Infrared harmonic features of collagen models at B3LYP-D3: From amide bands to the THz region

The Journal of Chemical Physics

Infrared harmonic features of collagen models at B3LYP-D3: From amide bands to the THz region

Cite as: J. Chem. Phys. **155**, 075102 (2021); https://doi.org/10.1063/5.0056422
Submitted: 10 May 2021 . Accepted: 02 August 2021 . Published Online: 19 August 2021

Michele Cutini, and Piero Ugliengo

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Infrared harmonic features of collagen models at B3LYP-D3: From amide bands to the THz region

Cite as: J. Chem. Phys. 155, 075102 (2021); doi: 10.1063/5.0056422
Submitted: 10 May 2021 • Accepted: 2 August 2021 •
Published Online: 19 August 2021

Michele Cutini[a] and Piero Ugliengo

AFFILIATIONS

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

[a]**Author to whom correspondence should be addressed:** michele.cutini@unito.it. **Telephone:** +39 011 670 4597

ABSTRACT

In this paper, we have studied the vibrational spectral features for the collagen triple helix using a dispersion corrected hybrid density functional theory (DFT-D) approach. The protein is simulated by an infinite extended polymer both in the gas phase and in a water micro-solvated environment. We have adopted proline-rich collagen models in line with the high content of proline in natural collagens. Our scaled harmonic vibrational spectra are in very good agreement with the experiments and allow for the peak assignment of the collagen amide I and III bands, supporting or questioning the experimental interpretation by means of vibrational normal modes analysis. Furthermore, we demonstrated that IR spectroscopy in the THz region can detect the small variations inherent to the triple helix helicity (10/3 over 7/2), thus elucidating the packing state of the collagen. So far, identifying the collagen helicity is only possible by means of crystal x-ray diffraction.

I. INTRODUCTION

Collagen protein is found in many different tissues with several functions in all vertebrates and it is a fundamental piece in bones and tendons. Its structural peculiarity is the geometrical motif in which three parallel polypeptide strands coil about each other to form a triple helix.[1,2] To sustain the collagen triple helix geometrical motif, the primary structure is forced to a triplet repeated sequence.[2] Each triplet always sports a glycine (Gly, G) in the first position of the triplet, and the second and third positions of the triplet are called X and Y, respectively. In the collagen triplet (GXY), proline (Pro, P) is usually in the X position and (2S, 4R)-4-hydroxyl-proline (Hyp, O) is in the Y position. Indeed, Pro and Hyp are found in X and Y positions in 28% and 38% of the cases, respectively. In the overall collagen residue count, Pro and Hyp contribute to 9.3% and 12.7% of the amino acid content, respectively. The Gly-Pro-Hyp (GPO) triplet occurs with the highest frequency in all collagens (10.5%).[3]

Many research lines are developed with the aim of facilitating collagen biomedical applications and guiding the design of new biomaterials for medicine,[4] such as the synthesis of hyper-stable collagen triple helices by replacing Gly with Aza-Gly,[5–9] by applying pendant hydrophobic moieties,[10] and by inter-strand cross-linking.[11,12]

Unfortunately, detailed structural information on the collagen structure is available in only a few cases.[9,13] Experimental structure crystallization and acquisition by high resolution x-ray diffraction patterns are the main challenges. The limitations are more stringent for flexible collagen aggregates.[14] In these, infrared spectroscopy can be of help. Indeed, it is a simple technique that can be applied to collagen based materials (not necessary as a crystalline phase), and the analysis of the IR spectrum can show different structural features of collagen based tissues.[15,16]

In general, the most interesting band for the analysis of the protein structure is the amide I band. It extends roughly in the interval between 1700 and 1600 cm^{-1}, mostly originating from stretching vibrations of the peptide C=O group with smaller components from both C–N stretching and N–H bending vibrations. This is an intense band, and it is quite sensitive to the changes in the conformational state of the polypeptide chain. The experimental amide I band of

the collagen protein in solution is quite broad,[16] not allowing to assign amide I peaks' contribution to vibration modes specific for each amino acid within the collagen triplet.[17] Conversely, Pro-rich mimetic polymers [Gly-Pro-Pro (GPP) composition] exhibit amide I band vibration characterized by three distinct peaks, located at 1664, 1644, and 1628 cm^{-1}.[18,19] In this case, it is possible to assign a C=O stretching vibration to each peak of the amide I region of the spectra (see Table S1 of the supplementary material). This analysis has been performed experimentally by Lazarev *et al.*[18] on collagen GPP models in D_2O. They assigned the low-frequency peak at 1628 cm^{-1} to Gly, the central one at 1644 cm^{-1} to Pro in the X position, Pro(X), and the last one at 1664 cm^{-1} to Pro in the Y position, Pro(Y). Remarkably, another work gave a different peak assignation using molecular mechanics simulations based on an *ad hoc* force field.[19]

A second notable band is the amide II band. This band is constituted by peptide C–N stretching and N–H bending vibrations. For dry collagen models, films are centered at ~1560 cm^{-1} and amide II band is usually less intense than amide I band.[16,20] Unfortunately, this IR region is perturbed by several peculiar bands of amino acid side chains, such as tyrosine, asparagine, and glutamine.[21]

Another interesting but less commonly studied IR band is the amide III band. Investigating the amide III band is a valuable complementary tool to amide I analysis in protein structural determination. Amide III bands fall in the 1200–1350 cm^{-1} region of the IR spectrum, which is mainly due to both C–N and C–C stretching and N–H and C–H bending modes.[22,23] One drawback of amide III is its relatively weak signal; recently, however, thanks to synchrotron radiation, it was possible to collect and analyze the amide III band for a collagen triple helix in water.[23] The spectra in that region exhibit three distinct peaks located roughly at 1202, 1240, and 1284 cm^{-1}. The experimental interpretation indicates that the higher frequency peak is generated by Hyp(Y) Cα–H bending mode. The central peak corresponds to Gly N–H bending mode and the lowest frequency peak comes from vibration modes located on the lateral chains of X and Y amino acids.

Another region of the IR spectra, which is gaining attention lately for structure recognition purposes, is the THz region. THz spectroscopy and THz simulated spectra are often employed for structural identification for several types of crystals, such as amino acids,[24] polypetides,[25] and polymers crystals, through their characteristic soft and collective motions occurring at very low wave numbers.[26,27] Unfortunately, for the collagen case, there are no experimental data available.

In the process of understanding the origin of the bands in the IR spectrum, experimentalists can get guidance from molecular simulations, particularly for complex spectra.[28] Indeed, the IR spectrum, while not being directly connected to the structure, can give fine details of the chemical environment of specific functional groups through their fingerprint in the vibrational spectrum. Due to the large size of the collagen, the most widespread techniques for its simulation are based on classical force fields.[29] These methods can be applied to realistic collagen models, but their use in the analysis of vibrational spectra is limited by their heavy parameterization.[19] Classical force fields were employed extensively in the last decade for simulating realistic collagen macrostructures (the so-called collagen fibrils). This has allowed for a better understanding of the atomic interactions,[30,31] as well as the mechanical properties,[32] of collagens within a realistic environment, which can also include the hydroxyapatite mineral to mimic the bone material.[33,34] Conversely, the use of *ab initio* techniques, essentially Density Functional Theory (DFT), ensures accurate and almost parameter free results not only for the structure but also for the vibrational feature prediction. DFT is also suitable to deal with hydrogen bond interactions between water and collagens. The subtle H-bond features are difficult to model by force fields, particularly when both the perturbation of the amide frequency bands and the THz spectra have to be modeled, as in this paper. Unfortunately, their use on collagens is rare, and in these few cases, the collagen is simply modeled as a short molecular tri-peptide.[35,36]

In this paper, for the first time at the full DFT level of theory, we report the vibrational spectrum features (in the harmonic approximation) for a collagen triple helix extended polymer either in the gas phase or in a water micro-solvation environment. We relied on the hybrid dispersion-corrected B3LYP-$D3^{ABC}$ functional,[37,38] coupled with an all-electron TZP polarized quality basis set,[39,40] which ensures a good description of hydrogen bonds while minimizing the basis set superposition error (see the supplementary material for further information). We run all simulations with the CRYSTAL17 code,[41] as our long experience in using it should ensure the needed accuracy and reproducibility of the results.[42–45] In our analysis, we have focused on the IR band region relevant for protein structure recognition, e.g., amide I, II, III, and THz. When possible, we have compared our results with the experiments, confirming or questioning the band interpretation.

II. COMPUTATIONAL METHODS

We computed relaxed geometries, energies, and vibrational frequencies with the CRYSTAL17 code.[41] Standard DFT simulations were run using the B3LYP hybrid functional,[37,38] corrected with the recent D3 scheme,[46] including the Axilrod–Teller–Muto (ATM)-three-body-term ($D3^{ABC}$).[47,48] Atomic positions and cell vectors optimization adopted the analytical gradient method. The Hessian was upgraded with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.[49–51] We set tolerances for the convergence of the maximum allowed gradient and the maximum atomic displacement to default values, e.g., 0.0009 and 0.0018 a.u., respectively.

To help convergence of the SCF, the Fock/KS matrix at a given cycle was mixed with 30% of one of the previous cycles.[52] The recently introduced DIIS extrapolator technique has been employed to speed up the SCF convergence.[53] Tolerances of the bi-electronic integral calculation were set to 10^{-6} for Coulomb overlap, Coulomb penetration, exchange overlap, and exchange pseudo-overlap in the direct space and 10^{-14} for exchange pseudo-overlap in the reciprocal space (CRYSTAL17 keyword: TOLINTEG 6 6 6 6 14).[52] The electronic shrink factor used in the calculations is set to 4, equivalent to 3 reciprocal k space points. For the vibrational frequency calculations, the mass-weighted force-constant matrix was computed at the gamma point by the numerical derivative of the analytic nuclear gradients. A value of 0.003 Å was chosen as the displacement of each atomic coordinate. The numerical first derivatives were calculated using the different quotient formula, e.g., one displacement for each atom along each Cartesian direction (CRYSTAL17

keyword: NUMDERIV 1). In the case of dry GPP 10/3 helix, negative frequencies appeared. In that case, we switched to the more accurate central-difference formula that uses two displacements for each atom along each Cartesian direction (CRYSTAL17 keyword: NUMDERIV 2). With that approach, all structures are true minima (all vibrational frequencies are positive). The IR intensity of each normal mode of vibration was computed using the Berry phase approach.[54] In frequency calculations, tolerance on the energy convergence was set to 10^{-10} hartree, which are calculated in full, including all the vibrational modes of the collagen polymeric models. We checked for the effect of the phonon dispersion on the frequency values for the considered C=O group stretching. Even for the smaller unit cells employed in this paper, the phonon dispersion effect was found negligible. B3LYP calculations were carried out using molecular all-electron Gaussian basis sets. A split valence basis with polarized function (SVP) set of 3-11G(p) was chosen for H atom and a more extended polarized VTZP basis set from Schäfer *et al.* for N, O, and C atoms (see the supplementary material).[40] The graphical visualization and structural manipulation of structures were performed with MOLDRAW version 2.0.[55] Images were rendered with VMD.[56]

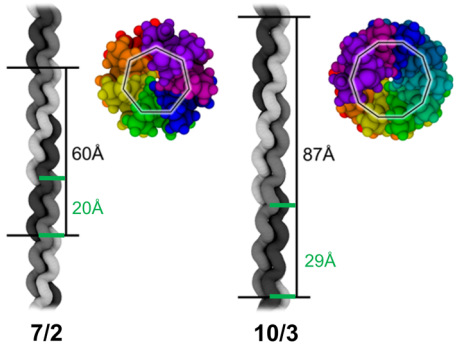

FIG. 1. Collagen models with 7/2 and 10/3 helicities. The lateral view is a tube representation in different levels of gray for each collagen single strand. The black segments delimit the translational repetition of the model. The green segments delimit the actual polymer unit cell length of the models employed in this paper. The length reduction is obtained, thanks to the imposed symmetry operators (see Ref. 57). The top view shows a single protein strand with all amino acid triplets within the unit cell depicted in different colors. The geometrical shapes follow the triplet wrapping.

III. COLLAGEN MODELS

We have adopted a symmetric and periodic collagen protein polymer (see Ref. 57). The main advantages of our approach with respect to literature molecular collagen models are as follows: (i) exploitation of the intrinsic roto-translational symmetry of collagen protein, thus reducing the computational cost of the simulations, and (ii) avoiding the spurious effects coming from the short length of molecular collagen models, so mimicking the length of the real collagen. It is worth noting that some structural effects cannot be taken into account in our collagen modeling, such as the structural deformations (protein bending and torsions) occurring for collagen type I within the fibrils of bones and tendons. We have already applied this modeling approach for studying the collagen helical structure[58] and for studying Pro and Hyp side-chain conformation within collagens.[57]

The collagen models employed in this paper are homo-trimeric, and they are built by the repetition of the same amino acid triplet. To allow some degree of variability, we have considered two collagen compositions, e.g., Gly-Pro-Hyp (GPO) and Gly-Pro-Pro (GPP), adopted for different purposes. In the following, we named the collagen polymer "COL" to refer to the GPO composition, but for the simulation of the vibrational spectrum in the THz region, where we adopted a COL with the GPP composition that allows for a cleaner understanding of the effect of different helicities on the collagen conformation compared to GPO. As for the helical packing, it is known that the protein amino acid sequence drives the collagen to pack with different helical geometries,[1] a fact long debated.[59] Currently, the agreement that holds is for a high content of Pros and derivatives, the collagen has a 7/2 helicity,[60,61] while Pro-poor collagens exhibit a 10/3 helicity (see Fig. 1).[62] The two helical geometries vary for the torsional degree of the amino acid triplets along the helical axes, which is higher for the 7/2 helix, providing a tighter and more compact helix than the 10/3. So far, evidence on the collagen helicity is accessible only by x-ray crystal diffraction. In this paper, we simulated both helicities, e.g., 7/2 and 10/3. The helical symmetry imposed to the models allows us to start from an asymmetric unit polymer cell content envisaging only one amino acid triplet. The resulting models have three peptide strands wrapped together for a total of seven (7/2) or ten (10/3) amino acid triplets in the unit cell. For the 7/2 case, the polymers have 245/252 atoms for GPP/GPO, while they rise to 350/360 atoms for the 10/3 case. For the adopted polymeric GPO and GPP protein models, we have characterized the geometry and assessed the stability ranking for all the gas-phase conformers in Ref. 57 and for the water micro-solvation environment in Ref. 63. In the gas-phase, the conformation variability arises from the side chain mobility of Pro and Hyp (puckering of the pyrrolidine ring and OH group orientation, respectively). In a water micro-solvation environment, along with the side-chain flexibility, different collagen models can be hydrated with a different number of water molecules with variable water molecule organization around the protein. In this regard, we have employed a micro-solvation approach using 5, 6, and 7 water molecules per amino acid triplet, for the GPO case. The resulting models have 357, 378, and 399 atoms, respectively. Five water molecules per triplet is the lowest level of hydration, which solvates each exposed C=O and OH group of a GPO collagen. In this paper, we present no results for solvated GPP. For reference, the solvation for GPP was described in Ref. 58. No mixing of different spectra is performed, but only one model in each case is employed. For gas phase results, the vibrational analysis is carried out on the most energetically stable collagen conformer for GPO and GPP (see Ref. 57). As for the GPO solvated collagen, we have employed the structure that has the solvation geometry (distance and angle of adsorbed water) in best agreement with the experimental data (crystals of collagen-like peptides with GPO composition) among all computed structures. Each vibrational computed spectrum includes all vibrational modes of the polymer.

To keep similarity with the experimental conditions, we have computed the IR spectrum of collagen in explicit water solvation. To simulate the water environment, while keeping the cost

of the calculation reasonable, we added 35 water molecules *per* polymer unit cell (5 water molecules per amino acid triplet, see Fig. 2).

Our approach does not allow (for computational reasons) to take anharmonicity into account and more importantly dynamic effects of the water interacting via the H-bond with specific groups (namely C=O) protruding out of the collagen framework. Nonetheless, the comparison of our predicted spectra with the experimental ones (*vide infra*) shows very good agreement and gives confidence that the neglected effects are overall relatively small and compensated by the small bandwidths associated with the harmonic frequency bands.

For hydration, we adopted a Gly-Pro-Hyp (GPO) collagen model with 7/2 helicity. We have chosen the GPO composition over GPP because GPO is the most common collagen triplet and makes a closer model for the natural collagen. Furthermore, the GPO difference in the amide I region compared to GPP is negligible. Indeed, we proved that the extra OH group of Hyp does not interfere directly with the C=O environment and therefore on the amide I band (see Fig. S1 of the supplementary material). For a better discussion of the results, we have also investigated a single collagen strand not folded into a triple helix. To model this, we have adopted the poly-proline type II (PPII) polymer. This helical geometry occurs often in the Pro-rich section of proteins, and it is the geometrical organization of peptides containing only proline in water.[64] Furthermore, the PPII structure is the geometry of each single collagen strand when packed together into a triple helix, due to geometrical restrictions imposed by the collagen helicity. Moreover, an all-Pro PPII polymer amide I response resembles that of a Pro-rich collagen strand.[19] Considering the high content of Pro and derivatives in our collagen models, we believe that the PPII geometry is an oversimplified but reasonable choice for modeling a single unfolded collagen strand. To simulate the water environment, we added 10 water molecules *per* polymer unit cell (see Fig. 2). Following the same procedure adopted for collagens, the most stable hydrated PPII conformer was selected for computing the IR spectrum (see Ref. 63).

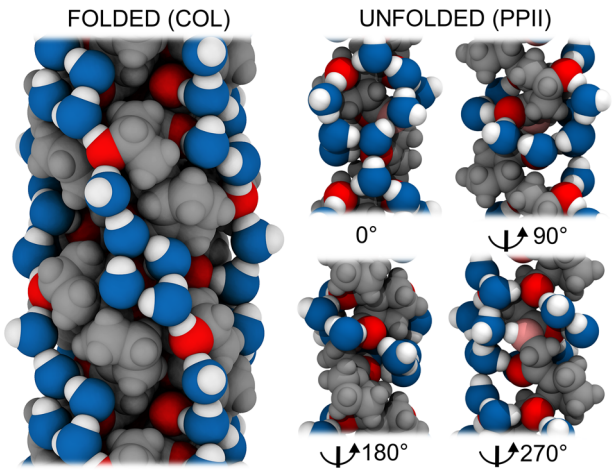

FIG. 2. Graphical representation (as van der Waals spheres) of the periodic collagen protein folded [COL(GPO)] and unfolded (PPII) in explicit water microsolvation. The PPII system is reported by four different orientations. Color code: Water oxygen is in blue, protein oxygen is in red, hydrogen is in white, and nitrogen (reported only for PPII) is in pink. The rest of the proteins is in gray.

IV. RESULTS AND DISCUSSION

First, we have made a detailed analysis on the amide I band of the collagen IR spectra varying as a function of collagen organization.[16] We have scaled (0.9815 scaling factor) all harmonic frequencies to account for systematic errors (DFT, basis set, and anharmonicity) so that the most intense computed amide I peak overlaps with the experimental one. The amide I band section of the computed spectrum is shown in Fig. 3(a), which superimposed with the experimental one: the agreement on relative stretching frequencies and intensities is remarkable. Figure 3(a) highlights the different contributions to the spectrum coming from the three nonequivalent C=O groups within the GPO triplet. Interestingly, the assignment indicates the Gly C=O stretching as the central peak, while that for Pro vibrates at the lowest frequency. This assignment is not in line with the analysis of Lazarev *et al.*,[18] which associates the low-frequency peak to Gly and the central one to Pro(X). Bryan *et al.* supported the assignment of the central peak to Pro(X) because its IR intensity is the most sensitive to the denaturation process, as summarized in Ref. 19. This is in agreement

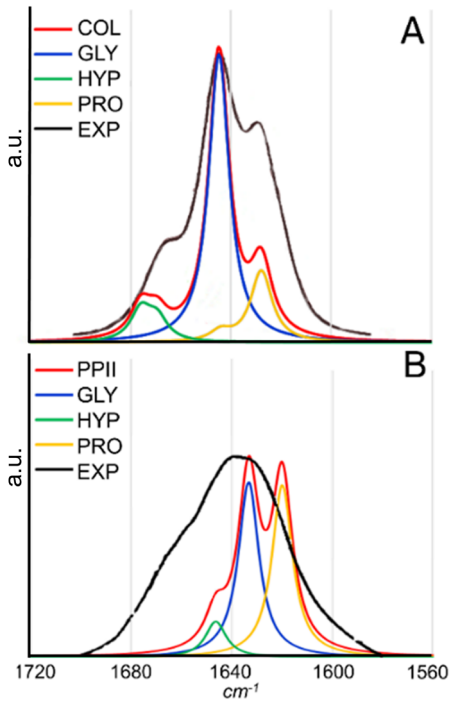

FIG. 3. Amide I region of the COL(GPO) (a) and PPII (b) IR spectra in D_2O (frequency scale in cm^{-1}). In black, the experimental data for collagen GPP (16 °C) and PPII GPO (80 °C) from Ref. 19. Simulated spectra are reported with separated contributions from the different amino acids. Computed intensities (in arbitrary units, a.u.) are scaled to match the experimental ones as close as possible. Full width at half maximum of the Gaussian functions is set to 10 cm^{-1} for a better comparison with the experimental results.

with the fact that the Pro(X) C=O group is buried within the protein and will become exposed to the water environment only during denaturation.

Therefore, to provide further clues on the arguments of Lazarev *et al.*,[18] we have simulated the IR spectrum of the unfolded collagen protein as a model for the denatured collagen in a heavy water (D_2O) environment. The results, reported in Fig. 3(b), deviate sensibly from the experiments. A better agreement would come enlarging the maximum width of the Gaussian functions used to build the spectra, in agreement with the high temperature of the reference experiments. Alternatively, considering a Boltzmann-weighted ensemble of PPII conformers, arising from high experimental temperature and the flexibility of the system, could reproduce the experiment. Nevertheless, the result still exposes the limits of the selected simplified polymeric model representing an unfolded collagen strand.

Interestingly, the computed amide I band central peak undergoes the highest variation in the denaturation process, in line with the experimental evidence. Its intensity is strongly reduced, leading both Gly and Pro peaks to have very similar frequencies and intensities. These peaks are not distinguishable experimentally [see Fig. 3(b)]. This was somehow expected, as both Gly and Pro C=O groups are within a tertiary amide peptide bond and both are engaged in H-bonds by two water molecules, as schematized in Fig. 4.

As we have already pointed out, the assignment of Pro(X) to the central peak is supported by the connection between its variation of atomic surroundings and the variation of the amide I band during the unfolding process. Actually, during unfolding, not only the Pro(X) C=O group but also the Gly one notably changes the environment (see Fig. 4). The C=O (Gly) changes its solvation status by passing from being engaged by one to two water molecules. The C=O (Pro) moves from a weakly bonded environment, due to C–H/N–H contacts, to two H-bonds with water molecules.

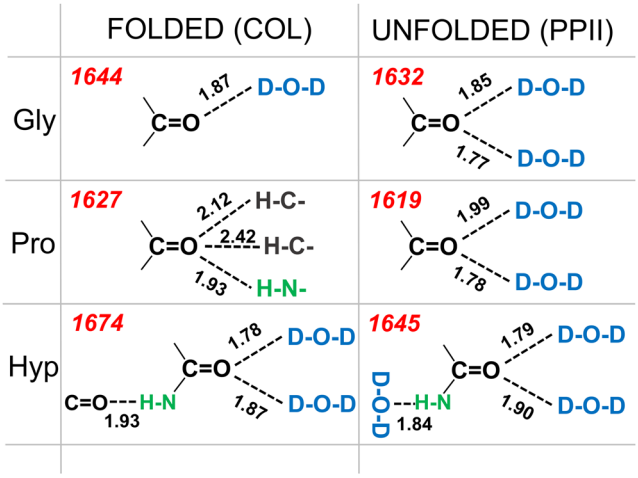

FIG. 4. C=O stretching frequencies (in red color and in cm^{-1}) for the folded [COL(GPO)] and unfolded (PPII) collagen in D_2O. The nearest intermolecular contacts (in Å) as dashed lines.

Conversely, the C=O (Hyp) is always solvated by two water molecules, but with a different geometrical organization.

Interestingly, during unfolding, the largest red shift (≈20 cm^{-1}) is computed for the Hyp residue whose C=O environment appears to remain the same, compared to that of Gly and Pro C=O groups (with a computed red shift on only ≈8 cm^{-1}) (see Fig. 4). To clarify the scenario, we have systematically decomposed the various effects driving the variations on C=O (Hyp) vibrational stretching for folded and unfolded collagens by adopting N-methyl acetamide as a simpler model to understand the solvation around the C=O group in the (periodic) COL and PPII cases (Fig. 5).

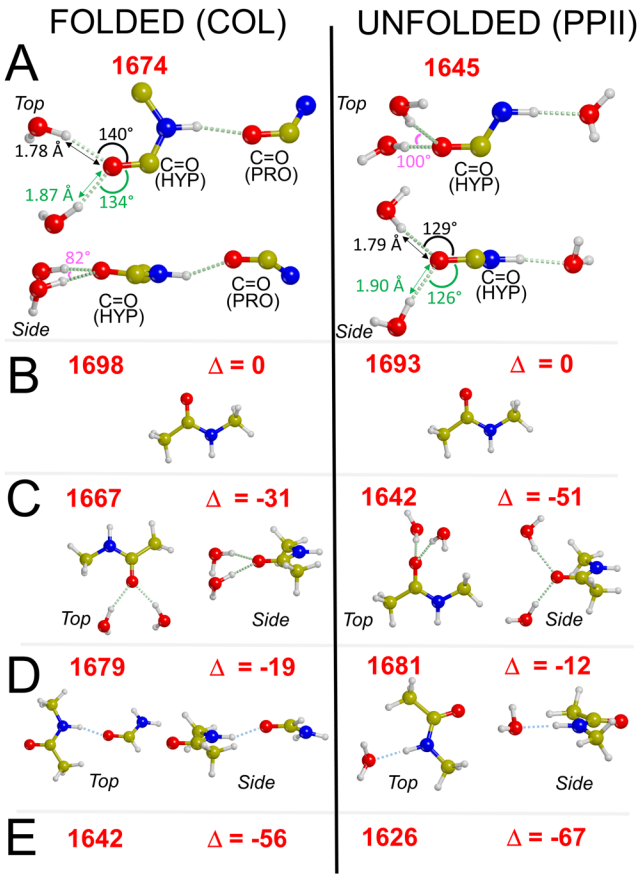

FIG. 5. Extensive analysis of the C=O stretching frequency of the Hyp residue for folded to unfolded collagens. In bold red, the computed C=O stretching frequencies (scaled) for the different chemical systems analyzed with the difference Δ with respect to the C=O frequency of the free molecular models (see the main text). (a) Folded [COL(GPO)] and unfolded (PPII) collagen protein models. In the image, we reported only the key geometrical contacts for describing the C=O groups' environment within the COL(GPO) and PPII geometries along with the H bond distance and angle. (b) Free N-methyl acetamide molecule. (c) Effect of direct hydration on C=O stretching frequency. (d) Effect of N–H group solvation on C=O stretching. For the COL case, to replicate the interaction of N–H with the amidic C=O group, we have employed the formamide molecule. (e) Resulting C=O stretching frequencies derived from the simultaneous C=O and N–H combination of the hydration effects [(c) and (d)]. Frequencies and shifts (Δ) are reported in cm^{-1}. All frequencies are computed considering water molecules as D_2O. Color code: oxygen is in red, hydrogen is in white, nitrogen is in blue, and carbon is in cyan.

To have results consistent with the optimized COL and PPII structures, we have constrained all atoms' positions of the reduced model to those of COL and PPII cases, but for the C=O group and the H atoms added to the terminal C atoms. Indeed, the H-bond distances and angles for the model deviate by only ±0.01 Å and ±2° from the full COL and PPII periodic models. The vibrational frequencies are computed only for the C=O group considered as a fragment; thus, a direct comparison with the case in Fig. 5(a) is not possible. Interestingly, the C=O (Hyp) for PPII makes notably stronger H-bonds with solvation waters, with respect to the COL case [see Figs. 5(b) and 5(c)]. The N–H is involved in an almost equally strong H-bond in COL and PPII, which, due to the electronic resonance within the peptide moiety, weakens the C=O bond strength and its stretching frequency [see Fig. 5(d)]. These geometrical variations lead to a relevant lowering of C=O (Hyp) stretching frequency in the unfolded polymer.

In summary, our simulation proves that the increased level of solvation at the C=O (Gly) can lead to the observed central peak intensity reduction observed during the collagen experimental denaturation.[18] These results are in line with the assignment of the central peak to the C=O stretching of the Gly amino acid. Interestingly, all computed vibrational C=O stretching frequencies suffer from a bathochromic shift in the unfolded state (see Fig. 4). This could indicate that C=O groups are involved in stronger H-bonds when in a single strand than when organized as a triple helix. The calculated shift is qualitatively in line with the observed experimental shift (see Fig. 3), while a point-to-point analysis of Lazarev *et al.* is available in the supplementary material. Re-analyzing their work in the light of our results validates our conclusion.

As the next point of analysis, we have focused on the amide II band of the collagen. Unfortunately, we could not find experimental results with high accuracy spectra on this interesting region to make a detailed comparison with our results. Nevertheless, our simulations indicate the amide II band to be less intense than the amide I band, in agreement with most of the experiments on collagen protein.[15] This band is mainly constituted by N–H bending modes, and its vibrational frequency is centered at 1616 cm^{-1}. Experimentally, this band falls into the 1600–1500 cm^{-1} region. The computed red shift is in line with the expected error by DFT simulations, as seen in the amide I case. Applying the same scaling factor used in the amide I band analysis (0.9815), this band peak falls in the expected region (1586 cm^{-1}). In Fig. S3, we have reported the whole IR spectra of COL 7/3 model in which the location of the amide II band is highlighted.[20]

As a further step, we have analyzed the amide III band of the collagen IR spectrum, as a valuable complementary tool to amide I in protein structural analysis. We have compared the computed amide III bands for the collagen in H_2O with the experimental value of the collagen protein in solution at 20 °C (see Fig. 6). The agreement on the relative peak intensity and frequency is remarkable (1.0118 scaling factor applied in this case). We believe that the very high percentage of Pro and Hyp in the collagen, in X and Y positions, i.e. 28% and 38%, respectively, is responsible for this very good agreement of amide III bands.

Indeed, more than half (55%) of all amino acids in collagen is constituted by only Gly, Pro, and Hyp. In agreement with Ref. 23,

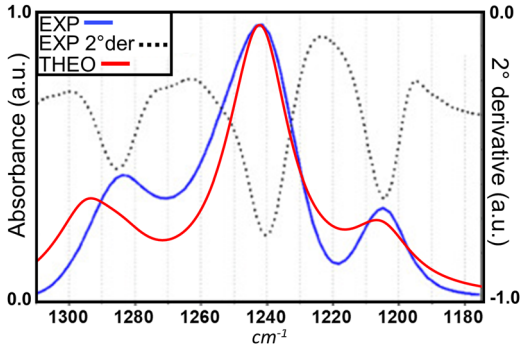

FIG. 6. Amide III region of the COL(GPO) IR spectra in H_2O (frequency scale in cm^{-1}). In blue, the experimental data for a real collagen protein (20 °C) from Ref. 23 with their relative second derivatives as dotted lines. The computed intensities are scaled to match the experimental ones as close as possible. Full width at half maximum of the Gaussian functions is set to 20 cm^{-1} to better mimic the experimental results. The computed frequencies are scaled by a factor of 1.0118 to overlap the highest intensity computed peak with the respective one in the experiments.

we have found that the higher frequency peak comes from the C_α–H bending of Hyp(Y) bonded to C=O of Pro(X). Experimentally, the lowest frequency peak comes from vibrations located on the lateral chains of X and Y amino acids. Our simulation agrees with this, showing that this peak arises from C–N and C–C stretching, and C–H bending of Pro(X) and Hyp(Y) side chains. The experimental analysis assigns the most intense peak (central in frequency) to the N–H bending. Our simulation indicates that this band has also a high component of N–C stretching [of Pro(X) and Hyp(Y)]. The shoulder located at 1268 cm^{-1} on the experimental spectra (more evident in the second derivative curve) is usually assigned to the random coil conformations (for globular proteins). In Ref. 23, they extend this assignation also to collagens. Our results disagree with this interpretation; indeed, our collagen model is crystalline but clearly shows this band, which arises from a combination of N–H and C_α–H bending modes.

Finally, we have analyzed the features of the collagen COL(GPP) IR spectrum in the THz region. In this case, there are no experimental data available to compare with. Therefore, we compared the THz spectra for different helical geometries with the purpose to support the assignment of future experiments. We compared both the high-frequency (amide I–III) and the low-frequency (THz) regions of the IR spectrum for 7/2 and 10/3 collagens (see Fig. 7). In this case, we have employed collagen models in the gas phase with GPP composition, whose helicity we have previously predicted is in agreement with experiments.[58] Our predictions indicate that the collagen helicity only minimally affects the amide I–III bands [Figs. 7(a) and 7(b)]. This arises from the small differences on the local geometry between the two helices. Conversely, the THz spectral features are very sensitive to the overall helical organization [Fig. 7(c)], giving distinct spectra features for 7/2 and 10/3 collagens. This is the first evidence that THz spectroscopy can distinguish the collagen helicity.

Furthermore, we have analyzed that the low frequency normal modes giving rise to the THz spectra in Figs. 7(a) and 7(b).

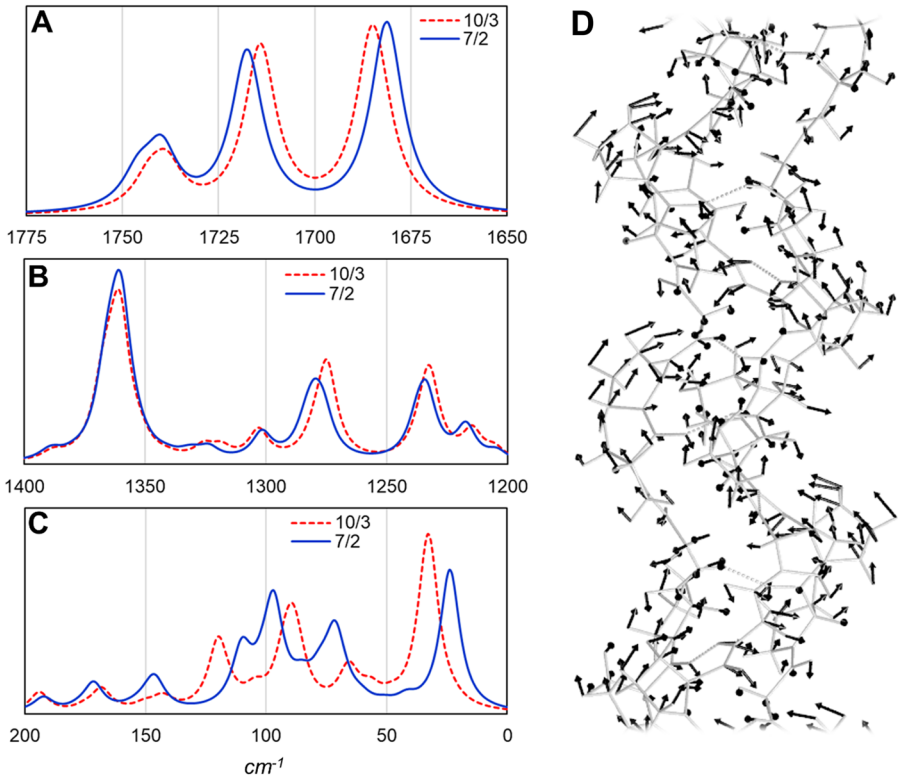

FIG. 7. Amide I (a), amide III (b), and THz (c) regions of the COL(GPP) IR spectra in the gas phase (frequency scale in cm^{-1}). In blue, the 7/2 helix and in red, the 10/3 helix. Full width at half maximum of the Gaussian functions is set to 10 cm^{-1}. The peak intensity is normalized for the number of triplets in the models. (d) Normal mode graphical representation of the vibration located at 33 cm^{-1} for the 10/3 GPP collagen model in the gas phase. The black arrows give an indication of the motion of the atoms within the normal vibrational mode.

The vibrations of this region of the IR spectra are collective motions, often of difficult interpretation due to mixed motion-types. Nevertheless, it is interesting to note that the normal mode corresponding to the first peak of the spectra of both helices [located at 33 and 24 cm^{-1} for 10/3 and 7/2 helices, former reported in Figs. 7(d) and S2] resembles a spring like motion in which the proteins elongate and compress along the periodic direction. This type of motion has already been detected by DFT simulations in poly-Pro crystals.[26] Other low frequency normal modes involve the motions of backbone dihedral angles often belonging to the Pro(Y) residue, the more exposed residue of the protein, which may possibly make it more flexible and thus more active in this spectral region. These complex collective motions sometimes invoke a simple interpretation, such as the 93 cm^{-1} peak of the 10/3 helix. In this collective motion, the three collagen chains vibrate rigidly in the direction normal to the polymer periodicity, expanding or compressing the helix radius.

In principle, the simulation of THz spectra could be employed to determine the collagen packing structure *a priori*. Unfortunately, the collective nature of the internal vibrations contributing to this spectral region is very sensitive to (i) the lateral interactions between protein chain residues and the backbone itself, (ii) the solvation structure of the water molecules, and (iii) the specific amino acid composition. Under this condition, we believe that the DFT simulation of THz spectra can be of greatest support in helping disentangling the distinct vibrational modes within a collagen molecular crystal.

V. CONCLUSIONS

In this paper, we adopted a triple helix infinite extended polymer envisaging the repetition of either Gly-Pro-Hyp (GPO) or Gly-Pro-Pro (GPP) amino acid triplet to elucidate its vibrational features at the DFT level on the most stable structures provided by previous studies.[57,63] Simulations were carried out for collagens, either in the gas phase or in a water micro-solvated environment. We used the hybrid B3LYP functional with a polarized triple-zeta quality Gaussian basis set supplemented by the London's dispersion contribution through the Grimme's D3 empirical correction. Our scaled harmonic spectra are in very good agreement with the experiments and allow for the peak assignment of the collagen amide I and III bands, supporting or questioning the experimental interpretation by means of a detailed vibrational normal mode analysis. The good agreement, despite the omission of both the anharmonicity and temperature dependent dynamical effects, probably means that both effects have a minor influence on the final spectra or, at least, are well within the scaling factor adopted to overcome anharmonicity and small empirical band width associated with each harmonic frequency. For the collagen GPP model, we showed that IR spectroscopy in the THz region can detect the small variations on the triple helix helicity (10/3 over 7/2), therefore elucidating the packing state of the collagen without resorting to x-ray diffraction experiments.

The same computational protocol and analysis of the computed results can be extended to collagen triple helices with different

compositions, helical packing,[58] solvation, applied compression, or stretching load as well as in interaction with surfaces, for instance, to shed some light on the complex collagen hydroxyapatite interface.[65]

SUPPLEMENTARY MATERIAL

In the supplementary material, we provide further details about the coefficients and exponents of the adopted Gaussian basis sets in the CRYSTAL17 code; the comparison of the IR amide I band for GPP and GPO in the gas phase (Fig. S1); a short paragraph expanding results, discussion, and direct analysis with the experimental data; and Fig. S2 showing the normal mode associated with the THz frequency region for the 10/3 GPP collagen model in the gas phase.

ACKNOWLEDGMENTS

M.C. acknowledges the generous allowance of CINECA computing time from ISCRA B (Project No. ISB16; Account ID: MACBONE, Origin ID: HP10BAL7D8), acknowledges the C3S Competence Centre for scientific computing of the University of Torino, and thanks B. Civalleri and A. Erba for supporting the usage of the CRYSTAL17 program.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its supplementary material.

REFERENCES

[1] J. Bella, Biochem. J. **473**, 1001 (2016).
[2] M. D. Shoulders and R. T. Raines, Annu. Rev. Biochem. **78**, 929 (2009).
[3] J. A. M. Ramshaw, N. K. Shah, and B. Brodsky, J. Struct. Biol. **122**, 86 (1998).
[4] P. M. Cowan, S. Mcgavin, and A. C. T. North, Nature **176**, 1062 (1955).
[5] Y. Zhang, M. Herling, and D. M. Chenoweth, J. Am. Chem. Soc. **138**, 9751 (2016).
[6] A. J. Kasznel, Y. Zhang, Y. Hai, and D. M. Chenoweth, J. Am. Chem. Soc. **139**, 9427–9430 (2017).
[7] Y. Zhang, R. M. Malamakal, and D. M. Chenoweth, J. Am. Chem. Soc. **137**, 12422–12425 (2015).
[8] A. J. Kasznel, T. Harris, N. J. Porter, Y. Zhang, and D. M. Chenoweth, Chem. Sci. **10**, 6979 (2019).
[9] T. Harris and D. M. Chenoweth, J. Am. Chem. Soc. **141**, 18021 (2019).
[10] J. Egli, C. Siebler, M. Köhler, R. Zenobi, and H. Wennemers, J. Am. Chem. Soc. **141**, 5607 (2019).
[11] I. C. Tanrikulu and R. T. Raines, J. Am. Chem. Soc. **136**, 13490 (2014).
[12] I. C. Tanrikulu, W. M. Westler, A. J. Ellison, J. L. Markley, R. T. Raines, I. C. Tanrikulu, W. M. Westler, A. J. Ellison, J. L. Markley, and R. T. Raines, J. Am. Chem. Soc. **142**, 1137 (2020).
[13] R. Berisio, L. Vitagliano, L. Mazzarella, and A. Zagari, Protein Sci. **11**, 262 (2002).
[14] J. P. R. O. Orgel, T. C. Irving, A. Miller, and T. J. Wess, Proc. Natl. Acad. Sci. U. S. A. **103**, 9001 (2006).
[15] T. Riaz, R. Zeeshan, F. Zarif, K. Ilyas, N. Muhammad, S. Z. Safi, A. Rahim, S. A. A. Rizvi, and I. U. Rehman, Appl. Spectrosc. Rev. **53**, 703 (2018).
[16] B. De Campos Vidal and M. L. S. Mello, Micron **42**, 283 (2011).
[17] K. J. Payne and A. Veis, Biopolymers **27**, 1749 (1988).
[18] Y. A. Lazarev, B. A. Grishkovsky, and T. B. Khromova, Biopolymers **24**, 1449 (1985).
[19] M. A. Bryan, J. W. Brauner, G. Anderle, C. R. Flach, B. Brodsky, and R. Mendelsohn, J. Am. Chem. Soc. **129**, 7877 (2007).
[20] J. Krishnamoorthi, P. Ramasamy, V. Shanmugam, and A. Shanmugam, Biochem. Biophys. Rep. **10**, 39 (2017).
[21] R. J. Jakobsen and F. M. Wasacz, Appl. Spectrosc. **44**, 1478 (1990).
[22] B. R. Singh, D. B. DeOliveira, F.-N. Fu, and M. P. Fuller, Proc. SPIE **1890**, 47 (1993).
[23] C. Stani, L. Vaccari, E. Mitri, and G. Birarda, Spectrochim. Acta, Part A **229**, 118006 (2020).
[24] T. M. Korter, R. Balu, M. B. Campbell, M. C. Beard, S. K. Gregurick, and E. J. Heilweil, Chem. Phys. Lett. **418**, 65 (2006).
[25] M. R. Kutteruf, C. M. Brown, L. K. Iwaki, M. B. Campbell, T. M. Korter, and E. J. Heilweil, Chem. Phys. Lett. **375**, 337 (2003).
[26] M. T. Ruggiero, J. Sibik, R. Orlando, J. A. Zeitler, and T. M. Korter, Angew. Chem., Int. Ed. **55**, 6877 (2016).
[27] H. Hoshina, S. Ishii, S. Yamamoto, Y. Morisawa, H. Sato, T. Uchiyama, Y. Ozaki, and C. Otani, IEEE Trans. Terahertz Sci. Technol. **3**, 248 (2013).
[28] A. Rimola, M. Fabbiani, M. Sodupe, P. Ugliengo, and G. Martra, ACS Catal. **8**, 4558 (2018).
[29] C. Domene, C. Jorgensen, and S. W. Abbasi, Phys. Chem. Chem. Phys. **18**, 24802 (2016).
[30] I. Streeter and N. H. De Leeuw, Soft Matter **7**, 3373 (2011).
[31] I. Streeter and N. H. De Leeuw, J. Phys. Chem. B **114**, 13263 (2010).
[32] A. Gautieri, S. Vesentini, A. Redaelli, and M. J. Buehler, Nano Lett. **11**, 757 (2011).
[33] A. K. Nair, A. Gautieri, and M. J. Buehler, Biomacromolecules **15**, 2494 (2014).
[34] A. Masic, L. Bertinetti, R. Schuetz, S.-W. Chang, T. H. Metzger, M. J. Buehler, and P. Fratzl, Nat. Commun. **6**, 5942 (2015).
[35] M. Tsai, Y. Xu, and J. J. Dannenberg, J. Am. Chem. Soc. **127**, 14130 (2005).
[36] I. Brand, F. Habecker, M. Ahlers, and T. Klüner, Spectrochim. Acta, Part A **138**, 216 (2015).
[37] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[38] C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
[39] A. Schäfer, C. Huber, and R. Ahlrichs, J. Chem. Phys. **100**, 5829 (1994).
[40] A. Schäfer, H. Horn, and R. Ahlrichs, J. Chem. Phys. **97**, 2571 (1992).
[41] R. Dovesi, A. Erba, R. Orlando, C. M. Zicovich-Wilson, B. Civalleri, L. Maschio, M. Rérat, S. Casassa, J. Baima, S. Salustro, and B. Kirtman, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**, e1360 (2018).
[42] M. Cutini, M. Corno, and P. Ugliengo, J. Chem. Theory Comput. **13**, 370–379 (2017).
[43] M. Cutini, B. Civalleri, and P. Ugliengo, ACS Omega **4**, 1838–1846 (2019).
[44] M. Cutini, I. Bechis, M. Corno, and P. Ugliengo, J. Chem. Theory Comput. **17**, 2566 (2021).
[45] M. Cutini, L. Maschio, and P. Ugliengo, J. Chem. Theory Comput. **16**, 5244 (2020).
[46] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).
[47] Y. Muto, Proc. Phys.-Math. Soc. Jpn. **17**, 629 (1943).
[48] B. M. Axilrod and E. Teller, J. Chem. Phys. **11**, 299 (1943).
[49] C. G. Broyden, IMA J. Appl. Math. **6**, 76 (1970).
[50] R. A. Fletcher, Comput. J. **13**, 317 (1970).
[51] D. F. Shanno and P. C. Kettler, Math. Comput. **24**, 657 (1970).
[52] R. Dovesi, V. R. Saunders, C. Roetti, R. Orlando, C. M. Zicovich-Wilson, F. Pascale, B. Civalleri, K. Doll, N. M. Harrison, I. J. Bush, P. D'Arco, M. Llunell, M. Causà, and Y. Noël, *CRYSTAL17 User's Manual* (Università di Torino, Torino, Italy, 2017).
[53] P. Pulay, J. Comput. Chem. **3**, 556 (1982).
[54] Y. Noel, C. M. Zicovich-Wilson, B. Civalleri, P. D'Arco, and R. Dovesi, Phys. Rev. B **65**, 014111 (2001).
[55] P. Ugliengo, D. Viterbo, and G. Chiari, Z. Kristallogr. -Cryst. Mater. **207**, 9 (1993).
[56] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).

[57]M. Cutini, M. Bocus, and P. Ugliengo, J. Phys. Chem. B **123**, 7354–7364 (2019).
[58]M. Cutini, S. Pantaleone, and P. Ugliengo, J. Phys. Chem. Lett. **10**, 7644 (2019).
[59]K. Okuyama, M. Takayanagi, T. Ashida, and M. Kakudo, Polym. J. **9**, 341 (1977).
[60]K. Okuyama, K. Miyama, K. Mizuno, and H. P. Bächinger, Biopolymers **97**, 607 (2012).
[61]K. Okuyama, Connect. Tissue Res. **49**, 299 (2008).
[62]J. Bella, J. Struct. Biol. **170**, 377 (2010).
[63]M. Cutini, M. Bocus, and P. Ugliengo, "Understanding the reasons of proline hydroxylation induced hyperstability in collagen triple helix by DFT simulations" (unpublished).
[64]M. Moradi, V. Babin, C. Roland, T. A. Darden, and C. Sagui, Proc. Natl. Acad. Sci. U. S. A. **106**, 20746 (2009).
[65]M. Cutini, M. Corno, D. Costa, and P. Ugliengo, J. Phys. Chem. C **123**, 7540–7550 (2019).