. the PBE functional and the two hybrid HSE06 and B3LYP functionals. The role of the basis set size is also analyzed, studying several types of basis sets (BSs), from contracted split valence type basis sets up to a large QZVP basis set. Within the cost-effective methods recently proposed, we check the HF-3c method and the semiempirical GFN1-xTB method. To our knowledge, the latter is for the first time applied to periodic biopolymers.

From a computational point of view, the main results of this work can be summarized as follows:

- The Ahlrichs TZVP outstands among the selected BSs, giving results in agreement with the more expensive QZVP (here considered as reference) at 1/3 of the computational cost. Unfortunately, the much cheaper Pople BSs give results which are not trustworthy, due to a significative BSSE. Even if all BSs compute helical propensity (ΔE-helix) in a similar way, the Pople BSs compute geometries and binding energies in strong disagreement with the reference. Therefore, we discourage the use of such BSs for collagen simulations and in general for organic systems in which non-covalent interactions are important.

- Regardless of the type of functional employed, all DFT-D approaches give coherent results for all the studied collagen features. The HF-3c method, underestimates the collagen inter-strand binding energy with respect to DFT-D methods. This may be due to the limited MINIX basis set encoded in HF-3c model chemistry. For error-balancing, HF-3c predictions on the helical features make only small deviations with respect to full DFT-D ones. These deviations are further reduced using energy estimations at the DFT-D level on relaxed HF-3c structures, e.g. DFT-D/HF-3c method. This hybrid method saves computational time of roughly one order of magnitude with respect to full DFT-D.

- Results obtained from GFN1-xTB are promising. It gives results in line with DFT-D, if the cell optimization is neglected. In principle, this approach is several orders of magnitude faster than the full DFT-D. Worth further future testing are the other flavors of the GFN-xTB method which currently are not implemented or still have limitations for computing systems with periodic boundary conditions.

Considering the computational advantage of using symmetry in the simulation, we suggest to use the DFT-D//HF-3c method for estimation of the helical propensity of homo-trimeric collagens, such as the one employed here. For hetero-trimetric collagens, in which no symmetry can be exploited, we suggest the GFN1-xTB approach, with the recipe of keeping the cell parameter fixed at the ideal rise per triplet of the 7/2 and 10/3 helices. These results are coherent with those obtained from DFT-D//DFT-D with a computational gain of several order of magnitude. As for the choice of the DFT functional, we suggest the PBE functional as the best accuracy-cost compromise.

Finally, from a biological point of view, we can state that GGO collagens prefer to pack in a tight helix, e.g. 7/2 helix. Similarly, GAO collagens seem to prefer the tight 7/2 helical

packing, but the energy stability of the 7/2 geometry is close to the 10/3 one. This suggests that increasing the size of the residue in the X position of the collagen triplet may induce collagen to pack into more loose and elongated helices to make room for larger lateral chains. A following work is under preparation on Gly-X-Hyp collagen triple helices with even larger residues in X, such as Leucine and Phenylalanine, to further explore this hypothesis. We believe that the present results provide computational recipes to allow exploring the helical propensity of collagen as a function of the residue content by a large screening workflow.

Associated Content

Supporting information content:

Table S1: 6-31G* basis set in CRYSTAL17 basis set format. Table S2: 6-311G* basis set in CRYSTAL17 basis set format. Table S3: TZP basis set in CRYSTAL17 basis set format. Table S4: QZP basis set in CRYSTAL17 basis set format. Figure S1: Torsional angles and electrostatic contacts definitions used in this work. Table S5: Inter-chains contacts for GGO 7/2 model obtained at the PBE-D2/QZP and PBE-D2/PW levels. Figure S2: Effect of basis set on the considered inter-chain contacts (EC1-5) between GGO COL strands. Figure S3: Effect of basis set on the considered inter-chain contacts (EC1-5) between GAO COL strands. Figure S4: Effect of BS on the torsional angle of X=Gly and Y=Hyp compared with the experimental average value on collagens. Figure S5: Effect of BS on the torsional angle of X=Ala and Y=Hyp compared with the experimental average value on collagens. Figure S6: Effect of the method on the considered inter-chain contacts (EC1-5) between GGO strands. Figure S7: Effect of the method on the considered inter-chain contacts (EC1-5) between GAO strands. Figure S8: Effect of the method on the dihedral angles of GGO collagen. Figure S9: Effect of the computational method on the dihedral angles of GAO collagen. Figure S10: Effect of the method on the BE^{*C} (in kJ·mol-1·triplet-1) of GGO and GAO collagen helices. Figure S11. Effect on ΔE-helix using different functionals. Table S6: Geometry, e.g. dihedral angles and electrostatic contacts, of the GGO-7/2 and GGO-10/3 triple helices relaxed using different methods. Table S7: BSSE corrected binding energy and BSSE correction for the GGO-7/2 and GGO-10/3 triple helices. Table S8: Evaluation of the energy differences (ΔE-helix) and contributions (ΔBE* and ΔE-ss, with ΔE-helix = -ΔBE* +ΔE-ss) between the 7/2 and 10/3 conformations of the GGO triple helix with different computational methods. Table S9: Geometry, e.g. dihedral angles and electrostatic contacts (see Figure S1), of the GAO-7/2 and GAO-10/3 triple helices relaxed using different methods. Table S10: BSSE corrected binding energy and BSSE corrections for the GAO-

7/2 and GAO-10/3 triple helices. Table S11: Evaluation of the energy differences (ΔE-helix) and contributions (ΔBE* and ΔE-ss, with ΔE-helix = -ΔBE* + ΔE-ss). Table S12: Effect of integral tolerance parameters on the collagen model geometry. Table S13: Effect of integral tolerance parameters on the collagen model BE*.

Author Information

Corresponding authors: Dr. Michele Cutini and Prof. Piero Ugliengo

NAME: ORCID

Dr. Michele Cutini: 0000-0001-6896-7005

Dr. Marta Corno: 0000-0001-7248-2705

Prof. Piero Ugliengo: 0000-0001-8886-9832

Dr. Irene Bechis: 0000-0001-7322-2446

Acknowledgments

Michele Cutini acknowledges the University of Torino for funding (grant agreement No CHI.2019.21/XXII), Massimo Bocus, M.Sc. for the efforts devoted to support the “Collagen Project”. PU acknowledges the generous allowance of CINECA computing time from ISCRA B (Project: ISB16; Account ID: MACBONE, Origin ID: HP10BAL7D8), C3S Competence Centre for scientific computing of the University of Torino and the CRYSTAL team for continuous support in the usage of the CRYSTAL program.

References

(1) Fratzl, P. *Collagen Structure and Mechanics - Unknown*; Fratzl, P., Ed.; Springer Berlin / Heidelberg: Postdam, Germany, 2008.

(2) Crick, F. H. C.; Rick, A. Structure of Polyglycine II. *Nature* **1955**, *176*, 780–781.

(3) Arnott, S.; Dower, S. D. The Structure of Poly-l-Proline II. *Acta Cryst. B* **1968**, *24*, 599–601.

(4) 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.* **1998**, *122* (1–2), 86–91.

(5) Bella, J. Collagen Structure: New Tricks from a Very Old Dog. *Biochem. J.* **2016**, *473* (8), 1001–1025.

(6) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu. Rev. Biochem.* **2009**, *78* (1), 929–958.

(7) Sakakibara, S.; Kishida, Y.; Okuyama, K.; Tanaka, N.; Ashida, T.; Kakudo, M. Single Crystals of (Pro-Pro-Gly)10, a Synthetic Polypeptide Model of Collagen. *J. Mol. Biol.* **1972**, *65* (2), 371–373.

(8) Cutini, M.; Bocus, M.; Ugliengo, P. Decoding Collagen Triple Helix Stability by Means of Hybrid DFT Simulations. *J. Phys. Chem. B* **2019**, *123*, 7354−7364.

(9) Cutini, M.; Pantaleone, S.; Ugliengo, P. Elucidating the Nature of Interactions in Collagen Triple-Helix Wrapping. *J. Phys. Chem. Lett.* **2019**, *10*, 7644–7649.

(10) 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.* **2010**, *170* (2), 377–391.

(11) Orgel, J. P. R. O.; Persikov, A. V.; Antipova, O. Variation in the Helical Structure of Native Collagen. *PLoS One* **2014**, *9* (2), 1–11.

(12) 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.* **2016**, *7* (2), 1496–1507.

(13) Christensen, A. S.; Kubař, T.; Cui, Q.; Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **2016**, *116* (9), 5301–5337.

(14) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.

(15) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the

Exchange Screening Parameter on the Performance of Screened Hybrid Functionals. *J. Chem. Phys.* **2006**, *125* (22), 224106.

(16) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.

(17) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.

(18) 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* **1988**, *37* (2), 785–789.

(19) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685.

(20) Cutini, M.; Civalleri, B.; Corno, M.; Orlando, R.; Brandenburg, J. G.; Maschio, L.; Ugliengo, P. Assessment of Different Quantum Mechanical Methods for the Prediction of Structure and Cohesive Energy of Molecular Crystals. *J. Chem. Theory Comput.* **2016**, *12* (7), 3340–3352.

(21) Cutini, M.; Corno, M.; Ugliengo, P. Method Dependence of Proline Ring Flexibility in the Poly-L-Proline Type II Polymer. *J. Chem. Theory Comput.* **2017**, *13*, 370−379.

(22) Cutini, M.; Civalleri, B.; Ugliengo, P. Cost-Effective Quantum Mechanical Approach for Predicting Thermodynamic and Mechanical Stability of Pure-Silica Zeolites. *ACS Omega* **2019**, *4*, 1838−1846.

(23) Cutini, M.; Maschio, L.; Ugliengo, P. Exfoliation Energy of Layered Materials by DFT-D: Beware of Dispersion! *JCTC*.

(24) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All Spd-Block Elements (Z = 1−86). *J. Chem. Theory Comput* **2017**, *13*, 1989−2009.

(25) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended Tight-Binding Quantum Chemistry Methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, No. April, 1–49.

(26) Hehre, W. J.; Ditchfield, K.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56* (5), 2257–2261.

(27) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets for Atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571–2577.

(28) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829.

(29) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(30) Dovesi, R.; Erba, A.; Orlando, R.; Zicovich-Wilson, C. M.; Civalleri, B.; Maschio, L.; Rérat, M.; Casassa, S.; Baima, J.; Salustro, S.; et al. Quantum-Mechanical Condensed Matter Simulations with CRYSTAL. *WIREs Comput Mol Sci.* 2018, pp 1–36.

(31) Brandenburg, J. G.; Alessio, M.; Civalleri, B.; Peintinger, M. F.; Bredow, T.; Grimme, S. Geometrical Correction for the Inter- and Intramolecular Basis Set Superposition Error in Periodic Density Functional Theory Calculations. *J. Phys. Chem. A* **2013**, *117* (38), 9282–9292.

(32) 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.* **2010**, *132* (15), 154104–154119.

(33) Muto, Y. The Force Between Nonpolar Molecules. *Proc. Phys. Math. Soc. Japan* **1944**, *17*, 629–31.

(34) Axilrod, B. M.; Teller, E. Interaction of the van Der Waals Type Between Three Atoms. *J. Chem. Phys.* **1943**, *11*, 299–300.

(35) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A Generally Applicable Atomic-Charge Dependent London Dispersion Correction. *J. Chem. Phys.* **2019**, *150* (15).

(36) Broyden, C. G. The Convergence of a Class of Double-Rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.* **1970**, *6* (1), 76–90.

(37) Fletcher, R. A. New Approach to Variable Metric Algorithms. *Comput. J.* **1970**, *13*, 317–322.

(38) Shanno, D. F.; Kettler, P. C. Optimal Conditioning of Quasi-Newton Methods. *Math. Comput.* **1970**, *24* (111), 657–664.

(39) Pulay, P. Improved SCF Convergence Acceleration. *J. Comput. Chem.* **1982**, *3* (4), 556–560.

(40) Dovesi, R.; Saunders, V. R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Pascale, F.; Civalleri, B.; Doll, K.; Harrison, N. M.; Bush, I. J.; et al. *CRYSTAL17 User's Manual*; Università di Torino: Torino, Italy, 2017.

(41) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamcis for Liquid Metals. *Phys. Rev. B* **1993**, *47* (1), 558.

(42) Kresse, G.; Furthmüller, J.; Hafner, J. Ab Initio Molecular-Dynamics Simulation of the Liquid-Metal–Amorphous-Semiconductor Transition in Germanium. *Phys. Rev. B* **1996**, *6* (1), 558–561.

(43) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1996**, *54* (16), 11169–11186.

(44) Kresse, G.; Furthmiiller, J. Water News Roundup. *J. / Am. Water Work. Assoc.* **2004**, *96* (10), 14–20.

(45) Grimme, S. Semiempirical GGA-Type Density Functional Constructed with a Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799.

(46) Kohlmeyer, A.; Mundy, C. J.; Mohamed, F.; Schiffmann, F.; Tabacchi, G.; Forbert, H.; Kuo, W.; Hutter, J.; Krack, M.; Iannuzzi, M.; et al. CP2K. 2004.

(47) Vandevondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a Mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167* (2), 103–128.

(48) Ugliengo, P.; Viterbo, D.; Chiari, G. MOLDRAW: Molecular Graphics on a Personal Computer. *Z. Krist.* **1993**, *207*, 9–23.

(49) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (1), 33–38.

(50) DeRider, M. L.; Wilkens, S. J.; Waddell, M. J.; Bretscher, L. E.; Weinhold, F.; Raines, R. T.; Markley, J. L. 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.* **2002**, *124* (11), 2497–2505.

(51) Panasik, N.; Eberhardt, E. s.; Edison, A. s.; Powell, D. r.; Raines, R. t. Inductive Effects on the Structure of Proline Residues. *Int. J. Pept. Protein Res.* **1994**, *44* (3), 262–269.

(52) Cutini, M.; Corno, M.; Costa, D.; Ugliengo, P. How Does Collagen Adsorb on Hydroxyapatite? Insights from Ab Initio Simulations on a Polyproline Type II Model. *J. Phys. Chem. C* **2019**, *123* (13).

(53) 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. **2015**, *143*, 054107.

(54) Piccardo, M.; Penocchio, E.; Puzzarini, C.; Biczysko, M.; Barone, V. Semi-Experimental Equilibrium Structure Determinations by Employing B3LYP/SNSD Anharmonic Force Fields: Validation and Application to Semirigid Organic Molecules. *J. Phys. Chem. A* **2015**, *119* (10), 2058–2082.

(55) Witte, J.; Goldey, M.; Neaton, J. B.; Head-Gordon, M. Beyond Energies: Geometries of Nonbonded Molecular Complexes as Metrics for Assessing Electronic Structure Approaches. *J. Chem. Theory Comput.* **2015**, *11* (4), 1481–1492.

(56) Caldeweyher, E.; Brandenburg, J. G. Simplified DFT Methods for Consistent Structures and Energies of Large Systems. *J. Phys. Condens. Matter* **2018**, *30* (21), aabcfb.