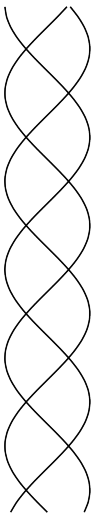
Physical Mechanisms of Collagen Self-Assembly:

From Residue Interactions to Microfibril Structural Polymorphism

Artom Zolotarjov



PhD

University of York
Mathematics

September 2025

Abstract

The structural protein collagen is often called the scaffold of multicellular life. From bone and cornea to the extracellular matrix and basement membranes of our vascular system, collagen is a key provider of structural support across a broad range of biological tissues.

The unique versatility of collagen as a structural protein is crucially aided by its ability to undergo self-assembly into intricate hierarchical structures. The hierarchical aggregates of collagen are characteristic in displaying various types of structural polymorphism. Yet, the biological functions of collagen are dependent on the structural features of its specific polymorphic aggregates.

The study of structural polymorphism of collagen constitutes the central topic of this thesis. Throughout the main chapters of this work, we aim to gain understanding of the physical interactions and mechanisms that govern the existence and competition between different polymorphic aggregates of collagen. To that end, we focus on the self-assembly of the smallest distinct unit in the collagen structural hierarchy - the microfibril.

Beginning with chapter 2, we show that the spatial organisation of interacting amino acids on the collagen molecular surface is chiral. We then demonstrate that this 3-dimensional spatial arrangement of amino acids is a crucial determinant of the structural features of collagen microfibrils as well as their polymorphisms. Continuing with chapter 3, we investigate the impact of ionic effects, namely pH and pKa on the aggregation of collagen microfibrils. We demonstrate how the aforementioned parameters as well as specific amino acid interactions stabilise different polymorphic aggregates of collagen under varying ionic conditions. Finally, in chapter 4, we study the polymorphisms of fibrous long-spacing (FLS) collagens. We determine the physical interactions that lead to their formation and predict the 3-dimensional molecular organisation within FLS collagen microfibrils.

Contents

List of Tables

List of Figures

Acknowledgements

The summer of 2019 was marked by my first encounter with collagen during a research project that I was doing at the time. I cannot say that my acquaintance with collagen can be described as "love at first sight", as at the end of the project I swore to never engage with the field again. Yet, in spite of my initial sentiments, here we are, some six years later with a PhD thesis titled "Physical Mechanisms of Collagen Self-Assembly". An acknowledgement of the various people that have supported and accompanied me on this academic journey is in order.

First and foremost, I would like to extend gratitude to my supervisor Dmitri Pushkin. Much of how I think about physics has been shaped and influenced by our discussions over the years. I greatly value the freedom that I have had to pursue topics of personal interest to me as well as your support and enthusiasm when it came to collaborating on the various research projects.

This PhD would have been a much more daunting experience without the support of my partner Sophie, who has been on the receiving end of many collagen-related trivia, windy ramblings as well as hour-long monologues about the technical details of numerical simulations. In all of those and other instances, you have treated me with love, kindness, patience and your undivided attention, for which I am forever grateful.

I want to also thank my parents Antonina and Oleg, my sister Maria and her husband Evgeny as well as my nephew Roman for showing genuine interest in my work as well as coming to visit me in the UK.

Finally, I want to acknowledge the community within the Mathematics Department that I have had the pleasure of being a part of. I thank my friend Kuntal for many intellectually stimulating conversations, humbling visits to the bouldering gym and too numerous to count pints of grapefruit Schöfferhofer. I also thank the members of the productivity "vortex" - Zane, Nick, Jamie and Simon, for providing the much needed levity to the PhD office. Last, but by no means least, I want to thank Ambroise for creating and sharing the LaTeX template that was used for writing this thesis.

Author's declaration

I declare that the work presented in this thesis, except where otherwise stated, is based on my own research carried out at the University of York and has not been submitted previously for any degree at this or any other university. Sources are acknowledged by explicit references.

The contents of chapter 2 are based on the arXiv pre-print: A. Zolotarjov, R. Kröger and D. O. Pushkin. Chiral interactions between tropocollagen molecules determine the collagen microfibril structure, arXiv:2504.21484, (2025). Chapters 3 and 4 have been written as for publication with intention of submission.

I am the sole author on all substantive chapters presented in this thesis. Project conceptualisation for chapter 2 is shared equally between me, my supervisor Dmitri Pushkin and collaborator Roland Kröger. Chapters 3 and 4 have been predominantly conceptualised by me, with valuable input from my supervisor. All numerical simulations and code development for chapters 2-4 was undertaken by me. Development of theoretical models was shared equally between me and my supervisor in chapter 2 and led primarily by me in chapters 3 and 4.

— 1 —

Introduction

1.1 A BRIEF INTRODUCTION TO COLLAGEN

The name of the structural protein collagen is assembled from the Ancient Greek words "κόλλα" ("glue") and "γενής" ("forming"), thus meaning "glue-producing" [14]. It is no exaggeration to say that the etymology accurately reflects the role of collagen in the bodies of living organisms as the "biological glue" that holds and connects them together. Collagen is the principal organic component of bone, imbuing it simultaneously with stiffness and toughness - a combination of mechanical properties that is often considered mutually exclusive [101]. At the same time, collagen is a major component of cornea, the outer layer of the eye, wherein it plays a vital role in maintaining both its shape and transparency for incoming light [73]. Collagen is also an important constituent of the biological network that surrounds cells - the extracellular matrix, in which it contributes to mediating cell activity and metabolism by acting as a bridge between mechanical and biochemical signalling pathways [31].

The unique structural versatility of collagen is attributed to its ability to aggregate into intricate hierarchical structures [101]. Individual collagen molecules, which in the first approximation can be thought of as long aspect ratio, semi-flexible cylinders of length $L \approx 300\,\text{nm}$ and diameter 1.5 nm, spontaneously self-assemble into cylindrically symmetric collagen fibrils, which span roughly 10-100 nm in diameter and at least several orders of magnitude more in length [99]. Collagen fibrils in turn assemble into higher-order hierarchical structures, which form the basis of biological tissues such as bone, cornea and the extracellular matrix [101].

Since at least the 1950s, a number of experimental studies have aspired to understand the structural versatility characteristic of collagen aggregates by studying the organisation of collagen molecules inside fibrils [91]. As anyone working with collagen knows, collagen fibrils famously display the D-banding pattern - a $D \approx 67\,\text{nm}$ periodic pattern of light and dark bands running along the long axis of the fibril (see top of Figure 1.1 and Figure 1.2B).

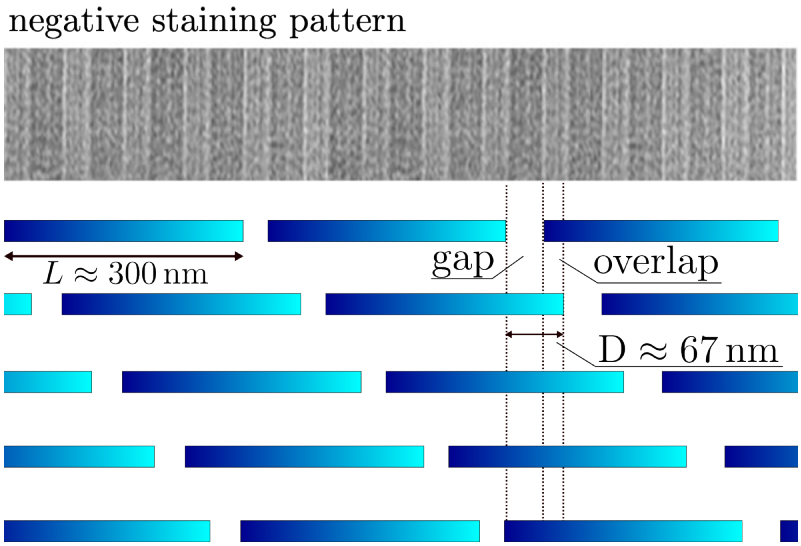


Figure 1.1: Hodge-Petruska scheme for a D-banded collagen microfibril. The negatively stained TEM image adapted from [118].

These dark and light bands appear in TEM (transmission electron microscopy) images with negative staining. To understand what the D-banding pattern conveys about the molecular organisation of collagen inside fibrils, we need to briefly recap the use of TEM for characterisation of biological tissues, which is accessibly detailed in [23]. TEM achieves contrast through differences in the electron scattering that arise from different portions of the analysed sample. Parts of the sample that heavily scatter the incoming electrons then appear darker than those that scatter more lightly. The elements comprising biological samples typically scatter electrons equally weakly, leading to poor image contrast. As such, staining agents which contain electron-dense elements (such as uranium or tungsten) and thus lead to strong scattering, are added to increase the image contrast. With negative staining, the stain is allowed to remain and dry in and around the specimen. This way, the stain can penetrate the spatial vacancies or in other words gaps inside the sample, creating a dark background in the image. On a few occasions, we will also reference positive staining, during which the sample is washed with a solvent following exposure to the stain. As a result, only the stain that reacts and binds to the sample is imaged (typically binding to charged amino acids).

Having the fundamentals of TEM imaging down, we now proceed to interpret the D-banding pattern in Figure 1.2B. This interpretation was first given by Hodge & Petruska in their seminal 1964 paper [90]. The authors proposed that collagen fibrils were comprised of fundamental periodic subunits consisting of 5 collagen molecules, which are now known as microfibrils. A 2-dimensional scheme of a single periodic microfibril is shown in Figure 1.1. The microfibril is "periodic" in the sense that a given collagen molecule of length

$L \approx 4.46$D is succeeded in the axial direction by another collagen molecule following a gap of size 0.54D. Each periodic molecular array is then translated (staggered) in the axial direction by the lengthscale D relative to its molecular neighbours, resulting in a series of alternating gap and overlap regions that appear with axial periodicity corresponding to the lengthscale D. The gaps in the microfibrillar structure are then interpreted to be the sites that accumulate the electron-dense stain, hence giving the D-banding pattern observed in TEM images.

Since the publication of Hodge & Petruska, multiple experimental studies found structural and mechanical evidence that supports the existence of 5-membered microfibrillar subunits in D-banded fibrils [83, 82, 43, 121]. Detailed X-ray scattering studies of microfibrillar structure by Orgel *et al.* [83, 82] have also revealed that collagen molecules inside a microfibril do not appear as straight cylinders, as suggested by Figure 1.1. Instead, they are supercoiled, meaning that the centrelines of the collagen molecules in a microfibril follow a right-handed helical path. It should be mentioned, however, that despite all the aforementioned experimental studies, isolated microfibrils are yet to be experimentally observed. Nevertheless, microfibrils have come to be widely accepted as the fundamental building block in the collagen structural hierarchy [82, 101, 10].

In the years following the work of Hodge & Petruska, it was increasingly realised that D-banded fibrils are just one member of a large family of polymorphic aggregates that collagen may form both *in vitro* and *in vivo* under varying ionic and biological conditions [120]. As can be seen in Figure 1.2, collagen fibrils may display a wide range of both axially periodic and non-periodic morphologies. Figures 1.2A, D-F illustrate four types of fibrous long-spacing (FLS) collagens, all of which are characterised by axial periodicities that significantly exceed the lengthscale D [22, 33]. This indicates that other stable periodic axial molecular arrangements are possible in collagen aggregates. Figure 1.2C shows a positively stained TEM image of segment-long-spacing (SLS) aggregates. The positive staining pattern can be exactly matched to the distribution of charged amino acids in the collagen sequence, leading to the conclusion that SLS aggregates are comprised of collagen molecules arranged in-register, with no axial stagger [120]. Finally, Figure 1.2G shows collagen fibrils that exhibit no periodic axial molecular order at all, demonstrating that periodic molecular arrangement is not necessarily required for fibril formation.

It is important to note that the majority of the aforementioned polymorphic collagen fibrils forms in presence of additional charged molecular species, such as ATP (adenosine triphosphate), chondroitin sulphate and proteoglycans among others [120]. This in turn indicates that whilst collagen-collagen interactions may encode the diverse fibril polymorphisms, additional interactions involving aforementioned charged molecular species also play a crucial role. The exceptions to this rule are the D-banded fibrils, which form spontaneously under physiological ionic conditions as well as the disordered fibrils in Figure

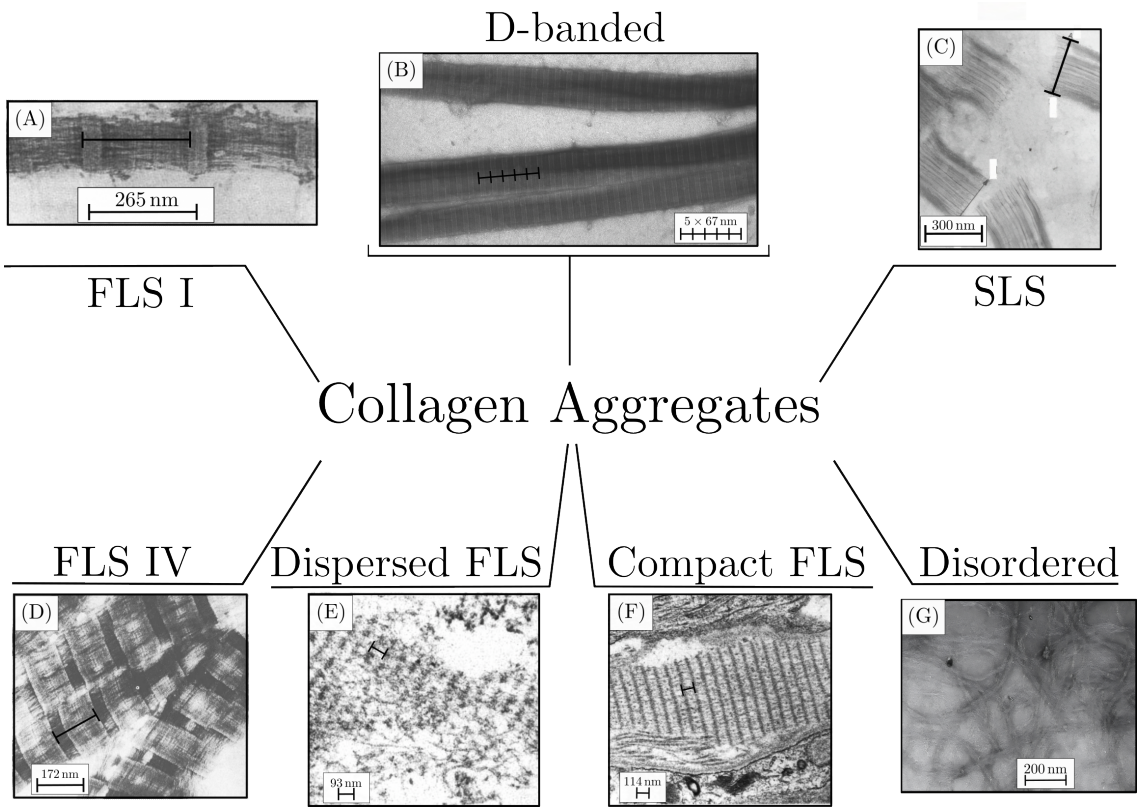


Figure 1.2: Polymorphic collagen aggregates: (A). FLS (fibrous long-spacing) I [22], (B). D-banded [51], (C). SLS (segment-long-spacing) [120], (D). FLS IV [3] , (E). Dispersed FLS [33], (F). Compact FLS [33], (G). Disordered [51] . Image in panel (C) is obtained using positively stained TEM. The remaining panels correspond to negatively stained TEM images.

1.2G, which require acidic pH but otherwise form under the same experimental conditions (see [51] for details).

As evidenced by Figure 1.2, collagen is a true polymorphic structure. Yet, the physical properties of collagen-containing tissues, such as bone, cornea and the extracellular matrix are crucially dependent on the specific molecular organisation observed in D-banded fibrils [30, 101]. This raises two crucial questions that we will return to repeatedly throughout this thesis. (1). What physical interactions lead to the emergence of structural polymorphism in collagen aggregates? (2). What physical mechanisms drive the competition between polymorphic aggregates of collagen? We now proceed to outline the broad theoretical approach that we take in tackling the aforementioned matters.

It would be a truly hopeless task to attempt to capture the complexity and diversity of the entire collagen structural hierarchy, as such we will focus on its lowest rung - the collagen microfibril. We will be interested in constructing a simple set of interaction

potentials between collagen molecules that will then be used in an equilibrium microfibril self-assembly model. Theoretical modelling of microfibril self-assembly is of utility for two distinct reasons. Firstly, due to the relatively small system size, it is feasible to directly link the physical interactions between collagen molecules with the emergent structural properties of microfibrils. Secondly, being the fundamental subunit in the collagen structural hierarchy, the structure of the microfibril directly informs both the morphology (such as D-banding) as well as the physical properties of higher-order assemblies in the collagen structural hierarchy. As such, the acquired structural understanding is not confined to the microfibrillar lengthscale.

There are three key structural characteristics of microfibrils that we will study with the aid of our self-assembly models. Firstly, it is the axial periodicity of the microfibril, instated by periodically alternating gap and overlap regions. We will determine how the interactions between collagen molecules encode the diverse set of axial periods observed in Figure 1.2. Secondly, we will be interested in understanding the spatial arrangement of molecules in a microfibril. Moving beyond the 2-dimensional Hodge-Petruska scheme of Figure 1.1, we will identify the physical principles that determine the number of molecules in polymorphic microfibrils as well as the 3-dimensional packing of collagen molecules therein. The final structural feature that will be of interest to us, is the molecular supercoiling of collagen in a microfibril. We will explain the physical origin of this molecular supercoiling as well as the preference for right-handedness. We now proceed to outline the content of each thesis chapter.

1.2 THESIS STRUCTURE

This thesis is written in a journal-style presentation and is comprised of three substantive chapters written in a paper format. Each chapter begins with a "preliminary remarks" section, which bridges the chapters and provides the background knowledge where necessary.

Chapter 2 introduces the intermolecular potentials that are used throughout the thesis to model the interactions between collagen molecules. In constructing these potentials, we introduce a novel approach to coarse-graining of the collagen molecular structure. Our approach retains the 3-dimensional organisation of amino acids that comprise collagen, enabling us to demonstrate the key role of their 3-dimensional spatial organisation in determining the structural features of D-banded microfibrils (Figure 1.2B). In particular, for the first time, we define physical interactions that give rise to molecular supercoiling of collagen in microfibrils. Finally, we introduce a novel type of emergent chirality that appears at the microfibrillar level and may be important for bone mineralisation.

Chapter 3 addresses the role of ionic conditions in microfibril self-assembly. We propose the first physical model that predicts phase transitions between the disordered microfibrils (see Figure 1.2G) and D-banded microfibrils. We demonstrate that this phase transition is controlled by interactions involving specific ionisable amino acids. Further, we show that under physiological ionic conditions, charged residue interactions preclude in-register aggregation of collagen molecules (see Figure 1.2C), instead favouring formation of D-banded microfibrils. The last result is significant, as a number of previous theoretical models have predicted that in-register aggregates of collagen represent a more energetically stable conformation than D-banded fibrils, which disagrees with existing experimental knowledge [53, 120].

Chapter 4 generalises the definition of axial periodicity to account for FLS collagens (see Figures 1.2A, D-F). We demonstrate that collagen-collagen interactions encode the FLS collagen axial periodicities, which is contrary to the existing knowledge in the field [34, 120]. Furthermore we predict, starting from pairwise interactions between collagen molecules, the 3-dimensional molecular packing for FLS I and FLS IV collagen microfibrils. This result is important, since 3-dimensional molecular packing models for FLS collagens are, to the best of our knowledge, completely absent from literature.

— 2 —

Chiral Interactions between Collagen Molecules Determine the Collagen Microfibril Structure

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 axially periodic collagen microfibril remain poorly understood. Here, we demonstrate that the 3-dimensional spatial organisation of amino acids in molecular collagen determines its supramolecular organisation. By combining statistical modelling of residue geometry with sequence-informed interaction potentials, we show that the chiral spatial arrangement of outward-facing residues and the elasticity of the triple helix control molecular supercoiling. We demonstrate that the geometric arrangement of residues determines the energetic selection of aggregate size in microfibrils. Finally, we define a novel type of emergent chirality at the microfibrillar level, which may play a role in directing the organisation of the mineral phase in bone.

2.1 PRELIMINARY REMARKS

The first step in modelling collagen self-assembly involves describing the physical interactions between collagen molecules. A single molecule of collagen is comprised of approximately 3000 amino acids, making it computationally intractable to directly evaluate pairwise molecular interactions through techniques such as molecular dynamics simulations. Coarse-grained theoretical approaches to modelling the energy of collagen molecular interactions and subsequently the energy of collagen aggregates offer a computationally tractable alternative and are one of the primary theoretical tools leveraged throughout this thesis. A crucial step in such models involves deciding on an appropriate coarse-grained representation of the collagen molecular structure. Discussing these coarse-graining methods necessitates first establishing the fundamental features of the collagen molecular structure, which will be often mentioned throughout this thesis.

Table 2.1: Naming conventions for the set of 20 proteinogenic residues as well as hydroxyproline.

[a]Residue Name	Three Letter Abbreviation
Arginine	**Arg**
Aspartic acid	**Asp**
Glutamic acid	**Glu**
Glycine	**[b]Gly**
Histidine	**His**
Lysine	**Lys**
Proline	**Pro**
Hydroxyproline	**Hyp**
Alanine	Ala
Asparagine	Asn
Cysteine	Cys
Glutamine	Gln
Isoleucine	Ile
Leucine	Leu
Methionine	Met
Phenylalanine	Phe
Serine	Ser
Threonine	Thr
Tryptophan	Trp
Tyrosine	Tyr
Valine	Val

[a] We have identified the residues that will be frequently mentioned throughout this thesis in a boldface font.
[b] For convenience, we will sometimes shorten Gly further to G.

We start by noting some elementary terminology pertaining to protein biochemistry. In line with conventions in the field, we will interchangeably refer to the amino acids that comprise collagen as "residues". There are 20 proteinogenic residues, which are commonly referred to by their 3-letter abbreviations, which are summarised in Table 2.1. In collagen, certain residues may undergo post-translational modifications, which alter the residue molecular structure. The most common type of post-translational modification is the hydroxylation of proline into hydroxyproline [10]. For the latter, we reserve a separate non-standard three-letter abbreviation "Hyp".

Residues are distinguished from one another by their side-chains, also often referred to as R-groups. Interactions between residue R-groups play a defining role in both stabilisation of the collagen molecular structure itself as well as in collagen-collagen interactions [53, 92, 14]. The specific type of side-chain determines the types of interactions (electrostatic, hydrophobic, covalent, hydrogen bonding etc.) that a given residue might partake in. Figure 2.1 illustrates an example of a polypeptide comprised of 6 residues. As can be seen with this example, the residue R-groups can differ drastically in terms of molecular structure, size and steric conformation. For the purposes of calculating residue-residue interactions, a common simplification is to represent the spatial coordinates of a residue via a single point, corresponding to the spatial coordinates of the residue's C_α atom. The C_α atoms alternating with peptide bonds, comprise the backbone of a given polypeptide, as can be seen in Figure 2.1 [37, 95]. In our calculations, we will often resort to this approximation, for the purpose of tractability.

Armed with biochemical vocabulary, we now proceed to describe the molecular structure of collagen [10]. Collagen is a fibrillar protein, characterised by a length $L \approx 300\,\text{nm}$ and a

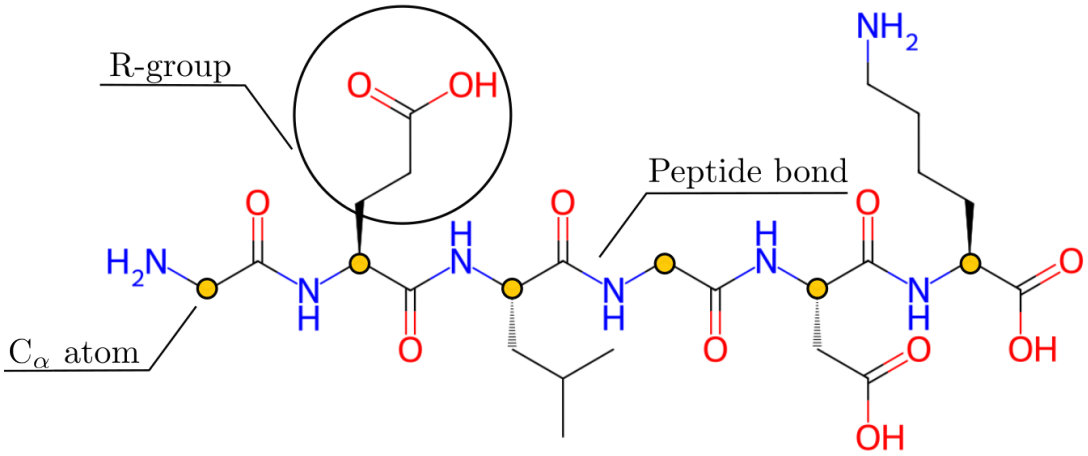
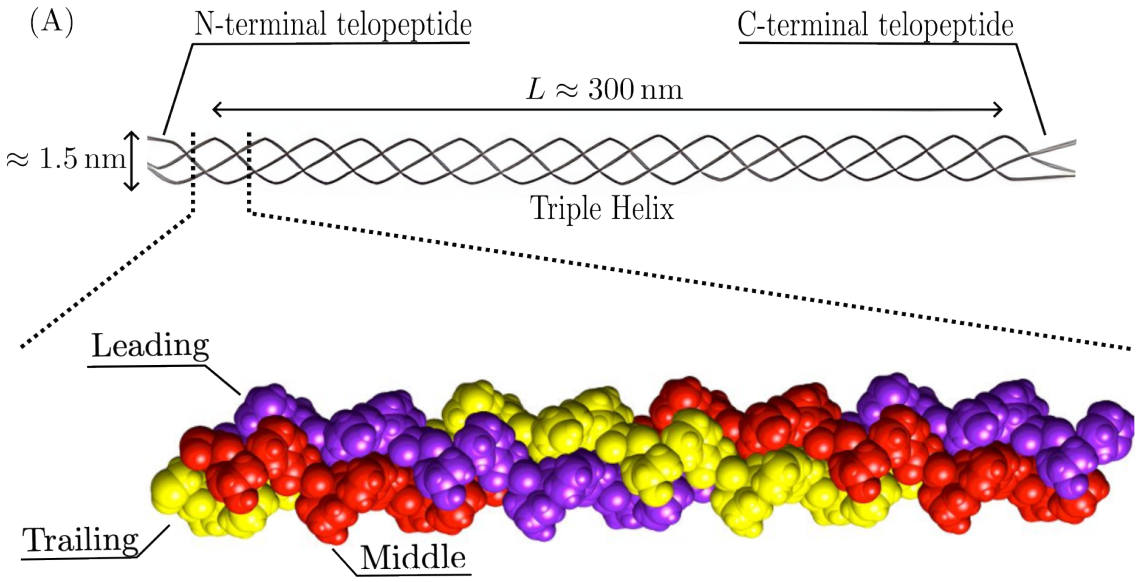


Figure 2.1: Molecular structure of a Gly-Glu-Leu-Gly-Asp-Lys polypeptide. The 2-dimensional structure has been generated using `pyPept` package [81].

diameter $\approx 1.5\,\text{nm}$, see top of Figure 2.2A. The collagen molecule itself is an aggregate of three left-handed helical polypeptides, known as α-chains. If the α-chains are identical, then the resulting collagen molecule is termed homotrimeric and, in the converse case, heterotrimeric. The characteristic feature of α-chains is the presence of a repetitive sequence motif [Gly-X-Y], where X and Y are often Pro or Hyp respectively. The presence of Gly residues in every third sequence position is crucial for the association of α-chains. During α-chain trimerisation, the centrelines of helical α-chains are deformed into right-handed helices in a process commonly referred to as supercoiling. Right-handed supercoiling of α-chains results in the burial of Gly residues on the inside of the resulting collagen molecule, enabling formation of inter-chain hydrogen bonds that stabilise the whole trimeric assembly [8]. The R-groups of residues in the X and Y sequence positions end up on the surface of the resulting polypeptide assembly, leaving them open for interactions [14]. The resulting trimerisation domain is known as the triple helix, which is shown in Figure 2.2A. In the



(B)

$$\begin{aligned} &G_1^t\text{-}X_2^t\text{-}Y_3^t\text{-}G_4^t\text{-}X_5^t\text{-}Y_6^t\text{-}G_7^t\text{-}\cdots \\ &G_1^m\text{-}X_2^m\text{-}Y_3^m\text{-}G_4^m\text{-}X_5^m\text{-}Y_6^m\text{-}\cdots \\ &G_1^l\text{-}X_2^l\text{-}Y_3^l\text{-}G_4^l\text{-}X_5^l\text{-}\cdots \end{aligned}$$

Figure 2.2: (A). (TOP) Schematic representation of the collagen triple helix with telopeptide domains. (BOTTOM) Space-filling diagram of a section of the triple helix. (B). Schematic representation of a triple helix section in terms of the comprising residues. Gly has been shortened to G. Images adapted from [118, 8].

majority of cases of interest to this thesis, the triple helical region is an uninterrupted domain that corresponds roughly to 95% of the collagen molecule. The rest of the collagen molecule is comprised of N- and C-terminal non-helical telopeptide domains. Although the telopeptide domains may play a catalytic role in collagen self-assembly, they are not essential for it [64]. Throughout this thesis we will focus on the interactions between triple helical regions of collagen molecules and will thus use the terms "collagen molecule" and "triple helix" interchangeably.

It is helpful to distinguish between the start and end of a collagen molecule, which we conventionally take to be the N-terminus (starting with the N-terminal telopeptide) and the C-terminus (ending with a C-terminal telopeptide). Within the triple helix, the α-chains are not in register, but instead have a single residue offset along the long axis of the triple helix, which can be seen at the bottom of Figure 2.2B [8]. The single residue offset is crucial for formation of stabilising hydrogen bonds involving Gly residues [8]. We conventionally refer to the most N-terminal α-chain as trailing, the most C-terminal α-chain as leading and the remaining α-chain as middle.

We finish the description of collagen molecular structure with some helpful nomenclature that we will often use. "Collagen" refers broadly to a family of proteins, with almost 30 distinct branches, which are commonly referred to as collagen types [10]. In all cases that will be of interest to us in this thesis, a collagen of a given type can be identified using the following nomenclature

$$\alpha_a(\mathcal{X})\alpha_b(\mathcal{X})\alpha_c(\mathcal{X}), \tag{2.1}$$

where the subscripts a, b, c are integers denoting the different α-chains and $\mathcal{X}$ is a Roman numeral that identifies the collagen type. Across the 28 distinct collagen types that have been identified thus far, the number of α-chains can vary significantly from just one in the case of type II and III collagens to six in collagen type IV [14]. Altogether, 44 distinct α-chains are known to be encoded across the collagen family [14]. The order of the α-chains is important, as it affects the 3-dimensional residue organisation. Notation in equation (2.1) specifies the α-chains from left to right as trailing, middle and leading respectively. Whenever two α-chains separated by 1 residue stagger are identical, without loss of meaning, we will use a compressed notation. For example, we will write $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ to mean $\alpha_2(\mathrm{I})\alpha_1(\mathrm{I})\alpha_1(\mathrm{I})$.

Having familiarised ourselves with the molecular structure of collagen, we proceed with discussion of molecular coarse-graining approaches. The most common coarse-grained representation of the collagen triple helix, that has been in use at least since the early 1970's [53], is the so-called "linear" collagen molecule, which is shown in Figure 2.3. We will base our description of the linear triple helix on the recent work of [95]. Residues located at approximately the same positions along the long axis of the triple helix are

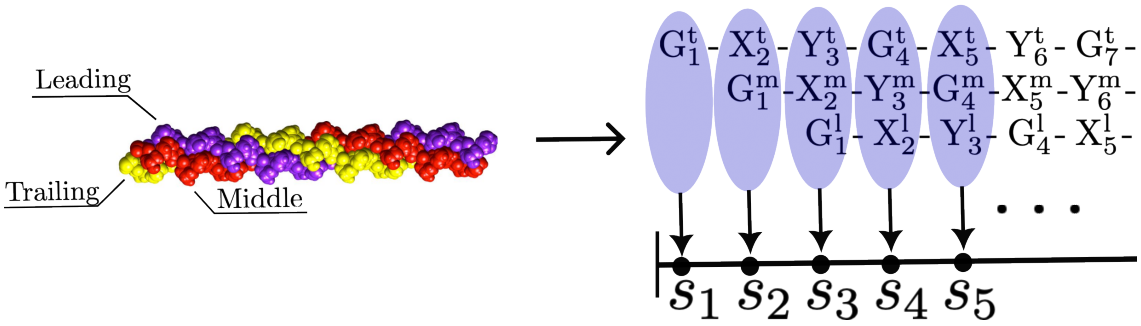


Figure 2.3: Linear coarse-graining approach to collagen molecular structure. Image adapted from [8].

grouped into segments s_i. An interaction between two such segments is then given by

$$\varepsilon(s_i, s_j) = \frac{1}{4} \sum_{\substack{\mathrm{R}_i \in s_i, \\ \mathrm{R}_j \in s_j}} \varepsilon(\mathrm{R}_i, \mathrm{R}_j), \tag{2.2}$$

where $\varepsilon(\mathrm{R}_i, \mathrm{R}_j)$ is the pairwise interaction energy between residues R_i and R_j and the summation is taken over the 4 possible pairs of residues from segments s_i and s_j (recall that Gly residues are buried inside the triple helix and are not available for interaction). The pairwise molecular interaction between two linear collagen triple helices is then calculated by summing up all pairwise segment interactions in a given geometrical arrangement.

The linear coarse-graining approach in Figure 2.3 imposes the assumption that collagen molecules are completely unconstrained when it comes to rotational degrees of freedom. In that instance, every pairwise residue interaction in equation (2.2) is equally likely to occur and thus the equal weighting of every $\varepsilon(\mathrm{R}_i, \mathrm{R}_j)$ term is indeed valid. Questioning the validity of the linear model for the collagen molecular structure was the starting point for much of the work that will be presented in this thesis.

Chapter 2 establishes one of the key ideas that will reappear throughout the other chapters of this thesis. We introduce a novel approach to calculating pairwise interactions between collagen molecules, which incorporates a coarse-grained 3-dimensional molecular structure of the collagen triple helix. We will show that the information encoded in the three-dimensional interaction potential between two collagen molecules can be represented in the form of 28 one-dimensional interaction potentials. These 28 interaction potentials arise due to the helical spatial organisation of the X and Y residues on the molecular surface of collagen. As we shall see in this chapter, our approach not only makes the problem of calculating the three-dimensional pairwise interaction potentials computationally tractable, but also explains for the first time, the key structural features of collagen aggregates, which most importantly cannot be understood using models that utilise linear coarse-graining for the collagen molecular structure.

2.2 INTRODUCTION

Collagen is by far the most abundant protein in the extracellular matrix, connective tissues, skin, and bones [10, 101]. 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 mineralisation to invasions of cancer cells [101, 62] and has even been linked to the Cambrian explosion of multicellular life [108, 38]. Since collagen is essential for maintaining tissue structure and function, it is a key focus in regenerative medicine, wound healing, orthopaedics, dermatology, and cardiovascular health. Recent advances in biochemical engineering have produced the amino acid sequences of thousands of natural collagens [80] and have enabled the design of synthetic collagen mimetic peptides [122]. The biocompatibility of collagen mimetic peptides, their tunable properties, and their potential to replicate natural collagen structures make them indispensable for tissue engineering [119].

At the physical level, the versatility of collagen as a structural protein arises from its propensity to assemble into fibrils, bundles of fibrils, and intricate hierarchical fibrillar matrices. Molecular collagen, the smallest unit in the fibrillar hierarchy, is a semi-flexible molecule approximately $L \approx 300\,\text{nm}$ long [100] and 1.5 nm in diameter. It is made up of three left-handed helical polypeptide strands (α-chains) supercoiled in a right-handed triple helix, see Figure 2.5A and Table 2.2. In each strand, about 1000 amino acids are arranged in a sequence with the regular sequence 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. Collagen is classified into nearly thirty distinct types, with each type varying in the amino acid composition and the hierarchical structures that it forms [10]. In this work, we focus primarily on the supramolecular structures formed by fibrillar collagens, of which type I collagen is the most abundant.

Collagen readily assembles *in vivo* and *in vitro* in fibrils, with typical diameters between 10 nm and 100 nm and lengths that are orders of magnitude larger than their diameters [99]. One salient feature of collagen fibrils is the periodic axial density modulations, which appear as regular alternating light and dark bands with period $\text{D} \approx 67\,\text{nm}$ in negatively stained TEM samples, see Figure 2.4A. The period value, D, is highly conserved across different collagen types, although notable variations occasionally occur, even within the same tissue [36, 10]. The regularity of the banding pattern is crucial for the mechanical strength of fibrils and collagen-rich tissues [26].

A simplified and widely used explanation for the banding pattern, known as the Hodge-Petruska scheme, was proposed in 1964 (see Figure 2.4B) [90]. 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

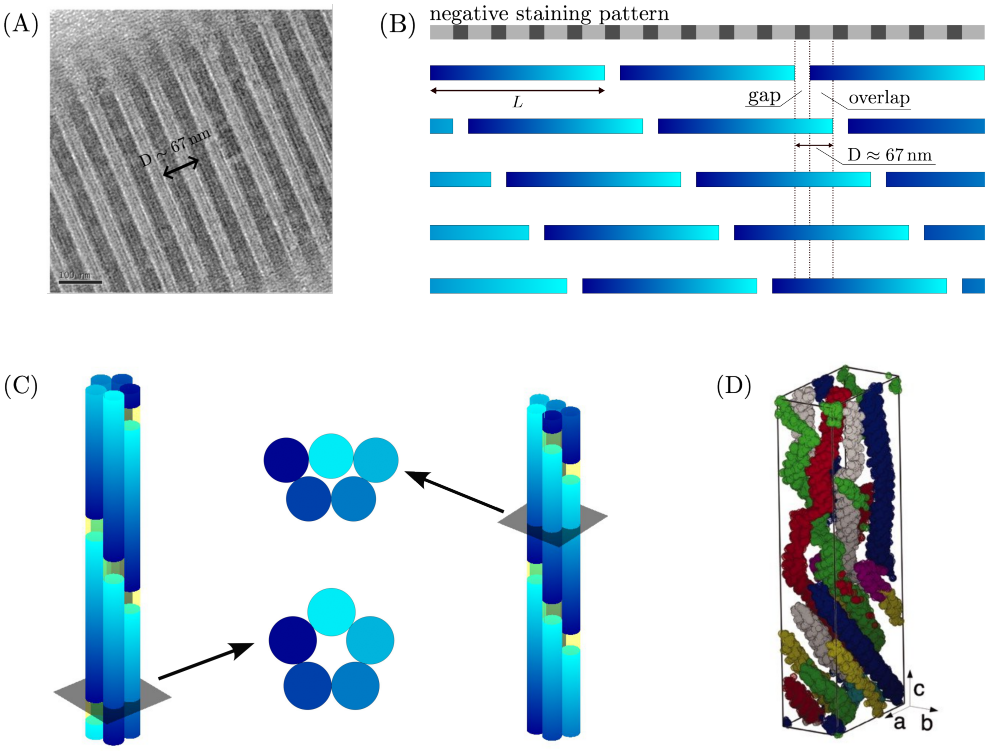


Figure 2.4: (A). TEM image of a negatively stained collagen fibril showing the D-banding pattern (image obtained from [118]). (B). 2D representation of a collagen microfibril according to the Hodge-Petruska scheme. (C). Schematic representations of the 3D microfibril models. Gap regions are highlighted in yellow. (LEFT) Smith microfibril. (RIGHT) Compressed microfibril. (D). Orgel model of the collagen unit cell (obtained from [82]).

are composed of units containing N = 5 collagen molecules. Since $L \approx 4.46$D, five collagen 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 Figure 2.4C) [104]. To further reconcile this model with the quasi-hexagonal lateral packing of individual collagen molecules observed in experiments, the compressed Smith's microfibril model was proposed (see Figure 2.4C) [109].

Isolated collagen microfibrils are yet to be observed experimentally, however evidence from structural and mechanical studies of collagen fibrils provides evidence supporting their existence [83, 82, 43, 121]. 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

Table 2.2: Different types of chirality in molecular collagen and its aggregates.

[a]Chirality Type	Handedness	Studied	[b]Mechanism
Residue Stereoisomerism	L/D Enantiomers [78]	[123]	[78]
α-Chain Chirality	Left-handed [10]	[1]	[28]
α-Chain Supercoil	Right-handed [10]	[10]	[79, 10]
Molecular Supercoil	Right-handed [82]	[20, 19]	unknown
[c]Microfibril Chirality	unknown	unknown	unknown

a The first three rows separate out the chirality types associated with a single collagen triple helix. The last two rows identify the chirality types that emerge in collagen aggregates.
b "Mechanism" is understood to mean a physical process that is hypothesised to give rise to the chirality itself as well as the associated handedness.
c We define microfibril chirality as the handedness of the helical path connecting adjacent gap regions in a collagen microfibril - see Figure 2.5E.

triple helices with a right-handed molecular supercoil are arranged to form a microfibril that interdigitates with neighbouring microfibrils, see Figure 2.4D. This interdigitation establishes the crystallographic superlattice, which is formed of quasi-hexagonally packed collagen molecules. It should be noted that the aforementioned crystalline ordering is not representative of the lateral molecular organisation within the entire fibril, which is characterised by a significant amount of positional disorder [55].

Despite the experimental progress, three distinct aspects of microfibrillar structure lack solid theoretical foundations. (i). Chirality, in its various manifestations is known to play a crucial role in collagen aggregation - see Table 2.2. At the single-molecular level, the chirality of the triple helix as well as the residue stereoisomerism have been widely studied and the physical mechanisms underlying the emergence of chirality have been extensively discussed [123, 78, 10, 79, 28]. At the level of collagen aggregates, the molecular supercoil has been suggested to play a crucial role in determination of the fibril diameter distribution [20, 19]. However, the physical mechanism that leads to the emergence of the molecular supercoil in collagen aggregates as well as selectivity for right-handedness remains unknown. (ii). D-banding and the underlying sequence-structure relationship has been characterised in a number of theoretical studies [53, 110, 92, 52, 25, 95]. It is well-known that the pairwise free energy of interaction between collagen molecules is minimised whenever two triple helices are either in-register or axially staggered by an integer multiple of the D-banding lengthscale [53, 110, 92, 52]. It remains to be understood why interactions between collagen triple helices lead to aggregation of collagen molecules in a D-staggered arrangement as opposed to an in-register arrangement [120]. (iii). A pentameric microfibril provides agreement with existing data from X-ray scattering studies of the crystalline regions within collagen fibrils as well as studies of fibril mechanical properties [104, 109,

82, 43, 121]. However, it has not been shown how pairwise interactions between collagen triple helices lead to the energetic selection of microfibrils comprised of N = 5 collagen molecules (as opposed to other values of N) in 3-dimensional space.

An important model predicting the axial stagger between pairs of collagen molecules comprising the microfibril using residue sequence data was recently suggested by Puszkarska *et al.* [95]. 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 Mijazawa-Jernigan approximation does not explicitly account for the role of water in collagen-collagen interactions, which has been shown previously to play an important role [9]. In spite of this, this theoretical approach has proven successful in predicting D-banding across a broad range of fibrillar collagens. The inter-residue contacts were determined based on the spatial proximity of residues, which was directly related to the sequence proximity: two residues close in a sequence were assumed to be 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 molecular supercoil and its handedness in collagen aggregates. Furthermore, the authors did not investigate the coupling between pairwise triple helix interactions and the 3-dimensional molecular packing [95].

In this article, we extend the approach of [95], combining empirical studies with theoretical arguments to quantify the interaction potential between pairs of parallel collagen 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 the triple helix. We demonstrate that the elasticity of the triple helix plays a crucial role in propagating this chirality to the level of the microfibril, resulting in the right-handed molecular supercoil of individual collagen molecules. Furthermore, we attribute the energetic selection of N = 5 in a microfibril to the geometric arrangement of residues. In agreement with the knowledge in the field [53, 110, 92, 52], we find that the optimal axial stagger, Δz, can assume different values corresponding to distinct local free energy minima, most notably $\Delta z \approx 0$ and $\Delta z = k\mathrm{D}$, where $k = 1, \ldots, 4$. 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* [50, 54]. We show that while the minimum at $\Delta z \approx 0$ is generally stronger than those at $\Delta z = k\mathrm{D}$, it is sensitive to perturbations in the residue-residue

interaction energies. Consequently, the axial staggers $\Delta z = k\mathrm{D}$ emerge as robust global optima under most conditions. Finally, we define microfibril chirality, which we believe to be a novel type of emergent chirality that is distinct from the molecular supercoil of the collagen molecules comprising the microfibril.

To compare our predictions with the available experimental data, we analyse amino acid sequences of more than 1000 known fibril-forming collagens of mammalian species across all known types of fibrillar collagens. 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 analysed sequences of heterotrimeric type I collagen result in D-banding, in agreement with the general knowledge in the field [107]. This agreement validates our methods and lends credibility to our predictions for other, less well-studied fibrillar collagen types.

2.3 RESULTS

2.3.1 Chiral Interactions and Helical Strip Organisation in Collagen

The observed molecular supercoil of collagen 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 collagen 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 [7]. Currently, no consensus has been reached on the preferred structural model for describing the collagen triple helix [7, 8]. Rather than making a subjective choice between different structural models, we use the best available experimental data on the structure of the collagen triple helix. In particular, we use a statistically-derived parametrisation of the triple helix based on the analysis of multiple high-resolution structures of shorter peptides modelling sections of the triple helix [96]. The statistical model accounts for differences in the imino acid content, resulting in two distinct triple helix parametrisations: 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 [7]. We note that in absence of high-resolution structures of the native collagen molecule, it is currently not possible to directly validate experimentally the accuracy of the aforementioned statistical model when applied to the entire triple helix.

We now demonstrate that each of the parametrisations gives rise to a helical arrangement of the outward-facing residues. Figure 2.5A shows the positions of the residues projected

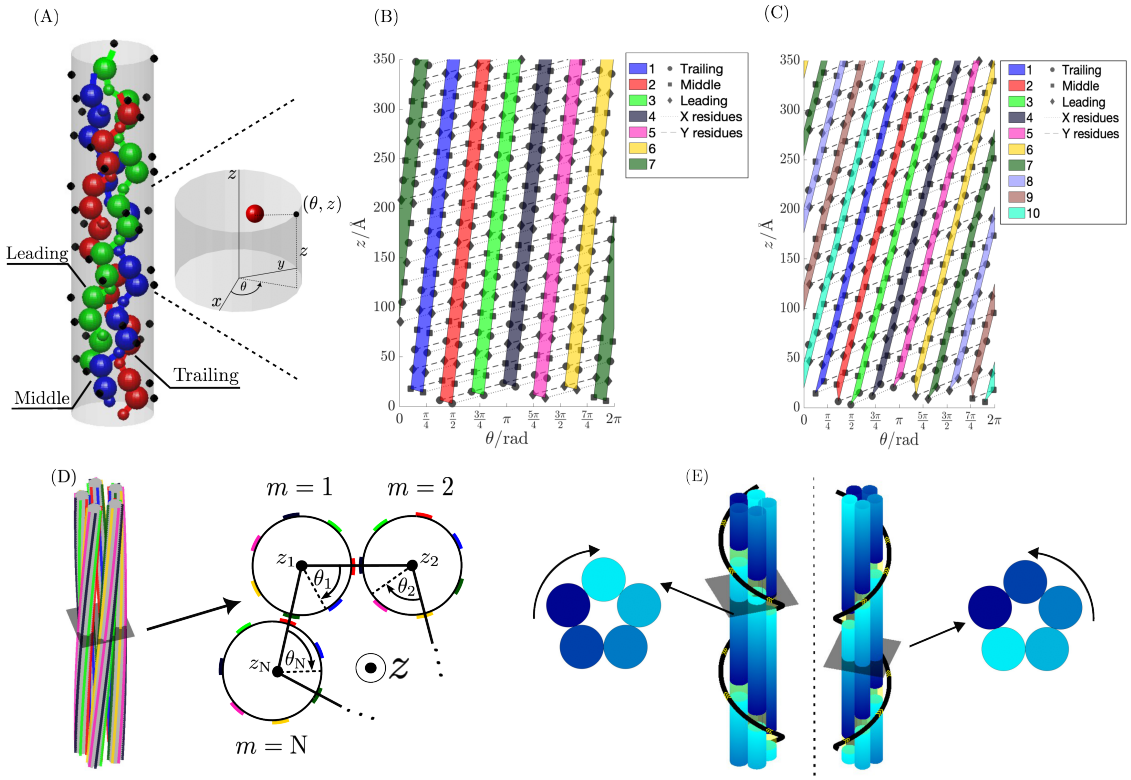


Figure 2.5: (A). Segment of the collagen triple helix bounded by a cylindrical surface onto which coordinates of each C_α residue atom are projected. The C_α atom positions are obtained using two statistically derived parametrisations based on analysis of (B) Pro-rich and (C) Pro-poor model peptides. 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 collagen 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). (E). Left-handed and right-handed microfibrils. Microfibril chirality is defined as the handedness of the helical path linking adjacent gaps (highlighted in yellow).

on the cylindrical surface of a molecule unwrapped in 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 2.5B and 2.5C. For the Pro-rich parametrisation, there are seven helical strips, each separated azimuthally by $2\pi/7\,\text{rad} \approx 51.4°$, with a pitch of approximately 200 nm. For the Pro-poor parametrisation, there are ten helical strips, with an azimuthal separation of $2\pi/10\,\text{rad} = 36°$ 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 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 collagen molecule varies as a function of the amino acid composition [84]. It is, clearly, chiral. This chirality generates a chiral interaction potential between pairs of parallel collagen molecules, leading to torques that can bend and twist them. We assume that the outward-facing residues of a collagen molecule can be represented as a superposition of helical strip families with varying pitches [96, 84], such as the Pro-rich and Pro-poor families described in [96]. When two parallel molecules are close enough for their outward-facing residues to interact, strong interactions between the residues comprising the helical strips generate torques that bend and twist the molecules to minimise the total 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 [79, 68], though collagen differs in having multiple helical strip families. For two triple helices to supercoil, the energy gained from residue interactions must outweigh the energy cost of elastic deformation. 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 2.6.1). 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 equation (2.5) in Materials and Methods, consistent with experimental observations in bone and tendon [82, 6, 98]. We have thus provided an elasticity-driven mechanism for the observed molecular supercoiling in collagen aggregates. This further emphasises the crucial importance of chain flexibility in collagen self-assembly, as has been

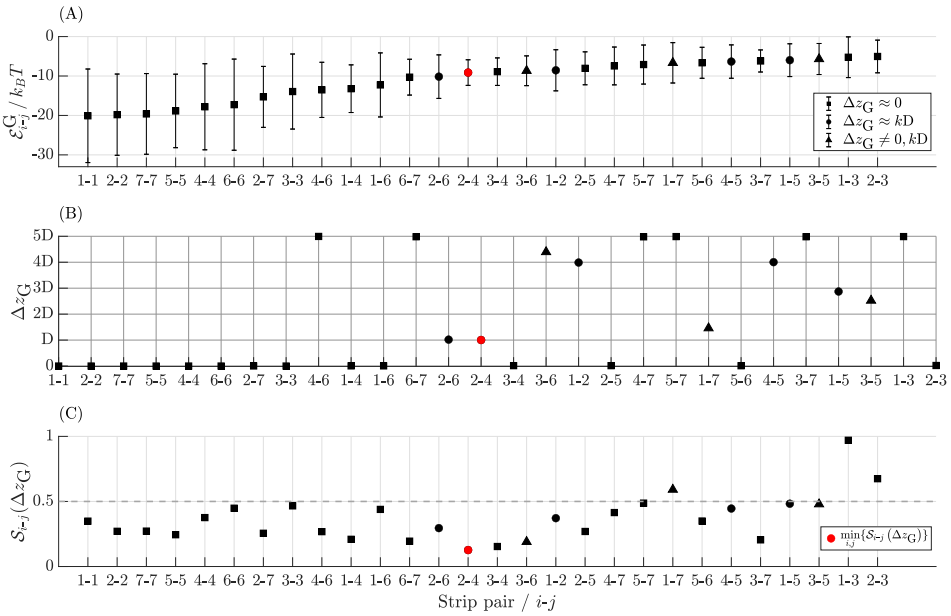


Figure 2.6: Global minima of the axially periodic strip-strip interaction energies and their sensitivity to perturbations in contact potential values. The presented results are obtained numerically for $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen. (A). Average values of the interaction energies at their global minima due to uniform uncorrelated perturbations in contact potential values. Error bars indicate the standard deviation in the perturbation-added energy values. (B). Locations of the global energy minima, Δz_{G}. (C) Perturbation 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 $\Delta z \approx \mathrm{D}$. In general, the former are stronger but more sensitive to perturbations than the latter. See Materials and Methods 2.6.3 for details of the procedures.

suggested by a number of other theoretical studies [20, 19, 58].

2.3.2 Energetics of Strip-Strip Interactions

Based on this, we hypothesise 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 $\mathrm{E}_{i\text{-}j}(\Delta z)$, 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 2.6.2. MJCP represent a first order approximation to the free energy of interaction between pairs of proteinogenic residues [95]. The entropic contributions arising from residue side-chain orientations are thus accounted by MJCP. We note that MJCP used in this study do not explicitly incorporate collagen-water interactions,

which have been shown to be important for collagen aggregation [9, 45]. Nevertheless, coarse-grained computational approaches utilising just protein-protein interactions have been successful in describing pairwise triple helix interactions as well as aggregation of collagen-like peptides [95, 45, 42, 58].

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: $\mathrm{E}^{p}_{i\text{-}j}(\Delta z) = \mathrm{E}_{i\text{-}j}(\Delta z) + \mathrm{E}_{i\text{-}j}(\Delta z - g - L)$, where $0 \leq \Delta z < L+g$. For consistency with the standard definitions, we define D in terms of g such that $5\mathrm{D} = L + g$[1]. While this definition anticipates the value of D, it does not constrain it. Deferring the discussion of how D is determined in simulations to Appendix 2.7.2.1, we find that the values of D that yield perfectly-staggered microfibrils fall within a narrow range, approximately $\mathrm{D} \approx (67 \pm 2)\,\mathrm{nm}$, see Figure 2.10. Therefore, when we next discuss the minima of $\mathrm{E}^{p}_{i\text{-}j}(\Delta z)$ over Δz, we will use the corresponding values of D as relevant lengthscales.

Figure 2.6B 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 $\Delta z \approx k\mathrm{D}$. 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 $\Delta z \approx 0$ [53, 110, 92, 52, 95]. Our results show that most, but not all, global minima fall into this class, see Figures 2.6A, B. This observation warrants further examination, 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 must be highly robust toward environmental variability.

To account for it, we add uniform, uncorrelated perturbations to the MJCP values and analyse the sensitivity of the intermolecular interactions and the emergent microfibrillar structures to these perturbations. We define the sensitivity of the pairwise strip-strip

[1] In chapter 4, we will derive this relationship for an arbitrary axial period.

interactions as the variance of the noisy potentials $\mathcal{E}^p_{i\text{-}j}(\Delta z)$ normalised by their mean, i.e.

$$\mathcal{S}_{i\text{-}j}(\Delta z) = \frac{\mathrm{Var}\left\{\mathcal{E}^p_{i\text{-}j}(\Delta z)\right\}}{\left|\mu\left\{\mathcal{E}^p_{i\text{-}j}(\Delta z)\right\}\right|^2}, \tag{2.3}$$

where $\mathrm{Var}\{\cdots\}$ and $\mu\{\cdots\}$ denote the variance and the mean, respectively. Figure 2.6C shows that, remarkably, the perturbation sensitivity turns out to be the smallest for the interaction potential minima at $\Delta z \approx \mathrm{D}$. In contrast, the energy minima at $\Delta z \approx 0$ are more sensitive to the perturbations.

We use the linear decomposition of $\mathcal{S}_{i\text{-}j}(\Delta z)$ into contributions from interacting pairs of residues (see Materials and Methods 2.6.3) to trace the high perturbation sensitivity of interactions at $\Delta z \approx 0$ to only two interacting pairs of highly abundant residues: Pro-Pro and Pro-Ala, see Figure 2.9. The axial staggers and strip combinations that achieve the strongest intermolecular interactions and minimise the number of interactions between abundant residues, end up being the least sensitive to the perturbations.

In the biological context, the perturbation 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 [10]. 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 perturbation 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.

2.3.3 EMERGENCE OF D-PERIODIC MICROFIBRILS

To determine whether the D-staggered microfibril emerges from the intermolecular interactions and explain why the microfibril is composed of $\mathrm{N} = 5$ molecules, we turn to numerical modelling. To keep the problem tractable, we assume an axially periodic microfibril comprised of aligned collagen 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 centrelines and be shifted along the microfibrillar axis by the distances z_m, see Figure 2.5D. 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 $\mathrm{E_M}$ is then given by the sum of N pairwise molecular interactions (see Materials and Methods 2.6.4). The equilibrium microfibrillar structure results from minimising the free energy of the system. As discussed previously in subsection 2.3.2, the entropic contributions due to residue side-chain orientations are taken into account by MJCP. The remaining entropic contributions associated with the axial

and azimuthal degrees of freedom of the triple helix are sub-extensive in microfibril length. Thus, they can be ignored in the present context and we can simply minimise E_M [95].

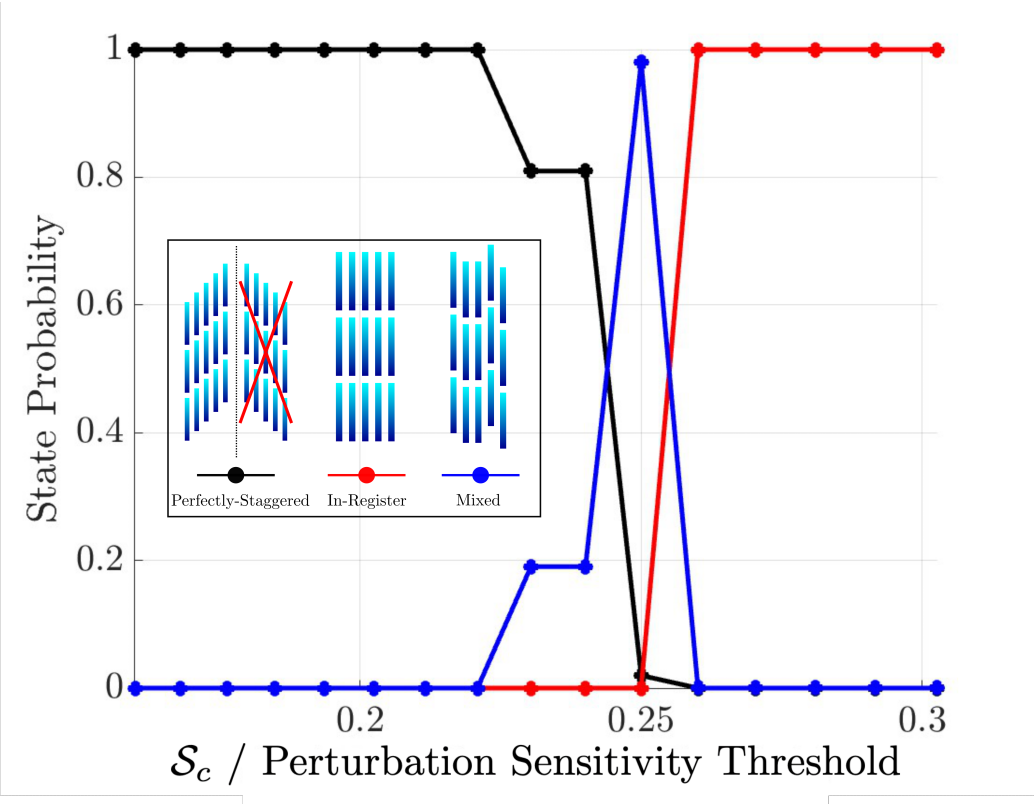


Figure 2.7: Equilibrium probabilities of different microfibrillar states as a function of perturbation sensitivity threshold.

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 $\Delta z \approx k\text{D}$ for $k = 1, 2, 3$ or 4, (2) in-register configurations with $\Delta z \approx 0$ for all molecules, and (3) mixed configurations, see Figure 2.7. When the sensitivity to the contact potential perturbations 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 perturbation sensitivity threshold is lowered, one of the perfectly-staggered configurations emerges at equilibrium, see Figure 2.7 and Figure 2.11. We remark that for certain values of $\mathcal{S}_c$, it is also possible to observe mixed microfibrillar conformations at equilibrium. This observation foreshadows the content of chapter 3, in which we will see that under specific ionic conditions, collagen preferentially aggregates into microfibrils with no axial banding pattern. Such microfibrils, we will argue, correspond to the disordered fibrils, which we first saw in Figure 1.2G.

Perfectly-staggered microfibrils form enantiomeric pairs (e.g., $\Delta z = \mathrm{D}$ and $\Delta z = 4\mathrm{D}$, or $\Delta z = 2\mathrm{D}$ and $\Delta z = 3\mathrm{D}$), as illustrated in Figure 2.5E for $\Delta z = \mathrm{D}$ and $\Delta z = 4\mathrm{D}$. While the handedness of the molecular supercoil of individual triple helices in a microfibril is set by the chirality of interacting residue strips, the microfibril chirality emerges from the collective interaction between the molecules. It corresponds to the handedness of the helical path linking adjacent gaps, see Figure 2.5E. To the authors' knowledge, this distinction between molecular and microfibrillar chirality has not been explicitly made in previous work.

In our model, enantiomeric fibrils in general have differing energies, 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 2.6.2.

2.3.4 AGGREGATE SIZE SPECIFICITY

Next, we consider why a microfibril is comprised of five molecules. Figure 2.8A shows the global minimum of microfibrillar energy per molecule for varying N. Notably, $\mathrm{N} = 5$ 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

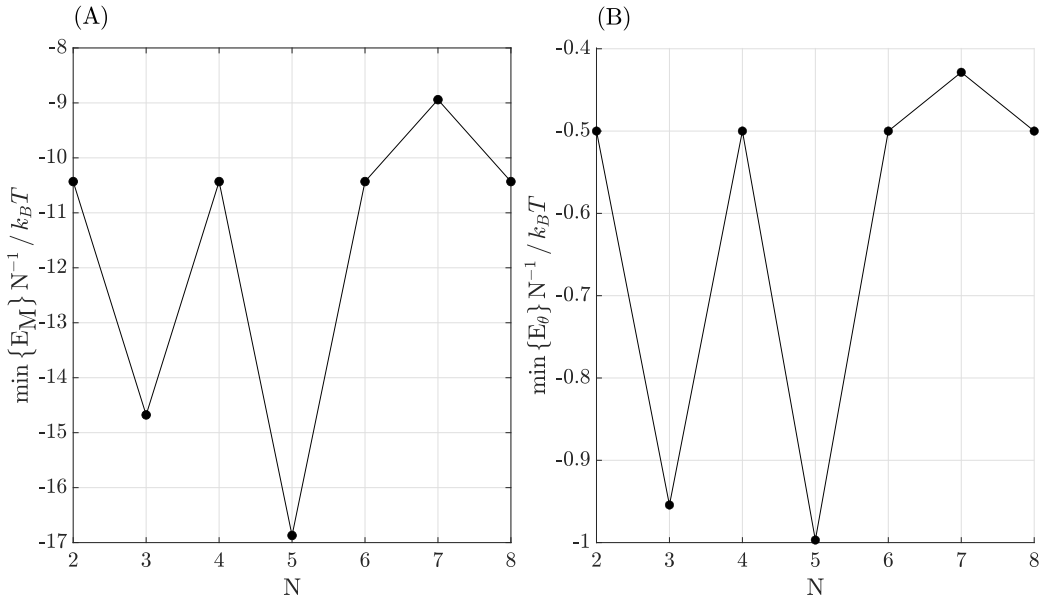


Figure 2.8: Global minimum of the microfibril energy per molecule as a function of aggregate size N in $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen. (A). Microfibril energy with empirically determined axial dependence. (B). Microfibril energy with no axial dependence, see Materials and Methods 2.6.4 for details of construction. Details of the global optimisation procedure can be found in Appendix 2.7.2.

Table 2.3: Minimum difference between the internal angle of N-membered microfibrils and the azimuthal inter-strip spacings.

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

face each another. When molecules are positioned at the vertices of a regular N-gon, the interior polygon angle $v(\mathrm{N})$ should approximate $m\alpha$, where $\alpha = 360°/7$ is the angle between the strips and m is an integer. It is easy to see that $v(5) = 108°$ closely approximates $2\alpha \approx 103°$, see Table 2.3. 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 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 Figure 2.8B 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 collagen residues alone can make a pentamer the preferred microfibrillar configuration, specific residue interactions are essential for stabilising it. In fact, in chapter 4, we will identify the specific interactions that selectively stabilise 5-membered microfibrils with D-banding among a significantly broader class of molecular packing models than considered in this chapter. It is also conceivable that some specific residue sequences might preferentially select for a trimeric microfibril.

2.4 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 arrangement emerges at the level of microfibrils – essential, if experimentally elusive, structures [82]. 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 spatial residue organisation guides the formation of this structure.

Focusing on collagen I, we found that the outward-facing collagen residues are arranged in sevenfold helical strips. This arrangement emerges from the supercoiling of collagen α-chains and is reminiscent of the hydrophobic strip that emerges in coiled coils due to a regular spatial arrangement of heptad repeats [71, 79]. 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 collagen triple helix.

We predict that the strip chirality is transmitted to the level of molecular supercoiling through the torques that arise from pairwise intermolecular interactions. We show that the resulting equilibrium helical angle ϕ^* is described by equation (2.5), which was first empirically obtained by Fraser and MacRae in the context of coiled coils [41]. For collagen type I, it predicts a right-handed molecular supercoil with the helical angle of approximately $5°$, in agreement with experimental observations [82]. Molecular supercoiling is, however, not uniform across different types of collagen aggregates. Heterotypic corneal collagen fibrils comprised of type I and type V collagens, are known to display a molecular supercoiling angle of $\approx 15\,°$ [98]. We remark that this helical angle can be explained as arising from strong chiral interactions between collagen triple helices that are characterised by the Pro-poor, rather than the Pro-rich parametrisation.

The role of molecular supercoil in collagen fibril self-assembly has been previously investigated using mesoscale models [18, 20]. Phenomenological liquid crystal models were used to describe the collagen fibril, wherein the molecular supercoil was accounted for by inclusion of a double-twist director field. Through minimisation of the resulting Frank free energy, the authors then obtained the fibril radius and fibril surface molecular supercoil angle at thermodynamic equilibrium. These mesoscale models of collagen fibril growth point to two distinct equilibrium modes of molecular supercoil angle dependence on the fibril radius [20]. The first equilibrium configuration is characterised by a constant positive growth rate of molecular supercoiling angle as a function of fibril radius. The resulting fibril surface molecular supercoil angle is $\approx 5°$ which is in agreement with experimental observations in tendon fibrils. The second equilibrium configuration exhibits a non-linear positive growth rate and results in collagen fibrils with a fibril surface molecular supercoiling angle of $\approx 15°$. This configuration is in good agreement with experimental characterisation of corneal fibrils.

The aforementioned results from mesoscale models of collagen fibril aggregation invite further investigation into the microscopic origins of the polymorphism associated with the growth rate of molecular supercoiling angle in collagen fibrils. Previously suggested microscopic models of collagen aggregation have not provided physical mechanisms that explain the emergence of molecular supercoiling in collagen fibrils [53, 110, 92, 52, 95].

Most recent microscopic models of collagen aggregation utilise a linear approximation for the collagen triple helix structure, wherein the detailed 3-dimensional conformation of the collagen triple helix is ignored and only the information about the sequential positions of residues is retained [95]. This approximation precludes such models from predicting 3-dimensional structural features of the collagen triple helix, such as its molecular supercoil. The approach developed in the current study fills this gap in the literature, thus offering an opportunity to bridge the microscopic mechanisms and interactions that determine collagen molecular supercoiling with mesoscopic phenomena such as the dependency of molecular supercoil on fibril radius. In particular, future research could elucidate the connection between the chiral 3-dimensional residue organisations of Pro-rich/Pro-poor regions of the triple helix and the constant/non-linear radial growth rates of molecular supercoil angle inside collagen fibrils that is predicted by mesoscopic models.

We have identified an emergent type of chirality - the microfibril chirality, which is different from the molecular supercoiling chirality. It can be defined as the chirality of the helical path connecting adjacent gap regions. Microfibril chirality emerges from the collective interaction of collagen molecules and can be sensitive to the collagen sequence details. In our model, accounting for the 3-dimensional residue organisation enables breaking the mirror symmetry of pairwise triple helix interactions, leading to energetic selection of microfibrils with a preferred handedness. This conclusion is particularly significant, given the critical role of gap regions in directing mineralisation during bone formation: experiments have shown that, at the smallest lengthscales, needle-shaped mineral crystals are arranged into left-handed helices [101]. Our results thus provide motivation for further study of the role of microfibril chirality in determining the helical architecture of the mineral phase in bone as well as the impact of collagen residue sequence variability and external conditions on bone mineralisation.

The stagger distance, $\mathrm{D} \approx 67\,\mathrm{nm}$, is encoded in local energy minima for the strip-strip interaction potentials, which occur at relative molecular stagger $\Delta z \approx k\mathrm{D}, \quad k = 1, \ldots, 4$ and at $\Delta z \approx 0$. While the minima at $\Delta z \approx 0$ are typically the strongest, they are sensitive to perturbations in the residue-residue interaction energies. This sensitivity is transmitted to the aggregate level. Upon introducing a perturbation sensitivity threshold that filters out perturbation-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, perturbation sensitivity points to a sensitive control that may be exerted on aggregates by local biochemical environments. For example, transitions between D-banded fibrils and SLS aggregates (corresponding to the dominant minima at $\Delta z \approx 0$) can be induced by introducing charged molecular species such as diazo dyes, which have been suggested to modify the interactions between charged residues [50]. Residue interactions may also be affected through post-translational enzymatic modification, such

as hydroxylation of Pro and Lys residues. Post-translational hydroxylation of Pro residues is known to significantly affect the temperature and ionic conditions required for D-banded fibril formation [88]. This observation aligns with our perturbation 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 [90] as well as Smith [104], that to conform to the regular axial D-banded pattern, collagen aggregates must be composed of pentamers. However, 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 collagen 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 2.4. Even though the mammalian collagen sequences are highly homologous, it is crucial to ascertain that the predictions of our model are not significantly affected by small changes in the amino acid sequences. High resolution *in situ* studies of microfibrillar structures have only been performed for $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ collagen originating from rat tail tendon [82]. 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. It should be noted that explicit measurements of D-banding have only been performed in a handful of commonly studied mammalian species, referenced in Table 2.4.

Our analysis predicts that for all sequences of $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ collagen examined, a perfectly-staggered microfibril is the most stable aggregate below some perturbation sensitivity threshold $\mathcal{S}_c$. This finding agrees with experimental observations of D-banding in collagens of this type. Furthermore, we find that for the majority of the sequences, a left-handed microfibril chirality is energetically selected (see Figure 2.11). A sceptical reader may object by noting that for a significant minority of sequences the microfibril chirality is right-handed, which does not appear to be in agreement with the left-handed chirality of the mineral phase in bone. To that end, we remark that in chapter 4, we will demonstrate that modifying the strength of charged residue interactions leads to a change

Table 2.4: Prediction of perfectly-staggered microfibrils in mammalian species for different collagen types.

Type	D-Banded Fibrils	№ Species	Predicted PS Microfibrils[a] (%)
$\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$	✓[6]	176	100
$[\alpha_1(\mathrm{I})]_3$	✓[72]	186	25.27
$[\alpha_2(\mathrm{I})]_3$	unknown[b]	197	94.92
$[\alpha_1(\mathrm{II})]_3$	✓[2]	191	100
$[\alpha_1(\mathrm{III})]_3$	✓[4, 17][c]	185	3.78
$[\alpha_1(\mathrm{V})]_3$	✗ [21]	167	86.83
$\alpha_1(\mathrm{V})\alpha_2(\mathrm{V})\alpha_1(\mathrm{V})$	✓[76]	151	78.81
$\alpha_1(\mathrm{V})\alpha_3(\mathrm{V})\alpha_2(\mathrm{V})$	✓[76]	124	98.39
$\alpha_3(\mathrm{XI})\alpha_2(\mathrm{XI})\alpha_1(\mathrm{XI})$	✓[49]	148	99.32
$[\alpha_1(\mathrm{XXIV})]_3$	unknown[d]	163	87.12
$[\alpha_1(\mathrm{XXVII}]_3$	✗ [93][d]	163	60.36

[a] The model is said to predict a perfectly-staggered (PS) microfibril, if there exists some value of $\mathcal{S}_c$, below which the perfectly-staggered state probability is unity.
[b] An α_2(I) homotrimer is not observed *in-vivo*. This homotrimer has been however observed in *in-vitro* refolding experiments [66]. Propensity of $[\alpha_2(\mathrm{I})]_3$ to undergo self-assembly into fibrils has not been investigated.
[c] The D-banding lengthscale 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.

in the preferred handedness of a microfibril from right- to left-handed and vice versa. As such, specific environmental or biochemical factors can change the energetically preferred microfibril chirality in our model. Similarly, our model indicates that over 99% of the tested sequences for collagens $[\alpha_1(\mathrm{II})]_3$, $\alpha_1(\mathrm{V})\alpha_3(\mathrm{V})\alpha_2(\mathrm{V})$ and $\alpha_3(\mathrm{XI})\alpha_2(\mathrm{XI})\alpha_1(\mathrm{XI})$ favour the formation of perfectly-staggered microfibrils.

However, for the homotrimeric collagens $[\alpha_1(\mathrm{I})]_3$ and $[\alpha_1(\mathrm{III})]_3$ our model predicts the formation of perfectly-staggered microfibrils only in a small fraction of sequences. It also predicts that the majority of the analysed homotrimeric collagens - $[\alpha_1(\mathrm{V})]_3$ and $[\alpha_1(\mathrm{XXVII})]_3$ - result in periodic microfibrils, which disagrees with the lack of periodic microfibril formation observed in some of these collagens, see Table 2.4. 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 [95]. This presents the 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. More broadly, this suggests fundamental differences in the self-assembly of heterotrimeric and homotrimeric collagens. Such differences may underlie serious medical conditions, such as the Ehlers-Danlos syndrome, which is characterised by joint hypermobility and cardiac valve abnormalities [40]. This condition arises from deleterious mutations in the COL1A2 gene, resulting in the production of homotrimeric type I collagen instead of the normal heterotrimeric form. The consequent misassembly alters the mechanical properties of collagen-rich tissues, leading to the observed pathology. Identifying the fundamental physical interactions that give rise to these mechanical alterations, such as those between Pro-poor strips, presents an exciting opportunity for further theoretical and experimental study.

In developmental collagens XXIV and XXVII, it is important to acknowledge the potential impact of the model's assumptions on the predictions of D-banding. For example, 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 [7]. 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 [56]. The details of the structural impact of sequence interruptions present in collagens XXIV and XXVII remain to be elucidated [7, 13, 61]. This calls for future research into the effect of sequence interruptions in developmental collagens on spatial residue organisation and microfibril self-assembly.

Collectively, 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. In addition to the polymorphisms in the molecular supercoiling angle [98], 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 as well as significantly greater than D-banding in FLS collagens [112, 4, 120]. Recent advances in the synthesis of collagen-mimetic peptides also suggest the possibility of microfibril aggregate sizes differing from $N = 5$ [24]. Crucially, we have demonstrated that the chiral 3-dimensional spatial residue organisation is central to determining the structural features associated with supramolecular aggregates of collagen. Thus, continued study of the spatial residue organisation within the triple helix is integral to the understanding of the various structural polymorphisms of collagen. 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 supramolecular collagen structures.

2.5 CONCLUSION

This study identifies chiral intermolecular interactions, rooted in the 3-dimensional spatial arrangement of outward-facing residues, as a fundamental mechanism driving the self-assembly of collagen into its characteristic D-periodic, pentameric microfibrils comprised of triple helices with a right-handed molecular supercoil. By integrating residue-level sequence data with a physically motivated interaction model, we demonstrate that the chiral spatial residue organisation determines the microfibrillar architecture. The right-handed molecular supercoil and staggered configuration predicted by our model, are not only energetically favoured but also robust to biochemical perturbations across diverse mammalian collagen sequences. These insights bridge the molecular and 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.

2.6 MATERIALS AND METHODS

2.6.1 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 each other, incurring the elastic energy

$$\mathrm{E_{el}} = \int_0^L \left(\frac{B}{2} \frac{\sin^4 \phi}{R^2} + \frac{C}{2} \left(\frac{\sin 2\phi}{2R} - \chi \frac{2\pi}{h} \right)^2 \right) ds, \tag{2.4}$$

where 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 [79]. Taking

R to be fixed, for a sufficiently small ratio $\epsilon = 2\pi R/h \ll 1$ and a finite ratio B/C (see Appendix 2.7.3 for the derivation), leads to the equilibrium helical angle ϕ^* given by the asymptotic expression

$$\phi^* \approx \chi\epsilon = \chi\frac{2\pi R}{h}. \tag{2.5}$$

This is the classical Fraser and MacRae formula widely used to analyse the triple-helical structure of collagen [41]. Neukirch *et al.* later extended it to include coiled coil proteins under non-zero external forces and torques [79], 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,

$$\frac{B}{C} \ll \epsilon^{-2}, \tag{2.6}$$

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

$$\mathrm{E}^*_{\mathrm{el}} \approx 8\pi^4\frac{BR^2L}{h^4} = 8\pi^4\frac{\xi_b R^2 L}{h^4}k_BT, \tag{2.7}$$

where ξ_b is the bending persistence length, k_B is the Boltzmann constant and T is the temperature. Condition (2.6) is easily satisfied in practice. The expression (2.7) shows that the energetic cost of the elastic deformation increases steeply as the helical pitch h decreases.

We estimate the molecular length as $L \sim 300\,\mathrm{nm}$, the microfibril radius as $R \sim 3\,\mathrm{nm}$, and the persistence length as $\xi_b \sim 120\,\mathrm{nm}$ at neutral pH and physiological salt concentration [100]. For the Pro-rich strips with $h \sim 200\,\mathrm{nm}$, the corresponding elastic energy cost is $\mathrm{E}^*_{\mathrm{el}} \sim 0.16\,k_BT$. This value is significantly smaller than the characteristic interaction energy of two D-staggered collagen molecules, approximately $10\,k_BT$ as can be seen in Figure 2.6A. In contrast, for the Pro-poor strips with $h \sim 75\,\mathrm{nm}$, the elastic energy cost is much higher at $\mathrm{E}^*_{\mathrm{el}} \sim 8\,k_BT$, 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 triple helix parametrisation when modelling the effective interaction potential between collagen molecules. The equilibrium coiling angle for the corresponding helical pitch h is estimated as $\phi^* \sim 5°$, which aligns well with the experimental observation in bone and tendon [6, 98].

2.6.2 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 [64]. Denote a pair of interacting strips as i-j, wherein $i, j = 1, 2, \ldots, 7$. Let $\{\mathbf{e}_\rho, \mathbf{e}_\theta, \mathbf{e}_z\}$ be the set of cylindrical basis vectors in the triple helix coordinate system. Let ${}^{\mathrm{q}}\mathbf{x}_j$ to be the position vector of the residues along strip j that are labelled in ascending order of axial coordinate by $\mathrm{q} \in \mathbb{Z}^+$. The staggered coordinates of ${}^{\mathrm{q}}\mathbf{x}_j$ are defined as

$$ {}^{\mathrm{q}}\mathbf{x}_j^{\mathrm{s}}(\Delta z) = {}^{\mathrm{q}}\mathbf{x}_j + 2\pi h^{-1} \left(\Delta z + c_j - c_i\right) \mathbf{e}_\theta + \Delta z \mathbf{e}_z, \tag{2.8} $$

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

$$ \mathrm{E}_{i\text{-}j}(\Delta z) = \sum_{\mathrm{p,q}} \varepsilon_{g(\mathrm{p})g(\mathrm{q})} \left[\Theta\left(r_{\mathrm{pq}}(\Delta z)\right) - \Theta\left(r_{\mathrm{pq}}(\Delta z) - l_c\right)\right], \tag{2.9} $$

where $r_{\mathrm{pq}}(\Delta z) = |{}^{\mathrm{p}}\mathbf{x}_i - {}^{\mathrm{q}}\mathbf{x}_j^{\mathrm{s}}(\Delta z)|$, Θ is the Heaviside step function, l_c is the interaction lengthscale and $g : \mathbb{Z}^+ \to \{1, 2, \ldots, 20\}$ maps the sequential residue position along a strip onto its integer designation.

The matrix $\varepsilon \in \mathbb{R}^{20\times 20}$ represents the free energies of the residue-residue interactions. We follow the method of Puszkarska *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 [60]. We take $l_c = 0.75\,\mathrm{nm}$, which is typically assumed to be the representative lengthscale at which a pair of residues is in contact [95].

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

$$ \mathrm{E}^p_{i\text{-}j}(\Delta z) = \mathrm{E}_{i\text{-}j}(\Delta z) + \mathrm{E}_{j\text{-}i}(\mathcal{T} - \Delta z), \tag{2.10} $$

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

$$ \mathrm{E}^p_{i\text{-}i}(\Delta z) = \mathrm{E}^p_{i\text{-}i}(\mathcal{T} - \Delta z), \tag{2.11} $$

meaning that the functions $\mathrm{E}^p_{i\text{-}i}$ possess a reflection symmetry with respect to $\Delta z = \mathcal{T}/2$ - see Figure 2.12 for visual intuition. Since previous studies [95] 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 $\mathcal{T} - \Delta z$. This property might lead to a degenerate ground

state, precluding formation of a well-defined axially periodic (D-banded) structure. In particular, perfectly-staggered microfibrils with left-handed and right-handed chiralities corresponding to the symmetric values of Δz cannot be differentiated. This symmetry is broken for interactions of different strips:

$$\mathrm{E}^{p}_{i\text{-}j}(\Delta z) \neq \mathrm{E}^{p}_{i\text{-}j}(\mathcal{T} - \Delta z), \quad i \neq j, \tag{2.12}$$

lifting the degeneracy.

2.6.3 PERTURBATION SENSITIVITY OF PAIRWISE INTERACTIONS

To account for uncertainty in the elements of residue interaction matrix ε, consider a perturbation-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 perturbation-added pairwise interaction energy $\mathcal{E}^{p}_{i\text{-}j}$ is then calculated according to equation (2.9) and equation (2.10) using the matrix ε^{*}.

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

$$\mathcal{S}_{i\text{-}j}(\Delta z) = \frac{12^{-1}\,|a-b|^{2} \sum\limits_{1\leq l\leq m\leq 20} N^{2}_{lm}}{\left|\mathrm{E}_{i\text{-}j}(\Delta z) + \frac{1}{2}\,(a+b) \sum\limits_{1\leq l\leq m\leq 20} N_{lm}\right|^{2}}, \tag{2.13}$$

where N_{lm} is the number of interacting residues with integer designations l and m at a given Δz. Importantly, equation (2.13) is a linear combination of the contributions due to pairs of interacting residues proportional to N^{2}_{lm}.

In addition to the analytical expression in equation (2.13), the perturbation sensitivity parameter can be computed numerically. The results presented in Figure 2.6 were performed numerically by constructing 50 perturbation-added interaction energy curves, with perturbations sampled from $U(-0.1\,k_B T, 0.1\,k_B T)$. The value of the perturbation 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 $|\varepsilon|$. Importantly, the relative perturbation sensitivity of the energies and, hence, all of our physical conclusions are independent of the chosen value.

2.6.4 MODEL OF A MICROFIBRIL

We parametrise the azimuthal component of pairwise molecular energy by

$$\Phi(\theta_m) = \left[1 + \exp\left(a\left|\sin\left(\frac{\pi(\theta_m - \theta_0)}{\theta_{\mathrm{d}}}\right)\right| - b\right)\right]^{-1}, \tag{2.14}$$

where the parameters $\theta_0, \theta_\mathrm{d}, a, b$ 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 Appendix 2.7.1). The pairwise molecular energy can be written as

$$P_m = \Phi(\theta_m)\Phi(\theta_{m+1} - v)\mathrm{E}^p_{\mathcal{F}(\theta_m)\text{-}\mathcal{F}(\theta_{m+1}-v)}(\Delta z_m), \tag{2.15}$$

where $\mathcal{F}(\theta_m) = \mathrm{nint}\left[(\theta_m - \theta_0)\,\theta_d^{-1}\right] \bmod 7{+}1$, $\mathrm{nint}[\cdots]$ rounds its argument to the nearest integer, v is the internal angle of the N-gon and $\Delta z_m = z_{m+1} - z_m$ is the relative axial translation between two collagen molecules. The energy of the whole microfibril is then simply

$$\mathrm{E_M} = \sum_{m=1}^{\mathrm{N}-1} P_m + \Phi(\theta_\mathrm{N})\Phi(\theta_1 - v)\mathrm{E}^p_{\mathcal{F}(\theta_\mathrm{N})\text{-}\mathcal{F}(\theta_1-v)}(\Delta z_\mathrm{N}). \tag{2.16}$$

Cyclical connectivity of the N-gon imposes a constraint

$$\Delta z_\mathrm{N} \equiv z_1 - z_\mathrm{N} = -\sum_{m=1}^{\text{N-1}} \Delta z_m. \tag{2.17}$$

We also define the microfibril energy E_θ, which is independent of the axial degrees of freedom z_m by excluding the $\mathrm{E}^p_{i\text{-}j}(\Delta z_m)$ terms from equation (2.16).

2.7 APPENDIX

2.7.1 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:

$$b = \frac{\log(q-1)f(\theta_\mathrm{f}) - \log(p-1)f(\theta_\mathrm{w})}{f(\theta_\mathrm{w}) - f(\theta_\mathrm{f})}, \quad a = \frac{\log(q-1) + b}{f(\theta_\mathrm{w})}, \quad f(t) = \sin\left(\frac{\pi t}{2\theta_\mathrm{d}}\right). \tag{2.18}$$

Parameters $(\theta_\mathrm{w}, \theta_\mathrm{f}, p, q)$ are defined via

$$\Phi(\theta_\mathrm{max} \pm \theta_\mathrm{f}) = p^{-1}, \;\; \Phi(\theta_\mathrm{max} \pm \theta_\mathrm{w}) = q^{-1}, \tag{2.19}$$

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

$$(\theta_0, \theta_\mathrm{d}, \theta_\mathrm{w}, \theta_\mathrm{f}, p, q) = \left(0.5, \frac{2\pi}{7}, \frac{\pi}{9}, 0.06, 1.0004, 100\right). \tag{2.20}$$

2.7.2 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.

2.7.2.1 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 $\mathrm{E}^{p}_{i\text{-}j}$, which determine the value of the microfibril energy - see equation (2.16). A priori we do not know the exact value of the parameter D. Based on the experimental measurements of the D-banding lengthscale, we require $\mathrm{D} \in [620, 700]$Å [10, 26]. Next, for each amino acid sequence, we construct a set of candidate values for the D-banding lengthscale, based on the axial staggers of the interaction energy minima of non-periodic pairwise potentials $\mathrm{E}_{i\text{-}j}$. The set of candidate values for D is defined as

$$S_{\mathrm{D}} = \left\{ \Delta\widetilde{z} \in [620, 720]\text{Å} \,\middle|\, \Delta\widetilde{z} = \arg\min_{\Delta z}\{\mathrm{E}_{i\text{-}j}(\Delta z)\},\, \mathcal{S}_{i\text{-}j}(\Delta\widetilde{z}) < \mathcal{S}_{\mathrm{thr}} \right\}, \tag{2.21}$$

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

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

$$I_{\mathrm{D}} = \bigcup_{\Delta\widetilde{z} \in S_{\mathrm{D}}} \left[\Delta\widetilde{z} - \delta z, \Delta\widetilde{z} + \delta z\right], \tag{2.22}$$

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

2.7.2.2 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 $2\mathrm{N} - 1$ degrees of freedom in our model. To aid us in finding the global minimum of E_{M}, we construct near-equilibrium states (NEqSs) which will give the largest energy contributions to the spectrum. NEqSs are members of the set $\mathrm{S}_{\mathrm{eq}} = \{(\vec{\theta}^{*}, \vec{\Delta z}^{*})\}$, in which the state vector $\vec{s} = (\vec{\theta}^{*}, \vec{\Delta z}^{*})$ specifies the microscopic state of a microfibril. The azimuthal components of

Table 2.5: Definitions of microfibril categories, based on the pattern of axial staggers.

Microfibril Category	**[a]Condition on Axial Staggers**	
Perfectly-staggered	$\Delta z_m^* = k\mathrm{D}, \text{ for all } m = 1, \ldots, \mathrm{N}-1$	(2.24)
In-register	$\Delta z_m^* \in [-\Delta_0, \Delta_0], \text{ for all } m = 1, \ldots, \mathrm{N}-1$	(2.25)
Mixed	Any other Δz_m^* that are not perfectly-staggered or in-register	

[a] We set the parameter $\Delta_0 = 5\,\mathrm{nm}$ and $k = 1, 2, 3$ or 4

the state vector maximise the strip overlap in a given N-gon:

$$\theta_l^* = \underset{\theta\in[0,2\pi)}{\arg\max}\{\Phi(\theta)\Phi(\theta - v)\}, \quad l = 1, \ldots, \mathrm{N}. \tag{2.23}$$

There are $\mathrm{N}-1$ axial components of the state vector $\vec{s}$ which correspond to the staggers that minimise the axial energy component for a given pair of strips in a microfibril. The total number of NEqSs is $7^{\mathrm{N}}M^{\mathrm{N}-1}$, where M is the number of minimisers for each pairwise interaction potential $\mathrm{E}_{i\text{-}j}^p$. 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 $[\theta_1^* - \delta\theta, \theta_1^* + \delta\theta] \times \cdots \times [\theta_{\mathrm{N}}^* - \delta\theta, \theta_{\mathrm{N}}^* + \delta\theta]$ with $\delta\theta = 0.15$.

2.7.2.3 Calculation of Equilibrium Probabilities with Perturbation Sensitivity

Finally, we calculate the equilibrium statistics of the collagen microfibril. We group the microscopic microfibril states into 3 categories based on the axial components of the state vector $\vec{s}$. The definitions of the microfibril categories are shown in Table 2.5 and Figure 2.7. The equilibrium probability in the canonical ensemble formalism for perfectly-staggered microfibrils is then

$$\mathcal{P}_{\mathrm{PS}} = \frac{\sum\limits_{\substack{\vec{s}\in \mathrm{S_{eq}},\\ \Delta z_m^* \text{ satisfy (2.24)}}} e^{-\beta \mathrm{E_M}}}{\sum\limits_{\vec{s}\in \mathrm{S_{eq}}} e^{-\beta \mathrm{E_M}}}, \tag{2.26}$$

where $\beta^{-1} = k_B T$ and the microfibril energy $\mathrm{E_M}$ is calculated using $g = 5\mathrm{D} - L$, where D is chosen according to the procedure outlined in subsection 2.7.2.1. Analogous formulae define the equilibrium probabilities of mixed and in-register states by suitably adjusting the condition on axial staggers Δz_m^* in equation (2.26) according to Table 2.5.

For a given perturbation sensitivity threshold $\mathcal{S}_c$, we include a NEqS in calculation of $\mathcal{P}_{\mathrm{PS}}$ if for a given $\vec{\theta}^*$, the components of the stagger vector satisfy

$$\mathcal{S}_{i\text{-}j}(\Delta z_l^*) < \mathcal{S}_c \text{ for all } l = 1, \ldots, \mathrm{N}, \tag{2.27}$$

where $\Delta z_{\mathrm{N}}^* = \left(-\sum_{m=1}^{\mathrm{N}-1} \Delta z_m^* \right) \bmod 5\mathrm{D}$. We note that equation (2.27) must hold for all strip pairs i-j which interact in a microfibril specified by the azimuthal components of the state vector $\vec{\theta}^*$.

If we find that there exists a value of $\mathcal{S}_c$, such that $\mathcal{P}_{\mathrm{PS}} \to 1$, 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.

2.7.3 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 arclength s in equation (2.4). Then, the elastic energy of deformation is minimised for an equilibrium helical angle $\phi = \phi^*$ which satisfies

$$2\gamma \sin^3 \phi^* \cos \phi^* + (\sin \phi^* \cos \phi^* - \chi\epsilon) \cos 2\phi^* = 0, \tag{2.28}$$

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

Our goal is to construct an asymptotic series for the equilibrium helical angle ϕ^* as a function of $\epsilon \ll 1$ and finite γ. To that end, we note that we can write ϵ as a function of ϕ^* in equation (2.28), obtaining

$$\bar{\epsilon} = (1 - \gamma) \sin \bar{\phi} + \gamma \tan \bar{\phi} \equiv f\left(\bar{\phi}\right), \tag{2.29}$$

where for convenience we have defined $\bar{\phi} = 2\phi^*$ and $\bar{\epsilon} = 2\chi\epsilon$. 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 [29] to obtain

$$\bar{\phi} \equiv g\left(\bar{\epsilon}\right) = \sum_{n=1}^{\infty} g_n \bar{\epsilon}^n, \tag{2.30}$$

where the expansion coefficients are given by

$$g_n = \frac{1}{n!} \lim_{\bar{\phi} \to 0} \frac{\mathrm{d}^{n-1}}{\mathrm{d}\bar{\phi}^{n-1}} \left(\frac{1}{h\left(\bar{\phi}\right)} \right)^n, \quad \text{where } h\left(\bar{\phi}\right) = \frac{f\left(\bar{\phi}\right)}{\bar{\phi}}. \tag{2.31}$$

We note that for $\overline{\phi} \ll 1$ we can expand $h\left(\overline{\phi}\right) = 1 + \frac{3\gamma-1}{6}\overline{\phi}^2 + O\left(\overline{\phi}^4\right)$. Using equation (2.31), the equilibrium helical angle is then asymptotically found to be

$$\phi^* = \chi\epsilon + \frac{2(1-3\gamma)}{3}(\chi\epsilon)^3 + O\left(\epsilon^5\right). \tag{2.32}$$

With the aid of the asymptotic expression in equation (2.32), we can estimate the equilibrium bend and twist energy contributions per unit length as

$$\mathrm{E}_{\mathrm{bend}} = \frac{B}{2}\frac{\sin^4\phi^\star}{R^2} \sim \frac{B\epsilon^4}{2R^2}, \qquad \mathrm{E}_{\mathrm{twist}} = \frac{C}{2R^2}\left(\frac{\sin 2\phi}{2} - \chi\epsilon\right)^2 \sim \frac{C\gamma^2\epsilon^6}{2R^2}. \tag{2.33}$$

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

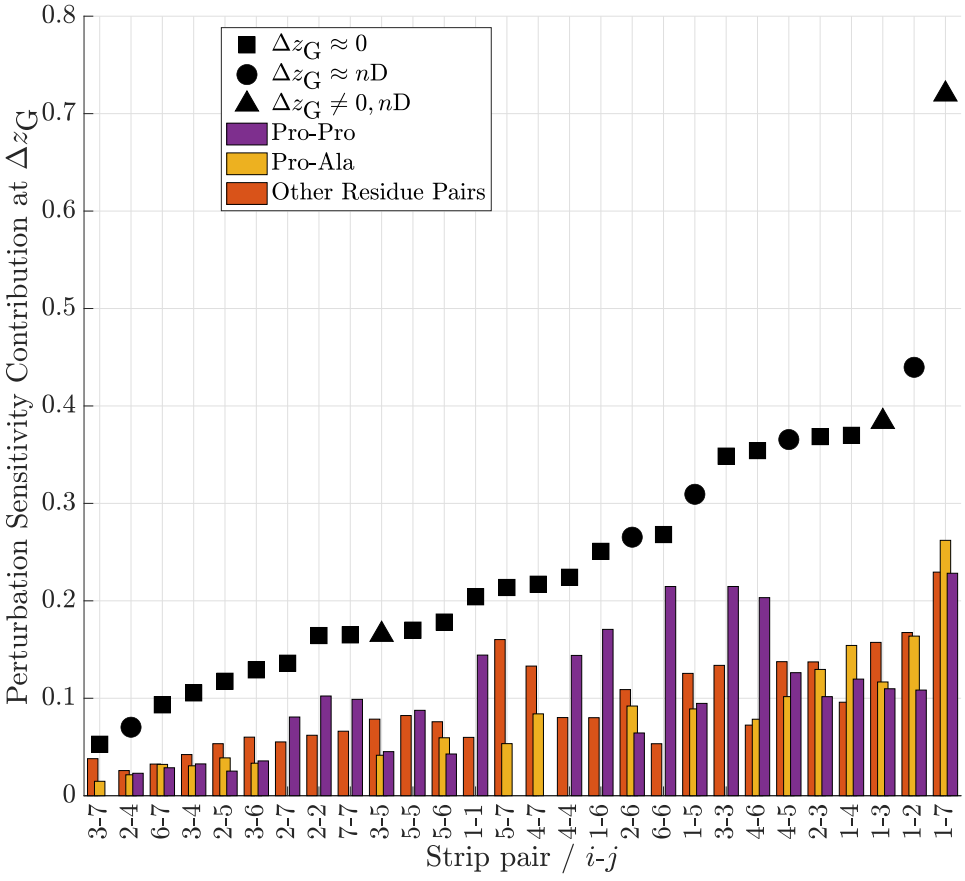


Figure 2.9: The largest contributions to the perturbation sensitivity of the global minima of the strip-strip energy $E^p_{i\text{-}j}(\Delta z_G)$ from the pairs of interacting residues. The black markers show the total perturbation sensitivity of each minimum. All residue pairs which contribute less than 20% to the total perturbation sensitivity, are categorised as "other residue pairs" (see Materials and Methods 2.6.3 for details). Typically, the highest contributions to the perturbation sensitivity come from two pairs of interacting residues, Pro-Pro and Pro-Ala.

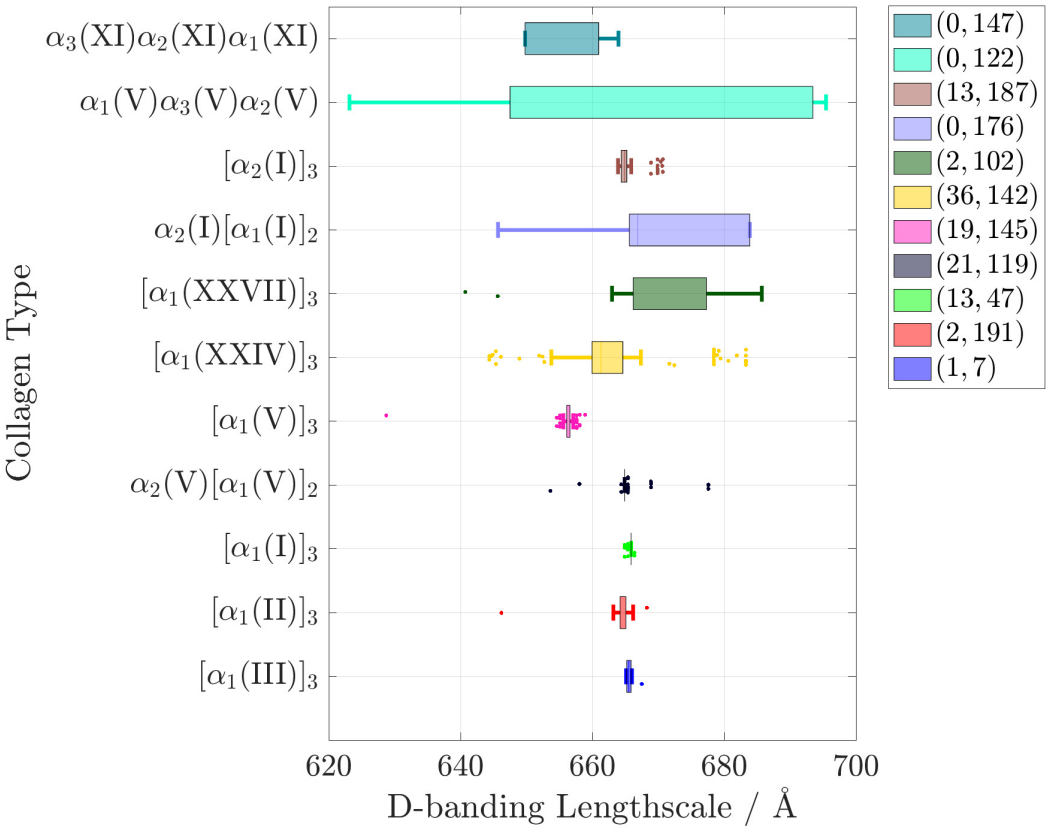


Figure 2.10: Box plot of the D-banding lengthscales across different mammalian species that give rise to stable perfectly-staggered microfibrils. The whiskers are drawn up to the largest/smallest data point that is within $1.5 \cdot \mathrm{IQR}$ (inter-quartile range) of the upper/lower quartile, indicated by the right/left edges of each box respectively. D-banding lengthscales that are a distance more than $1.5 \cdot \mathrm{IQR}$ from the right/left edges of a box are labelled as outliers and plotted as individual points. To each box, we associate an ordered pair $(N_{\mathrm{out}}, N_{\mathrm{tot}})$, 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.

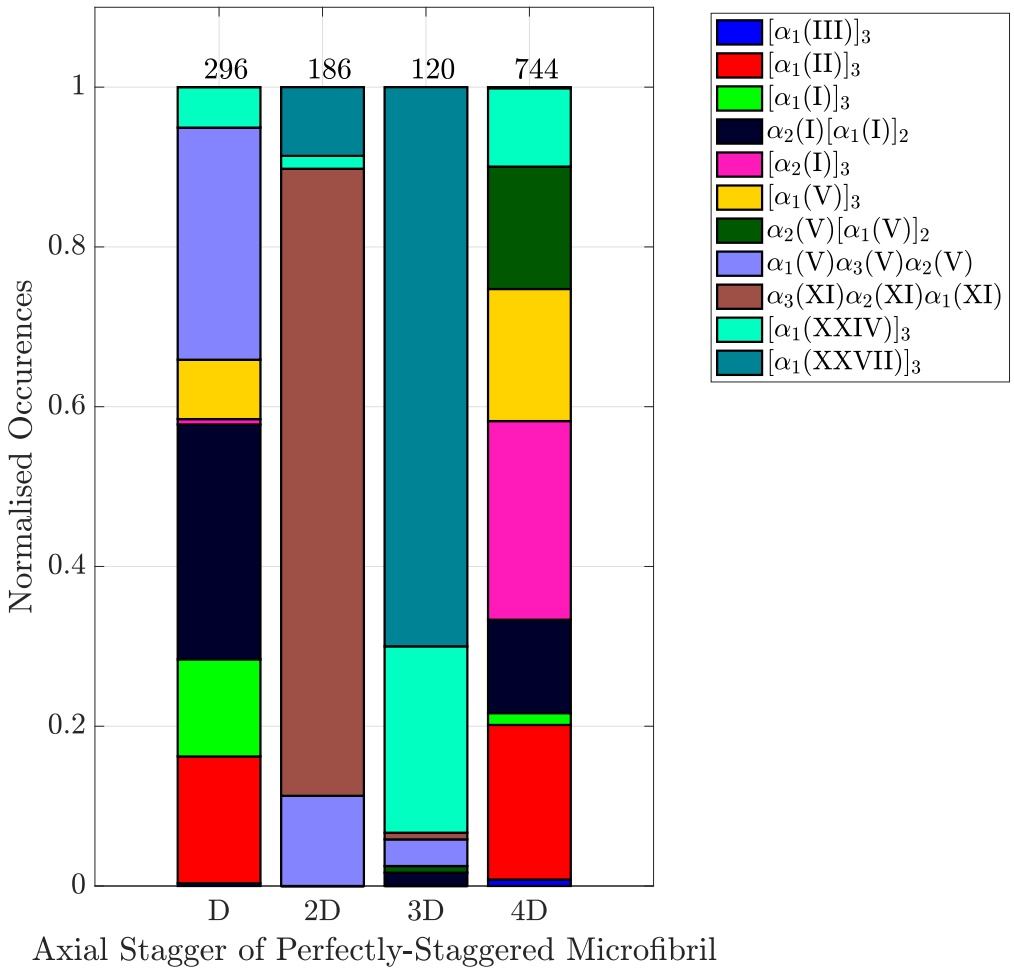


Figure 2.11: Histogram of the axial stagger value in stable perfectly-staggered microfibrils across different collagen types in mammalian species. The number of occurrences is normalised by the total number of stable perfectly-staggered microfibrils with a given axial stagger, which is shown at the top 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 2.10). Axial stagger values of D, 2D and 3D, 4D correspond to microfibrils with left-handed and right-handed chirality respectively.

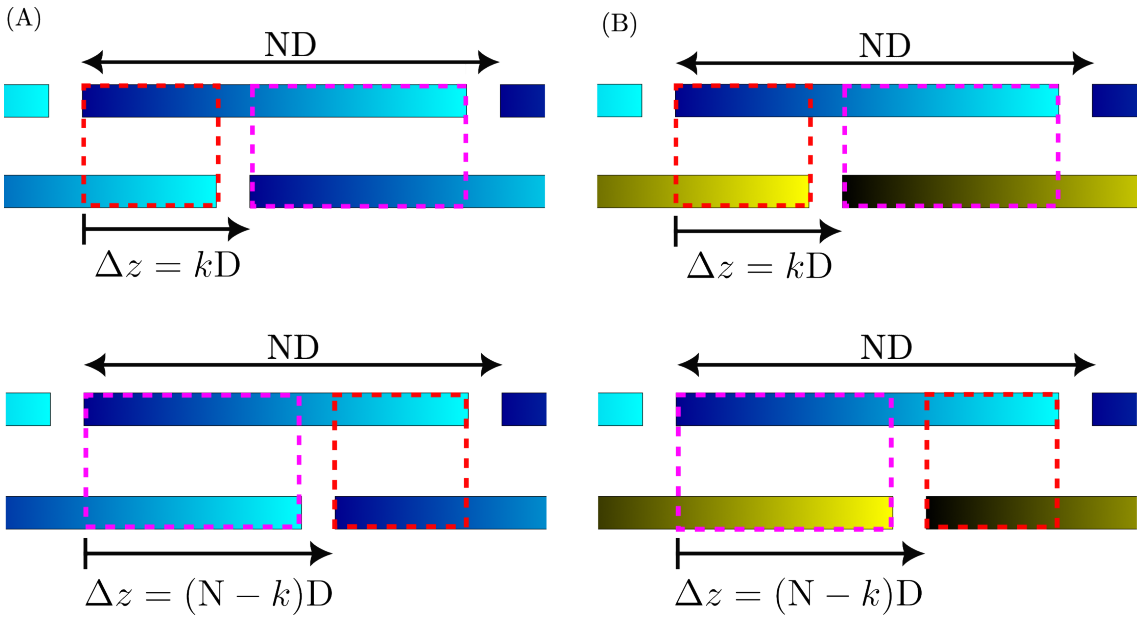


Figure 2.12: Schematic representation of pairwise, ND axially periodic collagen interactions showing: (A). Equivalence of interactions at $\Delta z = k\mathrm{D}$ and $\Delta z = (\mathrm{N} - k)\mathrm{D}$ for each molecule contributing an identical set of residues. (B). Non-equivalence of interactions at $\Delta z = k\mathrm{D}$ and $\Delta z = (\mathrm{N} - k)\mathrm{D}$ when each molecule contributes a distinct set of residues. Different colours represent distinct residue contributions.

— 3 —

Axial Order-Disorder Phase Transitions in Collagen Microfibrils are Controlled by pH and Residue pKa

Abstract

Collagen possesses a remarkable ability to spontaneously self-assemble in a hierarchy of fibrillar structures. The characteristic feature of collagen aggregates is the presence of periodic molecular ordering. This ordering is sensitive to pH variation, leading to the experimental observation of two distinct aggregated collagenous phases - an axially ordered and an axially disordered phase, which respectively possess and lack periodic molecular ordering. In this work, we use statistical mechanical models to study the effect of pH and residue pKa on collagen aggregation. We show that the existence of a thermodynamically stable axially disordered phase depends on the pKa values of ionisable residues. Further, we demonstrate that charged interactions involving Arg and Glu are vital for the formation of a thermodynamically stable axially ordered phase. On the other hand, we show that a loss of charged interactions involving Lys or Asp does not preclude formation of a stable axially ordered phase. In addition to being the first to predict phase transitions between axially ordered and disordered collagenous phases, our collagen self-assembly model is easily adaptable for theoretical study of site-specific residue modification in collagen, such as glycation. This opens a broad avenue for future quantitative investigation of the effect of age and pathology related residue modification in collagen on its self-assembly.

3.1 PRELIMINARY REMARKS

In Chapter 2, we have demonstrated that the chiral interactions between the helical residue strips on the collagen molecular surface determine the key structural features of the D-banded collagen microfibril. There are still, however, several important aspects of polymorphic collagen self-assembly that remain unaddressed.

Firstly, we saw in Figure 2.7 that there exist values of the perturbation sensitivity threshold, for which the "mixed" microfibrils corresponded to the globally stable conformation. This suggested the possibility of forming a microfibrillar aggregate that lacks axial periodicity, corresponding to the disordered fibril from Figure 1.2G. In fact, as we have already mentioned, unlike other polymorphic collagen fibrils, disordered aggregates are capable of forming under the same experimental conditions (temperature, type and concentration of dissolved solutes) as D-banded fibrils with the exception of the requirement for acidic pH for the former [51]. As such, we expect our microfibril self-assembly model to predict the existence of these two distinct aggregated phases at the relevant pH levels.

Secondly, although the perturbation sensitivity threshold approach to finding the lowest energy microfibrillar state did prove to be successful, it nevertheless indicated that the contact potentials do not fully encapsulate residue-residue interactions between triple

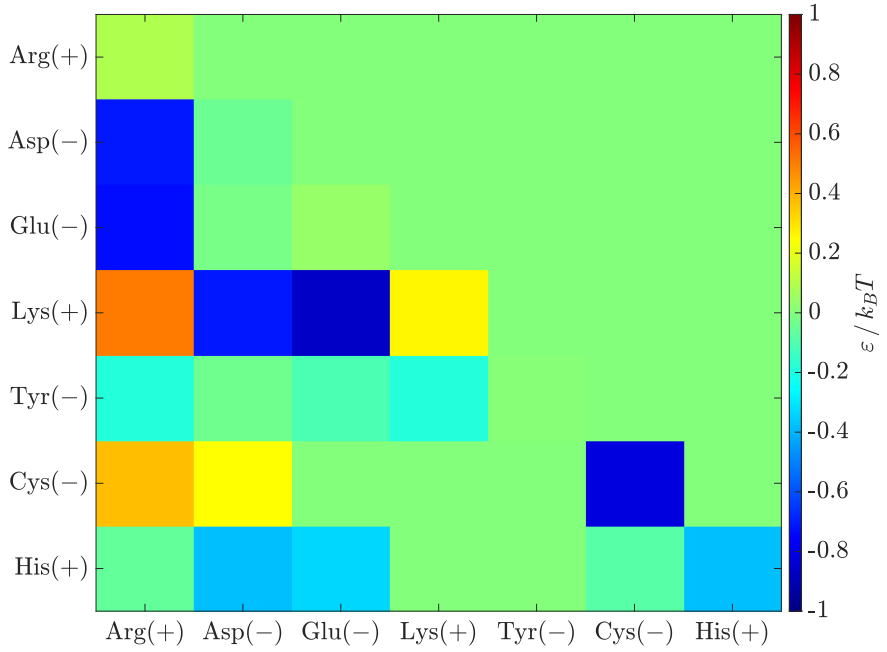


Figure 3.1: Miyazawa-Jernigan contact potentials for charged-charged residue interactions.

helices. We found that without resorting to perturbation sensitivity filtering of NEqSs, microfibrils corresponding to SLS aggregates were the most thermodynamically stable configuration. This does not agree with known experimental results, as collagen forms SLS aggregates often at acidic pH and further requires the presence of other charged molecules, such as ATP or diazo dyes [50, 120]. Our model in Chapter 2 accounted only for collagen-collagen interactions, which are expected to encode D-banded microfibrils as the globally stable conformation at physiological ionic conditions.

The aforementioned discrepancy between our theoretical predictions and experimental results is not surprising, as Miyazawa-Jernigan contact potentials are based primarily on the structures of globular proteins, which need not accurately describe the residue interactions between fibrillar proteins, such as collagen [75]. Evidence pointing to potential problems can be seen in Figure 3.1, which shows the strength of charged-charged interactions according to Miyazawa-Jernigan contact potentials. In particular, we notice that the interactions between identical ionisable residues are often not repulsive. This provides a poor model for charged residue interactions in collagen, as it is well-known that at highly acidic/basic pH collagen stops to aggregate due to acquisition of large positive/negative surface charge respectively, owing to residue ionisation [77].

The aforestated observations lead us to the main topic of Chapter 3 - studying the effect of ionic conditions on collagen self-assembly. We will see that accounting for the dependence of charged-charged residue interactions on the ionic conditions will remove the need for perturbation sensitivity analysis altogether. Under physiological ionic conditions, D-banded microfibrils will correspond to the most energetically stable state. On the other hand, at acidic pH, we will observe globally stable disordered microfibrils, corresponding to the fibrils lacking any banding pattern in Figure 1.2G. Most importantly, we will acquire novel understanding of the role played by charged residue interactions in controlling the emergence of axial molecular order in collagen microfibrils. If there was one thing to take away from Chapter 3, it would be the following fact: interactions of chemically similar charged residues can have diametrically opposite effects on the energetic stability of D-banded microfibrils. As we shall see, this unexpected result is directly relevant to understanding the effect of age- and pathology-induced chemical changes in tissues containing collagen.

3.2 INTRODUCTION

Collagens are a family of proteins that plays a crucial structural role across a range of biological contexts. The collagen polypeptide family is comprised of almost 30 different

collagen types that are distinguished by their amino acid composition and the biological functions served by the hierarchical structures that they form [10].

Fibrillar collagens of types I, II, III, V, and XI are distinguished by their ability to spontaneously aggregate under physiological ionic conditions into well-organised supramolecular aggregates known as D-banded collagen fibrils (see Figure 1.2B) [10]. The individual collagen molecules comprising the fibrils are made up of three helical polypeptide chains, known as α-chains, which are characterised by a repetitive sequence motif [Gly-X-Y]. Three left-handed α-chains supercoil into a right-handed triple helix. Supercoiling stabilises the triple helix via burial of the Gly residues, facing them towards the triple helical axis and enabling formation of hydrogen bonds [8]. The remaining X, Y residues are found on the collagen molecular surface and can engage in molecular interactions. The resulting triple helix has a molecular length $L \approx 300\,\text{nm}$ and a diameter of $\approx 1.5\,\text{nm}$.

Around 15-20% of all residues in fibrillar collagens are charged [113]. Collagen aggregation therefore has a strong dependence on pH. Across a broad range of pH $\approx$ 4-9, collagen exists in the form of an axially ordered aggregated phase, i.e. D-banded fibrils [77, 51]. These fibrils have lengths reaching several micrometers and diameters ranging from 30 nm to 200 nm depending on the collagen types comprising the fibril and the relevant biological context [99]. The overarching structural feature common to collagen fibrils is the presence of axial order in the comprising collagen molecules. The axial order manifests itself as D-banding - an alternating D $\approx 67\,\text{nm}$ periodic pattern of relatively high and low protein density regions along the fibril axis, visible in negatively stained TEM fibril samples. This observation was explained by Hodge & Petruska by envisioning the collagen fibril as being comprised of pentameric building blocks - microfibrils [90]. The molecules inside a microfibril are translated (staggered) along the long axis of the microfibril by integer multiples of the D-banding lengthscale, giving rise to the banded pattern seen experimentally. The notion of microfibrillar subunits has since been supported by X-ray diffraction as well as mechanical studies of collagen fibrils [83, 82, 43, 121].

Axial order is not common to all aggregated phases of collagen. At mildly acidic pH $\approx$ 2.5-3.5, collagen exists as an axially disordered aggregated phase, which manifests as a loose network comprised of axially disordered filaments (see Figure 1.2G). These long ($> 1\,\mu\text{m}$) filaments are $\approx 30\,\text{nm}$ in diameter and lack the characteristic D-banding pattern [51, 15]. At extreme pH < 2.0 and pH > 10.0, collagen aggregation stops altogether and collagen remains in a single molecule phase. This occurs due to a large net positive and negative charge respectively on the collagen molecular surface [77].

A detailed understanding of the effect of ionic conditions on collagen self-assembly is important for a range of biomedical applications. Ability of collagen to form axially ordered fibrils has broad biological significance through its correlation with fibril mechanical properties [26]. It has also been suggested to be important for several biological processes,

including biomineralisation as well as cell direction and alignment in the extracellular matrix [101, 94]. Pathological conditions such as chronic wounds and cancer are known to modify their local chemical environment through fluctuations in pH [12, 57]. Collagen surrounding cancer-affected tissues has been shown to possess altered morphology from that encountered in healthy tissues, leading to suggestions of its role in cancer cell invasions [48]. Ionisable residues on the collagen molecular surface may lose their ability to partake in charged interactions by undergoing irreversible non-enzymatic modification through processes like glycation. Glycation accelerates in presence of chronic diabetes, leading to deterioration in mechanical properties of collagen aggregates [5]. Prolonged glycation of molecular collagen has been shown to inhibit its ability to self-assemble altogether [103].

It is desirable to construct physical, quantitative models to establish the relationship between the amino acid sequence of molecular collagen and the various aggregated phases that it is capable of forming under varying external ionic conditions. The sequence-structure relationship of fibrillar collagens has been investigated in a number of theoretical works through construction of pairwise molecular interaction potentials for collagen molecules [53, 52, 92, 95]. A single collagen molecule contains in excess of 3000 residues, as such calculations of pairwise molecular interactions that balance computational tractability with sufficient physical detail of molecular interactions can prove to be a challenge. Recently, Puszkarska *et al.* have introduced a model for calculating pairwise molecular collagen interactions, which utilises Mijazawa-Jernigan contact potentials (MJCP) to model pairwise residue interactions [95]. MJCP are calculated on the basis of the spatial proximity of residues in known high-resolution protein structures. The resulting physical model has provided valuable insight into the sequence-structure relationship for a broad range of collagens that form D-banded fibrils. The downside of this approach is that MJCP for varying ionic conditions are not readily available, thus providing no understanding of the self-assembly dependence on them.

Other authors have introduced quantitative models that allow for studying the effect of pH and ionic strength on the electrostatic energy contribution to the collagen molecular interactions [116, 77]. Wallace proposed an entropy-driven mechanism to explain the phase transition between the single molecular (isotropic) and generic aggregated (anisotropic) phases, using Flory's theory for rod-like polymers [115, 116]. Within the scope of such phenomenological models, the free energy contributions of specific residue interactions are not predicted by the model and are instead inferred from fitting the model to experimental data. Morozova *et al.*, on the other hand, explicitly calculated the dependence on pH and ionic strength of the electrostatic interactions between discrete charged sites in molecular collagen [77]. The calculated pairwise electrostatic potentials were then used to justify the experimentally observed aggregation and morphology of collagen fibrils at varying pH and ionic strengths [77].

However, several important questions pertaining to the ionic dependence of collagen self-assembly remain unaddressed: (1). What is the physical origin of the axially disordered aggregated phase observed at mildly acidic pH? Previous works have focused on explaining the phase transition between the single molecular and generic aggregated phases [115, 116]. However, no explanation has been provided for the emergence of periodic molecular ordering in an aggregated collagenous phase. The existence of two aggregated phases that have distinct axial molecular ordering therefore remains unexplained. (2). What is the role of specific charged residues in driving collagen self-assembly? Morozova *et al.* did not elucidate the details of the effect that ionisation of specific charged residues has on collagen self-assembly, which thus remains to be investigated. (3). What is the effect of pKa variation on collagen self-assembly? This consideration is important, since residue pKa values can vary significantly between different proteins [85]. To our knowledge, no theoretical study has addressed this effect.

In this work, we combine and extend the approaches introduced by previous theoretical studies [95, 116, 77]. In section 3.3.1 we construct a mathematical model for pairwise collagen molecular interactions that accounts for the 3-dimensional spatial organisation of residues as well as the influence of ionic effects on residue-residue interactions. We then integrate these pairwise molecular interactions into our previously developed model of equilibrium microfibril self-assembly [124] in section 3.3.2. In sections 3.4.1 and 3.4.2 we use this model to predict and analyse a phase diagram for collagen aggregation as a function of pH and residue pKa. We demonstrate that the axially disordered aggregated phase is thermodynamically stable at mildly acidic pH, which has not been explained by prior theoretical works [115, 77]. Additionally, we show that its stability is strongly dependent on the residue pKa, the impact of which on collagen aggregation has not been studied previously [115, 77]. Finally, in section 3.4.3, we use our self-assembly model to infer the individual contributions of acidic and basic residues towards stabilisation of the axially ordered aggregated phase, extending the general findings of previous theoretical studies [77].

3.3 EQUILIBRIUM MODEL OF MICROFIBRIL SELF-ASSEMBLY

3.3.1 Residue Spatial Organisation & Pairwise Interactions

To proceed with calculations of pairwise collagen molecular potentials, we need to describe the residue spatial organisation and their pairwise interactions. Following our previous work [124], we use a statistical parametrisation of the collagen triple helix based on the high resolution structural data of short model peptides that are representative of short sections of the collagen's triple helix [96]. We use the parameter set that describes the Pro-rich

sections of the triple helix - see Figure 3.2B. The position of each X and Y residue within the [Gly-X-Y] triplet is represented by the coordinates of its C_α atom. The outward-facing residues organise into 7 right-handed helical strips of pitch $h \approx 200\,\text{nm}$ that lie on the collagen molecular surface.

To describe the pairwise residue interactions, we first split the standard set of 20 proteinogenic residues into two classes. The first class is the collection of all charged residues, which we denote by $\mathcal{C}$. $\mathcal{C}$ includes the acidic residues Asp, Glu, Tyr and Cys which acquire a unit negative electric charge upon ionisation and the basic residues Arg, Lys and His which have a unit positive electric charge in the ionised state. The remaining residues are denoted as hydrophobic.

For interactions of type hydrophobic-hydrophobic or hydrophobic-charged we follow the method of Puszkarska *et al.* and use statistical contact potentials to quantify the interaction strength [95]. We use the Miyazawa-Jernigan contact potentials, namely the

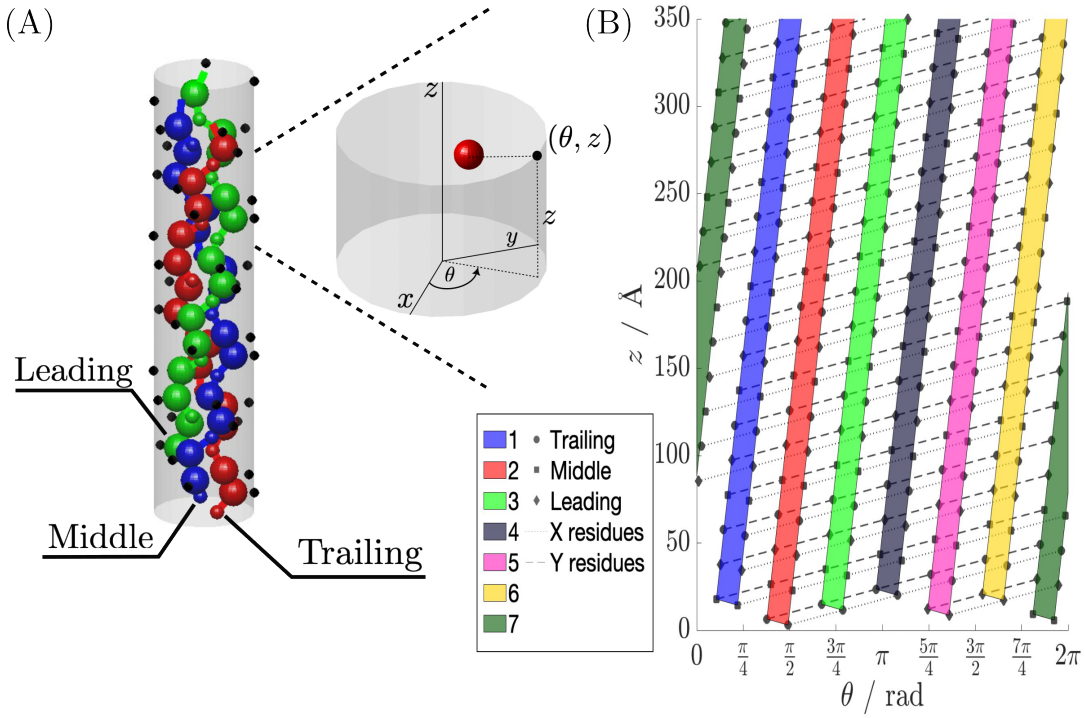


Figure 3.2: (A). Segment of the collagen triple helix bounded by a cylindrical surface onto which coordinates of each C_α residue atom are projected. The C_α atom positions are obtained using a statistically derived parametrisation based on analysis of Pro-rich model peptides [96]. (B) Helical residue organisation in molecular collagen. Dotted lines connect the residues that belong to the same α-chain. Solid lines indicate imaginary connections between residues that fall on one of the 7 right-handed spiral strips. We conventionally number the spiral strips in order of appearance when moving counter-clockwise around the molecular z-axis and assign the most N-terminal residue the azimuthal coordinate $\theta = \pi/2$.

entries MIYS850103, MIYS960102, MIYS990107 in the AAIndex database [60]. These can be interpreted as giving a first order approximation to the pairwise free energy of interaction between residues in an environment dominated by protein-protein interactions. Take two residues $\mathrm{R_p}$ and $\mathrm{R_q}$, where the positive integers p and q denote their positions in the sequence of the α-chain. Following other works [95], we use a step potential of width $l_c = 7.5\,\text{Å}$ to describe the dependence of the interaction energy between these residues on the separation r between them. The free energy of interaction energy is then given by

$$\Psi_{\mathrm{pq}}^{\mathrm{MJ}}(r) = \varepsilon_{f(\mathrm{p})f(\mathrm{q})}\left[\Theta\left(r\right) - \Theta\left(r - l_c\right)\right], \tag{3.1}$$

where Θ is the Heaviside function and ε is a 20×20 contact potentials matrix containing the free energies of interaction for each pair of residues. The function f maps the sequential positions p, q of the residues onto their integer residue designation from 1 to 20.

The electrostatic component of the pairwise collagen molecular interactions will in general involve several contributions. Apart from interactions between ionised residues, additional energetic contributions may arise from induced/permanent dipoles interacting with ionised residues [116]. Rigorous treatment of electrostatic interactions between helical biomolecules has been studied by other authors [63]. Here we pursue a simplified model of electrostatic interactions. In what follows, we only account for the energetic contribution arising from pairwise interactions of ionisable residues. Our task then is to determine the distribution of ionised residues on the collagen molecular surface, given the ionic conditions. At thermodynamic equilibrium, the propensity of a given residue's functional group to ionise is governed by its disassociation equilibrium constant K_{a}[1]. Assuming a well-mixed solution, the value of K_{a} is determined by a combination of spatially homogeneous parameters such as temperature, pH and I (ionic strength) as well as by site-specific factors. These site-specific factors include partial or complete water-sequestration of residues as well as local interactions between residues, which may favour either the ionised or the non-ionised form of the residue. As a consequence of these site-specific effects, the equilibrium disassociation constants K_{a} can vary by several orders of magnitude, depending on the specific protein and sequence position of the residue [85]. We choose a simple and analytically tractable approach, wherein we will treat each distinct ionisable residue along the triple helix as having an effective charge z^{eff}, which is independent of its position in the sequence and local environment. We define the effective charge as

$$z^{\mathrm{eff}}(\mathrm{R}) = \begin{cases} \alpha(\mathrm{R}) z_{\mathrm{R}}, \text{ if R has an acidic side chain,} \\ \left(1 - \alpha(\mathrm{R})\right) z_{\mathrm{R}}, \text{ if R has a basic side chain,} \end{cases} \tag{3.2}$$

where z_{R} is the electric charge of the side-chain of residue R in its ionised form and the pre-factor in front of z_{R} represents the equilibrium fraction of R with its side-chain in

[1] A common nomenclature, which we will use interchangeably is pKa, defined as $-\log_{10} K_{\mathrm{a}} \equiv \mathrm{pKa}$.

the ionised state. The pre-factor is dependent on the disassociation fraction $\alpha(\mathrm{R})$, which is determined by the pH, K_a of the residue R and I (see Appendix subsection 3.7.1 for detailed expressions). We describe the charged-charged interactions between two residues using a Debye-Hückel potential. Omitting physical constants, the potential is

$$\Psi^{\mathrm{DH}}_{\mathrm{pq}}(r) \propto \begin{cases} z^{\mathrm{eff}}(\mathrm{R_p}) z^{\mathrm{eff}}(\mathrm{R_q}) e^{-\kappa r}/r \equiv \psi(r), \text{ for } r > a_0, \\ \psi(a_0), \text{ for } r \leq a_0, \end{cases} \tag{3.3}$$

where $\kappa \propto \sqrt{I}$ is the inverse Debye length and a_0 is a tunable parameter that sets the maximum strength of electrostatic interactions $\psi(a_0)$. The ionic strength is given by $I = \frac{1}{2}\sum_\alpha n_\alpha z_\alpha^2$, with n_α being the solute number density and z_α the corresponding integer charge of the solute.

Throughout this work we set $I = 150\,\mathrm{mM}$, corresponding to a physiological salt concentration [77]. The distance of closest approach a_0 will in general be determined by the complex steric interactions between the residue side chains and is expected to be dependent on the residue sequence. Here, we will characterise all charged residue interactions by a single value of a_0. We estimate a_0 based on the closest lateral distance between collagen molecules in a native fibril and the Pro-rich statistical parametrisation of the triple helix. The lateral separation d_M between collagen molecules in native fibrils has been reported in the range 10-15 Å [32]. For a given value of d_M, we expect a_0 to be set by the residues with the largest radial coordinate. For the Pro-rich parametrisation of the triple helix, this corresponds to the X residues in the [Gly-X-Y] sequence motif with a radial coordinate $\rho_\mathrm{X} = 4.10\,\text{Å}$ [96]. Taking the inter-molecular separation to be $d_\mathrm{M} = 12.5\,\text{Å}$, we estimate $a_0 \approx d_\mathrm{M} - 2\rho_\mathrm{X} \approx 4\,\text{Å}$, which corresponds to a maximum charged interaction strength $\psi(a_0) \approx 1.2\,k_B T$.

3.3.2 Pairwise Molecular Interactions & Microfibril Energy

We take the total interaction energy between two collagen molecules to be the sum of pairwise interactions between spiral strips on the molecular surfaces of the interacting molecules. We denote a pair of interacting strips by i-j, where $i, j = 1, 2, \ldots, 7$. Let Δz denote the stagger of j relative to i. The strip-strip pairwise interaction energy is then given by

$$\mathrm{E}_{i\text{-}j}(\Delta z) = \sum_{\mathrm{R_p}, \mathrm{R_q} \in \mathcal{C}} \Psi^{\mathrm{DH}}_{\mathrm{pq}}(r_{\mathrm{pq}}(\Delta z)) + \sum_{\mathrm{R_p \text{ or } R_q} \notin \mathcal{C}} \Psi^{\mathrm{MJ}}_{\mathrm{pq}}(r_{\mathrm{pq}}(\Delta z)), \tag{3.4}$$

where $r_{\mathrm{pq}}(\Delta z)$ is the stagger dependent distance between interacting residues $\mathrm{R_p}$ and $\mathrm{R_q}$. The detailed expressions for $r_{\mathrm{pq}}(\Delta z)$ in terms of residue coordinates can be found in subsection 3.7.2 of the Appendix. It should be noted that not all charged residues may be able to partake in intermolecular interactions. Spatially proximal charged residues

may instead partake in inter-chain interactions within a single collagen molecule through formation of salt bridges [47]. This effect can be accounted for by excluding specific residues R_p, R_q from the first summation in equation (3.4). Here, we will work under the assumption that all charged residues may in principle contribute to inter-molecular interactions.

For a strip i, we also expect significant contributions to the pairwise molecular energy from the interactions involving its nearest neighbouring strips - see Table 3.4 of the Appendix for details. At this point, we introduce some convenient notation. Consider some indexing variable l, which may take values $l = 1, 2, \ldots, l_{\max}$, where $l_{\max}$ is some positive integer greater than 1. We then define $l' = (l \bmod l_{\max}) + 1$ and $l'' = ((l-2) \bmod l_{\max}) + 1$. With this notation in mind, we denote the nearest neighbouring strips of i as i' and i'', which correspond to the azimuthally closest strips in the anticlockwise and clockwise directions about the molecular axis O_1 respectively - see Figure 3.3A. Assuming that the nearest neighbour strips only interact with one another, the total molecular pairwise interaction energy is then

$$U_{i\text{-}j}(\Delta z) = E_{i\text{-}j}(\Delta z) + E_{i'\text{-}j''}(\Delta z) + E_{i''\text{-}j'}(\Delta z). \tag{3.5}$$

In this work, we will study the self-assembly of collagen microfibrils, which constitute the smallest unit in the collagen structural hierarchy. We have previously shown that spatial residue organisation energetically favours a 5-membered microfibril [124], which is in agreement with the microfibril structure observed in X-ray scattering studies [83, 82]. Following our previous work, we represent a microfibril as a regular pentagon with a collagen molecule placed at each vertex - see Figure 3.3A. The m^{th} molecule in the microfibril can stagger in and out of the packing plane by a distance z_m as well as rotate around its axis through an angle θ_m.

Collagen microfibrils are comprised of periodically repeating collagen molecules separated by gaps of length g - see Figure 3.3B. Throughout this work we will take gap lengths g to lie in the interval $[g_a, g_b]$, which is approximately centred on the experimentally measured value of gap size in collagen (see subsection 3.7.5 of the Appendix for details). The pairwise molecular interactions in a microfibril with a gap of length g are therefore given by the linear periodic potential

$$U^p_{i\text{-}j}(\Delta z_m, g) = U_{i\text{-}j}(\Delta z_m) + U_{i\text{-}j}(\Delta z_m - g - L), \tag{3.6}$$

where $\Delta z_m = z_{m\prime} - z_m$, for $m = 1, \ldots, 5$ and we restrict the axial stagger $\Delta z_m \in [0, g + L)$. The second term in equation (3.6) accounts for the axially periodic placement of molecules in the microfibril. We will assume that only the nearest neighbouring molecules connected by the edges of the polygon interact with one another. The microfibrillar energy is then

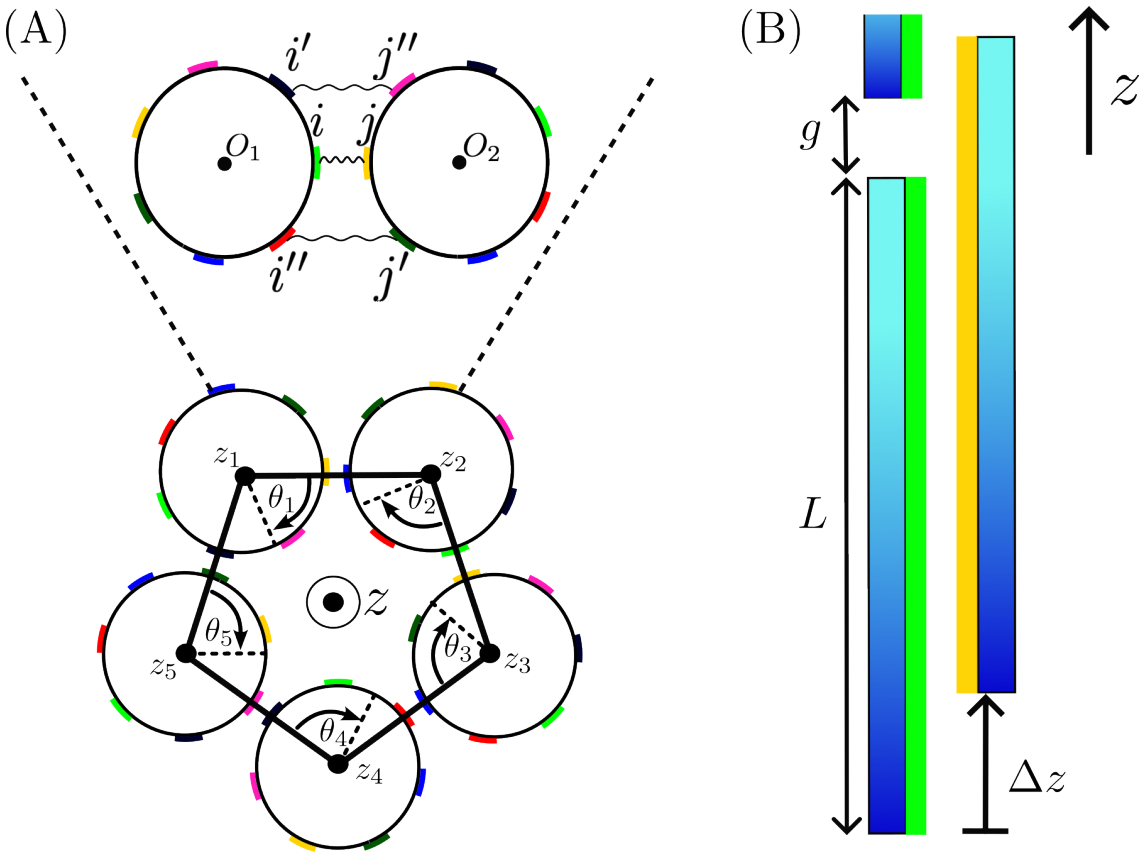


Figure 3.3: (A). Model of a collagen microfibril. Wavy lines indicate pairwise interactions between helical strips on the collagen molecular surface. (B). Side-view of a pairwise interaction between two spiral strips in a microfibril.

given by

$$\mathrm{E_M}(g) = \sum_{m=1}^{5} \mathrm{U}^{p}_{i_m\text{-}j_m}(\Delta z_m, g), \tag{3.7}$$

where i_m-j_m denotes a pair of interacting strips for molecules m and m' (the nearest anticlockwise neighbour of molecule m) for a given azimuthal configuration.

3.3.3 Axial Order in Microfibrils

Before proceeding further, it is crucial that we define precisely the notion of axial order and disorder in collagen microfibrils. Characterisation of collagen aggregates through negatively stained TEM produces regions of relatively high and low electron density that appear with periodicity D = 67 nm along the long axis of the collagen aggregate. The commonly accepted interpretation of this experimental result is that the high electron density regions correspond to axial gaps g between collagen molecules [90] - see Figure 2.4C. We remark that under the aforementioned interpretation, the existence of axial order

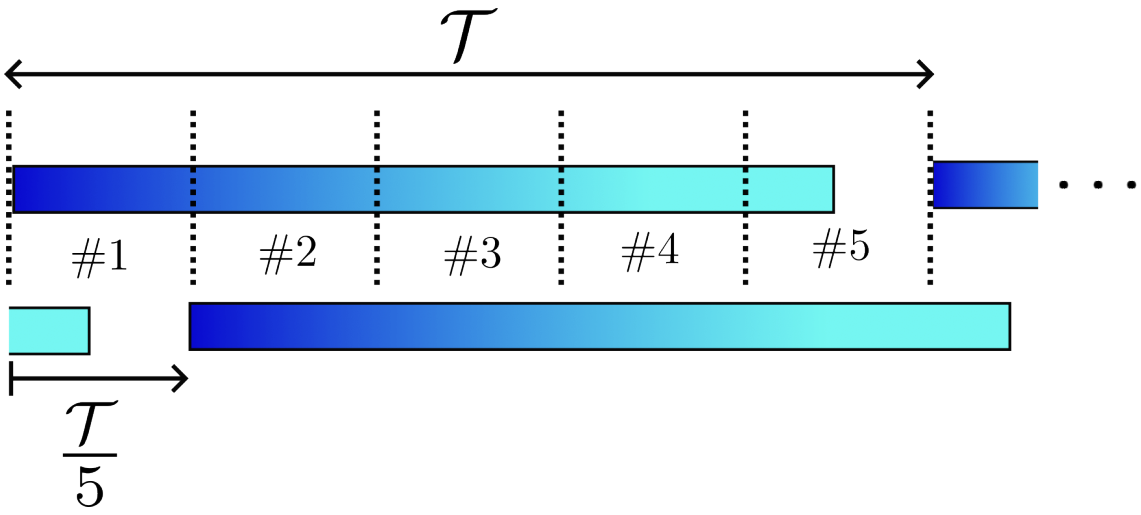


Figure 3.4: Partition of the axial periodicity lengthscale $\mathcal{T}$ into 5 equal-length segments numbered $\#1, \ldots, \#5$.

in collagen microfibrils is only a function of the axial degrees of freedom associated with the collagen molecules. Thus, the relative positions of collagen molecules in the plane normal to the microfibril axis are not relevant to the current discussion of axial order. In a microfibril comprised of 5 collagen molecules, each of length L, the axial period $\mathcal{D}$[2] is defined by

$$\mathcal{D} = \frac{L+g}{5} \equiv \frac{\mathcal{T}}{5}, \tag{3.8}$$

where for convenience, we have defined the axial periodicity lengthscale $\mathcal{T} = L + g$.

It remains determine for a microfibril in Figure 3.3A, the values of the axial staggers z_m that result in a periodic distribution of gaps along the long axis of the microfibril. We start by subdividing the axial periodicity lengthscale $\mathcal{T}$ into 5 intervals, as dictated by equation (3.8) - see Figure 3.4. Any axial stagger z_m corresponding to an integer multiple of the axial period $\mathcal{D}$ will result in a gap occupying one of the 5 intervals. To produce a periodic distribution of gaps, we require that there is exactly 1 gap in each of the aforementioned intervals. This can be achieved whenever

$$\left\{ \frac{z_1 \bmod \mathcal{T}}{\mathcal{D}}, \frac{z_2 \bmod \mathcal{T}}{\mathcal{D}}, \ldots, \frac{z_5 \bmod \mathcal{T}}{\mathcal{D}} \right\} = \{0, 1, \ldots, 4, \}\,, \tag{3.9}$$

where the curly braces $\{\ldots\}$ denote the fact that the elements inside the braces are unordered. Physically, this corresponds to the fact that the gap associated with a molecule in a microfibril can occupy any of the intervals resulting from partitioning of the axial periodicity lengthscale $\mathcal{T}$.

[2]Notice the subtle change in notation. $\mathcal{D}$ refers to any axial period in a microfibril, not just D = 67 nm.

We note that the definition of a $\mathcal{D}$-periodic microfibril according to equation (3.9) is subtly different from that of a perfectly-staggered microfibril in equation (2.24) of chapter 2. Namely, there are 4 perfectly-staggered (pentameric) microfibrils, whilst there are 24 $\mathcal{D}$-periodic microfibrils, including the 4 perfectly-staggered ones. Both $\mathcal{D}$-periodic and perfectly-staggered microfibrils by construction will result in a microfibril that is predicted to produce a $\mathcal{D}$-periodic banding pattern. The difference lies in the fact that only perfectly-staggered microfibrils have a clear notion of microfibril chirality. In this chapter, we are concerned more generally with the emergence of axial periodicity, rather than the different kinds of chirality as in chapter 2, hence the more general definition of equation (3.9).

3.3.4 Calculation of Equilibrium Statistics

In order to compute the spectrum of the collagen microfibril, for different values of g, we construct finite sets of near-equilibrium states (NEqSs) $\mathrm{S_{eq}}$. Each NEqS in $\mathrm{S_{eq}}$ corresponds to a local microfibril energy minimum in its $2\mathrm{N}-1$ dimensional phase space (N azimuthal and $\mathrm{N}-1$ axial degrees of freedom). We start by introducing discrete axial staggers $\Delta z^*_m = z^*_{m'} - z^*_m$, which correspond to the minima of the pairwise interaction potentials $\mathrm{U}^p_{i_m\text{-}j_m}$ in a microfibril. To describe a NEqS, let us define the vectors $\vec{i} = (i_1, \ldots, i_5)$, $\vec{j} = (j_1, \ldots, j_5)$ which characterise the azimuthal molecular orientations and the vector $\Delta\vec{z}^* = (\Delta z^*_1, \ldots, \Delta z^*_5)$ which describes the axial staggers in a microfibril. The phase space coordinates of a NEqS corresponding to a minimum of the microfibril energy are then given by $\vec{s} = (\vec{i}, \vec{j}, \Delta\vec{z}^*)$.

For each collection of NEqSs $\mathrm{S_{eq}}$, characterised by some fixed value of g, we can define the equilibrium probability of axially ordered microfibrils as

$$\mathcal{P}_{\mathrm{period}}(g) = \frac{\sum\limits_{\substack{\vec{s}\in \mathrm{S_{eq}},\\ z^*_m \text{ satisfy (3.9)}}} e^{-\beta \mathrm{E_M}(g)}}{\sum\limits_{\vec{s}\in \mathrm{S_{eq}}} e^{-\beta \mathrm{E_M}(g)}}, \tag{3.10}$$

where $\beta^{-1} = k_B T$ with T denoting the temperature. The construction of the set $\mathrm{S_{eq}}$ and the numerical implementation of condition (3.9) are detailed in subsection 3.7.4 of the Appendix. The equilibrium probability of the axially disordered microfibrils is similarly defined as

$$\mathcal{P}_{\mathrm{disorder}}(g) = \frac{\sum\limits_{\substack{\vec{s}\in \mathrm{S_{eq}},\\ \exists\, m, \text{s.t. } \Delta z^*_m \neq 0,\\ z^*_m \text{ do not satisfy (3.9)}}} e^{-\beta \mathrm{E_M}(g)}}{\sum\limits_{\vec{s}\in \mathrm{S_{eq}}} e^{-\beta \mathrm{E_M}(g)}}. \tag{3.11}$$

We remark that so far we have not defined the equilibrium probability of microfibrils with SLS axial staggers. However, as we shall see in section 3.4, we won't ever observe

Table 3.1: Classification of microfibrillar phases in the limit of an infinitely long microfibril.

Phase	$\mathrm{E}_{\mathrm{M}}^{\min} / k_B T$	$\lim_{L_{\mathrm{M}}\to\infty} \mathcal{P}_{\mathrm{period}}(g_{\min})$	[a]**Symbol**
Single Molecular	> 0	N/A	$\mid$
Axially Ordered	< 0	1	$\star$
Axially Disordered	< 0	0	$\times$

[a] The symbolic convention used as a shorthand representation of the relevant phase.

microfibrils with SLS axial staggers at thermodynamic equilibrium. As such, for the purposes of this chapter, it suffices to consider the equilibrium probabilities of just 2 distinct aggregated phases defined by equations (3.10) and (3.11).

A phase transition between axially ordered/disordered aggregated phases can occur in the zero-temperature limit $T \to 0$ through level crossing of two NEqS, with the requirement that only one of them satisfies condition (3.9) [46]. For collagen microfibrils, the zero-temperature limit is equivalent to a more physical limit of the infinite microfibril. We take an infinitely long microfibril to mean that it is comprised of $L_{\mathrm{M}} \to \infty$ identical microfibril segments, each of energy $\mathrm{E}_{\mathrm{M}}(g)$. This is a sensible limit to take, since the lengths of aggregated filaments comprising the axially ordered/disordered collagenous phases that are observed in fibrillogenesis experiments, significantly exceed the length of a single collagen molecule and, by extension, of a single microfibril segment [77, 51]. At thermodynamic equilibrium, we expect to observe the NEqS identified by the phase space point $\vec{s}_{\min}$, which corresponds to the minimum of microfibril energy $\mathrm{E}_{\mathrm{M}}^{\min} = \min_{g,\mathrm{S}_{\mathrm{eq}}}\{\mathrm{E}_{\mathrm{M}}(g)\}$ with gap length $g_{\min} = \operatorname{argmin}_{g,\mathrm{S}_{\mathrm{eq}}}\{\mathrm{E}_{\mathrm{M}}(g)\}$. The onset of aggregation will then occur precisely when $\mathrm{E}_{\mathrm{M}}^{\min} < 0\, k_B T$. Having reached the onset of aggregation, one of two aggregated phases can then form:

$$\begin{aligned}\lim_{L_{\mathrm{M}}\to\infty} \mathcal{P}_{\mathrm{period}}(g) &= \lim_{L_{\mathrm{M}}\to\infty} \frac{\sum\limits_{\substack{\vec{s}\in \mathrm{S}_{\mathrm{eq}},\\ z_m^* \text{ satisfies (3.9)}}} e^{-\beta \mathrm{E}_{\mathrm{M}}(g) L_{\mathrm{M}}}}{\sum\limits_{\vec{s}\in \mathrm{S}_{\mathrm{eq}}} e^{-\beta \mathrm{E}_{\mathrm{M}}(g) L_{\mathrm{M}}}} \\ &= \begin{cases} 1, \text{ if } \vec{s}_{\min} \text{ satisfies (3.9)}, \\ 0, \text{ if } \vec{s}_{\min} \text{ does not satisfy (3.9) and } \exists\, m \text{ s.t. } \Delta z_m^* \neq 0, \end{cases}\end{aligned} \tag{3.12}$$

The phase transition between the two aggregated phases is associated with a discontinuous change in the value of $\lim_{L_{\mathrm{M}}\to\infty} \mathcal{P}_{\mathrm{period}}(g)$ as a function of ionic parameters that affect the strength of residue-residue interactions, such as pH, K_{a} or I.

From the discussion above, we can construct a phase diagram for collagen microfibrils once $\mathrm{E}_{\mathrm{M}}^{\min}$ and $\lim_{L_{\mathrm{M}}\to\infty} \mathcal{P}_{\mathrm{period}}(g)$ are known. We then classify the microfibrils as one of

the 3 phases shown in Table 3.1.

3.3.5 Variation of Ionic Parameters

Our next step is to quantify the effect of the ionic parameters included in our model on microfibril self-assembly. These ionic parameters are pH, the 7 K_{a} values of the ionisable residues and the ionic strength I. Variation in I affects all charged residue interactions in the same way by changing the Debye length, since $\kappa \propto \sqrt{I}$. In turn, increasing I decreases the maximum attainable interaction strength $\psi(a_0)$ of a pairwise charged-charged residue interactions, resulting in a decreased contribution of electrostatic interactions to the pairwise molecular energy. This effect has been investigated by other authors and here we set the ionic strength at a constant physiological value $I = 150\,\mathrm{mM}$ [116, 77].

To incorporate variation in K_{a} values into our model, we construct a simple formalism that will enable us to describe the distinct ionisation regimes that collagen may take. We start by defining residue ionisation as beginning and ending at disassociation fractions $\alpha_1, \alpha_2 \in (0, 1)$ respectively such that $\alpha_2 > \alpha_1$. Each residue R can then be ascribed an ionisation window W_{R}, which is defined as the interval along the pH scale across which ionisation starts and finishes. The half width of the ionisation window can be shown to be

$$\Delta\mathrm{pH} = \frac{1}{2}\log_{10}\frac{\alpha_2(1-\alpha_1)}{\alpha_1(1-\alpha_2)}. \tag{3.13}$$

Throughout this work we take the onset/end of residue ionisation to correspond to starting/ending disassociation fractions $\alpha_1 = 0.05$ and $\alpha_2 = 0.95$ respectively. The most commonly encountered charged residues in fibrillar collagens occur roughly in abundance of 30 residues per spiral strip - see Figures 3.11A-E of the Appendix. With the aforestated choices of α_1 and α_2, we account for ionisation of all residues excluding at most ≈ 1.5 units of charge smeared across the surface of a given spiral strip. Outside of any given ionisation window we therefore expect ionisation to have a negligible effect on the self-assembly of fibrillar collagens. We note that for our choice of α_1, α_2, the ionisation window W_{R} is symmetric about the half-equivalence point $\mathrm{pH}_{\mathrm{eqv}}$ at which the disassociation fraction is exactly 1/2. We can therefore explicitly write

$$W_{\mathrm{R}} = \left[\mathrm{pH}_{\mathrm{eqv}}(\mathrm{R}) - \Delta\mathrm{pH},\, \mathrm{pH}_{\mathrm{eqv}}(\mathrm{R}) + \Delta\mathrm{pH}\right], \tag{3.14}$$

where $[\ldots,\ldots]$ denotes a closed interval along the pH scale.

It is convenient to introduce some additional notation. Let $\mathrm{R}_{(l)}$ denote one of 20 distinct proteinogenic residues, with the index l identifying the distinct residue. With this notation in mind, consider two distinct residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ with corresponding disassociation constants $K_{\mathrm{a}}(\mathrm{R}_{(1)})$ and $K_{\mathrm{a}}(\mathrm{R}_{(2)})$ respectively. We recognise that the ionisation behaviour of a polypeptide containing these residues will be determined by (i) the order in which the

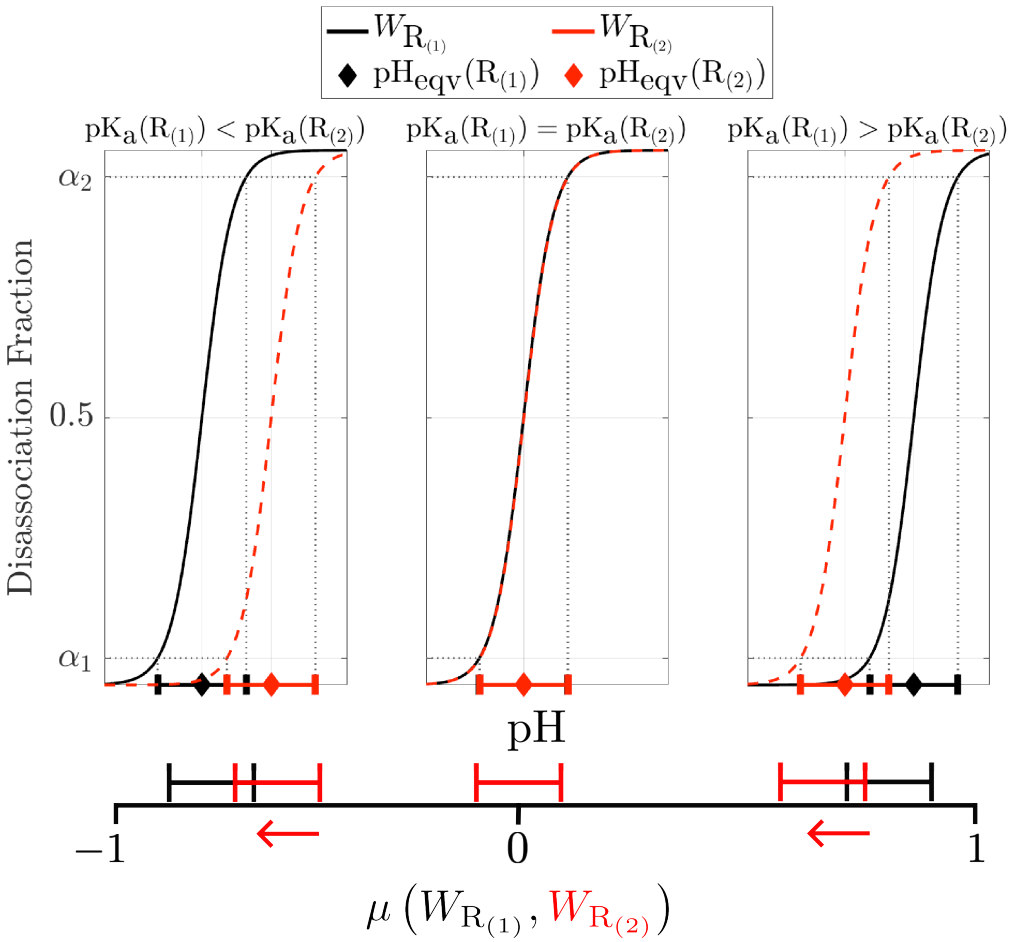


Figure 3.5: Illustration of the overlap parameter μ for two ionisation intervals $W_{\mathrm{R}_{(1)}}$ and $W_{\mathrm{R}_{(2)}}$ belonging to distinct residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ respectively with acidic side chains. The red arrow indicates the direction of movement for $W_{\mathrm{R}_{(2)}}$ as $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$ increases.

residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ are ionised and (ii) the extent of their simultaneous (co)-ionisation. We can quantify both of these factors by defining a signed overlap parameter between ionisation windows, which can be written down as

$$\begin{aligned}\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}}) &= \frac{\mathrm{sgn}[\mathrm{pKa}(\mathrm{R}_{(1)}) - \mathrm{pKa}(\mathrm{R}_{(2)})]\left(2\Delta\mathrm{pH} - \left|W_{\mathrm{R}_{(1)}} \cap W_{\mathrm{R}_{(2)}}\right|\right)}{2\Delta\mathrm{pH}} \\ &= \frac{\mathrm{sgn}[\mathrm{pKa}(\mathrm{R}_{(1)}) - \mathrm{pKa}(\mathrm{R}_{(2)})](\mathrm{pH}_{\mathrm{eqv}}^{\mathrm{max}} - \mathrm{pH}_{\mathrm{eqv}}^{\mathrm{min}})}{2\Delta\mathrm{pH}},\end{aligned} \tag{3.15}$$

where $\mathrm{sgn}[x]$ is the sign of x, conventionally taken to be 0 when $x = 0$, $|[a, b]| \equiv b - a$ denotes the length of the interval $[a, b]$ and $\mathrm{pH}_{\mathrm{eqv}}^{\mathrm{max}} = \max\{\mathrm{pH_{eqv}}(\mathrm{R}_{(1)}), \mathrm{pH_{eqv}}(\mathrm{R}_{(2)})\}$, with $\mathrm{pH}_{\mathrm{eqv}}^{\mathrm{min}}$ defined in an identical way by replacing the $\max\{\dots\}$ operation with $\min\{\dots\}$.

Physical interpretation of values of $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$ for ionisation of two acidic residues is

Table 3.2: Parameters used to describe the ionisation of type II mammalian collagen.

Parameter	Meaning	Parameter Range
$\mu\,(W_{\mathrm{Glu}}, W_{\mathrm{Asp}})$	Ionisation order and co-ionisation extent of Glu, Asp	$[-1, 1]$
$z^{\mathrm{ave}}\,(\mathrm{Glu}, \mathrm{Asp})$	Average normalised surface charge due to Glu and Asp	$[0, 1]$
$\mu\,(W_{\mathrm{Arg}}, W_{\mathrm{Lys}})$	Ionisation order and co-ionisation extent of Arg, Lys	$[-1, 1]$
$z^{\mathrm{ave}}\,(\mathrm{Arg}, \mathrm{Lys})$	Average normalised surface charge due to Arg and Lys	$[0, 1]$

illustrated in Figure 3.5. At $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}}) = \pm 1$, the ionisation windows reach minimum overlap, only sharing a single endpoint. This corresponds to the residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ ionising in sequence one after the other. At $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}}) = 0$ the ionisation windows overlap completely, meaning that the two residue types ionise simultaneously. Monotonic increases in values of $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$ can be interpreted as moving the ionisation window of $\mathrm{R}_{(2)}$ past that of $\mathrm{R}_{(1)}$ in the direction of decreasing pH. Physically, this process can be interpreted as an increase/decrease in simultaneous ionisation provided that $\mu < 0$ and $\mu > 0$ respectively. The aforementioned physical interpretation applies to basic residue ionisation in an identical manner.

For each set of K_{a} values we will vary the pH. The key physical characteristic of a polypeptide containing ionisable residues that is affected by pH is its surface charge. We therefore quantify the changes in pH using the average effective charge of the residues that are undergoing co-ionisation, which is defined as

$$z^{\mathrm{ave}}(\mathrm{R}_{(1)}, \ldots, \mathrm{R}_{(N_{\mathrm{co}})}) = \frac{\sum_{k=1}^{N_{\mathrm{co}}} \left| z^{\mathrm{eff}}(\mathrm{R}_{(k)}) \right|}{N_{\mathrm{co}}}, \tag{3.16}$$

where N_{co} is the number of distinct residues $\mathrm{R}_{(k)}$ that undergo co-ionisation. We take the meaning of co-ionisation to be that the residues $\mathrm{R}_{(k)}$ in equation (3.16) have a non-empty overlap between their ionisation windows, i.e. $\bigcap_k W_{\mathrm{R}_{(k)}} \neq \emptyset$. The parameter $z^{\mathrm{ave}}(\mathrm{R}_{(1)}, \ldots, \mathrm{R}_{(N_{\mathrm{co}})})$ in equation (3.16) can be interpreted as the mean surface charge due to the residues $\mathrm{R}_{(1)}, \ldots, \mathrm{R}_{(N_{\mathrm{co}})}$ normalised by the total number of co-ionising residues.

We now specialise our discussion of ionic parameter variation to a specific class of collagen proteins. Figures 3.11A-E of the Appendix illustrate the average abundances of ionisable residues across all mammalian species in fibrillar collagens of type I, II, III, V and XI respectively. Type II collagen stands out as unique, in that 3 of the 7 ionisable residues Cys, Tyr and His are either completely absent or occur on average once per spiral strip.

Due to the low abundances of the aforesaid residues, the variation in their K_a values is expected to have a negligible effect on the microfibril self-assembly. This is not true for the other collagen types I, III, V and XI, which are all characterised by non-negligible amounts of His. We will focus on studying the microfibril self-assembly of type II collagen, owing to its unique ionisable residue profile, which greatly simplifies systematic variation of ionic parameters. Henceforth, unless specified otherwise, we will illustrate all residue-dependent calculations for type II bovine collagen, which has been previously used in the study of fibrillogenesis and its dependence on ionic conditions [77].

As can be seen in Figure 3.11B, the remaining ionisable residues Asp, Glu, Arg and Lys occur roughly in abundance of 10 to 30 residues per strip and are thus expected to significantly affect the ionisation behaviour of type II collagen. Our goal now is to compute all distinct K_a combinations for Asp, Glu, Arg and Lys. We start by setting a baseline for the ranges of residue K_a values that may be observed in biological systems. Table 3.3 of the Appendix illustrates the average K_a values measured across a dataset of 78 proteins [85]. It is clear that the K_a values of acidic residues Glu and Asp are significantly different from those of basic residues Arg and Lys. We can quantify this difference in terms of the distance along the pH scale between the ionisation windows of acidic and basic residues. Using equation (3.13), we obtain $2\Delta\text{pH} \approx 2.56$, which corresponds to the width of the ionisation window for each residue. It is clear that ΔpH is several pH units smaller than the difference between the average pKa values of acidic residues Asp, Glu and basic residues Arg, Lys. We therefore expect that the acidic and basic residues will ionise independently of each other, and that their ionisations can be treated as independent physical processes. In other words, we can treat type II collagen as having 2 pairs of co-ionising residues - Arg, Lys and Glu, Asp. Based on the definitions of μ and z^{ave} in equations (3.15) and (3.16) respectively, we need 4 independent parameters to describe the ionisation of type II collagen as a function of pH and K_a - see Table 3.2. The detailed numerical procedure used for discretising the aforementioned ionic parameters can be found in subsection 3.7.6 of the Appendix.

To our knowledge, no direct measurement of K_a values have been performed for collagen. It is therefore helpful to define a set of reference equilibrium disassociation constants K_a^{ref}, which reflect the ionisation propensity of a given side chain in absence of site-specific factors that may affect it otherwise. A commonly used experimental system for that purpose are the alanine pentapeptides, which reflect the disassociation equilibrium constants in absence of pairwise residue interactions and side-chain burial [85]. These experimentally determined K_a^{ref} can be found in Table 3.3 of the Appendix.

3.4 RESULTS

3.4.1 PHASE DIAGRAM IN z^{AVE}-μ SPACE

The phase diagram for microfibril self-assembly as a function of pH and residue pKa is shown in Figure 3.6. We note that three distinct phases are observed at thermodynamic equilibrium: single molecular, axially ordered and axially disordered aggregated phases. We remark the absence of an SLS-like phase, which is in good agreement with experiment - presence of other molecules, such as ATP is required for its formation [120]. In Figure 3.12 of the Appendix, we calculate the axial period $\mathcal{D}$ of the axially ordered phase in the z^{ave}-μ plane. The value of $\mathcal{D}$ in the axially ordered phase depends weakly on both the residue pKa and pH, falling narrowly into the interval $[666, 668]$Å and thus always corresponds to a D-banded microfibril. This observation is particularly significant, since in Chapter 2, we found that the most thermodynamically stable microfibril corresponds to an "in-register" conformation, observed experimentally in SLS aggregates. Accounting for charged residue

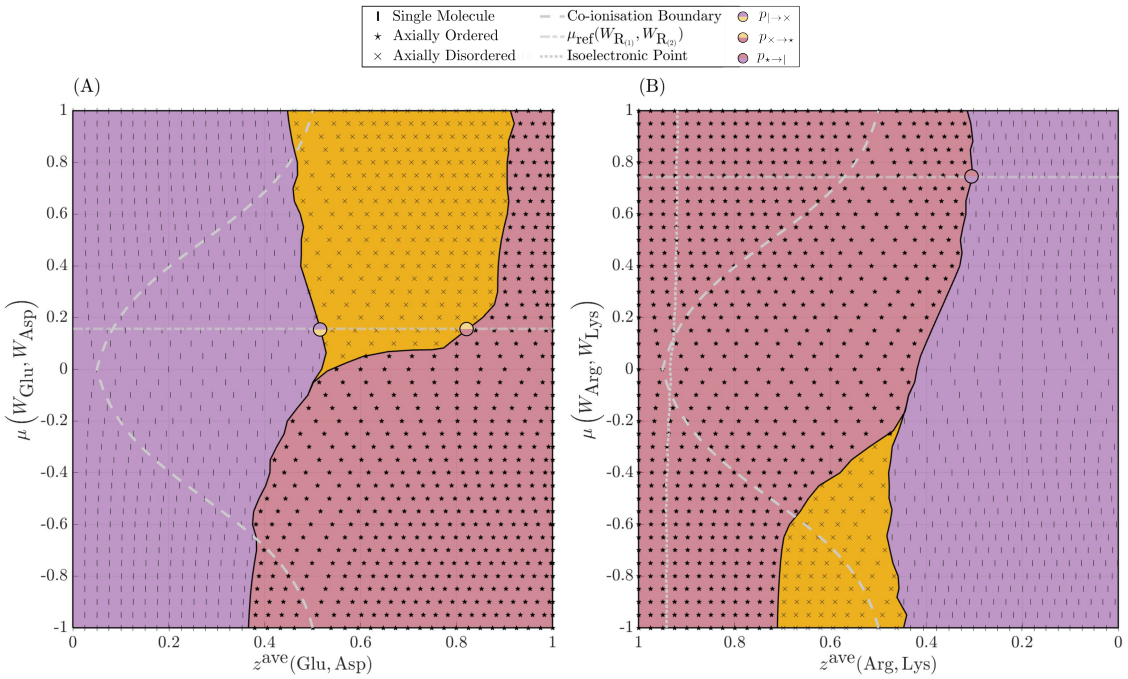


Figure 3.6: Phase diagram for collagen microfibril self-assembly in the z^{ave}-μ space. The phase transitions are shown for pH values corresponding to (A). ionisation of acidic residues and (B). de-ionisation of basic residues. The co-ionisation boundary separates the z^{ave}-μ plane into regions to the left and right of the boundary, where only one of the residues is ionised and both residues co-ionise respectively. The reference ionisation window overlap μ_{ref} is calculated based on the pKa values measured in alanine pentapeptides - see Table 3.3.

interactions has resolved this issue. We see this reflected in the most energetically stable global energy minimum among the pairwise interaction potentials $U^{p}_{i\text{-}j}$ in Figure 3.13 of the Appendix. This minimum now corresponds to $\Delta z = \mathrm{D}$, which is in stark contrast to Figure 2.6 from Chapter 2, in which we saw that the minimum at $\Delta z = 0$ was the most energetically stable. We now turn out attention to the detailed analysis of the phase diagram.

Figure 3.6A shows the phase diagram corresponding to increasing the pH from acidic towards isoelectronic. Two self-assembly regimes can be distinguished based on the value of the overlap parameter μ. We start by analysing the upper half-plane $\mu(W_{\mathrm{Glu}}, W_{\mathrm{Asp}}) \gtrsim 0$, which corresponds to Glu starting to ionise after Asp in the direction of increasing pH. In the upper half plane, we observe two phase transitions. For net ionisation fractions $z^{\mathrm{ave}}(\mathrm{Glu}, \mathrm{Asp}) \lesssim 0.5$, collagen exists in a single molecule phase. At $z^{\mathrm{ave}}(\mathrm{Glu}, \mathrm{Asp}) \approx 0.5$, a phase transition from single molecule to an axially disordered aggregated phase occurs. As the average effective charge of Glu and Asp continues to increase with the pH, a second phase transition occurs from an axially disordered to an axially ordered aggregated phase. Calculation of the overlap parameter using the reference equilibrium disassociation constants $K_{\mathrm{a}}^{\mathrm{ref}}$ yields $\mu_{\mathrm{ref}}(W_{\mathrm{Glu}}, W_{\mathrm{Asp}}) \approx 0.16$. Interpreting $K_{\mathrm{a}}^{\mathrm{ref}}$ as a first approximation of the actual K_{a} values of acidic residues Asp and Glu in molecular collagen, the predictions of our model agree with the phase transitions observed experimentally during collagen aggregation between acidic and isoelectronic pH [51, 15]. In the lower half-plane $\mu(W_{\mathrm{Glu}}, W_{\mathrm{Asp}}) \lesssim 0$, the ionisation order is reversed with Asp starting to ionise after Glu in the direction of increasing pH. In this region of the phase diagram there is one phase transition, namely between single molecule and axially ordered aggregated phases. This self-assembly regime, to our knowledge, has not been explicitly documented experimentally.

Figure 3.6B illustrates the phase diagram that corresponds to increasing pH towards and past isoelectronic to basic. We again observe two distinct aggregation regimes, which mimic those seen in Figure 3.6A. The region of the phase diagram corresponding to $\mu(W_{\mathrm{Arg}}, W_{\mathrm{Lys}}) \gtrsim -0.24$ is characterised by a single phase transition from an axially ordered aggregated phase to a single molecule phase. The reference overlap parameter $\mu_{\mathrm{ref}}(W_{\mathrm{Arg}}, W_{\mathrm{Lys}}) \approx 0.74$, falls firmly in this region of the phase diagram. The predictions of our model therefore again agree with the experimental data, since the disordered aggregated phase, to our knowledge, has only been reported at acidic pH conditions [51, 15]. Our model predicts that an axially disordered microfibrillar phase may exist as an intermediate phase between axially ordered and single molecular phases, provided that $\mu(W_{\mathrm{Arg}}, W_{\mathrm{Lys}}) \lesssim -0.24$. However, given the significantly higher value of $\mu_{\mathrm{ref}}(W_{\mathrm{Arg}}, W_{\mathrm{Lys}})$, it is unlikely that this aggregation regime can be observed experimentally.

3.4.2 PHASE TRANSITION MECHANISMS

We now address the physical mechanisms of the phase transitions described in the previous section. For convenience, we will denote the phase transitions in Figure 3.6 using symbolic notation "$\ldots \rightarrow \ldots$", where "$\ldots$" on the left and the right of the arrow are replaced by the symbolic representations of the phases in Figure 3.6 to the left and right of the phase boundary respectively. The coordinates of the points in the z^{ave}-μ space that lie on the phase boundary are denoted using $(z^{\text{ave}}_{\ldots\rightarrow\ldots}, \mu)$.

Figure 3.7 illustrates the evolution of $\lim_{L_{\text{M}}\rightarrow\infty}\mathcal{P}_{\text{period}}(g)$ and $\text{E}^{\text{min}}_{\text{M}}$ as a function of average effective charge of acidic and basic residues for two fixed values of the overlap parameter that are closest to the reference values $K^{\text{ref}}_{\text{a}}$: $\mu(W_{\text{Glu}}, W_{\text{Asp}}) = 0.15$ for acidic

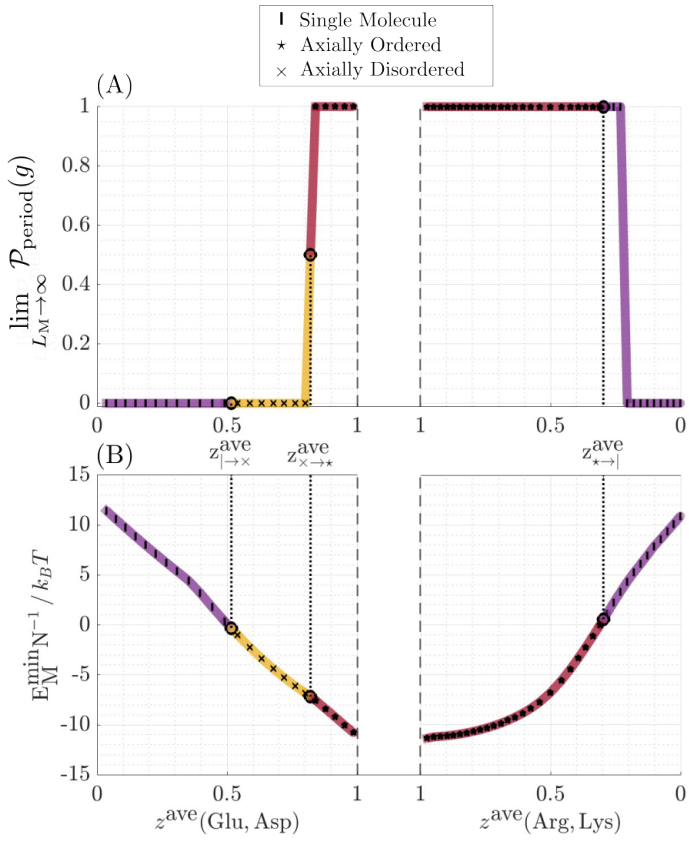


Figure 3.7: (A). Dependence of the equilibrium probability of axially ordered microfibrils on the mean charge of acidic/basic residues (z^{ave}) in the limit of an infinite microfibril. (B). The minimum microfibrillar energy as a function of z^{ave}. Dotted lines connecting to circled points show the location of the phase transitions. Dashed lines separate out the ionisation of acidic and basic residues. The results are illustrated for an overlap parameter values of $\mu(W_{\text{Glu}}, W_{\text{Asp}}) = 0.15$ for the two acidic residues and $\mu(W_{\text{Arg}}, W_{\text{Lys}}) = 0.75$ for the two basic residues.

residues and $\mu(W_{\mathrm{Arg}}, W_{\mathrm{Lys}}) = 0.75$ for the basic residues. We start by analysing the onset/offset of microfibril aggregation, which is determined entirely by the value of $\mathrm{E}_{\mathrm{M}}^{\mathrm{min}}$. Figure 3.7B shows that the microfibril energy monotonically decreases with increasing average effective charge of acidic/basic residues. Increasing the maximum strength of charged residue interactions is therefore favourable for collagen aggregation, allowing for eventual formation of axially disordered/ordered aggregated phases. This is in good agreement with existing knowledge in literature [51, 77].

Next, we turn to the axially disordered to axially ordered phase transition that occurs at the phase boundary point $p_{\times\to\star} = (z^{\mathrm{ave}}_{\times\to\star}(\mathrm{Glu}, \mathrm{Asp}), 0.15)$ in Figure 3.6A. To aid us in understanding the mechanism of the phase transition, we consider the microfibril energies $\mathrm{min}_{\mathrm{S_{eq}}}\{\mathrm{E_M}(g)\}$ and equilibrium probabilities $\mathcal{P}_{\mathrm{period}}(g)$ and of individual microfibril segments, see Figure 3.8. Together, these two quantities enable us to gauge the thermodynamic stability as well as the propensity of a microfibril with a given fixed gap size g to form axially ordered aggregates.

The emergence of a thermodynamically stable axially ordered phase can then be understood as a two-stage process. (i). The onset of ionisation of Glu and Asp to the left of the phase boundary point $p_{|\to\times}$ leads to the emergence of several microfibrils with different gap sizes, which correspond to the local minima in the $\mathrm{min}_{\mathrm{S_{eq}}}\{\mathrm{E_M}(g)\}$ profile - see Figure 3.8A. These microfibrils can be viewed as locally stable w.r.t. gap size perturbations. Gap sizes for which $\mathcal{P}_{\mathrm{period}}(g) \to 1$ already exist, however gap sizes with $\mathcal{P}_{\mathrm{period}}(g) \to 0$ result in more thermodynamically stable microfibrils, leading to formation of axially disordered aggregates at acidic pH. (ii). Continued ionisation of Asp and Glu leads to level-crossing at $p_{\times\to\star}$, and energetic selection of the microfibril with the gap size for which $\mathcal{P}_{\mathrm{period}}(g) \to 1$ - see Figure 3.8B. The gap size corresponding to the lowest energy NEqS $\vec{s}_{\mathrm{min}}$ at the phase space point $p_{\times\to\star}$ is found to be $g_{\mathrm{min}} = 33.2\,\mathrm{nm}$, corresponding to an axial period of $\mathcal{D} \approx 66.6\,\mathrm{nm}$, in good agreement with the experimental literature [77]. The discontinuity seen in Figure 3.7 therefore results from pH-induced level-crossing of two near-equilibrium states corresponding to microfibrils with different gap sizes. The effect of the increase in electrostatic interactions across the phase boundary point $p_{\times\to\star}$ is therefore selectively stabilising w.r.t. the axially ordered phase, as opposed to the axially disordered phase.

As was seen in Figure 3.6, for physiologically relevant K_{a} values, the axially disordered phase is observed only at acidic pH. At increasingly basic pH, de-ionisation of Arg and Lys does not lead to a level-crossing phase transition described above, as can be seen in Figure 3.8C. As a result, we observe a phase transition from an axially ordered to a single molecular phase at basic pH. The existence of a level crossing phase transition between axially ordered and axially disordered phases is thus dependent on both the residue ionisation order (determined by pKa values) as well as on the specific residues that are undergoing ionisation (controlled by pH).

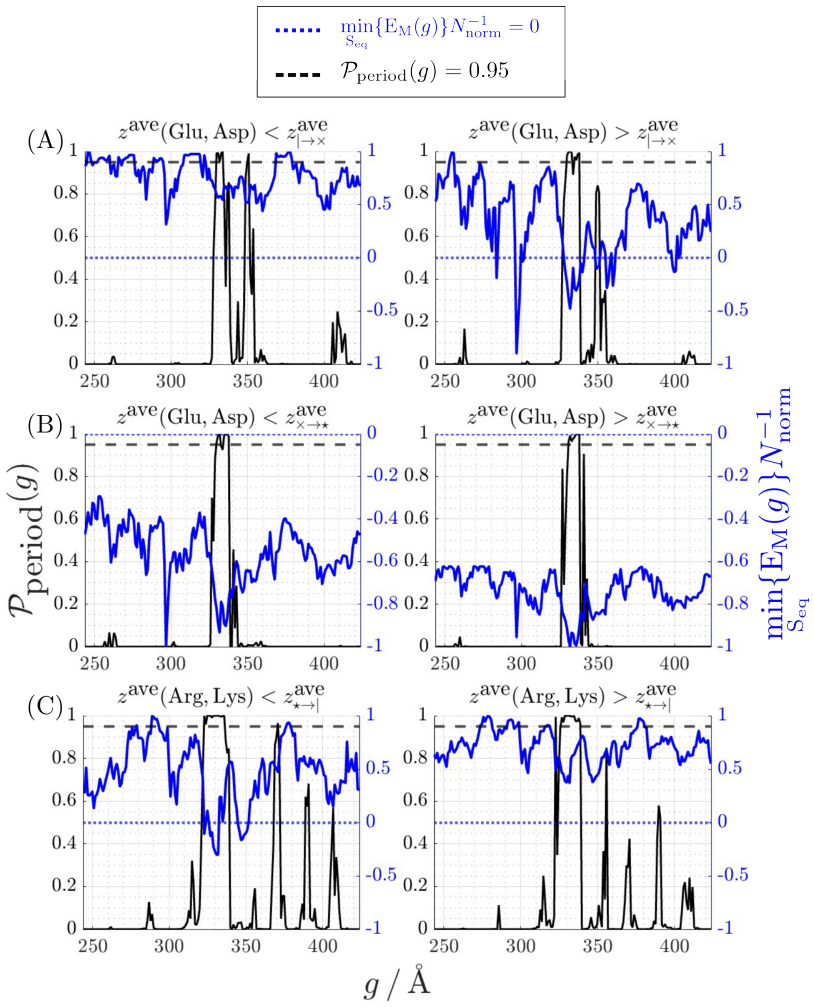


Figure 3.8: Gap size dependence of the equilibrium probability of axially ordered microfibrils for a single microfibril segment and normalised microfibril energy of the most stable NEqS. Within each row, the profiles are calculated on either side of the relevant phase boundary. Row (A). from single molecular (LEFT) to disordered aggregated phase (RIGHT), row (B). from disordered to axially ordered aggregated phase and row (C). from axially ordered to single molecular phase. The overlap parameter is fixed at $\mu(W_{\mathrm{Glu}}, W_{\mathrm{Asp}}) = 0.15$ for panels (A) and (B) and at $\mu(W_{\mathrm{Arg}}, W_{\mathrm{Lys}}) = 0.75$ for panel (C). N_{norm} normalises the microfibril energy to lie in the intervals $[-1, 1]$ and $[-1, 0]$ for panels (A), (C) and (B) respectively.

3.4.3 Role of Charged Residue Interactions in Self-Assembly

To infer the roles of specific charged residue interactions in microfibril self-assembly, we start by analysing the relative positions of the phase boundaries and the co-ionisation boundary $\widetilde{\alpha}(\mu)$ in Figure 3.6 (see subsection 3.7.3 of the Appendix for the derivation). For a given value of $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$, the co-ionisation boundary splits the phase plane into two parts. Whenever $\alpha_1 \leq z^{\mathrm{ave}}(\mathrm{R}_{(1)}, \mathrm{R}_{(2)}) \leq \widetilde{\alpha}(\mu)$ for a pair of acidic residues and

$\widetilde{a}(\mu) \leq z^{\text{ave}}(\mathrm{R}_{(1)}, \mathrm{R}_{(2)}) \leq \alpha_2$ for a pair of basic residues, just one of the residues $\mathrm{R}_{(1)}$, $\mathrm{R}_{(2)}$ is undergoing ionisation/de-ionisation respectively. In Figures 3.6A, B, these regions are located to the left of the co-ionisation boundary. To the right of $\widetilde{\alpha}(\mu)$, both residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ ionise simultaneously.

We focus on the region of the phase diagram with extremal values of the overlap parameter $\mu(\mathrm{R}_{(1)}, \mathrm{R}_{(2)}) = \pm 1$, which physically correspond to the sequential ionisation of residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$. "Sequential" in this context means that the ionisation of residue $\mathrm{R}_{(2)}$ starts once ionisation of residue $\mathrm{R}_{(1)}$ has finished and vice versa. Sequential ionisation processes serve as simple models for investigating the response of collagen self-assembly to the removal of specific charged residue interactions from the net pairwise intermolecular energy. For example, in Figure 3.6A, at $\mu(W_{\text{Glu}}, W_{\text{Asp}}) = -1$, we notice that $z^{\text{ave}}_{|\to\star}(\text{Glu}, \text{Asp}) < \widetilde{a}(\mu)$. Physically, this implies that complete deionisation of Asp does not destabilise the axially ordered phase. On the other hand, at $\mu(W_{\text{Glu}}, W_{\text{Asp}}) = 1$, we see that $z^{\text{ave}}_{\times\to\star}(\text{Glu}, \text{Asp}) > \widetilde{a}(\mu)$. In this instance, the relative positions of the phase and co-ionisation boundaries indicate that deionising $\approx 15\%$ of Glu is sufficient to destabilise the axially ordered phase. Collectively, the aforementioned results show that charged residue interactions involving Glu are crucial for the stabilisation of the axially ordered phase, whilst those involving Asp are not.

We can understand aforementioned results better by analysing the behaviour of maximum (across all gap sizes) equilibrium probability of the axially ordered phase across the z^{ave}-μ plane. In Figure 3.9A, for $\mu(W_{\text{Glu}}, W_{\text{Asp}}) \gtrsim 0.3$, we observe a significantly reduced propensity for formation of the axially ordered phase across all gaps sizes. We notice that along the line $\mu(W_{\text{Glu}}, W_{\text{Asp}}) = 1$, the minimum attained by $\max_g\{\mathcal{P}_{\text{period}}(g)\}$ coincides closely with $\widetilde{\alpha}(\mu)$. Noting that in this case the co-ionisation boundary separates regions of independent ionisations of Asp and Glu to the left and the right respectively, we infer that the sum of interactions of Asp with Arg, Lys and itself is destabilising towards the axially ordered phase. On the other hand, the net effect of Glu interactions with Arg, Lys, Asp and itself is stabilising towards formation of an axially ordered phase.

We now turn to sequential ionisation of basic residues Arg and Lys in Figure 3.6B. Looking at $\mu(W_{\text{Arg}}, W_{\text{Lys}}) = 1$, we see that $z^{\text{ave}}_{\star\to|}(\text{Arg}, \text{Lys}) < \widetilde{\alpha}(\mu)$. Physically, this means that with all Lys residues unable to partake in charged interactions, collagen remains in the axially ordered phase. On the other hand, at $\mu(W_{\text{Arg}}, W_{\text{Lys}}) = -1$, we see that $z^{\text{ave}}_{\star\to\times}(\text{Arg}, \text{Lys}) > \widetilde{\alpha}(\mu)$. In particular, an effective charge of less than 0.45 for Arg residues is sufficient to destabilise the axially ordered phase and initiate the phase transition to an axially disordered phase. We see that the sufficiency conditions for the self-assembly of the axially ordered phase in terms of basic residue ionisation are analogous to those of acidic residues. Presence of sufficiently high amounts of ionised Arg is required for self-assembly of the axially ordered phase, whilst the presence of ionised Lys is not required at all. In

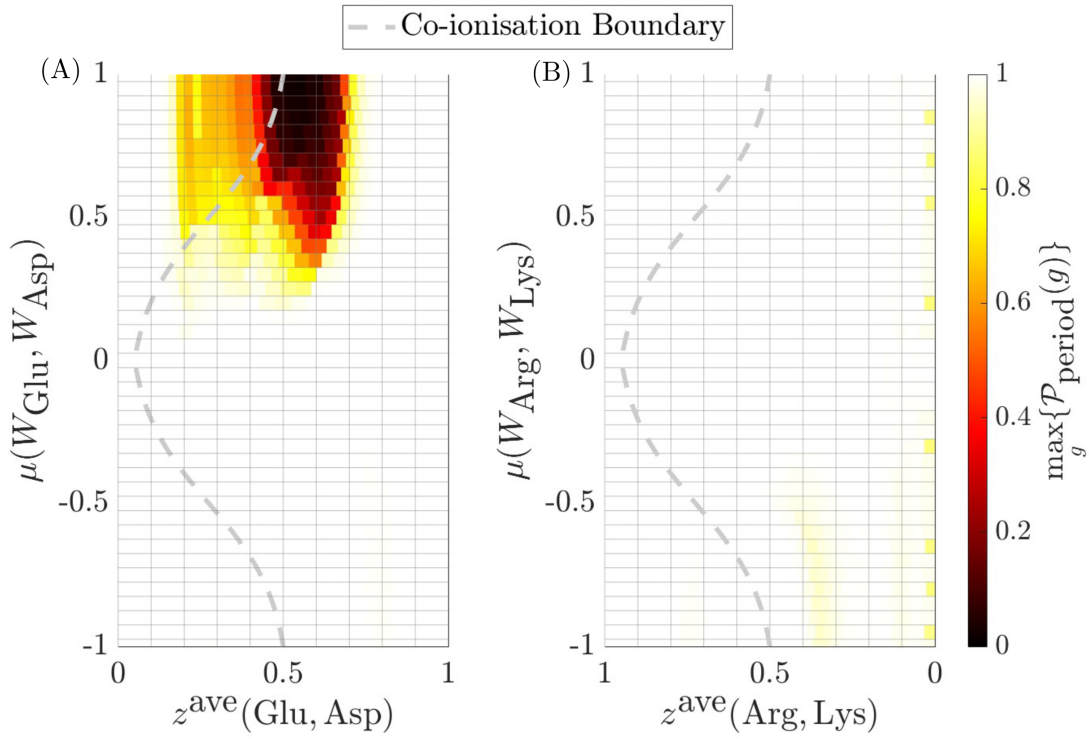


Figure 3.9: Maximum equilibrium probabilities of axially ordered microfibrils across all gap sizes in the z^{ave}-μ space for a single microfibril segment. The data is illustrated for (A). ionisation of acidic residues and (B). de-ionisation of basic residues.

contrast to acidic residues however, changing the relative proportion of charged Arg or Lys side chains does not have a significant adverse impact on the propensity of microfibrils to form the axially ordered phase - see Figure 3.9B.

3.5 DISCUSSION

Collagen self-assembly is intrinsically a multi-scale process in both space and time, with different physical mechanisms driving self-assembly at different levels of the collagen structural hierarchy. Experimental measurements of enthalpic and entropic contributions towards fibrillogenesis indicate that formation of D-banded fibrils is an entropy-driven process [27, 59]. These experimental results have led to a number of successful theoretical treatments of fibrillogenesis that use entropy-driven mechanisms to describe the concentration-dependent phase transition between molecular collagen and an aggregated fibrillar phase [115, 116]. Such approaches however offer no physical insight into whether the aggregated phase exhibits axial order. Axial order in aggregated collagenous phases exhibits strong dependence on the ionic conditions. At pH $\approx$ 2.5-3.5 collagen forms an

axially disordered phase, which lacks strict axial ordering that is associated with D-banding [51, 15]. At pH ≈ 4 the axially disordered phase transitions to an axially ordered phase with a well-defined D ≈ 67 nm axial periodicity [77].

In this work, we demonstrated that axial order in the aggregated collagenous phase emerges at the level of the microfibril and is mediated by electrostatic interactions. The notion that electrostatic interactions mediate collagen self-assembly has been previously suggested by other authors [102, 65, 77]. Studies of pairwise molecular interactions between collagen triple helices further indicate that charged residue interactions are maximised at D-staggered molecular arrangements [53]. However, the role of charged residue interactions in controlling the pH-dependent emergence of axial order, which manifests itself in existence of an axially ordered and disordered aggregated phases, has not been considered. To study the emergence of axial order in collagen aggregates, we used an equilibrium model for microfibrillar self-assembly, which incorporated dependence on the ionic conditions of specific residue-residue interactions as well as the 3-dimensional spatial residue organisation. Focusing on type II collagen, owing to its unique residue composition, we found that collagen microfibrils exhibit the same pH-dependent aggregation regime as the larger fibrillar aggregates observed experimentally [51, 15] - see Figure 3.6A. Namely, as pH increases from acidic to isoelectronic, a phase transition from the single molecular to disordered microfibril phase occurs, followed by a transition from a disordered to an axially ordered phase with a ≈ 67 nm axial banding pattern.

Our prediction of a thermodynamically stable axially ordered phase from pairwise collagen-collagen interactions is also significant in its own right. A number of previous theoretical studies of pairwise triple helix interactions have indicated that in-register association of triple helices is energetically preferred over D-staggering [53, 92]. Experimental evidence indicates, however, that in-register self-assembly observed in segment-long-spacing (SLS) aggregates typically necessitates additional charged interactions between collagen and molecules like ATP or diazo dyes, which are not accounted for in the aforementioned models [50, 120]. Other recent studies [95] have avoided addressing this problem by constraining axial staggers in an N-membered microfibril to lie in the interval $\Delta z \in [\frac{L}{\mathrm{N}}, \frac{L}{\mathrm{N}-1}]$. Our own study [124], presented in chapter 2, found that whilst molecular interactions at $\Delta z \approx 0$ were labile to perturbations in residue-residue interactions, they were nevertheless energetically the most favourable. The results of the current chapter suggest that this is an artifact of using models that poorly represent the residue-residue interactions between collagen triple helices. In particular, a separate and careful treatment of charged-charged residue interactions is necessary, as they appear to play a critical role in destabilising the in-register axial configuration and instead favouring a D-staggered arrangement, compare Figures 2.6A, B and Figures 3.13A, B. We note that Figure 3.13A suggests that thermodynamically stable in-register microfibrils are theoretically possible, however we do not

observe them at thermodynamic equilibrium, see Figure 3.6. In line with experimental evidence, this suggests the importance of interactions between collagen triple helices and charged molecules such as diazo dyes which have been shown to trigger a transition from an axially ordered aggregated phase to an SLS phase [50]. Understanding the molecular interactions that drive the aforementioned phase transition presents an interesting direction for future research.

Returning back to the transitions between axially ordered and disordered phases, we established the mechanism that controls the emergence of axial periodicity at the microfibrillar level. We found that microfibrils with different gap sizes g have different propensities for formation of the axially ordered phase, reflected in the values of the equilibrium probability of forming axially ordered aggregates $\mathcal{P}_{\text{period}}(g)$ in Figure 3.8. Some values of gap size lead to preferential aggregation into an axially disordered, rather than an axially ordered phase. We found that the electrostatic interactions significantly affect the relative energies of microfibrils with different gap sizes. Changes in pH can therefore lead to phase transitions between axially ordered and disordered aggregated phases - see Figure 3.8B. Partial ionisation of acidic residues Asp and Glu energetically favours microfibrils with the gap size that has a high propensity for formation of axially disordered microfibrils. On the other hand, complete ionisation of Asp and Glu energetically selects the gap size for which axially ordered microfibrils self-assemble at equilibrium.

Simultaneously with the effect of pH, we studied the influence of residue pKa on microfibril self-assembly - see the phase diagram in Figures 3.6A and 3.6B. Our model predicted that the microfibril self-assembly regime has a strong dependence on the relative pKa values of Asp and Glu. We identified two distinct pKa-dependent aggregation regimes. The first regime is characterised by pKa(Asp) $<$ pKa(Glu). Collagen microfibrils aggregate in the manner observed experimentally [51], existing either in a single molecule, disordered or axially ordered phase depending on the pH. The second regime entails no disordered phase and occurs whenever pKa(Asp) $>$ pKa(Glu). In this case, collagen microfibrils transition directly from a single molecular to an axially ordered phase. The average experimentally measured pKa values for Asp and Glu shown in Table 3.3, suggest that both aggregation regimes are physically feasible. Our findings indicate that systematic shifts in pKa values across multiple Glu, Asp residues can alter the microfibril aggregation regime. This raises questions regarding the biological role of different collagen aggregation regimes as well as their prevalence across different biological environments. To our knowledge, no experimental or theoretical study has been conducted that estimated site-specific pKa variation in collagen, thus providing an exciting avenue for further work.

We found that changing the pKa values of basic residues Arg and Lys altered the self-assembly regime of collagen in a similar way to that described for the acidic residues above. In experiments, only 2 phases are observed past the isoelectronic pH - the axially

ordered and the single molecular phase [51, 77]. We identified a region in the phase diagram in Figure 3.6B, in which the disordered phase exists at basic pH. This however requires Arg to have a pKa value substantially lower than that of Lys, which does not appear feasible, based on the average pKa values from Table 3.3 measured across different proteins.

Using the phase diagram in Figure 3.6, we ascertained the roles of specific charged residues in controlling the self-assembly of axially ordered microfibrils. We showed that the net effect of interactions of Asp with Arg, Lys and itself was globally destabilising towards formation of an axially ordered phase. We saw this reflected in significant reduction of maximum equilibrium probabilities associated with axially ordered microfibrils across all studied gap sizes in Figure 3.9A. On the other hand, the net effect of interactions of Glu with Arg, Lys, Asp and itself was found to be favourable for self-assembly of axially ordered microfibrils. As a consequences of this, we found that complete inability of Asp to partake in charged-charged interactions did not preclude formation of an axially ordered phase. On the other hand, reducing the fraction of ionised Glu by as little as 15% was sufficient to destabilise the axially ordered phase. For basic residues, we identified a similar relationship - reducing the fraction of ionised Arg by $\approx 55\%$ was sufficient to make the axially ordered phase thermodynamically unstable. On the other hand, reducing the fraction of ionised Lys did not lead to destabilisation of the axially ordered phase in favour of single molecular or axially disordered phases.

Throughout this work, we have focused on studying the impact of ionic conditions on the self-assembly of type II collagen. The primary reason for this choice is the unique residue composition of this collagen type, which enables one to characterise its full spectrum of ionisation behaviour by varying the pKa values of just 4 ionisable residues - Asp, Glu, Lys and Arg. Other fibrillar collagens of types I, III, V and XI necessitate accounting for variable pKa of His, which is found in negligible amounts in type II collagen - see Figure 3.11B. Ascertaining the role of His ionisation in the self-assembly of other fibrillar collagens presents an exciting avenue for application of the model introduced in this work.

Another area of biological significance that our model can provide insight into, is the study of detailed contributions of charged residues towards thermodynamic stability of axially ordered collagen aggregates. *In vivo*, the ability of Arg and Lys side chains to partake in electrostatic interactions may become diminished by irreversible modification of their side chains via glycation. These side chain modifications become more pronounced with ageing and in the presence of chronic diabetes. Continued glycation of molecular collagen has been shown to cause molecular reorganisation in collagen fibrils as well as to induce molecular stiffening and eventually preclude molecular collagen's ability to form any aggregated phase [5, 103]. Our results suggest that formation of sugar adducts with the guanidinium side chain of Arg may significantly contribute to inhibiting the formation of aggregated collagenous phases. An exciting avenue for extending these predictions,

would be to consider the influence of spatially and sequence dependent electrostatic effects. Throughout this work, we have assigned the same effective charge to residues with identical side chains. Glycation, however, has been shown to target residues in specific sequence positions [114]. Site-specific modification of charged residue side-chains may have different effects on collagen self-assembly, depending on whether the modified residue side chains ordinarily partake in intra- or inter-molecular electrostatic interactions, potentially leading to rich and complex behaviour. Our model offers an easily adaptable, computationally tractable and physically grounded method for future quantitative study of the effects of site-specific residue modifications on collagen self-assembly across different biological scenarios.

3.6 CONCLUSION

In this study, we investigated the role of ionic conditions as well as specific charged residue interactions in controlling the emergence of axial molecular order in aggregated collagen phases. Focusing on type II collagen, we used our equilibrium collagen self-assembly model to construct a phase diagram illustrating the impact of key ionic parameters, pH and residue pKa, on collagen aggregation. We were able to ascertain the physical origin of the disordered aggregated phase that has previously been reported in fibrillogenesis experiments [51, 15]. We argued that the disordered aggregated phase corresponds to thermodynamically stable collagen microfibrils with a gap size that is different from that observed in the axially ordered (D-banded) phase. Furthermore, our analysis indicated that the order and the extent of residue co-ionisation, controlled by the residues' pKa, is a crucial parameter that determines the existence of the axially disordered aggregated phase. Finally, we identified the presence of ionised Arg and Glu as a crucial factor in ensuring the thermodynamic stability of the axially ordered aggregated phase. On the other hand, we showed that the absence of either ionised Lys or Asp did not inhibit the formation of a stable axially ordered aggregated phase.

3.7 APPENDIX

3.7.1 Thermodynamics of Residue Ionisation

The contents of this section are based primarily on [106], with suitable notational changes where appropriate. The dependence of the residue charge on the ionic parameters can be obtained from the definition of the acid disassociation constant. Given a residue R, its

acid disassociation constant is given by

$$K_{\mathrm{a}}(\mathrm{R}) = \begin{cases} \dfrac{\{\mathrm{H_3O^+}\}\{\mathrm{R^-}\}}{\{\mathrm{RH}\}}, & \text{if } \mathrm{RH} + \mathrm{H_2O} \rightleftharpoons \mathrm{R^-} + \mathrm{H_3O^+}, \\ \dfrac{\{\mathrm{H_3O^+}\}\{\mathrm{R}\}}{\{\mathrm{RH^+}\}}, & \text{if } \mathrm{RH^+} + \mathrm{H_2O} \rightleftharpoons \mathrm{R} + \mathrm{H_3O^+}, \end{cases} \tag{3.17}$$

where $\{\mathrm{R}\}$ denotes the thermodynamic activity of the residue. The K_{a} values can be experimentally measured and are typically reported in the form $-\log_{10} K_{\mathrm{a}}(\mathrm{R}) \equiv \mathrm{pKa}(\mathrm{R})$. The pKa values of the 7 ionisable residues present in collagen are shown in Table 3.3.

The thermodynamic activity of R is proportional to its concentration [R] via the activity coefficient $f_{\mathrm{R}} = \{\mathrm{R}\}/[\mathrm{R}]$. The activity of a residue is dependent on the ionic strength I of the solution and in the regime $I \lesssim 100\,\mathrm{mM}$ an explicit expression can be derived directly from Debye-Hückel theory [106]. To extend the validity of the approximation to physiological $I \approx 150\,\mathrm{mM}$ we use the Davies semi-empirical approximation, which in the regime $I \lesssim 500\,\mathrm{mM}$ states that

$$\log_{10} f_{\mathrm{R}} = -A z_{\mathrm{R}}^2 \left(\frac{\sqrt{I}}{1+\sqrt{I}} - 0.2I \right), \tag{3.18}$$

where $A = 1.82 \times 10^6 (\epsilon_r T)^{-\frac{3}{2}}$, T is the temperature and ϵ_r is the relative permittivity of water.

To determine the fraction of residue R that exists in the ionised state at a given set of ionic conditions, we define its disassociation fraction

$$\alpha(\mathrm{R}) = \begin{cases} \dfrac{[\mathrm{R^-}]}{[\mathrm{R^-}] + [\mathrm{RH}]}, & \text{if R has an acidic side chain,} \\ \dfrac{[\mathrm{R}]}{[\mathrm{R}] + [\mathrm{RH^+}]}, & \text{if R has a basic side chain.} \end{cases} \tag{3.19}$$

Taking the logarithm of both sides of equation (3.17) and denoting $\mathrm{pH} \equiv -\log_{10}\{\mathrm{H_3O^+}\}$, we obtain

$$\alpha(\mathrm{R}) = \begin{cases} \dfrac{1}{1 + 10^{\mathrm{pKa(R)} - \mathrm{pH} + f_{\mathrm{R^-}}}}, & \text{if R has an acidic side chain,} \\ \dfrac{1}{1 + 10^{\mathrm{pKa(R)} - \mathrm{pH} - f_{\mathrm{RH^+}}}}, & \text{if R has a basic side chain,} \end{cases} \tag{3.20}$$

The full expression for the free energy of interaction between charged residues $\mathrm{R_p}$ and $\mathrm{R_q}$ can then be written down as

$$\Psi_{\mathrm{pq}}^{\mathrm{DH}}(r) = \begin{cases} \dfrac{z^{\mathrm{eff}}(\mathrm{R_p}) z^{\mathrm{eff}}(\mathrm{R_q}) l_B e^{\kappa(a_0 - r)}}{r(1+\kappa a_0)} k_B T, & \text{for } r > a_0, \\ \dfrac{z^{\mathrm{eff}}(\mathrm{R_p}) z^{\mathrm{eff}}(\mathrm{R_q}) l_B}{a_0(1+\kappa a_0)} k_B T, & \text{for } r \le a_0, \end{cases} \tag{3.21}$$

Table 3.3: Values of equilibrium constants for side-chain ionisation reactions of ionisable residues present in collagen.

Residue	[a]$\mathrm{pKa}^{\mathrm{ref}}$ **in alanine pentapeptides**	[b]**Average Value of** pKa
Arg	12.3	[c]$12.3 \ / \ 13.8 \pm 0.1$
Asp	3.9	3.5 ± 1.2
Cys	8.6	6.8 ± 2.7
Glu	4.3	4.2 ± 0.9
His	6.5	6.6 ± 1.0
Lys	10.4	10.5 ± 1.1
Tyr	9.8	10.3 ± 1.2

[a] We denote $-\log_{10} K_{\mathrm{a}}^{\mathrm{ref}} \equiv \mathrm{pKa}^{\mathrm{ref}}$, where the subscript refers to the fact that the disassociation reactions are understood to occur in an alanine pentapeptide.
[b] The values are averaged over all instances of a given ionisable residue across a database of 78 different proteins [85].
[c] To our knowledge, no measurements of Arg acid disassociation constant across different proteins and amino acid environments have been reported. Direct experimental measurements of the K_{a} value of Arg report two distinct values - 12.3 and 13.8 [85, 39].

where $l_B = e^2/(4\pi\epsilon_0\epsilon_r k_B T)$ is the Bjerrum length, ϵ_0 is the vacuum permittivity, $\kappa = \sqrt{8\pi l_B I}$ is the inverse Debye length and z^{eff} depends on the disassociation fraction α according to equation (3.2) of the main text.

3.7.2 INTERACTION INTERFACE COORDINATE TRANSFORMATIONS

Computation of the pairwise molecular interaction energy $\mathrm{U}_{i\text{-}j}(\Delta z)$ necessitates constructing an interaction interface between strips i and j. Interactions between chiral helical strips induce torques on the interacting collagen molecules. We have previously shown that the strength of inter-strip interactions is sufficiently high to enable the collagen triple helices to bend and twist in order to enable the interacting strips to face each other along a common axis [124]. The molecular conformation that minimises the energy of elastic deformation and enables formation of a strip-strip interface is a right-handed molecular supercoil [79, 124]. We will approximate the interaction interface in this molecular geometry with a simple coordinate transformation, without introducing additional geometrical parameters into the model, such as the supercoiling radius. We will do so by transforming the coordinates of the residues lying on strip j. Denote the position vector of the q^{th} residue lying on strip j by ${}^{\mathrm{q}}\mathbf{x}_j$. All coordinates are given in the standard cylindrical basis $\{\mathbf{e}_\rho, \mathbf{e}_\theta, \mathbf{e}_z\}$. The staggered coordinates of ${}^{\mathrm{q}}\mathbf{x}_j$ are prescribed by

$$ {}^{\mathrm{q}}\mathbf{x}_j^{\mathrm{s}}(\Delta z) = {}^{\mathrm{q}}\mathbf{x}_j + 2\pi h^{-1} \left(\Delta z + c_j - c_i\right) \mathbf{e}_\theta + \Delta z \mathbf{e}_z, \tag{3.22} $$

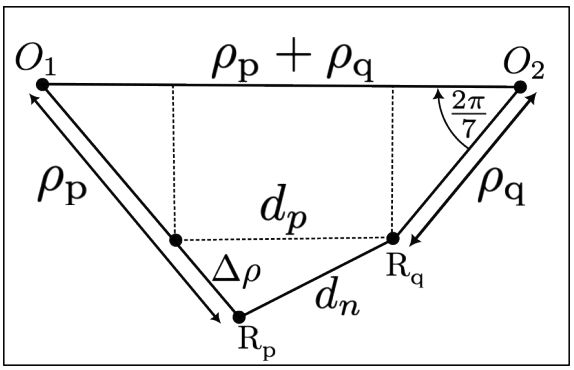


Figure 3.10: Geometric parameters used in calculating nearest neighbour strip separation.

where c_i are the constants that define the centreline equations of the strips given by $z = h\theta(2\pi)^{-1} + c_i$, where h is the pitch of the strip.

For the interaction interface between nearest neighbour strips, we need to account for the additional spatial separation between the residues. As shown in Figure 3.10, consider two residues $\mathrm{R_p}$ and $\mathrm{R_q}$ which lie on strips i'' and j' respectively. We denote the radii of $\mathrm{R_p}$ and $\mathrm{R_q}$ as ρ_p and ρ_q relative to their respective molecular axes O_1 and O_2. The molecular axes are separated by a perpendicular distance $\rho_\mathrm{p} + \rho_\mathrm{q}$. Denoting for convenience $\Delta\rho = |\rho_\mathrm{p} - \rho_\mathrm{q}|$, we find the planar distance d_n between the residues $\mathrm{R_p}$ and $\mathrm{R_q}$ to be

$$\begin{aligned} d_n &= \sqrt{\Delta\rho^2 + d_p^2 - 2\Delta\rho d_p \cos\frac{2\pi}{7}}, \\ d_p &= \rho_\mathrm{p} + \rho_\mathrm{q} - 2\min\{\rho_\mathrm{p}, \rho_\mathrm{q}\}\cos\frac{2\pi}{7}. \end{aligned} \tag{3.23}$$

We can now estimate the minimum separation between nearest neighbour residues

Table 3.4: Comparison of key lengthscales involved in pairwise molecular interactions between collagen molecules.

Lengthscale	**Magnitude / Å**
Debye length (κ^{-1})	[a]7.82
MJCP step potential width (l_c)	7.5
[b]Minimum nearest neighbour residue separation	2.36
[c] Minimum next nearest neighbour residue separation	7.68

[a] Calculated using $I = 150\,\mathrm{mM}$ and $T = 303.15\,\mathrm{K}$.
[b] This lengthscale is estimated using equation (3.23) by setting $\rho_\mathrm{p} = \rho_\mathrm{q} = 3.14\,\text{Å}$, which corresponds to the minimum residue radius in the Pro-rich statistical parametrisation of the triple helix [96].
[c] This is estimated similarly to [b], except the azimuthal separation between the next nearest neighbour residues is set to $4\pi/7$.

$\mathrm{R_p}$ and $\mathrm{R_q}$ - see Table 3.4. We see that it is substantially smaller than both the Debye length κ^{-1} and the MJCP interaction lengthscale l_c, and as such the nearest neighbour interactions are expected to contribute significantly to the pairwise molecular interactions. On the contrary, the minimum separation between next nearest neighbouring residues, which have an azimuthal separation of $4\pi/7$, is of the same magnitude as both κ^{-1} and l_c. On this basis we ignore next nearest neighbour contributions towards the pairwise molecular interactions.

Given a nearest neighbour strip pair i''-j', we can form an interaction interface by performing the following coordinate transformation:

$$ {}^{\mathrm{q}}\mathbf{x}^{\mathrm{nn}}_{j\prime}(\Delta z) = {}^{\mathrm{q}}\mathbf{x}^{\mathrm{s}}_{j\prime}(\Delta z) + (\mathbf{e}_\rho \cdot ({}^{\mathrm{p}}\mathbf{x}_{i\prime\prime} - {}^{\mathrm{q}}\mathbf{x}_{j\prime}) + d_n)\mathbf{e}_\rho. \tag{3.24} $$

We can repeat aforesaid argument for the other nearest neighbour strip pair i'-j''. The transformation defined by equation (3.24) can be directly applied to i'-j'' by replacing subscripts $j' \to j''$ and $i'' \to i'$. We can now write down an explicit expressions for the contributions to the net interaction energy $\mathrm{U}_{i\text{-}j}$ as

$$ \begin{aligned} \mathrm{E}_{i\text{-}j}(\Delta z) &= \sum_{\mathrm{R_p},\mathrm{R_q}\in\mathcal{C}} \Psi^{\mathrm{DH}}_{\mathrm{pq}}\left(|{}^{\mathrm{p}}\mathbf{x}_i - {}^{\mathrm{q}}\mathbf{x}^{\mathrm{s}}_j|\right) + \sum_{\mathrm{R_p\ or\ R_q}\notin\mathcal{C}} \Psi^{\mathrm{MJ}}_{\mathrm{pq}}\left(|{}^{\mathrm{p}}\mathbf{x}_i - {}^{\mathrm{q}}\mathbf{x}^{\mathrm{s}}_j|\right), \\ \mathrm{E}_{l_1\text{-}l_2}(\Delta z) &= \sum_{\mathrm{R_p},\mathrm{R_q}\in\mathcal{C}} \Psi^{\mathrm{DH}}_{\mathrm{pq}}\left(|{}^{\mathrm{p}}\mathbf{x}_{l_1} - {}^{\mathrm{q}}\mathbf{x}^{\mathrm{nn}}_{l_2}|\right) + \sum_{\mathrm{R_p\ or\ R_q}\notin\mathcal{C}} \Psi^{\mathrm{MJ}}_{\mathrm{pq}}\left(|{}^{\mathrm{p}}\mathbf{x}_{l_1} - {}^{\mathrm{q}}\mathbf{x}^{\mathrm{nn}}_{l_2}|\right), \end{aligned} \tag{3.25} $$

where l_1-l_2 denotes a nearest neighbour strip pair corresponding to either i'-j'' or i''-j', $\mathcal{C}$ is the set of all ionisable residues and the summation is taken over all pairs of interacting residues $\mathrm{R_p}, \mathrm{R_q}$ lying along a given strip pair.

3.7.3 CO-IONISATION BOUNDARY

In this section, we derive the analytic expression for the co-ionisation boundary. Consider two distinct acidic residues $\mathrm{R}_{(1)}$ and $\mathrm{R}_{(2)}$ with acid disassociation constants $-\log_{10} K_{\mathrm{a}}(\mathrm{R}_{(1)}) \equiv \mathrm{pKa}_1$ and $-\log_{10} K_{\mathrm{a}}(\mathrm{R}_{(2)}) \equiv \mathrm{pKa}_2$. Let us denote $\mathrm{pK}^{\mathrm{max}}_{\mathrm{a}} = \max\{\mathrm{pKa}_1, \mathrm{pKa}_2\}$ and $\mathrm{pK}^{\mathrm{min}}_{\mathrm{a}} = \min\{\mathrm{pKa}_1, \mathrm{pKa}_2\}$. We start by computing the disassociation fraction of the residue with $\mathrm{pK}^{\mathrm{min}}_{\mathrm{a}}$ at the value of pH such that the disassociation fraction of the residue with $\mathrm{pK}^{\mathrm{max}}_{\mathrm{a}}$ is exactly α_1. The corresponding average disassociation fraction is defined as the co-ionisation boundary and is given by

$$ \widetilde{\alpha} = \frac{1}{2}\left(\alpha_1 + \frac{\alpha_1}{\alpha_1 + 10^{\mathrm{pK}^{\mathrm{min}}_{\mathrm{a}} - \mathrm{pK}^{\mathrm{max}}_{\mathrm{a}}}(1-\alpha_1)}\right). \tag{3.26} $$

Noting that $\mathrm{pH}_{\mathrm{eqv}}^{\max} - \mathrm{pH}_{\mathrm{eqv}}^{\min} = \mathrm{pKa}_{\mathrm{eqv}}^{\max} - \mathrm{pKa}_{\mathrm{eqv}}^{\min}$, we use equation (3.15) to obtain an expression for $\widetilde{\alpha}$ in terms of μ:

$$\widetilde{\alpha}(\mu) = \begin{cases} \frac{1}{2}\left(\alpha_1 + \frac{\alpha_1}{\alpha_1 + (1-\alpha_1)10^{-2\Delta\mathrm{pH}|\mu|}}\right), \text{ if } \mathrm{R}_{(1)}, \mathrm{R}_{(2)} \text{ acidic} \\ \frac{1}{2}\left(\alpha_2 + \frac{\alpha_2}{\alpha_2 + (\alpha_2^{-1}-1)10^{2\Delta\mathrm{pH}|\mu|}}\right), \text{ if } \mathrm{R}_{(1)}, \mathrm{R}_{(2)} \text{ basic,} \end{cases} \tag{3.27}$$

where μ is understood to denote $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$. The second equation in (3.27) for basic residues can be derived by following the procedure outlined for acidic residues.

3.7.4 Near-Equilibrium States

In order to make predictions regarding the equilibrium behaviour of collagen microfibrils, we require an approximation to the microfibril spectrum. We construct such an approximation by considering the set of near-equilibrium states (NEqSs) S_{eq} - the collection of all microfibrils with locally minimal energies $\mathrm{E}_{\mathrm{M}}(g)$ in the microfibril phase space. We first note that the strongest inter-strip interactions are able to form whenever the spiral strips of two neighbouring molecules are able to face along the edges of the polygon in Figure 3.3. Failure to do so will incur an energy penalty due to azimuthal misalignment of the strips. In order for a molecule to satisfy this condition for both of its nearest neighbours, the internal angle of the polygon needs to be close to an integer multiple of the azimuthal distance between the strips. This condition is satisfied closely for a pentagonal microfibril, since $108^\circ - 2 \cdot 360^\circ/7 \approx 5^\circ$. We therefore have 7 distinct choices of azimuthal rotation angles θ_m that minimise the azimuthal misalignment for each collagen molecule in a microfibril. These distinct azimuthal orientations are more conveniently viewed as choices of an interacting strip pair i_m-j_m along the edge connecting molecule m and its nearest anticlockwise neighbour m'. The azimuthal orientations of the collagen molecules in a microfibril can then be characterised by two vectors $\vec{i} = (i_1, \ldots, i_5)$ and $\vec{j} = (j_1, \ldots, j_5)$. For axial degrees of freedom, we choose the axial staggers Δz_m^* that correspond to the minima of the axially periodic pairwise interaction potentials $\mathrm{U}_{i\text{-}j}^p$. The axial degrees of freedom of collagen molecules in a microfibril can then be described by the vector $\vec{\Delta z^*} = (\Delta z_1^*, \ldots, \Delta z_5^*)$. The set of NEqSs is defined as $\mathrm{S}_{\mathrm{eq}} = \{\vec{s} = (\vec{i}, \vec{j}, \vec{\Delta z^*})\}$, where the components of the state vector $\vec{s}$ satisfy

$$\begin{aligned} i_m &= 1, \ldots, 7 \text{ and } j_m = (i_{m'} - 3) \bmod 7 + 1, \\ \Delta z_k^* &= \underset{\Delta z}{\operatorname{argmin}}\{\mathrm{U}_{i_k\text{-}j_k}^p(\Delta z, g)\},\ k = 1, \ldots, 4, \end{aligned} \tag{3.28}$$

where the gap size g is a fixed value in the interval $[g_a, g_b]$ - see subsection 3.7.5.

We note that due to the cyclical connectivity of the microfibril, there are $\mathrm{N}-1$ independent axial degrees of freedom and the stagger Δz_5^* is constrained via

$$\Delta z_5^* = \left(-\sum_{m=1}^{4} \Delta z_m^* \right) \bmod \mathcal{T}. \tag{3.29}$$

In equation (3.28) we allow M stagger values Δz_k^* for each pairwise interaction energy $\mathrm{U}^p_{i_k\text{-}j_k}$, which correspond to the first M minima in each pairwise potential. The total number of NEqSs in S_{eq} is then $7^5 M^4$. To keep the problem tractable, we choose $M = 3$.

Once all NEqSs are calculated, it remains to determine the ones that satisfy the condition of being axially ordered - see equation (3.9). The parametrisation of hydrophobic-hydrophobic and charged-hydrophobic interactions with the step potential $\Psi^{\mathrm{MJ}}_{\mathrm{pq}}$ in equation (3.1) fundamentally limits the precision with which we can determine the pairwise interaction potential minima Δz_m^*. To circumvent this limitation when numerically evaluating whether a given microfibril is axially ordered, we utilise a weaker condition

$$\left| \left\{ \frac{z_1^* \bmod \mathcal{T}}{\mathcal{D}}, \ldots, \frac{z_5^* \bmod \mathcal{T}}{\mathcal{D}} \right\} - \{0, \ldots, 4, \} \right| < \{\epsilon_{\mathrm{tol}}, \ldots, \epsilon_{\mathrm{tol}}\}, \tag{3.30}$$

where ϵ_{tol} is the tolerance factor, which in all simulations we set to 0.05.

3.7.5 SELECTION OF GAP LENGTHSCALE RANGE

Our goal is to determine the interval of gap lengthscales $[g_a, g_b]$, over which we minimise the microfibrillar energy and study the phases that emerge at thermodynamic equilibrium. Based on the experimental studies of type II collagen, we expect the minimum of $\mathrm{E_M}(g)$ to appear around $g_{\mathrm{exp}} = 5\mathrm{D} - L$, where $\mathrm{D} \approx 670\,\text{Å}$ and L is the molecular length [2]. We perform a preliminary minimisation of $\mathrm{E_M}(g)$ on an interval centred around g_{exp} at physiological pH $= 7.40$, ionic strength $I = 150\,\mathrm{mM}$ and using $\mathrm{pKa}^{\mathrm{ref}}$ values for acid disassociation constants from Table 3.3. The gap length that minimises the microfibrillar energy is denoted by

$$g_{\mathrm{ref}} = \underset{\substack{g \in [(1-C)g_{\mathrm{exp}}, (1+C)g_{\mathrm{exp}}], \\ \mathrm{S_{eq}}}}{\operatorname{argmin}} \{\mathrm{E_M}(g)\},$$

where $C \in (0, 1)$. We perform all subsequent calculations of microfibrillar energy as well as of equilibrium probabilities on the interval

$$[g_a, g_b] = [(1-C)g_{\mathrm{ref}}, (1+C)g_{\mathrm{ref}}]\,. \tag{3.31}$$

To keep our calculations computationally tractable, we choose $C = 0.2$, which yields $g_{\mathrm{ref}} \approx 5 \cdot 667\,\text{Å} - L$, which is in good agreement with experimental measurements [2].

3.7.6 Discretisation of Ionic Parameters

In all calculations, the overlap measure μ is uniformly discretised with a step size of 0.05. For a given value of the overlap measure $\mu(W_{\mathrm{R}_{(1)}}, W_{\mathrm{R}_{(2)}})$, we want to study the self-assembly of collagen as a function of pH. The choice of pH discretisation becomes important, due to the fact that the disassociation fraction changes rapidly around the half-equivalence point $\mathrm{pH}_{\mathrm{eqv}}$. Insufficiently fine uniform discretisation can therefore end up inadvertently excluding disassociation fractions in the vicinity of $\mathrm{pH}_{\mathrm{eqv}}$. We circumvent this issue as well as the problem of computational cost associated with fine, uniform discretisations by constructing a non-uniform pH grid. The pH discretisation is such that it is finer around a given half-equivalence point and coarser as one moves towards the endpoints of a given ionisation window.

Suppose that we have N_{co} distinct co-ionising residues labelled $\mathrm{R}_{(1)}, \mathrm{R}_{(2)}, \ldots, \mathrm{R}_{(N_{\mathrm{co}})}$. Consider a closed interval of pH values given by $[\mathrm{pH}_{\mathrm{s}}, \mathrm{pH}_{\mathrm{e}}]$. For this interval we want to construct a pH grid of size $N_{\mathcal{G}}$. This means finding a finite set of points

$$\mathcal{G} = \{\mathrm{pH}_1 \equiv \mathrm{pH}_{\mathrm{s}}, \mathrm{pH}_2, \ldots, \mathrm{pH}_{N_{\mathcal{G}}} \equiv \mathrm{pH}_{\mathrm{e}}\}.$$

We will construct the grid $\mathcal{G}$ by requiring that given any two successive grid points $\mathrm{pH}_{i+1} > \mathrm{pH}_i$, for $i = 1, 2, \ldots, N_{\mathcal{G}} - 1$, the corresponding increment in the disassociation fractions of all residue types $\mathrm{R}_{(1)}, \mathrm{R}_{(2)}, \ldots, \mathrm{R}_{(N_{\mathrm{co}})}$ never exceeds $\Delta\alpha$. In all calculations we set $\Delta\alpha = 0.05$. We can achieve this by constructing a list of candidate pH grid points for every residue of type $\mathrm{R}_{(k)}$, where $k = 1, 2, \ldots, N_{\mathrm{co}}$. The candidate grid point is given by

$$\widetilde{\mathrm{pH}}_{i+1}(\mathrm{R}_{(k)}) = \mathrm{pH}_{\mathrm{eqv}}(\mathrm{R}_{(k)}) - \log_{10}\left(\frac{1 + 10^{\mathrm{pH}_{\mathrm{eqv}}(\mathrm{R}_{(k)}) - \mathrm{pH}_i}}{1 \pm \Delta\alpha\left(1 + 10^{\mathrm{pH}_{\mathrm{eqv}}(\mathrm{R}_{(k)}) - \mathrm{pH}_i}\right)} - 1\right), \tag{3.32}$$

where the symbol "$\pm$" is taken to be "$+$" when $\mathrm{R}_{(k)}$ is an acidic residue and "$-$" when $\mathrm{R}_{(k)}$ is a basic residue. We can ensure that the maximum disassociation fraction increment is $\Delta\alpha$ among all residues that are ionised on the interval $[\mathrm{pH}_{\mathrm{s}}, \mathrm{pH}_{\mathrm{e}}]$ by picking the smallest $\widetilde{\mathrm{pH}}_{i+1}(\mathrm{R}_{(k)})$ to be the next pH grid point

$$\mathrm{pH}_{i+1} = \min_{\mathrm{R}_{(k)}}\{\widetilde{\mathrm{pH}}_{i+1}(\mathrm{R}_{(k)})\}. \tag{3.33}$$

3.7.7 Residue Abundances across Fibrillar Collagens

The histogram of the frequencies of ionisable residues across different mammalian sequences of fibrillar collagens of types I, II, III, V and XI is illustrated in Figure 3.11. We note that type I and V collagens may form multiple isoforms comprised of different α-chains and therefore differing in residue composition. For these two collagen types, we illustrate

the data for the most commonly found isoforms - $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ and $\alpha_1(\mathrm{V})\alpha_2(\mathrm{V})\alpha_1(\mathrm{V})$ for types I and V respectively [11, 70]. We only account for the charged residues that occur in the sequence portions corresponding to the triple helix. We see that there are on average between 0 and 2 Cys/Tyr residue per spiral strip across all fibrillar collagens. This is substantially lower than the frequency of Arg, Asp, Glu and Lys, which on average occur 10-30 times per spiral strip. Type II collagen is unique in its residue composition in that the frequency of His residues per spiral strip is approximately between 0 and 1. We therefore expect that for type II collagen, the ionisation of residues Cys, His and Tyr will have a negligible effect on the microfibril self-assembly, due to their low abundance.

3.7.8 Axial Periods of Lowest Energy NEqSs

To further validate our model, we calculate the axial periods $\mathcal{D}$ of the lowest energy NEqSs $\vec{s}_{\min}$ as a function of residue pKa and pH - see Figure 3.12. We see that the axial periods fall narrowly in the interval $[666, 668]$Å, which is in good agreement with the experimental measurements of axial periodicity in the axially ordered phase [51]. To substantiate the aforementioned finding, in Figure 3.13 we include an example of the global energy minima for pairwise interaction potentials $\mathrm{U}^{p}_{i\text{-}j}$ corresponding to the region of the phase diagram in Figure 3.6 where we predict the axially ordered aggregated phase to form at thermodynamic equilibrium.

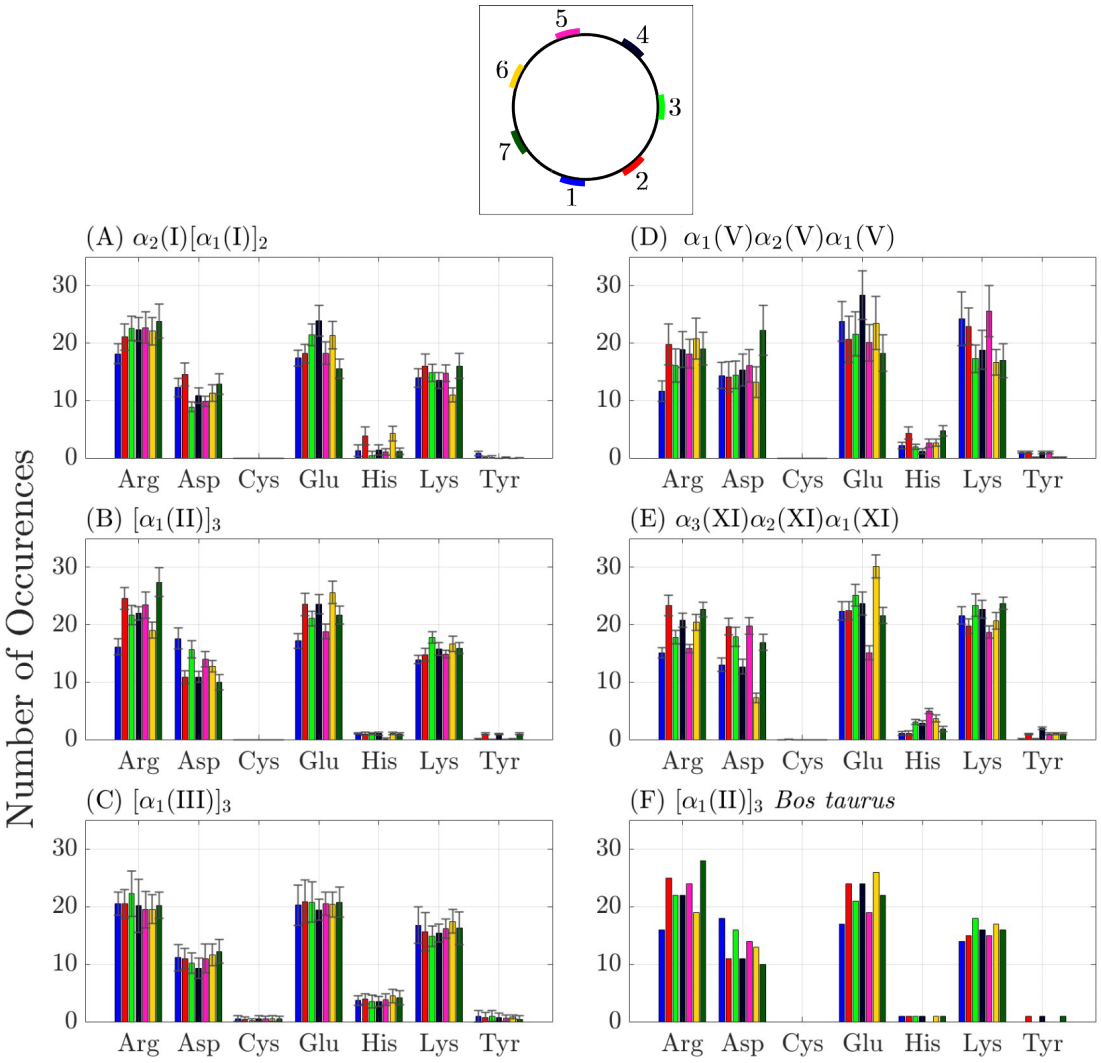


Figure 3.11: (A)-(E) Frequency of ionisable residues across spiral strips in fibrillar collagens of types I, II, III, V and XI. The bars and the associated error bars show the mean and the standard deviation of the number of residue occurrences across all mammalian fibrillar collagens in the NCBI RefSeq database [80]. (F) Frequency of ionisable residues across spiral strips in type II bovine collagen.

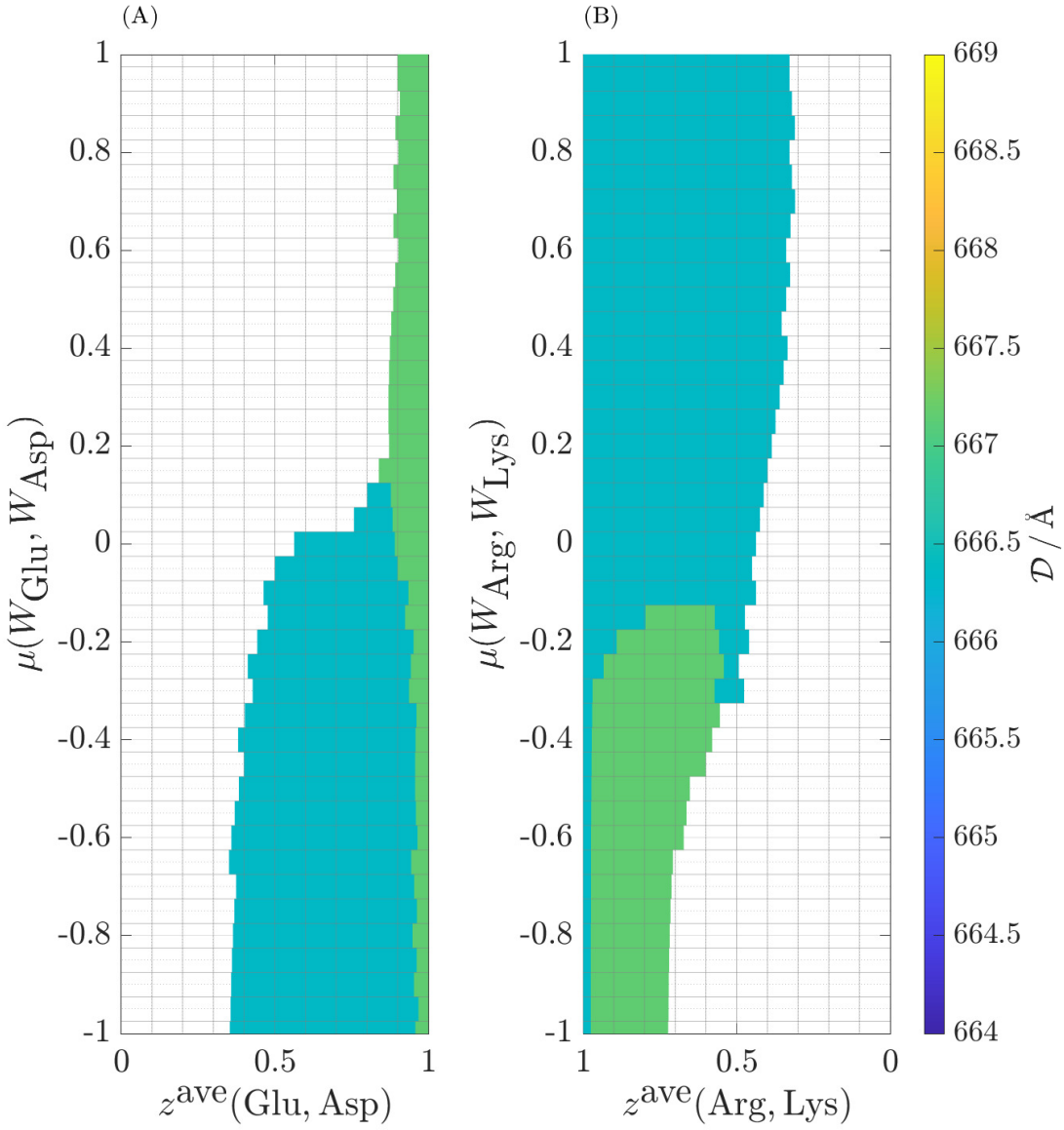


Figure 3.12: Axial periods $\mathcal{D}$ of the lowest energy near-equilibrium states $\vec{s}_{\mathrm{min}}$ in the z^{ave}-μ space. The data is shown for (A). ionisation of acidic residue and (B). de-ionisation of basic residues. The amino acid sequence is that of bovine collagen $[\alpha_1(\mathrm{II})]_3$.

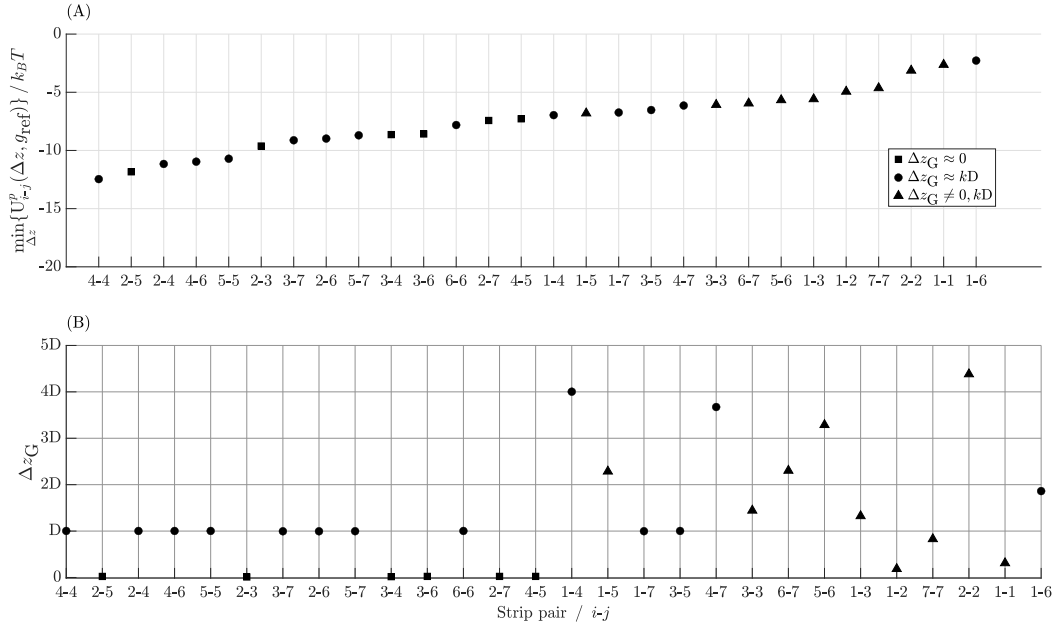


Figure 3.13: (A). Global energy minima of the interaction potentials $\mathrm{U}^p_{i\text{-}j}$. (B). Axial staggers corresponding to the global energy minima in (A). All calculations are performed for $\mu(W_{\rm Glu}, W_{\rm Asp}) = \mu(W_{\rm Arg}, W_{\rm Lys}) = 0.15$, $z^{\rm ave}({\rm Glu}, {\rm Asp}) = z^{\rm ave}({\rm Arg}, {\rm Lys}) = 1$ and $g = g_{\rm ref} \approx 5 \cdot 667 - L$. The amino acid sequence is that of bovine collagen $[\alpha_1({\rm II})]_3$.

— 4 —

Parallel Triple Helix Interactions Encode Fibrous Long-Spacing Collagen Axial Periodicities

Abstract

The notion that collagen is a polymorphic structure has been known at least since the 1950s. One such diverse group of polymorphic collagen aggregates are the fibrous long-spacing (FLS) collagens. FLS collagens are characterised by axial periodicities in the range $\approx$ 1000-2700 Å, which significantly exceeds the $\approx$ 670 Å native collagen D-period. In this work, we use a mathematical model of microfibril self-assembly to study the sequence-structure relationship of FLS collagens. In contrast with previous studies, we show that the axial periods of FLS collagens are determined by pairwise interactions between parallel collagen triple helices. We demonstrate that in spite of pairwise collagen interactions encoding a wide range of polymorphic structures, the native D-banded microfibril is selectively stabilised by specific inter-molecular interactions. Finally, for the first time, we use a sequence and molecular structure informed approach to predict the 3-dimensional organisation of triple helices inside FLS microfibrils. Our results offer a novel perspective on a well-known feature of collagen self-assembly and provide a rich ground for further experimental and theoretical study of collagen polymorphism.

4.1 PRELIMINARY REMARKS

So far in chapters 2 and 3, we have encountered a single type of axially periodic microfibril, namely the one characterised by D-banding. As we have seen in Figure 1.2, however, this is far from the only axial periodicity that collagen aggregates can display. Explaining the existence of other axial periods in collagen aggregates will require generalising two distinct aspects of the models that we have considered thus far.

Firstly, we need to determine in general what axial periodicities are admissible for a microfibril with N molecules of length L. As we shall see in chapter 4, in spite of the collagen microfibril being known since the 1960s, the aforementioned fundamental question has not been addressed in existing literature in its entirety.

The second aspect that requires generalisation is more subtle. Namely, it is crucial that we generalise the set of microfibril molecular packing models. In chapter 2, we established an important relationship between molecular packing models and pairwise collagen interactions - the molecular packing models with all internal polygon angles close to integer multiples of the azimuthal spacing between the spiral strips will be the most energetically stable. However, it is important to notice, that a given choice of an internal polygon angle imposes a significant constraint on the set of strip-strip interactions attainable in a microfibril (see equation (3.28) for an example of constraints on interacting strips pairs i_m-j_m). There is no reason to think that the same set of strip-strip interactions that favours formation of D-banded microfibrils will similarly favour formation of microfibrils with other axial periods. Considering a larger set of molecular packing models is therefore imperative.

Chapter 4 will focus on one class of polymorphic collagen aggregates, known as fibrous long-spacing collagens (FLS). FLS collagens form in the presence of large, flexible and typically highly charged macromolecules, resulting in aggregates with axial periods that significantly exceed D-banding [120]. The aforementioned results suggest that charged residue interactions may be important for encoding FLS axial periodicities in addition to the D-period. Rather unexpectedly, existing theoretical literature on FLS collagen self-assembly indicates that there is no preference for parallel association of collagen triple helices with axial staggers corresponding to FLS axial periods [34, 120]. If the reader were to take away a single idea from chapter 4, it would be that contrary to existing knowledge in the field, pairwise triple helix interactions do encode FLS collagen axial periodicities. As we shall see, aforementioned extensions to the theoretical methods introduced in chapters 2 and 3 as well as the fact that parallel triple helix interactions encode FLS axial periodicities, will lead us to predict, for the first time, 3-dimensional molecular packing for several different FLS collagen microfibrils.

4.2 INTRODUCTION

Fibrillar collagens are vital components of the extracellular matrix as well as different stress-bearing tissues, such as bone and tendon. Supramolecular aggregates of fibrillar collagens, known as fibrils, possess unique mechanical properties that enable them to serve diverse structural functions, ranging from mediating mechano-transduction across the extracellular matrix to providing resistance to repeated mechanical forces in bone [31, 101].

The mechanical properties of collagen fibrils result from the intricate periodic spatial arrangement of the comprising collagen molecules along the axial direction of the fibril. The axial molecular organisation of collagen molecules within fibrils manifests itself as D-banding, which corresponds to D $\approx$ 67 nm periodic collagen density modulations along the long axis of the fibril that are visible in negatively stained TEM images. The existence of D-banding is often explained with the aid of a simple 2-dimensional geometrical argument first introduced by Hodge and Petruska [90]. Collagen molecules of length $L \approx 300$ nm are grouped into pentameric subunits (microfibrils), wherein each molecule is translated (staggered) in the direction of the fibril axis by an integer multiple of D. The microfibrils are then arranged periodically along the fibril axis, separated by gaps of size 0.54D. Since $L \approx 4.46$D, the resulting molecular arrangement reproduces the D-banding pattern.

Later X-ray scattering studies of collagen fibrils led to the proposal of several models describing the 3-dimensional molecular arrangement inside the microfibril [74, 109, 82]. The unifying feature of these models is the presence of cyclical connectivity between the molecules comprising the microfibril, meaning that the molecular packing may be envisioned by placing the collagen molecules at the vertices of a simple polygon. More recently, 3-dimensional models of molecular organisation inside microfibrils have provided a basis for elucidating the relationship between the microscopic structure of D-banded microfibrils and the emergent mechanical properties of collagen-rich tissues [121, 30].

Collagen, however, displays a large degree of polymorphism in the spatial arrangement of molecules within fibrils. This is evidenced by the broad range of axial periodicities different from the usual D-banding that have been reported in literature [34, 120]. Fibrous long-spacing (FLS) collagens are one such diverse group of polymorphic structures, which is defined by the presence of axial periodicities greater than D. FLS aggregates self-assemble both *in vitro* and *in vivo* in presence of charged macromolecules, including proteoglycans, α_1-acid glycoprotein and chondroitin sulphate [120]. *In vitro* experiments have identified four distinct types of FLS aggregates designated FLS I-IV. With the exception of FLS III, which forms 1000 Å periodic aggregates, FLS aggregates formed *in vitro* are characterised by axial periods falling in the range $\approx$ 1700-2700 Å [22, 34, 97]. *In vivo*, FLS fibrils have been found across a broad range of both healthy and pathological tissues, however their biological function remains poorly understood [44, 120]. Curiously, FLS aggregates formed

in vivo most frequently display axial periodicities falling in the range ≈ 790-1500 Å that are distinct and not observed in *in vitro* aggregates - an observation that has been noted to be puzzling [87].

Unlike D-banded fibrils, relatively little is known from experimental literature about the spatial arrangement of collagen molecules in FLS aggregates or the specific interactions that drive FLS self-assembly at the microscale. AFM (atomic force microscopy) studies of FLS fibrillogenesis suggest that FLS aggregates undergo hierarchical self-assembly starting with an axially staggered microfibrillar structure [86, 97, 117]. Experimental limitations, however, did not allow for resolution and direct measurement of the axial stagger at the microfibrillar level. To our knowledge, no detailed studies of the 3-dimensional molecular organisation inside FLS microfibrils have been performed either.

Among the rich body of theoretical literature dedicated to collagen self-assembly, only a handful of studies over the years have focused on FLS collagens. As a result, some of the fundamental principles that underlie the observed axial periodicity polymorphism in aggregates of collagen remain only partially understood. (1). What is the general relationship between molecular length, axial periodicity and aggregate size in collagen microfibrils? Special cases of this problem are well-understood, such as the case of D-banded microfibrils [82, 95]. To explain the axial periodicities observed in FLS collagens, several 2-dimensional packing schemes akin to those of Hodge and Petruska have been suggested [22, 34]. Notably, there is no clear consensus regarding the number of molecules N comprising FLS microfibrils. This indicates a need for a general quantitative formalism relating the key structural descriptors of the microfibril. Existing suggestions for such a formalism impose mathematical constraints on the range of axial periodicities that may be predicted in microfibrils. The upper bound on axial periodicities is set at $L/2 \approx 150\,\text{nm}$ which renders the model unable to explain the self-assembly of FLS collagens observed *in vitro* [95]. (2). What is the sequence-structure relationship for FLS collagens? Previous studies of pairwise interactions between parallel collagen triple helices suggested that there was no energetic preference for axial staggers that could be used to explain axial periodicities in FLS collagens [34, 120]. Instead, it was suggested that negatively charged macromolecules necessary for FLS formation serve as interaction mediators between positively charged residues of collagen triple helices in anti-parallel arrangement [34, 120]. Such studies are however severely limited, as they were performed with only partial knowledge of the residue sequence of the triple helix. More modern studies of pairwise collagen interactions have so far only focused on prediction of D-banding in collagen aggregates [95]. (3). What 3-dimensional microfibrillar packing models admit axial periodicities observed in FLS aggregates? To our knowledge, no such models have been proposed in the literature.

In the first part of this work, corresponding to section 4.3, we present a physically motivated derivation for the relationship between the key structural descriptors of the

microfibril. Unlike previously stated formulations [95], our approach accommodates the existence of FLS axial periodicities at the microfibrillar level. Following this, in section 4.4, we outline the construction of 3-dimensional molecular packing models for collagen microfibrils which are based on the 3-dimensional residue organisation of the triple helix. In sections 4.5-4.6 we describe the mathematical procedure for comparing the thermodynamic stabilities of different axially periodic microfibrils. Next, in section 4.7, we use the mathematical models built up in the previous sections to study the sequence-structure relationship for FLS collagens and for the first time predict the 3-dimensional molecular organisation in FLS I and FLS IV microfibrils (Figure 4.5). Contrary to previous studies [34, 120], we show that the axial periodicities of FLS collagens are encoded by pairwise molecular interactions between parallel triple helices. We demonstrate that the existence of competing conformations in the form of FLS aggregates does not preclude formation of globally stable D-periodic microfibrils. Finally, we identify the physical mechanisms and specific residue interactions that allow for selective stabilisation of D-banded microfibrils as opposed to other polymorphic aggregates of collagen.

4.3 AXIAL PERIODICITY IN COLLAGEN MICROFIBRILS

Our first goal is to describe the influence of the main structural parameters of the microfibril on the emergent axial periodicity. Since we are currently concerned with axial molecular ordering in collagen aggregates, without loss of generality we can ignore the details of the 3-dimensional molecular packing. We will integrate 3-dimensional molecular packing into our microfibril self-assembly model in section 4.4. Figure 4.1A illustrates the 2-dimensional representation of a microfibril comprised of N collagen molecules, which we term an N-microfibril. An N-microfibril consists of identical collagen molecules of length $L \approx 300\,\mathrm{nm}$, which periodically repeat after a gap of length g. We associate a single axial degree of freedom to each molecule, represented by its axial stagger z_m for $m = 1, 2, \ldots, \mathrm{N}$.

The sum of L and g sets the axial periodicity lengthscale, which we denote by $\mathcal{T} = L + g$. We define a microfibril to be n-periodic, for a positive integer $n \geq 2$, with period $\mathcal{D}$, if the following two conditions are satisfied: (i). The lengthscale $\mathcal{T}$ divides into $\mathcal{D}$ exactly n times. (ii). Every time upon traversing a single axial period $\mathcal{D}$, one counts the same positive integer number N_g of gaps in a microfibril. Divisibility condition (i), can be stated as a linear constraint on the gap size

$$g = n\mathcal{D} - L. \tag{4.1}$$

A given N-microfibril is not generally n-periodic for any choice of n. To show this, we start by dividing $\mathcal{T}$ into n equal-length segments (condition (i) above), as shown in

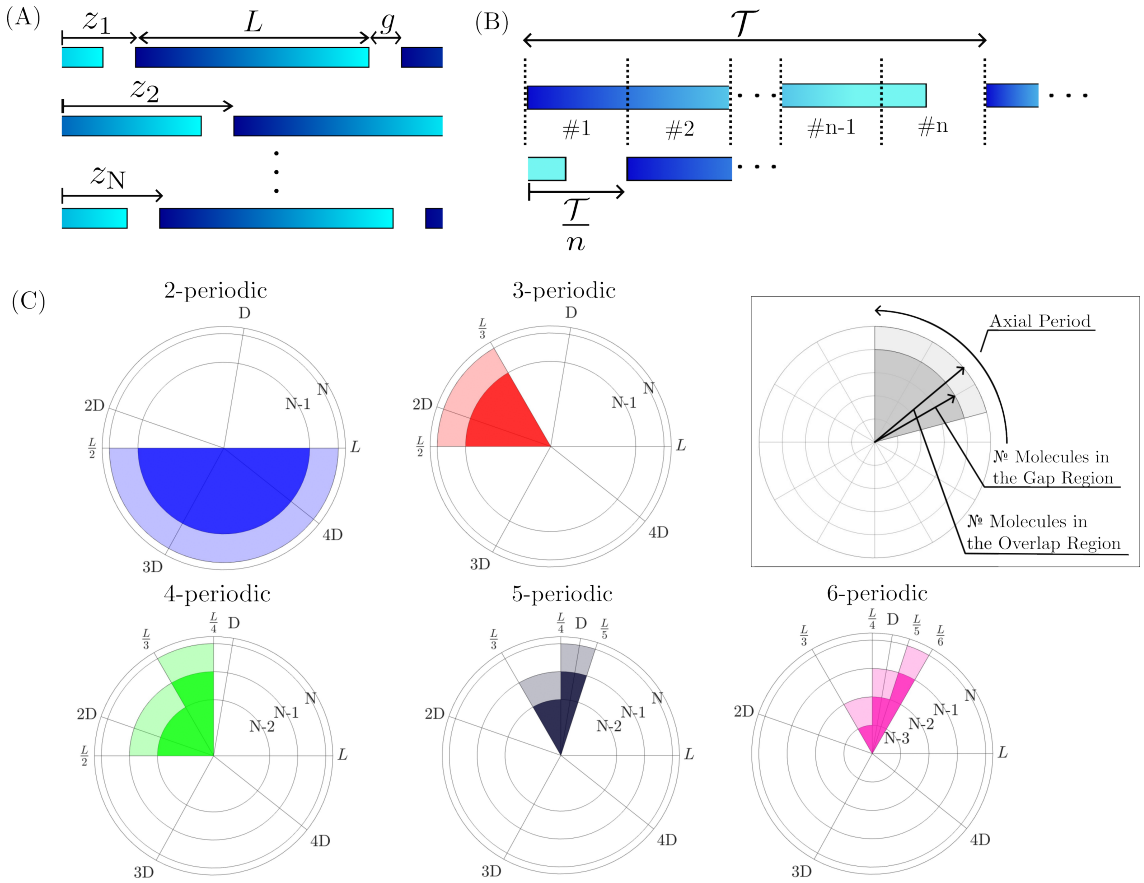


Figure 4.1: (A). 2-dimensional molecular packing scheme for an N-membered collagen microfibril. (B). Subdivision of axial periodicity lengthscale $\mathcal{T}$ into n equal-length segments $\#1, \ldots, \#n$. (C). Admissible axial periods and numbers of molecules in gap/overlap regions of n-periodic microfibrils for $n = 2, 3, \ldots, 6$.

Figure 4.1B. In order for an N-microfibril to be n-periodic, we further require that each equal-length segment in Figure 4.1B contains exactly $\mathrm{N}_g = \mathrm{N}/n$ gaps (condition (ii) above). This is equivalent to the requirement that N is divisible by n, which we denote as $n \mid \mathrm{N}$. Given that $n \mid \mathrm{N}$, every segment in Figure 4.1B will be occupied with a gap provided that each collagen molecule is axially staggered by an integer multiple of $\mathcal{D}$:

$$\{z_1 \bmod \mathcal{T}, \ldots, z_{\mathrm{N}} \bmod \mathcal{T}\} = \{\underbrace{0, \ldots, 0}_{\mathrm{N}_g \text{ times}}, \underbrace{\mathcal{D}, \ldots, \mathcal{D}}_{\mathrm{N}_g \text{ times}}, \ldots, \underbrace{(n-1)\mathcal{D}, \ldots, (n-1)\mathcal{D}}_{\mathrm{N}_g \text{ times}}\}, \quad (4.2)$$

where the curly braces $\{\ldots\}$ indicate the elements inside the braces are unordered. We note that equation (4.2) is a natural generalisation of our previously stated condition for a 5-microfibril to be axially periodic in equation (3.9) of chapter 3.

From current definitions, we can easily obtain a lower bound on axial periods that a given n-periodic microfibril may attain. It is reasonable to assume that $g > 0$, therefore

the lower bound on the axial period is given by

$$\mathcal{D} > \frac{L}{n}. \tag{4.3}$$

Ascertaining the upper bound on $\mathcal{D}$ necessitates introducing additional physical constraints. Consider a pair of interacting molecules in Figure 4.1A at axial positions z_m, z_{m^*} with $m \neq m^*$. The axial component of their pairwise interaction potential has the general form

$$\mathrm{E}^p(\Delta z_m, g) = \mathrm{E}(\Delta z_m) + \mathrm{E}(\Delta z_m - g - L), \tag{4.4}$$

where $\Delta z_m = z_m - z_{m^*}$ and $\mathrm{E}(\Delta z_m)$ is a pairwise interaction potential between two collagen molecules. In an n-periodic microfibril, we constrain the gap size and axial staggers according to equations (4.1) and (4.2) respectively to obtain

$$\mathrm{E}^p(k\mathcal{D}, n\mathcal{D} - L) = \mathrm{E}(k\mathcal{D}) + \mathrm{E}(k\mathcal{D} - n\mathcal{D}), \tag{4.5}$$

where $k = 1, 2, \ldots, n-1$. In this work, we only consider n-periodic microfibrils, in which there exists a pair of molecules such that their axially periodic energy $\mathrm{E}^p(k\mathcal{D}, n\mathcal{D} - L)$ results from two pairwise molecular interactions specified by equation (4.5). Table 4.2a (see Appendix 4.10.8) illustrates a pairwise interaction that violates this condition (top row) and one that satisfies it (bottom row). Since each collagen molecule in a microfibril is of length L, our condition on molecular interactions can be satisfied for a given value of k whenever

$$\mathcal{D} < \min\left\{\frac{L}{k}, \frac{L}{n-k}\right\}. \tag{4.6}$$

It is clear that the largest upper bound in equation (4.6) may be obtained by setting $k = \mathrm{ceil}\left(\frac{n}{2}\right)$ (by symmetry, we may also set $k = \mathrm{floor}\left(\frac{n}{2}\right)$). The value of the axial period in a given n-periodic microfibril is then constrained by

$$\frac{L}{n} < \mathcal{D} < \frac{L}{\mathrm{ceil}\left(\frac{n}{2}\right)}. \tag{4.7}$$

It is instructive to consider the physical interpretation of the upper bounds on the axial period $\mathcal{D}$ in equation (4.6) that correspond to other choices of k. Other choices of k generate a finite sequence of upper bounds on $\mathcal{D}$:

$$\left(\frac{L}{k}\right)_{k=\mathrm{ceil}\left(\frac{n}{2}\right)}^{n-1} = \left(\frac{L}{n-1}, \frac{L}{n-2}, \ldots, \frac{L}{\mathrm{ceil}\left(\frac{n}{2}\right)}\right), \tag{4.8}$$

where we have arranged the sequence elements in order from smallest to largest. Exceeding a given upper bound in sequence (4.8), results in one less collagen molecule per gap and overlap region of the microfibril - see Table 4.2b for an example. We therefore expect in general that depending on the value of gap length g, n-periodic microfibrils may produce

negative staining patterns of differing contrast to those observed in TEM images of D-banded collagen aggregates. Figure 4.1C illustrates the constraints on the axial period $\mathcal{D}$ for the first five n-periodic microfibrils and the associated numbers of molecules in the gap and overlap regions. Finally, we note that exceeding the smallest upper bound $\frac{L}{n-1}$ in equation (4.8) is equivalent to the condition that $\mathcal{D} < g$. This means that the length of the high electron density (gap) region L_g in the negatively stained pattern of a TEM image is no longer expected to be the same as the gap size g. To account for this, we distinguish between the geometric parameter g and the length of the gap region L_g in the negative staining pattern, which is given by the general expression

$$L_g = \mathcal{D}\left(1 - \text{frac}\left\{\frac{L}{\mathcal{D}}\right\}\right), \tag{4.9}$$

where frac$\{\dots\}$ returns the fractional part of its argument.

We conclude this section by applying our results to collagen aggregates of a specified size. Noting that the lower and upper bounds on $\mathcal{D}$ are both monotonic functions of n, the constraints on the value of $\mathcal{D}$ for a given N-microfibril are simply determined by the largest and smallest divisors of N. A given N-microfibril can then give rise to axial periods

$$\frac{L}{\mathrm{N}} < \mathcal{D} < \frac{L}{\text{ceil}\left(\min\limits_{n|\mathrm{N}}\{n\}/2\right)}, \tag{4.10}$$

where $\min_{n|\mathrm{N}}\{n\}$ corresponds to the smallest divisor of N that is greater than 1. A notable consequence of equation (4.10) is that microfibrils with axial periods observed in certain FLS collagens (FLS I and FLS IV), are necessarily comprised of an even number of molecules. This is by the virtue of the fact that only 2-periodic microfibrils are capable of producing aggregates with axial periods $\mathcal{D} \in (\frac{L}{2}, L)$. We will see this fact reflected later in section 4.7.

4.4 MOLECULAR PACKING MODELS

So far, all of our results have been independent of the 3-dimensional spatial molecule organisation inside a microfibril. The next step is to apply the principles that determine the axial periodicity in collagen aggregates to specific molecular packing models. To the best of the authors' knowledge, no experimental or theoretical studies have been conducted to ascertain the molecular organisation inside any of the FLS collagen aggregates. Inspired by the existing experimental and theoretical models for D-banded microfibrils [74, 109, 82], we will consider the family of cyclical molecular packing models.

We take a cyclical molecular packing model for a microfibril to be a simple (meaning non-self-intersecting) N-gon, with collagen molecules positioned at its vertices. We note that

for FLS collagens, the fine structure of the $\mathcal{D}$-period is characterised by dihedral symmetry, which has been suggested to be indicative of FLS fibrils possessing an approximately equal number of triple helices in parallel and antiparallel arrangement [22, 34]. Whether the dihedral symmetry arises at the microfibrillar or supramicrofibrillar scale remains to be elucidated. To keep our calculations tractable, we will assume that all collagen molecules in a microfibril are in parallel arrangement. In addition to axial degrees of freedom z_m, we ascribe each molecule an azimuthal degree of freedom θ_m, corresponding to a rotation around its molecular axis. We assume that the only intermolecular interactions contributing to the energy of the collagen microfibril are between the molecules that share an edge of the N-gon. To distinguish between the indices of the nearest neighbour molecules in a microfibril, we introduce the notation $m' = (m \bmod m^{\max}) + 1$ and $m'' = ((m-2) \bmod m^{\max}) + 1$, where $m^{\max}$ denotes the maximum admissible value taken by the index m.

For any given cyclical packing model, the pairwise interaction potential between two molecules sharing a polygon edge can be written as

$$P_m(\theta_m, \theta_{m'}, \Delta z_m, g) = \Phi(\theta_m)\Phi(\theta_{m'})\mathrm{E}^p(\Delta z_m, g), \tag{4.11}$$

where $\mathrm{E}^p(\Delta z_m, g)$ denotes the axial component of the pairwise interaction and $\Delta z_m = z_{m'} - z_m$. The azimuthal component of the interaction potential $\Phi(\theta_m) \in [0, 1]$, introduces an energetic penalty based on the azimuthal orientations of the interacting molecules. The exact form of $\Phi(\theta_m)$ depends on the detailed 3-dimensional spatial organisation of the residues comprising molecular collagen. In our previous work, we have shown that the residues on the collagen molecular surface organise into 7 equally-spaced, right-handed helical strips (see Figure 4.2A) [124]. Based on this spatial residue organisation, we proposed a simplified description for the azimuthal component of P_m which satisfies the following properties:

$$\Phi\left(\theta_m + \frac{2\pi}{n_s}\right) = \Phi(\theta_m), \quad \Phi(\theta_m) = \Phi\left(2\theta_0 + \frac{2\pi}{n_s} - \theta_m\right), \tag{4.12}$$

$$\Phi(\theta_m) \text{ is decreasing for } \theta_m \in \left(\theta_0, \theta_0 + \frac{\pi}{n_s}\right), \tag{4.13}$$

where n_s is the number of helical strips, $\theta_0 \in \left[0, \frac{2\pi}{n_s}\right)$ is such that $\Phi(\theta_0) = 1$ and $\theta_m \in [0, 2\pi)$. An example of a function Φ that satisfies conditions (4.12) and (4.13) can be seen in Figure 4.2B. Physically, assumption (4.12) indicates that all spiral strips are sterically equivalent, meaning that the azimuthal component of the pairwise molecular potential behaves identically for each spiral strip. Assumption (4.13) posits the existence of n_s^2 distinct optimal azimuthal orientations for a pair of interacting molecules, corresponding to the interacting spiral strips facing each other along an edge of the N-gon (see Figure 4.2C).

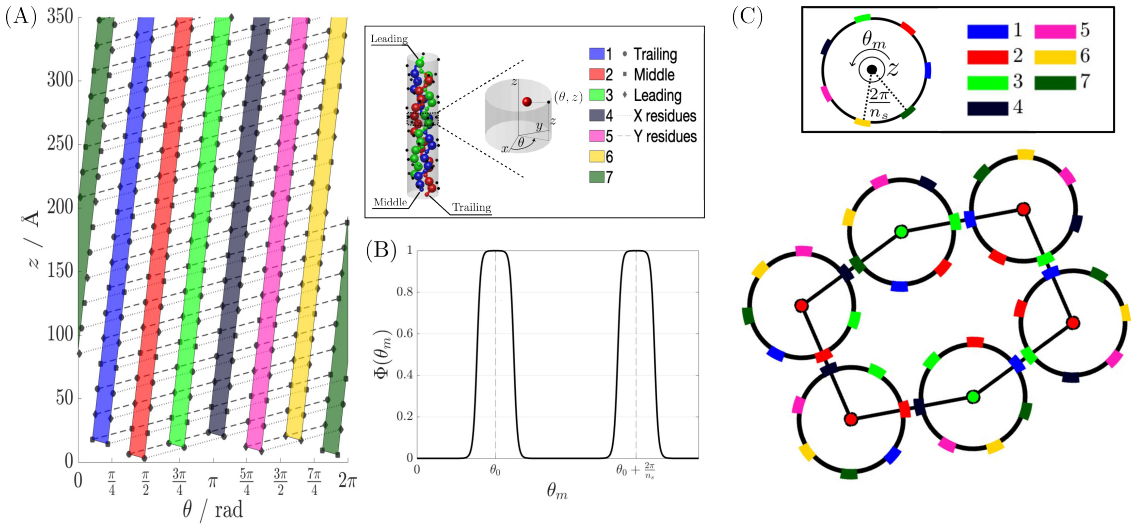


Figure 4.2: (A). Helical residue organisation of the collagen triple helix. Each point corresponds to the coordinates of a C_α atom of a residue. Dotted lines connect the residues that belong to the same α-chain. Solid lines indicate imaginary connections between residues that fall on one of the 7 right-handed spiral strips. (B). Example of the azimuthal component of pairwise interaction energy that satisfies criteria (4.12) and (4.13). (C). Example of a 3-dimensional molecular packing with 6 collagen triple helices that optimises the azimuthal component of pairwise interaction energy for each pair of molecules sharing a polygon edge.

Properties (4.12) and (4.13) enable us to significantly reduce the number of candidate molecular packing models. We neglect all N-gons with internal angles ψ_m that do not correspond to positive integer multiples of the azimuthal spacing between the spiral strips. Such molecular packing models will incur an azimuthal energy penalty in one or more pairwise potentials given by (4.11), regardless of the choice of θ_m. We therefore require the internal angles ψ_m to satisfy

$$\pi(\mathrm{N}-2) = \sum_{m=1}^{\mathrm{N}} \psi_m, \text{ for } \psi_m = \frac{2\pi n_m}{n_s}, \tag{4.14}$$

where $n_m = 1, 2, \ldots, n_s - 1$. One may question whether angles ψ_m satisfying equation (4.14) implies the existence of a simple polygon with internal angles ψ_m. The proof of this can be found in Appendix 4.10.1.

The constraint in equation (4.14) is equivalent to finding all positive integer partitions comprised of exactly N terms, namely

$$\frac{n_s}{2}(\mathrm{N}-2) = \sum_{m=1}^{\mathrm{N}} n_m. \tag{4.15}$$

It is clear that the integer partition in equation (4.15) may only be constructed for even aggregate sizes N whenever n_s is odd. For n_s and N both odd, we have to introduce some azimuthal misalignment between the helical strips of interacting molecules. For any choice of positive integers n_m, we find that $\sum_m \psi_m$ is at least $\frac{\pi}{n_s}$ away from the sum of internal angles in an N-gon. By setting $\psi_m = \frac{2\pi n_m}{n_s} \pm \frac{\pi}{\mathrm{N}n_s}$, we can satisfy the constraint (4.14), whilst minimising the azimuthal misalignment of each interacting molecular pair. This is equivalent to calculating the integer decomposition of

$$\frac{n_s}{2}(\mathrm{N}-2) \pm \frac{1}{2} = \sum_{m=1}^{\mathrm{N}} n_m, \quad \text{for N and } n_s \text{ odd.} \tag{4.16}$$

To keep the calculations tractable, we will consider molecular packing models for microfibril sizes $\mathrm{N} = 3, \ldots, 6$. Setting $n_s = 7$ in equations (4.15) and (4.16) yields 37 distinct solutions, which we denote as $\{n_1, \ldots, n_\mathrm{N}\}$ (see Table 4.3 of Appendix 4.10.8 for details). The curly braces indicate that the N-tuple is unordered, by the virtue of addition being commutative. These solutions however, do not account for all distinct molecular packing models. The order of the internal angles ψ_m in the N-gon will in general change the set of pairwise strip-strip interactions that may be obtained in a microfibril. Accounting for all distinct permutations of the internal angles yields a total of 270 distinct packing models that we analyse in this work - see Appendix 4.10.2 for details. Whenever referring to a molecular packing with an ordered set of vertices, we will use round brackets, i.e. $(n_1, \ldots, n_\mathrm{N})$ or $(\psi_1, \ldots, \psi_\mathrm{N})$.

4.5 STABILITY OF AXIALLY PERIODIC MICROFIBRILS

The next step is to quantify and compare the stability of different axially periodic microfibrils across the molecular packing models introduced in the previous section. A version of this problem has been previously considered by Puszkarska *et al.* [95]. Here, we extend their methodology to account for azimuthal degrees of freedom and to allow comparison between the thermodynamic stabilities of different 3-dimensional molecular packing models.

For a given molecular packing model, each collagen molecule has n_s azimuthal orientations that maximise the contact area between the interacting helical strips. For both odd and even numbered microfibrils we will assume that the azimuthal misalignment penalty in these conformations is negligible, meaning that $\Phi(\theta_m)\Phi(\theta_{m'}) \approx 1$. For two interacting molecules m and m' positioned in one of the optimal azimuthal orientations (as shown in Figure 4.2C), we can write the remaining axial energy contribution in equation (4.11) as

$$\begin{aligned} P_m(\theta_m, \theta_{m'}, \Delta z_m, g) &= \mathrm{E}^p_{i\text{-}j}(\Delta z_m, g) + \mathrm{E}^p_{i'\text{-}j''}(\Delta z_m, g) + \mathrm{E}^p_{i''\text{-}j'}(\Delta z_m, g) \\ &\equiv \mathrm{U}^p_{i\text{-}j}(\Delta z_m, g), \end{aligned} \tag{4.17}$$

where $i, j = 1, \ldots, n_s$ denote the spatially closest interacting spiral strips that correspond to the azimuthal orientations $\theta_m, \theta_{m'}$. We also account for the interactions between helical strips i', i'' and j', j'', which correspond to the azimuthally closest strips to i and j in the anticlockwise and clockwise sense respectively.

We have previously established in equations (4.1) and (4.2) that the requirement of axial periodicity in a microfibril imposes a constraint on the axial staggers of the comprising collagen molecules as well as the gap size. In order to compare the stabilities of axially periodic structures, we look for minima in the interaction potentials $\mathrm{U}^p_{i\text{-}j}$, subject to the following constraints

$$\begin{aligned} \underset{\mathcal{D}}{\text{minimize}} \quad & \mathrm{U}^p_{i\text{-}j}(\Delta z_m, g), \\ \text{subject to} \quad & \Delta z_m = k\mathcal{D}, \\ & g = n\mathcal{D} - L, \\ & \frac{L}{n} < \mathcal{D} < \min\left\{\frac{L}{n}, \frac{L}{n-k}\right\}, \end{aligned} \tag{4.18}$$

where $k = 1, \ldots, n-1$. The steps taken to numerically solve the minimisation problem (4.18) are detailed in Appendix 4.10.5.

Let us start by denoting the set of all solutions to the optimisation problem (4.18) by

$$\mathcal{M}^k_{i\text{-}j} = \left\{ (\Delta z^*_m, g^*) \,\middle|\, \begin{array}{l} (\Delta z^*_m, g^*) \text{ satisfy optimisation problem (4.18),} \\ \text{for } k = 1 \ldots, n-1 \text{ and } i, j = 1, \ldots, 7. \end{array} \right\}. \tag{4.19}$$

Following the work of Puszkarska *et al.*, we assign every ordered pair in $\mathcal{M}^k_{i\text{-}j}$ a significance score, which is defined as

$$\alpha^k_{i\text{-}j}(\Delta z^*_m, g^*) = \frac{\mu_{\mathcal{D}}\left[\mathrm{U}^p_{i\text{-}j}\right] - \mathrm{U}^p_{i\text{-}j}(\Delta z^*_m, g^*)}{\sigma_{\mathcal{D}}[\mathrm{U}^p_{i\text{-}j}]}, \quad \text{for } (\Delta z^*_m, g^*) \in \mathcal{M}^k_{i\text{-}j}, \tag{4.20}$$

where the operators $\mu_{\mathcal{D}}[\ldots]$ and $\sigma_{\mathcal{D}}[\ldots]$ denote taking the mean and standard deviation across all values of pairwise interaction energy specified by the constraints in equation (4.18). For each interaction potential $\mathrm{U}^p_{i\text{-}j}$, we use the significance scores to discard molecular interactions that are expected to lead to relatively unstable axially periodic microfibrils. For subsequent calculations, we only keep the ordered pairs that satisfy

$$\alpha^k_{i\text{-}j}(\Delta z^*_m, g^*) > \mu_\alpha\left[\alpha^k_{i\text{-}j}\right] + c_\alpha \sigma_\alpha\left[\alpha^k_{i\text{-}j}\right], \tag{4.21}$$

where c_a is a positive parameter, $\mu_\alpha[\ldots]$ and $\sigma_a[\ldots]$ denote taking the mean and standard deviation of the significance scores across all interaction strip pairs i-j and values of k. The value of the parameter c_α is selected empirically and in all simulations we set $c_\alpha = 1.73$ (see Appendix 4.10.7 for details of the selection procedure).

For significant interaction minima of $\mathrm{U}^{p}_{i\text{-}j}$ we proceed to calculate the microfibrillar energies. Let us denote the spatially closest interacting pair of helical strips between molecules m and m' by i_m-j_m. For a given molecular packing model, specified by the ordered N-tuple $(n_1, n_2, \ldots, n_{\mathrm{N}})$, the set of interacting strips pairs i_m-j_m that may be present in a given microfibril, is constrained by

$$j_{m''} = ((i_m - n_m - 1) \bmod n_s) + 1. \tag{4.22}$$

For each one of the molecular packing models shown in Table 4.3, we exhaustively determine all axially periodic microfibrils that may be constructed from the significant pairwise molecular interactions. Such microfibrils have energies

$$\mathrm{E_M} = \sum_{m=1}^{\mathrm{N}} \mathrm{U}^{p}_{i_m\text{-}j_m}(\Delta z^*_m, g^*), \tag{4.23}$$

where each $(\Delta z^*_m, g^*)$ and i_m-j_m satisfy conditions (4.21) and (4.22) respectively. In the instance that n_m is either 1 or 6, some of the nearest neighbour interactions $\mathrm{E}_{i''\text{-}j''}$ and $\mathrm{E}_{i''\text{-}j'}$ are expected to be sterically screened. To account for this, we introduce a 25% energy penalty on the interaction potentials $\mathrm{U}^{p}_{i_m\text{-}j_m}$ in equation (4.23) that are subject to steric screening (see Appendix 4.10.6 for details).

4.6 PAIRWISE RESIDUE INTERACTIONS

To ascertain the stability of axially periodic microfibrils, it remains to specify the potentials $\mathrm{U}^{p}_{i\text{-}j}$. The detailed construction of $\mathrm{U}^{p}_{i\text{-}j}$ has been recounted in chapter 3, here we provide only the necessary details.

The net interaction energy of a spiral strip pair i-j is given by the sum of pairwise interactions between the residues that comprise the spiral strips. Among the 20 proteinogenic residues, we identify Asp, Glu Tyr, Cys, Arg, Lys and His as the set of charged residues $\mathcal{C}$, which possess ionisable side chains. The remaining residues are designated as non-charged. Let $\mathrm{R_p}$, $\mathrm{R_q}$ denote the residues on strips i, j respectively with integers p, q denoting the sequential residue positions. The pairwise interaction energy is then written as a sum of two contributions:

$$\mathrm{E}_{i\text{-}j}(\Delta z) = \sum_{\mathrm{R_p\ or\ R_q} \notin \mathcal{C}} \Psi^{\mathrm{MJ}}_{\mathrm{pq}}(r_{\mathrm{pq}}(\Delta z)) + \sum_{\mathrm{R_p, R_q} \in \mathcal{C}} \Psi^{\mathrm{DH}}_{\mathrm{pq}}(r_{\mathrm{pq}}(\Delta z)), \tag{4.24}$$

where $r_{\mathrm{pq}}(\Delta z)$ is the inter-residue distance and the summations are taken over all pairs of interacting residues $\mathrm{R_p}$, $\mathrm{R_q}$ at a given Δz.

Pairwise residue interactions that involve at least one uncharged residue are described by a step potential $\Psi^{\mathrm{MJ}}_{\mathrm{pq}}$. The interaction strengths between the residues are given by the

Mijazawa-Jernigan contact potentials (MJCP), which correspond to the first order approximation to the free energy of interaction between residues in an environment dominated by protein-protein interactions. Remaining charged-charged interactions between residues distance r apart are described by a potential of the Debye-Hückel form

$$\Psi_{\mathrm{pq}}^{\mathrm{DH}}(r) \propto \begin{cases} z^{\mathrm{eff}}(\mathrm{R_p}) z^{\mathrm{eff}}(\mathrm{R_q}) e^{-\kappa r}/r \equiv \psi(r), \text{ for } r > a_0, \\ \psi(a_0), \text{ for } r \leq a_0, \end{cases} \tag{4.25}$$

where κ is the inverse Debye length and $z^{\mathrm{eff}}(\mathrm{R_p})$ denotes the effective charge of residue $\mathrm{R_p}$, which depends on the external ionic conditions and thermodynamic properties of the residue side chain. The detailed dependence of charged-charged residue interactions on ionic conditions has been recounted in chapter 3. Here, we note that throughout all calculations, we set the ionic strength $I = 150\,\mathrm{mM}$ and $\mathrm{pH} = 7.40$, corresponding to physiological conditions [77]. The residue pKa values are set to the reference values measured in alanine pentapeptides [85]. The parameter a_0 may be interpreted as controlling the strength of charged-charged residue interactions relative to those given by MJCP. Various FLS forms of collagen typically form in presence of charged macromolecules, suggesting that charged-charged interactions may play an important role in controlling their emergent axial periodicity [34, 120]. Addition of charged solutes can in principle introduce a range of complex interactions, including pairwise collagen-solute as well as many-body bridging interactions in which the solute mediates electrostatic interactions between several collagen molecules [34, 120]. Other studies have successfully modelled the effect of changing collagen-solvent interactions by varying effective interaction strengths between residues [45]. For the purposes of this study, we adopt a similar simplified approach, wherein we reflect the complex changes in collagen-solute electrostatic interactions through variations in the parameter a_0. We express the magnitude of charged-charged interactions in relation to the average strength of interactions involving at least one uncharged residue:

$$\chi_{\mathcal{C}} = \frac{\psi(a_0)}{\mu_w \left[\left| \Psi_{\mathrm{pq}}^{\mathrm{MJ}}(0) \right| \right]}, \quad \text{for} \quad \mathrm{R_p \, or \, R_q} \notin \mathcal{C}, \tag{4.26}$$

where $\mu_w[\dots]$ denotes averaging over interaction strengths of distinct residue pairs weighted by their abundance in the collagen amino acid sequence (see Appendix 4.10.3 for details). Throughout this study, all of the pairwise molecular interactions will be calculated for rat $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ collagen.

4.7 RESULTS

STABILITY OF D-BANDED MICROFIBRILS

We now proceed to quantify the thermodynamic stability of axially periodic microfibrils across the family of cyclic molecular packing models introduced in section 4.4. To that end, we start by determining the most energetically stable molecular packings at different maximum charged-charged residue interaction strengths $\psi(a_0)$ for each microfibril size $\mathrm{N} = 3, \ldots, 6$.

An important test for our model is that it should predict thermodynamically stable D-periodic microfibrils, which are ordinarily observed during collagen self-assembly. Figure 4.3C illustrates how different values of $\psi(a_0)$ affect the axial periods $\mathcal{D}$ of the most energetically stable 5-membered molecular packing models, which are commonly used to model D-periodic collagen microfibrils [104, 109, 82]. We notice that the characteristic D-periodicity of $\approx 667\,\text{Å}$ appears in the most stable 5-membered microfibrils for sufficiently strong charged-charged residue interaction strengths, namely for $\psi(a_0) \geq 0.6\,k_BT$. This result agrees with the general understanding that charged residue interactions are vital for the self-assembly of collagen fibrils with D-banding [77]. Furthermore, we find that the molecular packing in all D-banded microfibrils in Figure 4.3C, corresponds to that of the Smith microfibril [104], which is identified with the solution $\{n_1, \ldots, n_5\} = \{2, \ldots, 2\}$ of equation (4.16).

Next, we compare the thermodynamic stabilities of microfibrils with different numbers of molecules - see Figure 4.3E. The 5-membered cyclical molecular packings minimise the microfibril energy across all aggregate sizes $\mathrm{N} = 3, \ldots, 6$, whenever the maximum strength of charged-charged residue interactions falls into one of three intervals: $[0.15\,k_BT, 0.3\,k_BT]$, $[0.8\,k_BT, 1.0\,k_BT]$ or $[1.4\,k_BT, 1.5\,k_BT]$. Along the first interval, as discussed previously, we do not observe D-banding in the most stable axially periodic microfibrils with $\mathrm{N} = 5$. On the other hand, along the two other intervals $[0.8\,k_BT, 1.0\,k_BT]$ and $[1.4\,k_BT, 1.5\,k_BT]$, the charged-charged residue interactions are sufficiently strong such that the most stable 5-membered axially periodic microfibril displays D-banding. As can be seen in Figure 4.3E, for the remaining values of $\psi(a_0)$ outside of the three aforementioned intervals, we find that the most stable axially periodic microfibrils are 6-membered. Microfibrils comprised of 3 or 4 collagen molecules never minimise the microfibril energy across all studied values of aggregate size N and generally have higher microfibril energies than microfibrils containing 5 or 6 triple helices. This finding is not surprising - there are only 7 distinct molecular packing models for $\mathrm{N} = 3, 4$, with the remaining 263 distinct models describing microfibrils with $\mathrm{N} = 5, 6$ (see Table 4.3). The molecular packing-dependent constraint (4.22) thus severely limits the set of admissible pairs of interacting spiral strips that may appear in

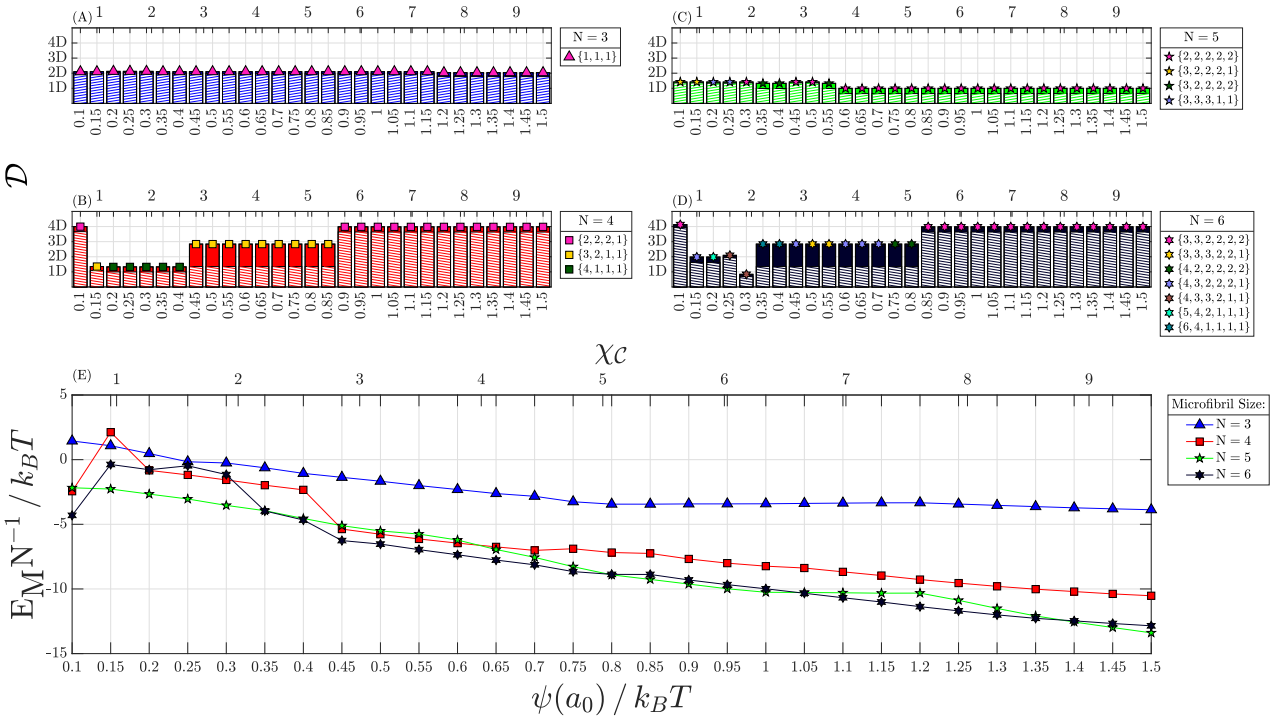


Figure 4.3: (A)-(D). Axial periods of the most energetically stable molecular packing models as a function of maximum charged-charged interaction strength $\psi(a_0)$ for microfibril sizes $\mathrm{N} = 3, \ldots, 6$. The legend groups together microfibrils which have the same combinations of internal angles $\psi_m = 2\pi n_m/7$ without regard for the order of the vertices. The internal polygon angles are written in the form $\{n_1, \ldots, n_\mathrm{N}\}$ (curly braces indicate that the N-tuple is unordered). The hatched and filled portions of each bar represent the length of the gap and overlap regions respectively. (E). Microfibrillar energy per molecule of the most stable molecular packing models for each N as a function of $\psi(a_0)$. All data is illustrated for $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen.

microfibrils with $\mathrm{N} = 3, 4$, leading to less energetically stable aggregates.

Our next goal is to understand the physical mechanisms that lead to the energetic selection of the Smith microfibril as the most thermodynamically stable molecular packing model. Figures 4.12A-D and 4.13A-D of Appendix 4.10.8 characterise the azimuthal and axial degrees of freedom respectively that are associated with the intermolecular interactions of the most thermodynamically stable microfibrils in Figures 4.3A-D respectively. We first notice in Figure 4.12C that all 5-membered microfibrils with D-banding are characterised by 5 pairwise interactions between strips 6 and 4. For convenience, we will denote this set of interacting strips by $(6\text{-}4)_5$, where the subscript identifies the number of times a given strip-strip interaction is repeated in a microfibril. The set of strips $(6\text{-}4)_5$ is in fact unique to the Smith microfibril, and does not appear by itself in any other molecular packing

model. To see this, recall that for a given molecular packing model, the set of interacting spiral strips is determined by the constraint (4.22). Substituting $i_m = 6$ and $j_{m''} = 4$, the only admissible solution to equation (4.22) is $n_m = 2$ for all m, which uniquely corresponds to the Smith microfibril - see Table 4.3 for an exhaustive list of molecular packing models.

We have now established that the 5-membered Smith microfibril has a unique set of interacting strips $(6\text{-}4)_5$. We can, however, make an even stronger statement - the axial staggers Δz_m and gap sizes g in 5-membered axially periodic microfibrils are unique and cannot be realised in axially periodic microfibrils with $\mathrm{N} = 3, 4$ or 6. This can be easily seen from equation (4.18). For any given value of Δz_m, two axially periodic microfibrils with different numbers of molecules N_1 and N_2 can have the same gap size provided that N_1 and N_2 share at least a single divisor that is not unity (recall from section 4.3 that we require that $n \mid \mathrm{N}$ for $n \geq 2$ and that $g = n\mathcal{D} - L$). Since 5 is relatively prime to $3, 4$ and 6, the values of Δz_m and g that characterise 5-membered, axially periodic microfibrils are unique among axially periodic microfibrils with $\mathrm{N} = 3, \ldots, 6$. This in turn means that the set of residue-residue interactions that stabilises the 5-membered axially periodic microfibrils in Figure 4.3E is unique as well. Figure 4.14A of Appendix 4.10.8 shows the physical consequence of this fact: for $\psi(a_0) \geq 0.7\, k_B T$, the global energy minimum of

Table 4.1: Summary of model predictions for different axial periodicities observed in collagen aggregates.

Classification	Axial Period / Å	[a]Stable Microfibril?	N
[c]Short D-periodic	≈ 500-600 [67]	✓	6
D-periodic	≈ 620-670 [67, 16]	✓	5, 6
[d] FLS III	≈ 1000 [22, 34]	N/A	N/A
[e]Dispersed FLS	≈ 790-1000 [33]	✓	4, 5
[e]Compact FLS	≈ 1000-1500 [44, 69, 33]	✓	3, 6
FLS IV	≈ 1700 [22, 34]	✓	6
FLS I	≈ 2300-2700 [22, 34, 97]	✓	4, 6
[f]FLS II	≈ 2300-2600 [22, 34]	N/A	N/A

[a] A stable microfibril is defined in the sense specified by equations (4.21) and (4.22). Furthermore, we exclude any aggregates for which our model predicts a net repulsive microfibrillar energy.
[b] We report all distinct aggregate sizes predicted by our model for stable microfibrils.
[c] We note that short D-periodic fibrils listed in this table are homotypic and comprised of type I collagen [67]. These fibrils are distinct from heterotypic fibrils incorporating type I and type III collagens, which also display unusually short D-spacing [17, 105].
[d] FLS III is distinct from FLS observed *in vivo* by the virtue of having approximately equal proportions of gap and overlap regions, which cannot be rationalised as arising at the level of a single microfibril [22].
[e] These forms of collagen are observed predominantly *in vivo*.
[f] FLS II is distinct from FLS I in having two overlap regions per $\mathcal{D}$-period.

potential $U^{p}_{4\text{-}6}$ is the most stable for N = 5. As can be further seen in Figure 4.14B, this global energy minimum corresponds to the axial stagger in D-banded microfibrils. We therefore conclude that strong interactions between spiral strips 6 and 4 selectively stabilise 5-membered D-banded microfibrils with the molecular packing of the Smith microfibril, which can then lead to their global stability among microfibrils with other values of N. This finding complements our result in section 2.3.4 of chapter 2. There, we demonstrated that the spatial residue organisation leads to preferential aggregation into microfibrils with N = 5, whilst noting that specific residue interactions are required to stabilise pentameric microfibrils. Here, we have identified the aforesaid specific residue interactions in addition to showing the energetic selection of 5-membered microfibrils across a significantly broader class of molecular packing models.

Next, we ascertain further details of the interactions that selectively stabilise the 5-membered Smith microfibril in Figure 4.13C, which illustrates the axial staggers associated with thermodynamically stable microfibrils in Figure 4.3C. We see that 5-membered microfibrils with D-banding are characterised by intermolecular interactions at a single stagger: either $\Delta z_m = 667\,\text{Å}$ for $0.6\,k_BT \leq \psi(a_0) \leq 1.2\,k_BT$ or $\Delta z_m = 2673\,\text{Å}$ for $1.25\,k_BT \leq \psi(a_0) \leq 1.5\,k_BT$. This observation is significant, as the typical criterion used as a proxy for *in silico* prediction of D-banding in collagen aggregates, is the presence of minima in the pairwise triple helix interaction potential at $\Delta z = k\mathcal{D}$, for all $k = 1, 2, \ldots, n-1$ [95]. As we have demonstrated, specific 3-dimensional molecular packing models, such as the Smith microfibril, can significantly relax this requirement, necessitating that there is a single interaction energy minimum between some pair of interacting strips i-j at $\Delta z = k\mathcal{D}$ for just one of $k = 1, 2, \ldots, n-1$.

Finally, we return to Figure 4.13C to remark that at $\psi(a_0) = 1.25\,k_BT$ we observe a switch from thermodynamically stable D-periodic microfibrils having axial staggers $\Delta z_m = 667\,\text{Å}$ to those with $\Delta z_m = 2673\,\text{Å}$. This corresponds to a change from left-handed to right-handed microfibril chirality that we have defined previously in chapter 2 [124]. Microfibril chirality reflects the chiral spatial organisation of gap regions, which have been suggested to act as nucleation sites for the mineral phase in bone, which itself is characterised by left-handed chirality [101]. We have previously suggested that microfibril chirality may be important for controlling the emergent architecture of the mineral phase in bone [124]. Our current finding is thus significant, as it demonstrates that the microfibril chirality is controlled by electrostatic interactions.

STABILITY OF FLS MICROFIBRILS

We now turn our attention to the broad class of collagen aggregates that are characterised by axial periodicities that are different from D-banding - see Table 4.1. Firstly, we identify

the types of polymorphic collagenous aggregates that can arise from a single underlying microfibrillar structure. For that, the axial period $\mathcal{D}$ and the length of the gap region L_{g} must obey the linear relationship (4.9), which is visualised in Figure 4.4A. For FLS III fibrils, we see that the ratio of L_g to the $\mathcal{D}$-period cannot be explained by equation (4.9), suggesting that the axial period emerges at supramicrofibrillar lengthscales, which is in line with previous studies [22, 34]. FLS II fibrils are characterised by two overlap regions per $\mathcal{D}$-period and have been similarly suggested to form at the supramicrofibrillar level [22]. The values of $\mathcal{D}$ and L_{g} have been experimentally measured in FLS I and FLS IV collagens and are consistent with equation (4.9) and can thus be explained with a microfibrillar structure [22]. For short D-banding as well as compact and dispersed FLS collagens we are not aware of any measurements of gap/overlap lengths that were performed on large populations of fibrils - only measurements of the overall $\mathcal{D}$-period have been reported. As such, for the rest of this work, we will assume that the short D-banding, dispersed and compact FLS collagens can be associated to microfibrils with axial periods falling in the relevant experimentally measured range in Table 4.1.

As can be seen in Figure 4.3E, for maximum charged-charged interaction strengths in the interval $[1.05\,k_BT, 1.35\,k_BT]$ as well as for $\psi(a_0) = 0.1\,k_BT$, the most thermodynamically stable collagen microfibril is 6-membered and is characterised by an axial period $\mathcal{D} \approx 2668\,\text{Å}$, which corresponds closely to the FLS I periodicity observed experimentally. Microfibrils with N = 6 are also energetically selected as the most stable aggregate size for $\psi(a_0)$ in the interval $[0.35\,k_BT, 0.75\,k_BT]$. Interestingly, in this range of charged-charged interaction strengths, they are characterised by an axial period $\mathcal{D} \approx 1902\,\text{Å}$, which to our knowledge, has not been observed experimentally.

Next, we extend the analysis of the previous sections by determining for each parameter value N and $\psi(a_0)$ all distinct axial periods predicted by our model, not just those that correspond to axially periodic microfibrils that are globally stable for some choices of the aforementioned parameter values. For a given value of N, we will analyse distinct axial periods that consistently appear across multiple values of maximum charged-charged interaction strength $\psi(a_0)$ (see Appendix 4.10.4 for details). Figures 4.4B-E illustrate all distinct axial periods across different microfibril sizes N in the L_g-$\mathcal{D}$ plane. The size of each point set to be proportional to the magnitude of microfibril energy per unit molecule. We find that the axial periodicities of FLS I, FLS IV, dispersed FLS and compact FLS collagens are all predicted by our model - see Table 4.4 of Appendix 4.10.8 for exact numerical values. Physically, this means that the majority of the FLS periodicities are encoded by pairwise interactions between parallel collagen molecules. This finding is important, as it is contrary to previous analyses of the sequence-structure relationship for FLS collagens, which found no energetic preference for FLS-like axial arrangements of parallel triple helices [34, 120]. As mentioned in the previous section, we also find D-periodicities. Curiously,

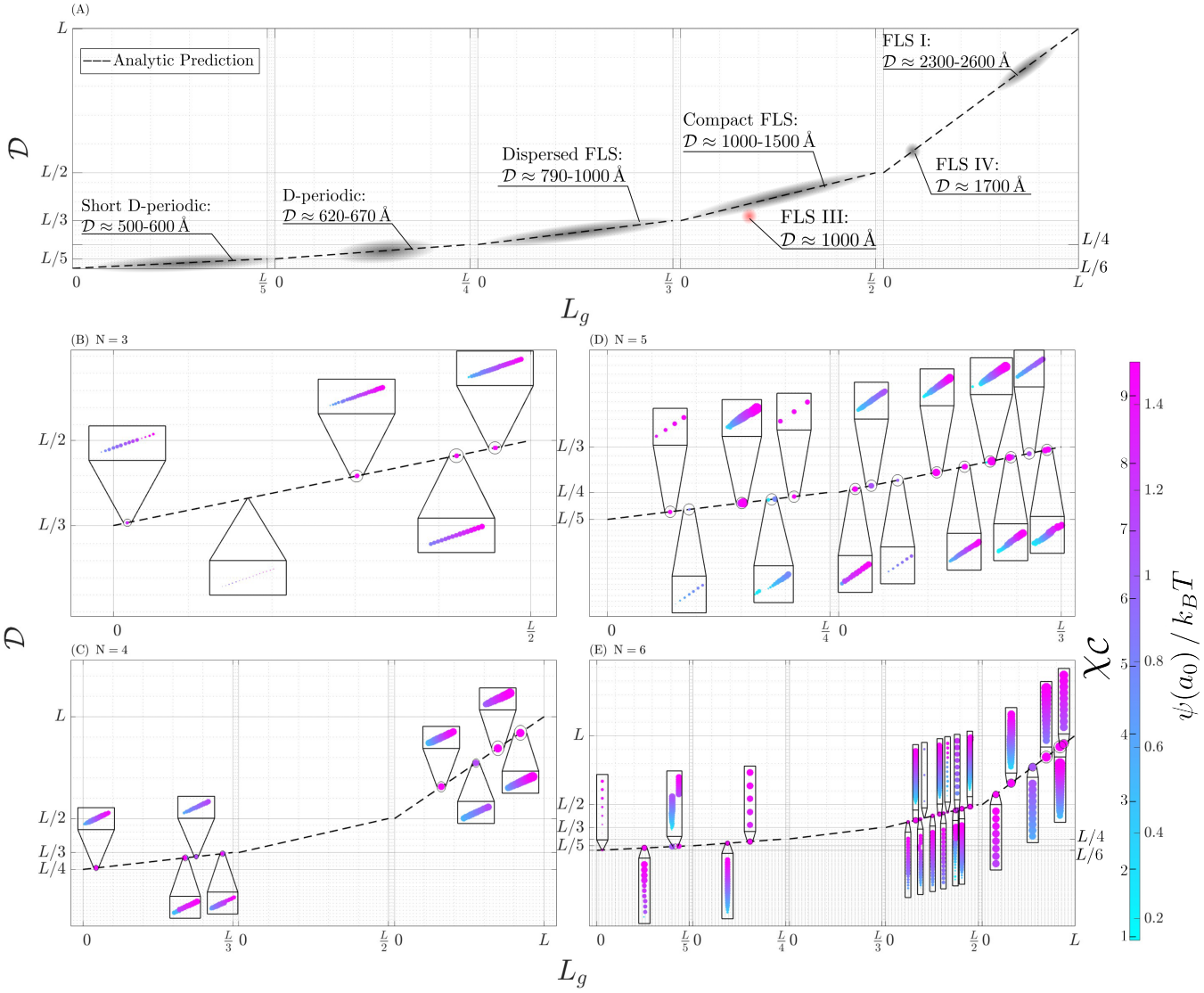


Figure 4.4: (A). Summary of experimentally observed periodicities in collagen aggregates. (B)-(E). All distinct axial periods and corresponding gap region lengths in α_2(I)[α_1(I)]$_2$ rat collagen for microfibril sizes N = 3, . . . , 6. Black circles show the average $\mathcal{D}$-periods across multiple values of $\psi(a_0)$ (periodic signatures). The size of each marker is proportional to the magnitude of the microfibril energy per unit molecule $|\mathrm{E_M}|\,\mathrm{N}^{-1}$ (excluding microfibrils with net repulsive interactions). The rectangular boxes emanating from the periodic signatures illustrate the value of $|\mathrm{E_M}|\,\mathrm{N}^{-1}$ for microfibrils with a given axial period as a function of charged-charged interaction strength. The exact numerical values of the periodic signatures are summarised in Table 4.4 of Appendix 4.10.8. All data is illustrated for α_2(I)[α_1(I)]$_2$ rat collagen.

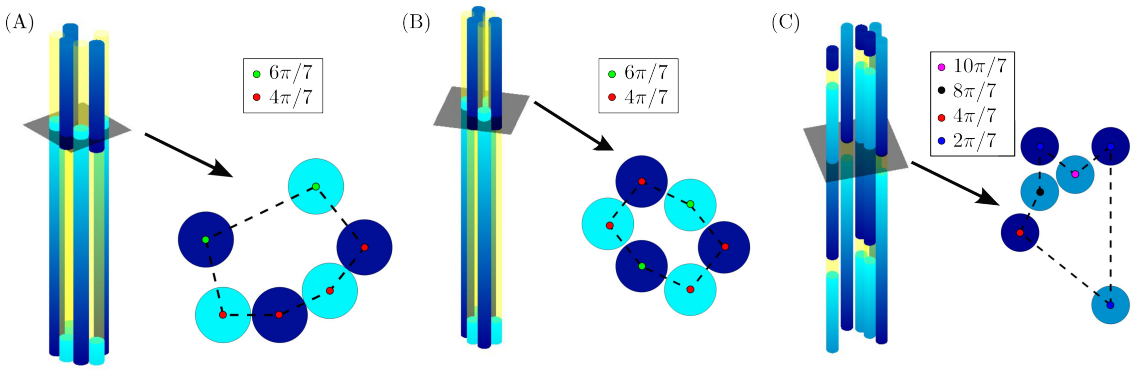


Figure 4.5: 3-dimensional molecular packings predicted by our model for (A)-(B). FLS I and (C). FLS IV collagen microfibrils.

as can be seen in Figure 4.3D and Table 4.4, 6-membered microfibrils also display axial periods $\mathcal{D} \approx 669\,\text{Å}$. However, such microfibrils are substantially less stable in comparison to N = 5 microfibrils with D-banding. Finally, we also predict the existence of 6-membered microfibrils with short D-spacing, which have been previously reported during early stages of fibrillogenesis of type I collagen [67].

FLS I and FLS IV collagens can form either 4 or 6-membered microfibrils (see Figures 4.4C, E), with a strong energetic preference for 6-membered packings. Among the dispersed and compact FLS that predominantly form *in vivo*, we notice a clear separation in terms of the molecular packing models that are associated with each FLS collagen type - see Figures 4.4B-E. Dispersed FLS collagen microfibrils are always comprised of $N = 4, 5$ molecules, whilst those with compact FLS axial periodicities are characterised by $N = 3, 6$.

Unlike other FLS microfibrils, FLS IV microfibrils are unique in that their energy is largely independent of the maximum strength of charged-charged interactions - see Figure 4.4E. This suggests that hydrophobic-hydrophobic and/or hydrophobic-charged residue interactions may play an important role in encoding FLS IV periodicities. For FLS I, dispersed/compact FLS, D-banded and short D-banded microfibrils, we generally find that the microfibril energy becomes more stable with increasing strength of charged-charged interactions, suggesting their importance for stabilising these polymorphic collagen forms.

Finally, we are able to determine details of the 3-dimensional molecular packing in FLS collagens. As can be seen in Figures 4.15A-D, dispersed/compact FLS are associated with a range of different molecular packing models, with no single one standing out as preferential. For FLS I and IV collagens, however, there is a clear preference in molecular packing models with internal angles $\{6\pi/7, 6\pi/7, 4\pi/7, 4\pi/7, 4\pi/7, 4\pi/7\}$ and $\{10\pi/7, 8\pi/7, 4\pi/7, 2\pi/7, 2\pi/7, 2\pi/7\}$ respectively - see Figures 4.15D and 4.3D. There are a total of 23 distinct molecular packing models with these internal angles, as can be seen

in Table 4.3. Out of these 23 molecular packing models, we find that only 3 distinct ones are realised in the most energetically stable FLS I and FLS IV microfibrils - see Figure 4.5 above and Figure 4.16 of Appendix 4.10.8.

4.8 DISCUSSION

Collagen possesses a remarkable ability to aggregate into structures with periodic molecular ordering along the axial direction. It has been known since at least the 1950s that the axial molecular ordering in collagen fibrils is highly polymorphic [22, 34, 120]. This is best exemplified by FLS (fibrous long-spacing) fibrils, which display axial periodicities in the range $\approx$ 1000-2700 Å, thus significantly exceeding the well-known D $\approx$ 670 Å axial period. Past and current experimental and theoretical literature has heavily focused on D-periodic collagen fibrils, leaving unaddressed several important questions pertaining to the polymorphism associated with axial periodicity in collagen fibrils.

In this work, we formulated a physically motivated quantitative relationship between the key structural descriptors of the microfibril - length of the triple helix L, number of triple helices in a microfibril N and the generalised axial periodicity lengthscale $\mathcal{D}$, see equation (4.10). Less general formulations of equation (4.10) are well-known and widely used across literature pertaining to collagen self-assembly. However, crucial details that are essential for understanding the broad spectrum of axial periodicities observed in collagen aggregates are missing. This is best exemplified by the bounds on the admissible $\mathcal{D}$-period values, which have been previously stated as $\frac{L}{\mathrm{N}} \leq \mathcal{D} \leq \frac{L}{\mathrm{N}-1}$ in microfibrils comprised of N molecules [95]. This notably precludes one from explaining the self-assembly of FLS collagens observed *in vitro*, which possess axial periods in the range between 1700 Å and 2700 Å [22, 34]. Our formulation rectifies this shortcoming, allowing for a largest upper bound of $\mathcal{D} < L$ in microfibrils with an even number of molecules. This allows us to explain the existence of FLS axial periodicities observed *in vitro* at the microfibrillar level. Since current experimental evidence points towards existence of a staggered microfibrillar substructure in FLS collagens, it is crucial that theoretical models of axial periodicity account for it [97, 117].

Using our previously developed microfibril self-assembly model (see chapters 2 and 3) [124], we investigated the thermodynamic stability of different axially periodic microfibrils in $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen across different charged-charged residue interaction strengths and 3-dimensional molecular packing models. We established an important fact: the axial periodicities of FLS collagens are encoded by pairwise interactions between parallel triple helices, see Table 4.1 and Figures 4.4B-E. This observation applies both to FLS I and FLS IV collagen , which are observed *in vitro* as well as to dispersed and compact FLS

collagen which are predominantly seen *in vivo*. Our findings are in direct contrast to previous studies, which found no evidence for FLS periodicities in parallel collagen-collagen interactions [34, 120]. We rationalise this discrepancy as arising from the lack of physical detail used in previous studies, as the full residue sequence of the triple helix was not known at the time.

In addition to new fundamental understanding of the sequence-structure relationship of FLS collagens, we have proposed, for the first time, an interaction-informed model for 3-dimensional molecular organisation inside FLS microfibrils. In our analysis, we included 270 distinct 3-dimensional molecular packing models, spanning microfibrils with $N = 3, \ldots, 6$. Our molecular packing models are based on the 3-dimensional residue organisation of the triple helix. In particular, they represent the geometric arrangements of triple helices that maximise the contact area between residue interaction sites on the collagen molecular surfaces. We have identified 2 distinct models that are energetically favoured by FLS I microfibrils (see Figures 4.5A,B) and one that is favoured by FLS IV aggregates (see Figure 4.5C). These results provide an excellent opportunity for experimental validation of our theoretical model.

Unlike FLS I and FLS IV microfibrils, for dispersed and compact FLS microfibrils we did not find a single energetically preferential molecular packing model. Instead, we observed a broad range of molecular packing models with comparable microfibril energies, see Figure 4.15. We did, however, observe a clear preference for aggregate size N, with dispersed and compact FLS microfibrils forming whenever $N = 4, 5$ and $N = 3, 6$ respectively, see Table 4.1. Dispersed and compact FLS collagens are known to differ in their morphology, with the former characterised by significant molecular disorder [33]. Our results in turn suggest that the morphological differences may be due to the differences in the underlying microfibrillar structure. The origin of these structural differences can be plausibly explained by the fact that dispersed FLS fibrils are found exclusively in tissues that are associated with active breakdown and/or high turnover of collagen [33]. These structural differences may then arise as a result microfibril remodelling in tissues that are undergoing degradation.

We found that the pairwise molecular interactions that encoded FLS axial staggers generally became stronger with increasing strength of charged-charged residue interactions - see Figures 4.4B-E. The primary exception to this rule were FLS IV microfibrils, the energy of which is largely independent of the strength of electrostatic interactions - see Figure 4.4E. This in turn suggests that hydrophobic-hydrophobic and/or hydrophobic-charged residue interactions play a more important role in formation of FLS IV aggregates, as opposed to FLS I, dispersed and compact FLS where electrostatic interactions are more important. Further study of the role of electrostatic interactions in FLS collagen self-assembly may involve accounting for salt-bridge formation, which may lead certain ionisable residues to partake in intra-helical, rather than inter-helical interactions [89].

The fact that parallel collagen-collagen interactions encode FLS periodicities raises a question of the relative thermodynamic stability of FLS and D-banded aggregates. We found that 5-membered D-banded microfibrils corresponded to a globally stable conformation for charged-charged residue interaction strength $\psi(a_0)$ falling in the ranges $[0.8\,k_BT, 1.0\,k_BT]$ or $[1.4\,k_BT, 1.5\,k_BT]$, see Figure 4.3E. In all such cases we found the underlying molecular packing to be that of the 5-membered Smith microfibril, see Figures 4.3C, E. For the remaining values of $\psi(a_0)$ between $0.1\,k_BT$ and $1.5\,k_BT$, we found that the most thermodynamically stable microfibrils were 6-membered with an axial period $\mathcal{D} \approx 2668\,\text{Å}$ (FLS I) or $\mathcal{D} \approx 1902\,\text{Å}$. The latter axial period appeared only in the interaction regime associated with relatively weak electrostatic interactions and has not been reported in experimental literature to our knowledge. The strong dependence of the relative thermodynamic stability of D-banded and FLS microfibrils is in line with the current understanding in literature that electrostatic interactions play an important role in formation of both aggregates [77, 34].

It remained to explain how the D-banded Smith microfibril may be energetically selected as the globally stable conformation among other microfibrils with $\mathrm{N} = 3, \ldots, 6$. We demonstrated that the set of interacting helical strips observed in the D-banded Smith microfibril was unique, by the virtue of the geometric constraint (4.22) imposed by the 3-dimensional organisation of residues on the collagen molecular surface - see Figure 4.12 and section 4.7. Furthermore, we have shown that the set of admissible axial staggers and gap lengths in 5-membered axially periodic microfibrils is also unique. This is a specific example of a more general fact - two different axially periodic microfibrils comprised of N_1 and N_2 molecules respectively, can only be characterised by the same axial staggers and gaps if $\gcd(\mathrm{N}_1, \mathrm{N}_2) > 1$. This can be easily seen by noting in equation (4.18) that the constraint on the gap size of axially periodic interactions can be simultaneously satisfied for two microfibrils with different numbers of molecules only if they share a common divisor greater than 1. Collectively, the uniqueness of interacting helical strips and their axial conformations implies that an axially periodic Smith microfibril admits a unique set of residue-residue interactions that selectively stabilises the D-banded microfibril as opposed to FLS or any other axially periodic microfibrils - see Figure 4.14.

The results presented in this work utilise a number of simplifying assumptions in construction of the molecular packing models. Our models for the 3-dimensional molecular organisation in microfibrils account for the azimuthal component of the pairwise interaction energy between collagen triple helices. Further refinements to our predictions can be made by fully accounting for the component of the pairwise molecular interactions that depends on the inter-molecular separation in the packing plane. To be precise, when calculating the microfibril energy for a given molecular packing model, we assumed that all interacting triple helices in a microfibril can get within the same distance of closest

approach a_0 for any axial/azimuthal orientations of the interacting collagen molecules. As can be seen from the molecular packing models in Figure 4.5, this is true in the case of molecular packing 4.5B, but not for 4.5A or 4.5C. In practice, this means that our approach may overestimate the microfibril energy for the molecular packing models that are associated with non-equilateral polygons. Accounting for this may further increase the energy difference between the globally stable 5-membered microfibrils with D-banding and 6-membered FLS I microfibrils in Figure 4.3E. Another refinement to our calculations may take form of relaxing the assumption that only nearest-neighbouring molecules connected by polygon edges significantly contribute to the microfibril energy. Indeed, in molecular packing models such as the one in Figure 4.5C, some of the next nearest neighbour interactions are expected to make a significant energetic contribution.

The results of our study raise a number of exciting questions pertaining to the sequence-structure relationship of FLS collagens. Whilst we found that the axial periods $\mathcal{D}$ of compact and dispersed FLS collagens are encoded by pairwise interactions between parallel triple helices, we were not able to ascertain whether the ratio of gap region length to the $\mathcal{D}$-period is in agreement with the measurements of the negative staining patterns in TEM images. This is due to a lack of the aforementioned experimental measurements, which provides an opportunity to test our model's predictions for the microfibrillar structure that underlies the axial periodicities observed in *in vivo* FLS collagens.

Another notable aspect of FLS collagen structure is the fact that their negatively stained banding patterns possess dihedral symmetry, which has led to suggestions that FLS fibrils are comprised of equal proportions of collagen molecules in parallel and anti-parallel axial arrangements [22]. In this study, we considered parallel arrangements of interacting triple helices in a microfibril, which we have shown to encode the axial staggers corresponding to FLS periodicities. Ascertaining the exact role of anti-parallel interactions and indeed the lengthscales at which they first appear during FLS self-assembly presents an exciting avenue for further theoretical and experimental study.

Another crucial aspect of FLS self-assembly is the requirement for long, charged, flexible polymers such as chondroitin sulfate to be present during their formation [22, 16]. In this work, we did not account explicitly for collagen-solute interactions, instead opting for a coarse-grained approach of varying the strength of charged-charged residue interactions between collagen triple helices. Previous studies have suggested that important contributors to the stability of FLS aggregates are the bridging interactions between anti-parallel triple helices, wherein the aforesaid flexible, charged polymers mediate the electrostatic interactions between residues of the same charge [34, 120]. However, these findings have only been established for a single α-chain, rather than the entire triple helix. Furthermore, the impact of the relative 3-dimensional positions of molecules in a microfibril on these bridging interactions remains to be elucidated. It is also possible

that collagen-solute interactions are important for supramicrofibrillar association, wherein they decrease the flexibility of microfibrils with large gap regions, such as those of FLS I collagen [97, 117]. Future research could shed new light on the relative contributions of collagen-collagen and solute bridging interactions towards stabilisation of FLS aggregates.

4.9 CONCLUSION

In this work, we investigated the sequence-structure relationship of fibrous long-spacing (FLS) collagens, which are characterised by axial periods that significantly exceed the D-banding lengthscale. We used our equilibrium model of microfibril self-assembly to demonstrate that pairwise interactions between parallel collagen triple helices encode the axial periodicities of FLS collagens that are observed both *in vitro* and *in vivo*. This result is in direct contrast with existing knowledge in the field [22, 34, 120], which is based on numerical studies that were carried out using a partial residue sequence of the collagen triple helix. Further, we were able, for the first time, to construct 3-dimensional molecular packing models for FLS microfibrils, based on the pairwise triple helix interactions. Finally, we demonstrated that the 3-dimensional spatial organisation of the residues comprising the triple helix allows for selective energetic stabilisation of D-banded microfibrils, as opposed to other axial periodicities encoded in the collagen amino acid sequence. Collectively, our results provide a novel understanding as well as a fresh mechanistic perspective on the polymorphism associated with the axial periodicities of collagen aggregates.

4.10 APPENDIX

4.10.1 Proof that Equation (4.14) Implies Existence of a Simple Polygon with Internal Angles ψ_m

The goal of this section is to answer the following question: given a set of angles ψ_m that satisfies equation (4.14), what is the minimum number of self-intersections that a polygon with internal angles ψ_m has? A general version of this problem has been previously studied by [35]. The authors demonstrated that for a sequence of turning (external) angles $\alpha_1, \ldots, \alpha_{\mathrm{N}}$ of a polygon P with $\alpha_m \in (-\pi, \pi)$, the number of self-intersections of P is given by

$$\mathrm{cr}(P) = |\kappa| - 1, \text{ if } \sum_{m=1}^{\mathrm{N}} \alpha_m = 2\kappa\pi \text{ for } \kappa = 1, 2, \ldots. \tag{4.27}$$

Following the convention of [35], the turning angles are given by $\alpha_m = \pi - \psi_m$. We then immediately find that $\sum_m \alpha_m = 2\pi$. According to equation (4.27) we find that $\mathrm{cr}(P) = 0$

or equivalently that there exists a simple polygon P with internal angles ψ_m given that ψ_m satisfy equation (4.14), answering the question posed at the start of the section.

4.10.2 VERTEX ORDER IN MOLECULAR PACKING MODELS

Each positive integer decomposition satisfying equation (4.15) provides us with an unordered set of internal angles ψ_m that describe the polygon onto which collagen triple helices are packed. The next step is to associate the angles ψ_m to specific vertices of the polygon, or in other words to order ψ_m. For a given set of angles ψ_m, let us define the packing type of a microfibril to be the unordered l-tuple $\{\psi_1^{r_1}, \psi_2^{r_2}, \ldots, \psi_l^{r_l}\}$, where $r_l = 1, \ldots, \mathrm{N}$ is the number of times a distinct internal angle ψ_l repeats in a given packing. It should be noted that multiple positive integer decompositions and hence distinct packings may correspond to the same packing type. As we shall see, however, simple polygons with the same packing type share useful combinatorial properties.

The number of ways to arrange the internal angles ψ_m in a simple polygon is given by the multinomial coefficient $\binom{\mathrm{N}}{r_1,\ldots,r_l} = \mathrm{N}!/\prod_{i=1}^{l} r_i!$. It is clear that not all of these arrangements will correspond to distinct microfibrils. Namely, all arrangements that are equivalent under rotation will have identical sets of pairwise molecular interactions. The problem of enumerating all such arrangements that result in distinct microfibrils is then equivalent to the combinatorial problem of counting the number of necklaces comprised of distinct beads numbered $1, \ldots, l$ that repeat $r_1, \ldots, r_l$ times respectively. For a given microfibril type $\{\psi_1^{r_1}, \psi_2^{r_2}, \ldots, \psi_l^{r_l}\}$, the number of distinct ways to arrange the polygon vertices, whilst taking all rotations to be equivalent, is given by

$$\mathcal{N}(r_1, r_2, \ldots, r_l) = \frac{1}{\mathrm{N}} \sum_{d|\gcd(r_1,\ldots,r_l)} \binom{\mathrm{N}}{\frac{r_1}{d}, \ldots, \frac{r_l}{d}} \phi(d), \tag{4.28}$$

where $\phi(d)$ is the Euler's totient function and d are the divisors of $\gcd(r_1, \ldots, r_l)$, including unity [111]. Table 4.3 shows the values of $\mathcal{N}(r_1, \ldots, r_l)$ for different cyclical polygonal packings between $\mathrm{N} = 3$ and $\mathrm{N} = 6$.

4.10.3 ESTIMATION OF RESIDUE PAIRWISE INTERACTION ENERGIES

To determine the range of values for the parameter a_0 in equation (4.25), we start by estimating the magnitude of pairwise residue interactions. For that purpose, we utilise the MJCP MIYS850103, MIYS960102, MIYS990107 that have been previously applied to the study of collagen self-assembly [60, 95]. Figure 4.6A illustrates the magnitude of MJCP for two classes of pairwise residue interactions that we identify for type I rat collagen. For both interaction classes, we observe substantial variation in the interaction strength. To

broadly quantify their interaction strength we define the mean interaction strength

$$\mu_w \left[\left|\Psi_{\mathrm{pq}}^{\mathrm{MJ}}(0)\right|\right] = \frac{\sum\limits_{l\leq m} w_{lm}\,|\varepsilon_{lm}|}{\sum\limits_{l\leq m} w_{lm}}, \tag{4.29}$$

where ε_{lm} is the matrix element of a 20×20 MJCP matrix, which describes the interaction strength for each distinct residue pair. The weights w_{lm} are determined on the basis of the combined abundance of the interacting residues:

$$w_{lm} = \begin{cases} A_l + A_m, & \text{if } A_l, A_m \neq 0, \\ 0, & \text{if } A_l \text{ or } A_m \text{ is } 0, \end{cases} \tag{4.30}$$

where A_l is the number of times residue l occurs in the collagen amino acid sequence.

Using the interaction strength metric in equation (4.29), we find that for MJCP, the magnitudes of charged-charged interactions and those that involve at least one non-charged

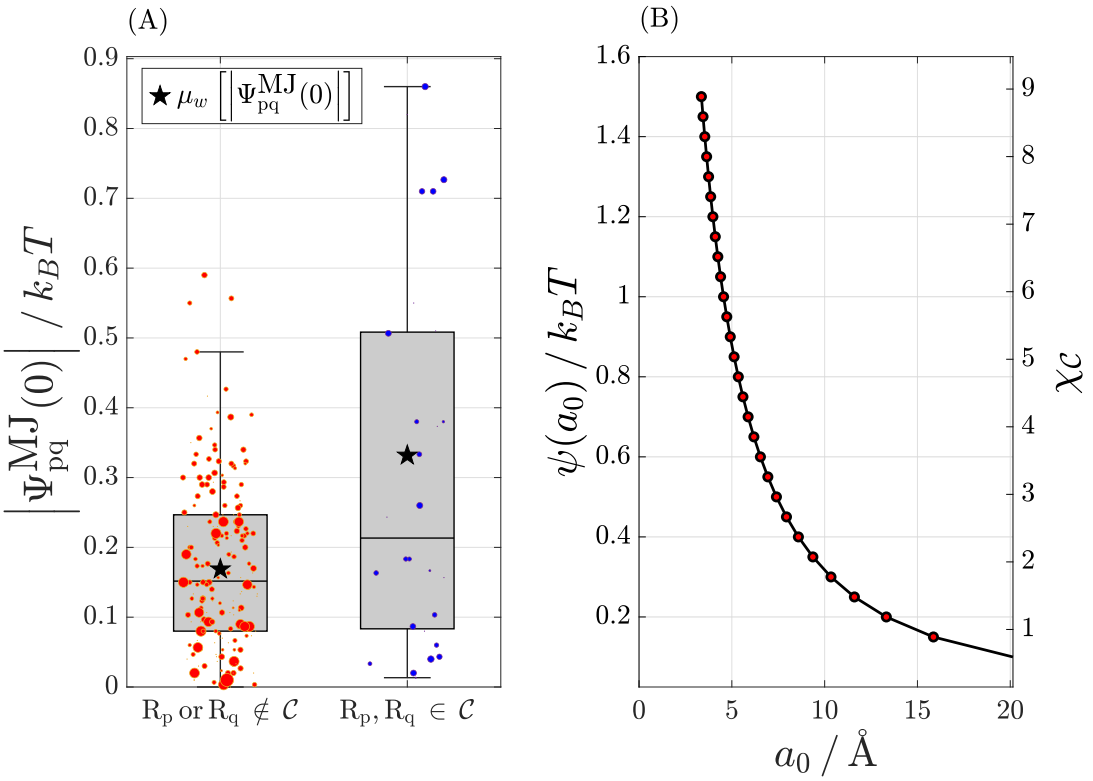


Figure 4.6: (A). Box plot illustrating the magnitude of Miyazawa-Jernigan contact potentials for distinct pairs of interacting residues. The size of each point is proportional to the combined abundance of interacting residues in rat type I collagen. The star shows the mean interaction strength weighted by the number of residues. (B). (LEFT) Magnitude of Debye-Hückel interaction energy at the distance of closest approach $\psi(a_0)$ and (RIGHT) its corresponding ratio to the mean interaction strength for residues $\mathrm{R_p}$ or $\mathrm{R_q}$ non-ionisable.

residue are approximately $0.33\,k_BT$ and $0.16\,k_BT$ respectively. Based on this, we choose a_0 values corresponding $\psi(a_0) \in [0.1\,k_BT, 1.5\,k_BT]$, in increments of $0.05\,k_BT$ (see Figure 4.6B).

4.10.4 CLASSIFICATION OF ALL DISTINCT AXIAL PERIODS

Exhaustive calculation of all axially periodic aggregates that satisfy constraints (4.21) and (4.22) can yield in excess of 4 000 000 distinct microfibrils, depending on the values of N and a_0. The axial periods of the resulting microfibrils do not form a continuous distribution, but instead have a tendency to cluster - see Figure 4.7. We can greatly simplify the analysis of the axial periods emerging from our model by identifying these clusters for each value of parameters N and a_0.

For each N and a_0, we construct clusters as follows:

1. Cluster microfibrils with axial periods $\mathcal{D}$ that are at most within 20 Å of each other.
   a) Use hierarchical clustering with complete linkage in order to constrain the maximum separation between elements of each cluster.
2. Assign a single representative to each cluster generated during step (i).
   a) Discard all microfibrils for which $\mathrm{E_M} > 0$.
   b) If the cluster is non-empty, choose the cluster representative to be the microfibril with axial period ${}^{l}\mathcal{D}_{\mathrm{rep}}$ that minimises $\mathrm{E_M}$ across all members of the l^{th} cluster.

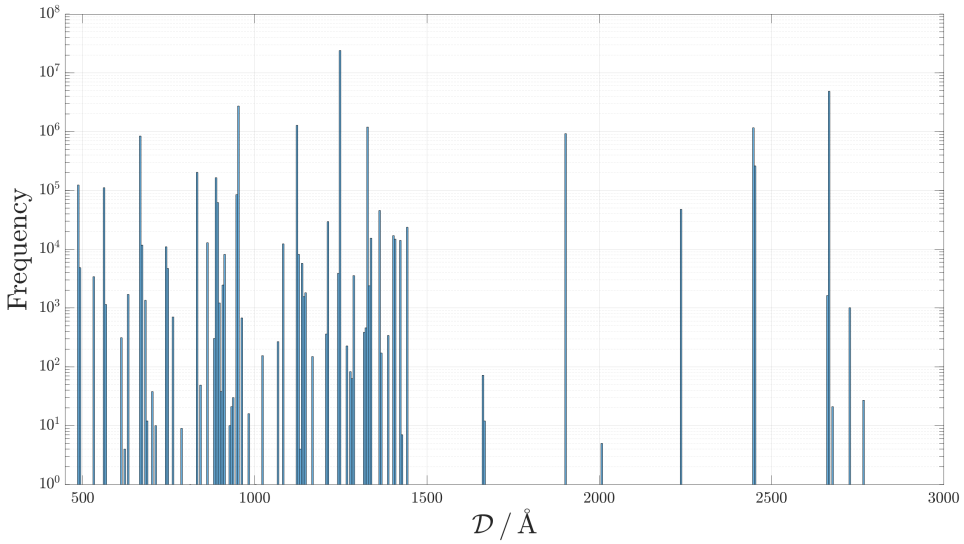


Figure 4.7: Histogram of axial periods of all microfibrils satisfying the constraints (4.21) and (4.22) for $\mathrm{N} = 3, \ldots, 6$ and $\psi(a_0) \in [0.1\,k_BT, 1.5\,k_BT]$. The significance score cut-off parameter is set to $c_\alpha = 1.73$.

The values of ${}^{l}\mathcal{D}_{\text{rep}}$ are illustrated in Figures 4.4B-E. We notice that the axial periods ${}^{l}\mathcal{D}_{\text{rep}}$ again have a strong tendency to cluster around certain values of $\mathcal{D}$. It is sufficient to identify all clusters in Figures 4.4B-E by repeating the hierarchical clustering procedure described above with maximum admissible separation between cluster elements set to 10 Å. We call the mean axial period across each resulting cluster, its periodic signature. The resulting periodic signatures are illustrated in Table 4.4 and are classified according to the experimental measurements of axial periods in Table 4.1. One point necessitates further clarification, namely the distinction between compact and dispersed FLS aggregates. The experimentally measured ranges of axial periods for these collagen aggregates overlap around $\mathcal{D} \approx 1000$ Å. For the purposes of assigning a periodicity classification, we set the upper and lower bounds for the axial periods of dispersed and compact FLS aggregates at $\frac{L}{3}$, for $L = 2900$ Å.

4.10.5 Minimisation of Pairwise Interaction Potentials $\mathrm{U}^{p}_{i\text{-}j}$

The aim of this section is to outline the steps taken to solve the optimisation problem (4.18) of the main text. The first step is to calculate the value of the pairwise interaction energy for each admissible axial period $\mathcal{D}$. We note from equation (4.24) that calculations of pairwise residue interactions involve using the step potential $\Psi^{\text{MJ}}_{\text{pq}}$ of finite width $l_c = 7.5$ Å.

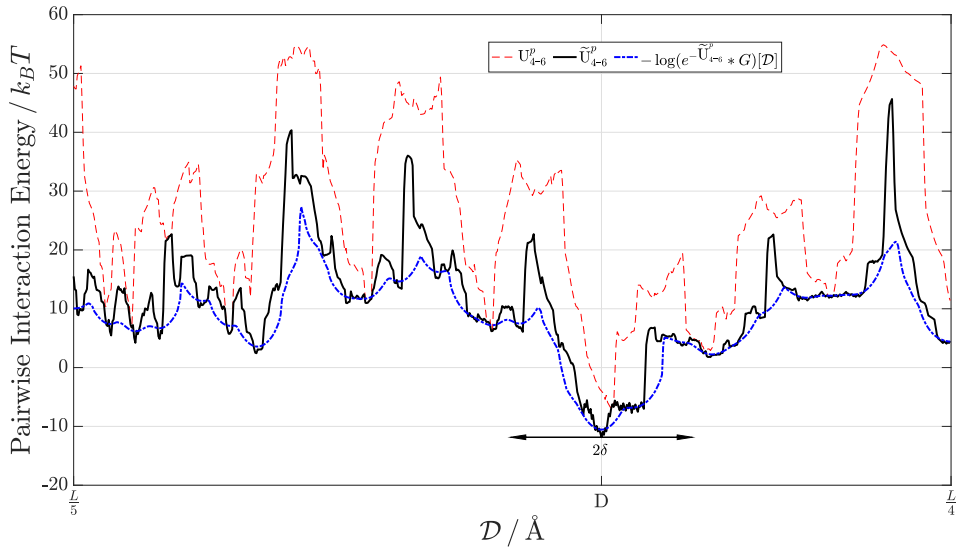


Figure 4.8: Effect of relaxation and smoothing on the constrained pairwise interaction energy $\mathrm{U}^{p}_{4\text{-}6}(k\mathcal{D}, n\mathcal{D} - L)$. The relaxed and smoothed interaction potentials are denoted by $\widetilde{\mathrm{U}}^{p}_{4\text{-}6}$ and $-\log(e^{-\widetilde{\mathrm{U}}^{p}_{4\text{-}6}} * G)[\mathcal{D}]$ respectively. The double-headed arrow illustrates the smoothing lengthscale $\delta = 15$ Å. The pairwise interaction energies are calculated for $n = 5$, $k = 4$ and $\psi(a_0) = 1.05\, k_B T$.

As a result, the energetic contributions arising from charged-hydrophobic and hydrophobic-hydrophobic residue interactions are characterised by discontinuous jumps in the values of the free energy. In the context of the constraint optimisation problem (4.18), this means that direct evaluation of the pairwise interaction energy can lead to overestimating the repulsive contributions to the free energy. To circumvent this issue, for each value of $\mathcal{D}$, we calculate the pairwise interaction energy by minimising over small perturbations δz to the axial stagger:

$$\widetilde{\mathrm{U}}^{p}_{i\text{-}j}(\mathcal{D}) = \min_{\delta z \in [-\Delta_0, \Delta_0]} \left\{ \mathrm{U}^{p}_{i\text{-}j}(k\mathcal{D} + \delta z, n\mathcal{D} - L) \right\}, \tag{4.31}$$

where $\widetilde{\mathrm{U}}^{p}_{i\text{-}j}(\mathcal{D})$ denotes the relaxed pairwise interaction energy. In all our calculations, we choose the perturbation size to be comparable to the step potential width l_c, namely we set $\Delta_0 = 5\,\text{Å}$. The relaxed interaction potential resulting from equation (4.31) can be seen in Figure 4.8.

It remains to identify the local minima in the potential $\widetilde{\mathrm{U}}^{p}_{i\text{-}j}(\mathcal{D})$. As can be seen upon close inspection in Figure 4.8, $\widetilde{\mathrm{U}}^{p}_{i\text{-}j}(\mathcal{D})$ can have several local minima of comparable energy in close proximity. To simplify the analysis, we apply Gaussian-weighted smoothing transformation to the relaxed potential. We then utilise the smoothed potential in order to identify the interaction energy minima:

$$\begin{aligned} \underset{\mathcal{D}}{\text{minimize}} \quad & -\log\left(e^{-\widetilde{\mathrm{U}}^{p}_{i\text{-}j}} * G\right)[\mathcal{D}], \\ \text{subject to} \quad & \frac{L}{n} < \mathcal{D} < \min\left\{\frac{L}{n}, \frac{L}{n-k}\right\}, \end{aligned} \tag{4.32}$$

where the symbol $*$ denotes a convolution defined as

$$\left(e^{-\widetilde{\mathrm{U}}^{p}_{i\text{-}j}} * G\right)[\mathcal{D}] = \frac{1}{2\delta}\int_{-\delta}^{\delta} e^{-\widetilde{\mathrm{U}}^{p}_{i\text{-}j}(\mathcal{D}-s)} G(\mathcal{D}, \sigma)\, ds, \tag{4.33}$$

where $G(\mathcal{D}, \sigma)$ is a Gaussian function given by

$$G(\mathcal{D}, \sigma) = Ce^{-\left(\frac{\mathcal{D}}{\sqrt{2}\sigma}\right)^2}, \quad \text{where} \quad \int_{-\delta}^{\delta} G(\mathcal{D}, \sigma)\, d\mathcal{D} = 1. \tag{4.34}$$

In all calculations, we set $\delta = 15\,\text{Å}$ and $\sigma = 0.4\delta$. Denoting all values of $\mathcal{D}$ that satisfy the minimisation problem (4.32) for a given value of k by $\mathcal{D}^k_{\mathrm{s}}$, we take the solution to the original problem (4.18) to be

$$\Delta z^*_m = k\mathcal{D}^k_{\mathrm{s}}, \qquad g^* = n\mathcal{D}^k_{\mathrm{s}} - L. \tag{4.35}$$

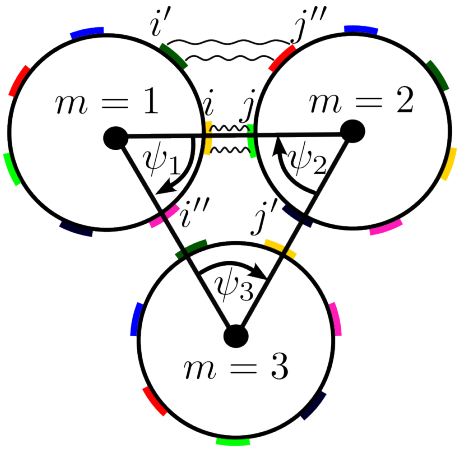


Figure 4.9: Schematic illustration of the steric screening in a triangular molecular packing model.

4.10.6 Magnitude of Steric Screening

In equation (4.17) we have identified three distinct contributions towards the pairwise interaction energy of two interacting collagen molecules. The three contributions arise from the interactions of the spatially closest strip pair i-j as well the interactions of the nearest neighbour strip pairs i'-j'' and i''-j'. Equation (4.17), however, does not hold for all molecular packing models and may require modification. To see this, consider the equilateral triangle packing model in Figure 4.9. It is clear that the interactions between strips i'' and j' will be screened, due to interactions of the first and the second molecules with the third one. The exact magnitude of the screening will depend on the 3-dimensional conformation and complex steric effects associated with the side chains of interacting amino acids. Here, we aim to provide a rough estimate for the impact of this steric screening on the pairwise interaction potentials $\mathrm{U}^{p}_{i\text{-}j}$.

We first note that the steric screening illustrated in Figure 4.9 will arise in all molecular packing models in which at least one internal polygon angle ψ_m is either $2\pi/7$ or $12\pi/7$. This is also true for odd-numbered microfibrils, except the internal angle also includes an azimuthal penalty of $\frac{\pi}{\mathrm{N}n_s}$. In these packing models, steric screening will affect the energy of all pairwise molecular interactions $\mathrm{U}^{p}_{i_m\text{-}j_m}$ that occur along the polygon edges that are adjacent to the vertex associated with the internal angle $\psi_m = 2\pi/7$ or $12\pi/7$. To estimate the impact of steric screening on the microfibrillar energy in equation (4.23), we start by calculating the contributions of the nearest neighbour interactions towards potentials $\mathrm{U}^{p}_{i_m\text{-}j_m}$ at the minima $(\Delta z^*_m, g^*)$, which satisfy equation (4.18). We will express

these contributions in the form

$$\mu_{\mathrm{U}}(\mathrm{E}_{i'\text{-}j''}) = \frac{\sum\limits_{i<j}\sum\limits_{(\Delta z_m^*, g^*)} \frac{\mathrm{E}_{i'\text{-}j''}(\Delta z_m^*, g^*)}{\mathrm{U}^p_{i\text{-}j}(\Delta z_m^*, g^*)} \left|\mathrm{U}^p_{i\text{-}j}(\Delta z_m^*, g^*)\right|}{\sum\limits_{i<j}\sum\limits_{(\Delta z_m^*, g^*)} \left|\mathrm{U}^p_{i\text{-}j}(\Delta z_m^*, g^*)\right|}, \tag{4.36}$$

where the summations are taken over all minima $(\Delta z_m^*, g^*)$ that are considered significant according to equation (4.21). An estimate for the contribution towards $\mathrm{U}^p_{i\text{-}j}$ of the other nearest neighbour interaction $\mathrm{E}_{i''\text{-}j'}$ is defined analogously. We put an upper bound $\mathrm{U}^p_{i\text{-}j}(\Delta z_m^*, g^*) < -1\, k_B T$ on the minima that are used in the calculation of the weighted average in equation (4.36), since the fractional energy contributions of the nearest neighbour interactions can become heavily skewed for weak molecular interactions.

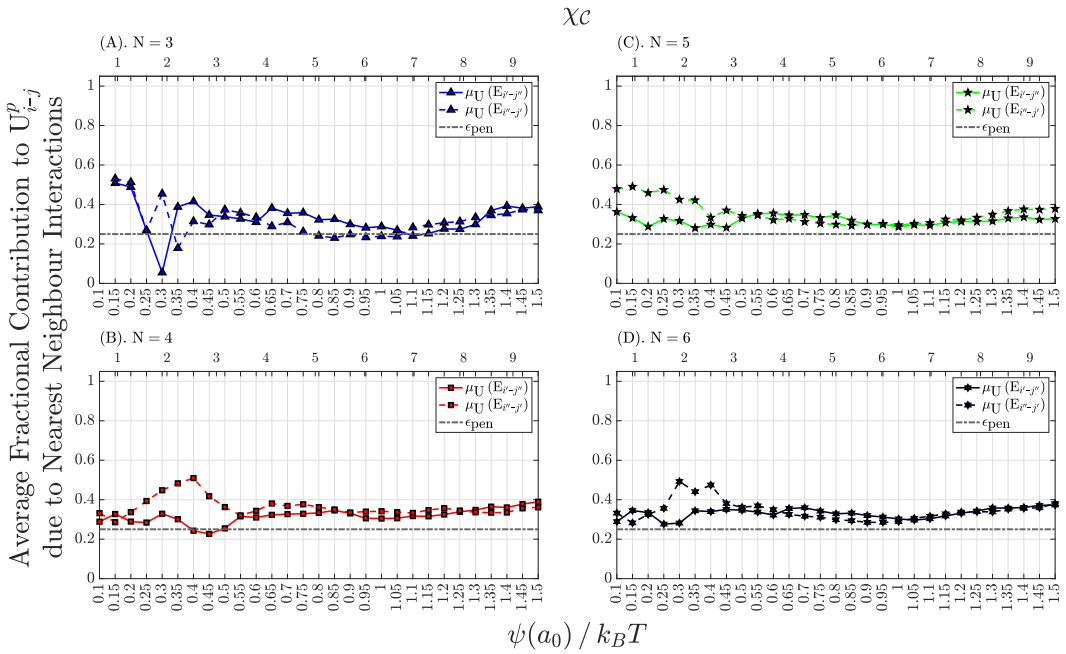


Figure 4.10: (A)-(D) Mean fractional contributions of nearest neighbour interactions towards net pairwise molecular interaction energy $\mathrm{U}^p_{i\text{-}j}$ as a function of charged-charged residue interaction strength. The four panels illustrate the data for aggregate sizes $\mathrm{N} = 3, \ldots, 6$. The horizontal line $\epsilon_{\mathrm{pen}} = 0.25$ represents the estimate for the fractional energy penalty due to steric screening.

The values of the weighted average μ_{U} for different aggregate sizes N and strengths of charged-charged residue interactions $\psi(a_0)$ are shown in Figures 4.10A-D. We first note that on average, the nearest neighbour interactions stabilise the net pairwise molecular interactions. We see that the majority of the average fractional contributions fall between 30-40% of the total pairwise molecular interaction. Based on this, we conservatively estimate that steric screening imposes a 25% energetic penalty on each affected pairwise

molecular potential $\mathrm{U}^{p}_{i_m\text{-}j_m}$ in a microfibril. The microfibrillar energy is then calculated using

$$\mathrm{E_M} = \sum_{m=1}^{\mathrm{N}} \mathrm{Q}_m \mathrm{U}^{p}_{i_m\text{-}j_m}. \tag{4.37}$$

We define the energy penalty factor Q_m as

$$\mathrm{Q}_m = \begin{cases} 1+\epsilon_{\mathrm{pen}}, \text{ if } n_{m'} \text{ or } n_{m''} \text{ is 1 or 6 , and } \mathrm{U}^{p}_{i_m\text{-}j_m}(\Delta z^*_m, g^*) > 0\, k_B T, \\ 1-\epsilon_{\mathrm{pen}}, \text{ if } n_{m'} \text{ or } n_{m''} \text{ is 1 or 6 , and } \mathrm{U}^{p}_{i_m\text{-}j_m}(\Delta z^*_m, g^*) < 0\, k_B T, \\ 1, \text{ otherwise,} \end{cases} \tag{4.38}$$

where $\epsilon_{\mathrm{pen}} = 0.25$. We note that in equation (4.38), we have assumed that for molecular interactions with net repulsive energy, steric screening further destabilises the interaction. It is plausible that the opposite may occur, in the instance that one of the nearest neighbour contributions towards $\mathrm{U}^{p}_{i_m\text{-}j_m}$ is repulsive. This, however, will not bear any importance for our calculations, since we are primarily interested in predicting thermodynamically stable microfibrils and the magnitude of the aforementioned stabilisation is unlikely to change the signature of $\mathrm{E_M}$.

4.10.7 Selection of Significance Score Cut-off

Selection of the significance score cut-off parameter c_α in equation (4.21) is crucial for interpretation of the outputs arising from our model. We note that the significance score of each pairwise molecular interaction in equation (4.20) is calculated for a fixed value of k. For different values of k, a given significance score can in principle correspond to different strengths of pairwise molecular interactions. As such, when implementing the significance score cut-off condition (4.21), we must ensure that the minimum energy of the pairwise molecular interactions that are deemed insignificant, is higher, than that of significant molecular interactions. Our goal then is to find the largest value of c_α such that the aforementioned requirement is satisfied for all values of parameters N and $\psi(a_0)$. We specifically look for the largest value of c_α in order to keep the computations tractable. Starting from a significance score cut-off value of 2.15 suggested by Puszkarska *et al.* [95], we incrementally decrease the value of c_α until for all N and $\psi(a_0)$ we satisfy the requirement that

$$\min_{\substack{\text{all } (\Delta z^*_m, g^*) \\ \text{satisfying (4.21)}}} \left\{ \mathrm{U}^{p}_{i\text{-}j}(\Delta z^*_m, g^*) \right\} < \min_{\substack{\text{all } (\Delta z^*_m, g^*) \\ \text{not satisfying (4.21)}}} \left\{ \mathrm{U}^{p}_{i\text{-}j}(\Delta z^*_m, g^*) \right\}, \tag{4.39}$$

where the $\min\{\dots\}$ operation is calculated across all spiral strips and corresponding $(\Delta z^*_m, g^*)$. The largest value of the significance score cut-off parameter that satisfies equation (4.39) for all N and $\psi(a_0)$ is $c_\alpha = 1.73$, which we use in all calculations - see Figures 4.11A-D.

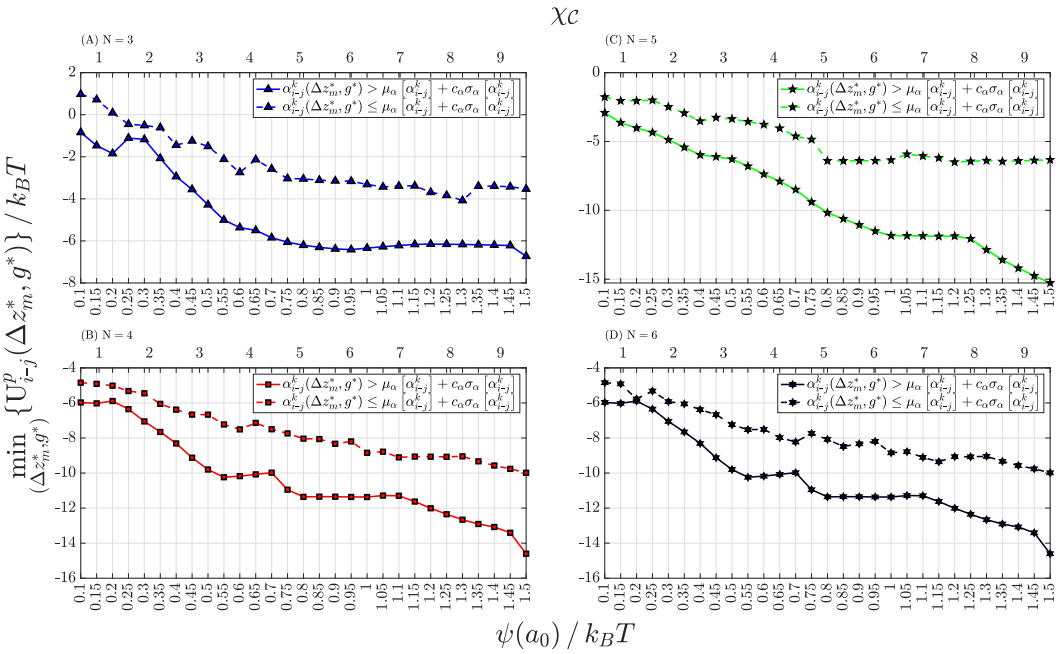


Figure 4.11: (A)-(D) Minimum energy of pairwise molecular interactions satisfying the optimisation problem (4.18) as a function of charged-charged residue interaction strength. The uninterrupted/dashed lines correspond to the pairwise interactions that do/do not satisfy the significance score cut-off respectively. The four panels illustrate the data for aggregate sizes $\mathrm{N} = 3, \ldots, 6$. In all four panels, we set $c_\alpha = 1.73$.

4.10.8 Additional Tables and Figures

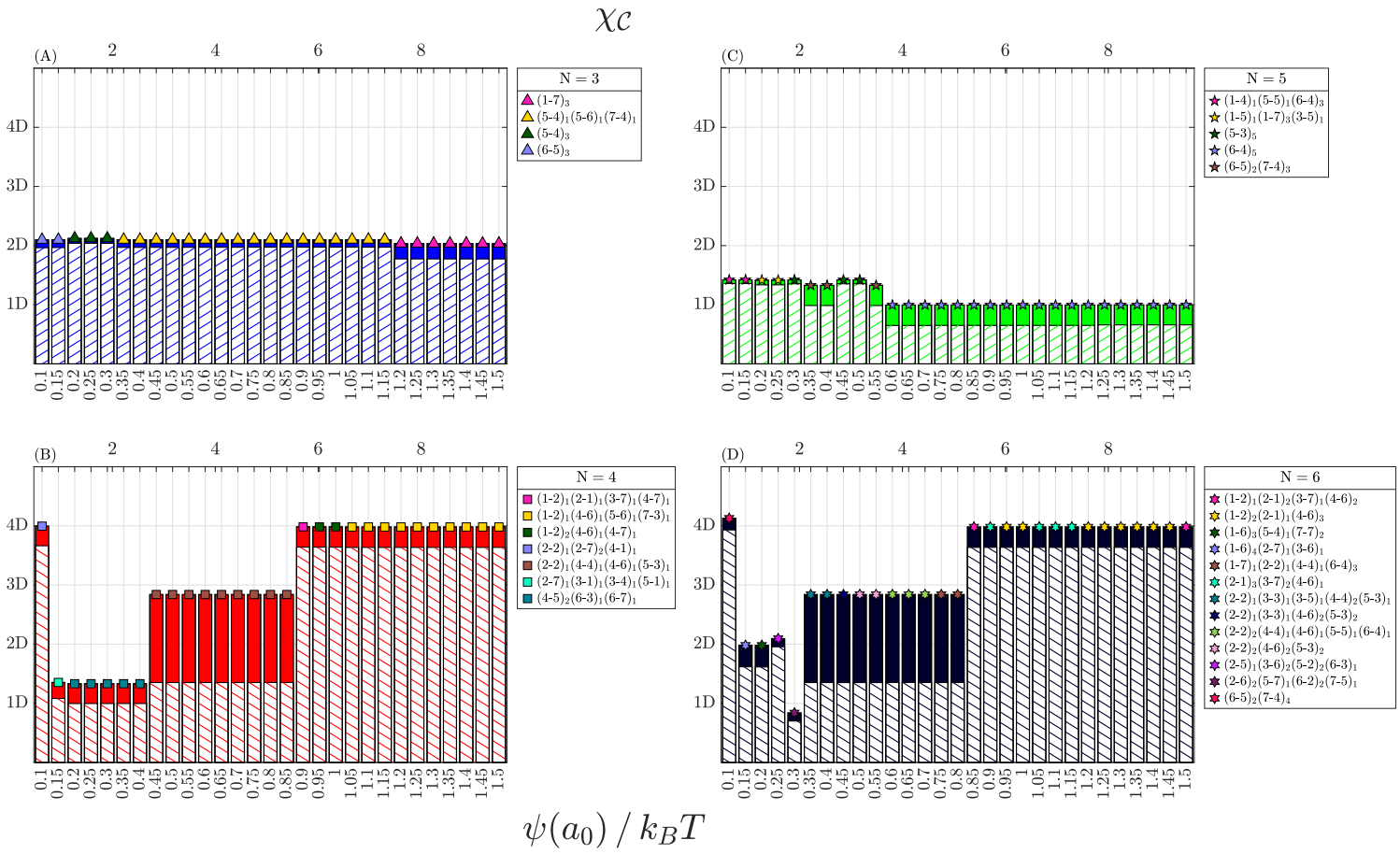


Figure 4.12: (A)-(D) Interacting strip pairs of the most energetically stable microfibrils as a function of charged-charged residue interaction strength $\psi(a_0)$ for microfibril sizes $N = 3, \ldots, 6$. A given microfibril is identified with interacting spiral strip pairs i_1-j_1, i_2-j_2, ..., i_N-j_N, which we write in compressed notation. For example, $(6\text{-}5)_2(7\text{-}4)_3$ denotes a 5-membered microfibril comprised of 2 and 3 interacting strip pairs 6-5 and 7-4 respectively.

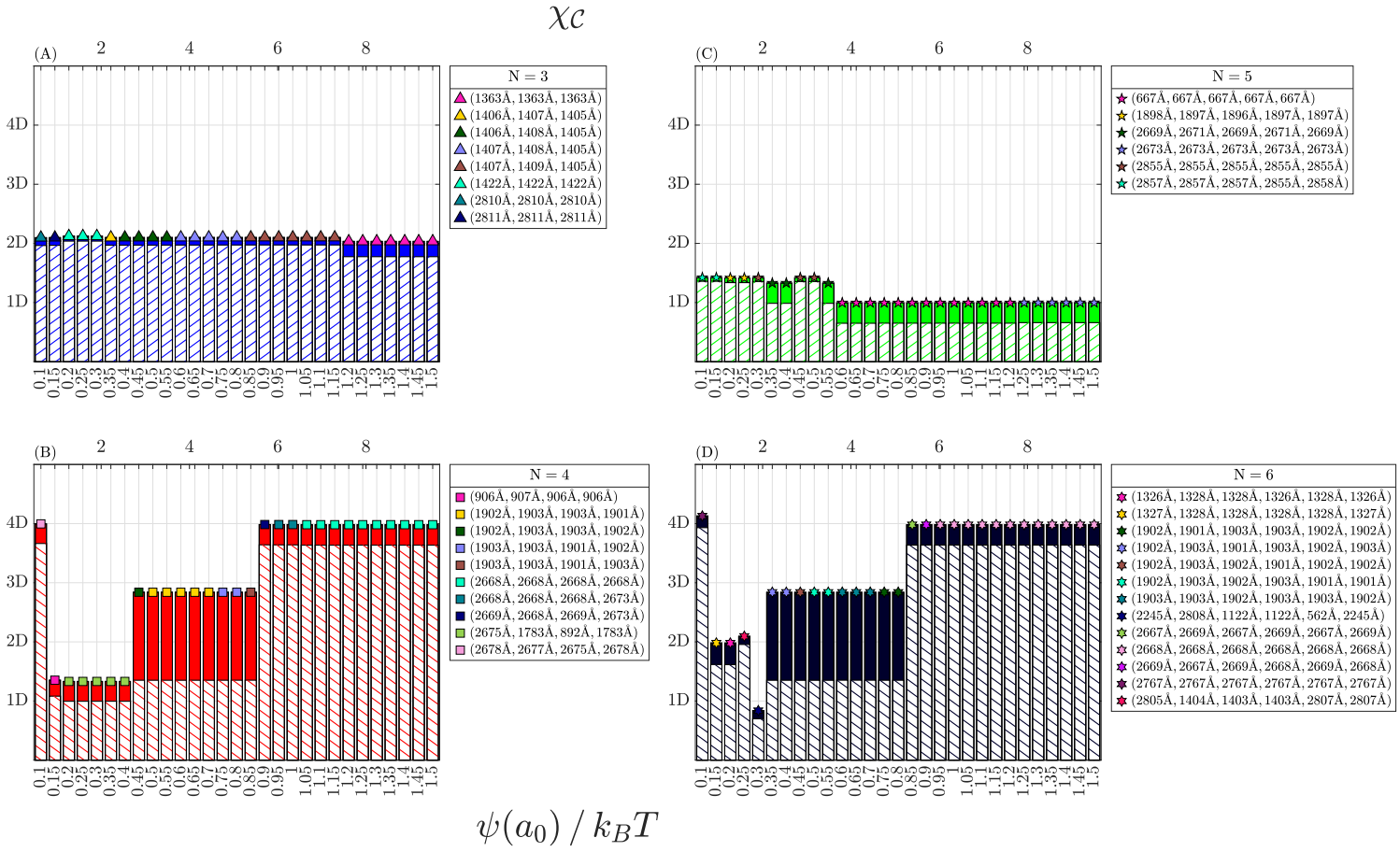


Figure 4.13: (A)-(D) Axial molecular staggers of the most energetically stable microfibrils as a function of charged-charged interaction strength $\psi(a_0)$ for microfibril sizes $\mathrm{N} = 3, \ldots, 6$. The axial staggers of the molecules in a microfibril are written in the form $(\Delta z_1, \Delta z_2, \ldots, \Delta z_{\mathrm{N}})$. The notation Δz_m denotes the relative stagger between the molecule $m = 1, \ldots, \mathrm{N}$ and its nearest neighbouring molecule m' in the clockwise direction about the long axis of the microfibril.

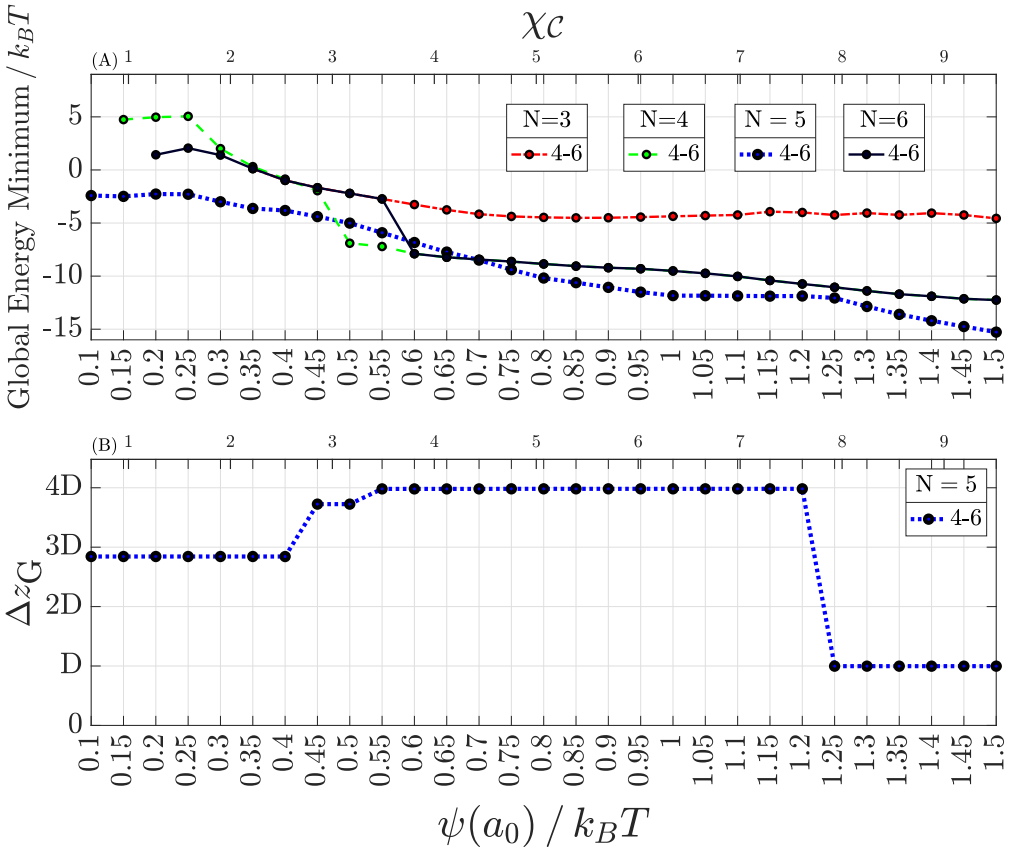


Figure 4.14: (A). Global energy minimum of the pairwise strip-strip potential $U^p_{4\text{-}6}$ subject to the constraints in equation (4.18) for $N = 3, \ldots, 6$ across different values of maximum charged-charged interaction strength $\psi(a_0)$. (B). Axial stagger at the global energy minimum in units of the D-banding lengthscale.

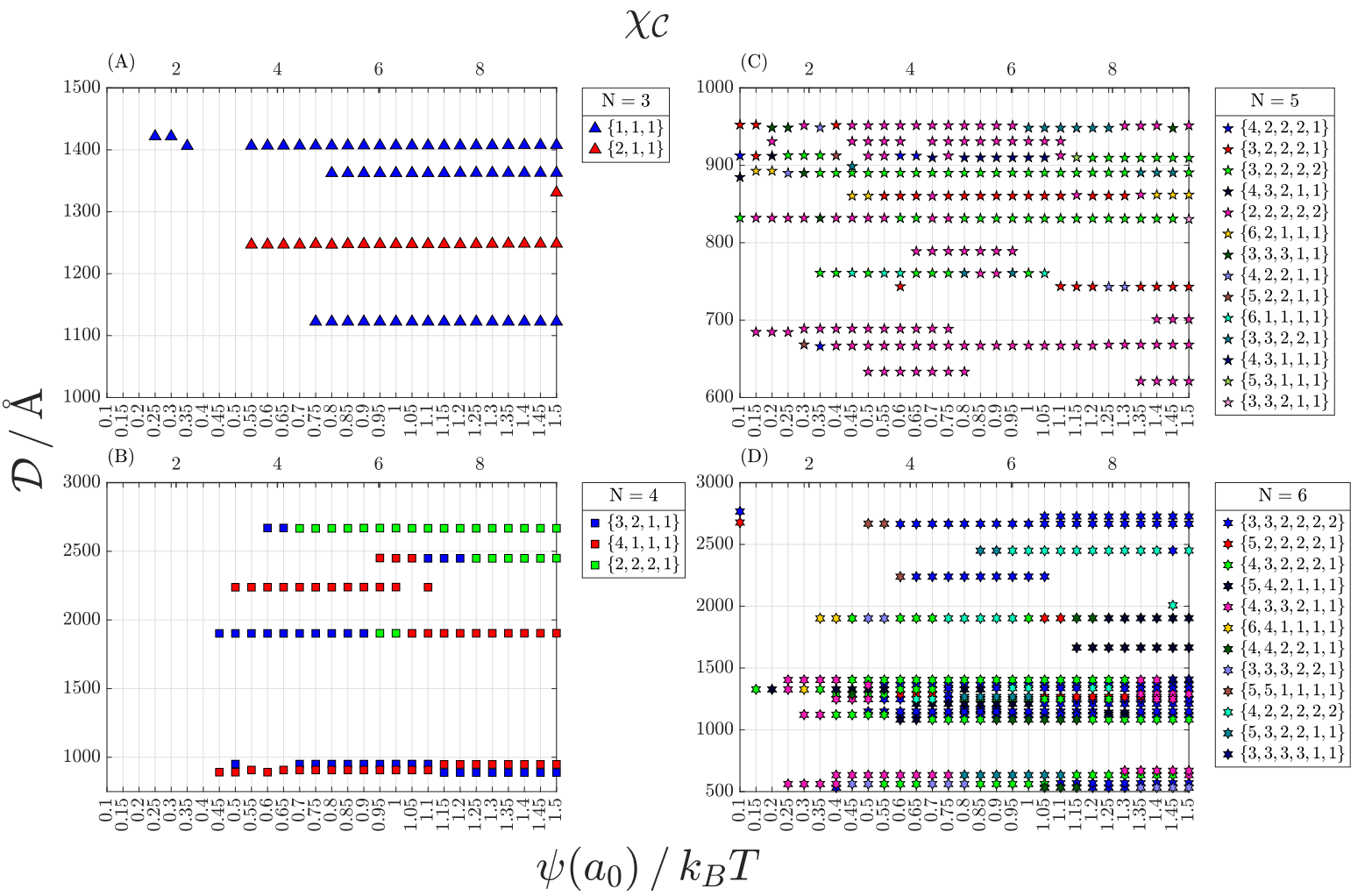


Figure 4.15: (A)-(D) Molecular packing models associated with distinct axial periods as a function of maximum charged-charged interaction strength $\psi(a_0)$ for $\mathrm{N} = 3, \ldots, 6$. Distinct axial periods correspond to ${}^{l}\mathcal{D}_{\mathrm{rep}}$ as defined in Appendix 4.10.4. The molecular packing models are identified by an unordered N-tuple $\{n_1, \ldots, n_{\mathrm{N}}\}$ which corresponds to an unordered set of internal polygon angles $\{2\pi n_1/7, \ldots, 2\pi n_{\mathrm{N}}/7\}$.

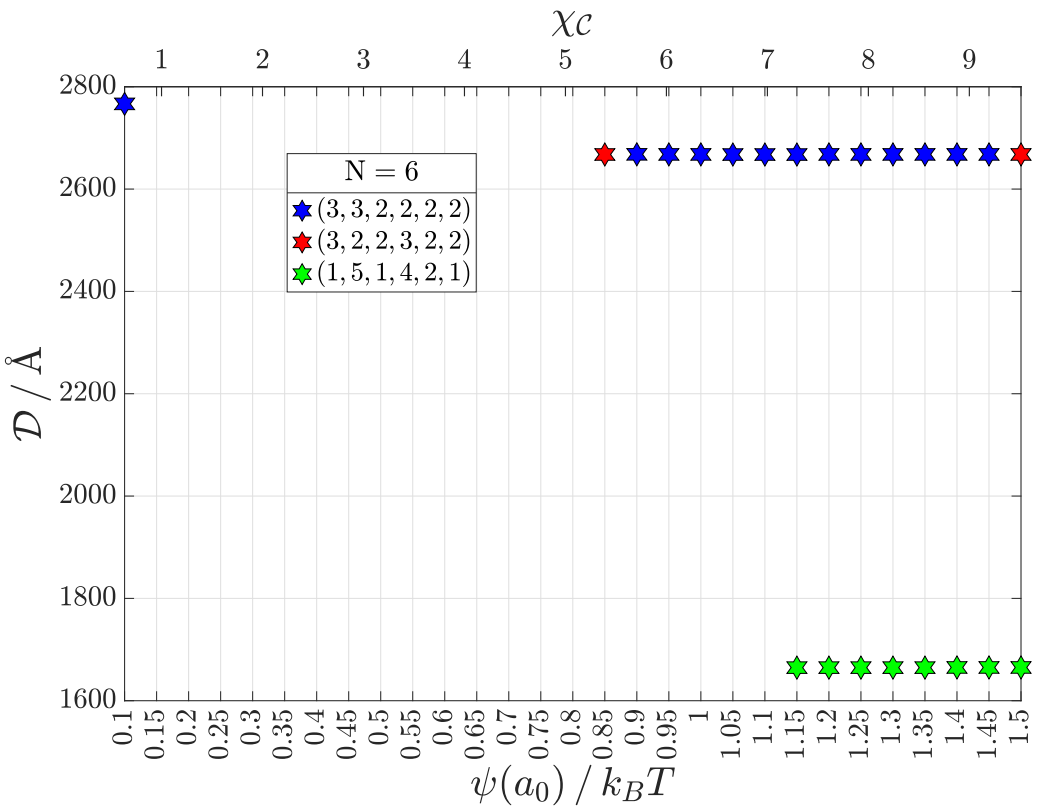
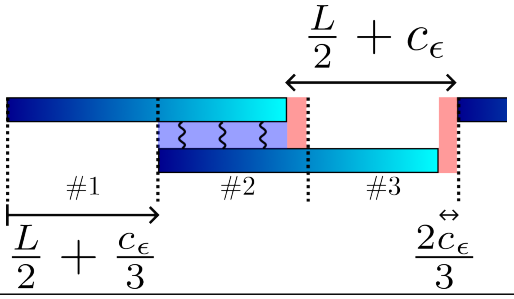
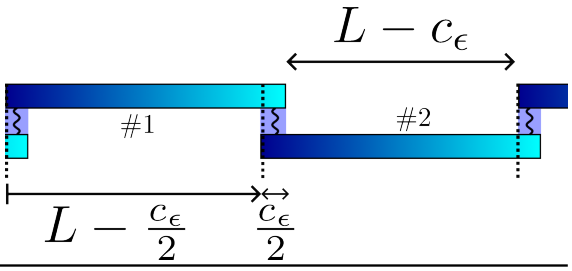


Figure 4.16: Detailed molecular packing models in the most energetically stable FLS I and FLS IV collagen microfibrils. The molecular packing is identified by an ordered N-tuple $(n_1, \ldots, n_{\mathrm{N}})$ which corresponds to an ordered set of internal polygon angles $(2\pi n_1/7, \ldots, 2\pi n_{\mathrm{N}}/7)$ illustrated in Figure 4.5. For FLS I microfibrils, the detailed molecular packing is shown for values of $\psi(a_0)$ such that they are the most stable 6-membered aggregates (see Figure 4.3D). For FLS IV microfibrils, the most stable molecular packing is determined among all other microfibrils with FLS IV $\mathcal{D}$-period.

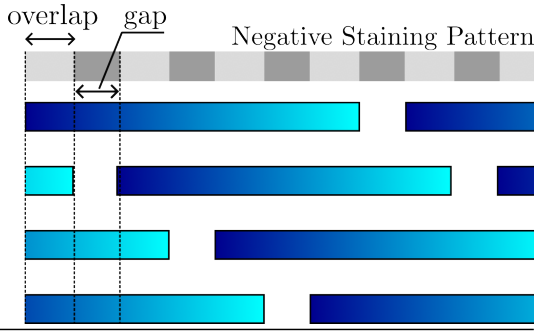
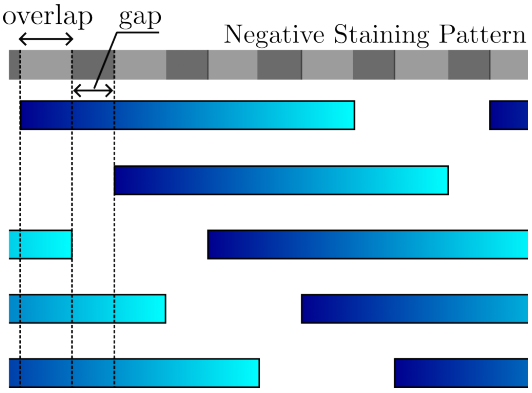
Table 4.2: Illustrative molecular packing schemes for different n-periodic microfibrils.

(a) Upper bounds on axial periodicity values.

Microfibril Descriptors	Pairwise Interaction Scheme
Microfibril size: any 3 | N **Periodicity lengthscale divisor:** $n = 3$ [a]**Gap size:** $g = \frac{L}{2} + c_\epsilon$	
Microfibril size: any 2 | N **Periodicity lengthscale divisor:** $n = 2$ **Gap size:** $g = L - c_\epsilon$	

[a] We define $c_\epsilon > 0$.

(b) Negative staining patterns of different contrast.

Microfibril Descriptors	Microfibril Packing Scheme
Microfibril size: N = 4 **Periodicity lengthscale divisor:** $n = 4$ **Gap size:** $g = \frac{L}{7}$ [a]**Parameter regime:** $0 < g < \frac{L}{3}$	
Microfibril size: N = 5 **Periodicity lengthscale divisor:** $n = 5$ **Gap size:** $g = \frac{2L}{5}$ **Parameter regime:** $\frac{L}{4} < g < \frac{2L}{3}$	

[a] Within the gap range, the contrast between the gap and overlap regions with negative staining is predicted to be the same.

Table 4.3: Number of distinct molecular packings resulting in microfibrils with distinct sets of pairwise molecular interactions.

№ Molecules / N	Packing Type	[a]Polygon Angles	[b]№ Packings / $\mathcal{N}(r_1, \ldots, r_l)$
3	$\{\psi_1^3\}$	$\{1,1,1\}$	1
	$\{\psi_1, \psi_2^2\}$	$\{2,1,1\}$	1
4	$\{\psi_1, \psi_2^3\}$	$\{4,1,1,1\}$ $\{1,2,2,2\}$	1
	$\{\psi_1, \psi_2, \psi_3^2\}$	$\{3,2,1,1\}$	3
5	$\{\psi_1, \psi_2^4\}$	$\{6,1,1,1,1\}$ $\{3,2,2,2,2\}$	1
	$\{\psi_1^5\}$	$\{2,2,2,2,2\}$	1
	$\{\psi_1^2, \psi_2^3\}$	$\{4,4,1,1,1\}$ $\{1,1,3,3,3\}$	2
	$\{\psi_1, \psi_2, \psi_3^3\}$	$\{5,2,1,1,1\}$ $\{4,3,1,1,1\}$ $\{3,1,2,2,2\}$ $\{6,2,1,1,1\}$ $\{5,3,1,1,1\}$ $\{4,1,2,2,2\}$	4
	$\{\psi_1, \psi_2^2, \psi_3^2\}$	$\{4,2,2,1,1\}$ $\{2,3,3,1,1\}$ $\{5,2,2,1,1\}$ $\{1,3,3,2,2\}$	6
	$\{\psi_1, \psi_2, \psi_3, \psi_4^2\}$	$\{4,3,2,1,1\}$	12
6	$\{\psi_1, \psi_2^5\}$	$\{4,2,2,2,2,2\}$	1
	$\{\psi_1^2, \psi_2^4\}$	$\{5,5,1,1,1,1\}$ $\{1,1,3,3,3,3\}$ $\{3,3,2,2,2,2\}$	3
	$\{\psi_1, \psi_2, \psi_3^4\}$	$\{5,1,2,2,2,2\}$ $\{6,4,1,1,1,1\}$	5
	$\{\psi_1, \psi_2^2, \psi_3^3\}$	$\{5,3,3,1,1,1\}$ $\{3,4,4,1,1,1\}$	10

	$\{6,1,1,2,2,2\}$ $\{1,2,2,3,3,3\}$	
$\{\psi_1^2,\psi_2^2,\psi_3^2\}$	$\{4,4,2,2,1,1\}$	16
$\{\psi_1,\psi_2,\psi_3,\psi_4^3\}$	$\{4,3,1,2,2,2\}$ $\{6,3,2,1,1,1\}$ $\{5,4,2,1,1,1\}$	20
$\{\psi_1,\psi_2,\psi_3^2,\psi_4^2\}$	$\{4,2,3,3,1,1\}$ $\{5,3,2,2,1,1\}$	30

[a] The internal angles of the molecular packing are written in the form of an unordered N-tuple $\{n_1,n_2,\ldots,n_{\mathrm{N}}\}$, where n_m are the positive integer multiples satisfying equations (4.15) or (4.16). For each packing type, we illustrate only a single molecular packing model.

[b] For each ordered N-tuple of polygon angles, this refers to the number of ways to arrange the angles ψ_m of the polygon, taking rotations to be equivalent.

Table 4.4: Classification of periodic signatures illustrated in Figure 4.4.

№ Molecules / N	[a]Periodic Signature / Å	[b]Periodicity Classification
3	983 1122 1248 1362 1407	Compact FLS
4	747 889 907 940	Dispersed FLS
	1903	[c]Not observed
	2239 2449 2668	FLS I
5	621 633 667 688	D-periodic
	701 743 761 789 831 861 891 911 931 950	Dispersed FLS
6	489 533 564	Short D-banding
	634 669	D-periodic

1082	Compact FLS
1122	
1147	
1167	
1210	
1248	
1266	
1287	
1329	
1338	
1362	
1404	
1665	FLS IV
1903	[c]Not observed
2239	FLS I
2449	
2668	
2726	

a Periodic signatures in Figure 4.4 are shown with black circles connected to rectangular boxes. The coordinates of the black circles along the $\mathcal{D}$-axis are shown in this Table. The detailed construction of the periodic signatures is discussed in Appendix 4.10.4.

b Experimentally measured ranges of admissible axial periods for each distinct periodicity class are listed in Table 4.1.

c To the best of the authors' knowledge.

— 5 —

Conclusions

5.1 OVERVIEW OF KEY FINDINGS

At the start of this thesis in chapter 1, we posed two broad questions pertaining to the polymorphic self-assembly of collagen: (1). What physical interactions lead to the emergence of structural polymorphism in collagen aggregates? (2). What physical mechanisms drive the competition between polymorphic aggregates of collagen? We draw attention to the phrase "structural polymorphism" which we take to be the three key structural characteristics of collagen microfibrils that, throughout this thesis, we saw exhibit polymorphism, namely: chirality, axial periodicity and 3-dimensional molecular packing. We now summarise our contributions towards answering the questions posed in chapter 1 with emphasis on their novelty in context of existing knowledge in the literature.

5.1.1 What Physical Interactions Lead to the Emergence of Structural Polymorphism in Collagen Aggregates?

Chirality

A significant portion of chapter 2 was devoted to understanding the physical origin of molecular supercoiling that is observed at the level of collagen microfibrils [82, 98]. Our contribution was demonstrating that the interactions between collagen triple helices are themselves chiral, in that they correspond to interactions between residues that organise into right-handed helical strips on the surfaces of the collagen molecules. We then provided an elasticity-driven mechanism, that showed that the chirality of collagen-collagen interactions is transmitted to the molecular level, thus explaining the experimentally observed right-handed molecular supercoiling of the triple helix in microfibrils. Previous models of collagen self-assembly did not identify this mechanism due to the choice of the molecular coarse-graining models for the triple helix. Linear models of the collagen molecule (see Figure 2.3), which are ubiquitous in literature (see [53, 92, 95, 45, 42] just for a few

examples), completely ignore the 3-dimensional residue organisation and are thus unable to explain the emergence of molecular supercoiling. Other coarse-grained approaches, such as those that apply the liquid crystal formalism to fibril self-assembly, necessitate postulating the existence of a director field, which subsequently models the molecular supercoil of the collagen molecules [20, 19]. The molecular-level insight into the origin of supercoiling at the microfibrillar level is again, absent in such models.

As it usually happens with structural features of collagen aggregates, molecular supercoiling displays at least two distinct types of polymorphism [98]. The first, is observed in tissues like bone and tendon, wherein the molecular supercoil angle is $\approx 5^\circ$. In tissues like cornea, on the other hand, the molecular supercoiling angle is significantly larger at $\approx 15^\circ$. We have shown that both of these types of molecular supercoiling can be accounted for using our theoretical approach by invoking different statistical parametrisations of the collagen triple helix that have been identified from the analysis of model peptide structural data [96]. The $\approx 5^\circ$ and $\approx 15^\circ$ molecular supercoiling angles can then be rationalised as arising from interactions of helically organised residues with the Pro-rich and the Pro-poor statistical parametrisations respectively.

We also defined microfibril chirality, which we believe to be a novel type of emergent chirality in collagen aggregates. Microfibril chirality corresponds to the handedness of the imaginary line that connects together the nearest gap regions in a microfibril, as can be seen in Figure 2.5E. We have shown that non-degeneracy of left-handed and right-handed microfibrils, according to the definition above, can only be achieved when the pairwise interaction potentials P_m between collagen triple helices m and m' break the symmetry

$$P_m(\theta_m, \theta_{m'}, \Delta z_m) = P_m(\theta_m, \theta_{m'}, \mathcal{T} - \Delta z_m), \tag{5.1}$$

where $\mathcal{T} = L + g$ is the sum of molecular and gap lengths, $\theta_{m'}$ and θ_m are the azimuthal molecular orientations and Δz_m is the axial stagger between the molecules. The illustration in Figure 2.12 showed that this symmetry arises from interactions between identical residue sequences and is thus inherent to models that coarse-grain the collagen triple helix as a linear molecule. In our model, the reside sequences of the interacting helical strips need not be the same, thus breaking the symmetry (5.1) and lifting the energy degeneracy between the left-handed and right-handed microfibril chiralities. We suggested that microfibril chirality plays an important role in bone mineralisation. This is due to the mineral phase in bone existing as left-handed helical crystals, which are known to preferentially nucleate in the gap regions of collagen fibrils [101]. Non-degeneracy of microfibrils with different microfibril chiralities then constitutes the first step towards understanding the enantioselectivity of the mineral phase in bone.

Axial Periodicity

Chapter 4 was dedicated to studying the interactions that give rise to the broad range of axial periods experimentally observed in fibrous long-spacing (FLS) collagens. We demonstrated a rather surprising and fundamental fact - axial periods of FLS collagens are encoded by interactions between parallel triple helices. This result is in direct opposition to previous studies of parallel triple helix interactions, which found no evidence for parallel association of collagen in FLS-staggered conformations [34, 120]. We remark that the aforementioned results were obtained using the sequence of a single α_1-chain, rather than the residue sequence of the entire triple helix. As such, the discrepancy with our results is not surprising, as we account for both the full residue sequence as well as the molecular structure of the triple helix in our calculations. We identified charged-charged residue interactions as being particularly important for stabilisation of FLS I, compact and dispersed FLS collagen microfibrils, but not FLS IV, for which hydrophobic-charged and/or hydrophobic-hydrophobic interactions were suggested to be more important by our calculations.

3-Dimensional Molecular Packing Models

Another important contribution of chapter 4 was our interaction-based prediction of the 3-dimensional packing of collagen molecules in microfibrils of FLS I and FLS IV collagens. To our knowledge, only 2-dimensional models, akin to those of Hodge-Petruska have been previously proposed [22, 34], making our result the first of its kind. We have first shown that axial periods of FLS I and FLS IV collagens may only be obtained in microfibrils with an even number of molecules by deriving a general relationship between microfibril size N and axial period $\mathcal{D}$ in equation (4.10). We then demonstrated that pairwise molecular interactions between triple helices leading to FLS I and FLS IV collagen axial periods, were the most energetically stable in one of the three 6-membered microfibrils with molecular packings shown in Figure 4.5. This illustrates another type of theoretical prediction that cannot be replicated with linear models of the collagen triple helix, as they exclude any azimuthal dependence in the pairwise interaction potential between collagen molecules.

5.1.2 What Physical Mechanisms Drive the Competition Between Polymorphic Aggregates of Collagen?

Chirality

In chapter 4, we showed that altering the strength of charged-charged residue interactions in D-banded microfibrils can act as a physical mechanism to change the energetically preferable

microfibril chirality from left-handed to right-handed and vice versa. In $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen, we demonstrated that whenever charged-charged residue interaction energies exceeded $1.20\,k_BT$, collagen microfibrils preferentially formed right-handed rather than left-handed assemblies. In terms of practical outcomes, this finding suggests a pathway for theoretical prediction of how experimentally tunable parameters such as pH and ionic strength may affect the microfibril chirality and subsequently the mineralisation of collagen fibrils.

The conspicuous absence of discussion pertaining to the molecular-level mechanisms that underlie the selection of different polymorphisms associated with molecular supercoiling points towards the current lack of understanding of this topic. We will return to this matter in section 5.2.

Axial Periodicity

Mechanisms that lead to the energetic selection of D-banded microfibrils as opposed to microfibrils which either lack or have different axial periods were the topic that reappeared throughout all substantive chapters of this thesis. In chapter 2, we first showed that the D-staggered conformation was the least sensitive to introducing perturbations into the values of residue-residue interaction energies. This suggested the possibility of biochemical control over the energies of other types of polymorphic microfibrils, which we studied in detail in chapter 3.

In chapter 3, we demonstrated that charged residue interactions generally favour the aggregation of D-banded microfibrils over the microfibrils in which all molecules are in-register. The latter microfibrils correspond to the experimentally observed SLS aggregates [120]. Previous theoretical works (including our own [124]) that studied collagen self-assembly using coarse-grained calculations of pairwise triple helix interactions found that in-register staggers were energetically more preferable than D-staggers [53, 92, 120]. Our approach to calculating pairwise collagen interactions presented in chapter 3, which utilised both Miyazawa-Jernigan contact potentials as well as a screened electrostatic potential (dependent on pH, ionic strength and residue pKa) for charged residue interactions, resolved this issue.

Another important result of chapter 3, centred around the phase transitions between axially disordered and D-banded collagen aggregates, which exist around mildly acidic and neutral pH respectively [51]. Previous studies of collagen aggregation as a function of ionic conditions did not account for the possibility of several different aggregated phases that differ in the axial molecular order, thus making our analysis novel [116, 115, 77]. We demonstrated that the charged interactions involving ionisable residues Glu as well as those incorporating Arg, selectively stabilised the D-banded phase, as opposed to the disordered

phase. Furthermore, we identified the order of residue ionisation, controlled by residue pKa values, to be a crucial parameter in determining the existence of the disordered phase. In particular, we demonstrated that a disordered phase exists at acidic pH only if Asp starts to ionise before Glu, i.e. in the regime pKa(Asp) < pKa(Glu). We are not aware of previous studies that identified the role of residue pKa in controlling the competition between different polymorphic aggregates of collagen, thus leading us to emphasise the novelty of the aforementioned insight.

3-Dimensional Molecular Packing Models

On several occasions in this thesis, we tackled the question of the energetic selection of the pentameric microfibril, as opposed to other molecular packing models with different numbers of triple helices and/or different spatial arrangement of collagen molecules. In chapter 2, we demonstrated that among other regular polygon packings, the pentagonal microfibril was energetically selected by the virtue of two distinct mechanisms. The first mechanism was geometric in nature. Positioning the triple helices at the vertices of a pentagon allowed for maximum contract area between the interacting residues on the collagen molecular surfaces. The second mechanism involved selective stabilisation of the pentameric microfibril by specific residue interactions.

In chapter 4, we delved deeper into the specific mechanisms that enabled selective energetic stabilisation of the pentameric microfibril. To do so, we constructed a new family of molecular packing models, in which every pair of nearest neighbouring triple helices was able to attain maximum contact area between their interacting residues, thereby removing the aforestated geometric mechanism from our calculations. For this new family of molecular packing models, we demonstrated that the residue interactions stabilising the pentameric D-banded microfibrils are unique and may not be attained in other axially periodic microfibrils with FLS or any other axial periodicities. The combined uniqueness of the interacting helical strips as well as the associated axial staggers and gap size, provided a mechanism for the selective energetic stabilisation of a pentameric D-banded microfibril. Using the labelling convention for spiral strips on the collagen molecular surface introduced in Figure 2.5B, we demonstrated that the aforesaid unique residue interactions corresponded to the strip pair 6-4 in $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ rat collagen. Yet again, the aforestated predictions cannot be made with the commonly used linear models of the triple helix, as establishing the coupling between specific residue interactions and molecular packing geometry necessitates inclusion of the 3-dimensional structure of the collagen triple helix.

5.2 FUTURE WORK

We now proceed to outline several suggestions for future developments of the theoretical models introduced in this thesis. A more comprehensive discussion of further work has already been recounted in each substantive chapter of this thesis. Here, the author merely highlights the areas of particular personal interest.

CHIRALITY

The study of the molecular-level mechanisms that lead to chirality polymorphism in collagen assemblies presents an exciting theoretical challenge. We have shown in chapter 2 that the two experimentally measured distinct molecular supercoiling angles of $\approx 5^\circ$ and $\approx 15^\circ$ can be explained as arising from chiral interactions between Pro-rich and Pro-poor residues strips respectively. This, however, leaves several important questions to be answered. (1). What physical factors control the selection of different chiral symmetries associated with the 3-dimensional residue organisation of the triple helix? As suggested by their names, Pro-poor and Pro-rich statistical parametrisations of the triple helix are associated with regions of the triple helix that contain varying levels of Pro [83]. Both bone and cornea, however, are primarily composed of $\alpha_2(\mathrm{I})[\alpha_1(\mathrm{I})]_2$ collagen, yet the molecular supercoiling angles in these biological tissues are best explained by the interactions of the Pro-rich and

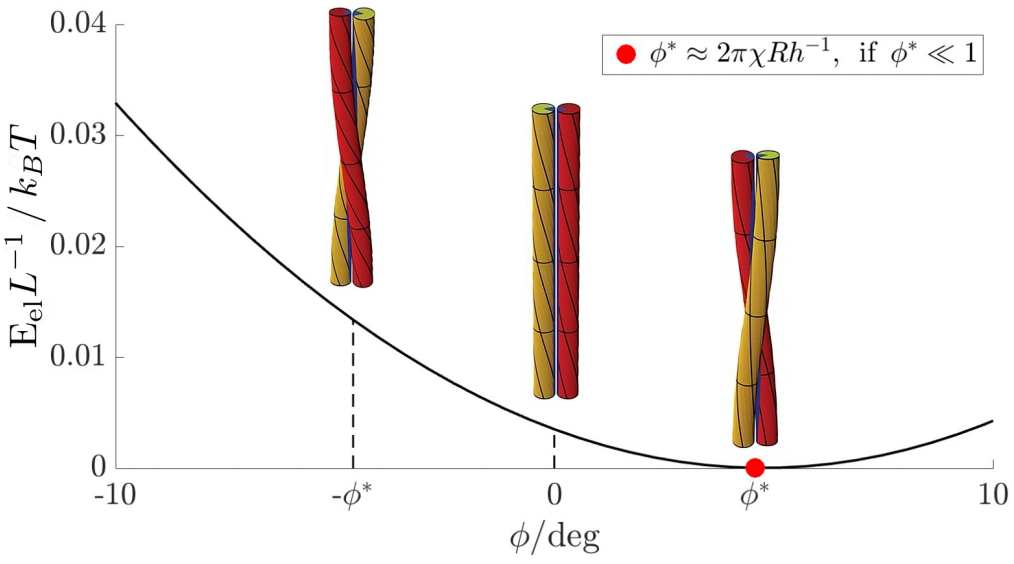


Figure 5.1: Cost of elastic deformation for different values of molecular supercoiling angle. The blue lines on each cylinder are representative of helical residue strips on the collagen molecular surface. The cost of elastic deformation is calculated using equation (2.4) for Pro-rich helical strips. Images of the supercoiled cylinders are adapted from [79].

Pro-poor helical residue strips respectively. Other factors, such as the presence of fibrillar collagens of types III and V in cornea may be an important factor [98], yet the definitive answer is not clear. (2). What molecular-level mechanisms control the propagation of molecular supercoiling to larger lengthscales during fibrillogenesis? Coarse-grained studies of collagen fibril growth have argued that the molecular supercoiling angle dependence of fibril radius is mildly non-linear with zero supercoiling angle at the centre of the collagen fibril [20, 19]. Zero angle molecular supercoil can be accommodated within our triple helix interaction model, by invoking a pure twist deformation of the interacting collagen molecules, albeit at some additional energetic cost as compared to a bend-dominated deformation that leads to a right-handed helical supercoil - see Figure 5.1. It is then of interest to bridge the predictions arising from coarse-grained fibril growth models with our proposed molecular-level mechanism for the emergence of molecular supercoiling angle.

Another avenue for further research involves ascertaining the role of microfibril chirality in mineralisation of bone. The left-handed helical crystals found in mineralised bone have a characteristic lateral dimension of 5-10 nm [101], corresponding to the lateral dimension of several collagen microfibrils. The next step would be to determine the chiral spatial distribution of gap regions in aggregates of microfibrils, which have been suggested to act as nucleation sites for the mineral phase [101]. The spatial organisation of gap regions can then be used to predict the architecture of the mineral phase in bone. Our triple helix interaction model developed in chapter 3 can supplement the aforementioned study through investigating the effect of experimentally adjustable parameters, such as pH and ionic strength, on microfibril chirality and the emergent architecture of the mineral phase.

Axial Periodicity

The results of chapter 4 invite a number of further questions pertaining to axially periodic microfibrils. One such topic of interest is the origin of dihedral symmetry in the fine structure of the banding pattern in FLS collagens. Several studies have suggested that the dihedral symmetry originates from the presence of equal fractions of collagen triple helices in parallel and anti-parallel arrangements [22, 34, 120]. It is however unclear, at which stage of hierarchical aggregation do anti-parallel triple helix interactions appear.

A more fundamental question, could entail determining the role of pairwise interactions at $\Delta z = 0$ in aggregation of axially periodic microfibrils. We recall equation (4.2) of chapter 4, which stated the constraints on the axial degrees of freedom for an axially periodic microfibril. We notice for $N_g > 1$, that molecules with axial staggers $z_m = 0$ may stabilise axially periodic microfibrils. The physical interpretation of this fact is illustrated in Figure 5.2 for a 6-membered microfibril with an axial period $\mathcal{D}$, such that $g = 2\mathcal{D} - L$ (for instance an FLS I microfibril). In this case, the microfibril can be thought of as a trimer comprised

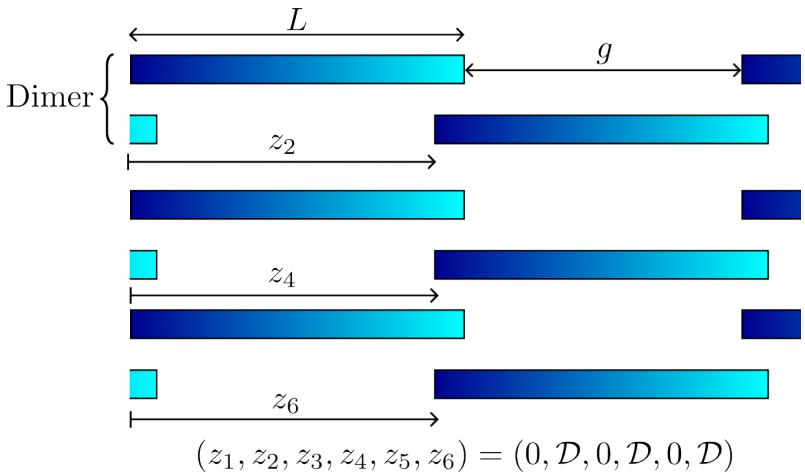


Figure 5.2: A 2-dimensional scheme of a 6-membered trimer of $\mathcal{D}$-staggered dimers. We use z_m to denote the axial stagger of the m^{th} molecular array. All dimers are arranged in-register with one another, i.e. $z_1 = z_3 = z_5 = 0$. One-sided arrows illustrate non-zero axial staggers in the aggregate.

of $\mathcal{D}$-staggered dimers that are arranged in-register, i.e. the axial staggers of the molecules are $z_{2m-1} = 0$ and $z_{2m} = \mathcal{D}$. Notice that the molecular arrangement shown in Figure 5.2 produces an identical $\mathcal{D}$-period to the one in which every molecule is $\mathcal{D}$-staggered, i.e. $z_m = \mathcal{D}$ for all m. The molecular interactions that stabilise the two aforementioned microfibrils are however clearly different. In addition to stabilising the microfibrillar structure itself, pairwise molecular interactions with $\Delta z = 0$ may be particularly important in FLS collagens for subsequent hierarchical aggregation of microfibrils. Evidence from AFM studies suggests that axially staggered FLS microfibrils aggregate in-register into higher-order structures [117].

3-Dimensional Molecular Packing Models

The 3-dimensional molecular packing models introduced in chapter 4 provide rich grounds for further theoretical study. Our predictions of the individual molecular coordinates within microfibrils as well as their energies can be further refined by accounting for two types of physical effects. Firstly, since certain molecular packing models may correspond to concave polygons (see Figure 4.5C), non-nearest neighbour interactions may contribute significantly to the microfibril energy. Secondly, we performed our calculations under the assumption that triple helices connected by polygon edges are able to get within the same distance of each other. However, as illustrated by Figures 4.5A, C, not every molecular packing model admits an equal separation between nearest neighbouring triple helices. Accounting for these physical effects would enable producing refined predictions of molecular organisation within FLS microfibrils for subsequent experimental validation.

Commonly Used Symbols

L length of the triple helix, 13

D axial periodicity in native collagen, 13

N № molecules comprising a microfibril, 14

Δz axial stagger, 16

k integer multiple of the axial periodicity lengthscale, 16

θ azimuthal angle, 19

z axial position, 19

$\mathrm{E}_{i\text{-}j}$ pairwise interaction energy between strips i and j, 20

g gap length, 21

$\mathrm{E}^{p}_{i\text{-}j}$ axially periodic pairwise interaction energy between strips i and j, 21

θ_m azimuthal angle of the m^{th} molecule, 22

z_m axial position of the m^{th} molecule, 22

E_{M} energy of the microfibril, 22

h helical pitch of the spiral strips, 31

T temperature, 32

Φ azimuthal component of the pairwise triple helix interaction, 34

S_{eq} the set of near-equilibrium states, 36

$\mathcal{C}$ set of all ionisable residues, 50

R_{p} residue with sequential position p along the α-chain, 51

I ionic strength, 51

z^{eff} effective residue charge, 51

$\alpha(\mathrm{R})$ disassociation fraction of residue R, 52

κ inverse Debye length, 52

a_0 distance of closest approach between two residues, 52

$\psi(a_0)$ electrostatic interaction energy at a distance a_0, 52

$\mathrm{U}_{i\text{-}j}$ pairwise interaction energy between strips i and j and their nearest neighbouring strips, 53

$\mathrm{U}^{p}_{i\text{-}j}$ axially periodic pairwise interaction energy between strips i and j and their nearest neighbouring strips, 53

i_m-j_m pair of interacting strips between molecule m and its nearest clockwise neighbour, 54

$\mathcal{D}$ generalised axial period, 55

$\mathcal{P}_{\text{period}}$ equilibrium probability of the axially ordered phase, 56

W_{R} ionisation window of residue R, 58

μ overlap measure between two residue ionisation windows, 59

z^{ave} average effective charge of co-ionising residues, 60

n number of partitions of the axial periodicity lengthscale $\mathcal{T}$, 88

ψ_m internal angle of a simple N-gon, 93

List of Acronyms

TEM	transmission electron microscopy 2
FLS	fibrous long-spacing 3
SLS	segment-long-spacing 3
ATP	adenosine triphosphate 3
MJCP	Miyazawa-Jernigan contact potentials 20
PS	perfectly-staggered 29
NEqSs	near-equilibrium states 36
AFM	atomic force microscopy 87

References

[1] A. A. Adzhubei, M. J. Sternberg, and A. A. Makarov. “Polyproline-II helix in proteins: structure and function”. In: *J. Mol. Biol.* 425.12 (2013), pp. 2100–2132.

[2] O. Antipova and J. P. Orgel. “In situ D-periodic molecular structure of type II collagen”. In: *J. Biol. Chem.* 285.10 (2010), pp. 7087–7096.

[3] P. M. Armitage and J. A. Chapman. “New fibrous long spacing form of collagen”. In: *Nature New Biol.* 229.5 (1971), pp. 151–152.

[4] M. Asgari, N. Latifi, H. K. Heris, H. Vali, and L. Mongeau. “In vitro fibrillogenesis of tropocollagen type III in collagen type I affects its relative fibrillar topology and mechanics”. In: *Sci. Rep.* 7.1 (2017), p. 1392.

[5] S. Bansode, U. Bashtanova, R. Li, J. Clark, K. H. Müller, A. Puszkarska, I. Goldberga, H. H. Chetwood, D. G. Reid, L. J. Colwell, et al. “Glycation changes molecular organization and charge distribution in type I collagen fibrils”. In: *Sci. Rep.* 10.1 (2020), p. 3397.

[6] D. R. Baselt, J.-P. Revel, and J. D. Baldeschwieler. “Subfibrillar structure of type I collagen observed by atomic force microscopy”. In: *Biophys. J.* 65.6 (1993), pp. 2644–2655.

[7] J. Bella. “A new method for describing the helical conformation of collagen: Dependence of the triple helical twist on amino acid sequence”. In: *J. Struct. Biol.* 170.2 (2010), pp. 377–391.

[8] J. Bella. “Collagen structure: new tricks from a very old dog”. In: *Biochem. J.* 473.8 (2016), pp. 1001–1025.

[9] J. Bella, B. Brodsky, and H. M. Berman. “Hydration structure of a collagen peptide”. In: *Structure* 3.9 (1995), pp. 893–906.

[10] J. Bella and D. J. Hulmes. “"Fibrillar collagens"”. In: *Fibrous proteins: structures and mechanisms*. Springer, 2017, pp. 457–490.

[11] D. E. Birk. “Type V collagen: heterotypic type I/V collagen interactions in the regulation of fibril assembly”. In: *Micron* 32.3 (2001), pp. 223–237.

[12] E. Boedtkjer and S. F. Pedersen. “The acidic tumor microenvironment as a driver of cancer”. In: *Annu. Rev. Physiol.* 82.1 (2020), pp. 103–126.

[13] R. P. Boot-Handford, D. S. Tuckwell, D. A. Plumb, C. F. Rock, and R. Poulsom. “A novel and highly conserved collagen (proα1 (XXVII)) with a unique expression pattern and unusual molecular characteristics establishes a new clade within the vertebrate fibrillar collagen family”. In: *J. Biol. Chem.* 278.33 (2003), pp. 31067–31077.

[14] S. P. Boudko. “Around the collagen triple helix: an introduction to studying associated genetic and acquired diseases”. In: *Matrix Biol.* (2025).

[15] J. Bowden, J. Chapman, and C. Wynn. “Precipitation of collagen in the segmented long-spacing form by various organic and inorganic compounds”. In: *Biochim. Biophys. Acta - Prot. Struct.* 154.1 (1968), pp. 190–195.

[16] B. Brodsky and E. F. Eikenberry. “Characterization of fibrous forms of collagen”. In: *Methods Enzymol.* Vol. 82. Elsevier, 1982, pp. 127–174.

[17] B. Brodsky, E. F. Eikenberry, and K. Cassidy. “An unusual collagen periodicity in skin”. In: *Biochim. Biophys. Acta - Prot. Struct.* 621.1 (1980), pp. 162–166.

[18] A. I. Brown, L. Kreplak, and A. D. Rutenberg. “An equilibrium double-twist model for the radial structure of collagen fibrils”. In: *Soft Matter* 10.42 (2014), pp. 8500–8511.

[19] S. Cameron, L. Kreplak, and A. D. Rutenberg. “Phase-field collagen fibrils: Coupling chirality and density modulations”. In: *Phys. Rev. Res.* 2.1 (2020), p. 012070.

[20] S. Cameron, L. Kreplak, and A. D. Rutenberg. “Polymorphism of stable collagen fibrils”. In: *Soft Matter* 14.23 (2018), pp. 4772–4783.

[21] H. Chanut-Delalande, A. Fichard, S. Bernocco, R. Garrone, D. J. Hulmes, and F. Ruggiero. “Control of heterotypic fibril formation by collagen V is determined by chain stoichiometry”. In: *J. Biol. Chem.* 276.26 (2001), pp. 24352–24359.

[22] J. A. Chapman and P. M. Armitage. “An analysis of fibrous long spacing forms of collagen”. In: *Connect. Tissue Res.* 1.1 (1972), pp. 31–37.

[23] J. A. Chapman, M. Tzaphlidou, K. M. Meek, and K. E. Kadler. “The collagen fibril—a model system for studying the staining and fixation of a protein”. In: *Electron Microsc. Rev.* 3.1 (1990), pp. 143–182.

[24] F. Chen, R. Strawn, and Y. Xu. "The predominant roles of the sequence periodicity in the self-assembly of collagen-mimetic mini-fibrils". In: *Protein Sci.* 28.9 (Sept. 2019), pp. 1640–1651.

[25] J. M. Chen, C. E. Kung, S. H. Feairheller, and E. M. Brown. "An energetic evaluation of a "Smith" collagen microfibril model". In: *J. Protein Chem.* 10 (1991), pp. 535–552.

[26] J. Chen, T. Ahn, I. D. Colón-Bernal, J. Kim, and M. M. Banaszak Holl. "The relationship of collagen structural and compositional heterogeneity to tissue mechanical properties: a chemical perspective". In: *ACS Nano* 11.11 (2017), pp. 10665–10671.

[27] A. Cooper. "Thermodynamic studies of the assembly in vitro of native collagen fibrils". In: *Biochem. J.* 118.3 (1970), pp. 355–365.

[28] T. P. Creamer and M. N. Campbell. "Determinants of the polyproline II helix from modeling studies". In: *Adv. Protein Chem.* 62 (2002), pp. 263–282.

[29] N. G. De Bruijn. *Asymptotic methods in analysis*. Courier Corporation, 2014.

[30] B. Depalle, Z. Qin, S. J. Shefelbine, and M. J. Buehler. "Influence of cross-link structure, density and mechanical properties in the mesoscale deformation mechanisms of collagen fibrils". In: *J. Mech. Behav. Biomed. Mater.* 52 (2015), pp. 1–13.

[31] X. Di, X. Gao, L. Peng, J. Ai, X. Jin, S. Qi, H. Li, K. Wang, and D. Luo. "Cellular mechanotransduction in health and diseases: from molecular mechanism to therapeutic targets". In: *Signal Transduct. Target. Ther.* 8.1 (2023), p. 282.

[32] G. A. Di Lullo, S. M. Sweeney, J. Korkko, L. Ala-Kokko, and J. D. San Antonio. "Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen". In: *J. Biol. Chem.* 277.6 (2002), pp. 4223–4231.

[33] K. P. Dingemans and P. Teeling. "Long-spacing collagen and proteoglycans in pathologic tissues". In: *Ultrastruct. Pathol.* 18.6 (1994), pp. 539–547.

[34] B. B. Doyle, D. W. Hukins, D. J. Hulmes, A. Miller, and J. Woodhead-Galloway. "Collagen polymorphism: its origins in the amino acid sequence". In: *J. Mol. Biol.* 91.1 (1975), pp. 79–99.

[35] A. Efrat, R. Fulek, S. Kobourov, and C. D. Tóth. "Polygons with Prescribed Angles in 2D and 3D". In: *International Symposium on Graph Drawing and Network Visualization*. Springer. 2020, pp. 135–147.

[36] M. Fang, E. L. Goldstein, A. S. Turner, C. M. Les, B. G. Orr, G. J. Fisher, K. B. Welch, E. D. Rothman, and M. M. Banaszak Holl. "Type I collagen D-spacing in fibril bundles of dermis, tendon, and bone: bridging between nano-and micro-level tissue hierarchy". In: *ACS nano* 6.11 (2012), pp. 9503–9514.

[37] G. Faure, A. Bornot, and A. G. de Brevern. "Protein contacts, inter-residue interactions and side-chain modelling". In: *Biochimie* 90.4 (2008), pp. 626–639.

[38] A. L. Fidler, C. E. Darris, S. V. Chetyrkin, V. K. Pedchenko, S. P. Boudko, K. L. Brown, W. Gray Jerome, J. K. Hudson, A. Rokas, and B. G. Hudson. "Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues". In: *eLife* 6 (2017), e24176.

[39] C. A. Fitch, G. Platzer, M. Okon, B. Garcia-Moreno E, and L. P. McIntosh. "Arginine: Its pKa value revisited". In: *Protein Sci.* 24.5 (2015), pp. 752–761.

[40] A. Forlino, W. A. Cabral, A. M. Barnes, and J. C. Marini. "New perspectives on osteogenesis imperfecta". In: *Nat. Rev. Endocrinol.* 7.9 (2011), pp. 540–557.

[41] R. Fraser and M. TP. *Conformation in fibrous proteins and related synthetic polypeptides*. New York: Academic Press, 1973.

[42] C. Garcia-Sacristan, V. G. Gisbert, K. Klein, A. Saric, and R. Garcia. "In Operando Imaging Electrostatic-Driven Disassembly and Reassembly of Collagen Nanostructures". In: *ACS nano* 18.28 (2024), pp. 18485–18492.

[43] A. Gautieri, S. Vesentini, A. Redaelli, and M. J. Buehler. "Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up". In: *Nano Lett.* 11.2 (2011), pp. 757–766.

[44] F. N. Ghadially. *Ultrastructural pathology of the cell and matrix: a text and atlas of physiological and pathological alterations in the fine structure of cellular and extracellular components*. Butterworth-Heinemann, 2013.

[45] G. Giubertoni, L. Feng, K. Klein, G. Giannetti, L. Rutten, Y. Choi, A. Van Der Net, G. Castro-Linares, F. Caporaletti, D. Micha, et al. "Elucidating the role of water in collagen self-assembly by isotopically modulating collagen hydration". In: *PNAS* 121.11 (2024), e2313162121.

[46] N. Goldenfeld. *Lectures on phase transitions and the renormalization group*. CRC Press, 2018.

[47] T. Gurry, P. S. Nerenberg, and C. M. Stultz. "The contribution of interchain salt bridges to triple-helical stability in collagen". In: *Biophys. J.* 98.11 (2010), pp. 2634–2643.

[48] W. Han, S. Chen, W. Yuan, Q. Fan, J. Tian, X. Wang, L. Chen, X. Zhang, W. Wei, R. Liu, et al. "Oriented collagen fibers direct tumor cell intravasation". In: *PNAS* 113.40 (2016), pp. 11208–11213.

[49] U. Hansen and P. Bruckner. "Macromolecular specificity of collagen fibrillogenesis: fibrils of collagens I and XI contain a heterotypic alloyed core and a collagen I sheath". In: *J. Biol. Chem.* 278.39 (2003), pp. 37352–37359.

[50] J. R. Harris and R. J. Lewis. "The collagen type I segment long spacing (SLS) and fibrillar forms: Formation by ATP and sulphonated diazo dyes". In: *Micron* 86 (2016), pp. 36–47.

[51] J. R. Harris and A. Reiber. "Influence of saline and pH on collagen type I fibrillogenesis in vitro: fibril polymorphism and colloidal gold labelling". In: *Micron* 38.5 (2007), pp. 513–521.

[52] H. Hofmann, P. 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". In: *J. Mol. Biol.* 125.2 (1978), pp. 137–165.

[53] D. J. Hulmes, A. Miller, D. A. Parry, K. A. Piez, and J. Woodhead-Galloway. "Analysis of the primary structure of collagen for the origins of molecular packing". In: *J. Mol. Biol.* 79.1 (1973), pp. 137–148.

[54] D. Hulmes, R. R. Bruns, and J. Gross. "On the state of aggregation of newly secreted procollagen." In: *PNAS* 80.2 (1983), pp. 388–392.

[55] D. Hulmes, T. J. Wess, D. J. Prockop, and P. Fratzl. "Radial packing, order, and disorder in collagen fibrils". In: *Biophys. J.* 68.5 (1995), pp. 1661–1670.

[56] E. S. Hwang, G. Thiagarajan, A. S. Parmar, and B. Brodsky. "Interruptions in the collagen repeating tripeptide pattern can promote supramolecular association". In: *Protein Sci.* 19.5 (2010), pp. 1053–1064.

[57] E. M. Jones, C. A. Cochrane, and S. L. Percival. "The effect of pH on the extracellular matrix and biofilms". In: *Adv. Wound Care.* 4.7 (2015), pp. 431–439.

[58] A. Jongeling, C. Svaneborg, and R. d. Vries. "Interaction patterns for staggered assembly of fibrils from semiflexible chains". In: *Symmetry* 12.11 (2020), p. 1926.

[59] K. E. Kadler, Y. Hojima, and D. Prockop. "Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase. Assay of critical concentration demonstrates that collagen self-assembly is a classical example of an entropy-driven process." In: *J. Biol. Chem.* 262.32 (1987), pp. 15696–15701.

[60] S. Kawashima and M. Kanehisa. "AAindex: amino acid index database". In: *Nucleic Acids Res.* 28.1 (2000), pp. 374–374.

[61] M. Koch, F. Laub, P. Zhou, R. A. Hahn, S. Tanaka, R. E. Burgeson, D. R. Gerecke, F. Ramirez, and M. K. Gordon. "Collagen XXIV, a vertebrate fibrillar collagen with structural features of invertebrate collagens: selective expression in developing cornea and bone". In: *J. Biol. Chem.* 278.44 (2003), pp. 43236–43244.

[62] T. Koorman, K. A. Jansen, A. Khalil, P. D. Haughton, D. Visser, M. A. Rätze, W. E. Haakma, G. Sakalauskaitè, P. J. van Diest, J. de Rooij, et al. "Spatial collagen stiffening promotes collective breast cancer cell invasion by reinforcing extracellular matrix alignment". In: *Oncogene* 41.17 (2022), pp. 2458–2469.

[63] A. A. Kornyshev, D. J. Lee, S. Leikin, and A. Wynveen. "Structure and interactions of biological helices". In: *Reviews of Modern Physics* 79.3 (2007), pp. 943–996.

[64] N. Kuznetsova and S. 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". In: *J. Biol. Chem.* 274.51 (1999), pp. 36083–36088.

[65] S. Leikin, D. Rau, and V. Parsegian. "Temperature-favoured assembly of collagen is driven by hydrophilic not hydrophobic interactions". In: *Nature structural biology* 2.3 (1995), pp. 205–210.

[66] E. Leikina, M. Mertts, N. Kuznetsova, and S. Leikin. "Type I collagen is thermally unstable at body temperature". In: *PNAS* 99.3 (2002), pp. 1314–1318.

[67] Y. Li, A. Asadi, M. R. Monroe, and E. P. Douglas. "pH effects on collagen fibrillogenesis in vitro: Electrostatic interactions and phosphate binding". In: *Mater. Sci. Eng. C* 29.5 (2009), pp. 1643–1649.

[68] J. Liu, W. Yong, Y. Deng, N. R. Kallenbach, and M. Lu. "Atomic structure of a tryptophan-zipper pentamer". In: *PNAS* 101.46 (2004), pp. 16156–16161.

[69] M. Maeda, T. Matsuzaki, F. Akai, S. Hashimoto, and H. Takagi. "Occurrence of long-spacing collagen in the intramuscular nerves of biopsied muscle tissues". In: *Med. Electron Microsc.* 29 (1996), pp. 124–128.

[70] E. Makareeva and S. Leikin. "Collagen structure, folding and function". In: *Osteogenesis imperfecta*. Elsevier, 2014, pp. 71–84.

[71] J. M. Mason and K. M. Arndt. "Coiled coil domains: stability, specificity, and biological implications". In: *ChemBioChem* 5.2 (2004), pp. 170–176.

[72] D. J. McBride Jr, V. Choe, J. R. Shapiro, and B. Brodsky. "Altered collagen structure in mouse tail tendon lacking the α2 (I) chain". In: *J. Mol. Biol.* 270.2 (1997), pp. 275–284.

[73] K. M. Meek. "Corneal collagen—its role in maintaining corneal shape and transparency". In: *Biophys. Rev.* 1.2 (2009), pp. 83–93.

[74] A. Miller and J. Wray. "Molecular packing in collagen". In: *Nature* 230.5294 (1971), pp. 437–439.

[75] S. Miyazawa and R. L. Jernigan. "Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading". In: *J. Mol. Biol.* 256.3 (1996), pp. 623–644.

[76] K. Mizuno, H. P. Bächinger, Y. Imamura, T. Hayashi, and E. Adachi. "Fragility of reconstituted type V collagen fibrils with the chain composition of α1 (V) α2 (V) α3 (V) respective of the D-periodic banding pattern". In: *Connect. Tissue Res.* 54.1 (2013), pp. 41–48.

[77] S. Morozova and M. Muthukumar. "Electrostatic effects in collagen fibril formation". In: *J. Chem. Phys* 149.16 (2018), pp. 163333–163333-9.

[78] M. Morvan and I. Mikšík. "The chiral proteomic analysis applied to aging collagens by LC-MS: Amino acid racemization, post-translational modifications, and sequence degradations during the aging process". In: *Anal. Chim. Acta* 1262 (2023), p. 341260.

[79] S. Neukirch, A. Goriely, and A. C. Hausrath. "Chirality of coiled coils: elasticity matters". In: *Phys. Rev. Lett.* 100.3 (2008), p. 038105.

[80] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Res.* 44.D1 (2016), pp. D733–D745.

[81] R. Ochoa, B. J B, and F. Thomas. "pyPept: a python library to generate atomistic 2D and 3D representations of peptides". In: *J. Cheminform.* 15.79 (2023), pp. 1–10.

[82] J. P. Orgel, T. C. Irving, A. Miller, and T. J. Wess. "Microfibrillar structure of type I collagen in situ". In: *PNAS* 103.24 (2006), pp. 9001–9005.

[83] J. P. Orgel, A. Miller, T. C. Irving, R. F. Fischetti, A. P. Hammersley, and T. J. Wess. "The in situ supermolecular structure of type I collagen". In: *Structure* 9.11 (2001), pp. 1061–1069.

[84] J. P. Orgel, A. V. Persikov, and O. Antipova. "Variation in the helical structure of native collagen". In: *PLoS One* 9.2 (2014), e89519.

[85] C. N. Pace, G. R. Grimsley, and J. M. Scholtz. "Protein ionizable groups: pK values and their contribution to protein stability and solubility". In: *J. Biol. Chem.* 284.20 (2009), pp. 13285–13289.

[86] M. F. Paige, J. K. Rainey, and M. C. Goh. “Fibrous long spacing collagen ultrastructure elucidated by atomic force microscopy”. In: *Biophys. J.* 74.6 (1998), pp. 3211–3216.

[87] M. Paige, J. Rainey, and M. Goh. “A study of fibrous long spacing collagen ultrastructure and assembly by atomic force microscopy”. In: *Micron* 32.3 (2001), pp. 341–353.

[88] S. Perret, C. Merle, S. Bernocco, P. Berland, R. Garrone, D. J. Hulmes, M. Theisen, and F. Ruggiero. “Unhydroxylated triple helical collagen I produced in transgenic plants provides new clues on the role of hydroxyproline in collagen folding and fibril formation”. In: *J. Biol. Chem.* 276.47 (2001), pp. 43693–43698.

[89] A. V. Persikov, J. A. Ramshaw, A. Kirkpatrick, and B. Brodsky. “Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability”. In: *Biochemistry* 44.5 (2005), pp. 1414–1422.

[90] J. A. Petruska and A. J. Hodge. “A subunit model for the tropocollagen macromolecule”. In: *PNAS* 51.5 (1964), pp. 871–876.

[91] K. A. Piez. “History of extracellular matrix: a personal view”. In: *Matrix Biol.* 16.3 (1997), pp. 85–92.

[92] K. A. Piez and B. L. Trus. “Sequence regularities and packing of collagen molecules”. In: *J. Mol. Biol.* 122.4 (1978), pp. 419–432.

[93] D. A. Plumb, V. Dhir, A. Mironov, L. Ferrara, R. Poulsom, K. E. Kadler, D. J. Thornton, M. D. Briggs, and R. P. Boot-Handford. “Collagen XXVII is developmentally regulated and forms thin fibrillar structures distinct from those of classical vertebrate fibrillar collagens”. In: *J. Biol. Chem.* 282.17 (2007), pp. 12791–12795.

[94] K. Poole, K. Khairy, J. Friedrichs, C. Franz, D. A. Cisneros, J. Howard, and D. Mueller. “Molecular-scale topographic cues induce the orientation and directional movement of fibroblasts on two-dimensional collagen surfaces”. In: *J. Mol. Biol.* 349.2 (2005), pp. 380–386.

[95] A. M. Puszkarska, D. Frenkel, L. J. Colwell, and M. J. Duer. “Using sequence data to predict the self-assembly of supramolecular collagen structures”. In: *Biophys. J.* 121.16 (2022), pp. 3023–3033.

[96] J. K. Rainey and M. C. Goh. “A statistically derived parameterization for the collagen triple-helix”. In: *Protein Sci.* 11.11 (2002), pp. 2748–2754.

[97] J. K. Rainey, C. K. Wen, and M. C. Goh. “Hierarchical assembly and the onset of banding in fibrous long spacing collagen revealed by atomic force microscopy”. In: *Matrix Biol.* 21.8 (2002), pp. 647–660.

[98] M. Raspanti, M. Reguzzoni, M. Protasoni, and P. Basso. "Not only tendons: The other architecture of collagen fibrils". In: *Int. J. Biol. Macromol.* 107 (2018), pp. 1668–1674.

[99] C. K. Revell, O. E. Jensen, T. Shearer, Y. Lu, D. F. Holmes, and K. E. Kadler. "Collagen fibril assembly: New approaches to unanswered questions". In: *Matrix Biol. Plus* 12 (2021), p. 100079.

[100] N. Rezaei, A. Lyons, and N. R. Forde. "Environmentally controlled curvature of single collagen proteins". In: *Biophys. J.* 115.8 (2018), pp. 1457–1469.

[101] N. Reznikov, M. Bilton, L. Lari, M. M. Stevens, and R. Kröger. "Fractal-like hierarchical organization of bone begins at the nanoscale". In: *Science* 360.6388 (2018), eaao2189.

[102] F. H. Silver. "A molecular model for linear and lateral growth of type I collagen fibrils". In: *Coll. Relat. Res.* 2.3 (1982), pp. 219–229.

[103] D. Sloseris and N. R. Forde. "AGEing of collagen: The effects of glycation on collagen's stability, mechanics and assembly". In: *Matrix Biol.* 135 (2025), pp. 153–160.

[104] J. Smith. "Molecular pattern in native collagen". In: *Nature* 219.5150 (1968), pp. 157–158.

[105] R. H. Stinson and P. R. Sweeny. "Skin collagen has an unusual d-spacing". In: *Biochim. Biophys. Acta - Prot. Struct.* 621.1 (1980), pp. 158–161.

[106] W. Stumm and J. J. Morgan. *Aquatic chemistry: chemical equilibria and rates in natural waters*. John Wiley & Sons, 2013.

[107] A. Stylianou. "Assessing collagen D-band periodicity with atomic force microscopy". In: *Materials* 15.4 (2022), p. 1608.

[108] K. M. Towe. "Oxygen-collagen priority and the early metazoan fossil record". In: *PNAS* 65.4 (1970), pp. 781–788.

[109] B. L. Trus and K. A. Piez. "Compressed microfibril models of the native collagen fibril". In: *Nature* 286.5770 (1980), pp. 300–301.

[110] B. L. Trus and K. A. Piez. "Molecular packing of collagen: three-dimensional analysis of electrostatic interactions". In: *J. Mol. Biol.* 108.4 (1976), pp. 705–732.

[111] J. H. Van Lint and R. M. Wilson. *A course in combinatorics*. Cambridge university press, 2001.

[112] M. Venturoni, T. Gutsmann, G. E. Fantner, J. H. Kindt, and P. K. Hansma. "Investigations into the polymorphism of rat tail tendon fibrils using atomic force microscopy". In: *Biochem. Biophys. Res. Commun.* 303.2 (2003), pp. 508–513.

[113] M. G. Venugopal, J. A. Ramshaw, E. Braswell, D. Zhu, and B. Brodsky. "Electrostatic interactions in collagen-like triple-helical peptides". In: *Biochemistry* 33.25 (1994), pp. 7948–7956.

[114] P. Voziyan, S. Uppuganti, M. Leser, K. L. Rose, and J. S. Nyman. "Mapping glycation and glycoxidation sites in collagen I of human cortical bone". In: *BBA Adv.* 3 (2023), p. 100079.

[115] D. Wallace. "The role of hydrophobic bonding in collagen fibril formation: a quantitative model". In: *Biopolymers* 24.9 (1985), pp. 1705–1720.

[116] D. G. Wallace. "The relative contribution of electrostatic interactions to stabilization of collagen fibrils". In: *Biopolymers* 29.6-7 (1990), pp. 1015–1026.

[117] C. K. Wen and M. C. Goh. "Fibrous long spacing type collagen fibrils have a hierarchical internal structure". In: *Proteins: Struct. Funct. Bioinf.* 64.1 (2006), pp. 227–233.

[118] A. Wieczorek, N. Rezaei, C. K. Chan, C. Xu, P. Panwar, D. Brömme, E. F. Merschrod S, and N. R. Forde. "Development and characterization of a eukaryotic expression system for human type II procollagen". In: *BMC Biotechnol.* 15 (2015), pp. 1–17.

[119] Y. Xu and M. Kirchner. "Collagen mimetic peptides". In: *Bioengineering* 8.1 (2021), p. 5.

[120] Y. Xu and M. Kirchner. "Segment-Long-Spacing (SLS) and the Polymorphic Structures of Fibrillar Collagen". In: *Macromolecular Protein Complexes IV: Structure and Function*. Springer, 2022, pp. 495–521.

[121] L. Yang, K. Van der Werf, P. J. Dijkstra, J. Feijen, and M. L. Bennink. "Micromechanical analysis of native and cross-linked collagen type I fibrils supports the existence of microfibrils". In: *J. Mech. Behav. Biomed. Mater.* 6 (2012), pp. 148–158.

[122] S. M. Yu, Y. Li, and D. Kim. "Collagen mimetic peptides: progress towards functional applications". In: *Soft Matter* 7.18 (2011), pp. 7927–7938.

[123] W. Zhu, K. Li, Q. Liu, H. Zhong, C. Xu, J. Zhang, H. Kou, B. Wei, and H. Wang. "Effect of molecular chirality on the collagen self-assembly". In: *New J. Chem.* 45.35 (2021), pp. 15863–15868.

[124] A. Zolotarjov, R. Kröger, and D. O. Pushkin. "Chiral interactions between tropocollagen molecules determine the collagen microfibril structure". In: *arXiv preprint arXiv:2504.21484* (2025).