

This is a repository copy of *Chiral interactions between tropocollagen molecules determine the collagen microfibril structure*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/226282/>

Version: Other

Preprint:

Zolotarjov, Art'om, Kröger, Roland orcid.org/0000-0002-5070-0297 and Pushkin, Dmitri O. (2025) Chiral interactions between tropocollagen molecules determine the collagen microfibril structure. [Preprint]

<https://doi.org/10.48550/arXiv.2504.21484>

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Chiral interactions between tropocollagen molecules determine the collagen microfibril structure

Art'om Zolotarjov^a, Roland Kröger^{b,2}, and Dmitri O. Pushkin^{a,3}

^aDepartment of Mathematics, University of York, York

^bSchool of PET, University of York, York

ABSTRACT

Collagen is the most abundant structural protein in animals, forming hierarchically organised fibrils that provide mechanical support to tissues. Despite detailed structural studies, the physical principles that govern the formation of the characteristic D-periodic collagen microfibril remain poorly understood. Here, we present a theoretical framework that links the amino acid sequence of tropocollagen to its supramolecular organisation. By combining statistical modeling of residue geometry with sequence-informed interaction potentials, we show that the chiral arrangement of outward-facing residues induces directional intermolecular interactions that drive molecular supercoiling. These interactions favour the formation of right-handed, pentameric microfibrils with a staggered axial periodicity of approximately 67 nm. Our simulations reveal that this structure emerges across a wide range of mammalian collagen sequences as a global energy minimum robust to biochemical noise. These findings provide a mechanistic explanation for collagen's supramolecular chirality and offer design principles for engineering synthetic collagen-mimetic materials.

Keywords: chiral self-assembly, collagen microfibrils, elastic biomaterials, sequence-encoded assembly

INTRODUCTION

Collagen is by far the most abundant protein in the extracellular matrix, connective tissues, skin, and bones (1, 2). It provides the scaffold that enables the organisation of cells into tissues. As a key structural biomaterial, it influences a multitude of multicellular processes, from bone mineralization to invasions of cancer cells (2, 3) and has even been linked to the Cambrian explosion of multicellular life (4, 5). Since collagen is essential for maintaining tissue structure and function, it is a key focus in regenerative medicine, wound healing, orthopedics, dermatology, and cardiovascular health. Recent advances in biochemical engineering have produced the amino acid sequences of thousands of natural collagens (6) and have enabled the design of synthetic collagen mimetic peptides (CMPs)(7). The biocompatibility of CMPs, their tunable properties, and their potential to replicate natural collagen structures make them indispensable for tissue engineering (8).

At the physical level, the versatility of collagen as a structural protein arises from its propensity

Correspondence E-mail: ³mitya.pushkin@york.ac.uk and ²roland.kroger@york.ac.uk

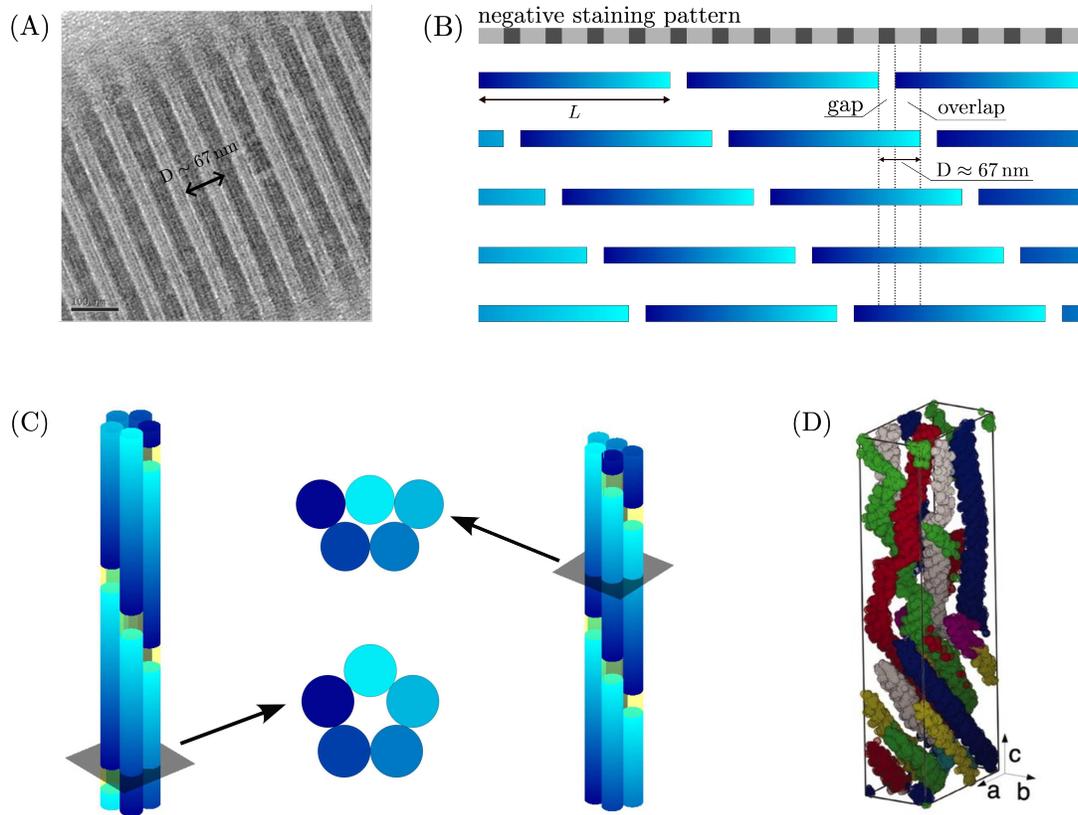


Figure 1. (A). TEM image of a negatively stained collagen fibril showing the D-banding pattern (image obtained from (11)). (B). 2D representation of a collagen microfibril according to the Hodge-Petruska scheme. (C). Schematic representations of the 3D microfibril models. Gap regions are highlighted in yellow. (LEFT) Smith microfibril. (RIGHT) Compressed microfibril. (D). Orgel model of the collagen unit cell (obtained from (10)).

to assemble into fibrils, bundles of fibrils, and intricate hierarchical fibrillar matrices. Despite its importance, the physical principles of collagen self-assembly and their link to the amino acid sequences remain poorly understood. Tropocollagen, the smallest unit in the fibrillar hierarchy, is a semiflexible molecule approximately $L \approx 300\text{ nm}$ long (9) and 1.5 nm in diameter. It is made up of three polypeptide strands (α chains) wrapped around in a right-handed helix, see Fig. 2(A). In each strand, about 1000 amino acids are arranged in a sequence with the regular motif of [Gly-X-Y], where Gly is glycine, and X and Y may be various amino acids but most often proline (Pro) and hydroxyproline (Hyp), respectively. Tropocollagen is classified into nearly thirty distinct types, each varying in amino acid composition and the hierarchical structures they form (1). In this work, we focus on the structures formed by fibrillar collagens, with type I collagen being the most abundant. It has been extensively studied in experimental literature (10).

Tropocollagen readily assembles *in vivo* and *in vitro* in fibrils, with typical diameters between 10 nm and 100 nm (12). Their length is orders of magnitude larger than their diameters. One of their salient features is the periodic axial density modulations, which appear as regular alternating light and dark bands with period $D \approx 67\text{ nm}$ in negatively stained TEM samples, see Fig. 1A. D is often

called the native collagen period. The period value, D , is highly conserved across different collagen types, although notable variations occasionally occur, even within the same tissue (1, 13). The regularity of the banding pattern is crucial for the mechanical strength of fibrils and collagen-rich tissues (14).

A simplified and widely used explanation for the banding pattern, known as the Hodge-Petruska scheme, was proposed in 1964 (see Fig. 1B) (15). It envisages a two-dimensional stack of aligned straight molecules of length L , each shifted by the distance D relative to its neighbour. The banding pattern is explained by the assumption that fibrils are composed of pentameric units. Since $L \approx 4.46D$, five tropocollagen molecules staggered according to the Hodge-Petruska scheme create alternating 'overlap' regions of length $0.46D$ that contain five molecules and 'gap' regions of length $0.54D$ that contain four molecules. The later Smith's microfibril model expanded on this scheme by positioning the five neighbouring molecules at the vertices of a regular pentagon in a plane normal to the fibrillar axis (see Fig. 1C) (16). To further reconcile this model with the quasi-hexagonal lateral packing of individual tropocollagen molecules observed in experiments, the compressed Smith's microfibril model was proposed (see Fig. 1C) (17).

Since 1964, the reality of microfibrils and their role as basic blocks of the fibril, have been experimentally confirmed (10, 18). In a series of papers, Orgel and co-workers have resolved *in situ* the molecular structure of the microfibril using multiple isomorphous replacement and X-ray diffraction experiments. In particular, they showed that five neighbouring molecules are arranged to form a supertwisted right-handed microfibril that interdigitates with neighbouring microfibrils, see Fig. 1D. This interdigitation establishes the crystallographic superlattice, which is formed of quasi-hexagonally packed collagen molecules.

Despite this progress, the theoretical foundations of these packing schemes remain unclear. What physical interactions result in a pentameric microfibril? What determines the axial stagger distance D ? What governs the handedness of the microfibril? And how is this information encoded within the amino acid sequence? While some of these questions have occasionally been addressed in theoretical studies (19–24), to the best of the authors' knowledge, the microfibrillar structure has not been explained as an emergent phenomenon arising from the fundamental molecular interactions.

Orthologous α chain sequences have been extensively documented for all known fibrillar collagens (6). An important model predicting the axial stagger between pairs of tropocollagen molecules comprising the microfibril using residue sequence data was recently suggested by Puzkarska *et al.* (24). In this approach, the D -stagger emerges as the equilibrium microfibril configuration corresponding to a local minimum of the free energy. The interaction potential between pairs of collagen molecules is calculated using the Miyazawa-Jernigan approximation for the contact interaction energy between amino acids, averaged over all possible inter-residue contacts. The latter were determined based on the spatial proximity of residues, which is directly related to the sequence proximity: two residues close in a sequence are necessarily close in space. Hence, one can think of this algorithm as calculating interactions between linear sequences of residues. Consequently, this algorithm ignores the angular dependence of the interaction potential between two collagen molecules. This drawback, in particular, precludes explaining the emergent supertwist and handedness of the microfibril.

In this article, we extend the approach of (24), combining empirical studies with theoretical arguments to quantify the interaction potential between pairs of parallel tropocollagen molecules. Using numerical simulations, we investigate which features of this potential, and under what conditions, give rise to the microfibrillar structure. In particular, we demonstrate that the pairwise interactions

between collagen molecules are chiral due to the chiral spatial arrangement of the outward-facing residues of tropocollagen. This chirality is propagated to the level of the microfibrillar aggregate, resulting in the right-handed supertwist of individual tropocollagen molecules. Furthermore, we attribute the pentameric nature of the microfibril to the geometric arrangement of residues. Our findings reveal that the optimal axial stagger, Δz , can assume different values corresponding to distinct local free energy minima, most notably $\Delta z \approx 0$ and $\Delta z = nD$, where $n = 1, \dots, 4$. The local minimum at $\Delta z \approx 0$ has until now been largely overlooked in theoretical discussions of microfibrillar structures despite experimental evidence for the existence of segment-long spacing (SLS) aggregates both *in vitro* and *in vivo* (25, 26). We show that while the minimum at $\Delta z \approx 0$ is generally stronger than those at $\Delta z = nD$, it is sensitive to noise in the residue-residue interaction energies. Consequently, the axial staggers $\Delta z = nD$ emerge as robust global optima under most conditions.

To compare our predictions with the available experimental data, we analyze amino acid sequences of more than 1000 known fibril-forming collagens of mammalian species. In the absence of detailed studies of the microfibrillar structure for most of the collagens, we take the experimentally observed D-banding pattern in macroscopic aggregates as a proxy for the formation of the D-staggered microfibril. Under this assumption, we predict that all 176 analyzed sequences of heterotrimeric type I collagen result in D-banding, in agreement with the general knowledge in the field (27). This agreement validates our methods and lends credibility to our predictions for other, less well-studied collagen types.

RESULTS

Chiral Interactions and Helical Strip Organisation in Tropocollagen

The observed molecular supertwist of tropocollagen molecules within the microfibril points to the chiral nature of intermolecular interactions. We trace this chirality to the spatial arrangement of the outward-facing residues of the tropocollagen molecule.

High-resolution data on the spatial organisation of collagen residues is currently unavailable, and several distinct structural models attempt to describe it on average (28). To avoid choosing between the models, we use a statistically-derived parametrisation of the triple helix based on the analysis of multiple high-resolution structures of shorter peptides modeling sections of the triple helix (29). This statistical model accounts for differences in the imino acid content, resulting in two distinct triple helix parameterizations: Pro-rich and Pro-poor. These parametrisations can also be viewed as limiting cases, representing the helical parameters of the average collagen structure corresponding to triple-helix segments that are entirely saturated or completely free of Pro and/or Hyp residues (28).

We now demonstrate that each of the parametrisations gives rise to a helical arrangement of the outward-facing residues. Fig. 2A shows the positions of the residues projected on the cylindrical surface of a molecule unwrapped on the (θ, z) -plane, where θ represents the azimuthal angle and z is the axial position for each parametrisation. The position of each residue is indicated by the location of its C_α atom. The outward-facing residues cluster into two families of equally spaced, right-handed helical strips, as shown in Figures 2B and 2C. For the Pro-rich parametrisation, there are seven helical strips, each separated azimuthally by $2\pi/7$ rad $\approx 51.4^\circ$, with a pitch of approximately 200 nm. For the Pro-poor parametrisation, there are ten helical strips, with an azimuthal separation of $2\pi/10$ rad = 36° and a pitch of approximately 75 nm. These emergent helical strips are distinct from the helices formed by the sequential positions of residues. The helical strips have a finite width

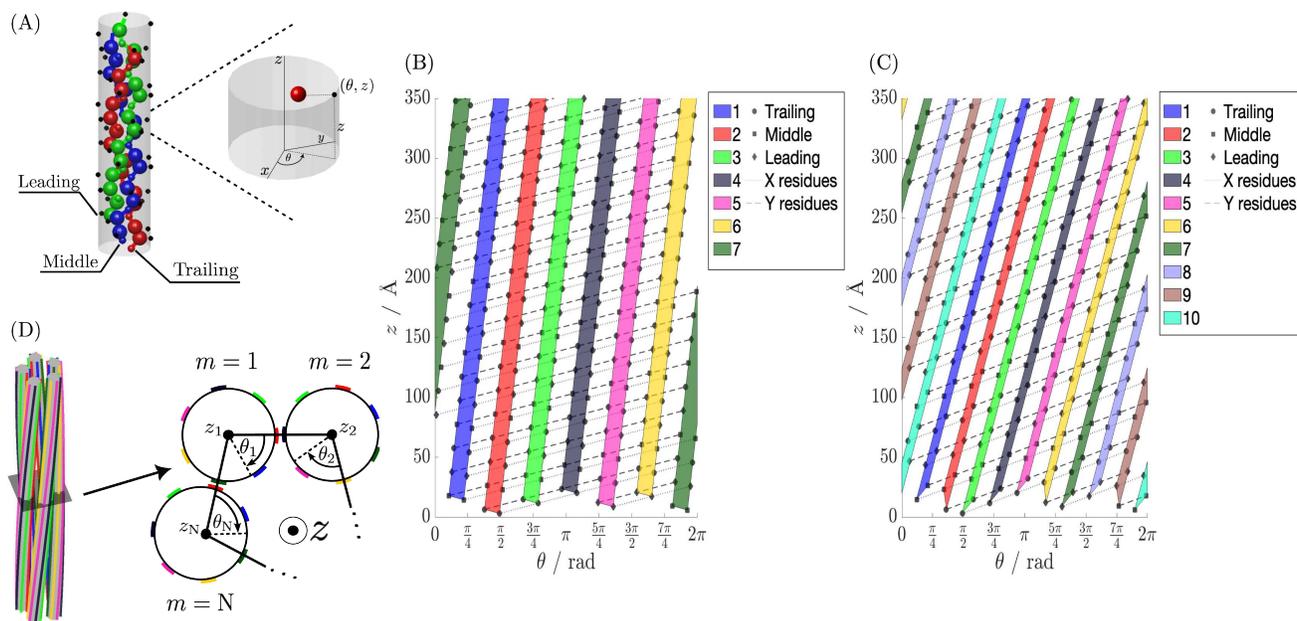


Figure 2. (A). Segment of the collagen triple helix bounded by a cylindrical surface onto which coordinates of each C_α residue atom are projected. The C_α -atom positions are obtained using two statistically derived parametrisations based on analysis of (B) Pro-rich and (C) Pro-poor model peptides. Dotted lines connect the residues that belong to the same α chain. We conventionally denote the most N-terminal α chain as trailing. Solid lines indicate imaginary connections between residues that fall on the same spiral strip. The spiral strips are numbered in order of appearance when moving counter-clockwise around the molecular z -axis and the most N-terminal residue is assigned the azimuthal coordinate $\theta = \frac{\pi}{2}$. The Pro-rich and Pro-poor parametrisations give rise to the two families of right-handed helical strips of amino acids with 7 and 10 helices in each family, respectively. (D). N-membered collagen microfibril model. An axially periodic microfibril is comprised of aligned tropocollagen molecules placed at the vertices of a regular N-gon in the azimuthal plane. The coloured segments on the molecular surfaces correspond to the Pro-rich strips shown in (B).

of approximately 21° for the Pro-rich case and 16° for the Pro-poor case. This width arises from the constant azimuthal coordinate difference between the left and right edges of each strip, which are uniformly composed of X and Y residues, respectively. It is important to note that within a given strip, the spatially nearest X and Y residues do not belong to the same [Gly-X-Y] triplet.

The actual spatial distribution of outward-facing residues on the surface of the tropocollagen molecule varies as a function of the amino acid composition (30). It is, clearly, chiral (non-superimposable with its mirror image). This chirality generates a chiral interaction potential between pairs of parallel tropocollagen molecules, leading to torques that can bend and twist them. While the potential may exhibit complex dependencies on the relative orientation of the molecules, we show that only certain features of this potential are essential for forming the axially periodic, helical microfibril. This enables the development of a simplified model sufficient to predict the microfibril structure. We assume that the outward-facing residues of a tropocollagen molecule can be represented as a superposition of helical strip families with varying pitches (29, 30), such as the

Pro-rich and Pro-poor families described in (29). When two parallel molecules are close enough for their outward-facing residues to interact, strong interactions between the residues in their helical strips generate torques that bend and twist the molecules to minimize interaction energy, aligning the strips along a common axis. This process is analogous to molecular supercoiling observed in coiled coils, which arises from chiral interactions between hydrophobic strips on α -helices (31, 32), though collagen differs in having multiple helical strip families. For the interaction to dominate, the energy gained must outweigh the bending energy cost. This condition is easily met for the 7-strip family but is highly restrictive for the 10-strip family due to their smaller pitch and a strong dependence of the elastic deformation cost on the pitch (see Materials and Methods). Consequently, interactions from the 7-strip family are energetically favoured.

Thus, we model the effective interaction potential between collagen molecules based on the 7 helical strips from the Pro-rich parameterization. This leads to the prediction that collagen molecules in a microfibril form right-handed helices with a helical angle of approximately 5° , see Eq. [3] in Materials and Methods, consistent with experimental observations in bone and tendon (33, 34).

Energetics of Strip-Strip Interactions

Based on this, we hypothesize that interactions between aligned collagen molecules in a microfibril are primarily driven by opposing strip-strip interactions. With 7 strips, there are 28 potential strip-strip interactions, denoted as $E_{i,j}(\Delta z)$, where $-L \leq \Delta z \leq L$ is the axial stagger of strip j relative to strip i . We calculate them using the empirically determined Miyazawa-Jernigan contact potentials (MJCP) for residue-residue interactions, see Materials and Methods.

Assuming that collagen molecules form axially periodic arrays separated by gaps of length g , the interactions between opposing arrays are described by 28 $(L + g)$ -periodic potentials: $E_{i,j}^p(\Delta z) = E_{i,j}(\Delta z) + E_{i,j}(\Delta z - g - L)$, where $0 \leq \Delta z < L + g$. For consistency with the standard definitions, we define D in terms of g such that $5D = L + g$. While this definition anticipates the value of D , it does not constrain it. Deferring the discussion of how D is determined in simulations to the Appendix, we find that the values of D that yield perfectly-staggered microfibrils fall within a narrow range, approximately $D \approx (67 \pm 2)$ nm, see Fig. S2. Therefore, when we next discuss the minima of $E_{i,j}^p(\Delta z)$ over Δz , we will use the corresponding values of D as relevant length scales.

Fig. 3B shows that the global minima of the strip-strip interaction potentials typically belong to two classes: the minima at $\Delta z \approx 0$ and the minima at $\Delta z \approx nD$. Motivated by the experimentally observed D -banded structures, previous studies have focused on local energy minima at positive multiples of D overlooking the possibility of a global minimum at $\Delta z \approx 0$ (19–22, 24). Our results show that, unexpectedly, most, but not all, global minima fall into this class, see Figs. 3A, B. This conclusion is surprising since the dominance of the minima at $\Delta z \approx 0$ would lead to an ‘in-register’ arrangement of the molecules in a microfibril, precluding the formation of the D -staggered microfibril. This raises the question of what conditions warrant the formation of D -staggered microfibrils.

To address this question, we note that the MJCP values used for calculating the interaction energies are subject to significant uncertainties due to experimental errors and high variability in biochemical environments. These uncertainties arise from neglecting the specifics of factors such as electrostatics, solvent effects, molecular crowding, and post-translational modifications (e.g., hydroxylation of Pro/Lys and glycosylation), as well as assuming sequence-independent interactions. Nevertheless, D -banded collagen fibrils do form under diverse conditions, including in *in vitro* environments (which lack biological regulatory factors). This suggests that the emergent structures

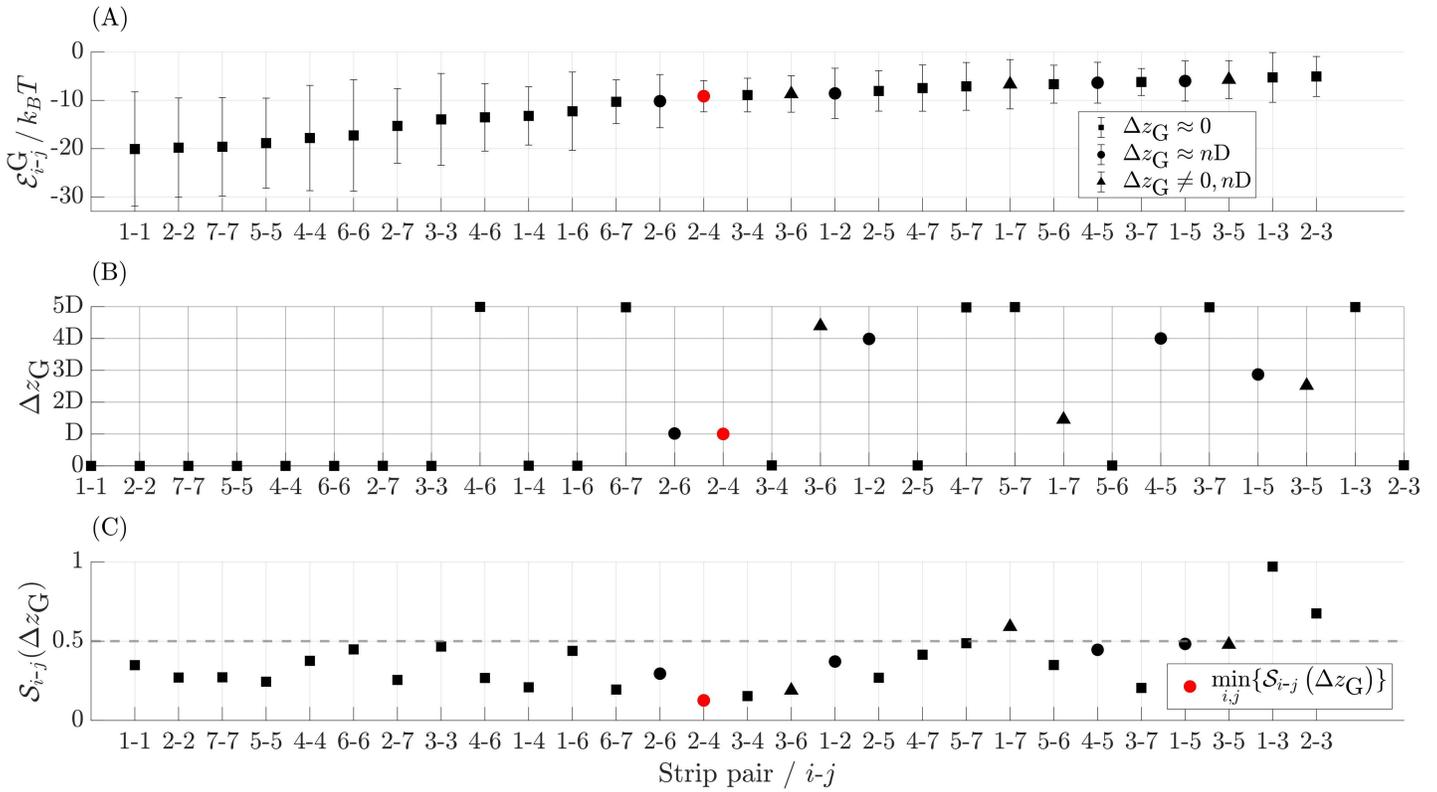


Figure 3. Global minima of the axially periodic strip-strip interaction energies and their sensitivity to noise in contact potential values. The presented results are obtained numerically for $\alpha_2(\mathbf{I})[\alpha_1(\mathbf{I})]_2$ rat collagen. (A). Average values of the interaction energies at their global minima due to random noise in contact potential values. Error bars indicate the standard deviation in the noise-added energy values. (B). Locations of the global energy minima, Δz_G . (C) Noise sensitivity of the global energy minima. The most pronounced minima belong to two classes: the minima at $\Delta z \approx 0$ and the minima at $\Delta z \approx D$. In general, the former are stronger but more sensitive to noise than the latter. See Materials and Methods for details of the procedures.

must be highly robust toward environmental variability.

To account for it, we add random noise to the MJCP values and analyze the sensitivity of the intermolecular interactions and the emergent microfibrillar structures to this noise. We define the noise sensitivity of the pairwise strip-strip interactions as the variance of the noisy potentials $\mathcal{E}_{i-j}^p(\Delta z)$ normalised by their mean, i.e.

$$\mathcal{S}_{i-j}(\Delta z) = \frac{\text{Var} \left\{ \mathcal{E}_{i-j}^p(\Delta z) \right\}}{\left| \mu \left\{ \mathcal{E}_{i-j}^p(\Delta z) \right\} \right|^2}, \quad [1]$$

where $\text{Var}\{\dots\}$ and $\mu\{\dots\}$ denote the variance and the mean, respectively. Fig. 3C shows that, remarkably, the noise sensitivity turns out to be the smallest for the interaction potential minima at $\Delta z \approx D$. In contrast, the energy minima at $\Delta z \approx 0$ are more sensitive to the noise.

We use the linear decomposition of $\mathcal{S}_{i,j}(\Delta z)$ into contributions from interacting pairs of residues (see Materials and Methods) to trace the high noise sensitivity of interactions at $\Delta z \approx 0$ to only two interacting pairs of highly abundant residues: Pro-Pro and Pro-Ala, see Fig. S1. The axial staggers and strip combinations that achieve the strongest inter-molecular interactions and minimise the number of interactions between abundant residues, end up being the least sensitive to the noise.

In the biological context, the noise sensitivity implies that the majority of strip-strip interaction energies at $\Delta z \approx 0$ may be strongly affected by such factors as variations in pH and temperature or the post-translational hydroxylation of Pro residues (1). We hypothesize that this feature may form the basis of a sensitive biochemical control over the emergent structures. It requires a separate investigation in each biochemical context. For the present study, we simply assume that if the noise sensitivity of a minimum turns out to be higher than a chosen threshold \mathcal{S}_c , the minimum can be disregarded from the microfibril energy calculation.

Emergence of D-periodic Microfibrils

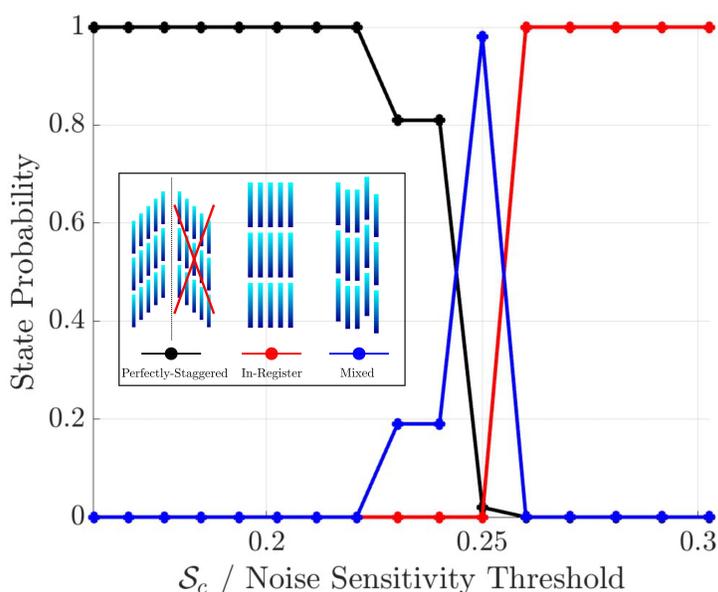


Figure 4. Equilibrium probabilities of different microfibrillar states as a function of noise sensitivity threshold.

To determine whether the D-staggered microfibril emerges from the intermolecular interactions, elucidate the role of the strip-strip interactions, and explain why the microfibril is composed of $N = 5$ molecules, we turn to numerical modeling. To keep the problem tractable, we assume an axially periodic microfibril comprised of aligned tropocollagen molecules placed at the vertices of a regular N -gon in the azimuthal plane. Individual molecules may rotate by the angles θ_m around their centerlines and be shifted along the microfibrillar axis by the distances z_m , see Fig. 2D. We assume that the only interacting molecules are the nearest neighbours that share a polygon edge. Any such pair of molecules is assumed to interact only via a single pair of strips at a time. The microfibril energy E_M is then given by the sum of N pairwise molecular interactions (see Materials and Methods). The equilibrium microfibrillar structure results from minimising the free energy of

the system. Since the microfibril entropy is sub-extensive in microfibril length, it can be ignored in the present context and we can simply minimise E_M (24).

For $N = 5$, we identify three types of emergent microfibrillar configurations: (1) four perfectly-staggered ones, where each molecule is shifted relative to its right neighbour by the same value of $\Delta z \approx nD$ for $n = 1, 2, 3$ or 4 , (2) in-register configurations with $\Delta z \approx 0$ for all molecules, and (3) mixed configurations, see Fig. 4. When the sensitivity to the contact potential noise is disregarded, i.e. for high values of \mathcal{S}_c , simulations predict that the equilibrium microfibrils adopt the ‘in-register’ configuration, consistent with the energetical dominance of the interaction minima at $\Delta z \approx 0$. As the acceptable noise sensitivity threshold is lowered, one of the perfectly-staggered configurations emerges at equilibrium, see Fig. 4 and Fig. S3. Perfectly-staggered microfibrils forming enantiomeric pairs ($\Delta z = D$ and $\Delta z = 4D$, $\Delta z = 2D$ and $\Delta z = 3D$) have differing energies in our model, so only one is selected at equilibrium. Without accounting for the chiral strip organisation of interacting residues, they would be energetically indistinguishable, see Materials and Methods.

Aggregate Size Specificity

Next, we consider why a microfibril is comprised of five molecules. Fig. 5A shows the global minimum of microfibrillar energy per molecule for varying N . Notably, $N = 5$ gives the lowest energy, thus being selected at equilibrium. This fact has a simple geometrical explanation: for strong interactions, the helical strips of the neighbouring molecules must face each another. When molecules are positioned at the vertices of a regular N -gon, the interior polygon angle $\nu(N)$ should approximate $m\alpha$, where $\alpha = 360^\circ/7$ is the angle between the strips and m is an integer. It is easy to see that $\nu(5) = 108^\circ$ closely approximates $2\alpha \approx 103^\circ$, see Table 1. For other values of N , some strips would always face away from their neighbours, reducing energetic gain.

This argument relies on the spatial organisation of the residues in seven spirals, independent

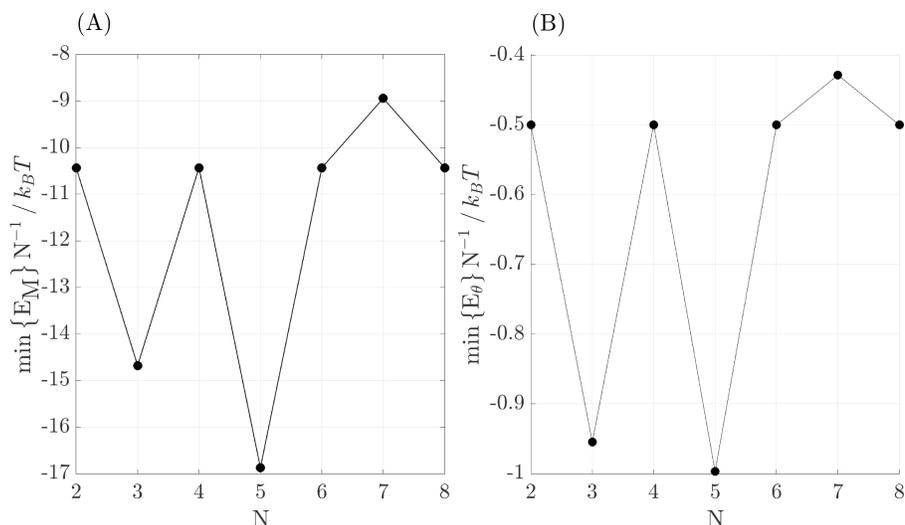


Figure 5. Global minimum of the microfibril energy per molecule as a function of cluster size N in $\alpha_2(I)[\alpha_1(I)]_2$ rat collagen. (A). Microfibril energy with empirically determined axial dependence. (B). Microfibril energy with no axial dependence. Details of the global optimisation procedure can be found in SI.

of their specific sequences. To substantiate this hypothesis, we perform simulations using intermolecular potentials that maintain azimuthal dependence due to the 7 helical strips but ignore axial dependence sensitive to sequence details. The results shown in Fig. 5B indicate that the pentameric microfibril still has an energetic advantage over the trimeric aggregate, but this difference is reduced. Thus, although the spatial organisation of tropocollagen residues alone can make a pentamer the preferred microfibrillar configuration, specific residue interactions are essential for stabilising it. It is also conceivable that some specific residue sequences might preferentially select for a trimeric microfibril.

Table 1. Minimum difference between the internal angle of N-membered microfibrils and the azimuthal inter-strip spacings.

Internal polygon angle $v(N)$	Minimum difference between $v(N)$ and $m\alpha$ / deg
60	8.6
90	12.9
108	5.1
120	17.1
128.6	25.7
135	19.3

DISCUSSION

The exquisite, axially periodic and helically entwined arrangement of collagen molecules in self-assembling fibrils and bundles of fibrils lies at the heart of collagen’s versatility as a structural protein. This ordering emerges at the level of microfibrils – essential, if experimentally elusive, structures (10). In microfibrils, five supercoiled molecules are staggered relative to their neighbours by a fixed distance D , and stacked to form an axially periodic structure. In this work, we investigate how amino acid sequences guide the formation of this structure.

Focusing on collagen I, we found that the outward-facing tropocollagen residues are arranged in sevenfold helical strips. This arrangement emerges from the supercoiling of tropocollagen α -chains and is reminiscent of the hydrophobic strip that emerges in coiled coils due to a regular spatial arrangement of heptad repeats (31, 35). There are, however, important differences: there are seven, rather than just one, interaction strips, the residues forming the strips are, in general, not hydrophobic, and, most importantly, the seven-fold chiral arrangement emerges as a result of energetic selection favouring the spatial arrangement of residues described by the Pro-rich parametrisation of the tropocollagen triple helix.

We predict that strip handedness and chirality are transmitted to the level of molecular collagen conformation through the torques that arise from pairwise intermolecular interactions. The resulting equilibrium helical angle ϕ^* is described by Eq. [3], which was first empirically obtained by Fraser and MacRae in the context of coiled coils (36).

Table 2. Prediction of perfectly-staggered microfibrils in mammalian species for different collagen types.

Type	D-Banded Fibrils	N ^o Species	Predicted FS Microfibrils ^a (%)
$\alpha_2(\text{I})[\alpha_1(\text{I})]_2$	✓ (33)	176	100
$[\alpha_1(\text{I})]_3$	✓ (37)	186	25.27
$[\alpha_2(\text{I})]_3$	unknown ^b	197	94.92
$[\alpha_1(\text{II})]_3$	✓ (38)	191	100
$[\alpha_1(\text{III})]_3$	✓ (39, 40) ^c	185	3.78
$[\alpha_1(\text{V})]_3$	X (41)	167	86.83
$\alpha_1(\text{V})\alpha_2(\text{V})\alpha_1(\text{V})$	✓ (42)	151	78.81
$\alpha_1(\text{V})\alpha_3(\text{V})\alpha_2(\text{V})$	✓ (42)	124	98.39
$\alpha_3(\text{XI})\alpha_2(\text{XI})\alpha_1(\text{XI})$	✓ (43)	148	99.32
$[\alpha_1(\text{XXIV})]_3$	unknown ^d	163	87.12
$[\alpha_1(\text{XXVII})]_3$	X (44) ^d	163	60.36

^a The model is said to predict a perfectly-staggered microfibril, if there exists some value of S_c , below which the perfectly-staggered state probability is unity.

^b An $\alpha_2(\text{I})$ homotrimer is not observed *in-vivo*. This homotrimer has been however observed in *in-vitro* refolding experiments (45). Propensity of $[\alpha_2(\text{I})]_3$ to undergo self-assembly into fibrils has not been investigated to our knowledge.

^c The D-banding length scale of reprecipitated type III collagen fibrils has been reported as (66.7 ± 0.2) nm and (25 ± 10) nm.

^d Developmental collagens are characterised by presence of highly-conserved sequence interruptions. In this work, we do not account for their effect.

The stagger distance, $D \approx 67$ nm, is encoded in local energy minima for the strip-strip interaction potentials, which occur at relative molecular stagger $\Delta z \approx nD$, $n = 1, \dots, 4$ and at $\Delta z \approx 0$. While the minima at $\Delta z \approx 0$ are typically the strongest, they are sensitive to noise in the residue-residue interaction energies. This sensitivity is transmitted to the aggregate level. Upon introducing a noise sensitivity threshold that filters out noise-labile microfibrils, we find that the perfectly-staggered D-periodic microfibrils are the robust global minimisers of the microfibrillar free energy.

In the biophysical context, noise sensitivity points to a sensitive control that may be exerted on aggregates by local biochemical environments. For example, transitions between perfectly-staggered microfibrils and SLS aggregates (corresponding to the dominant minima at $\Delta z \approx 0$) can be induced by varying the interactions between charged residues (46). Residue interactions may also be affected through post-translational enzymatic modification, such as hydroxylation of Pro and Lys residues. Post-translational hydroxylation is known to significantly affect the temperature and ionic conditions required for D-banded fibril formation (47). This observation aligns with our noise sensitivity analysis, which indicates that Pro-containing residue interactions, specifically Pro-Pro and Pro-Ala, have the largest effect on pairwise molecular energy.

It has been understood at least since the work of Hodge and Petruska (15) and Smith (16), that to conform to the regular axial D-banded pattern, collagen aggregates must be composed of pentamers. However, to the best of our knowledge, the physical interactions that could warrant this have not been discussed. We find a strong energetic minimum at $N = 5$ for an axially periodic N-membered

microfibril. We show that for typical intermolecular interactions, the geometric condition that the strips of neighbour molecules face each other alone may select a pentameric microfibril. Yet, specific residue sequences are required for stabilising it and ensuring D-banding.

Our analysis relies on many assumptions and simplifications: we assume that the microfibril is a regular N-gon, that tropocollagen molecules are perfectly aligned, neglect the influence of collagen molecules external to the microfibril, and assume the validity of the contact potentials approach. Furthermore, we disregard the role of post-translational modifications, collagen telopeptide domains, and biological regulation among other factors. To validate our model, we have tested its predictions for all known mammalian sequences of fibril-forming collagens documented in the NCBI RefSeq database, see Table 2.

High resolution *in situ* studies of microfibrillar structures have only been performed for $\alpha_2(\text{I})[\alpha_1(\text{I})]_2$ collagen originating from rat tail tendon (10). Due to the limited availability of such detailed structural data for other collagen types, we use the available measurements of D-banding in different collagen types as a proxy for the emergence of the D-staggered microfibrils. Our analysis predicts that for all sequences of $\alpha_2(\text{I})[\alpha_1(\text{I})]_2$ collagen examined, a perfectly-staggered microfibril is the most stable aggregate below some noise sensitivity threshold \mathcal{S}_c . This finding agrees with experimental observations of D-banding in collagens of this type. Similarly, our model indicates that over 99% of the tested sequences for collagens $[\alpha_1(\text{II})]_3$, $\alpha_1(\text{V})\alpha_3(\text{V})\alpha_2(\text{V})$ and $\alpha_3(\text{XI})\alpha_2(\text{XI})\alpha_1(\text{XI})$ favour the formation of perfectly-staggered microfibrils under the same conditions.

However, our model's results for some homotrimeric collagens $[\alpha_1(\text{I})]_3$, $[\alpha_1(\text{III})]_3$, $[\alpha_1(\text{V})]_3$ and $[\alpha_1(\text{XXVII})]_3$ seem to be at odds with the available experimental data. These discrepancies point to the limitations of the model assumptions used in this study and require future studies. For now, we note the possible sources of these discrepancies. First, in our model, pairwise molecular interactions between homotrimeric collagens of types I, III, and V are particularly strong, consistent with previous studies (24). This presents the tantalizing possibility that interactions of Pro-poor rather than Pro-rich strips might be selected despite the associated higher cost of elastic deformation. In addition to the selection of a different chiral symmetry, it is plausible that the molecular organisation or indeed the number of molecules comprising the microfibril may vary across different collagen types. In particular, specific residue sequences could in principle favour alternative microfibril configurations, such as a three-membered microfibril. In such cases, D-banding may first appear at the level of supramicrofibrillar structures, like pentameric aggregates composed of trimeric microfibrils.

In our analysis of developmental collagens XXIV and XXVII, we did not account for sequence interruptions within their triple-helical domains. Studies using model peptides have demonstrated that deviations from the typical [Gly-X-Y] motif can cause localised unwinding or overwinding of the triple helix at the interruption site (28). These structural perturbations can significantly alter the amino acid composition of the helical strips following the interruption, thereby influencing the interaction potentials between these strips. Consequently, sequence interruptions can markedly affect the stability and assembly of collagen microfibrils (48). To our knowledge, the details of the structural impact of sequence interruptions present in collagens XXIV and XXVII remain to be elucidated (28, 49, 50). This calls for future research into the effect of sequence interruptions in developmental collagens on spatial residue organisation and microfibril self-assembly.

Finally, it should be noted that explicit measurements of D-banding have only been performed in a handful of commonly studied mammalian species. This raises the question of whether in certain

species fibrillar collagens may aggregate into structures that lack D-banding.

Understanding the differences in the self-assembly of heterotrimeric and homotrimeric collagens is crucial for uncovering the fundamental principles that underlie certain medical conditions. Deleterious mutations in the COL1A2 gene are known to lead to the production of homotrimeric type I collagen instead of the normal heterotrimeric form. Clinically, this mutation manifests itself as the Ehlers-Danlos syndrome, which is characterised by altered mechanical properties of collagen-containing tissues, leading to joint hypermobility and cardiac valve abnormalities (51).

Our findings indicate that microfibrillar structural features are not uniform across all fibril-forming collagens. This is in good agreement with established knowledge in the field. For instance, corneal collagen fibrils exhibit a helical angle of approximately 15° , significantly larger than that observed in bone or tendon tissues (34). Additionally, the characteristic periodic banding pattern can manifest at the fibrillar level with periodicities less than the typical D-spacing in type I and III collagens (39, 52). Recent advances in the synthesis of collagen-mimetic peptides also suggest the possibility of microfibril aggregate sizes differing from $N = 5$ (53). We hypothesize that the nonuniformity of structural features among supramolecular collagen aggregates is a crucial characteristic that ensures collagen's structural versatility across diverse biological environments. Studying these diverse self-assembly scenarios offers valuable opportunities for applying our theoretical methods to understand collagen structures.

CONCLUSIONS

This study identifies chiral intermolecular interactions, rooted in the spatial arrangement of outward-facing residues, as a fundamental mechanism driving the self-assembly of collagen into its characteristic D-periodic, supercoiled, pentameric microfibrils. By integrating residue-level sequence data with a physically motivated interaction model, we demonstrate that molecular chirality and microfibrillar architecture are intrinsically linked. The predicted right-handed supercoiling and staggered configuration are not only energetically favoured but also robust to biochemical noise across diverse mammalian collagen sequences. These insights bridge molecular sequence with mesoscale structure, offering a quantitative framework to understand fibrillar collagen assembly. Beyond elucidating a long-standing biophysical question, our approach provides guiding principles for the rational design of collagen mimetic materials, with potential applications in tissue engineering and synthetic extracellular matrices.

MATERIALS AND METHODS

Mechanism of Molecular Supercoiling and Energy-Driven Strip Selection

We model the semi-flexible collagen molecule as an inextensible circular elastic rod, with residues organised on its surface in a family of helical strips with a pitch h . For a pair of molecules interacting via these helical strips, each molecule bends and twists into a (super)helical shape with the radius R and a helical angle ϕ of the filament centreline. This configuration aligns the residue strips of the interacting molecules to face toward other, incurring the elastic energy

$$E_{\text{el}} = \int_0^L \left(\frac{B \sin^4 \phi}{2 R^2} + \frac{C}{2} \left(\frac{\sin 2\phi}{2R} - \chi \frac{2\pi}{h} \right)^2 \right) ds. \quad [2]$$

Here s is the arclength, B and C are the effective bending and twisting rigidities, respectively, $\chi = 1$ for a right-handed strip while $\chi = -1$ for a left-handed strip (31). Taking R to be fixed, for a sufficiently small ratio $\varepsilon = 2\pi R/h \ll 1$ and a finite ratio B/C (see Appendix for the derivation), leads to the equilibrium helical angle ϕ^* given by the asymptotic expression

$$\phi^* \approx \chi \varepsilon = \chi \frac{2\pi R}{h}. \quad [3]$$

This is the classical Fraser and MacRae formula widely used to analyse the triple-helical structure of tropocollagen (36). Neukirch *et al.* later extended it to include coiled-coil proteins under non-zero external forces and torques (31), albeit deriving it in a more restrictive limit where bending rigidity is much smaller than twisting rigidity $B/C \ll 1$, as opposed to a finite ratio of the two.

Furthermore, we show that if, additionally,

$$B/C \ll \varepsilon^{-2}, \quad [4]$$

the elastic equilibrium energy becomes dominated by the bending deformation component and is given by

$$E_{\text{el}}^* \approx 8\pi^4 \frac{BR^2L}{h^4} = 8\pi^4 \frac{\xi_b R^2L}{h^4} k_B T, \quad [5]$$

where ξ_b is the bending persistence length. Condition [4] is easily satisfied in practice. The expression [5] shows that the energetic cost of the elastic deformation increases steeply as the helical pitch h decreases.

We estimate the molecular length as $L \sim 300$ nm, the microfibril radius as $R \sim 3$ nm, and the persistence length as $\xi_b \sim 120$ nm at neutral pH and physiological salt concentration (9). For the Pro-rich strips with $h \sim 200$ nm, the corresponding elastic energy cost is $E_{\text{el}}^* \sim 0.16 k_B T$. This value is significantly smaller than the characteristic interaction energy of two D-staggered collagen molecules, approximately $10 k_B T$. In contrast, for the Pro-poor strips with $h \sim 75$ nm, the elastic energy cost is much higher at $E_{\text{el}}^* \sim 8 k_B T$, approximately 50 times higher than that of the Pro-rich strips.

Thus, the interactions between outward-lying residues that cluster along spirals with the larger pitch h are energetically strongly favoured. Therefore, it is sufficient to account only for the interactions between the seven right-handed helical strips originating from the Pro-rich tropocollagen parametrisation when modelling the effective interaction potential collagen molecules. The equilibrium coiling angle for the corresponding helical pitch h is estimated as $\phi^* \sim 5^\circ$, which aligns well with the experimental observation in bone and tendon (33, 34).

Axial Dependence of Pairwise Molecular Interactions

In calculating pairwise molecular interactions, we will disregard the interactions involving the N/C-telopeptides, which whilst kinetically important, are not necessary for collagen self-assembly into D-banded fibrils (54). Denote a pair of interacting strips as i - j , wherein $i, j = 1, 2, \dots, 7$. Let $\{\mathbf{e}_\rho, \mathbf{e}_\theta, \mathbf{e}_z\}$ be the set of cylindrical basis vectors in the triple helix coordinate system. Let ${}^q \mathbf{x}_j$ to be the position vector of the residues along strip j that are labeled in ascending order of axial coordinate by $q \in \mathbb{Z}^+$. The staggered coordinates of ${}^q \mathbf{x}_j$ are defined as

$${}^q \mathbf{x}'_j(\Delta z) = {}^q \mathbf{x}_j + 2\pi h^{-1} (\Delta z + c_j - c_i) \mathbf{e}_\theta + \Delta z \mathbf{e}_z, \quad [6]$$

where Δz is the axial stagger of strip j relative to strip i and c_l are the constants that define the centerline equations of the strips $z = \frac{h\theta}{2\pi} + c_l$, for $l \in \{1, 2, \dots, 7\}$. The pairwise interaction energy for a staggered strip pair i - j is then

$$E_{i-j}(\Delta z) = \sum_{p,q} \varepsilon_{g(p)g(q)} \left[\Theta({}^{pq}r_{ij}) - \Theta({}^{pq}r_{ij} - l_c) \right], \quad [7]$$

where ${}^{pq}r_{ij} = |{}^p\mathbf{x}_i - {}^q\mathbf{x}'_j|$, Θ is the Heaviside step function, l_c is the interaction length scale and $g: \mathbb{Z}^+ \rightarrow \{1, 2, \dots, 20\}$ maps the sequential residue position along a strip onto its integer designation.

The matrix $\varepsilon \in \mathbb{R}^{20 \times 20}$ represents the energies of the residue-residue interactions. We follow the method of Puzkarska *et al.* and take the values of ε to be the empirically determined Miyazawa-Jernigan contact potentials, namely the entries MIYS850103, MIYS960102, MIYS990107 in the AAIndex database (55). We take $l_c = 0.75$ nm, which is typically assumed to be the representative length scale at which a pair of residues is in contact (24).

Interactions between axially periodic arrays of parallel collagen molecules separated by gaps of length g are described by the 28 T -periodic potentials, where $T = L + g$:

$$E_{i-j}^p(\Delta z) = E_{i-j}(\Delta z) + E_{j-i}(T - \Delta z), \quad [8]$$

where $0 \leq \Delta z < T$. When $i = j$, the sequences of the opposing strips are identical and

$$E_{i-i}^p(\Delta z) = E_{i-i}^p(T - \Delta z), \quad [9]$$

i.e. the functions E_{i-i}^p possess a reflection symmetry with respect to $\Delta z = T/2$. Since previous studies (24) did not differentiate between residue strips, their interaction potentials inherently exhibit this property. In particular, this means that such physical interactions do not distinguish between the enantiomeric pairs corresponding to Δz and $T - \Delta z$. This property might lead to a degenerate ground state, precluding formation of a well-defined axially periodic (D-banding) structure. In particular, perfectly-staggered right-handed and left-handed microfibrils corresponding to the symmetric values of Δz could not be differentiated. This symmetry is broken for interactions of different strips,

$$E_{i-j}^p(\Delta z) \neq E_{i-j}^p(T - \Delta z), \quad i \neq j, \quad [10]$$

lifting the degeneracy.

Noise Sensitivity of Pairwise Interactions

To account for uncertainty in the elements of ε , consider a noise-added residue interaction matrix with elements $\varepsilon_{lm}^* = \varepsilon_{lm} + u_{lm}$. We choose $u_{lm} \sim U(a, b)$, where $U(a, b)$ is the continuous uniform distribution on the interval (a, b) . The noise-added pairwise interaction energy \mathcal{E}_{i-j}^p is then calculated according to Eq. [7] and Eq. [8] using the matrix ε^* .

The noise sensitivity parameter can be analytically evaluated for \mathcal{E}_{i-j}^p using the following expression

$$S_{i-j}(\Delta z) = \frac{12^{-1} |a - b|^2 \sum_{1 \leq l \leq m \leq 20} N_{lm}^2}{\left| E_{i-j}(\Delta z) + \frac{1}{2} (a + b) \sum_{1 \leq l \leq m \leq 20} N_{lm} \right|^2}, \quad [11]$$

where N_{lm} is the number of interacting residues with integer designations l and m at a given Δz . Importantly, Eq. [11] is a linear combination of the contributions due to pairs of interacting residues proportional to N_{lm}^2 .

In addition to the analytical expression in Eq. [11], the noise sensitivity parameter can be computed numerically. The results presented in Fig. 3 were performed numerically by constructing $K = 50$ noise-added interaction energy curves, with noise sampled from $U(-0.1k_B T, 0.1k_B T)$. The value of the noise amplitude $|a|$ ($= |b|$) is unknown, as such for convenience we chose it to be $\approx 10\%$ of the maximum value of the matrix elements in $|\epsilon|$. Importantly, the relative noise sensitivity of the energies and, hence, all of our physical conclusions are independent of the chosen value.

Model of a Microfibril

We parametrise the azimuthal component of pairwise molecular energy by

$$\Phi(\theta_m) = \left[1 + \exp \left(a \left| \sin \left(\frac{\pi(\theta_m - \theta_0)}{\theta_d} \right) \right| - b \right) \right]^{-1}, \quad [12]$$

where the parameters θ_0, θ_d, a, b are chosen to produce 7 equally-spaced maxima for $\theta_m \in [0, 2\pi)$ with the same width as the Pro-rich strips (further details can be found in SI). The pairwise molecular energy can be written as

$$P_m = \Phi(\theta_m) \Phi(\theta_{m+1} - \nu) E_{n(\theta_m) - n(\theta_{m+1} - \nu)}^p(\Delta z_m), \quad [13]$$

where $n(\theta_m) = \text{nint} \left[(\theta_m - \theta_0) \theta_d^{-1} \right] \bmod 7 + 1$, $\text{nint}[\dots]$ rounds its argument to the nearest integer and ν is the internal angle of the N-gon. The energy of the whole microfibril is then simply

$$E_M = \sum_{m=1}^{N-1} P_m + \Phi(\theta_N) \Phi(\theta_1 - \nu) E_{n(\theta_N) - n(\theta_1 - \nu)}^p(\Delta z_N). \quad [14]$$

Cyclical connectivity of the N-gon constrains $\Delta z_N \equiv z_1 - z_N = -\sum_{m=1}^{N-1} \Delta z_m$, where $\Delta z_m = z_{m+1} - z_m$ is the relative axial translation between two molecules.

AUTHOR CONTRIBUTIONS

A.Z., R.K. and D.O.P. conceptualised the project. A.Z. and D.O.P. developed theoretical models. A.Z. carried out empirical and numerical analysis. A.Z., R.K. and D.O.P. wrote the paper.

APPENDIX

Detailed Parametrisation of the Azimuthal Energy Component

Parameters a and b that are used in the definition of the azimuthal energy component Φ are parametrised as follows:

$$b = \frac{\log(q-1)f(\theta_f) - \log(p-1)f(\theta_w)}{f(\theta_w) - f(\theta_f)}, \quad a = \frac{\log(q-1) + b}{f(\theta_w)}, \quad f(t) = \sin \left(\frac{\pi t}{2\theta_d} \right). \quad [15]$$

Parameters $(\theta_w, \theta_f, p, q)$ are defined via

$$\Phi(\theta_{\max} \pm \theta_f) = p^{-1}, \quad \Phi(\theta_{\max} \pm \theta_w) = q^{-1}, \quad [16]$$

where θ_{\max} maximises $\Phi(\theta_m)$ for $\theta_m \in [0, 2\pi)$. In all calculations we set

$$(\theta_0, \theta_d, \theta_w, \theta_f, p, q) = \left(0.5, \frac{2\pi}{7}, \frac{\pi}{9}, 0.06, 1.0004, 100\right). \quad [17]$$

Global Optimisation of Microfibrillar Energy & Calculation of Equilibrium Statistics

In this section we outline the algorithm for global optimisation of the microfibril energy and subsequent calculation of equilibrium microfibril statistics.

Selection of the D-banding Lengthscale

The first step in calculating the possible values of the microfibril energy is deciding on the value of the D-banding lengthscale. This then allows for construction of axially-periodic pairwise potentials E_{i-j}^p , which determine the value of the microfibril energy - see equation [15] of the main text. A priori we do not know the exact value of the parameter D. We start by constraining $D \in [620, 700]\text{\AA}$, based on the experimental measurements of D-banding (1, 14). Next, for each amino acid sequence, we construct a set of candidate values for the D-banding lengthscale, based on the axial stagger of the interaction energy minima of non-periodic pairwise potentials E_{i-j} . The set of candidate values for D is defined as

$$S_D = \left\{ \Delta\tilde{z} \in [620, 720]\text{\AA} \mid \Delta\tilde{z} = \arg \min_{\Delta z} \{E_{i-j}(\Delta z)\}, \mathcal{S}_{i-j}(\Delta\tilde{z}) < \mathcal{S}_{\text{thr}} \right\}, \quad [18]$$

where $\mathcal{S}_{\text{thr}} = 0.49$ is the threshold value of noise sensitivity, below which the minimum is considered a candidate value. \mathcal{S}_{thr} serves as means of roughly filtering out candidate values of D that are unlikely to give rise to interactions with low noise sensitivity. For practical calculations, we restrict the number of elements in S_D by further requiring that $\Delta\tilde{z}$ only correspond to global, secondary or tertiary minima of the pairwise potentials E_{i-j} .

We can now construct a numerical grid of candidate D values to be used for further calculations. The grid points are sampled from

$$I_D = \bigcup_{\Delta\tilde{z} \in S_D} [\Delta\tilde{z} - \delta z, \Delta\tilde{z} + \delta z], \quad [19]$$

where we pick $\delta z = 3\text{\AA}$. We sample candidate values of D by first discretising each closed interval comprising I_D into a uniformly-spaced grid with spacing of 0.5\AA . If we have an overlap between intervals, we pick the grid points for the discretisation that are associated with the least noise sensitive $\Delta\tilde{z}$. We now construct axially-periodic pairwise potentials E_{i-j}^p using a candidate D value that is generated from $\Delta\tilde{z}$ with the lowest noise sensitivity.

Construction of Near-Equilibrium States

The next step is constructing an approximation to the spectrum of the microfibril. Studying the predictions of our model at thermal equilibrium necessitates performing global optimisation of the microfibril energy E_M . An N-membered collagen microfibril has $2N - 1$ degrees of freedom in our model. To aid us in finding the global minimum of E_M , we construct near-equilibrium states (NEqS) which will give the largest energy contributions to the spectrum. NEqS are members of the set $S_{\text{eq}} = \{(\theta^{\text{eq}}, \Delta z^{\text{eq}})\}$, in which the pair of state vectors $(\theta^{\text{eq}}, \Delta z^{\text{eq}})$ specifies the microscopic state

Table 3. Definitions of microfibril categories, based on the pattern of axial staggers.

Microfibril Category	Category Index	^a Condition on Axial Staggers
Perfectly-staggered	A	$\Delta z_m^{\text{eq}} = nD$, for all $m = 1, \dots, N-1$
In-register	B	$\Delta z_m^{\text{eq}} \in [-\Delta_0, \Delta_0]$ for all $m = 1, \dots, N-1$
Mixed	C	Any other Δz_m^{eq} that are not perfectly-staggered or in-register

^a We set the parameter $\Delta_0 = 5 \text{ nm}$ and $n = 1, 2, 3$ or 4

of a microfibril. The components of azimuthal state vector maximise the strip overlap in a given N-gon:

$$\theta_l^{\text{eq}} \in \arg \max_{\theta \in [0, 2\pi)} \{\Phi(\theta)\Phi(\theta - \nu)\}, \quad l = 1, \dots, N. \quad [20]$$

The axial state vector contains $N-1$ components which correspond to the staggers that minimise the axial energy component for a given pair of strips in a microfibril. The total number of NEqS is $7^N M^{N-1}$, where M is the number of minimisers for each interaction curve E_{i-j}^P . To keep the problem tractable, we choose $M = 3$.

In a given microfibril the axial energy components of P_m in general will not be the same. To account for this, we relax the azimuthal degrees of freedom using the sequential quadratic programming algorithm over the domain $[\theta_1^{\text{eq}} - \delta\theta, \theta_1^{\text{eq}} + \delta\theta] \times \dots \times [\theta_N^{\text{eq}} - \delta\theta, \theta_N^{\text{eq}} + \delta\theta]$ with $\delta\theta = 0.15$.

Calculation of Equilibrium Probabilities with Noise Sensitivity

Finally, we calculate the equilibrium statistics of the collagen microfibril. We group the microscopic microfibril states into 3 categories based on the components of Δz^{eq} . The definitions of the microfibril categories are shown in Table 3 and Fig. 4 of the main text. Let s_k denote the k^{th} microfibril in one of these three categories, which we will denote by $s \in \{A, B, C\}$. The equilibrium probability in the canonical ensemble formalism for perfectly-staggered microfibrils (category A) is then

$$\mathcal{P}_A = \frac{\sum_k \exp(-\beta^{A_k} E_M^{\text{eq}})}{\sum_s \sum_k \exp(-\beta^{s_k} E_M^{\text{eq}})}, \quad [21]$$

where $^{A_k} E_M^{\text{eq}}$ is the microfibrillar energy of the k^{th} perfectly-staggered microfibril and $\beta^{-1} = k_B T$. Analogous formulae define the equilibrium probabilities of mixed and in-register states.

For a given noise sensitivity threshold \mathcal{S}_c , we include a NEqS in calculation of \mathcal{P}_s if for a given θ^{eq} , the components of the stagger vector satisfy

$$\mathcal{S}_{i-j}(\Delta z_l^{\text{eq}}) < \mathcal{S}_c \quad \text{for all } l = 1, \dots, N, \quad [22]$$

where $\Delta z_N = \left(- \sum_{m=1}^{N-1} \Delta z_m^{\text{eq}} \right) \text{ mod } 5D$. We note that Eq. [22] must hold for all strip pairs $i-j$ which interact in a microfibril specified by the azimuthal state vector θ^{eq} .

If we find that there exists a value of \mathcal{S}_c , such that $\mathcal{P}_A \rightarrow 1$, we say that our model predicts perfectly-staggered microfibrils at thermal equilibrium. If such a value of \mathcal{S}_c does not exist, we

repeat our calculations with a different candidate value for the D-banding lengthscale. If all such candidate values are exhausted, we conclude that perfectly-staggered microfibrils are not expected at equilibrium within our modelling framework.

Derivation of the Asymptotic Expression for Equilibrium Helical Angle ϕ^*

Let us assume that the supercoiling radius R has a fixed value and the helical angle ϕ is independent of the arcelength s in Eq. [2] of the main text. Then, the elastic energy of deformation is minimised for an equilibrium helical angle $\phi = \phi^*$ which satisfies

$$2\gamma \sin^3 \phi^* \cos \phi^* + (\sin \phi^* \cos \phi^* - \chi \varepsilon) \cos 2\phi^* = 0, \quad [23]$$

where we have defined $\gamma = B/C$ and $\varepsilon = 2\pi R/h$.

Our goal is to construct an asymptotic series for the equilibrium helical angle ϕ^* as a function of $\varepsilon \ll 1$ and finite γ . To that end, we note that we can write ε as a function of ϕ^* in Eq. [23], obtaining

$$\bar{\varepsilon} = (1 - \gamma) \sin \bar{\phi} + \gamma \tan \bar{\phi} \equiv f(\bar{\phi}), \quad [24]$$

where for convenience we have defined $\bar{\phi} = 2\phi^*$ and $\bar{\varepsilon} = 2\chi\varepsilon$. The desired asymptotic expression for ϕ^* is therefore equivalent to finding the series expansion of the inverse function $g \equiv f^{-1}$. Noting that f is analytic at $\bar{\phi} = 0$ and that $f'(0) = 1$, we can apply the Lagrange inversion formula (56) to obtain

$$\bar{\phi} \equiv g(\bar{\varepsilon}) = \sum_{n=1}^{\infty} g_n \bar{\varepsilon}^n, \quad [25]$$

where the expansion coefficients are given by

$$g_n = \frac{1}{n!} \lim_{\bar{\phi} \rightarrow 0} \frac{d^{n-1}}{d\bar{\phi}^{n-1}} \left(\frac{1}{h(\bar{\phi})} \right)^n, \quad \text{where } h(\bar{\phi}) = \frac{f(\bar{\phi})}{\bar{\phi}}. \quad [26]$$

We note that for $\bar{\phi} \ll 1$ we can expand $h(\bar{\phi}) = 1 + \frac{3\gamma-1}{6}\bar{\phi}^2 + O(\bar{\phi}^4)$. Using Eq. [26], the equilibrium helical angle is then asymptotically found to be

$$\phi^* = \chi\varepsilon + \frac{2(1-3\gamma)}{3}(\chi\varepsilon)^3 + O(\varepsilon^5). \quad [27]$$

With the aid of the asymptotic expression in Eq. [27], we can estimate the equilibrium bend and twist energy contributions per unit length as

$$E_{\text{bend}} = \frac{B \sin^4 \phi^*}{2 R^2} \sim \frac{B \varepsilon^4}{2 R^2}, \quad E_{\text{twist}} = \frac{C}{2 R^2} \left(\frac{\sin 2\phi}{2} - \varepsilon \right)^2 \sim \frac{C \gamma^2 \varepsilon^6}{2 R^2}. \quad [28]$$

We therefore conclude that in the limit $\varepsilon^{-2} \gg \gamma$, the bend contribution to the total equilibrium elastic deformation energy is dominant over the twist contribution.

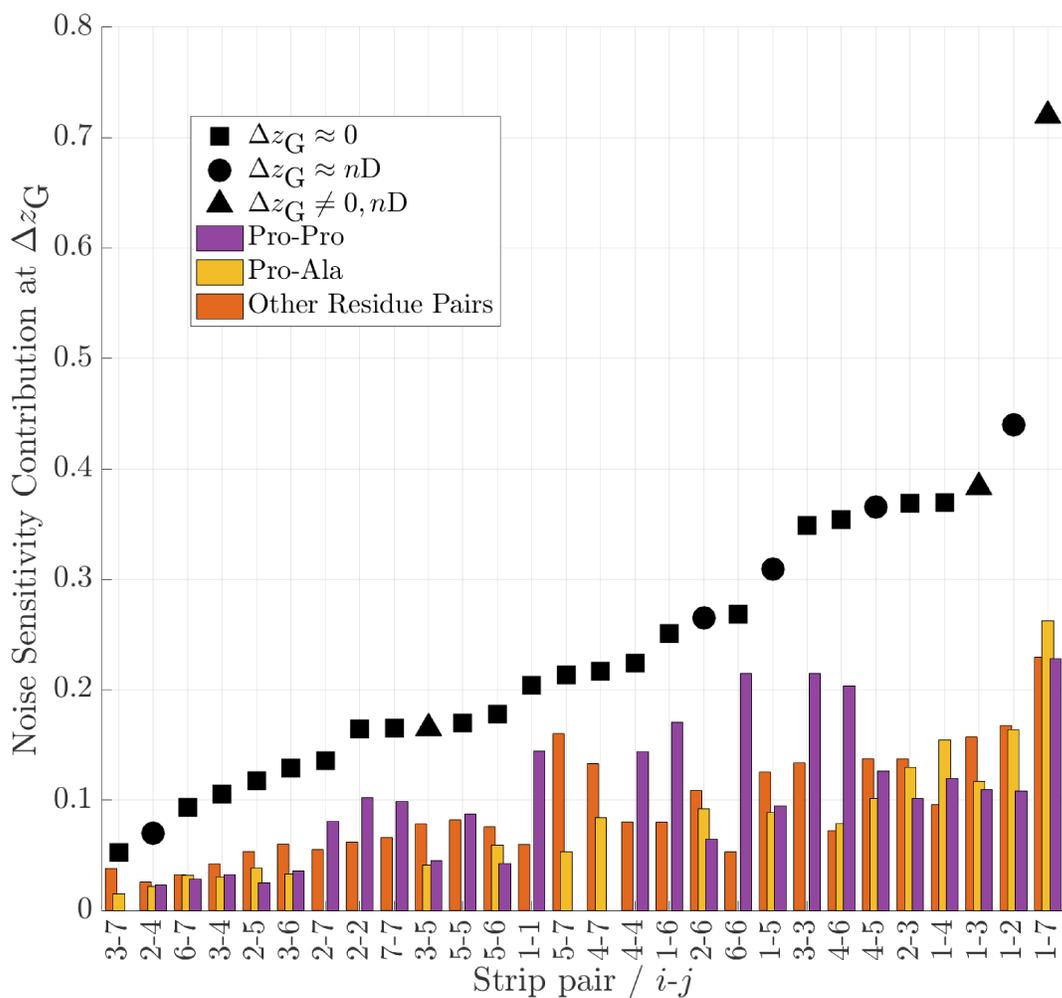


Figure S1. The largest contributions to the noise sensitivity of the global minima of the strip-strip energy $E_{i,j}^p(\Delta z_G)$ from the pairs of interacting residues. The black markers show the total noise sensitivity of each minimum. All residue pairs which contribute less than 20% to the total noise sensitivity, are categorised as ‘other residue pairs’ (see Materials and Methods for details). Typically, the highest contributions to the noise sensitivity come from two pairs of interacting residues, Pro-Pro and Pro-Ala.

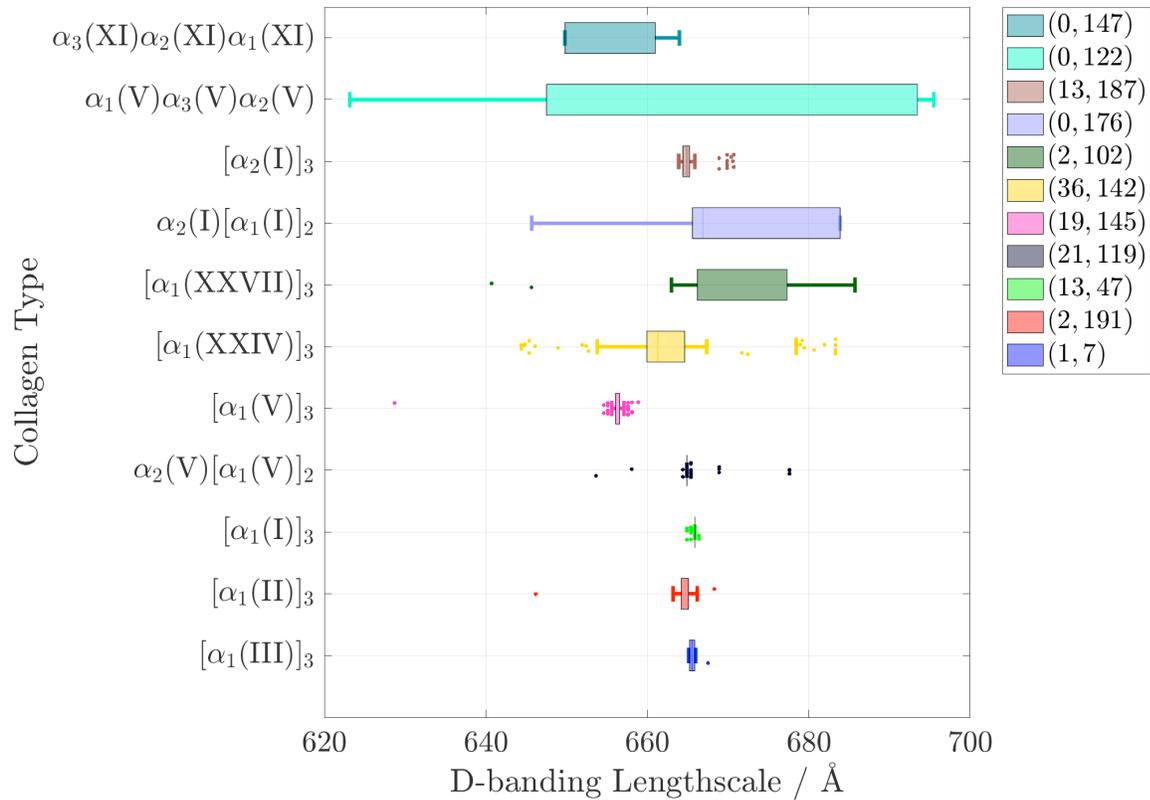


Figure S2. Box plot of the D-banding lengthscales across different mammalian species that give rise to stable perfectly-staggered microfibrils. The whiskers are drawn up to the largest/smallest data point that is within $1.5 \cdot \text{IQR}$ (inter-quartile range) of the upper/lower quartile, indicated by the top/bottom edges of each box respectively. D-banding lengthscales that are a distance more than $1.5 \cdot \text{IQR}$ from the top/bottom of a box are labelled as outliers and plotted as individual points. To each box, we associate an ordered pair (N_{out}, N_{tot}) , where N_{out} denotes the number of outliers and N_{tot} denotes the total number of species for which stable perfectly-staggered microfibrils were found.

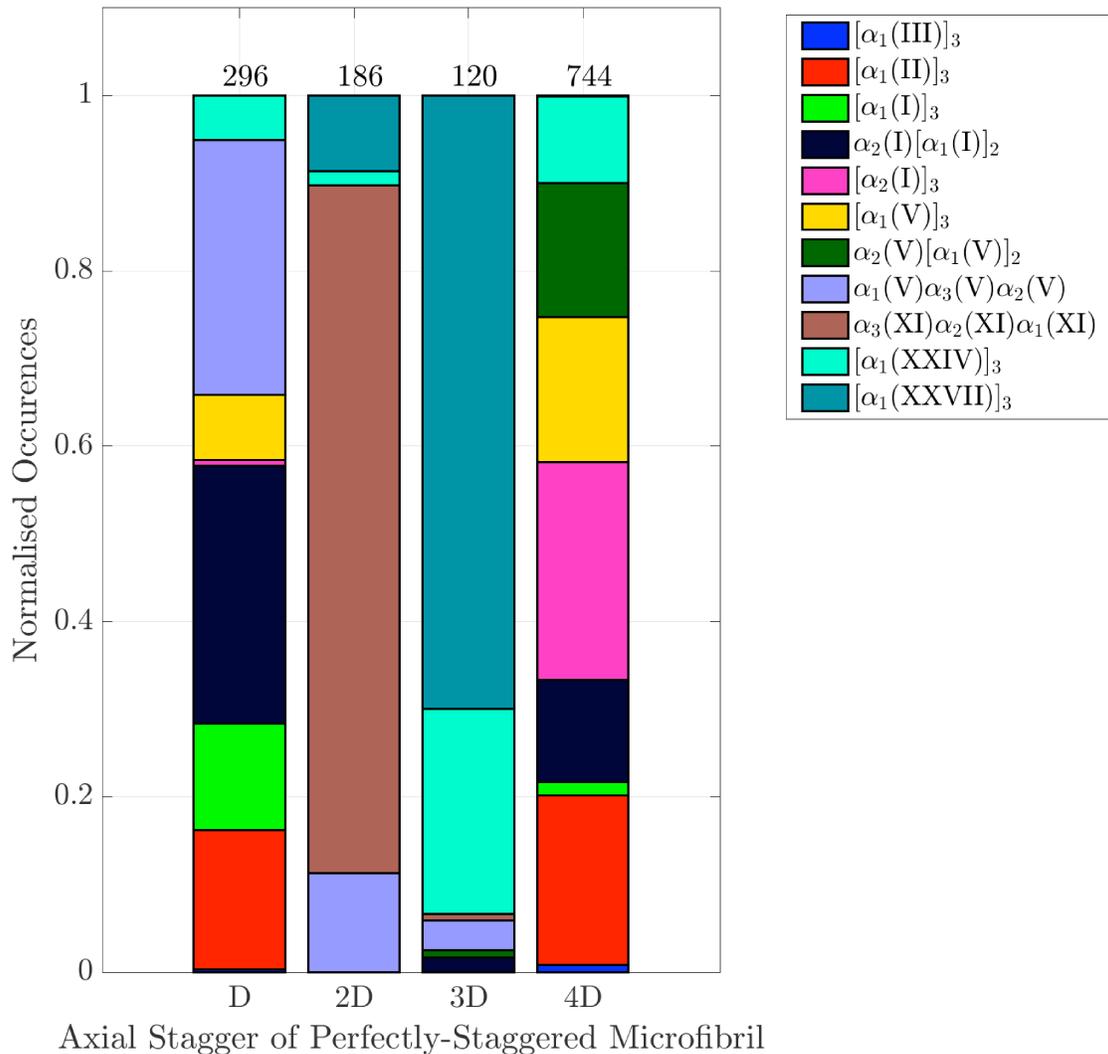


Figure S3. Histogram of the axial stagger value in stable perfectly-staggered microfibrils across different collagen types in mammalian species. The number of occurrences is normalised by the total number of stable perfectly-staggered microfibrils with a given axial stagger, which is shown at the top of each bar. A microfibril is deemed perfectly-staggered, provided that each axial stagger is within 5% of the same integer multiple of the D-banding lengthscale (values used are those shown in Figure S2).

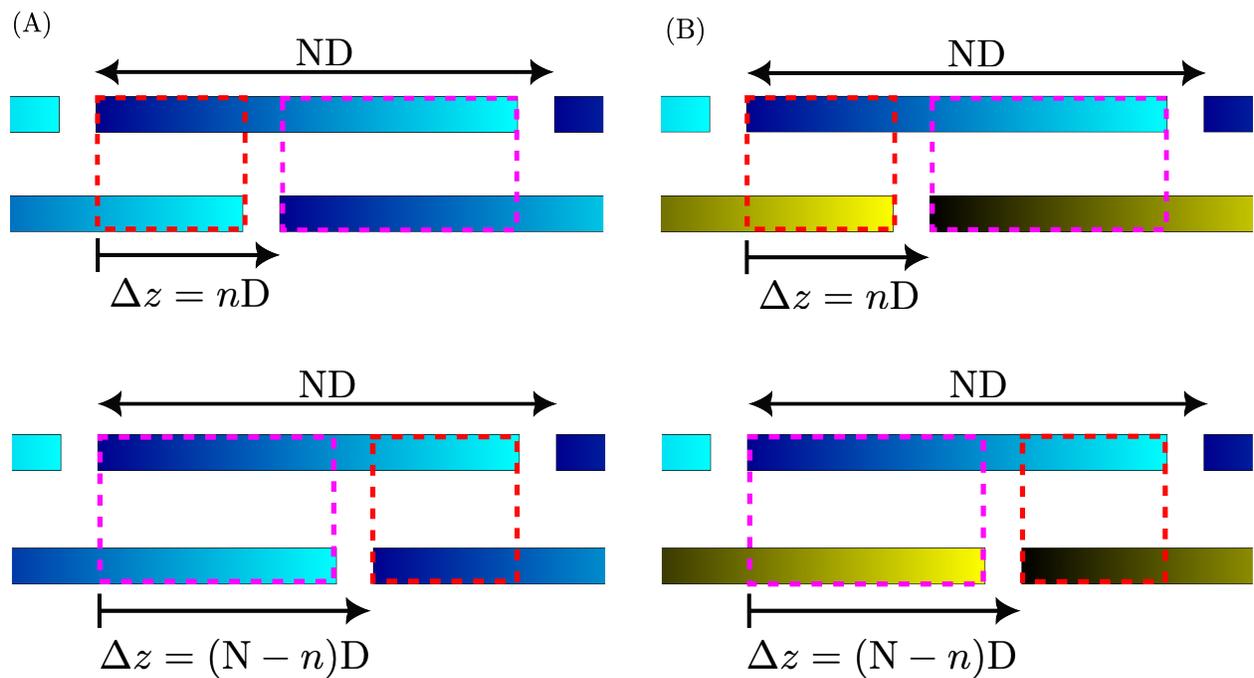


Figure S4. Schematic representation of pairwise, ND axially periodic tropocollagen interactions showing: (A). Equivalence of interactions at $\Delta z = nD$ and $\Delta z = (N - n)D$ for each molecule contributing an identical set of residues. (B). Non-equivalence of interactions at $\Delta z = nD$ and $\Delta z = (N - n)D$ when each molecule contributes a distinct set of residues. Different colours represent distinct residue contributions.

REFERENCES

- [1] Jordi Bella and David JS Hulmes. "fibrillar collagens". In *Fibrous proteins: structures and mechanisms*, pages 457–490. Springer, 2017.
- [2] Natalie Reznikov, Matthew Bilton, Leonardo Lari, Molly M Stevens, and Roland Kröger. Fractal-like hierarchical organization of bone begins at the nanoscale. *Science*, 360(6388): eaao2189, 2018.
- [3] Thijs Koorman, Karin A Jansen, Antoine Khalil, Peter D Haughton, Daan Visser, Max AK Rätze, Wisse E Haakma, Gabrièlè Sakalauskaitè, Paul J van Diest, Johan de Rooij, et al. Spatial collagen stiffening promotes collective breast cancer cell invasion by reinforcing extracellular matrix alignment. *Oncogene*, 41(17):2458–2469, 2022.
- [4] Kenneth M Towe. Oxygen-collagen priority and the early metazoan fossil record. *Proceedings of the National Academy of Sciences*, 65(4):781–788, 1970.
- [5] Aaron L Fidler, Carl E Darris, Sergei V Chetyrkin, Vadim K Pedchenko, Sergei P Boudko, Kyle L Brown, W Gray Jerome, Julie K Hudson, Antonis Rokas, and Billy G Hudson. Collagen iv and basement membrane at the evolutionary dawn of metazoan tissues. *eLife*, 6:e24176, 2017.
- [6] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, 2016.
- [7] S Michael Yu, Yang Li, and Daniel Kim. Collagen mimetic peptides: progress towards functional applications. *Soft Matter*, 7(18):7927–7938, 2011.
- [8] Yujia Xu and Michele Kirchner. Collagen mimetic peptides. *Bioengineering*, 8(1):5, 2021.
- [9] Nagmeh Rezaei, Aaron Lyons, and Nancy R Forde. Environmentally controlled curvature of single collagen proteins. *Biophysical Journal*, 115(8):1457–1469, 2018.
- [10] Joseph PRO Orgel, Thomas C Irving, Andrew Miller, and Tim J Wess. Microfibrillar structure of type i collagen in situ. *Proceedings of the National Academy of Sciences*, 103(24):9001–9005, 2006.
- [11] Andrew Wieczorek, Naghmeh Rezaei, Clara K Chan, Chuan Xu, Preeti Panwar, Dieter Brömme, Erika F Merschrod S, and Nancy R Forde. Development and characterization of a eukaryotic expression system for human type ii procollagen. *BMC biotechnology*, 15:1–17, 2015.
- [12] Christopher K Revell, Oliver E Jensen, Tom Shearer, Yinhui Lu, David F Holmes, and Karl E Kadler. Collagen fibril assembly: New approaches to unanswered questions. *Matrix Biology Plus*, 12:100079, 2021.
- [13] Ming Fang, Elizabeth L Goldstein, A Simon Turner, Clifford M Les, Bradford G Orr, Gary J Fisher, Kathleen B Welch, Edward D Rothman, and Mark M Banaszak Holl. Type i collagen d-spacing in fibril bundles of dermis, tendon, and bone: bridging between nano-and micro-level tissue hierarchy. *ACS nano*, 6(11):9503–9514, 2012.
- [14] Junjie Chen, Taeyong Ahn, Isabel D Colón-Bernal, Jinhee Kim, and Mark M Banaszak Holl. The relationship of collagen structural and compositional heterogeneity to tissue mechanical properties: a chemical perspective. *ACS nano*, 11(11):10665–10671, 2017.
- [15] John A Petruska and Alan J Hodge. A subunit model for the tropocollagen macromolecule. *Proceedings of the National Academy of Sciences*, 51(5):871–876, 1964.

- [16] JW Smith. Molecular pattern in native collagen. *Nature*, 219(5150):157–158, 1968.
- [17] Benes L Trus and Karl A Piez. Compressed microfibril models of the native collagen fibril. *Nature*, 286(5770):300–301, 1980.
- [18] Joseph PRO Orgel, Andrew Miller, Thomas C Irving, Robert F Fischetti, Andrew P Hammersley, and Tim J Wess. The in situ supermolecular structure of type i collagen. *Structure*, 9(11):1061–1069, 2001.
- [19] David JS Hulmes, Andrew Miller, David AD Parry, Karl A Piez, and John Woodhead-Galloway. Analysis of the primary structure of collagen for the origins of molecular packing. *Journal of Molecular Biology*, 79(1):137–148, 1973.
- [20] Benes L Trus and Karl A Piez. Molecular packing of collagen: three-dimensional analysis of electrostatic interactions. *Journal of Molecular Biology*, 108(4):705–732, 1976.
- [21] Karl A Piez and Benes L Trus. Sequence regularities and packing of collagen molecules. *Journal of Molecular Biology*, 122(4):419–432, 1978.
- [22] H Hofmann, PP Fietzek, and K Kühn. The role of polar and hydrophobic interactions for the molecular packing of type i collagen: a three-dimensional evaluation of the amino acid sequence. *Journal of Molecular Biology*, 125(2):137–165, 1978.
- [23] James M Chen, Chun E Kung, Stephen H Fairheller, and Eleanor M Brown. An energetic evaluation of a “smith” collagen microfibril model. *Journal of Protein Chemistry*, 10:535–552, 1991.
- [24] Anna M Puzzkarska, Daan Frenkel, Lucy J Colwell, and Melinda J Duer. Using sequence data to predict the self-assembly of supramolecular collagen structures. *Biophysical Journal*, 121(16):3023–3033, 2022.
- [25] J Robin Harris and Richard J Lewis. The collagen type i segment long spacing (sls) and fibrillar forms: Formation by atp and sulphonated diazo dyes. *Micron*, 86:36–47, 2016.
- [26] DJ Hulmes, Romaine R Bruns, and Jerome Gross. On the state of aggregation of newly secreted procollagen. *Proceedings of the National Academy of Sciences*, 80(2):388–392, 1983.
- [27] Andreas Stylianou. Assessing collagen d-band periodicity with atomic force microscopy. *Materials*, 15(4):1608, 2022.
- [28] Jordi Bella. A new method for describing the helical conformation of collagen: Dependence of the triple helical twist on amino acid sequence. *Journal of Structural Biology*, 170(2):377–391, 2010.
- [29] Jan K Rainey and M Cynthia Goh. A statistically derived parameterization for the collagen triple-helix. *Protein Science*, 11(11):2748–2754, 2002.
- [30] Joseph PRO Orgel, Anton V Persikov, and Olga Antipova. Variation in the helical structure of native collagen. *PLoS One*, 9(2):e89519, 2014.
- [31] Sébastien Neukirch, Alain Goriely, and Andrew C Hausrath. Chirality of coiled coils: elasticity matters. *Physical Review Letters*, 100(3):038105, 2008.
- [32] Jie Liu, Wei Yong, Yiqun Deng, Neville R Kallenbach, and Min Lu. Atomic structure of a tryptophan-zipper pentamer. *Proceedings of the National Academy of Sciences*, 101(46):16156–16161, 2004.
- [33] David R Baselt, Jean-Paul Revel, and John D Baldeschwieler. Subfibrillar structure of type i collagen observed by atomic force microscopy. *Biophysical Journal*, 65(6):2644–2655, 1993.
- [34] Mario Raspanti, Marcella Reguzzoni, Marina Protasoni, and Petra Basso. Not only tendons: The other architecture of collagen fibrils. *International Journal of Biological Macromolecules*, 107:1668–1674, 2018.

- [35] Jody M Mason and Katja M Arndt. Coiled coil domains: stability, specificity, and biological implications. *ChemBioChem*, 5(2):170–176, 2004.
- [36] RDBM Fraser and MacRae TP. *Conformation in fibrous proteins and related synthetic polypeptides*. New York: Academic Press, 1973.
- [37] Daniel J McBride Jr, Vincent Choe, Jay R Shapiro, and Barbara Brodsky. Altered collagen structure in mouse tail tendon lacking the $\alpha 2$ (i) chain. *Journal of Molecular Biology*, 270(2): 275–284, 1997.
- [38] Olga Antipova and Joseph PRO Orgel. In situ d-periodic molecular structure of type ii collagen. *Journal of Biological Chemistry*, 285(10):7087–7096, 2010.
- [39] Meisam Asgari, Neda Latifi, Hossein K Heris, Hojatollah Vali, and Luc Mongeau. In vitro fibrillogenesis of tropocollagen type iii in collagen type i affects its relative fibrillar topology and mechanics. *Scientific Reports*, 7(1):1392, 2017.
- [40] Barbara Brodsky, Eric F Eikenberry, and Kathleen Cassidy. An unusual collagen periodicity in skin. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 621(1):162–166, 1980.
- [41] H elene Chanut-Delalande, Agnes Fichard, Simonetta Bernocco, Robert Garrone, David JS Hulmes, and Florence Ruggiero. Control of heterotypic fibril formation by collagen v is determined by chain stoichiometry. *Journal of Biological Chemistry*, 276(26):24352–24359, 2001.
- [42] Kazunori Mizuno, Hans Peter B achinger, Yasutada Imamura, Toshihiko Hayashi, and Eijiro Adachi. Fragility of reconstituted type v collagen fibrils with the chain composition of $\alpha 1$ (v) $\alpha 2$ (v) $\alpha 3$ (v) respective of the d-periodic banding pattern. *Connective Tissue Research*, 54(1): 41–48, 2013.
- [43] Uwe Hansen and Peter Bruckner. Macromolecular specificity of collagen fibrillogenesis: fibrils of collagens i and xi contain a heterotypic alloyed core and a collagen i sheath. *Journal of Biological Chemistry*, 278(39):37352–37359, 2003.
- [44] Darren A Plumb, Vivek Dhir, Aleksandr Mironov, Laila Ferrara, Richard Poulson, Karl E Kadler, David J Thornton, Michael D Briggs, and Raymond P Boot-Handford. Collagen xxvii is developmentally regulated and forms thin fibrillar structures distinct from those of classical vertebrate fibrillar collagens. *Journal of Biological Chemistry*, 282(17):12791–12795, 2007.
- [45] E Leikina, MV Merts, N Kuznetsova, and S Leikin. Type i collagen is thermally unstable at body temperature. *Proceedings of the National Academy of Sciences*, 99(3):1314–1318, 2002.
- [46] Yujia Xu and Michele Kirchner. "segment-long-spacing (sls) and the polymorphic structures of fibrillar collagen". In *Macromolecular Protein Complexes IV: Structure and Function*, pages 495–521. Springer, 2022.
- [47] St ephanie Perret, Christine Merle, Simonetta Bernocco, Patricia Berland, Robert Garrone, David JS Hulmes, Manfred Theisen, and Florence Ruggiero. Unhydroxylated triple helical collagen i produced in transgenic plants provides new clues on the role of hydroxyproline in collagen folding and fibril formation. *Journal of Biological Chemistry*, 276(47):43693–43698, 2001.
- [48] Eileen S Hwang, Geetha Thiagarajan, Avanish S Parmar, and Barbara Brodsky. Interruptions in the collagen repeating tripeptide pattern can promote supramolecular association. *Protein Science*, 19(5):1053–1064, 2010.
- [49] Raymond P Boot-Handford, Danny S Tuckwell, Darren A Plumb, Claire Farrington Rock, and Richard Poulson. A novel and highly conserved collagen (pro $\alpha 1$ (xxvii)) with a unique expression pattern and unusual molecular characteristics establishes a new clade within the

- vertebrate fibrillar collagen family. *Journal of Biological Chemistry*, 278(33):31067–31077, 2003.
- [50] Manuel Koch, Friedrich Laub, Peihong Zhou, Rita A Hahn, Shizuko Tanaka, Robert E Burgeson, Donald R Gerecke, Francesco Ramirez, and Marion K Gordon. Collagen xxiv, a vertebrate fibrillar collagen with structural features of invertebrate collagens: selective expression in developing cornea and bone. *Journal of Biological Chemistry*, 278(44):43236–43244, 2003.
- [51] Antonella Forlino, Wayne A Cabral, Aileen M Barnes, and Joan C Marini. New perspectives on osteogenesis imperfecta. *Nature Reviews Endocrinology*, 7(9):540–557, 2011.
- [52] Manuela Venturoni, Thomas Gutschmann, Georg E Fantner, Johannes H Kindt, and Paul K Hansma. Investigations into the polymorphism of rat tail tendon fibrils using atomic force microscopy. *Biochemical and Biophysical Research Communications*, 303(2):508–513, 2003.
- [53] Fangfang Chen, Rebecca Strawn, and Yujia Xu. The predominant roles of the sequence periodicity in the self-assembly of collagen-mimetic mini-fibrils. *Protein Sci.*, 28(9):1640–1651, September 2019.
- [54] Natalia Kuznetsova and Sergey Leikin. Does the triple helical domain of type i collagen encode molecular recognition and fiber assembly while telopeptides serve as catalytic domains?: effect of proteolytic cleavage on fibrillogenesis and on collagen-collagen interaction in fibers. *Journal of Biological Chemistry*, 274(51):36083–36088, 1999.
- [55] Shuichi Kawashima and Minoru Kanehisa. Aaindex: amino acid index database. *Nucleic Acids Research*, 28(1):374–374, 2000.
- [56] Nicolaas Govert De Bruijn. *Asymptotic methods in analysis*. Courier Corporation, 2014.