Frenet frame analysis of protein geometry: Hints for secondary structure assignments

M. Prados,* M. D. Hernández,† and F. de Soto‡
Departamento de Sistemas Físicos, Químicos y Naturales, Pablo de Olavide University, Seville, Spain

(Received 9 December 2025; revised 15 April 2026; accepted 23 July 2026; published 2 September 2026)

We analyze the protein secondary structure using the Frenet frame to describe the curvature and torsion of the discrete curve formed by the protein α-carbons. We show how a simple criterion based on the evaluation of Frenet frame rotation along the discrete curve supports identification of alpha-helices, beta-strands, and selected supersecondary structures in proteins. Moreover, the description of proteins as fixed points of an effective action inspired by a $U(1)$ gauge model is supported by the analysis of a large dataset of proteins in the Protein Data Bank.

DOI: 10.1103/zs7r-sql6

I. INTRODUCTION

Proteins are one of the four major macromolecules that direct life. The protein structure is described at four different levels; the arrangement of amino acids in a polypeptide chain is called its primary structure. The protein secondary structure refers to the way the primary structure of a protein arranges itself as a result of regular hydrogen bonds forming between the backbone carbonyl and amino groups of each peptide bond. The secondary structure corresponds to local conformations stabilized mainly by hydrogen bonds, such as α-helices or β-strands. The tertiary structure represents the overall three-dimensional folding of a single polypeptide chain, including interactions between secondary structural elements. Finally, the quaternary structure involves the assembly of multiple polypeptide chains into a functional protein complex [1].

Knowing the protein structure is crucial for understanding its functions. Experimental characterization of protein structure is based on techniques such as x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (Cryo-EM). These methods provide high-resolution insights into the three-dimensional conformation of proteins [2].

However, predicting protein secondary and tertiary structure from its primary structure remains a major scientific challenge due to the immense conformational space and the complexity of intramolecular interactions. Despite remarkable progress, accurately modeling secondary, tertiary, and quaternary structures continues to be difficult because of the dynamic nature of proteins and the influence of the environment [3]. In recent years, advances in computational biology have transformed the field. Deep learning algorithms such as AlphaFold have dramatically improved the accuracy of structure prediction by leveraging large datasets and sophisticated neural network architectures to predict protein conformations with unprecedented precision [4].

Secondary structures like α-helices and β-strands are fundamental motifs. In our current understanding of proteins, the assignment of secondary structures relies on the atomic coordinates determined experimentally, using algorithms such as the Dictionary of Secondary Structure in Proteins (DSSP) [5], that has become a standard in the field. The DSSP identifies structural elements based on hydrogen bond patterns and backbone geometry as characterized by the electrostatic energy between $C=O$ and $N-H$ groups. Specific bonding patterns define α-helices ($i \to i+4$), 3_{10}-helices ($i \to i+3$), π-helices ($i \to i+5$), and β-sheets (interstrand bonds). Ramachandran angles (ϕ, ψ) are used as complementary criteria, and a minimum number of consecutive residues must follow the same pattern to be considered a stable secondary structure.

A large dataset of protein data is stored in the Protein Data Bank (PDB) [6], including atomic positions for each amino acid in the protein, as well as annotations including the information on the secondary structure of the protein, typically generated using standard algorithms, mainly DSSP. Other methods for assigning secondary structures use geometric data of the protein backbone; for example, King and Johnson [7] introduced a geometric approach to secondary structure assignment based on backbone dihedral angles and interatomic distances, rather than hydrogen-bond energies. Their method analyzes characteristic angular and distance patterns along the peptide backbone to classify regions as α-helices, 3_{10}-helices, β-strands, or turns.

A different perspective is provided by the Frenet-frame analysis of polymers [8–12]. This idea has been applied to proteins by A.J. Niemi and collaborators in a series of papers [13–20] where they analyzed protein characteristics based on the geometry of the alpha-carbon backbone using the Frenet frame. Following these works, protein backbones can be understood as links whose degrees of freedom are curvature and torsion. This offers a simplified frame to analyze, from a purely geometric point of view, the polymer structure, where specific secondary structures can be linked to the fixed points of a gauge model of proteins.

*Contact author: manuprad2003@gmail.com
†Contact author: mdhernandezdelatorre@gmail.com
‡Contact author: fcsotbor@upo.es

Our goal in this paper is to use the Frenet frame description of protein backbone geometry to delve into protein structure characterization, showing that this framework is useful for characterizing secondary structures such as α-helices, β-strands, or turns. The structure of the paper is as follows: in Sec. II, we review a three-dimensional curve description in terms of the Frenet frame. In Sec. III, we apply this frame to analyze typical protein secondary structures, delve into the understanding of this curve as a gauge model in terms of its curvature and torsion, and propose a method based solely on the geometry of the α-carbon backbone structure to identify secondary structures. Finally, we present our conclusions in Sec. IV.

II. FRENET FRAME

A. Frenet frame in the continuum

The Frenet trihedron provides a framework for describing three-dimensional curves in terms of their curvature and torsion, parameters that quantify the variations of the trihedron vectors along the curve. The trihedron is formed by three orthogonal vectors: the tangent vector $\vec{T}$, the normal vector $\vec{N}$, and the binormal vector $\vec{B}$. Their derivatives with respect to the arc length s are given by

$$\frac{d\vec{T}}{ds} = \kappa\vec{N}, \tag{1}$$

$$\frac{d\vec{N}}{ds} = -\kappa\vec{T} + \tau\vec{B}, \tag{2}$$

$$\frac{d\vec{B}}{ds} = -\tau\vec{N}, \tag{3}$$

that can be compactly written as

$$\frac{d}{ds}\begin{pmatrix}\vec{T}\\ \vec{N}\\ \vec{B}\end{pmatrix} = \begin{pmatrix}0 & \kappa & 0\\ -\kappa & 0 & \tau\\ 0 & -\tau & 0\end{pmatrix}\begin{pmatrix}\vec{T}\\ \vec{N}\\ \vec{B}\end{pmatrix}. \tag{4}$$

The curve is uniquely defined in three-dimensional space, up to a rigid rotation, given the curvature and torsion along the curve, $\kappa(s)$ and $\tau(s)$.

The Frenet frame described by Eq. (4) can be understood as a particular choice of a larger class of trihedrons describing the curve, where the vectors belonging to the orthogonal plane of the curve are written as the complex combinations

$$\vec{E}^{\theta}_{\pm} = e^{i\theta}(\vec{N} \pm i\vec{B}). \tag{5}$$

The angle of rotation θ redefines the normal and binormal vectors of the Frenet frame, keeping the curve unchanged, hence redefining the curvature and torsion as $k_\theta = e^{i\theta}\kappa$ and $\tau_\theta = \tau + \frac{d\theta}{ds}$. In this sense, the rotation can be understood as a U(1) gauge transformation that leaves the curve unchanged. The existence of this gauge symmetry led the authors of [13–15] to propose a Landau-Ginzburg energy functional $E = \int_0^L ds\, \mathcal{E}(\kappa[s], \tau[s])$:

$$\mathcal{E}(\kappa, \tau) = \left(\left(\frac{d}{ds} - i\tau_\theta\right)\kappa_\theta\right)^2 + c\left(\kappa_\theta^2 - \mu^2\right)^2 + d\tau_\theta + \cdots, \tag{6}$$

analogous to an Abelian Higgs model where κ_θ and τ_θ play the role of scalar and gauge fields, respectively.

The curve described by the polymer backbone, such as a protein, is nonetheless discrete, with C_α–C_α distances showing a sharp distribution around 3.8 Å. We will therefore resort to the description of this discrete curve within the Frenet frame.

B. Frenet frame for a discrete curve

The discrete set of positions occupied by the protein alpha carbons C_α defines a discrete curve in three-dimensional space that can be described using the Frenet frame. At this level, the protein backbone is simplified to a set of discrete links connecting each pair of consecutive carbons C_α. Each link is defined by the angles that measure how this link bends with respect to the preceding one. The protein can therefore be understood as a gauged model of links whose relative orientations give rise to the three-dimensional geometry of the curve.

A discrete version of Eqs. (1)–(3) can be obtained by replacing the derivatives with finite differences. This leads to the following relations between the Frenet vectors at consecutive points along the discrete curve:

$$\vec{T}_{i+1} \propto \vec{T}_i + s_i\kappa_i\vec{N}_i, \tag{7}$$

$$\vec{N}_{i+1} \propto \vec{N}_i - s_i\kappa_i\vec{T}_i + s_i\tau_i\vec{B}_i, \tag{8}$$

$$\vec{B}_{i+1} \propto \vec{B}_i - s_i\tau_i\vec{N}_i, \tag{9}$$

where $\vec{T}_i$ defines the direction of the vector between the two adjacent positions i and $i+1$ of the discrete curve $\vec{R}_i$ and $\vec{R}_{i+1}$, i.e.,

$$\vec{T}_i = \frac{1}{s_i}(\vec{R}_{i+1} - \vec{R}_i), \tag{10}$$

and $s_i = |\vec{R}_{i+1} - \vec{R}_i|$ is the distance between two consecutive points. The normal and binormal vectors can then be defined from the tangent vectors $\vec{T}_i$ as

$$\vec{B}_i = \frac{\vec{T}_{i-1} \times \vec{T}_i}{|\vec{T}_{i-1} \times \vec{T}_i|}, \tag{11}$$

and

$$\vec{N}_i = \vec{B}_i \times \vec{T}_i. \tag{12}$$

In Eqs. (7)–(9), the proportionality symbol has been used to emphasize that the vectors at position $i+1$ are normalized, and in the discrete case ($s \neq 0$) an explicit normalization factor is therefore required.

The finite rotation of the Frenet frame from bond i to $i+1$ can be characterized using two angles, as described in Appendix A. Relying on Eqs. (7)–(9), we have chosen to characterize the three-dimensional curve using

$$\overline{\kappa}_i \equiv -\vec{T}_i \cdot \vec{N}_{i+1}, \tag{13}$$

$$\overline{\tau}_i \equiv -\vec{N}_i \cdot \vec{B}_{i+1}, \tag{14}$$

that, for a discrete but smooth curve characterized by Frenet vectors almost parallel at neighboring sites (or $s_i\kappa_i \ll 1$, $s_i\tau_i \ll 1$), would correspond to the dimensionless products of

TABLE I. Values of the curvature $\overline{\kappa}$ and torsion $\overline{\tau}$ and the corresponding bond ψ and dihedral θ angles in typical secondary structures in proteins [17].

Structure	Curvature	Torsion
α-helix	$\overline{\kappa} = \pm 1$ $(\psi = \pm\pi/2)$	$\overline{\tau} = \sin(1) \approx 0.84$ $(\theta = 1)$
β-strand	$\overline{\kappa} = \pm\sin(1) \approx \pm 0.84$ $(\psi = 1)$	$\overline{\tau} = 0$ $(\theta = \pi)$

1: Compute $\vec{T}_i$ using Eq. (10).
2: Compute $\vec{B}_i$ using Eq. (11).
3: **if** $\vec{B}_{i-1} \cdot \vec{B}_i < 0$ **then**
4: $\quad \vec{B}_i \to -\vec{B}_i$
5: **end if**
6: Compute $\vec{N}_i$ using Eq. (12).
7: Compute $\overline{\kappa}_i$, $\overline{\tau}_i$ using Eqs. (13,14).

FIG. 1. Algorithm followed to compute the Frenet frame vectors, including the sign flip in the binormal vectors.

the curvature and torsion times the bond length: $\overline{\kappa}_i \approx s_i\kappa_i$ and $\overline{\tau}_i \approx s_i\tau_i$ (see Appendix B).

As $s_i \approx s = 3.80$ Å for the vast majority of $C_\alpha - C_\alpha$ links, as discussed in Appendix B, we will work with the dimensionless parameters $\overline{\kappa}$ and $\overline{\tau}$, and the dimensions of κ and τ are provided by this distance that sets the typical length of the curve's characteristics. It is important to remark that, despite using the symbols $\overline{\kappa}$ and $\overline{\tau}$, and referring to these parameters as dimensionless curvature and torsion, they correspond to the continuum ones only in the limit of smooth curves.

It is interesting, before moving on, to consider the meaning of these parameters for a polymer structure. In the case of two consecutive tangent vectors that are almost in the same direction $\vec{T}_{i+1} \approx \vec{T}_i$, the normal vector at position i, $\vec{N}_i$ will be almost orthogonal to the tangent vector at the next one $\vec{T}_{i+1}$, and thus $\overline{\kappa} \approx 0$ according to Eq. (13). The same will occur for tangent vectors that are opposite to each other. This means that small values of $\overline{\kappa}$ (in absolute value) will correspond to either noncurved geometries or to very large bending, which will be forbidden by molecules' self-avoidance. Concerning torsion, if four consecutive C_α were contained in the same plane, the two binormal vectors corresponding to two central positions would be identical (orthogonal to that plane) and therefore the dimensionless torsion $\overline{\tau}_i$ defined in Eq. (14) would be equal to zero. More details on the angles among the trihedron vectors at consecutive positions of the discrete curve have been included for completeness in Appendix A.

Note that the dimensionless parameters $\overline{\kappa}_i$ and $\overline{\tau}_i$ are closely related to the bond or turning angle ψ_i and torsion or dihedral angle θ_i defined in [17] [see Eqs. (21) and (22) therein]:

$$\cos\psi_i = (\vec{T}_{i+1} \cdot \vec{T}_i) \equiv \rho_i, \tag{15}$$

$$\cos\theta_i = (\vec{B}_{i+1} \cdot \vec{B}_i) \equiv \chi_i. \tag{16}$$

As detailed in Appendix A, our definition in Eqs. (13) and (14) is related to the angles in Eqs. (15) and (16) by $\overline{\kappa}_i = \sin\psi_i$, $\overline{\tau}_i = \sin\theta_i$. A combination of the scalar products in Eqs. (13) and (14) with the ones in (15) and (16), that we denoted ρ_i and χ_i allows a characterization of the angles ψ_i and θ_i without sign ambiguities. These equations imply that for a straight line characterized in the continuum by zero curvature and torsion, one also has that $\overline{\kappa}_i$, $\overline{\tau}_i$, ψ_i, and θ_i vanish, and $\rho_i = \chi_i = 1$, while for the case of typical secondary structures, the values of $\overline{\kappa}$ and $\overline{\tau}$ and their respective bond and dihedral angles have been included in Table I.

The most important difference between the definitions of the bond and dihedral angles in Eqs. (15) and (16) and the curvature/torsion of Eqs. (13) and (14) is that the definition in Eq. (15) is positive definite, while the new definition in Eq. (13) is not. As will be shown below, the sign, combined with a consistent prescription for the orientation of the binormal vectors, provides a powerful tool for identifying secondary structures.

One of the more subtle issues when computing curvature and torsion for a discrete set of points is handling cases where there is an inflection point, where either the normal or binormal vectors change sense (see Fig. 6 of Ref. [16]). Inflection points are inherently difficult to identify and lead to ambiguities in the sign of the curvature [14]. Indeed, since it is not well defined, there is no way of deciding whether the result of Eq. (11) is the most natural choice or a binormal vector in the opposite direction. Both definitions describe the same curve once the curvature and torsion have been computed accordingly using Eqs. (13) and (14), and the different prescriptions amount to deciding between a binormal vector that forms an angle θ larger or smaller than $\pi/2$ with the preceding one. Here, we will adopt the prescription of modifying the definition of $\vec{B}_i$ in Eq. (11), altering its sense so that two consecutive binormal vectors always form an angle smaller than $\pi/2$, and subsequently define the normal vector using Eq. (12). This redefinition amounts to a sign change in the curvature as we reverse the vectors $\vec{N}_i$ and $\vec{B}_i$. This change in the curvature with respect to the direct application of Eqs. (13) and (14) will last until the next change in the definition of a binormal vector, while for the torsion it amounts to a change of sign for amino acid $i+1$. In terms of (ρ_i, χ_i), we restrict ourselves to $\chi_i \geqslant 0$. The algorithm is sketched in Fig. 1.

Although the sign of the curvature is itself not well defined, we have found this definition useful for characterizing secondary structures in proteins. For a helix, for example, the curvature remains either positive or negative along the helix except for the rare cases where the helix has a large bending at some point and appears therefore distorted. For β-strands, in contrast, the curvature obtained has alternating signs at adjacent positions of the curve; this alternating pattern is strong evidence for the presence of β-strands (as will be seen in the next section). As in the case of helices, strongly bent β-strands may not strictly exhibit this alternating sign pattern in $\overline{\kappa}_i$, since the adopted convention for the binormal vector does not always guarantee it under large deformations.

In this work, we will proceed as follows: we use the Protein Data Bank to download the PDB file that contains the positions $\vec{R}_i$ of all the atoms in each one of the polypeptide chains that constitute the protein, and compute the Frenet vectors for each position of the chain, flipping the sense of the

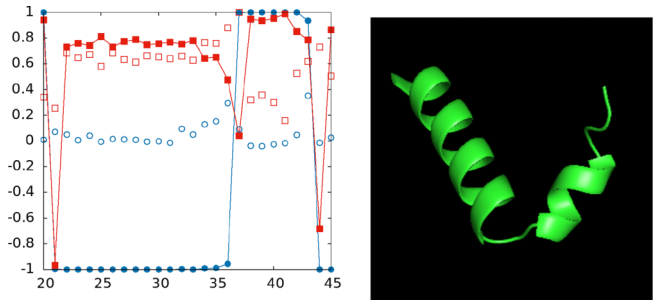

FIG. 2. Transition between an α-helix and a 3_{10} helix in protein 1A3N (deoxy human hemoglobin) around amino acid number 36 as characterized in terms of $\overline{\kappa}_i$, ρ_i (blue-filled and empty circles, respectively) and $\overline{\tau}_i$, χ_i (red-filled and empty squares, respectively) (left) and its 3D counterpart (right).

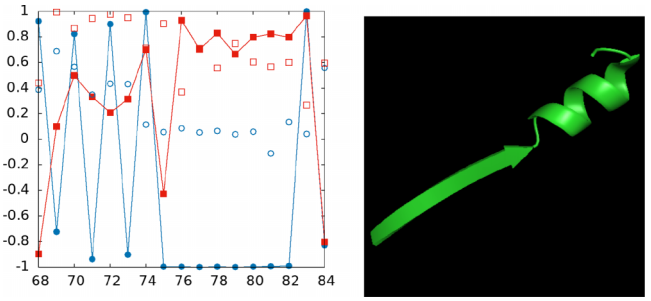

FIG. 3. Transition between a β-strand and an α-helix in protein 2PAB (prealbumin) around amino acid number 75 as characterized in terms of $\overline{\kappa}_i$, ρ_i (blue-filled and empty circles, respectively) and $\overline{\tau}_i$, χ_i (red-filled and empty squares, respectively) (left) and its 3D counterpart (right).

binormal vector following the criterion explained above. Then, we apply our definition for the calculation of the dimensionless curvature and torsion that will characterize the discrete three-dimensional curve. This approach is complementary to the one employed in [21], where it is shown that two variables, related to the bond and torsion angles, are sufficient to satisfactorily reproduce the backbone main characteristics, in contrast to the use of Ramachandran angles ϕ/ψ.

III. RESULTS

A. Secondary and super-secondary structures

The application of our prescription for calculating curvature and torsion allows the identification of some of the secondary structures in proteins, as well as the most common transitions among them. The geometric pattern of helices corresponds to $\overline{\kappa} \approx \pm 1$, with a torsion $\overline{\tau} \approx \sin(1) \approx 0.84$ for α-helices and a larger value for 3_{10}-helices. For β-strands, they will be characterized by a small torsion (theoretically zero for a completely flat β-strand) and a curvature that has alternating signs with $|\overline{\kappa}| \approx \sin(1)$, according to the values in Table I. In this sense, our prescription (13), together with the definition of the binormal vector designed to disentangle inflection points, enables the identification of β-strands, whose characteristic zigzag structure is associated with alternating signs of the curvature.

In Figs. 2–4, we show some of the typical secondary and supersecondary structures that appear in proteins, showing both our results for dimensionless curvature $\overline{\kappa}$ and torsion $\overline{\tau}$ and the three-dimensional representation of these structures, and they exemplify the main secondary structures found in proteins and some of the transitions among them. For comparison, we have included in these plots ρ_i and χ_i.

In Fig. 2, we represent the curvature and torsion of the protein deoxy human hemoglobin (with PDB ID 1A3N) between residues 20 and 45, according to the PDB labeling. The presence of an α-helix spanning residues 21–36 can be clearly identified, characterized by a curvature $\overline{\kappa} \approx -1$ and a positive torsion of about $\overline{\tau} \sim 0.7$. Another helix with $\overline{\kappa} \approx +1$ appears between amino acids 38 and 43. This second helix is characterized by a larger torsion, signaling the fact that this second structure is a 3_{10} helix, as identified in PDB. The transition between the two helices occurs at residues 36–37 and is remarkably fast, in the sense that one helix starts immediately after the previous one ends, without a smooth transition between them.

Figure 3 provides an archetypal example of the transition between a β-strand and an α-helix in prealbumin (protein ID 2PAB in PDB). Before residue number 75, the alternating signs of the curvature indicate the presence of a β-strand,

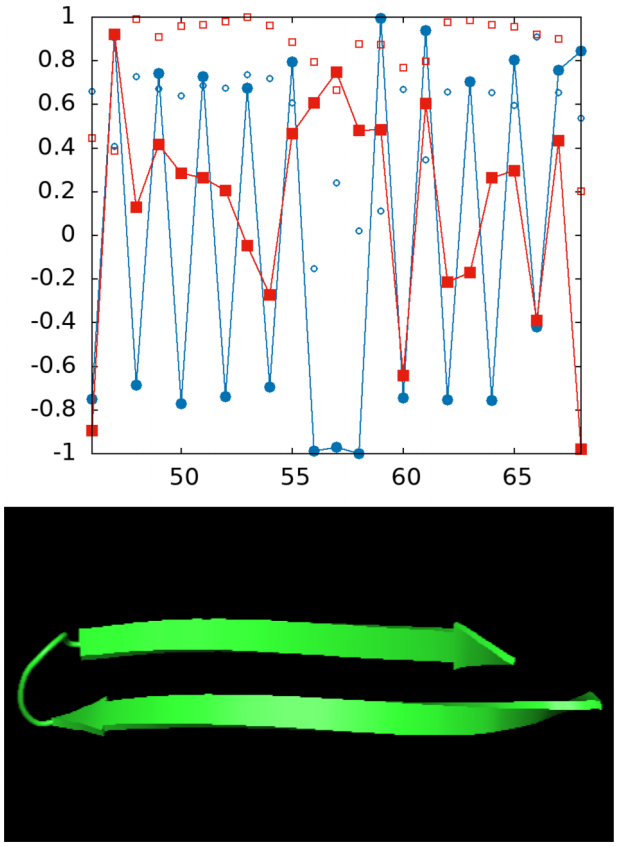

FIG. 4. Hairpin transition between two antiparallel β-strands in protein 3CNA (concanavalin) as characterized in terms of $\overline{\kappa}_i$, ρ_i (blue-filled and empty circles, respectively) and $\overline{\tau}_i$, χ_i (red-filled and empty squares, respectively) (top) and the corresponding three-dimensional structure (bottom).

while after residue 75 the curvature becomes constant, with $\overline{\kappa} \approx -1$, characteristic of a helix.

Finally, in Fig. 4, we have represented an example of the hairpin transition between two antiparallel β-strands in the protein concanavalin (3CNA). The two adjacent antiparallel β-strands are separated by a short turn that connects both strands running in opposite directions, causing the polypeptide chain to bend by approximately 180°. This short loop is crucial for the formation of structures like β-hairpins and enables the antiparallel arrangement of the strands. Many such supersecondary structures can be easily found in a search over PDB files. If one observes the curvature of the residues in the transition region, it is quite similar in all of them, and the data show that they adopt a conformation that is compatible with an α-helix ($\overline{\kappa} \approx -1$ and $\overline{\tau}$ consistent with $\sin(1) \approx 0.84$).

In all these structures, both the presence of secondary structures and the transitions among them can be characterized in terms of $\overline{\kappa}$ and $\overline{\tau}$. The authors of [13,15–19] studied these transitions as soliton solutions of an energy function (17) that includes nearest-neighbor interactions for $\overline{\kappa}$ and on-site interactions for $\overline{\tau}$. The transitions exemplified by Figs. 2–4, however, are too abrupt for the smooth-transition interpretation proposed in Refs. [13,15–19]. Indeed, Fig. 4 shows that the transition region between two β-strands corresponds to another minimum of the energy functional (17). The hairpin region in Fig. 4 is characterized by a curvature close to $|\overline{\kappa}| = 1$, showing that even a small loop can adopt this stable configuration even without hydrogen bonds that are responsible for the stability of helices in proteins.

Of particular interest are supersecondary structures such as α-helice turns or $\alpha - \alpha$ or α–3_{10} transitions such as the one shown in Fig. 2, $\alpha - \beta$ transitions such as the one represented in Fig. 3, or the β-hairpin structure between antiparallel β-strands shown in Fig. 4. It is important to note that identifying these structures in terms of curvature and torsion is made possible by our prescription for the sense of the binormal vector and the definition in Eqs. (13) and (14), which provides additional information with respect to Eqs. (15) and (16) associated with the sign of the curvature. In the present prescription, the curvature–sign ambiguity still persists, but only in the cases of β-strands that suffer a large bending at some point or bent helices; the results deviate from the ones presented above.

B. Distribution of curvature and torsion

The data indicate that the $(\overline{\kappa}, \overline{\tau}) = (\pm 1, -\sin 1)$ configuration is the most stable combination in proteins. This is evidenced by the following analysis: we have randomly downloaded more than 18 000 protein structures from PDB, and computed the curvature and torsion for each of the polypeptide chains contained in these files. We have discarded the chains with distances between consecutive C_α larger than 5 Å or, in general, when the atomic coordinates of residues are missing in the PDB file. This leaves us with more than 70 000 polypeptide chains to analyze. The histogram of the curvature and torsion values obtained for each residue of each chain has been represented in Fig. 5. These histograms show several interesting characteristics: (i) the distribution of curvatures is symmetric in $\overline{\kappa}$, as expected, and (ii) almost 60% of the values

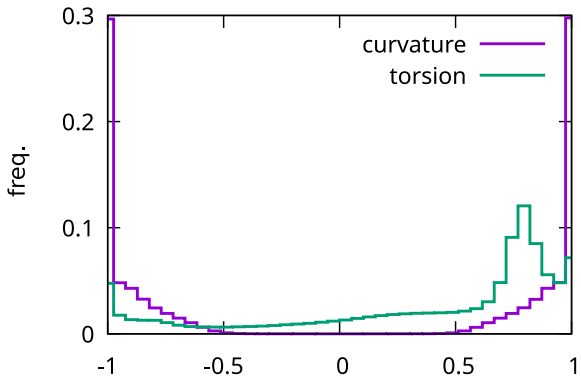

FIG. 5. Histogram of the values of curvatures and torsion obtained over a large number of randomly chosen proteins from PDB.

correspond to $\overline{\kappa} = \pm 1$, which is typical of helices. The large fraction of values with a curvature close to $|\overline{\kappa}| = 1$ manifests that this configuration is very stable and should correspond, in Eq. (17), to its lowest energy minimum.

The histogram of torsion values in Fig. 5 is asymmetric, in contrast to the curvature distribution, showing a preference for positive values of the torsion. A distinct peak at positive values of $\overline{\tau}$ shows that about 40% of the values are contained between $\overline{\tau} = 0.65$ and $\overline{\tau} = 0.95$, compatible with the values of α-helices according to the values in Table I. In any case, the fact that the probability of finding curvature $|\overline{\kappa}| \approx 1$ is much larger than $\overline{\tau} \approx \sin 1$ would indicate that the curve tends to have a large curvature even in geometries that are not helices.

The signature of β-strands in Fig. 5 is less clear than for helices. They would appear as the tails for $\overline{\kappa}$ slightly larger than -1 or slightly smaller than $+1$. In terms of torsion, they would correspond to the region of small $\overline{\tau}$, although this region does not exhibit a distinct peak but rather a smooth distribution.

One more characteristic emerges from Fig. 5: the existence of peaks at $\overline{\tau} = -1$ and $\overline{\tau} = +1$. These peaks are associated with kinks in the curve, where there is a strong bending of the protein, and appear mainly in the turn among different secondary structures.

C. A criterion for identifying secondary structures

Based on the observation that helices are typically characterized by a curvature $\overline{\kappa} = \pm 1$ and large torsion, we define a criterion to identify the presence of such structures based solely on their geometrical features: curvature and torsion. We will adopt the criterion that a helix (either α or 3_{10}) is present whenever the amino acid chain is characterized by three or more amino acids with curvature $|\overline{\kappa}| \in (0.95, 1)$ and torsion $\overline{\tau} > 0.5$. Although 3_{10}-helices are characterized by larger torsion values than α-helices, as illustrated in Fig. 2, it is difficult to establish a criterion to distinguish between them, and therefore we will analyze both types of helices together.

Thus, to identify helices, we search along each protein chain for a sequence of three or more consecutive bonds with $|\overline{\kappa}_i| > 0.95$ and $\tau_i > 0.5$. The whole sequence of consecutive bonds satisfying both conditions will be labeled as a helix (see Fig. 6). For example, for the protein 1A3N represented in Fig. 2, the application of this algorithm identifies helices

for each maximal block B such that $|\overline{\kappa}_i| > 0.95$ and $\overline{\tau}_i > 0.5$ **do**
  if $|B| \geq 3$ **then**
    classify B as helix
  end if
end for

FIG. 6. Minimal set of conditions used to identify helices based on their curvature and torsion.

among residues 20–33 and 36–40, which is consistent with the DSSP assignment. The second helix is a 3_{10}-helix according to DSSP; however, we do not distinguish between α- and 3_{10}-helices. In contrast, the PDB annotates the first helix ending at residue 34 and the second one as spanning residues 36–41.

For β-strands, our criterion requires that $\overline{\kappa}$ alternates in sign over four or more consecutive amino acids, with an average curvature (in absolute value) of $\langle|\overline{\kappa}|\rangle > 0.6$, where $\langle\cdots\rangle$ stands for the average value. With respect to the torsion, for a flat β-strand it would have a vanishing value, and in our criterion we will accept β-strands when the average value of the torsion satisfies: $\langle\overline{\tau}\rangle \in (-0.5, 0.5)$ in the section of the chain with alternating curvature sign. This criterion for the torsion forbids the identification of strands with a large bending, but in any case, these strands are difficult to identify because our criterion for detecting inflection points also fails, and we do not have alternating sign curvature. It is interesting to note that this criterion could have been written in terms of $\chi_i = \pm\sqrt{1-\overline{\tau}_i^2}$ as $|\chi_i| > 0.86$, with the extra information provided by the sign of χ_i, positive in β-strands, but in practice its contribution to the identification does not change the overall results presented throughout the paper.

For identifying β-strands, we start searching a block of four or more consecutive bonds satisfying $k_i \cdot k_{i+1} < 0$ (i.e., alternating signs). Then, for the sequence of consecutive bonds that have alternate signs of $\overline{\kappa}$, compute the average value of this parameter in absolute value $\langle|\overline{\kappa}|\rangle$, and the average value of the torsion in the same sequence of amino acids $\langle\overline{\tau}\rangle$. We label this sequence as a β-strand if $\langle|\overline{\kappa}|\rangle > 0.6$ and $|\langle\overline{\tau}\rangle| < 0.5$ (see Fig. 7). The application of our algorithm to protein 2PAB represented in Fig. 3 finds a β-strand between amino acids 67–74 and a helix between residues 75 and 81, while DSSP identifies a β-strand among residues 67–73 and an α-helix at 75–82. PDB annotations include a β-strand at residues 67–74 and a helix at 75–84. For the hairpin structure in Fig. 4, DSSP identifies β-strands among amino acids 47–55 and 60–66 and our algorithm among amino acids 45–54 and 57–65. In all these cases, we consider that our algorithm identifies within the stated ± 2 residue tolerance the different secondary structures, despite small differences in the starting/ending residues. A full comparison of our method with DSSP will be done in a forthcoming publication.

for each maximal block B with $\overline{\kappa}_i\overline{\kappa}_{i+1} < 0$ **do**
  if $|B| \geq 4$, $\langle|\overline{\kappa}|\rangle_B > 0.6$, $|\langle\overline{\tau}\rangle_B| < 0.5$ **then**
    classify B as β-strand
  end if
end for

FIG. 7. Conditions used to identify β-strands using the backbone curvature and torsion.

These criteria have been established by analyzing a small group of proteins, and later applied to the large statistical set of proteins downloaded randomly from the Protein Data Bank and used to produce Fig. 5, consisting of over 70 000 polymer chains from ∼18 000 proteins. We have checked whether our criteria based solely on the curvature and torsion of the C_α backbone are capable of detecting the presence of the main secondary structures in proteins. To assess the quality of the assignments, we will consider that a positive identification is obtained whenever our criteria lead to the identification of the secondary structure at the same residues labeled in the PDB file, admitting a ± 2 shift in the starting/ending residue. The results show that our criterion identifies correctly 92.6% of helices (jointly including α and 3_{10}), i.e., for 92.6% of the structures labeled either as α or 3_{10} helices in PDB, our algorithm detects the presence of a helix. Concerning β-strands labeled in PDB files, our algorithm is capable of detecting 93.3% of them (again accepting a positive identification regardless of the existence of small differences in the initial and ending residue). The high ratio of β-strands identified is due to the alternating sign $\overline{\kappa}$ discussed in the previous sections and is hence due to our criterion for choosing the sense of the binormal vector. We have checked that varying the threshold for $\overline{\kappa}$ for helices between 0.92 and 0.97 and simultaneously for the torsion between 0.4 and 0.6, the fraction of helices correctly recognized by our method is in the range $90.4\% - 93.8\%$. For β-strands, a similar sensitivity test with curvature thresholds varying between 0.4 and 0.6 and torsion between 0.4 and 0.6 implies a positive identification of strands in the range $84.0\% - 95.0\%$. This high rate of positive identification shows that our definition of curvature and torsion in Eqs. (13) and (14) can be useful for characterizing three-dimensional polymer structure, enabling a simple and fast recognition of common secondary structures. The identification of less common secondary structures as well as a refinement of our criterion for detecting inflection points, taking into account the identification of secondary structures, is left for a future analysis. In a future paper, we will also present a detailed comparison (per residue) of the identification based on our criteria compared to standard DSSP and STRIDE methods.

In order to assess the quality of the proposed criterion, we have selected the work [7] that presents the global amount of residues that belong to each secondary structure for a particular set of proteins. In Table II, we compare the percentage of amino acids belonging to helices (either α or 3_{10}) and β-strands for each one of these proteins, and compare our results with those of [7] and with the corresponding HELIX/SHEET records in the Protein Data Bank. Our method is at least as good as the one in [7], which considers not only the C_α positions, but also geometric parameters derived from backbone dihedral angles and interatomic distances. In their study, they defined secondary structures by analyzing five geometric parameters for each residue: two dihedral angles, $\zeta(i)$, which describe the orientation between consecutive carbonyl groups, and tau $\tau(i)$, which describes the relative position of neighboring C_α atoms; and three interatomic distances, O(i)–N($i+3$), O(i)–N($i+4$), and C(i)–N($i+3$). These values capture the

TABLE II. Percentage of amino acids in the protein that belong either to helices or β-strands, as given by PDB labeling, Ref. [7] and this work.

		Helices			Strands		
Protein	ID	PDB	KJ	This work	PDB	KJ	This work
Azurin	5azu	24.8	17.8	8.0	31.6	33.7	38.8
Bence Jones	1rei	4.8	0.0	0.0	46.2	33.5	50.2
α-Chymotrypsin-a	5cha	11.3	11.0	8.8	45.1	15.8	38.1
Concanavalin-a	1nls	7.3	2.6	1.3	44.0	36.4	59.4
Cytochrome-c	5cyt	51.0	41.7	38.0	0.0	7.7	10.5
Elastase	1esa	7.6	10.1	7.6	50.6	21.3	37.1
Flavodoxin	5fx2	39.6	30.8	38.2	31.9	17.1	33.0
Hemerythrin	1hmd	63.6	71.2	69.1	0.0	3.6	8.4
Hemoglobin	2mhb	77.6	74.6	75.8	0.0	0.0	1.8
Lactate dehydrogenase	6ldh	38.0	41.9	43.3	21.8	14.0	22.2
β-Lactoglobulin	1beb	17.3	12.5	10.1	37.6	33.6	53.1
Lysozyme	4lzt	43.7	41.4	38.1	4.8	3.9	15.1
Myoglobin	1vxf	80.7	80.9	76.7	0.0	0.0	2.7
Papain	9pap	26.3	28.9	22.9	20.5	15.2	16.6
Pepsinogen	3psg	31.1	17.3	10.5	14.9	27.1	41.6
Prealbumin	2pab	8.1	8.5	6.3	50.5	35.3	63.3
Ribonuclease-a	1rbx	21.5	22.0	21.5	29.8	19.1	33.1
Superoxide dismutase	1sxn	6.4	4.3	3.0	36.5	24.0	39.2
T4 lysozyme	5lzm	67.3	66.8	65.4	11.3	6.1	3.8
Thermolysin	1lnf	41.2	39.8	38.7	16.6	13.6	24.0
Triose phosphate isomerase	1ypi	54.5	39.5	40.4	18.9	14.8	23.2
Trypsin	1tng	5.1	11.9	0.0	37.8	19.3	46.9

local conformation and potential hydrogen-bonding patterns of the peptide backbone. Based on specific numerical ranges of these parameters, the algorithm classifies residues into α-helices, 3_{10}-helices, β-strands and turns.

Although individually for each protein, both the results of Ref. [7] and our results differ from the PDB assignments, our criterion, being simple and with only the information on the C_α positions, is comparable by this aggregate metric to theirs, according to the results obtained. Overall, our results differ approximately in 8% for helices and 10% for β-strand localization with respect to the data consigned in PDB, slightly better than theirs. On the other hand, they can distinguish between α and 3_{10}-helices, while based solely on the torsion of the curve, we have difficulties in setting a quantitative criterion for their distinction. In both cases, the comparison is made with respect to the helices and strands labeled in PDB files. It is important to note that some secondary structures may not be annotated in PDB files. For this reason, the analysis presented here focuses on sensitivity, namely the fraction of PDB-annotated secondary structures recovered by our geometrical criterion. A more comprehensive benchmark, including an assessment of false positives and the corresponding confusion matrices, will be the subject of future work using DSSP or STRIDE as the reference.

Finally, it is important to note that the results presented in this section are completely independent of the individual signs of $\overline{k}$ that are not well defined. The identification of β-strands relies on the sign-flip induced by our criterion for selecting the direction of the binormal vectors (Fig. 1), while our criteria for helices are independent of the curvature signs.

D. A gauge model of protein structure

The gauge model of polymers inspired by the functional form of Eq. (6) is relatively general, and supports the description of the three-dimensional curves described by protein chains based on their geometric structure, without any reference to the underlying atomic nature of the molecule.

For the discrete chain formed by the α-carbons in polypeptide structures, one can resort to a discrete functional energy [14–17]:

$$E = \sum_i -2\overline{\kappa}_{i+1}\overline{\kappa}_i + 2\overline{\kappa}_i^2 + U(\overline{\kappa}_i, \overline{\tau}_i), \tag{17}$$

with $\overline{\kappa}_i, \overline{\tau}_i \in [-1, 1]$ and where $U(\overline{\kappa}_i, \overline{\tau}_i)$ is a generalized potential. This model incorporates first-neighbor coupling for $\overline{\kappa}_i$ and only on-site interactions for $\overline{\tau}_i$. This first-neighbor coupling arises from a discretization of the derivative $(\frac{d\kappa}{ds})^2$ in Eq. (6) as a finite difference $(\kappa_{i+1} - \kappa_i)^2/s_i^2$, which accounts for the terms $\sum_i -2\overline{\kappa}_{i+1}\overline{\kappa}_i + 2\overline{\kappa}_i^2$. The overall s_i^{-2} is constant in proteins, with $s_i \approx 3.8$ Å, and is thus absorbed into the definition of E.

For illustrative purposes, we can think of a two-dimensional discrete curve, described only in terms of its curvature, with a Higgs-inspired potential:

$$U^{2D}(\overline{\kappa}_i) = \tfrac{1}{4}\lambda\overline{\kappa}_i^4 - \tfrac{1}{2}\mu^2\overline{\kappa}_i^2, \tag{18}$$

that supports the existence of two different minima (as does the Higgs potential with the existence of a symmetric and a broken phase).

It can be easily shown that this discrete model supports (i) constant curvature solutions $k_i = k_{i+1} = \pm\sqrt{\frac{\mu^2}{\lambda}}$

corresponding for $\mu \neq 0$ to the minima of the Higgs broken phase in the continuum that would represent a circumference (the 2D counterpart of helices), and (ii) alternating sign solutions $k_{i+1} = -k_i = \pm\sqrt{\frac{\mu^2-8}{\lambda}}$ for $\mu^2 > 8$ corresponding to the typical zig-zag structure of β-strands.

For the three-dimensional curves, we may devise a functional form of $U(\overline{\kappa}_i, \overline{\tau}_i)$ that departs from the two-dimensional case and contains more general terms for the dependence on the torsion, $\overline{\tau}_i$. We can quite generally write it as an expansion in powers of $\overline{\kappa}_i$ and $\overline{\tau}_i$ as

$$U(\overline{\kappa}_i, \overline{\tau}_i) = U^{2D}(\overline{\kappa}_i) + \sum_{n,m=0}^{\infty} c_{nm} \overline{\kappa}_i^{2n} \overline{\tau}_i^{m+1}, \tag{19}$$

where only even powers in $\overline{\kappa}_i$ have been included to support the symmetry $\overline{\kappa}_i \to -\overline{\kappa}_i$. The lowest-order terms of this expansion correspond to $\overline{\tau}_i$, $\overline{\tau}_i\overline{\kappa}_i^2$ (breaking the symmetry in $\overline{\tau}_i$ and thus supporting solutions with either left-handed or right-handed chirality), $\overline{\tau}_i^2$ (a mass-like term for $\overline{\tau}_i$), and $\overline{\tau}_i^2\overline{\kappa}_i^2$ (similar to the Higgs mechanism). These terms are presented in Eq. (3) of [20] as well as in [14,15], but indeed they are the natural extension of the two-dimensional energy functional to three dimensions, as discussed above; they are just the first terms of an expansion in $\overline{\kappa}_i$ and $\overline{\tau}_i$. The values of the coefficients c_{nm} of the terms mentioned above, or even higher-order terms, can be tuned to support the existence of fixed points corresponding to the secondary structures of proteins or to other possible characteristics of other three-dimensional curves. In particular, for proteins, one can choose values of the coefficients c_{nm} that support the existence of fixed points of Eq. (17) for the values of $\overline{\kappa}_i$ and $\overline{\tau}_i$ corresponding to the typical secondary structures of proteins such as the ones in Table I.

In a series of papers [14–16], the secondary structures and transitions between them have been studied as stationary points of the energy functional in Eq. (17). For any general form of U, the application of Euler-Lagrange equations to this discrete case leads to

$$\overline{\kappa}_{i+1} + \overline{\kappa}_{i-1} - 2\overline{\kappa}_i = \frac{1}{2}\frac{dU_i}{d\overline{\kappa}_i} = \overline{\kappa}_i \frac{dU_i}{d\overline{\kappa}_i^2}, \tag{20}$$

and

$$\frac{dU_i}{d\overline{\tau}_i} = 0, \tag{21}$$

where, for simplicity, we wrote $U_i \equiv U(\overline{\kappa}_i, \overline{\tau}_i)$.

The authors of [20] used the discrete Euler-Lagrange equations (20) and (21) to write the equivalent of a discrete non-linear Schrödinger Hamiltonian in $\overline{\kappa}$ that is later used to describe transitions among secondary structures as solitons. Our approach here is different, and in some sense complementary to [20]; we will instead use Eq. (20) to investigate the dependency of $U(\overline{\kappa}, \overline{\tau})$ on curvature and torsion and the determination of the secondary structure of amino acid chains.

We will exploit here the set of more than 18 000 proteins used in the previous section to analyze the dependency of the energy functional with $\overline{\kappa}$ and $\overline{\tau}$. To this aim, we have plotted in Fig. 8 a histogram of the values of the quantity,

$$\frac{\overline{\kappa}_{i+1} + \overline{\kappa}_{i-1}}{\overline{\kappa}_i} - 2, \tag{22}$$

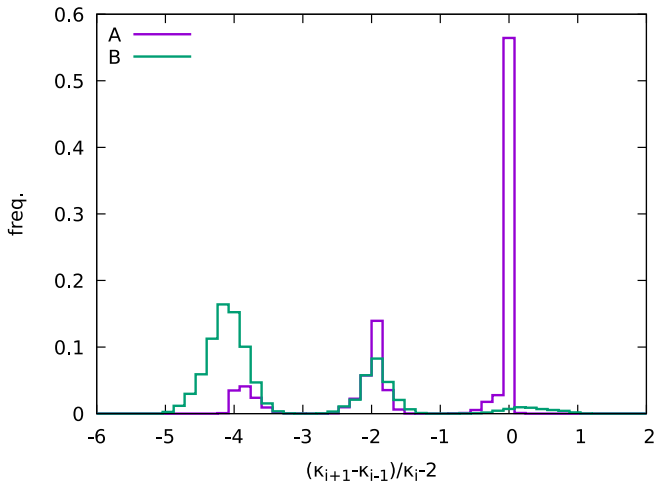

FIG. 8. Histogram of the values of $\frac{dU}{d\overline{\kappa}^2}$ as given by Eq. (22) for residues with $|\overline{\kappa}| > 0.95$ (set A) and for $|\overline{\kappa}| \in (0.65, 0.90)$ (set B).

that, according to Eq. (20), would be equal to $\frac{dU_i}{d\overline{\kappa}_i^2}$. In this analysis, therefore, we do not assume the form of Eq. (17) but rather investigate whether this functional form for the energy is reasonable and the shape of U under the hypothesis that the protein backbone curves result as fixed points of Eq. (17). By selecting the positions of the curve characterized by $|\overline{\kappa}| > 0.95$ (set A), we observe that in most cases $\frac{dU}{d\overline{\kappa}^2} \approx 0$, indicating that these points correspond to a minimum of the energy functional, such as the one in the Higgs broken phase.

In terms of curvature, we can consider three possible scenarios for residues with approximate constant $|\overline{\kappa}|$:

(1) $\overline{\kappa}_{i-1} = \overline{\kappa}_i = \overline{\kappa}_{i+1}$, as in the case of a helix, which would imply $\frac{dU_i}{d\overline{\kappa}_i^2} = 0$;

(2) $\overline{\kappa}_{i-1} = -\overline{\kappa}_i = \overline{\kappa}_{i+1}$, corresponding to an oscillating curvature, for which $\frac{dU_i}{d\overline{\kappa}_i^2} = -4$; or

(3) $\overline{\kappa}_{i-1} = \overline{\kappa}_i = -\overline{\kappa}_{i+1}$, i.e., where a sign change occurs, such as at the end of a helix, leading to $\frac{dU_i}{d\overline{\kappa}_i^2} = -2$.

Set A in Fig. 8 shows the appearance of three peaks, as expected, with the following values of $\frac{dU_i}{d\overline{\kappa}_i^2}$: 0, which is the most frequent value, and indicating that we are at the minimum of the action; -2 which may appear, for example, at the end of a helix; and -4, where there is an oscillating sign that may appear in strongly curved β-strands such as collagen.

For the data with $|\overline{\kappa}| \in (0.65, 0.90)$ (set B), which would correspond (at least in some cases) to β-strands, the situation is quite different: they do not correspond to minima of U, as the number of points with $(k_{i+1} + k_{i-1})/k_i - 2 \approx 0$ is rather small, and most of them correspond to the oscillating case described above, for which we have $(k_{i+1} + k_{i-1})/k_i - 2 \approx -4$.

The meaning of these results has to be understood as follows: if we assume the three-dimensional curves drawn by proteins as stationary points of the energy functional of the general form of Eq. (17), there are stationary points corresponding to the minima of the on-site term, $U(\overline{\kappa}, \overline{\tau})$, as appears to be the case for helices, while other stationary points are dynamical, associated with the kinetic term in Eq. (17), which describes β-strands. It is important to emphasize that

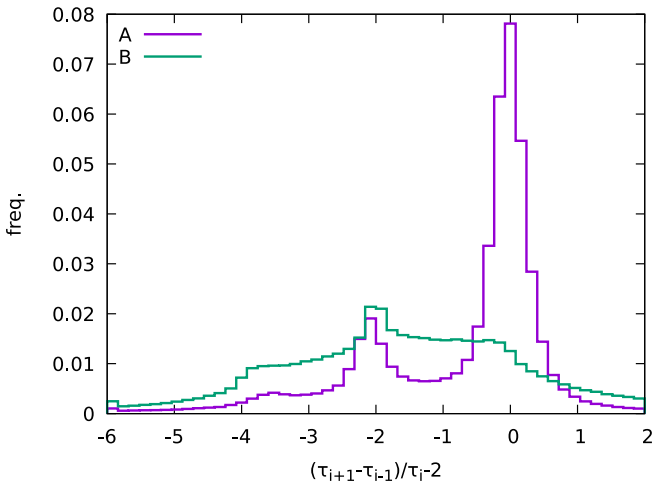

FIG. 9. Histogram of the values of $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}_i-2$ for residues with $|\overline{\kappa}|>0.95$ (set A) and for $|\overline{\kappa}|\in(0.65,0.90)$ (set B).

these findings are independent of the functional form chosen for $U(\overline{\kappa},\overline{\tau})$.

The cases with $\frac{dU_i}{d\overline{\kappa}_i}=-2$ in Fig. 8 would correspond either to a transition between the two minima of the action or to a position where our criterion for the sign correction failed and corresponds to an inner position of a helix or a β-strand with a large bending. Naturally, we have only analyzed the most frequent structures in proteins, and the patterns that we found might also correspond to other less frequent secondary structures such as π-helices, polyproline, Ω-loops, etc. Had we used the traditional description in terms of ρ_i, the general shape of the corresponding plot, $(\rho_{i+1}+\rho_{i-1})/\rho_i-2$, would have shown no major structure, as, for example, α-helices correspond to $\rho\approx 0$.

Concerning the $\overline{\tau}$ dependence, an analogous plot has been included in Fig. 9 for $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}_i-2$. In this case, for set A, corresponding to a large curvature, we observe a maximum around $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}_i-2=0$, which corresponds to $\frac{dU_i}{d\overline{\tau}_i}=0$, according to the interpretation in terms of the action (6) or its discrete version of Eq. (17). Although the peaks are not as clear as in Fig. 8, some structures can still be observed at $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}_i-2\approx-4$ and, more prominently at $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}-2\approx-2$. Based on the histograms obtained for sets A and B (containing protein helices and β-strands), this plot would exclude the presence of a kinetic term in the action for the torsion $\overline{\tau}$, as there seems to be no dynamical fixed point of the action characterized by $(\overline{\tau}_{i+1}+\overline{\tau}_{i-1})/\overline{\tau}_i-2\approx-4$, as in the case of the curvature.

Finally, based on the observed protein backbone structures in nature, a description based on Eq. (17), with a potential term constructed from scratch as the one proposed in Eq. (19), that we can rewrite as

$$U[\overline{\kappa},\overline{\tau}]=\sum_{m,n}c_{m,n}\overline{\kappa}^{2m}\overline{\tau}^{n}, \tag{23}$$

is consistent with the observed polymer structure. The values of the coefficients $c_{m,n}$ are determined by the interactions of the proteins and the structure of the monomers, and knowing them would help describe the three-dimensional curves adopted by proteins.

IV. CONCLUSIONS

The geometric pattern drawn by α-carbons in proteins has been used to compute the curvature and torsion along the curve in a discrete Frenet frame approach. We have proposed an algorithm for computing such values that provides valuable information for identifying and understanding the most frequently found secondary structures in proteins and the transitions among them. The information on curvature and torsion has served as a primary indicator of secondary structures, which has proven useful when compared to more sophisticated methods involving hydrogen bonds and Ramachandran angles (ϕ,ψ). Overall, the simple method proposed can recognize most of the helices and β-strands labeled in the Protein Data Bank files.

We have analyzed the backbone curve drawn by protein α carbons as the result of a $U(1)$ gauge model, whose energy depends solely on the computed dimensionless curvature and torsion. From the analysis of a large collection of proteins, we have justified the presence of a kinetic term for the curvature, while the kinetic term for the torsion may be absent. Overall, this analysis makes it possible to write an energy functional in terms of curvature and torsion, which includes a kinetic term for the curvature and a potential term and Higgs terms, as well as terms that break the symmetry between positive and negative values of the torsion.

The picture of proteins as fixed points of a functional action in terms of their curvature and torsion, whose functional form appears naturally, is an interesting starting point to analyze the properties of proteins. The analysis of protein secondary structure based on this assumption will be the subject of a forthcoming publication.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the Spanish MICINN research Grant No. PID2022-140440NB-C2. All calculations have been performed at the UPO computing center, C3UPO.

DATA AVAILABILITY

The data that support the findings of this article are openly available at [22].

APPENDIX A: DISCRETE FRENET FRAME ANGLES

We are considering the three-dimensional curve formed by the positions of the protein α carbons, which are separated by distances s_i along the direction of the tangent vectors $\vec{T}_i$, which indicate the direction of each segment. The set of distances and tangent vectors then fully characterizes the curve. Alternatively, the directions of the tangent vectors can be expressed via a couple of angles (because they are normalized by construction). In this appendix, we review the possible definitions of these two angles.

Once we have the vectors $\vec{T}_i$, we can define the discrete Frenet frame vectors $\vec{N}_i$ and $\vec{B}_i$ according to Eqs. (11) and (12),

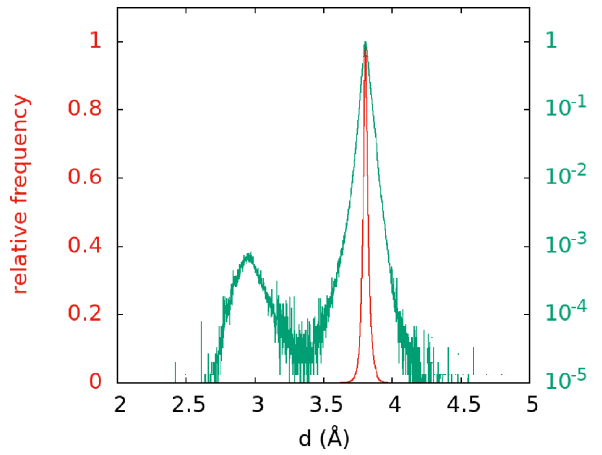

FIG. 10. Histogram of distances between consecutive C_α obtained for over 70 000 protein residue positions extracted from the Protein Data Bank. The histogram has been normalized to 1 at its maximum to better illustrate the relative abundance of bonds for each distance. The red line shows the sharp distribution around $s = 3.80$ Å, while the green one corresponds to a logarithmic scale showing the presence of a second peak for distances of approximately 3 Å that represents less than 0.3% of all amino acid bonds.

and we have an orthogonal direct base defined for each one of the bonds of the polygonal curve. In terms of these vectors, one can compute the turning angle ψ_i of the curve defined in Eq. (15), which measures the angle between two consecutive segments, and the dihedral angle θ_i in Eq. (16), which is the angle between the plane containing the positions $(i-1, i, i+1)$ and the one containing the positions $(i, i+1, i+2)$.

Indeed, any other product among the discrete Frenet frame vectors at consecutive bonds can in principle be used to characterize the discrete three-dimensional rotation of the frame between two adjacent positions. We have used the alternative description in Eqs. (13) and (14) based on the angles between the normal vector and the tangent one for the following site, $\vec{T}_{i+1} \cdot \vec{N}_i$, and the angle between the binormal vector and the normal one at the following site, $\vec{N}_{i+1} \cdot \vec{B}_i$.

The product of Frenet frame vectors at consecutive sites of the chain can be easily obtained in terms of the turning, ψ_i, and dihedral θ_i, angles. For example, for the product of normal and tangent vectors, one obtains

$$\vec{T}_{i+1} \cdot \vec{N}_i = \sin\psi_i \cos\theta_i, \tag{A1}$$

$$\vec{T}_i \cdot \vec{N}_{i+1} = -\sin\psi_i. \tag{A2}$$

The products between normal and binormal vectors are written in terms of the turning and dihedral angles as

$$\vec{N}_{i+1} \cdot \vec{N}_i = \cos\psi_i \sin\theta_i, \tag{A3}$$

$$\vec{N}_i \cdot \vec{B}_{i+1} = -\sin\theta_i, \tag{A4}$$

while, by construction, the products among tangent and binormal vectors vanish.

Each one of the scalar products above defines an angle, and the full characterization of the curve's rotation along each segment can be obtained once one determines the angle based on the sin() and cos() of the angle. It is interesting to note, with respect to this, that the definitions of $\overline{\kappa}_i$ and ρ_i (and the equivalent is true for $\overline{\tau}_i$ and χ_i) satisfy $\overline{\kappa}_i^2 + \rho_i^2 = 1$, and thus given one of them the second is known except for a sign that indicates the relative orientation of the polygonal line segments.

APPENDIX B: PROTEIN'S BOND LENGTH AND CONTINUUM LIMIT

Using the large dataset of proteins discussed in the text, we have computed the distance between consecutive C_α for the ∼70 000 protein chains analyzed. In the analysis, we discarded those chains where there are missing residues or, in general, when the distance between α-carbons is detected to be larger than 5 Å. A histogram of distances has been represented in Fig. 10, normalized to its maximum. The distribution is characterized by a sharp peak centered at $s = 3.80$ Å with a width of 0.02 Å, with over 66% data points lying in the range (3.78, 3.82) Å and over 97% between 3.72 Å and 3.88 Å, results compatible with Fig. 3 of [21] that also shows the distribution of $C_\alpha - C_\alpha$ distances but is limited to the main peak. The log-scale plot (green line in Fig. 10) shows the appearance of a second peak corresponding to distances below 3 Å that is known to be associated to proline's cis configuration, but that overall contains less than 0.3% of the amino acid bonds in the protein backbone.

In our work, we have used the fact that in almost all cases the distances among C'_αs are close to $s = 3.8$ Å to resort to the dimensionless quantities $\overline{\kappa}_i$, $\overline{\tau}_i$, with the typical length-scale of the curve being set by s. Let us analyze the meaning of these parameters in the limit of a continuum curve, $s \to 0$. In this limit, using Eqs. (13) and (14), and the discrete Frenet frame equations (7)–(9), one has $\overline{\kappa}_i \to s_i\kappa_i + O(s^3)$, $\overline{\tau}_i \to s_i\tau_i + O(s^3)$, while for the alternative description in terms of ρ_i and χ_i defined in Eqs. (15) and (16) one has $\rho_i = 1 - \frac{s^2\kappa^2}{2} + O(s^4)$, $\chi_i = 1 - \frac{s^2\tau^2}{2} + O(s^4)$. Although both $\overline{\kappa}_i$ and ρ_i can be used to compute κ, we have chosen the former because in the continuum limit, it provides a measure of the curvature in units of the inverse bond length.

In the limit of a continuum curve, the bond and dihedral angles would be small, while in the polygonal line drawn by proteins, the angles are far from being small. For example, for the case of α-helices, bond angles are close to $\pi/2$, the largest possible values, and thus, the actual curvature of α-helices cannot be accurately estimated from our values of $\overline{\kappa}$. At first order, if we admitted $\overline{\kappa} = s\kappa$, and take the value $|\overline{\kappa}| \approx 1$ from Table I, one would obtain $\kappa \approx 0.26\,\text{Å}^{-1}$, which differs sizably from the actual value of $0.38\,\text{Å}^{-1}$. For the torsion, taking again the theoretical values in Table I, one would obtain $\tau \approx 0.22\,\text{Å}^{-1}$, which again differs from the value $0.14\,\text{Å}^{-1}$ that one can compute from the α-helix radius and pitch. In any case, the notion of curvature and torsion is yet captured by the parameters $\overline{\kappa}$ and $\overline{\tau}$, and the results indicate that the description in terms of them is useful for the study of polygonal curves even when the curve is far from the continuum limit.

[1] T. E. Creighton, *Proteins Structures and Molecular Properties* (W. H. Freeman, New York, 1993).
[2] T. Blundell and L. N. Johnson, Protein crystallography, in *Chemical Crystallography*, edited by J. M. Robertson (Butterworths, London, UK, 1972).
[3] K. A. Dill and J. L. MacCallum, The protein-folding problem, 50 years on, Science **338**, 1042 (2012).
[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, Highly accurate protein structure prediction with AlphaFold, Nature (London) **596**, 583 (2021).
[5] W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers **22**, 2577 (1983).
[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, Nucl. Acids Res. **28**, 235 (2000).
[7] S. M. King and W. C. Johnson, Assigning secondary structure from protein coordinate data, Proteins **35**, 313 (1999).
[8] S. M. Rappaport and Y. Rabin, Differential geometry of polymer models: Worm-like chains, ribbons and Fourier knots, J. Phys. A: Math. Theor. **40**, 4455 (2007).
[9] S. Rappaport, Y. Rabin, and A. Y. Grosberg, Worm-like polymer loops and Fourier knots, J. Phys. A: Math. Gen. **39**, L507 (2006).
[10] Y. Kats, D. A. Kessler, and Y. Rabin, Frenet algorithm for simulations of fluctuating continuous elastic filaments, Phys. Rev. E **65**, 020801(R) (2002).
[11] L. Giomi and L. Mahadevan, Statistical mechanics of developable ribbons, Phys. Rev. Lett. **104**, 238104 (2010).
[12] P. Castro-Villarreal and J. E. Ramírez, Stochastic curvature of enclosed semiflexible polymers, Phys. Rev. E **100**, 012503 (2019).
[13] A. J. Niemi, Phases of bosonic strings and two-dimensional gauge theories, Phys. Rev. D **67**, 106004 (2003).
[14] U. H. Danielsson, M. Lundgren, and A. J. Niemi, A gauge field theory of chirally folded homopolymers with applications to folded proteins, Phys. Rev. E **82**, 021910 (2010).
[15] M. Chernodub, S. Hu, and A. J. Niemi, Topological solitons and folded proteins, Phys. Rev. E **82**, 011916 (2010).
[16] S. Hu, M. Lundgren, and A. J. Niemi, The discrete Frenet frame, inflection point solitons and curve visualization with applications to folded proteins, Phys. Rev. E **83**, 061908 (2011).
[17] A. Krokhotin, A. J. Niemi, and X. Peng, Soliton concepts and protein structure, Phys. Rev. E **85**, 031906 (2012).
[18] A. Molochkov, A. Begun, and A. Niemi, Gauge symmetries and structure of proteins, EPJ Web Conf. **137**, 04004 (2017).
[19] D. Melnikov, A. J. Niemi, and A. Sedrakyan, Topological indices of proteins, Sci. Rep. **9**, 14641 (2019).
[20] N. Molkenthin, S. Hu, and A. J. Niemi, Discrete nonlinear Schrödinger equation and polygonal solitons with applications to collapsed proteins, Phys. Rev. Lett. **106**, 078102 (2011).
[21] K. Hinsen, S. Hu, G. R. Kneller, and A. J. Niemi, A comparison of reduced coordinate sets for describing protein structure, J. Chem. Phys. **139**, 124115 (2013).
[22] M. Prados, M. D. Hernández, and F. de Soto, Dataset de A Frenet frame analysis of protein geometry: Hints for secondary structure assignments, RIO, 2026, https://hdl.handle.net/10433/27289.