

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Elucidating the Nature of Interactions in Collagen Triple-Helix Wrapping

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/1728925> since 2020-02-19T20:33:10Z

Published version:

DOI:10.1021/acs.jpcllett.9b03125

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

Elucidating the Nature of Interactions in Collagen Triple Helix Wrapping

Michele Cutini, Stefano Pantaleone, and Piero Ugliengo

J. Phys. Chem. Lett., **Just Accepted Manuscript** • DOI: 10.1021/acs.jpcllett.9b03125 • Publication Date (Web): 18 Nov 2019

Downloaded from pubs.acs.org on November 24, 2019

Just Accepted

“Just Accepted” manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides “Just Accepted” as a service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. “Just Accepted” manuscripts appear in full in PDF format accompanied by an HTML abstract. “Just Accepted” manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are citable by the Digital Object Identifier (DOI®). “Just Accepted” is an optional service offered to authors. Therefore, the “Just Accepted” Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the “Just Accepted” Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these “Just Accepted” manuscripts.

Elucidating the Nature of Interactions in Collagen Triple

Helix Wrapping

Michele Cutini,^{a} Stefano Pantaleone,^b and Piero Ugliengo^{a*}*

^aUniversity of Torino, Department of Chemistry and NIS (Nanostructured Interfaces and Surfaces) Center, Via P. Giuria 7, 10125 Turin - Italy

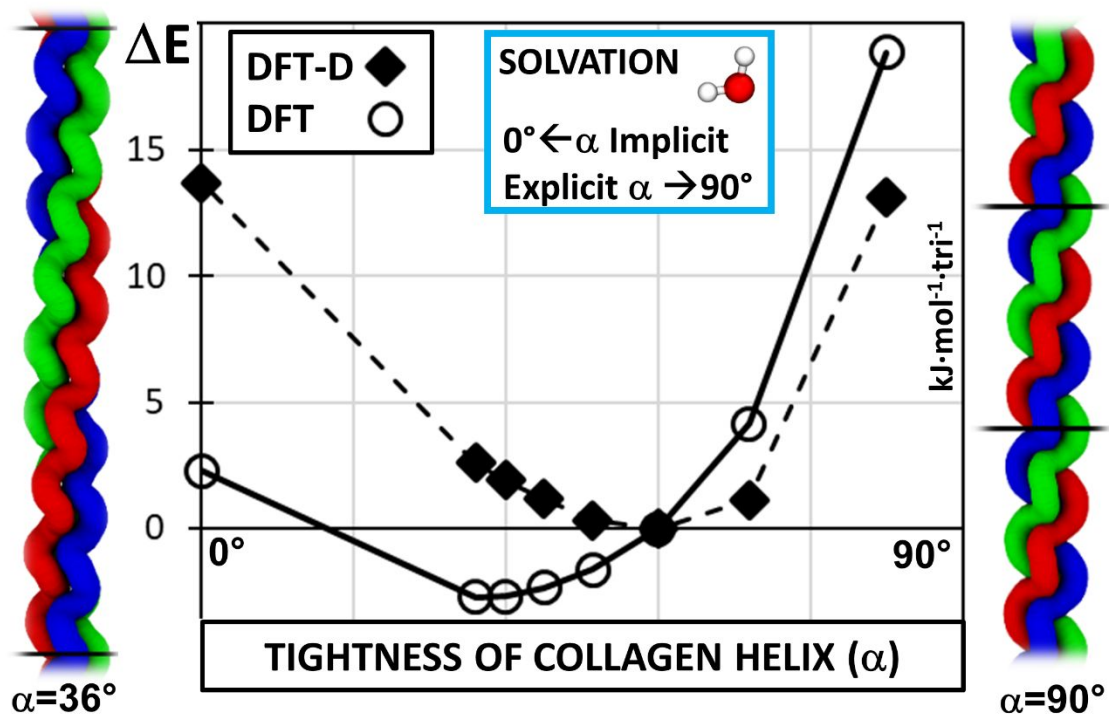
^bGrenoble Alps University, CNRS, Institute of Planetary Sciences and Astrophysics, Grenoble (IPAG), 38000 Grenoble - France

*Corresponding Authors: michele.cutini@unito.it, piero.ugliengo@unito.it

1
2
3
4 Collagen is the most abundant protein family in the animal kingdom. Its structural motif
5
6
7 envisages three polypeptide chains coiled in the so-called collagen triple helix. Depending
8
9
10 on the triplet amino acidic sequence of the chains, collagen has different helical
11
12
13 arrangements. Such atomic-scale structural variations have large impact on the large-scale
14
15
16 structure of collagen. In this letter, we elucidate the interactions responsible of a specific
17
18
19 helical pattern of the collagen protein by means of DFT-D based computer simulations. We
20
21
22 demonstrate that inter-chains interactions and solvation effects stabilize compact helices
23
24
25 over elongated ones. Conversely, elongated helices are stabilized by less geometrical strain
26
27
28 and entropic factors. Our computational procedure predicts the collagen helical pattern in
29
30
31 agreement with the experimental evidences.
32
33
34
35
36
37
38

39 TOC GRAPHICS

40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



KEYWORDS: collagen, triple helix, DFT, micro-solvation, dispersion interactions

1
2
3 Protein structure prediction is a broadly studied problem since long ago.¹⁻⁴ In particular, due
4
5
6
7 to the structure-property one-to-one correspondence, its understanding at atomistic level is
8
9
10 very useful in the protein engineering field. Many factors contribute to the protein structure:
11
12
13 amino acidic sequence, temperature, pressure, pH, presence and nature of the solvent,
14
15
16
17 intra- and inter-molecular interactions.^{1,5,6} Due to the large size of proteins, the multiple weak
18
19
20
21 interactions (H-bonds and dispersive forces) both within the protein itself, and with the
22
23
24 solvent, play a delicate role in determining the protein geometry.⁷ The experimental
25
26
27
28 approach to elucidate the weight of each component of these interactions is hindered by
29
30
31 their intermingled nature. In this note, we show that modern density functional theory based
32
33
34
35 on accurate hybrid functionals and large variational basis set is capable to cope with this
36
37
38 problem.
39
40

41
42 Among all proteins, collagen is the main structural protein in mammals. It is found in skin,
43
44
45 tendons, cartilage and bones, with the role of providing mechanical support to tissues.
46
47
48
49 Collagen has a triple helical structure, in which three parallel polypeptide strands wrap
50
51
52 together.⁸⁻¹⁰ This geometrical organization imposes severe condition on the protein amino
53
54
55
56 acidic content. Indeed, the amino acidic sequence rests on a specific triplet pattern
57
58
59
60

1
2
3 characterized by the presence of Glycine (Gly) every three amino acids (G-X-Y).^{11,12} Proline
4
5
6
7 (Pro) and (2S,4R)-4-Hydroxyl-Proline (Hyp) are the most common residues and occupy
8
9
10 positions X and Y of the G-X-Y triplet, respectively. Therefore, Gly-Pro-Hyp is the most
11
12
13 frequent amino acidic triplet in collagen.¹³
14
15
16

17 Collagen can organize in different helical geometries depending on the amino acidic
18
19
20 sequence.¹¹ The helicity-composition relationship has been a matter of debate since long
21
22
23 time ago.¹⁴ A high amount of experimental evidences have proven that, for high content of
24
25
26 Pros and derivatives, collagen exhibits a 7/2 helicity.^{15,16} Conversely, in Pro-free zone,
27
28
29 collagen seems to prefer a 10/3 helix.¹⁷ The former case is a tight helix, in which 7 residues
30
31
32 fit into two helical turns, thus there are 3.5 residues per turn, with amino acidic triplet rotation
33
34
35 angle $\alpha = 51.4^\circ$. The latter is a loose helix, in which 10 residues fit into three helical turns
36
37
38 (3.33 residues per turn with $\alpha = 36.0^\circ$). Even if the differences between these helices seem
39
40
41 to be small at the atomic scale, they affect significantly the long triple helical domains.¹¹
42
43
44
45
46
47

48 The purpose of the present letter is to clarify the interactions responsible to bring collagen
49
50
51 in a specific helical geometry. This is of interest from both fundamental and applied points
52
53
54 of view. Indeed, the present work also aims to provide useful insights for collagen protein
55
56
57
58
59
60

1
2
3 engineering.¹⁷⁻²⁶ To achieve our goals we rely on computer simulations, based on DFT and
4
5
6
7 HF theories (see computational method section and SI for further details). We have analyzed
8
9
10 seven collagen helices, which differ for the amino acidic triplet rotational angle (α), i.e. $\alpha =$
11
12
13
14 90° , 72° , 60° , 51.4° , 40° , 36° and 0° . The higher the α value, the tighter the collagen helix is
15
16
17 (see Figure 1). As shown in Figure 1, we have used a periodic model, in which the chosen
18
19
20 motif is repeated infinitely. The advantages and details of this particular choice to represent
21
22
23
24 collagen compared to finite models has been recently discussed in Ref. 27.
25
26

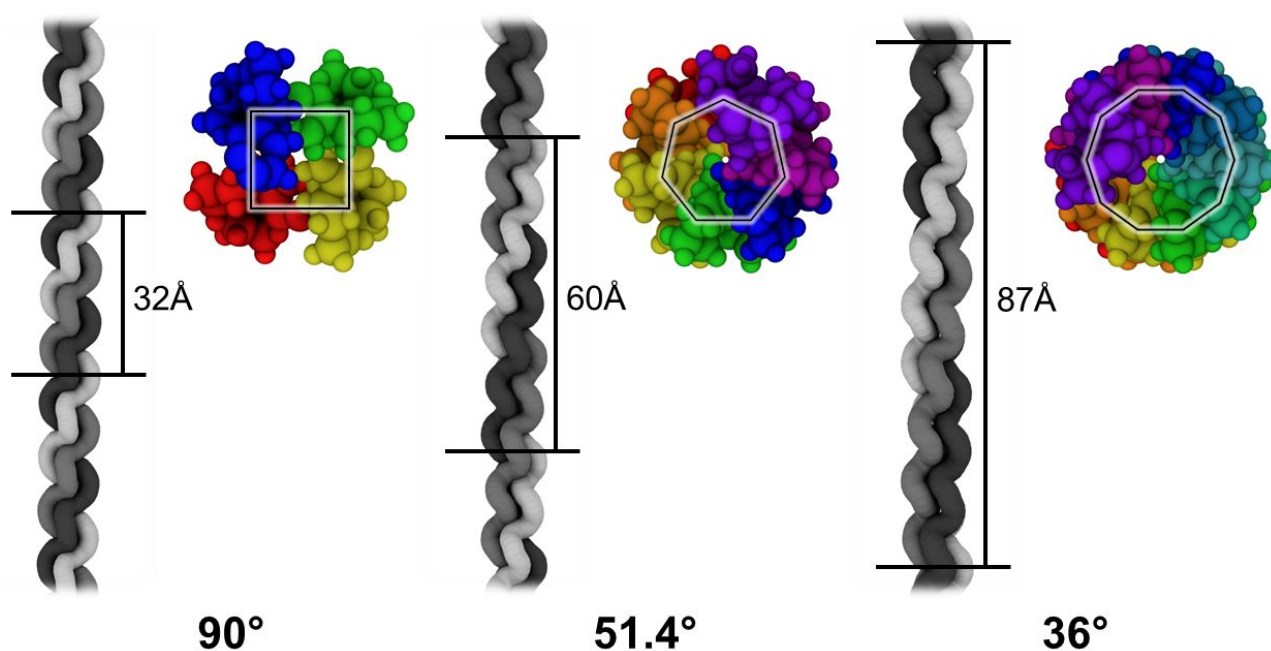


Figure 1. Collagen models with $\alpha = 90^\circ$, 51.4° and 36° . *Lateral view*: tube representation in different grey levels for each collagen single strand. The black segments delimitate the unit cell of the model. *Top view*: a single protein strand is reported with all amino acidic triplet

1
2
3 within the unit cell depicted in different colors. The geometrical shapes follow the triplets
4
5
6
7 wrapping.
8
9

10
11 For the $\alpha = 0^\circ$ case, the three collagen strands are no longer wrapped together, as they
12
13 stand parallel to each other, while keeping the peculiar collagen inter-strand H-bond pattern.
14
15
16
17

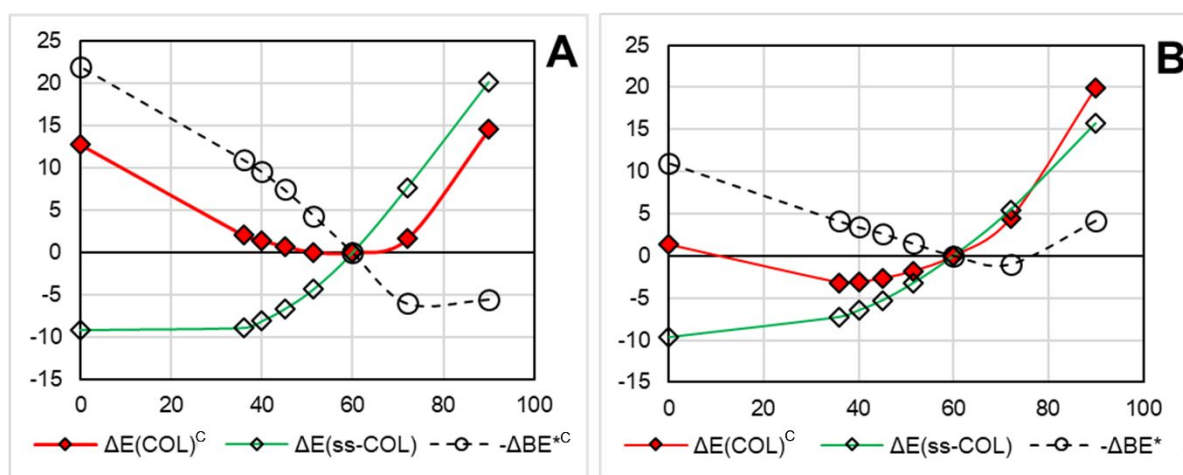
18 Regarding the composition, we have simulated homo-trimeric collagens, in which each
19
20 strand is made of the repetition of the Gly-Pro-Pro triplet. We chose this composition
21
22 because it resembles the most diffuse one in collagens, i.e. Gly-Pro-Hyp, but it is simpler to
23
24 handle, see Ref. 27 for further details and a precise description of the models. As we want
25
26 to compare the reasons of the relative energy stability, $\Delta E(\text{COL})^{\text{C}}$, between different
27
28 collagen wrappings, ultimately controlled by the α values, we define the following
29
30 relationship:
31
32
33
34
35
36
37
38
39
40
41
42

$$\Delta E(\text{COL})^{\text{C}} = E(\text{COL}_\alpha)^{\text{C}} - E(\text{COL}_{60.0^\circ})^{\text{C}} = \Delta E(\text{ss-COL}) - \Delta \text{BE}^{\text{C}} \quad (1)$$

43
44
45
46 Where $E(\text{COL}_\alpha)^{\text{C}}$ is the BSSE-corrected (“C” apex) energy of a fully relaxed collagen triple
47
48 helix (COL_α) with triplet rotation equal to α minus the energy $E(\text{COL}_{60.0^\circ})^{\text{C}}$ of the COL with
49
50 $\alpha = 60.0^\circ$ set as a reference state. The $\text{COL}_{60.0^\circ}$ structure is, indeed, the most stable wrapped
51
52 conformation with the whole set of the quantum mechanical methods here adopted (see
53
54
55
56
57
58
59
60

points i-ii) in the computational method section). As there are no conflicting results, we do not discuss here the details, which can be found in the SI (see Figure S2 and S3 and Table S4). Therefore, from now on, the discussion will be focused only on the results obtained at the B3LYP-D3^{ABC}/VTZP levels, with some comparison with B3LYP/VTZP to assess the role of dispersion.

As anticipated, the $\Delta E(\text{COL})^{\text{C}}$ has a minimum value for $\alpha = 60^\circ$, see Figure 2A, in good agreement with the experimentally reported value of $\alpha = 51.4^\circ$.¹¹ The discrepancy between the two cases ($\alpha = 60^\circ$ and 51.4°) are fully justifiable, as the $\Delta E(\text{COL})^{\text{C}}$ minimum is very shallow, with energy difference between the two structures of only $0.04 \text{ kJ mol}^{-1} \text{ triplet}^{-1}$.



1
2
3 Figure 2. Energy decomposition of $\Delta E(\text{COL})^{\text{C}}$ (vertical axis, kJ mol^{-1} triplet⁻¹) vs α (horizontal
4 axes, degrees). **A:** Results at B3LYP-D3^{ABC}/VTZP level. **B:** Results at B3LYP/VTZP level.
5
6
7
8

9
10 Lines reported to guide the readers eye only.
11
12
13

14 $\Delta E(\text{COL})^{\text{C}}$ can also be dissected following the second term of equation 1), in which $\Delta E(\text{ss-}$
15 $\text{COL}) = E(\text{ss-COL}_{\alpha} // \text{COL}_{\alpha}) - E(\text{ss-COL}_{60.0} // \text{COL}_{60.0})$ is the energy difference of a single
16
17 collagen strand, fixed at the geometry assumed in the triple helix structure for an angle α ,
18
19 and the corresponding value for the collagen structure at $\alpha = 60.0^{\circ}$ (the minimum). $\Delta E(\text{ss-}$
20
21 $\text{COL})$ is, therefore, a measure of the geometrical distortions of a single collagen strand due
22
23 to a different angle wrapping compared to the most stable one. Its nature is obviously
24
25 dominated by covalent intra-strand contributions (bond distances, angles and torsions
26
27 changes). The term $\Delta \text{BE}^{*\text{C}}$ is the difference between $\text{BE}_{\alpha}^{*\text{C}} = E(\text{ss-COL}_{\alpha} // \text{COL}_{\alpha}) -$
28
29 $E(\text{COL}_{\alpha} // \text{COL}_{\alpha})$ and the corresponding $\text{BE}_{60.0}^{*\text{C}}$ one. It measures the BSSE-corrected inter-
30
31 strands interaction energy within a triple helix with wrapping angle α with respect to the
32
33 energy of the most stable one (wrapping angle equal to 60.0°). Interestingly, within the
34
35 $\Delta E(\text{COL})^{\text{C}}$ these two terms act against each other: indeed, $\Delta E(\text{ss-COL})$ favors less packed
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 structures with less twisted geometries at variance with ΔBE^{*C} which favors more packed
4
5
6
7 ones (see Fig. 2).
8
9

10 We also split the pure DFT contribution (electrostatic, exchange, polarization and charge
11
12 transfer) from the dispersive (D) contributions from $\Delta E(ss-COL)$ and ΔBE^{*C} terms. The
13
14
15
16
17 results can be summarized as:
18
19

- 20
21 1) the ΔBE^{*C} term is dominated by the dispersion interactions (Figure S4) in good
22
23 correlation with the compactness of the helix (Figure S5). The pure DFT contribution
24
25 to BE^{*C} comes from the interactions of N-H and C-H groups with the C=O group of Pro
26
27
28 in X position within the protein core (Figure S6).
29
30
31
32
- 33
34 2) the $\Delta E(ss-COL)$ term is dominated by the DFT contribution (Figure S4), mainly arising
35
36 from the dihedrals, angles and bond distances deformation occurring in the collagen
37
38
39
40
41
42 helix.
43
44

45 Therefore, if we exclude the D term from our computational set-up, *i.e.* running the
46
47 calculations with the bare B3LYP method (or any other pure GGA/hybrid functional), the
48
49 $\Delta E(ss-COL)$ term will dominate the $\Delta E(COL)$ expression, leading to an artificial stabilization
50
51
52 of more elongated helical geometries. This is clearly shown in Figure 2B, in which the
53
54
55
56
57
58
59
60

1
2
3 B3LYP/VTZP simulation, leads to a shift of the minimum region from $\alpha = 60^\circ$ to $0^\circ < \alpha < 36^\circ$.
4
5

6
7 Therefore, lowering inter-strand interactions loosens the collagen triple helix structure and
8
9
10 *vice versa*.
11
12

13
14 To better mimic the collagen environment, we should take into account dynamic and
15
16
17 solvent effects. Within the quantum mechanical framework, they can be rigorously included
18
19
20 with *Ab-Initio* Molecular Dynamics (AIMD) simulations in explicit solvent. Unfortunately, this
21
22
23 is undoable at present, due to the large size of the protein models. A possible and popular
24
25
26 alternative, is to rely on classical Force-Field Molecular Dynamics (FFMD) simulations.
27
28
29 Unfortunately, the very tiny energy differences characterizing the various helix
30
31
32 conformations (see Figure 2A-B) are at the limits of the FFMD accuracy.
33
34
35

36
37
38 We choose a more pragmatic and simpler approach to keep the cost of the calculations
39
40
41 reasonable while enforcing good accuracy: i) dynamical effects are accounted for by
42
43
44 classical statistical thermodynamic based on the harmonic approximation of the vibrational
45
46
47 contributions, which define the entropic contribution to the Gibbs free energy. As the
48
49
50 calculation of the whole set of frequencies is time consuming for large structures, we relied
51
52
53 on the fast HF-3c method,²⁸ ii) solvent effects are estimated through the Polarizable
54
55
56
57
58
59
60

1
2
3
4 Continuum Model (PCM) which brings to the free energy $G(\text{COL})_{\text{sol}}$ definition, computed at
5
6
7 room T and in water solvent:
8
9

$$10 \quad G(\text{COL})_{\text{sol}} = E(\text{COL})^{\text{C}} + \text{HYD} + \text{ZPE} + \text{H} - \text{TS} \quad (2)$$

11
12
13 Here, the $E(\text{COL})^{\text{C}}$ is the BSSE-free total energy of collagen (*vide supra*), the HYD term
14
15
16 is the interaction energy of the protein with the continuum medium along with the cavitation
17
18
19 energy, the ZPE term is the protein vibrational zero-point energy computed within the
20
21
22 harmonic approximation, the H term is the protein thermal vibrational energy correction at
23
24
25 room T and S is the protein vibrational entropic contribution to the Gibbs free energy. The
26
27
28 vibrational quantities are computed correcting the HF-3c harmonic frequencies with a 0.86
29
30
31 scaling factor, as suggested in the original paper.²⁸ The relative collagen free energy,
32
33
34
35
36
37
38 $\Delta G(\text{COL})_{\text{sol}}$, is defined along the line of eq. (1) as:
39
40

$$41 \quad \Delta G(\text{COL})_{\text{sol}} = G(\text{COL}_{\alpha})_{\text{sol}} - G(\text{COL}_{60.0^{\circ}})_{\text{sol}} \quad (3)$$

42
43
44 and following eq. (2), $\Delta G(\text{COL})_{\text{sol}}$ can be decomposed in its components:
45
46
47

$$48 \quad \Delta G(\text{COL})_{\text{sol}} = \Delta E(\text{COL})^{\text{C}} + \Delta \text{HYD} + \Delta \text{ZPE} + \Delta \text{H} - \text{T}\Delta \text{S} \quad (4)$$

49
50
51 The terms of eq. (4) are shown as a function of α in Figure 3.
52
53
54
55
56
57
58
59
60

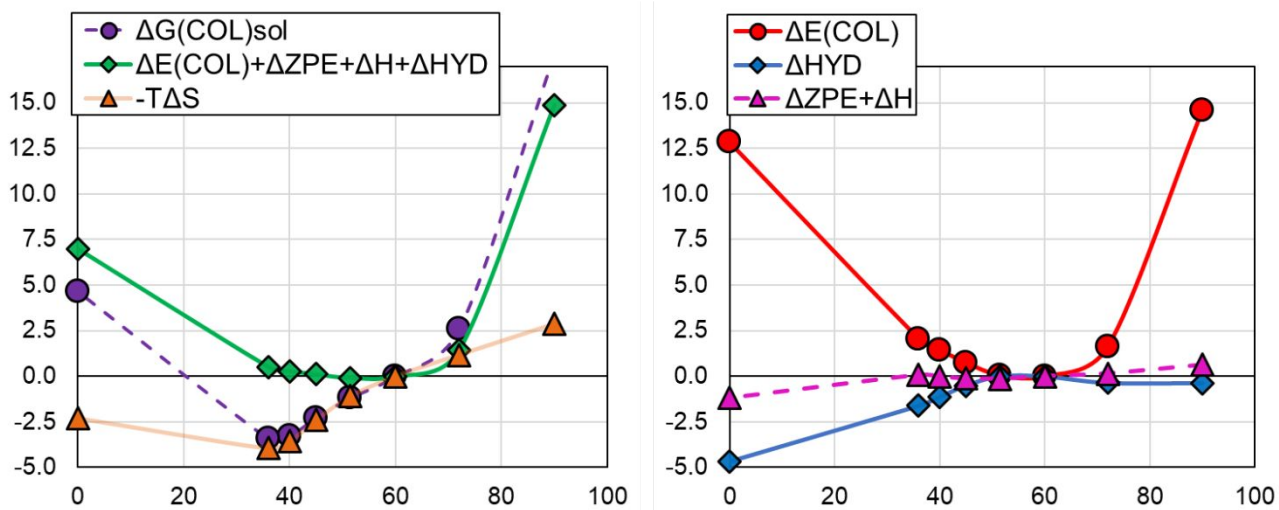


Figure 3. Decomposition of $\Delta G(\text{COL})_{\text{sol}}$ computed at room T vs α . Energy (vertical axis) in $\text{kJ mol}^{-1} \text{ triplet}^{-1}$ and α (horizontal axis) in degrees.

1
2
3 Remarkably, by including the dynamic and solvent effects, the minimum of the energy
4
5
6
7 curve moves to lower $\alpha = 36^\circ$ values, a clear indication of the triple helix unwrapping. The
8
9
10 change is mainly due to entropic effects, which stabilize more elongated structures. The
11
12
13 continuum solvation model gives solvent-protein interaction similar for most of the cases.
14
15
16
17 Only for the most elongated structures (small α values) hydration energy HYD is higher. We
18
19
20 believe this is due to the more exposed Gly and Pro (Y) C=O groups, which maximize the
21
22
23 protein-continuum solvent interaction. Zero-point and thermal vibrational energies do not
24
25
26
27 vary notably with α , thus we have grouped them together in Figure 3.
28
29
30

31
32 The over-stabilization of the elongated collagen helices (small α values), due to hydration
33
34 and entropic effects, leads to results in disagreement with the experimental evidences, for
35
36
37 which $\alpha = 51.4^\circ$. We believe the problem is due to: i) the accuracy of the HF-3c computed
38
39
40 vibration frequencies; ii) the implicit solvation approach, not accurate enough when specific
41
42
43 hydrogen bond interactions are important as in the present case. As for point i), we
44
45
46 computed, for few cases only, the vibrational frequencies at the more accurate B3LYP-
47
48
49 D3^{ABC}/VTZP level. The vibrational corrections, reported in Table S3, indicate that the HF-3c
50
51
52 level is capable of computing the right trend, but, indeed, over-stabilizes the elongated
53
54
55
56
57
58
59
60

1
2
3 structures. As for point ii), we improved the solvation model via explicit micro-solvation. In
4
5
6
7 micro-solvation, few critically important water molecules are explicitly added to the models,
8
9
10 with the purpose of representing the most important H-bonds between explicit water
11
12
13 molecules and hydrophilic groups of the collagen polymer. On top of the micro-solvated
14
15
16
17 model we run the PCM approach to account for long range bulk solvation effects.
18
19
20 Unfortunately, describing the solvent through the micro-solvation approach increases the
21
22
23 cost of the simulations notably. Therefore, we have applied it only to the two most
24
25
26
27 representative cases, i.e. collagens with $\alpha = 36^\circ$ and 51.4° . We have analyzed three levels
28
29
30 of micro-solvation with 3, 4 and 5 water molecules per amino acidic triplet, named as 3w,
31
32
33
34
35 4w and 5w, in the following. The models are reported graphically in Figure 4.
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

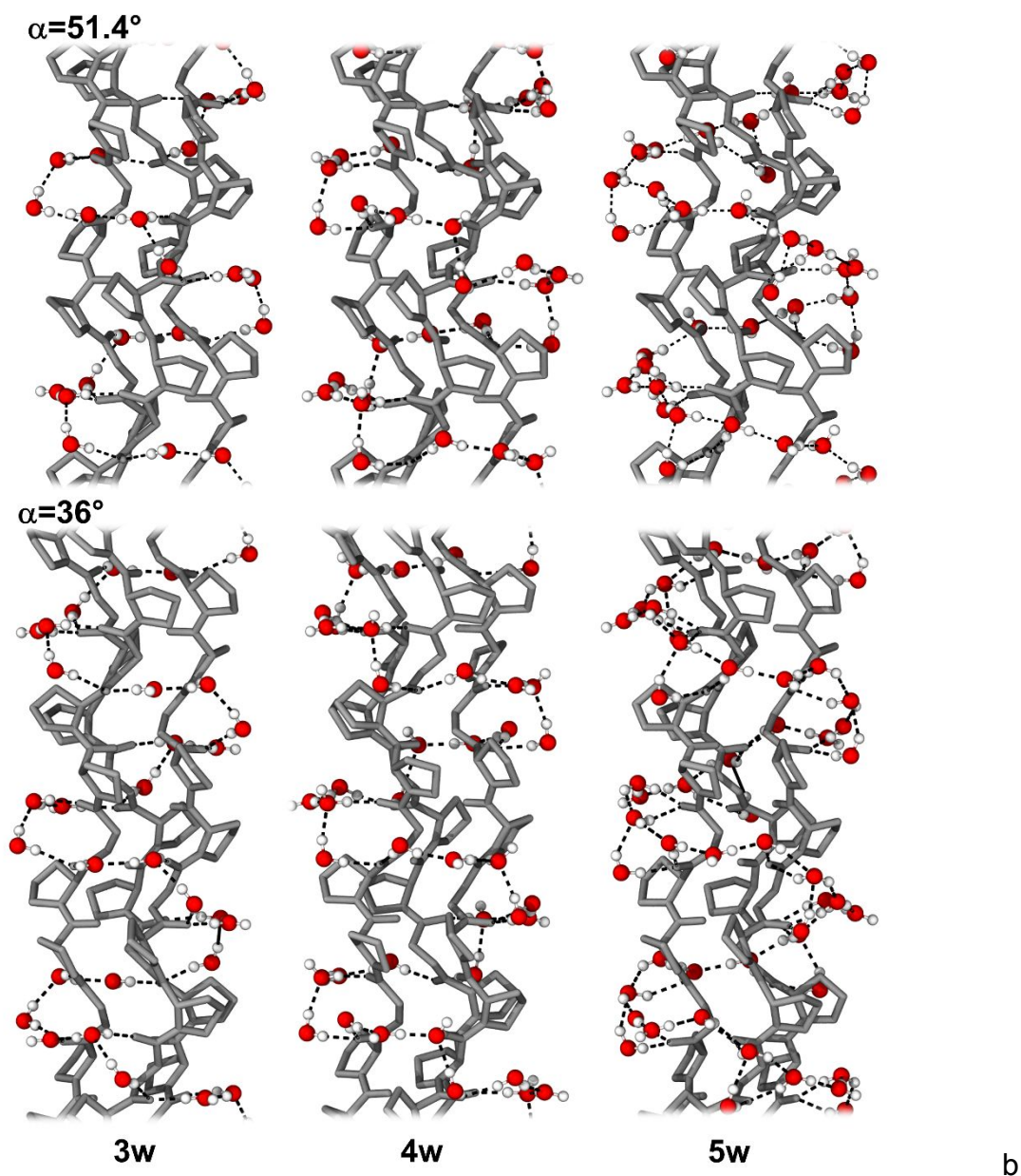


Figure 4. Micro-solvated collagen triple helices (full unit cell shown) with $\alpha = 51.4^\circ$ (TOP) and $\alpha = 36^\circ$ (BOTTOM).

We have computed the $G(\text{COL})_{\text{sol}}$ for all models as defined in eq. (3). In this case, the BSSE-corrected total energy of the micro solvated collagens ($E(\text{COL})^{\text{C}}$) is decomposed as follows:

$$E(\text{COL})^{\text{C}} = -\text{BE}^{\text{C}} + E(\text{ss-COL}) - \text{BE}^{\text{C}}_{\text{hyd}} - \text{BE}^{\text{C}}_{\text{wat}} + E(\text{W}) \quad (5)$$

Where the BE^{C} (i), $\text{BE}^{\text{C}}_{\text{hyd}}$ (ii), and $\text{BE}^{\text{C}}_{\text{wat}}$ (iii) are the BSSE corrected interaction energies between: i) collagen strands; ii) the solvated waters and the protein; iii) water molecules. The $E(\text{ss-COL})$ term, as defined above, is the energy of an isolated single collagen strand and the $E(\text{W})$ term is the sum of the energies of each isolate water molecule, both energy terms considered at the relaxed micro solvated collagen geometry. Substituting eq (5) into eq (3), and then subtracting the corresponding free energies between helices, we obtain the $\Delta G(\text{COL})_{\text{sol}}$ expression for the micro solvated case:

$$\Delta G(\text{COL})_{\text{sol}} = G(\text{COL}_{36^\circ})_{\text{sol}} - G(\text{COL}_{51.4^\circ})_{\text{sol}} = -\Delta \text{BE}^{\text{C}} + \Delta E(\text{ss-COL}) - \Delta \text{BE}^{\text{C}}_{\text{hyd}} - \Delta \text{BE}^{\text{C}}_{\text{wat}} + \Delta E(\text{W}) + \Delta \text{HYD} + \Delta \text{ZPE} + \Delta \text{H} - \text{T}\Delta \text{S} \quad (6)$$

We reported in Figure 5 the above-mentioned terms contributing to the $\Delta G(\text{COL})_{\text{sol}}$ definition for micro solvated (3w, 4w and 5w) implicit solvated (PCM) collagens.

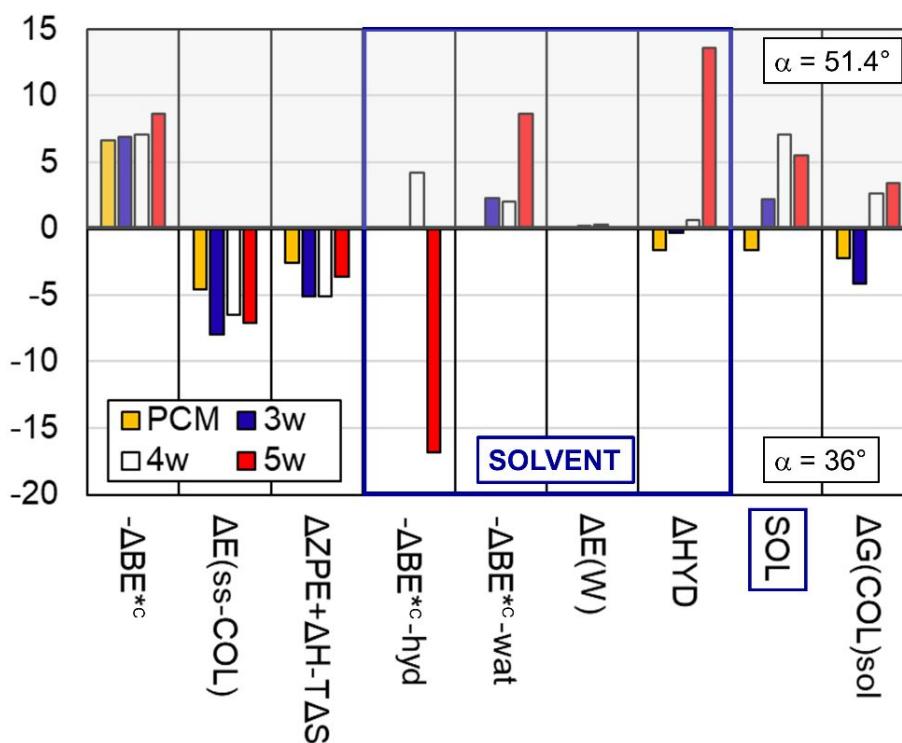


Figure 5. $\Delta G(\text{COL})_{\text{sol}}$ (vertical axis, $\text{kJ mol}^{-1} \text{ triplet}^{-1}$) contributions (see Equation 6) for micro solvated (3w, 4w and 5w) and implicit solvated (PCM) collagens. Positive/negative $\Delta G(\text{COL})_{\text{sol}}$ values stabilize/destabilize the $\alpha = 51.4^\circ$ wrapping.

As in gas-phase, also in micro solvation, the $-\Delta \text{BE}^*_{\text{C}}$ term stabilizes tight collagens ($\alpha = 51.4^\circ$), as shown by the bars in the positive region of Figure 5; on the contrary, both the $\Delta E(\text{ss-COL})$ and the vibrational/entropic contributions counterbalance the $-\Delta \text{BE}^*_{\text{C}}$ term (bars in the negative region of Figure 5). Differently from gas-phase cases, when explicit solvation is taken into account the $-\Delta \text{BE}^*_{\text{C}}$ term has a balanced contribution from DFT and D components due to the shortening of the inter-chains $\text{N-H} \cdots \text{O}=\text{C}$ H-bond (Table S10-11).

1
2
3 The $-\Delta BE^*C_{\text{hyd}}$, $-\Delta BE^*C_{\text{wat}}$, $\Delta E(W)$ and ΔHYD terms, highlighted in a blue frame in Figure 5,
4
5
6 describe the solvent contributions to the free energy difference. All solvent effects
7
8
9
10 contributions are grouped in the SOL term of Figure 5. Each single component, as well as
11
12
13 the overall value, vary notably depending on the level and geometry of solvation. Notable is
14
15
16 the 5w case; in this case the $\alpha = 51.4^\circ$ collagen has a low value for the water-protein
17
18
19 interaction ($-\Delta BE^*C_{\text{hyd}}$) which is balanced by a high inter-waters ($-\Delta BE^*C_{\text{wat}}$) and water-bulk
20
21
22 (ΔHYD) interaction. The comprehensive SOL term amounts to -1.6, 2.2, 7.1 and 5.5 kJ mol⁻¹
23
24
25 triplet⁻¹ for the PCM, 3w, 4w and 5w cases. As expected, the sole PCM stabilizes the
26
27
28 elongated collagen and it is inaccurate to simulate the water-collagen case.
29
30
31
32
33

34 The computed interaction energy felt by each water molecules, which comes from the bulk
35
36
37 and the inter-waters interaction, increases with the increasing number of water molecules.
38
39
40
41 This is a consequence of the shortening of the inter-water H-bond length, as shown in Figure
42
43
44 S10, due to enhanced H-bond cooperativity. An exception to this trend is the 5w case with
45
46
47
48 $\alpha = 36^\circ$, in which an addition water-protein H-bond reduces the inter-waters and water-bulk
49
50
51
52 interaction with respect to the other cases. Interestingly, the inter-waters cooperative effects
53
54
55 also affect the H-bond strength with the protein, which increases along with the inter-waters
56
57
58
59
60

and water-bulk interaction, see Figure 6. This is due to a shortening of the water-protein H-bond length as reported in Figure S10.

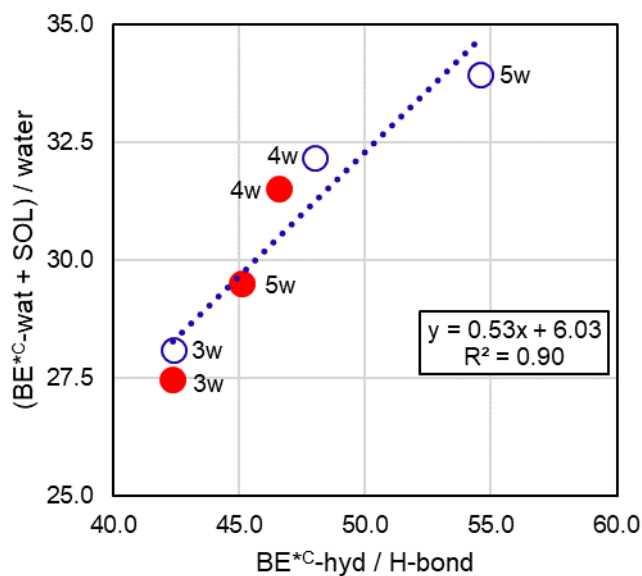


Figure 6. Correlation between average water-protein (vertical axis) and water-water binding energy (horizontal axis), normalized by the number of water-protein H-bonds and the number of water molecules, respectively. Empty blue circle for $\alpha=51.4^\circ$, red filled ones for $\alpha=36^\circ$. Energy in kJ mol^{-1} triplet⁻¹.

Overall, the computed $\Delta G(\text{COL})_{\text{sol}}$ is -2.2, -4.1, 2.6 and 3.4 kJ mol^{-1} triplet⁻¹ for the PCM, 3w, 4w and 5w cases, respectively (see Fig. 5). Only for the collagens with high solvation, i.e. 4w and 5w case, the $\alpha = 51.4^\circ$ helix conformation is favored, in agreement with the

1
2
3 experiments. This indicate that the minimum water number to solvate accurately Gly-Pro-
4
5
6
7 Pro collagen models is 4 water per aminoacidic triplet.
8
9

10 In summary, we have built up a robust, controllable, reproducible and relatively cheap
11
12
13 computational procedure capable of reproducing the experimental evidences about collagen
14
15
16
17 helical wrapping. The fine physico-chemical features driving collagen to specific helical
18
19
20 wrappings are captured by accurate hybrid DFT simulations on reliable collagen protein
21
22
23
24 polymer-like models. The models included also explicit water micro solvation supplemented
25
26
27 by PCM solvation to mimic bulk water effects. Through these simulations we have dissected
28
29
30 the energetic terms and discovered those favoring more compact/elongated collagen
31
32
33
34 helices. We demonstrated that a tight collagen helices wrapping is favored by: i) a stronger
35
36
37 inter-chains interaction, not only coming from the dispersion interaction, but also from inter-
38
39
40 chains electrostatic/polarization/charge-transfer terms; ii) a favorable water-collagen
41
42
43
44 interaction. On the contrary, a loose collagen helices wrapping is stabilized by: i) a more
45
46
47 stable polymer structure, resembling closely the expected geometry of a free Pro rich
48
49
50 peptide (poly-Pro type II); ii) entropic factors. The delicate balance of these factors induces
51
52
53
54
55
56
57
58
59
60

1
2
3 more compact helices to be more stable for the more realistic micro solvated collagens. In
4
5
6
7 these cases, the helical features are in agreement with the experimental evidences.
8
9

10 The approach presented here is tested on the Gly-Pro-Pro collagen composition, which is
11
12
13 well known in literature, but it can be extended to any other collagen composition.
14
15
16

17 Preliminary results on collagens with Gly-Leu-Hyp and Gly-Phe-Hyp compositions indicate
18
19
20 a preference for more elongated helices with respect to Gly-Pro-Pro, in agreement with the
21
22
23 experimental evidences.¹⁷ The proven predictivity of this approach can be useful for the
24
25
26
27 design of collagen like peptides with specific helical features.
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Computational Methods

We relaxed the geometry for all models using the CRYSTAL17 suit,²⁹ relying on several *ab-initio* methodologies. Within the DFT framework we employed the B3LYP-D,³⁰⁻³² and PBE-D,³³ functionals with a large Gaussian VTZP quality basis set.³⁴ The adoption of localized Gaussian function implies that many quantities are affected by the basis set superposition error (BSSE) which has, therefore, taken into proper account. Due to the very large size of the systems we run frequencies calculations with the cost-effective HF-3c method.²⁸ Furthermore, we run single point energy calculations to assess the effect of: i) the basis set quality, ii) the dispersion forces description, and iii) the water solvation. To do so we also employed the VASP code.³⁵⁻³⁸ Further details on the computational approach are reported in the SI.

Associated Content

Supporting information content: Computational Methods; Collagen Models Description; Definition of the Computed Quantities; Table S1-2: Basis sets employed; Table S3: Vibrational corrections to the energy for gas phase collagens at the HF-3c and DFT-D levels; Table S4: Relative total energies between helices in function of the method; Table S5: BE*, $\Delta E(\text{ss-COL})$ and $\Delta E(\text{COL})$ and all their contributions for DFT and DFT-D results; Table S6: Computing BSSE with the counter poise method and a plane waves basis sets; Figure S2-3: Dispersion scheme and geometry relaxation effect of energy ranking; Figure S4: Energy decomposition of $\Delta E(\text{ss-COL})$ and ΔBE^* terms; Figure S5: Height per triplet and dispersion component of BE correlation; Figure S6: Average inter-strands electrostatic contacts lengths in function of α ; Figure S7: Correlation between $\Delta E(\text{ss-COL})$ and $-\Delta \text{BE}^{*C}$ terms with α ; Figure S8: Main torsional angles at the HF-3c level of theory in function of α ; Figure S9: Correlation of $T\Delta S$, ΔE_T and ΔE_0 contributions with the axial rise per residue; Energy Decomposition in Micro-Solvated Collagens (Table S7-10); Table S11: Geometrical analysis of the COL-W models, Figure S10: Correlation energetic and geometry within micro solvated collagen models.

1
2
3 Author Information
4
5

6
7 Corresponding authors: Dr. Michele Cutini and Prof. Piero Ugliengo
8
9

10 NAME: ORCID
11
12

13
14 Dr. Michele Cutini: 0000-0001-6896-7005
15
16

17 Dr. Stefano Pantaleone: 0000-0002-2457-1065
18
19

20 Prof. Piero Ugliengo: 0000-0001-8886-9832
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Acknowledgments
4
5
6
7

8 SP acknowledges funding from the European Research Council (ERC) under the
9
10 European Union's Horizon 2020 research and innovation program, for the Project “The
11
12 Dawn of Organic Chemistry” (DOC), grant agreement No 741002. MC acknowledges
13
14 Massimo Bocus, M.Sc. and Irene Bechis, M.Sc for the efforts devoted to support the
15
16 “Collagen Project”. PU and MC acknowledge the University of Torino for funding (grant
17
18 agreement No CHI.2019.21/XXII) and the CRYSTAL team for continuous support in the
19
20 usage of the CRYSTAL program.
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

References

- (1) Dobson, C. M. Protein Folding and Misfolding. *Nature* **2003**, *426*, 884.
- (2) Dill, K. A.; Maccallum, J. L.; Folding, P. The Protein-Folding Problem , 50 Years On. **2012**, No. NOVEMBER, 1042–1047.
- (3) Onuchic, J. N.; Wolynes, P. G. Theory of Protein Folding. *Curr. Opin. Struct. Biol.* **2004**, *14* (1), 70–75.
- (4) Dill, K. A.; Ozkan, B. S.; Schell, S. M.; Thomas, W. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **2008**, *37*, 289–316.
- (5) Gething MJ; J, S. Protein Folding in the Cell. *Nature* **1992**, *355*, 33–45.
- (6) Ulrich Hartl, F. Molecular Chaperones in Cellular Protein Folding. *Nature* **2010**, *3* (9), 571–580.
- (7) Dill, K. A. Dominant Forces in Protein Folding. *Biochemistry* **1990**, *29* (31), 7133–7155.
- (8) Rich, A. and Crick, F. H. C. The Structure of Collagen. *Nature* **1955**, *176*, 915–916.

- 1
2
3
4 (9) Ramachandran, G. N.; Kartha, G. Structure of Collagen. *Nature* **1955**, *176*, 593.
5
6
7
8 (10) Cowan, P. M.; Mcgavin, S.; North, A. C. T. The Polypeptide Chain Configuration of
9
10
11 Collagen. *Nature* **1955**, *176* (4492), 1062–1064.
12
13
14
15 (11) Bella, J. Collagen Structure: New Tricks from a Very Old Dog. *Biochem. J.* **2016**, *473*
16
17
18 (8), 1001–1025.
19
20
21
22
23 (12) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu. Rev. Biochem.*
24
25
26 **2009**, *78* (1), 929–958.
27
28
29
30
31 (13) Ramshaw, J. A. M.; Shah, N. K.; Brodsky, B. Gly-X-Y Tripeptide Frequencies in
32
33
34 Collagen: A Context for Host-Guest Triple-Helical Peptides. *J. Struct. Biol.* **1998**, *122*
35
36
37 (1–2), 86–91.
38
39
40
41
42 (14) Okuyama, K.; Takayanagi, M.; Ashida, T.; Kakudo, M. A New Structural Model for
43
44
45 Collagen. *Polym. J.* **1977**, *3*, 341–343.
46
47
48
49
50 (15) Okuyama, K.; Miyama, K.; Mizuno, K.; Bächinger, H. P. Crystal Structure of (Gly-Pro-
51
52
53 Hyp)₉: Implications for the Collagen Molecular Model. *Biopolymers* **2012**, *97* (8), 607–
54
55
56
57 616.
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

- (16) Okuyama, K. Revisiting the Molecular Structure of Collagen. *Connect. Tissue Res.* **2008**, *49* (5), 299–310.
- (17) 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.
- (18) Luo, T.; Kiick, K. L. Collagen-like Peptides and Peptide-Polymer Conjugates in the Design of Assembled Materials. *Eur. Polym. J.* **2013**, *49*, 2998–3009.
- (19) Fields, G. B. Synthesis and Biological Applications of Collagen-Model Triple-Helical Peptides. *Org. Biomol. Chem.* **2010**, *8* (6), 1237.
- (20) Fallas, J. A.; O’Leary, L. E. R.; Hartgerink, J. D. Synthetic Collagen Mimics: Self-Assembly of Homotrimers, Heterotrimers and Higher Order Structures. *Chem. Soc. Rev.* **2010**, *39* (9), 3510–3527.
- (21) Rele, S.; Song, Y.; Apkarian, R. P.; Qu, Z.; Conticello, V. P.; Chaikof, E. L. D-Periodic Collagen-Mimetic Microfibers. *J. Am. Chem. Soc.* **2007**, *129* (47), 14780–14787.
- (22) Kusebauch, U.; Cadamuro, S. A.; Musiol, H. J.; Lenz, M. O.; Wachtveitl, J.; Moroder,

- 1
2
3 L.; Renner, C. Photocontrolled Folding and Unfolding of a Collagen Triple Helix.
4
5
6
7 *Angew. Chemie - Int. Ed.* **2006**, *45*(42), 7015–7018.
8
9
10
11 (23) Wang, A. Y.; Mo, X.; Chen, C. S.; Yu, S. M. Facile Modification of Collagen Directed
12
13
14 by Collagen Mimetic Peptides. *J. Am. Chem. Soc.* **2005**, *127*(12), 4130–4131.
15
16
17
18 (24) Fallas, J. A.; Gauba, V.; Hartgerink, J. D. Solution Structure of an ABC Collagen
19
20
21
22 Heterotrimer Reveals a Single-Register Helix Stabilized by Electrostatic Interactions.
23
24
25
26 *J. Biol. Chem.* **2009**, *284*(39), 26851–26859.
27
28
29
30 (25) Newberry, R. W.; VanVeller, B.; Raines, R. T. Thioamides in the Collagen Triple Helix.
31
32
33
34 *Chem. Commun. (Camb)*. **2015**, *51*(47), 9624–9627.
35
36
37
38 (26) Egli, J.; Siebler, C.; Köhler, M.; Zenobi, R.; Wennemers, H. Hydrophobic Moieties
39
40
41
42 Bestow Fast-Folding and Hyperstability on Collagen Triple Helices. *J. Am. Chem. Soc.*
43
44
45 **2019**, jacs.8b13871.
46
47
48
49 (27) Cutini, M.; Bocus, M.; Ugliengo, P. Decoding Collagen Triple Helix Stability by Means
50
51
52
53 of Hybrid DFT Simulations. *J. Phys. Chem. B*.
54
55
56
57 (28) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685.

- (29) 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.
- (30) Becke, A. D. Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. *Phys. Rev. A* **1988**, *38* (6), 3098–3100.
- (31) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652.
- (32) 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.
- (33) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.
- (34) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully Optimized Contracted Gaussian Basis Sets

1
2
3 of Triple Zeta Valence Quality for Atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829.
4
5
6

7
8 (35) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamcis for Liquid Metals. *Phys. Rev. B*
9
10
11 **1993**, *47*(1), 558.
12

13
14
15 (36) Kresse, G.; Furthmüller, J.; Hafner, J. Ab Initio Molecular-Dynamics Simulation of the
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
Liquid-Metal–Amorphous-Semiconductor Transition in Germanium. *Phys. Rev. B*
1996, *6*(1), 558–561.

(37) 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.

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