# Dissertation

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the Ruperto Carola University of Heidelberg, Germany

for the degree of

**Doctor of Natural Sciences**

Put forward by:

**Benedikt Rennekamp**

born in Mainz, Germany

Oral examination: November 7th, 2023

# Multi-Scale Simulations of Collagen Failure and Mechanoradicals

Referees:      Prof. Dr. Frauke Gräter

           Prof. Dr. Tristan Bereau

# Multi-Scale Simulations of Collagen Failure and Mechanoradicals

Collagen, the most abundant protein in the human body, must withstand high mechanical loads due to its structural role in tendons, skin, bones, and other connective tissue. It was recently found that tensed collagen creates mechanoradicals by homolytic bond scission in the sub-failure regime. The locations and types of initial rupture sites critically decide on both the mechanical and chemical impact of these micro-ruptures on the tissue, but are yet to be explored.

We here employ hybrid scale-bridging simulations to determine these first breakage points in collagen, combining existing and newly developed methods tailored towards collagen's hierarchical structure. We improved our Kinetic Monte Carlo/Molecular Dynamics scheme to simulate bond scissions at the all-atom level, and also developed a mesoscopic ultra coarse-grained description of a collagen fibril.

We find collagen crosslinks to rupture first, and identify individual sacrificial bonds in trivalent crosslinks that break preferentially, without compromising structural integrity. Collagen's weak bonds funnel ruptures such that the potentially harmful mechanoradicals are readily stabilized. Our simulations further suggest the length of helices between pairs of crosslinks to determine the trade-off between overall strength and breakage specificity. The combined results suggest this unique failure mode of collagen to be tailored towards combatting an early onset of macroscopic failure and material ageing.

## Skalenübergreifende Simulationen von Rissen und mechanisch induzierten Radikalen in Kollagen

Kollagen, das häufigste Protein im menschlichen Körper, muss aufgrund seiner strukturellen Rolle in Sehnen, Haut, Knochen und anderen Bindegeweben hohen mechanischen Belastungen standhalten. Kürzlich wurde entdeckt, dass Kollagen unter Spannung durch homolytische Bindungsbrüche Mechanoradikale erzeugt, aber ohne makroskopisch zu reißen. Die Positionen und Arten dieser initialen Bruchstellen bestimmen maßgeblich die mechanischen und chemischen Auswirkungen dieser Mikrobrüche auf das Gewebe, und sind bisher noch nicht ausreichend erforscht.

In dieser Arbeit verwenden wir skalenübergreifende Simulationen, um diese ersten Bruchstellen in Kollagen zu bestimmen. Dabei kombinieren wir bestehende und neue, auf die hierarchische Struktur von Kollagen zugeschnittene Methoden. Wir haben unser Kinetic Monte Carlo/Molecular Dynamics Verfahren verbessert, das Bindungsbrüche auf atomarer Ebene simuliert, und auch eine mesoskopische grobkörnige Beschreibung einer Kollagenfibrille entwickelt.

Unsere Ergebnisse zeigen, dass Querverbindungen in Kollagen (sogennante Crosslinks) zuerst brechen. Wir identifizieren insbesondere überzählige Verbindungen in dreiarmigen Crosslinks, die bevorzugt brechen, ohne dass dies die strukturelle Integrität beeinträchtigt. Die Brüche werden in diese schwachen Bindungen gelenkt, sodass die potenziell schädlichen Mechanoradikale leicht stabilisiert werden können. Unsere Simulationen legen weiterhin nahe, dass die Länge der helikalen Moleküle zwischen Paaren von Crosslinks den Kompromiss zwischen Gesamtfestigkeit und Bruchspezifität bestimmt. Unsere Ergebnisse zeigen, dass dieser einzigartige Bruchvorgang von Kollagen einem frühen makroskopischen Versagen und Materialalterung entgegenwirkt.

# Related publications

Parts of this thesis (mainly results of section 3.1 and chapter 4) have already been published. Throughout the whole thesis, including introduction and discussion, text and figures of these publications will be reused, both unaltered or adapted:

> A. Obarska-Kosinska, **B. Rennekamp**, A. Ünal, and F. Gräter: "Colbuilder: a server to build collagen fibril models." *Biophysical Journal* 120, 3544–3549 (2021)

> **B. Rennekamp**, C. Karfusehr, M. Kurth, A. Ünal, D. Monego, K. Riedmiller, G. Gryn'ova, D. M. Hudson and F. Gräter: "Collagen breaks at weak sacrificial bonds taming its mechanoradicals." *Nature Communications* 14, 2075 (2023).

In addition, we expect two collaborative projects (sections 3.2.3 and 3.3) to result in publications with contributing authorships, and also currently prepare to publish chapter 5:

> **B. Rennekamp**, H. Grubmüller and F. Gräter: "Collagen is a buffer of mechanical and oxidative stress." (in preparation).

For transparency, we also published other related work during the PhD, but conducted most of our contributions for this already *in a previous Master's thesis [1]*:

> C. Zapp*, A. Obarska-Kosinska*, **B. Rennekamp**, M. Kurth, D. M. Hudson, D. Mercadante, U. Barayeu, T. P. Dick, V. Denysenkov, T. Prisner, M. Bennati, C. Daday, R. Kappl and F. Gräter: "Mechanoradicals in tensed tendon collagen as a source of oxidative stress." *Nature Communications* 11, 2315 (2020).

> **B. Rennekamp**, F. Kutzki, A. Obarska-Kosinska, C. Zapp, and F. Gräter: "Hybrid kinetic Monte Carlo/molecular dynamics simulations of bond scissions in proteins. *Journal of Chemical Theory and Computation* 16: 553-563 (2020)

Additionally, this work has received a computing time grant of 16 million core-h on the GCS Supercomputer SuperMUC-NG at Leibniz Supercomputing Centre. For this reason, we already wrote a short related report for their public results book.

# Contents

# Introduction

Mechanical forces acting on proteins are known to have a large range of implications, from conformational changes to regulation of protein activity to signalling [2, 3, 4]. These forces can arise due to interactions with binding partners, thermal fluctuations, external stresses or can even be actively generated like in muscle. A prime example of a force-bearing biomolecule is collagen. It provides strength to a wide range of connective tissue types: bones, skins, ligaments, cartilage, and tendons [5]. For this reason, collagen has been studied extensively from a mostly mechanical perspective both in theory and experiments, including also its failure in high stress situations [6, 7]. Only recently, however, a new view brought up by our group emerged, connecting mechanical load also to chemical stress: Upon loading in a sub-failure regime, collagen can create mechanoradicals by homolytic bond rupture [8, 9]. Due to their high reactivity, radicals are often considered potentially damaging, though also other functions as in ribonucleotide reductases (RNR) are known [10]. In collagen, we found the mechanoradicals to readily stabilize on dihydroxy-phenylalanines (DOPAs), with these highly efficient radical-scavenging residues built into the collagen protein [8, 9]. Subsequently, $H_2O_2$ is created, a biochemical signalling molecule that could initiate further physiological process to deal with the high stress situation, as it is known from human exercise [11].

Interestingly, the observation of radicals forming from homolytic bond rupture in polymers dates back almost a century to Staudinger [12]. Since then, the creation of so-called mechanoradicals (radicals generated by mechanical loading) has been studied intensively [13]. In the field of polymer mechanochemistry, selective bond rupture has been achieved by synthetically incorporating weak bonds into the polymer. They serve as sacrificial bonds to increase the toughness of the material [14], or as mechanophores to

1

report on mechanical force [15, 16, 17]. Sacrificial bonds that so far have been observed in biological systems are non-covalent in nature, such as hydrogen bonds or ionic bonds in bone collagen [18, 19]. If protein materials harness the virtue of mechanically weak covalent bonds as sacrificial bonds or mechanophores, has been, to our knowledge, unknown before. In collagen, selective bond scission could critically determine mechanical stability, further radical-mediated damage, and modes of mechano-sensing.
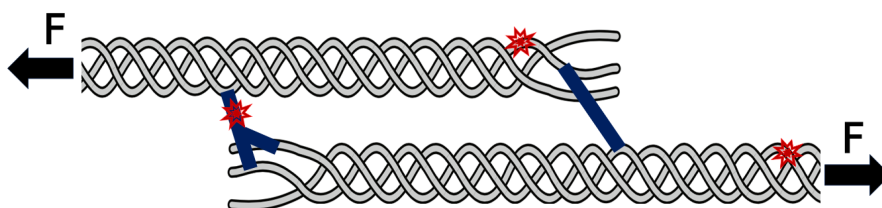


**Figure 1.1: Introductory scheme on ruptures in tensed crosslinked collagen molecules.** Collagen triple helices are connected via crosslinks (blue) at their terminal regions to other molecules; the crosslinks can be either trivalent with two arms (left) or divalent (right). In this simple representation with just two molecules under force, we show three breakage possibilities (red, from left to right): in the crosslink, in the backbone nearby a crosslink, or elsewhere in the backbone.

The first determinant of rupture is the relative strength of the chemical bonds. Any protein backbone features three types of bonds: the C-N peptide bond, the $C_\alpha$-N bond and the $C_\alpha$-C bond. Due to its partial double bond character, we can exclude the peptide bond from the list of likely rupture candidates. Previous studies suggest bond dissociation energies for these two bond types to be in a similar range [8, 20], so the chemical environment matters. Collagen also features enzymatically derived crosslinks connecting the individual triple helices, as depicted schematically in Fig. 1.1. A variety of crosslink chemistries has evolved, which can be categorized into divalent (or premature) and trivalent (or mature) crosslinks. All of them are derived from lysines, and as a consequence feature again single bonds of the C-C and C-N type.

The second determinant of rupture is how force distributes through the complex hierarchical structure of collagen and loads its chemical bonds. In collagen, $\alpha$-chains wind up to triple helices that are packed in a quasi-hexagonal shape in cross-section. They are braided into each other in a non-trivial way along the fibril axis, forming fibrils on a larger scale [21]. This three-dimensional structure, combined with the crosslinking in between the triple helices, results in a molecular force distribution network that

critically decides on the rupture propensity of an individual bond. In the simplest model, the Bell-Evans model, force acting on a bond exponentially increases its rate of rupture [22, 23, 24]. Several previous studies have suggested that crosslinks are likely rupture sites in collagen: Crosslinks have been reported as responding to stress and suggested as rupture candidates in smaller scale molecular simulations [25] as well as in coarse-grained models [26]. We previously identified crosslinks as well as adjacent backbone bonds as the most strongly loaded links [8, 27]. Finally, recent semi-empirical calculations identified $C_\alpha$-C bonds at X-positions (in particular of prolines) in the typical GLY-X-Y amino acid pattern of collagen as rupture candidates [28], but also suggested one exemplary crosslink as another preferred rupture point. None of the previous studies allowed quantitative conclusions on specific bond rupture sites within the collagen fibril.

A powerful tool to gain insights into such processes, and more generally into the time traces of proteins and other molecules at an atomistic level of detail, are Molecular Dynamics (MD) simulations. The time evolution of the system is simulated by classically integrating Newton's equation of motion for each particle. After decades of refinement of the empirical force fields that describe the system's energy, MD simulations have reached predictive power with widespread usage. Application of (external) steering forces also allows us to study mechanical properties of proteins under load [29]. However, in classical MD simulations, the bonded interactions are usually fixed and reactions such as bond rupture cannot occur. Whereas there are several hybrid methods like quantum mechanics / molecular mechanics (QM/MM) that are capable of simulating bond ruptures, these require the definition of a quantum mechanical (QM) region beforehand. As collagen might break throughout the fibril, this is not suitable for us. Further, reactive force fields like ReaxFF exist [30], but they have not (yet) reached the accuracy for biomolecules as classical force fields and the efficiency to handle large systems like collagen [31, 32]. In the future, machine-learning based methods like the ANI force field [33] might be able to achieve this, and already match quantum-chemical accuracy for some chemical problems, but are not yet at a stage to handle large folded proteins. For these reasons, we have designed a hybrid simulation scheme Kinetic Monte Carlo/Molecular Dynamics (KIMMDY) in previous work [27]. In short, it takes the bond elongations from the MD simulations as input for a Bell-type model to calculate rupture rates. Bonds are modelled with Morse potentials, and the initial barrier height is effectively lowered by the work that acts on the bond. The individual bond rupture rates are then obtained from the barrier crossing problem of this new effective potential

and are used in a Kinetic Monte Carlo step determining the scission site.

A precondition to investigate such processes with MD simulations is an atomistic collagen I structure as it occurs in native tissue. Before, only individual triple helices were available at the level of atomistic detail, but no collagen fibril model. To this end, in our group, all-atom models of cross-linked collagen fibrils were built by integrating the low-resolution structure of collagen fibril available from X-ray fiber diffraction with high-resolution structures of short collagen-like peptides from X-ray crystallography and mass spectrometry (MS) data. We created a web resource ColBuilder [34] of collagen models for 20 different species with a large variety of crosslink types and localization within the fibril to facilitate structure-based analyses and simulations of type I collagen in health and disease. To easily enable simulations, parameters of the modelled crosslinks are provided for an Amber force field as well. These models and parameters needed yet to be validated, to prove that they accurately reproduce the properties of a collagen microfibril.

Further, collagen is a structural protein with characteristic features across many length scales. The triple helices, known as collagen molecules, are the building blocks of that complex hierarchical structure collagen exhibits on the mesoscale. These individual molecules already comprise about 3,000 amino acids and have a length of about 300 nm. They wind up to microfibrils and, finally, macroscopic fibers. Despite the advancement of computational power, there is obviously a limit on the level of detail, and especially on the length and time scales, that can be simulated in an all-atom MD approach. For this reason, a variety of coarser and coarser descriptions has been developed: Starting from Martini (about four heavy atoms per bead) to mesoscopic (continuum) models [35, 36, 37], but without the focus on failure mechanisms.

## 1.1   Aims of this thesis

Taking all the above together, where collagen, and potentially other protein materials, rupture under mechanical load remains to be clarified. The microscopic failure mechanisms, however, define macroscopic mechanical properties such as ultimate strain and toughness. Further, the positions of the weakest links within the structure determine where and what types of mechanoradicals are formed and how they will propagate through and potentially damage the system until they get scavenged. A tempting hypothesis for collagen is, therefore, that rupture occurs in the vicinity of DOPAs to rapidly stabilize the primary radical and thereby prevent uncontrolled radical reactions such

as migration and recombination. Nailing down the rupture process of collagen, being the major substituent of tissues, is a key necessity for addressing tissue degradation and ageing, and for guiding tissue engineering.

Hence, the central questions of this thesis include:

- Where does collagen break under tension?

- What are the major rupture mechanisms?

- Which consequences do these micro-ruptures have on the material, both mechanically and (subsequently) chemically?

To achieve the aim of answering these questions, we took a scale-bridging simulation approach combining both existing and newly developed methods tailored towards the hierarchical structure of collagen spanning many length scales (see Fig. 1.2). First and most importantly, we utilized our reactive MD scheme KIMMDY, into which we incorporated new bond dissociation energies (BDEs) as input from high-level QM, to get rupture propensities throughout our collagen microfibril. For this, we made comprehensive use of different variants (both with regard to species and to crosslink types) of the new large atomistic models. Secondly, we developed an ultra coarse-grained model on the mesoscale, parametrized bottom-up from our MD simulations, that is able to simulate fibril failure on larger length scales and bolsters our understanding of the underlying mechanisms.

## 1.2   Outline of this thesis

This thesis is structured such that chapter 2 gives an introduction both about the current collagen research, in more detail, and about the theories that are fundamental to the here employed methods.

Chapter 3 deals with our contributions to the methodological challenges in modelling and simulating tensed collagen, and, more generally, in simulating bond ruptures, on the atomistic level. During this thesis, we especially validated the ColBuilder models comparing the structural conformations obtained with MD with available experimental data, as we elucidate in section 3.1. Further, we incorporated new bond dissociation energies (obtained with comprehensive and high-level quantum mechanical calculations by other people in our group) into KIMMDY [27] to assess collagen's mechanochemical rupture with steered MD simulations. We especially updated our hybrid simulation
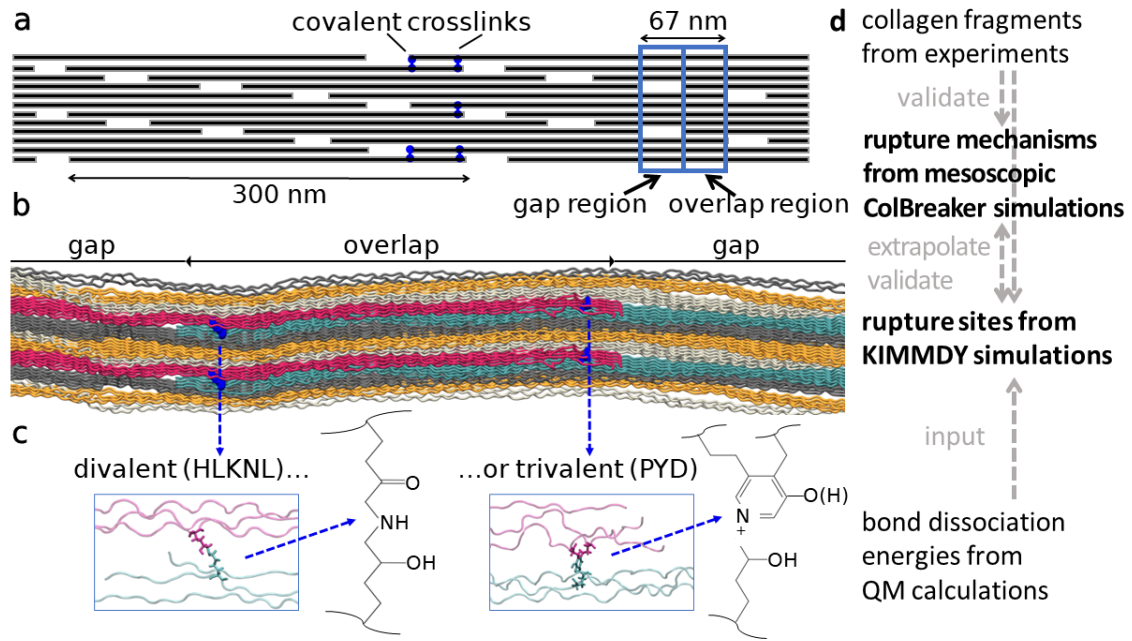
**Figure 1.2: Collagen structure, spanning multiple length scales, and our corresponding methods. a** 2D projection of the staggered arrangement of collagen triple helices, each 300 nm. This results in the typical overlap and gap regions of collagen, but cannot accurately depict the more complex thee-dimensional braiding of triple helices around each other. **b** Our atomistic model spanning one overlap (middle) and about one gap region (split into two parts) of collagen. **c** Zoom in on enzymatic crosslinks connecting the triple helices. Different chemistries (divalent or trivalent) are possible at these positions, for example, Hydroxylysino-keto-norleucine (HLKNL) or Pyridinoline (PYD). **d** Our workflow combines different methods, as collagen spans multiple length scales. In particular, KIMMDY and ColBreaker simulations (both in bold) were conducted during the course of this thesis. Figure adapted from our own publication [38].

scheme with respect to the crosslink chemistry described above. These methodological advances, among others, are discussed in section 3.2. Beside the barrier height, we also carried out a thorough investigation of limitations and potential improvements of the force calculation from MD as the second major contribution to the rupture rate. Moreover, we present our preliminary work on hydrogen atom transfer (HAT), a natural follow-up reaction for the highly reactive radicals that arise due to the homolytic bond rupture. We also shortly discuss our collaboration on a project using KIMMDY for bond rupture in DNA. Alongside in the same chapter in section 3.3, we consider our contributions to even larger models and MD simulations, with more than 45 million atoms, in another ongoing collaboration.

Having a validated model and fine-tuned our computational methods, in chapter 4 we present our KIMMDY simulations for a wide range of models, varying species and crosslink types, to obtain rupture propensities. We observe a large concentration of ruptures of $\sim$70 % in crosslink bonds, the two key factors of which are up to $\sim$120 kJ/mol lower bond dissociation energies (BDEs) of crosslink compared to backbone bonds (see section 4.1), and the higher stress concentration in this region. As a result, analysed in detail in section 4.2, most rupture sites are in direct vicinity to DOPA residues which can prevent uncontrolled radical reactions by rapid scavenging. However, there is a strong competition with unspecific breakages in the vast amount of backbone bonds throughout the fibril. In section 4.3, we show that our simulations find a particular bond in one of the two 'arms' of the trivalent crosslink Pyridinoline (PYD) and Deoxypyridinoline (DPD) to be particularly primed for rupture, due to a stabilizing conjugation of the resulting radicals. Its rupture does not compromise collagen integrity and leads to local stress relaxation, a hallmark of a sacrificial bond. In addition to the simulations, we analysed polyacrylamide gel electrophoresis (SDS-PAGE) coupled to mass spectrometry (MS) experiments of our collaborators (see section 4.4). They support altered collagen crosslinking in stressed tissue compared to control samples, as well as a widespread rupture within both $\alpha$1 and $\alpha$2-chains in collagen I, as evident from the redistribution of molecular masses in collagen upon mechanical stress.

Turning away from individual bonds to the (micro-)fibrillar features in chapter 5, we observe that there is a sequential rupturing between the crosslinks at the two possible connection sites in the same overlap region (presented in section 5.1). We find one connection to be consistently shorter, leading to a release of hidden length after the first ruptures – yet another property that is well-known from polymer networks with

sacrificial bonds (see section 5.2). In this way, additional energy can be dissipated and the toughness of the material is increased. At this level, however, it is not an individual sacrificial bond within an otherwise intact crosslink, but a whole crosslink that can break and has a 'back-up' at the other connection site.

Furthermore, we developed an ultra coarse particle-based model ColBreaker that we introduce in section 5.3. Guided by our atomistic results, we aimed to extrapolate to the mesoscopic scale. For this reason, ColBreaker abstracts complete gap and overlap regions of collagen into extensible worm-like chain (WLC) models that are connected with crosslinks, such that we arrive at only 10 beads per triple helix. We use our MD simulations to thoroughly bottom-up parametrize the interaction potentials, but also the breaking process and the topological features such as the aforementioned lengths of the force transmission pathways. In spite of the model's simplicity, we are able to recover important features, such as the force-extension curve. Moreover, we empirically fitted the force dependency of breakages rates to the results of our KIMMDY simulations.

Lastly, with ColBreaker, we explore the influence of larger scale topological features on the rupture propensities, far further than the MD accessible range, e.g. for the connections between crosslink sections in the fibril (section 5.4). We observe again a trade-off on how much comparatively weaker sections can direct rupture locations in a controlled manner, at the cost of an overall increased failure rate. ColBreaker also lets us compare this to hypothetical scenarios if collagen might have evolved differently. Our calculations imply that collagen might have found a sweet-spot in that compromise.

All results, together with their relevance, their limitations and an outlook to future work, are discussed in a combined fashion in chapter 6.

# CHAPTER 2

# Background, Theory and Methods

This chapter provides the theoretical foundation that is behind the biological questions and methods presented in this thesis. It includes a broader introduction to the current collagen research in the context of this thesis, as well as an overview of established methods like Molecular Dynamics (MD). We further introduce the basics of our previously developed bond rupture simulation scheme KIMMDY, often referencing back to that work [1, 27]. With respect to new self-developed methods or improvements of the existing schemes like KIMMDY, we refer to chapter 3 where we present our own contributions during the course of this thesis to this matter.

## 2.1   Collagen I: A hierarchical structural biopolymer

Collagen is the most abundant protein by molecular weight in the human body [41]. Even before humans, it evolved in the extracellular matrices of sponges, vertebrates and other mammals [42]. It has various functions, mostly in connective tissue from cartilage to bones to skin to tendon. By far (over 90% [43, 5]) the most common is collagen type I, but at least 28 types have been identified in humans alone [44]. In this thesis, we are focused on stretched collagen tendons, which are constituted mostly of type I collagen. For this reason, we use sequences of type I collagen in the atomistic models. For the minor percentages of other types, especially types III and V, a multiple sequence alignment indicates the transferability of our results [38]. In particular, it finds crosslinks and potential sites of dihydroxy-phenylalanine (DOPA) to be well conserved, which are most important for our results.
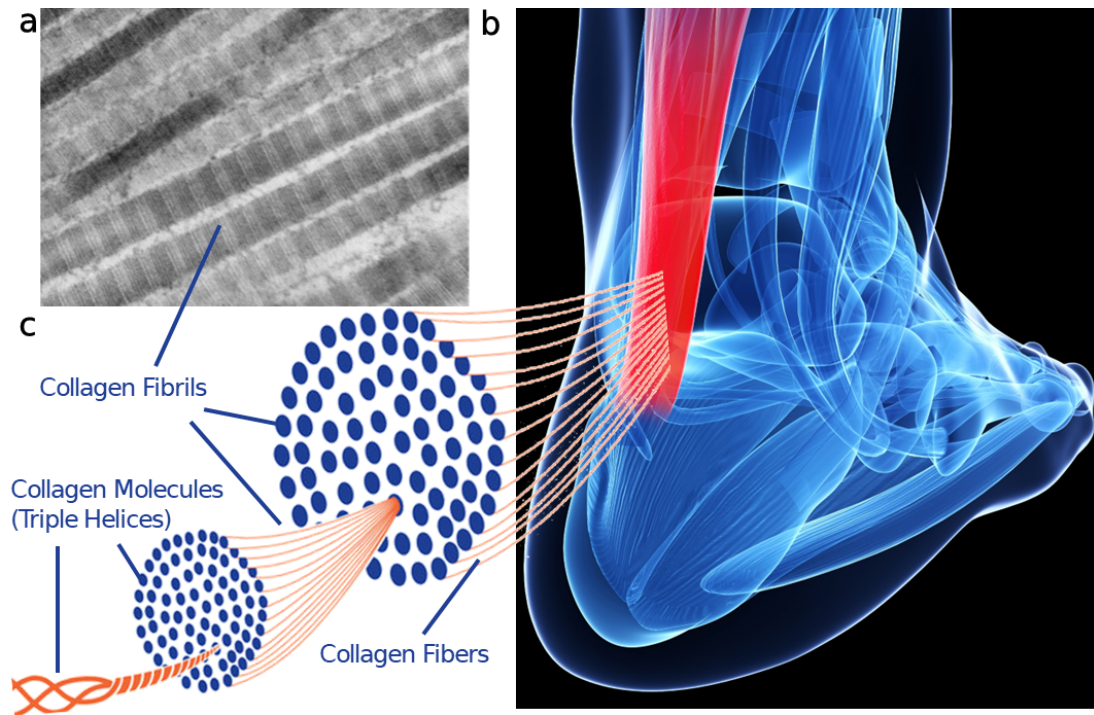
**Figure 2.1: Collagen across length scales**. **a** Transmission electron microscope image of connective tissue showing collagen type I fibrils. The stripes in the fibrils, the so-called D-pattern, have a periodicity of about 67 nm. **b** Artistic image of an Achilles tendon under stress. **c** Collagen triple helices wind up to fibrils on the micrometer length scale and to large macroscopic fibers, e.g. in tendon. Image Credits: **a** public domain, **b** bought from istockphoto.com/SciePro, **c** adapted from [40]

## 2.1.1   Collagen structure

Collagen features characteristic structural properties throughout many length scales, as shown in Figs. 2.1 and 2.2. On the molecular level in collagen type I, there are two types of chains, called $\alpha 1$(I) and $\alpha 2$(I), or Col1a1 and Col1a2, which differ slightly in amino acid sequence and length. They have the typical amino acid pattern GLY-X-Y in common, where proline is often found at the X position and hydroxyproline at the Y position. Two $\alpha 1$-chains and one $\alpha 2$-chain together wind up to a heterotrimer in the form of an extended right-handed triple helix. This structure is stabilized by hydrogen bonds between the glycine residues (only their side chains are small enough to point inside the triple helix) and CO groups of residues on opposite chains [45]. This superhelix has about 3.3 residues per turn, with a rise of almost 3 Å per residue; this means that the collagen helix is much more extended than the alpha helices of other proteins. One full collagen helix, also referred to as collagen molecule, has a diameter of about 1.5 nm and a length of 300 nm.

On the following length scale, a fibril is formed with a characteristic stacking, the D-pattern, as can be seen in the electron microscope image in Fig. 2.1a. These D-periods are about 67 nm long and contain two subunits: The overlap region, as the name suggests, is where beginning and ending collagen molecules overlap; this is shown in a simplified 2D schematic representation in Fig. 2.2a. In contrast, the gap region is less dense (for every 5 molecules in overlap, there are 4 in gap) and also less ordered. The 5 overlap regions that each collagen molecule comprises are often referred to as phases. In the cross-section, collagen is arranged in a (quasi-)hexagonal pattern [46]. In the gap region, however, the triple helices are intertwined and exchange positions with respect to their neighbors on the cross-section [21], as can be seen in Fig. 2.2b. If you follow the colored strands in this cross-section, it is evident that this position shift can make up a considerable lateral distance, which results in some collagen molecules being kinked in the gap region. The three-dimensional structure is, hence, more complicated than what the above two-dimensional scheme might suggest. This observation will also become important when building our collagen models for this thesis.

## 2.1.2   Covalent crosslinks in collagen

At the beginning and ending of collagen molecules, enzymatically derived crosslinks connect the different triple helices, as shown exemplary in blue in Fig. 2.2a. These cova-
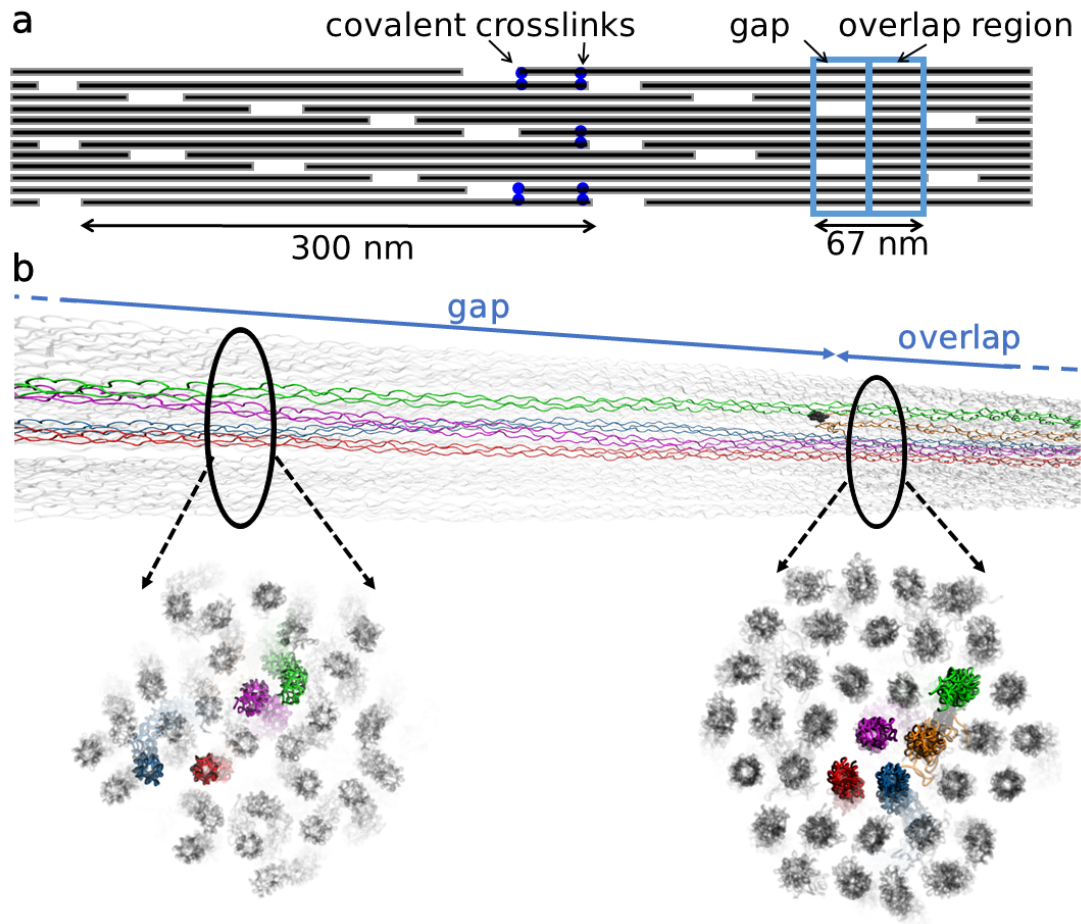
**Figure 2.2: Structure of a collagen (micro-)fibril. a** Two-dimensional scheme of staggered collagen molecules leading to the gap and overlap pattern. The triple helices are covalently connected with enzymatically derived crosslinks at their N- and C-terminal ends (a few exemplary positions are marked in blue). **b** Three-dimensional view along the fibril (top) and the cross-section (below) in the gap and overlap region of our atomistic model. The quasi-hexagonal packing can be seen in the overlap region, but also the intertwined strands in the overlap region. We colored one strand from each phase (more details in text), such that the shift in relative position can be followed.

lent connections provide strength to the network and are a major focus of this thesis, as we have already previously identified them to be highly stressed [8, 27]. They are formed with the help of lysyl oxidases, that create aldehydes from lysines or hydroxylysines. In turn, they can react with other hydroxylysines to form divalent connections between two triple helices. These are called immature crosslinks, in contrast to the mature (trivalent) crosslinks that develop over time due to spontaneous connection of divalent crosslinks with a third residue [47, 48]; see Fig. 2.3 for an example. Note that collagen overall has a low turnover rate, up to several years depending on tissue type. For this reason, the percentage of trivalent crosslinks increases during the lifespan.
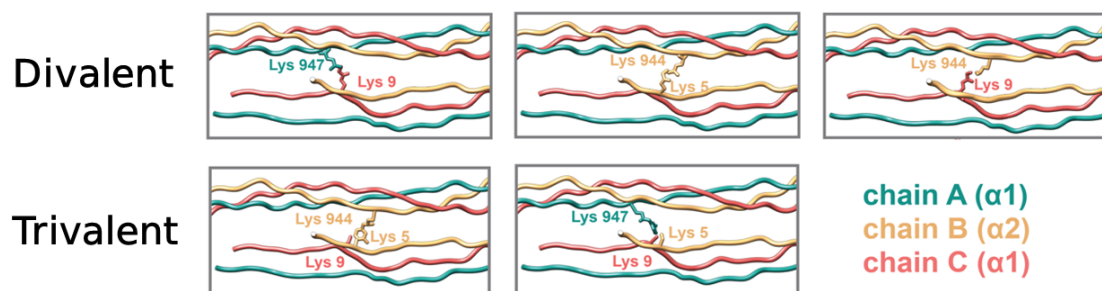


**Figure 2.3: Enzymatic crosslinks at the N-telopeptide region.** In each triple helix, there are two $\alpha 1$-chains (green and red) and one $\alpha 2$-chain (yellow) that contain lysines both at the N-terminal and C-terminal site. We depict the N-terminal region and show different types of crosslinks: Divalent crosslinks connecting to the same or a different chain of another triple helix, or trivalent crosslinks connecting two chains from one helix to another collagen molecule. Figure adapted from our own publication [34].

As all enzymatic crosslinks are derived from lysines, only certain positions in the opposing triple helices allow for enzymatic crosslinking. Still, there are various possibilities in the terminal regions, as illustrated in Fig. 2.3 for the N-telopeptide region. Another implication of their chemical pathway is that enzymatically-derived crosslinks consist of C-C and C-N bonds with a thermodynamic strength comparable to the protein backbone; which we will shed more light on in the results chapters. Elastin as another structural protein possesses similar crosslinking chemistry, but otherwise this feature is quite unique for biological material, where crosslinking is mostly achieved via weaker disulphide bridges (for example in keratin).

Beside these enzymatically driven crosslinks that have clearly defined positions, a different type of more randomly placed crosslinks is added over time with ageing.

Their formation is driven by oxidative reactions of sugars with collagen, which is why there are called advanced glycation endproducts (AGEs), with a very telling acronym. AGEs can stiffen the tissue, making old collagen less flexible, and might contribute to diseases [49, 50]. In this thesis, however, we mostly focus on healthy young tissue that is subjected to load and, therefore, we have not included AGE crosslinks in our models.

## 2.1.3   Mechanoradicals

In polymer research, radical generation upon mechanical stress is a well established and well studied phenomenon [13, 51, 52]. The discovery of homolytic bond rupture in polymers, for instance, dates back to Staudinger in 1930 [12]. We call the force-induced radicals "mechanoradicals". In biological material, radicals generally have received a lot of attention as well, e.g. in ribonucleotide reductases [53, 54, 10]. Radical generation was also reported in collagen, e.g. after exposure to gamma rays [55]. "Free radicals" made it even into popular science, with the (simplified) belief that they are particularly dangerous due to their high reactivity. This created a whole industry of food supplements containing antioxidants, extracted for example from green tea [56, 57].
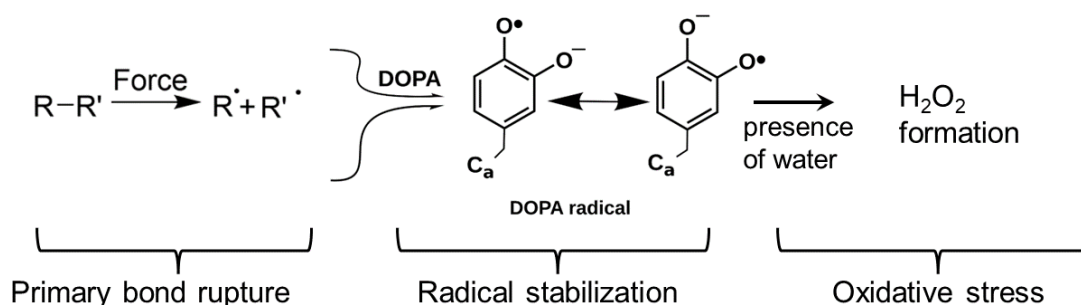


**Figure 2.4: Mechanoradicals in collagen can produce hydrogen peroxide.** Collagen tendons under force have been shown to produce mechanoradicals that can stablize on dihydroxyphenylalanine (DOPA). These can react further with water, producing hydrogen peroxide. Figure adapted from our own publication [8].

In contrast to polymers, however, there have been only very few examples showing that radicals can originate from mechanical loading in biological materials. Radical formation by mechanical force in biomaterials has been reported in 1987 when cutting keratin [58], or in 1996 when grinding bone [59]. Whereas these two early examples indeed apply mechanical procedures, they are far from force regimes that are biologically relevant under physiological conditions. However, in our group, we have recently established

that mechanoradicals can also arise in collagen in the sub-failure regime [8]. The key experiment was to detect an increased signal in electron paramagnetic resonance (EPR) of stressed vs. unstressed rat tail tendon. As the EPR can detect unpaired electrons only, we could conclude that homolytic bond rupture must have happened, i.e. the two electrons that made up the bond split up to the two created fragments. Subsequently, we could detect $H_2O_2$, a biological signalling molecule. This novel conversion from mechanical to oxidative stress is shown in Fig. 2.4 and is a main motivator for the research within this thesis. Rupture type and location will influence the subsequent radical generation, which is why their identification is important. Migration of the highly reactive radicals can impact the material in various potentially damaging ways before these are finally stabilized. In order to obtain detailed insights into these processes, simulations as we will introduce below are valuable tools.

## 2.2   Molecular modelling and simulations

When experiments come to their limits in terms of accuracy (e.g. resolution), available details (e.g. time traces) or just feasibility (e.g. scanning a whole library of compounds), modelling and simulation methods can provide insights. In this section, we will deal with Molecular Dynamics (MD) simulations. In there, particles are treated classically and their time trajectory is obtained by solving Newton's equation of motion. MD simulations can be placed in the regime where quantum-mechanical calculations are not feasible any more (due to the system complexity and the required computing times), but molecular details are still needed.

The first MD simulations can be dated back to the 1950s; for example, with Fermi reporting in 1955 that a computer (MANIAC I at Los Alamos) solved the equations of motion for a many body system containing 64 particles subjected to non-linear forces in one dimension. Hard sphere collisions were published in 1957 by Alder and Wainwright using MD [60, 61], and in 1964 Rahman simulated the first Lennard-Jones gas [62]. They have paved the ground for a whole field that grew continuously over time. MD is widely used both in biological [63] and material sciences [64]. There are also hybrid methods, for example combining a small region that is more accurately treated in a quantum-mechanical (QM) description with a larger classical region (MM) around – such hybrid methods were pioneered by Warshel, Levitt and Karplus [65, 66]. Similarly, to overcome the limitation that chemical reactions cannot be accounted for, other hybrid methods like the Empirical Valence Bond (EVB) model by Warshel et al. can

be used [67]. Levitt, Warshel and Karplus together received the Nobel Prize in Chemistry in 2013 [68] "for the development of multiscale models for complex chemical systems". Furthermore, a whole field exists that coarse-grains atoms to enhance sampling, in the trade-off between accuracy and computational cost [69]. A prominent representative is the MARTINI force field [70], that on average maps four heavy atoms onto a pre-defined set of beads. However, probably still the most common approach and also a major focus of this thesis are all-atom MD simulations, which we introduce in more detail below. After decades of developments, their power has been proven in a wide range of applications [71]. Using special purpose codes like GENESIS or machines like ANTON, system sizes of up to a few hundred million atoms [72] or time scales of up to milliseconds [73] are reachable, respectively. Beside these efforts to push the limits, standard high-performance computing clusters enable simulations that are currently typically in the order of million atoms in size and micro-seconds in time.

## 2.2.1   Molecular Dynamics simulations

For a molecular system, we ideally start the description at the time-dependent Schröder equation. As electrons are much lighter and, hence, much faster, we can assume that they follow the nuclei instantaneously. This is known as the Born-Oppenheimer approximation [74] and allows to decouple the wave functions of electrons and nuclei. Taking the expectation value of the nucleic wave functions and considering that they obey the Ehrenfest theorem for their time evolution, we are back to classical mechanics describing the atoms' movements.

Hence, to study the dynamics of the molecule of interest, in MD simulations Newton's equations of motion are numerically solved for each particle (i = 1 ,..., N):

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i = -\frac{\partial V_i}{\partial \mathbf{r}_i} \ . \tag{2.1}$$

The potential energies $V_i$ are a function of the atomic positions $\mathbf{r}_i$ and are given by a set of interactions, termed "force field". Within the force field, different interactions (like bonded and non-bonded) are parametrized, as we will describe below. Decades of effort went into the improvements of force fields, with many tailored towards certain applications. For proteins, Charmm [75] and Amber force fields are nowadays probably most common. For collagen, the amber99sb*-ildnp [76, 77] force field has been proven most reliable and is therefore used throughout this thesis.

There are several algorithms available to solve the arising set of differential equations from (2.1) in discrete time steps $\Delta t$. Commonly used is the so-called "leap-frog" integrator, as is conserves energy during the advance in time. It is mathematically equivalent to the 'Verlet' scheme [78], with the name-giving algorithmic difference that velocities are updated at $t + \frac{\Delta t}{2}$ (instead of at $t + \Delta t$, as it is the case for the positions) to make it computationally more efficient. Hence, on the timeline, updates of positions and velocities leap over each other. Using a Taylor expansion up to the second order on positions and velocity, the resulting equations read:

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \frac{1}{2}\Delta t) \cdot \Delta t \tag{2.2}$$

$$\mathbf{v}_i(t + \frac{1}{2}\Delta t) = \mathbf{v}_i(t - \frac{1}{2}\Delta t) + \frac{\mathbf{F}_i(t)}{m} \cdot \Delta t . \tag{2.3}$$

In here, the time step $\Delta t$ is chosen as large as possible to increase simulation speed, but as small as necessary to capture the fasted movements in the system. Typically, this is given by high-frequency interatomic bond vibrations in the order of femtoseconds. The light hydrogen atoms would have the fasted vibrations, but these vibrations would not influence the trajectory considerably. Because of this, they are often constrained to increase the possible time step from 1 fs to 2 fs. In this thesis, we employ LINCS constraints to achieve this [79].

Several highly optimized software packages are available for MD simulations. We utilize GROMACS [80], an open source software that impresses by its speed and is particularly tailored towards biological systems.

## 2.2.2   Molecular interactions and force fields

In MD, both the largest computational effort during simulations and the most sophisticated calibration of parameters beforehand are hidden in the inconspicuous forces $\mathbf{F_i}$ in eq. (2.1). The contributions to the potential energy can be categorized into bonded and non-bonded interactions. A typical functional form of a force field then reads:

$$V = V_{\text{bonded}} + V_{\text{non-bonded}} \tag{2.4}$$

$$= (V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}}) + (V_{\text{Lennard-Jones}} + V_{\text{Coulomb}}). \tag{2.5}$$

In this example taken from GROMACS [81], the bonded interactions are split into 2-body (bonds), 3-body (angles) and 4-body (dihedrals) terms. A standard choice is to

use harmonic potentials for bonds and angles. Dihedrals are split again into improper dihedrals, which avoid out-of-plane bending via harmonic potentials (e.g. of aromatic groups), and proper dihedrals with periodic torsion potentials with periodicity $n$. In total, we have [81]:

$$V_{\text{bonded}} = \sum_{\text{bonds (i,j)}} k_{ij}^b (b_{ij} - b_{ij}^0)^2 \tag{2.6}$$

$$+ \sum_{\text{angles (i,j,k)}} \frac{k_{ijk}^\theta}{2} (\theta_{ijk} - \theta_{ijk}^0)^2 \tag{2.7}$$

$$+ \sum_{\text{proper (i,j,k,l)}} \sum_{n=0}^{N} k_{ijkl}^\Phi (1 + \cos(n\Phi_{ijkl} - \Phi_{ijkl}^0)) \tag{2.8}$$

$$+ \sum_{\text{improper (i,j,k,l)}} \frac{k_{ijkl}^\xi}{2} (\xi_{ijkl} - \xi_{ijkl}^0)^2. \tag{2.9}$$

Here, the various $k$ denote the (spring) constants of the respective potentials. The equilibrium values of the bonds lengths $b_{ij}$, bond angles $\theta_{ijk}$ and twist angles between proper $\Phi_{ijkl}$ and improper $\xi_{ijkl}$ dihedral planes are indicated with superscript $0$ each. All constants have to be provided as parameters of the force fields. Note that from a quantum-mechanical perspective, however, this handy split is arbitrary as there are only electron densities, but no such things as bond or angle potentials.

In this thesis, we will mostly deviate from the default harmonic bond potentials and use Morse potentials instead, see Fig. 2.5. The reason for this is that the infinite barrier of a harmonic potential does not allow bond ruptures. We are, however, interested in the high-force regime where covalent bonds might break and a Morse potential is more accurate. In our case, the bond potential then reads:

$$V_{\text{morse}}(b_{ij}) = E_{\text{dis}}[1 - \exp(-\beta_{ij}(b_{ij} - b_{ij}^0))]^2. \tag{2.10}$$

The Morse potential requires the dissociation energy $E_{\text{dis}}$ as a third parameter in comparison to the standard biparametric Harmonic approximation. This parameter will be explored in more detail in this thesis for the various bond types present in collagen. The width $\beta$ of the potential can be expressed in terms of the harmonic force constant $k_{ij}^b$
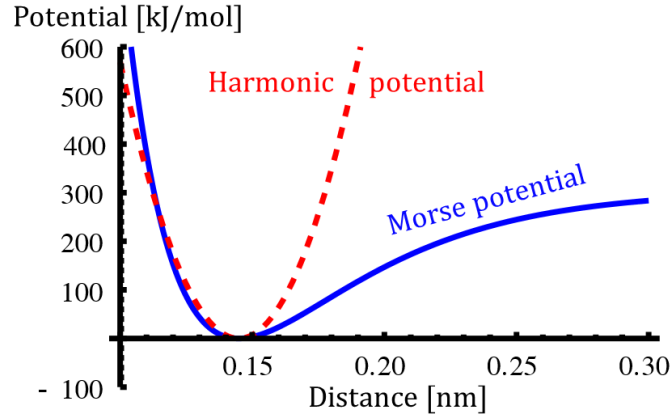
**Figure 2.5: The Morse potential and its harmonic approximation.** In contrast to the Harmonic potential (red), the Morse potential (blue) is more accurate for stretched bonds and has a finite barrier that allows for bond rupturing. Figure from previous work [1].

and, hence, obtained from the default force field parameters:

$$\beta_{ij} = \sqrt{\frac{k_{ij}^b}{2E_{\text{dis}}}}. \tag{2.11}$$

If $|b_{ij} - b_{ij}^0| \ll 1$, it is possible to regain the Harmonic potential by applying a Taylor expansion of $\exp(-x) \approx 1 - x$. This shows that the Harmonic potential is consistent with the Morse potential for small deviations from the equilibrium, as also Fig. 2.5 suggests.

The non-bonded part of the potential energy consists of an electrostatic interaction $V_{\text{coulomb}}$ and a Lennard-Jones term $V_{\text{LJ}}$ that combines the short-ranged Pauli repulsion with the long-ranged van-der-Waals attraction:

$$V_{\text{non-bonded}} = \sum_{\text{pairs (i,j)}} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.12}$$

$$+ \sum_{\text{pairs (i,j)}} \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}}. \tag{2.13}$$

In the first term, $\epsilon_{ij}$ corresponds to the depth of Lennard-Jones potential, and $\sigma_{ij}$ to its width. This van-der-Waals term combines various effects. Its $\sim r^{-6}$ dependency can be derived when considering two interacting dipoles or from London dispersion forces. In contrast, the standard choice of a $r^{-12}$ term for the repulsive Pauli part is motivated by computational efficiency (since it is just $r^{-6}$ squared) and accepted as convention. In the

coulombic term, $q_i, q_j$ denote the charge of the respective particle, $\epsilon_0$ the electric constant and $\epsilon_r$ the relative permittivity.

Both Lennard-Jones and Coulomb interactions arise in principle between all possible atomic pairs. Their complete calculation would, therefore, require large computational efforts with complexity $\mathcal{O}(n^2)$. To avoid this, a different treatment of the short-ranged and long-ranged parts of the interactions makes sense.

For the Lennard-Jones term, the calculation of the long-range part is easily circumvented by setting the Lennard-Jones potential to zero at a defined cut-off, as the potential quickly decays with distance. In general, an interaction needs to decay with at least $\sim \frac{1}{r^3}$ to make sure that only the local environment contributes significantly, which is given here. Still, the computational cost of calculating all pairwise distances remains. To reduce this expense, only interactions within a list of neighbours are calculated, and this list is only updated in certain intervals [78].

For the Coulomb part, once more short-ranged electrostatics are calculated explicitly pairwise only up to a cut-off. In addition, however, long-range electrostatics are calculated with mesh-based methods that sum the charge densities on a discrete lattice. The commonly used Particle Mesh Ewald (PME) scheme has the convenient property that the summation can happen in Fourier space to reach convergence faster [82].

Both the cut-offs, combined with neighbour lists, and the PME then lead to a much more feasible complexity of $\mathcal{O}(n \cdot \log n)$.

## 2.2.3 External (biasing) forces

On top of these naturally arising forces that are parametrized in the force field, we have the option to add external forces acting on individual atoms or atom groups in the MD simulations. In GROMACS, there are two implementations that differ in the way the force is applied: First, in a scheme called constant force, a predefined force is simply added to the defined atoms. We will mainly utilize this option, as it mimics a tensed collagen fibril well. Secondly, in what is known as constant velocity pulling, a virtual spring is added to the atom that is pulled away at a constant rate $v$, such that $x_{\text{spring}} = x_0 + v \cdot t$. Defining a spring constant $k$, the force acting upon that atom is then given by $F = k(x_{\text{atom}} - x_{\text{spring}})$ leading to either an increase in force the farther away the spring gets pulled or to the atoms following the spring and the protein unfolding, if possible. This is used, for instance, for unfolding processes, or to mimic force-probe ex-

periments like Atomic Force Microscopy (AFM) or Optical Tweezers that work similarly, which is also why MD simulations with external forces can be referred to as force-probe MD [83, 84, 29].

Additionally, biasing forces are used to accelerate processes that would otherwise not be observable in MD accessible time scales, giving pulling simulations yet another name: steered MD. Overall, due to the limited computational resources, the forces in MD are often higher than in experiments, to speed up the underlying processes. Recently, however, for the case of streptavidin-biotin unbinding, an overlap between force-probe simulations and high-speed force spectroscopy was achieved that underlines the predictive power of MD simulations [85].

## 2.2.4   Simulation conditions and ensembles

The molecular system of interest is usually placed in a finite box of solvent (water + ions) that mimics physiological conditions. To avoid edge effects of the finite box size, but also to enable the infinite grid that is needed for the aforementioned PME scheme, periodic boundary conditions (PBCs) are usually applied. To this end, an infinite series of virtual copies of the system is placed on each side of the box. If a particle leaves the box on one side, it will virtually enter again on the other side. For PBCs, the box type needs to be space-filling, and the box large enough such that there are no interactions of the protein with its own virtual image, i.e. a minimal distance of the protein to the virtual copies that is larger than the interaction cut-offs needs to be maintained.

Another aspect to consider is the ensemble in which the simulations are conducted. Simply solving Newton's equation in the presented framework would lead to a micro-canonical ensemble (NVE) with particle number $N$, box volume $V$ and system energy $E$ preserved. However, most experiments under laboratory conditions work at fixed temperature $T$ and pressure $P$. To achieve this grand-canonical NPT ensemble in MD simulations, temperature and pressure need to be actively controlled.
For the temperature coupling, different thermostats have been developed. A simple approach known as Berendsen method [86] is to rescale the particles velocities (and, hence, the system temperature) after a time constant $\tau$ proportional to the deviation from the reference temperature $T_0$:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau} \tag{2.14}$$

This results in an exponential decay of the temperature deviation with time, but also suppresses thermal fluctuations. To maintain the correct ensemble properties, a stochastic term can be added [87]. This v-rescale protocol is also employed in this thesis. Similar schemes can be applied to control the pressure by rescaling the box dimensions. Of these, the simpler Berendsen algorithm is usually applied in the equilibration process [86]. It achieves the reference pressure $p_0$ more reliably, but also does not fulfil all ensemble properties. For this reason, we employ the more elaborate Parinello-Rahman barostat [88] during production runs in this thesis.

## 2.3   Kinetic Monte Carlo and KIMMDY

Another key method employed in this thesis is a kinetic Monte Carlo (kMC) scheme applied to invoke covalent bond ruptures. The principle behind kMC is to model the behaviour of a system using a set of "states" that represent different configurations of the system. The system is allowed to evolve by making a transition from one state to another according to transition rates that are determined by the system's properties. In our case, the states would be the collagen system with different bonds broken, and the transition probabilities are the bond rupture rates $k_{ij}$ that we obtained from the Molecular Dynamics (MD) simulations, depending in particular on the bond elongations. For now, we do not consider bond formations, so this process is irreversible. This combination of **Ki**netic **M**onte Carlo and **M**olecular **Dy**namis (KIMMDY) was developed in our previous work [27]. In KIMMDY, we apply the following algorithm of rejection-free kMC:
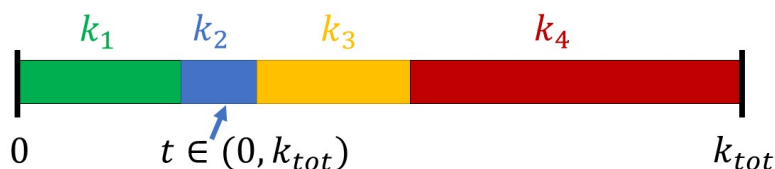


**Figure 2.6: Illustration of Kinetic Monte Carlo.** For this exemplary step with four possible transition states, state 2 is chosen since $k_1 < t < k_1 + k_2$. Figure from previous work [1].

Starting with the current state $i$ and a list of all $j = 1, ..., N$ bonds that could reasonably rupture, we calculate the scission rates $k_{ij}$ (details will be given later on). Next, we draw a random number $t \in (0, k_{\text{tot}})$ from an equal distribution, where $k_{\text{tot}} = \sum_{j=1}^{N} k_{ij}$

is the sum of all rates. The criterion that decides which breakage $f$ occurs is given by $\sum_{j=1}^{f-1} k_{ij} < t < \sum_{j=1}^{f} k_{ij}$. This principle is illustrated in Fig. 2.6. We get to the final state by removing the bond from the system's topology. Lastly, we draw another random number $u \in (0,1)$ to determine the time step from the distribution that corresponds to the chosen transition, using $\Delta t = \frac{1}{k_{if}} \ln(\frac{1}{u})$. This is the time that the system has evolved due to the kMC step.
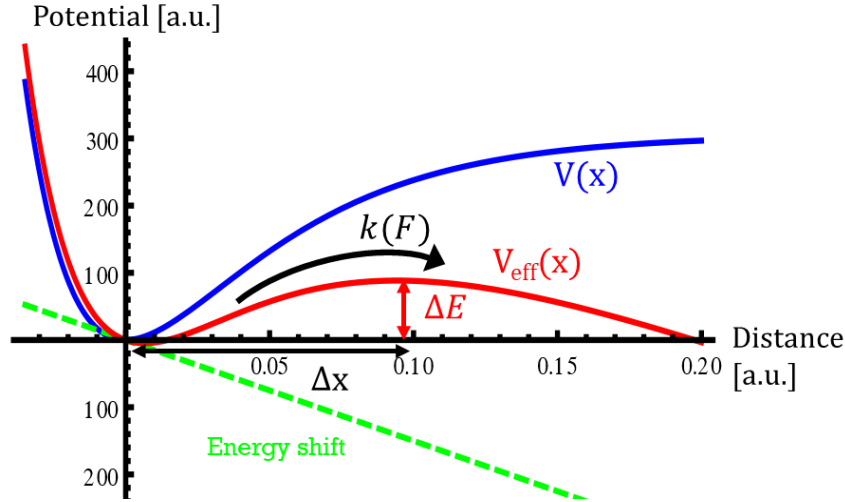


**Figure 2.7: The Bell model yields an effective potential barrier under load.** The original (Morse) potential in blue is tilted due to work done by the external constant force (green) leading to an effective potential (red). The transition rate k(F) is calculated based on the effective energy landscape. Figure from previous work [1].

A decisive step here is the calculation of bond rupture rates based on their elongation in the MD simulations. Once more, we outline the key steps of the previously established protocol [1, 27]. As stated above, we employ Morse potentials for the bond interactions. Starting from that, we use a theory introduced by Bell [22] and later on improved by Evans and Ritchie [23] to describe how the applied force effects the bond potential. The key point for our model is that the external force leads to a shift in the potential energy landscape (see Fig. 2.7). The additional energy term, which can be interpreted as the work that has been performed by the external force against the bond, and the original potential are added up to a new effective energy potential:

$$V_{\text{eff}}(x) = V(x) - F_{\text{bond}} \cdot x \tag{2.15}$$
$$= V(x) + V'(x_{\text{av}}) \cdot x \tag{2.16}$$

It is the $F_{\text{bond}}$ term where the dynamic situation of the MD simulation comes into play: We use the average bond elongation $x_{\text{av}}$ to calculate the internal bond force as the gradient of the Morse potential at the new equilibrium position. Note that for convenience we defined the minimum of the potential to be at zero here, while in the simulation this is of course at the equilibrium bond length.

The barrier height is then given by:

$$\Delta E = V_{\text{eff}}(x_{\text{max}}) - V_{\text{eff}}(x_{\text{min}}). \tag{2.17}$$

To get to a rate, we apply Kramers theory [89]. It coincides with earlier results by Arrhenius [90, 91] and states that for chemical reactions the empirical rate constant is given by:

$$k = A \cdot \exp\left(\frac{-\Delta E}{k_b T}\right). \tag{2.18}$$

The pre-factor $A$, also known as attempt frequency, was in our case obtained by adjusting the theoretical predictions to bond rupture events in MD simulations as described earlier [27]. Note that in the formulation of $\Delta E$, the reaction distance $\Delta x = x_{\text{max}} - x_{\text{min}}$ between the minimum and maximum of the potential can be calculated as:

$$x_{\text{min/max}} = x_0 - \frac{1}{\beta} \ln\left(\frac{\beta \cdot E_{\text{dis}} \pm \sqrt{\beta^2 E_{\text{dis}}^2 - 2E_{\text{dis}}\beta F_{\text{bond}}}}{2\beta E_{\text{dis}}}\right). \tag{2.19}$$

We like to report that in the KIMMDY publication [27], there is a typo in this formula concerning shifted brackets of the square-root (but not in the code base, fortunately). This can be seen when testing the case $F = 0$. The reaction distance decreases with higher applied forces, as the positions of $x_{\text{min}}$ and $x_{\text{max}}$ depend on $F_{\text{bond}}$. In contrast, the original proposal by Bell and Evans assumed a constant reaction distance. Our model takes this dependency into account and is, for this reason, consistent with the more elaborate formulation of Dudko, Hummer and Szabo [24, 92].

Overall, KIMMDY combines the thermodynamic strength $E_{\text{dis}}$ of a bond with the dynamic force distribution in the molecular system to obtain a stochastic sampling of bond rupture events.

## 2.4 Theories for larger-scale fibril modelling

Lastly, we turn to theories that are useful for a mesoscopic description of a collagen fiber. They are employed in our ultra coarse-grained model that we developed during the course of this thesis and present in more detail in section 5.3.

### 2.4.1 Extensible worm-like chain models for collagen

Beside the formation of mechanoradicals, another common ground of collagen and polymers is that due to its building block nature, collagen can be viewed as a biopolymer. Many models have been developed in physics to describe polymers as chains of N microscopic building units. Typically, corresponding continuum versions of these discrete formulations exist, which can, for instance, be derived by considering the limits of many but small blocks. The different types of models depend on the way the segments are connected and on the way they can rotate with respect to each other, e.g. in the freely-jointed chain and the freely-rotating chain. For semi-flexible polymers that are rather stiff, and for collagen in particular, the description of collagen as an elastic rod in worm-like chain (WLC) models has been proven very useful. A WLC can be defined by the fact that the next segment points in a similar direction as the previous one. This leads to the fact that the directions of neighbouring segments are correlated up to a characteristic length scale, the persistence length $l_p$. It is a measure of the polymer stiffness and can be defined as the decay constant of the expected correlation of the angle $\theta(s)$ along the contour $s$ :

$$\langle \cos(\theta(s)) \rangle = \exp(-s/l_p) \tag{2.20}$$

or via the end-to-end distance $R$ in relation to the contour length $L_c$:

$$\langle R^2 \rangle = 2l_p L_c [1 - \frac{l_p}{L_c}(1 - \exp(-L_c/l_p))]. \tag{2.21}$$

Note that sometimes instead of the contour length $L_c$ a relative contour factor $c_f$ is used to describe the system; with $L_c = c_f \cdot L_0$ it yields the maximal available strain in relation to the equilibrium length $L_0$. We use this dimensionless description when comparing models of different lengths in section 5.3.

Another possibility to calculate the persistence length $l_p$ is to consider the WLC under force when stretching a polymer in (without loss of generality) the z-direction. When

the chain extends, the number of possible configurations and, hence, the entropy of the system decreases. The resulting counteracting entropic force can be calculated from the total (bending plus pulling) energy. There is no exact solution for the force-extension relation, but several interpolation formulae exist. Probably most known is the version from Marko and Siggia [93, 94]:

$$F(z) = \frac{k_B T}{l_p} \left[ \frac{1}{4} \left( 1 - \frac{z}{L_c} \right)^{-2} + \frac{z}{L_c} - \frac{1}{4} \right].$$
(2.22)

They investigated DNA, which is the one of the most famous examples for the application of the WLC to a biopolymer, measured with atomic force microscopy or optical tweezers [93, 95, 96].

Turning back to collagen, the WLC model has also been applied to a range of experimental data, obtained by various techniques from rheology to viscometry to AFM to optical tweezers, as reviewed in [97, 98]. These reviews point out that the measured persistence lengths range from 5 nm to 167 nm and vary depending on the utilized method, as well as on other factors like ion concentration.

In addition to these uncertainties for the experimental values, the force-extension formula in eq. 2.22 by definition only covers entropic contributions, but in the high-force regime collagen also exhibits structural changes like (un-)twisting and enthalpic stretching of internal degrees of freedom, like molecular bonds [99]. As we are mostly interested in this high-force regime where occasional covalent bond rupture is possible, more sophisticated models that add an enthalpic contribution (usually an extra harmonic term with spring constant $K_0$) are more applicable. One early approximation formula for the extension including the high force regime was derived by Odijk [100]:

$$z = L_c \left[ 1 - \frac{1}{2} \left( \frac{k_B T}{F l_p} \right)^{0.5} + \frac{F}{K_0} \right].$$
(2.23)

We refer to this as extensible worm-like chain (eWLC) model. As noted in an article reviewing this and other more advanced WLC models, the eWLC description is comparatively accurate in bridging between the entropic lower force regime and the enthalpic stretching, but has the disadvantage that it cannot be inverted to a form that could be used in simulations to obtain forces at given extensions [101]. To circumvent this, the authors of that study suggest a piece-wise worm-like chain (pwWLC), defining the two

regimes separately:

$$F(z) = \begin{cases} F_s(z) = \frac{k_B T}{4 l_p}\left(1 - \frac{z}{L_c}\right)^{-2}, & \text{if } z \leq z^* \\ F_e(z) = \frac{K_0}{L_0}(z - z^*) + F_s(z^*), & \text{if } z > z^*. \end{cases} \tag{2.24}$$

In there, the transition happens at the inflection point of the eWLC, given by:

$$z^* = L_0\left[1 - \frac{1}{2}\left(\frac{k_B T}{l_p F^*}\right)^{0.5}\right] \text{ with} \tag{2.25}$$

$$F^* = \frac{1}{4}\left(\frac{k_B T K_0^2}{l_p}\right)^{1/3}. \tag{2.26}$$

This pwWLC provides a computationally accessible formulation at the cost of a kink at the transition point and a worse approximation in the intermediate regime. We have, in the course of this thesis, tested the various models against our own all-atom MD data and present the results in section 5.3.5.

## 2.4.2 Fiber bundle models
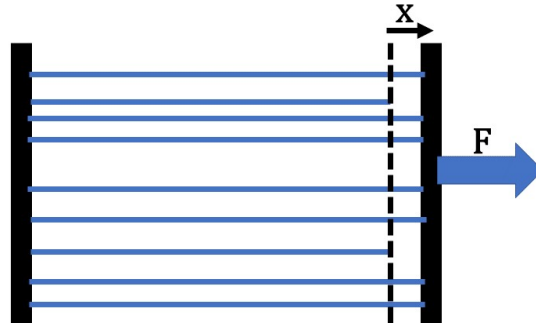


**Figure 2.8: Fiber Bundle Model.** $N$ elastic fibers are connected with a (rigid) platform. While one side is fixed, a force $F$ is applied to the other side, leading to an elongation of the fibers (indicated by the dashed line). The length of the fibers increases until the individual threshold of a fiber is reached and it fails. The load will be redistributed among the remaining fibers, which might lead to a successive failure of fibers that cannot hold the additional load. Figure from previous work [1].

Having a collagen fiber in mind, there is also some apparent resemblance to so-called Fiber Bundle Models (FBMs). The FBM is a widely-used framework for studying the mechanical behaviour of composite materials [32, 102]. It is based on the idea that

the material can be represented as a bundle of fibers, each of which behaves like an individual elastic rod. In this model, the fibers are assumed to be perfectly aligned, and the mechanical response of the composite material is determined by the interactions between the fibers, see Fig. 2.8 for a simple case where the force gets equally distributed among the fibers. In contrast to MD, the models are usually not dynamic as the force is just stepwise increased, and the fibers have an individual (randomly) chosen stress threshold $\sigma_c$ they can sustain. In a comprehensive review by Pradhan et al. [102], also models that integrate thermal fluctuations are introduced as a special case of the FBM. There the maximal sustained stress $\sigma_i$ also depends on an added noise term $N(k_B T)$:

$$\sigma_i = \sigma_c + N(k_B T). \tag{2.27}$$

In this case, the failure time $\tau \sim e^{f/k_B T}$ depends exponentially on the thermal noise for an applied constant stress f, which yields a similar behaviour as the Kramers rate theory discussed above in eq. (2.18). Still, as there are neither interactions nor connections between the fibers in this parallel setup (so no crosslinks), it is useful in yielding a basic framework, but the need for more sophisticated three-dimensional collagen network models for our purpose remains.

## 2.4.3  Smoluchowski drift-diffusion equation

In order to combine the relevant interactions in collagen, a particle-based approach appears as a natural choice: Somewhat similar to our atomistic description, but on a much coarser level with the abstraction of collagen backbones and crosslinks into continuum physics-based potentials as the WLC that we discuss above.

In a mesoscopic particle-based description, a common approach is to combine the effects of drift (directional forces, e.g. from external potentials acting on the particles) and diffusion (random forces, e.g. from thermal noise). This can generally be done with the Fokker-Planck equation that yields the time evolution of the probability density of such an object. We are interested in the special case of particles upon which external forces F(x) and thermal fluctuations with a constant diffusion coefficient $D$ act. This is also known as the Smoluchowski equation (or as Kolmogorov equation, as well):

$$\partial_t P(r,t|r_0,t_0) = \partial_x \left[ -\beta F(r) + D \partial_x \right] P(r,t|r_0,t_0). \tag{2.28}$$

We later-on consider a three-dimensional system, but with forces $F(x)$ acting only along

the collagen fibril axis. For this purpose, we can integrate the one-dimensional Langevin equation, with fluctuations $\xi(t)$ having an amplitude $\sigma$:

$$m\ddot{x} = -\gamma\dot{x} + F(x) + \sigma\xi(t). \tag{2.29}$$

In the overdamped limit of strong friction, we have $|\gamma\dot{x}| \gg |m\ddot{x}|$ and therefore:

$$\gamma\dot{x} = F(x) + \sigma\xi(t). \tag{2.30}$$

Assuming $\xi(t)$ to be white Gaussian noise $N$ with the amplitude being related to the temperature via $\sigma = \sqrt{2\gamma k_B T}$, and making use of the Einstein relation $D = k_B T/\gamma$, we obtain:

$$dx = \sqrt{2Ddt} \cdot N + D\beta F dt. \tag{2.31}$$

This is the core equation governing the dynamics of our developed mesoscopic model that complements the MD and KIMMDY simulations (presented in section 5.3).

# CHAPTER 3

# Advancing Atomistic Models and Simulations of Collagen

As we have seen, collagen is a hierarchical protein spanning multiple length scales. This poses a particular challenge for computational research, given that we are interested in making very accurate predictions at the molecular level, while the mechanical response depends on the force distribution across many structural scales. The collagen fibers are simply too large to be exclusively simulated in all-atom MD. Further, classical MD simulations do not incorporate reactions like bond scissions. It is apparent, then, that only a combination of methods can feasibly handle this spread of tasks. Even if it was possible, however, and despite the importance of collagen research overall, no atomistic model at the fibril level was available before.

In this chapter, we present our method improvements with regard to (hybrid) all-atom simulations of collagen failure. Note that, as we introduce the newly built mesoscopic models in detail in chapter 5, we here start with the molecular level. We shortly describe the atomistic collagen fibril models built by others in our group, as well as how we validated them and obtained the necessary force field parametrizations. Then, we discuss the incorporation of bond dissociation energies into our previously developed simulation scheme KIMMDY (Kinetic Monte Carlo / Molecular Dynamics) [27] that is used to enable bond scissions on all-level without prior knowledge about the reaction sites. In this section, we also revisit how KIMMDY calculates bond forces, for which a detailed investigation was carried during this thesis as prerequisite for obtaining the final MD and KIMMDY results presented in chapter 4. Further, we discuss our preliminary work on a proof of concept for hydrogen atom transfer (HAT) reactions and see how KIMMDY can be used for bond ruptures in DNA as well. Lastly, to examine on a larger

scale how forces distribute through collagen and lead to bond rupture, we set out to build a 45 million all-atom model of a tensed 335 nm collagen fibril, investigating for the first time full 300 nm collagen molecules at this level of detail. Through an ongoing collaboration with the RIKEN institute in Japan, we are running simulations of this model on FUGAKU – currently the second-largest supercomputer in the world.

## 3.1 ColBuilder: a webserver for collagen models

The first step of any molecular simulation study is to build (or have at hand) an atomistic structure that serves as initial configuration. For collagen, the best available experimental data is a 5.4 Å resolution fiber diffraction electron density from 2006 [21], and on the atomistic level only short idealized triple helices were available before our group ventured into the collagen research. Here we briefly describe the modelling process (done by Dr. Agnieszka Obarska-Kosinska) and our validation of the models using MD simulations and, for this, adapt parts of texts and figures of our published work [34]. To make the final models available for a broad community, our group created Col-Builder [34], a web resource that provides collagen models for 20 different species with a large variety of crosslink types and localizations within the fibril. This resource aims to facilitate structure-based analyses and simulations of type I collagen in health and disease mutants research. To this end, it comes along with parameters of the modelled crosslinks for an Amber force field and is available at https://colbuilder.h-its.org.

As the models rely on low resolution data, it has uncertainties in particular with respect to crosslink positions and to the more disordered telopeptide region at the end of the collagen molecules. Expanding on our work reported in [34], we here also discuss alternative modelling options with respect to crosslinking. These choices of the fibrillar topology can be relevant for the mechanical response on a larger length scale, which we investigate in chapter 5.

### 3.1.1 Building atomistic collagen models

A colleague in our group used the integrative modelling approach that is depicted in Fig. 3.1 to build our atomistic microfibrils. It starts from the best resolution structure that is available of collagen type I by Orgel et al. from 2006 [21]. This data includes the three-dimensional packing of collagen molecules and suggests at which positions crosslinks might be possible. Unfortunately, the resolution of this X-ray fiber diffraction data is
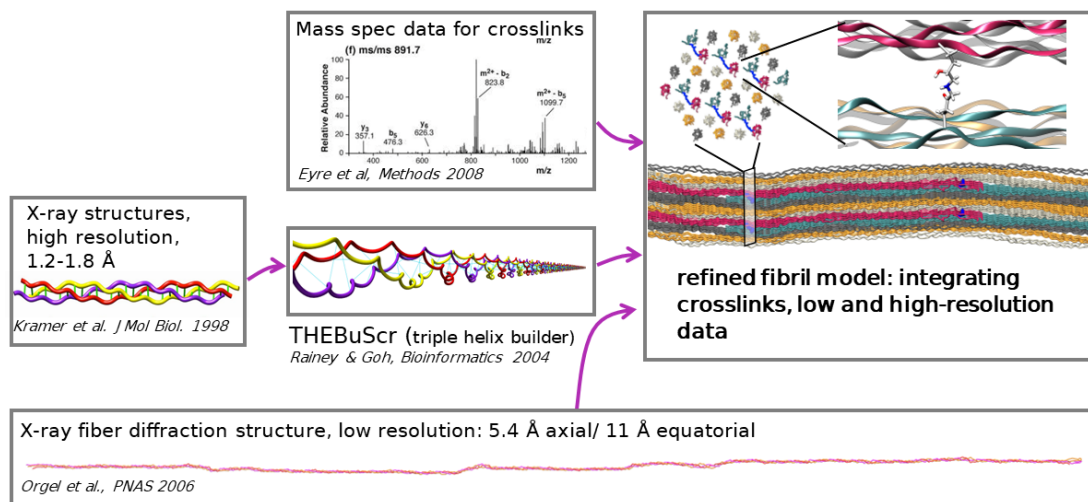
**Figure 3.1: Modelling atomistic collagen models for ColBuilder**. Modelling process as conducted by Dr. Agnieszka Obarska-Kosinska: High resolution data of individual triple helices is combined with the structural X-ray fiber diffraction data. Taking information about crosslinks into account and remodelling the telopeptide regions to accommodate the crosslinks, we arrive at the final refined fibril model.

only 5.4 Å. However, more refined data of shorter (idealized) collagen triple helices is available [103]. This biochemical data was then fitted into the lower-resolution structure. Specifically, the more flexible telopeptide regions were remodelled to accommodate different crosslink types, the potential positions of which are known from mass spectrometry [48]. The first constructed model was a representative species *Rattus Norvegicus*, which is the same species studied in earlier experiments on the rat tail tendons [8]. With homology modelling, models for 19 more species could be obtained. All models, accompanied by Amber force field parameters of the crosslinks, are available online at ColBuilder [34].

## 3.1.2   Validating our collagen models

We validated the models by comparing MD simulations to additional structural data that was not used for modelling. To this end, we conducted MD simulations of a representative models set: one model with divalent and one model with trivalent crosslinks for each one of three species (*Rattus Novergicus*, *Pongo Abelii* and *Loxodonta Africana*). We simulated duplets for these 6 models, leading to 12 independent simulations of at least 100 ns equilibrium MD each.
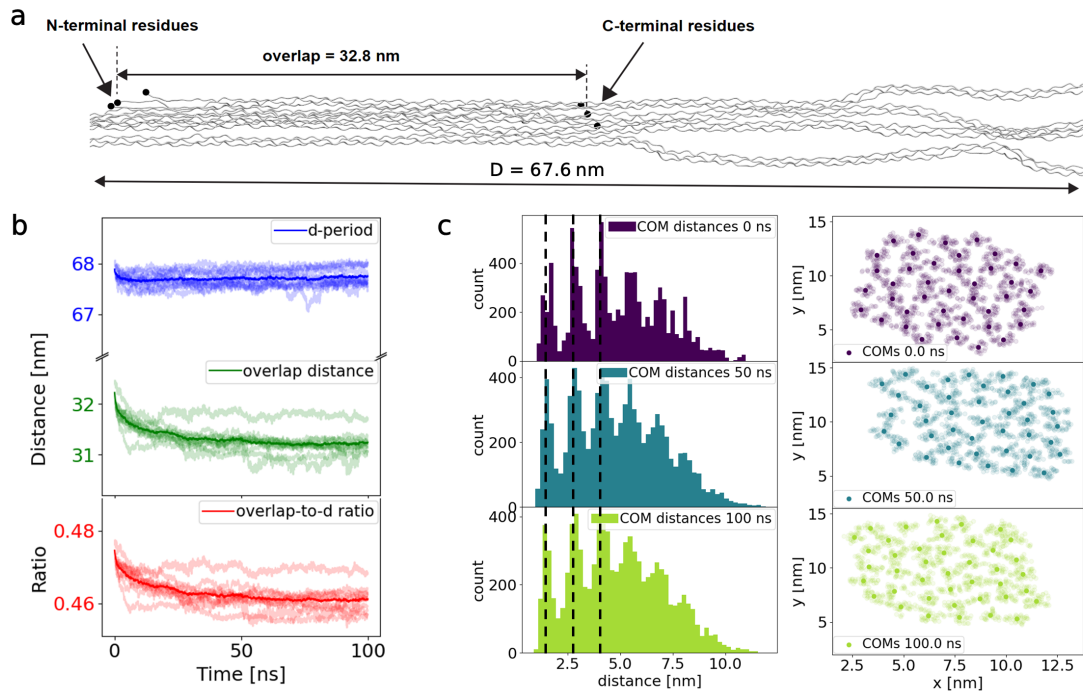
**Figure 3.2: Comparison of our simulated model configurations to experimental structural data**. **a** Overlap/gap ratio as given by the length of overlap region and overall D-period is at most, taking full telopeptide region into account, 0.485/0.515. **b** Gap and overlap lengths for the complete set of our 12 simulations: Starting from a configuration as shown in panel a, the overlap distances and, hence, the average ratio decreases to about 0.465/0.535 (bold line: average, transparent lines: individual replica) **c** All-to-all distances between the center-of-masses of the triple helix (a cross-sectional slice of the overlap region thereof) at 0 ns, 50 ns and 100 ns simulation time. The counts in the histograms are summed up across all our models, whereas an exemplary slice for one individual replica (*Rattus Norvegicus*, HLKNL crosslink) is shown on the right. The dashed lines in the histogram indicate the average distance to nearest, next, and over-next nearest neighbours. The figure and its caption are adapted from our own publication [34].

Figure 3.2 shows the comparison of the obtained simulated configurations with different experimentally accessible properties of collagen. The characteristic banding pattern (D-period), consisting of a gap and an overlap region as explained in section 2.1, is a well-studied structural feature of collagen. At first, our modelling as depicted in Fig. 3.2a leads to a 32.8 nm long overlap region, which amounts to 0.485 D given that we have a 67.6 nm long D-period. The telopeptide conformation is particularly extended

at the beginning due to the accommodation of the crosslinks. Over the course of our simulations, as depicted in panel b, the overlap distance shrinks and, therefore, also the overlap-to-D ratio. The resulting average of $\approx 0.465$ is in line with previous experimental estimates that range between 0.40–0.48 D [46, 104, 105, 106, 107].

In addition to this feature in axial fibril direction, we also investigated the cross-sectional packing of the collagen fibrils. On the left column of Fig. 3.2c, we display the all-to-all distances between the center of masses (COMs) of the triple helices from a slice of the middle of the system, at different simulation times. Hence, the histogram shows the count for binned distances between COMs of all molecules at one fixed cross-sectional position, aggregated for all 12 models. One example of these 12 slices is shown on the right column, at the respective simulation times. We observe that the average distances of nearest and next-nearest neighbours (dashed lines) stay approximately constant over time and, importantly, in line with the original position inferred from the experimental X-ray data that was used for modelling. Taken together, our atomistic simulations explore a conformational space that is close to experimental data both in lateral and axial fibril direction.

### 3.1.3  Force field parametrization of non-standard residues

We have seen in section 2.1.2 that collagen features different crosslink chemistries. These special lysine-derived residues are not part of standard force fields, which is why it was necessary to parametrize them for our MD simulations. The parameters for the crosslinks were obtained by Dr. Obarska-Kosinska in the context of ColBuilder [34] and can be found in the Supplementary Information of that article. In this thesis, we encountered the same situation for dihydroxy-phenylalanine (DOPA), which is a post-translational modification that we found to scavenge the arising mechanoradicals in collagen [8, 9]. It resembles tyrosine but contains an extra OH-group that also influences nearby atoms and their partial charges, see Fig. 3.3, and was to our knowledge not available in any Amber force field. For a later comparison with the Charmm36 force field [75], we additionally parametrized hydroxyproline (HYP), which is a common residue in the collagen triple helix. Previously, it was only available in the Amber and not the Charmm force field. We adopted the procedure for DOPA and HYP that Dr. Obarska-Kosinska used for the crosslink parameters. With this, we complemented our Amber and Charmm force fields with them, respectively, during the course of this thesis.
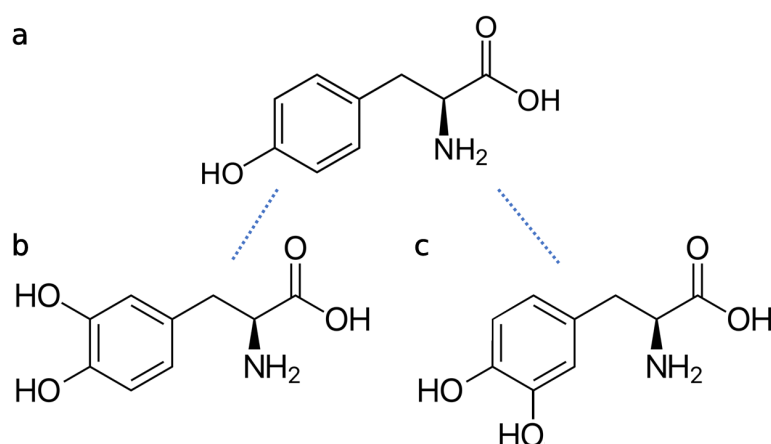
**Figure 3.3: Parametrizing DOPA in our atomistic force fields, starting from tyrosine**. **a** Structure of tryrosine. **b** Structure of DOPA, with the extra OH group added to the CE1 carbon. **c** Structure of DOPA, with the extra OH group added to the CE2 carbon. As the ring can rotate, this is chemically equivalent to panel b, but requires a own residue in the force field due to the distinct atom naming.

First, we conducted QM calculations using Gaussian09 [108] and then used Antechamber [109] to convert the charge densities to partial charges that are assigned to the individual atoms as necessary for a classical force field. Lastly, with the help of ACPYPE [110], we converted them to the format of the force field. For the partial charges, we kept the original backbone values of the force field and only took the newly obtained values for the side chain, as the backbone is identically parametrized across amino acids and maintaining force field self-consistency is essential. However, this can result in a small off-set in the total charge of the residue, with the sum of backbone and crosslinks not adding up to an integer, which is why we also re-distributed this small deviation uniformly on the side chain. Additionally, for DOPA we had to make a choice on where to put the extra hydroxy group; we opted for making the CE2-carbon the default, but also provide the CE2-carbon as an alternative in our force field, compare Fig. 3.3. There is no chemical difference between them, as the ring can rotate, but due to the distinct atom naming in the force field they are different residues from a purely technical perspective. Both parameter sets for DOPA and HYP are now part of the force fields that are used within our group and might be made publicly available with the next update of ColBreaker, which is supposed to be a further evolving web resource for collagen models. We have compiled the obtained parameters in the appendix, tables A.2 to A.5, including the definition of bonded interactions as introduced in section 2.2.1.

## 3.1.4   Model quality and alternatives

We have shown above that the developed models are consistent with current experimental data. It is important to note, however, that a few choices had to be made during the modelling process that were best guesses with different options available. Overall, the underlying trace of $C_\alpha$-atoms from the fiber diffraction data has only been solved at a resolution of 5.4 Å. Especially in the more disordered telopeptide regions, the individual helices cannot be clearly identified. In our group, we opted for accommodating the triple helices such that the divalent and trivalent crosslinks could be included at their positions known from mass spectrometry.
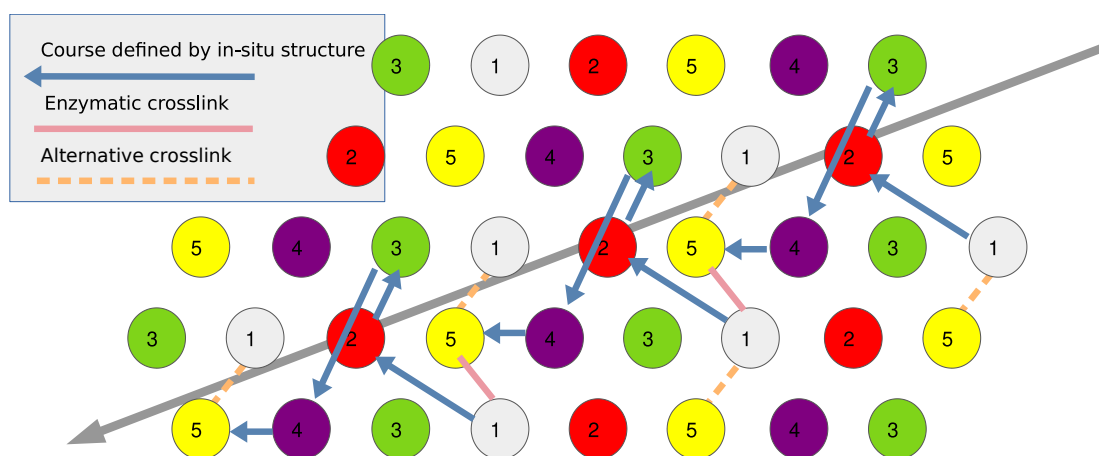


**Figure 3.4: Potential crosslink configurations and collagen braiding along the fibril.**
The different phases (1-5) of D-periods of a collagen triple helix are shown in the (quasi-) hexagonal cross-section. The fibril continues along the large grey arrow, giving this 3D-to-2D projection its directionality. An exemplary course for a triple helix can be followed along the blue arrows. As previously suggested [21], the crosslink pattern (in pink) leads to a continuation of the fibril, even if a triple helix ends in phase 5. However, this would create unconnected layers that might be able to slide against each other. An alternative/additional crosslinking site (orange) could prevent this, at the cost of not having locally fully connected molecules any more. Figure adapted from [21, 111].

Furthermore, for the braiding of the triple helices and the directionality of the crosslinks connections, we kept the scheme that can be found in the Supplementary Information of the article of Orgel et al. [21] for our atomistic models. The triple helices exchange lateral positions on the (quasi-)hexagonal grid in each gap region, resulting in 5 phases (for each gap/overlap region) that a collagen molecule goes through. Enzymatic crosslinks

connect the molecules at their start and end, so phase 1 triple helices with phase 5. In Fig. 3.4, both the implemented (according to Orgel et al. [21]) and an alternative option for the crosslinking in a projected cross-sectional view is shown. In a Master's project that was partially supervised during this thesis, it was found that, for random alteration between the crosslink configurations, the fibrillar network is locally not fully connected any more [111]. Only on a larger length scale, this might provide a stable configuration, with the advantage that a more heterogeneous network avoids unconnected layers that could potentially slide along each other.

Finally, we would like to shed more light onto the question of chain registration of the two $\alpha(1)$- and one $\alpha(2)$- chains in the triple helices (that differ in sequence), which has been discussed in the literature for several decades already [112, 113, 114, 115, 116]. In the heterotrimer that aligns the GLY-X-Y pattern of the collagen sequence, this arrangement question could influence which residues come close to each other. In ColBuilder, we again followed Orgel et al. [21] with an $\alpha(1)$-$\alpha(2)$-$\alpha(1)$ sequence. More recent evidence from Jalan et al. [116], who computationally investigated protein recognition properties of the three possible heterotrimers, suggest an $\alpha(1)$-$\alpha(1)$-$\alpha(2)$ configuration with the $\alpha(2)$-chain in the trailing position to be preferred. Whereas the question of chain registration might not influence the rupture sites and mechanics that we are mostly interested in, we recommend considering different registries in future work.

### 3.1.5 Comparing the model with new experimental data

As collagen is so abundant and of high interest, it also seems apparent that better structural data (almost 20 years after Orgel et al. [21]) could have a high impact on the field. We have seen above that several properties remain at least ambiguous on the fibril level, and the direct identification of crosslinks or a detailed study of the positioning of individual triple helices throughout a fiber are lacking. In our group, a colleague is currently in the process of obtaining high resolution data by cryogenic electron microscopy (cryo-EM) and we facilitate that process by comparisons with our existing models. In cryo-EM, the signal strength correlates with the electron density. A small section of an image of such a fiber from rat tail tendon can be seen in Fig. 3.5a. We observe the characteristic D-band pattern along the vertical fiber axis, in which the smaller but denser overlap regions appear darker. Further, we observe periodic lines in particular at the boundaries, but also in the gap region. For comparison, we can obtain
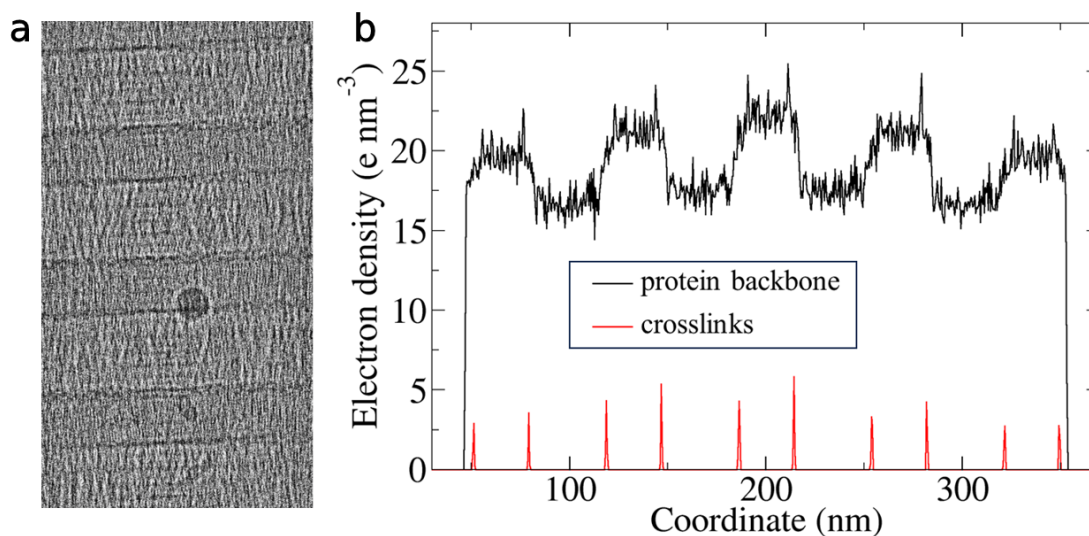
**Figure 3.5: Comparing experimental electron densities of collagen fibers with our all-atom models**. **a** Enlarged view of a cryogenic electron microscopy (cryo-EM) image of a collagen fiber obtained from rat tail tendon, the fiber axis is (almost) vertical here. This picture is background corrected and provided by the courtesy of our colleague Aysecan Ünal. **b** Using our atomistic model, we assume that the electrons are at the atom nuclei, calculate the electron density and show it here binned along the fibril axis. In black, we depicted the resulting density of the backbone (so without side chains) and in red the (divalent) crosslinks. In both panels, the characteristic gap-overlap pattern, but also some smaller peaks can be recognized.

electron densities from MD simulations as well, by assuming that the electrons are at the center of the nuclei. The comparison to our atomistic data can help in two ways here: On the one hand, as one may expect, the experimental data shall in the end lead to more accurate models. On the other hand, however, in the process of validating the experimental procedure, the atomistic details that we already have available might help to explain the origin of certain features in the data. For this reason, we used our 300 nm collagen model (more information on that in section 3.3) and show the backbone and crosslinks electron density in Fig. 3.5b, binned along the fibril axis. The most apparent feature is again the gap and overlap pattern, with an expected density ratio of about 4:5. A closer look reveals that there are also peaks at the end of the overlap region, in particular on the C-terminal side. This can partially be explained by the crosslinks there, but not fully, as then it should be more even on both sides of the region. Further, to exclude that the amino acid sequence is the origin of these peaks, we have only depicted the density of the backbone without side chains. We hypothesize that, as the

C-telopeptide region is more flexible and not triple helical any more, the ending strands fold back. Once the cryo-EM experiments progress, it is worth investigating this in more detail.

## 3.2 KIMMDY: improvements and limitations

Having at hand atomistic models of collagen I, we next focus on the simulation method to investigate rupture during the dynamics of stretched collagen. Our approach to simulate covalent bond ruptures in KIMMDY, developed in previous work [27] and described in section 2.3, was refined and expanded during the course of this thesis. We know of no other existing method that is able to make precise yet computationally feasible predictions of bond rupture in a large protein all-atom system without prior knowledge of reaction sites. The achieved advancements of KIMMDY, but also more recent insights into its limitations, are described in this section.

The existing homolytic bond rupture scheme requires two sorts of main inputs. First, the initial barrier height of the bond (Morse) potential, also known as bond dissociation energy (BDE), needs to be provided. To be specifically tailored for different amino acids and crosslinks, we have updated KIMMDY to enable different values for the same bond types, see section 3.2.1. Second, the initial barrier is then lowered by the force acting upon that bond in a bell-type approach as described before. This force is obtained from the bond elongations in the MD simulations. For this process of force calculation, we have also had a closer look into its limitations in section 3.2.2.

Lastly, beside these improvements of the existing KIMMDY scheme, we have also explored the technical feasibilities of other reactions, namely DNA bond rupture in section 3.2.3 and hydrogen atom transfer in section 3.2.4, a likely follow-up reaction after homolytic bond rupture.

### 3.2.1 Enabling different bond dissociation energies in KIMMDY

In KIMMDY, the first input to the calculation of rupture rates is the relative strength of the chemical bonds. Beside the protein backbone that is stressed under external load, collagen also features the aforementioned enzymatically derived crosslinks connecting the individual triple helices. Due to their position in the network, they carry a heavy load as well, as we have seen previously [8, 27]. The enzymatic crosslinks are derived from lysines, and as a consequence feature single bonds of the C-C and C-N type, as

in the protein backbone. Hence, it is a priori unclear which are the weakest bonds, but the thermodynamic differences in homolytic scission of collagen's backbone and crosslink bonds are decisive for collagen mechanics. During this work, we have updated the KIMMDY code to feature different bond types for the crosslinks such that we can adjust their BDEs individually – in contrast to the standard force field description. More precisely, we introduced new atom names for each crosslink and also for different amino acids, even if they have the same atom types, i.e. had the same (bonded) force field parameters. Utilizing a separate file, we can provide more specific BDEs for both the production MD in GROMACS and the evaluation in KIMMDY thereafter. In order for GROMACS to take up these values, moreover, the atom names in the topology file need to be adjusted accordingly. The updated code was then used to obtain the results that are presented in chapter 4.

## 3.2.2 Force calculation in KIMMDY

The main input from the MD simulation to KIMMDY are the forces that act upon the individual bonds and, measured by their elongations, how much they have weakened the bonds. In the original KIMMDY work [27], we discussed that this one-dimensional reaction coordinate – while reproducing the overall behaviour of our validation system – might have limitations as it neglects other degrees of freedom such as bond angles. Only quantum-chemical calculations would be exact, but as they are prohibitively expensive it is important to understand the applicability (and its limits) of other methods like KIMMDY. In the context of this thesis, we noticed differences in rupture propensities for different amino acids that were difficult to explain. While these differences did not qualitatively change our results, we took this as a starting point for a more detailed investigation about their origin and, potentially, for improvements of KIMMDY in that regard.

To this end, we investigated simple polypeptide chains under force. The naïve expectation here is that, if we assume that amino acids behave like a chain of Hookean springs, the measured bond force on their inside when pulled from the outside should be the same regardless of their molecular structure. However, what we observe in Figure 3.6 is a different picture: Depending on the amino acid type and position, force levels deviated measurably. For two different external pulling levels of 1 nN and 1.5 nN, we first tested 4 configurations of polypeptide chains containing all amino acids in a random order (blue to yellow points in the figure). For the cases with an inconsistent picture of the results, we extended this set with homopeptides of one amino acid type (in red) and, lastly,
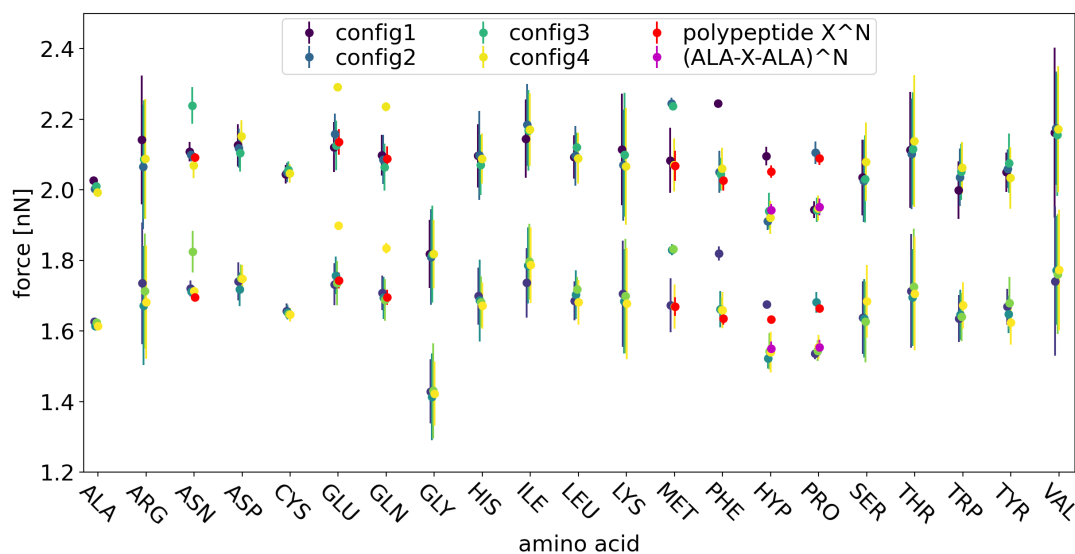
**Figure 3.6: Force levels in different stretched polypeptides**. We created different configurations of polypeptides. For all of them, we conducted force-probe simulations at two force levels of 1 nN and 1.5 nN, respectively, and measured the forces with KIMMDY via the bond elongations. The first set (configs 1-4) contains four random arrangements of all 20 amino acids. For amino acid types with inconsistent results in this first set, we added polypeptide chains that either consist of only one amino acid type (red data), e.g. $PRO^{20}$, or of a sequence of that amino acid with ALA as neutral neighbours in between (violet data), e.g. $(ALA-PRO-ALA)^{10}$. Forces in $C_\alpha$-N and $C_\alpha$-C bonds are averaged and displayed per amino acid type.

with a setup where we altered that amino acid with alanine (ALA) as a more neutral neighbour, e.g. $(ALA-X-ALA)^N$ (in purple). We included these last configurations as we found that depending on the amino acid sequence in the polypeptide, i.e. the neighbouring amino acids, the force levels changed more for these amino acids. Inspecting again Figure 3.6, we see that glycine has the largest deviation from the mean, which is particular of interest as it is the most common amino acid in collagen. Furthermore, throughout all the amino acid types, we can not only observe varying average force levels, but, in particular, also large variances.

These observations point us to the first idea that the side chains might also influence the backbone bonds, e.g. through interactions with neighbours leading to a stretch towards one direction. Hence, one might conclude that there are other interactions beyond the covalent backbone in a simple polypeptide chain, and the assumption of a Hookean

spring does not fully hold. In this sense, such differences would lead to different configurations and resulting stress levels. However, repeating the simulations with the Charmm36 force field [75] instead of our default Amber force field showed yet again different results between the force levels in amino acids. This suggests that at least parts of these differences could be attributed to force field artefacts that should be corrected for.
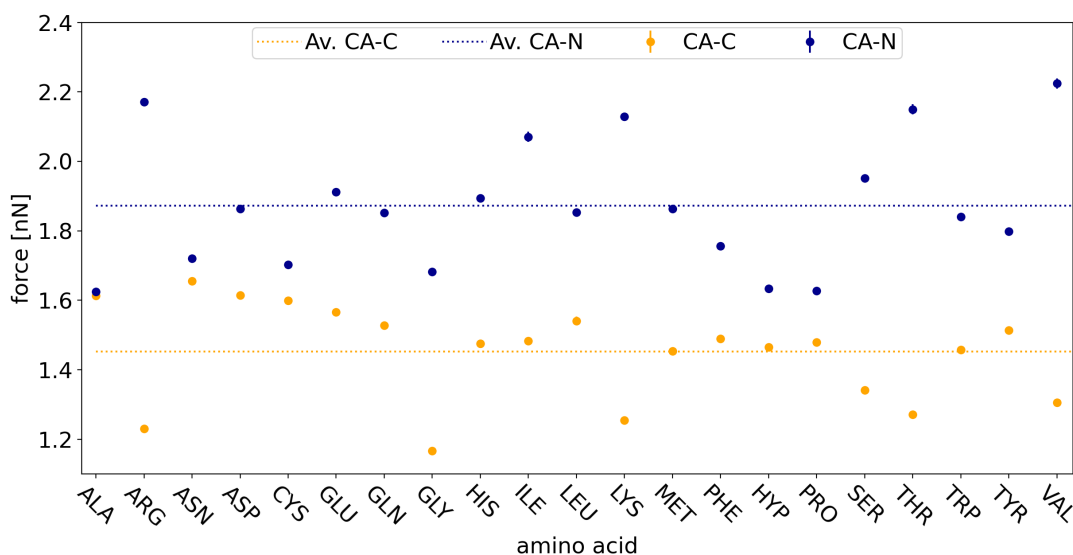


**Figure 3.7: Force levels vary for different bond and amino acid type**. For one polypeptide configuration and the external force level of 1 nN, we show the average force inside the amino acid types, split up by bond type $C_\alpha$-N and $C_\alpha$-C.

Secondly, we had a deeper look at the backbone bond types in Figure 3.7. For clarity, this analysis is only depicted for the force level of 1 nN. The general trend is that KIMMDY considers $C_\alpha$-C to be less loaded than $C_\alpha$-N bonds. Moreover, if there is a large discrepancy between the two types, like in the cases of ARG, ILE, THR and VAL, this leads to these amino acids breaking more often than amino acids with more balanced force levels. Since the bond rupture rates depend exponentially on the forces, the more extreme cases dominate the rupture distribution. We also note that in all cases the measured forces are higher than the external load. Even in equilibrium MD, it was found both here and previously in our group that bonds are on average stretched compared to their equilibrium bond length $r_0$, without external force application. Various effects might contribute to this stretch, starting from repulsive electrostatic interactions over thermal fluctuations to an entropic contribution: If the polypeptide stretches, its configurational

space increases. This interesting topic is, however, outside the scope of this thesis and should not influence the relative breakage propensities in collagen.
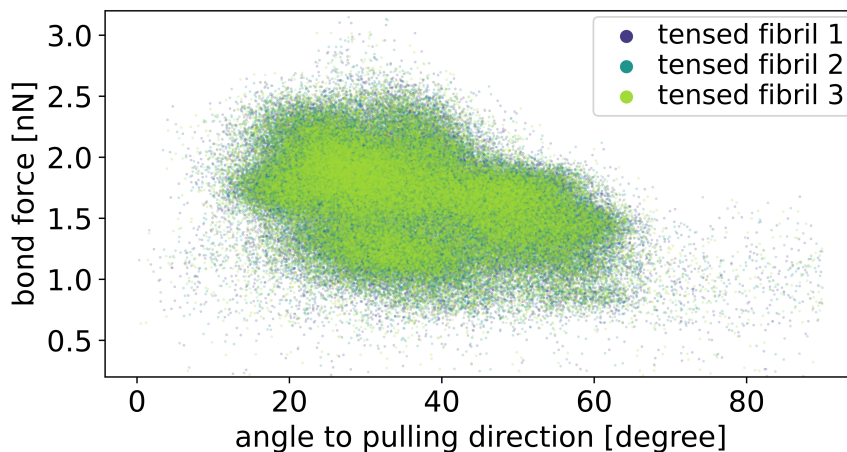


**Figure 3.8: Bond force levels can depend on the bond angle to the pulling direction.** For three different simulations of our collagen microfibril under force, we show the average bond force against the angle of the bond to the fibril axis. Each dot represents one of the about 50,000 bonds that are rupture candidates in the simulation.

A third related consideration are other bonded force field terms (angle, dihedrals) that are neglected in KIMMDY and could be heterogeneous among amino acids. To investigate the impact of angles, we correlated the measured forces with the angle of the bond with respect to the pulling direction in Figure 3.8. As expected, there is a negative correlation showing that more tilted bonds take up less force; a first order explanation would be a simple vector addition of forces, leading to a factor of $\cos(\alpha)$. However, the observed correlation is weaker, presumably as other structural rearrangements are also at work.

Taking all these different aspects together, there is both an error due to the simplifying nature of KIMMDY that only utilizes the bond distances in the atomic force field description (instead of all bonded interactions or, even better, electronic contributions on a QM level) and a real contribution due to interactions within different amino acid types. These two points can hardly be separated, by construction of the protocol. Even though these differences in forces are not large compared to their absolute levels, this error propagates strongly into the rate, as rates depend exponentially on the effective energy barrier. To estimate an upper bound of the effect, we used our polypeptide simulations as a baseline
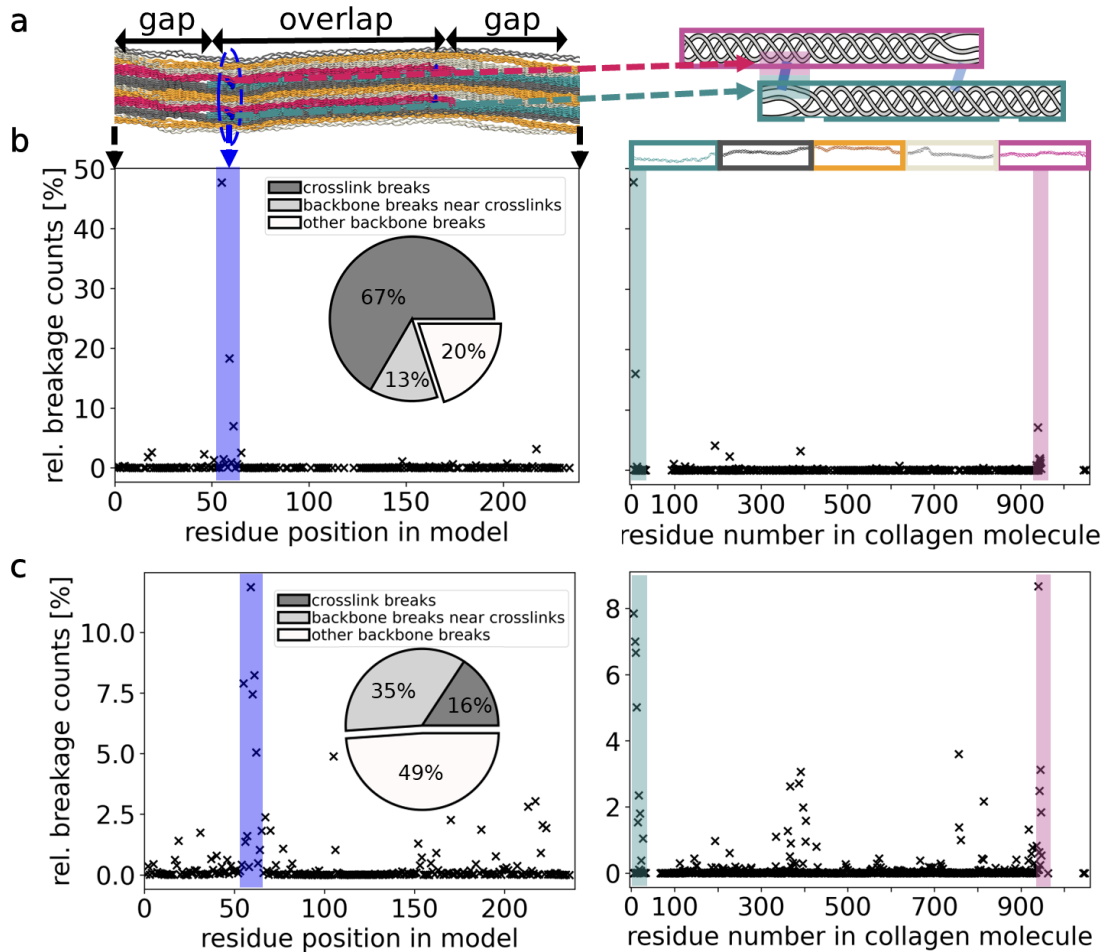
**Figure 3.9: Comparison of breakage type ratios using different methods to calculate bond forces**. This analysis was performed for data of divalent (HLKNL) crosslinks only. **a** Location inside the model (left) and along a collagen triple helix (right) as will be used in b and c. **b** Data for default force calculation. **c** Data using (maximal) correction factor. The concentration happens at the same regions, but is more spread out among different residues (also note the different y-axis). Figure from Supplementary Information of our own publication [38].

and added a new hypothetical force correction factor into KIMMDY. Assuming, for testing, that the Hookean spring picture did hold and that all the observed differences between amino acids were artefacts, we could balance the deviation from the average force level depending on each amino acid type. This correction factor from the simple polypeptide simulations could then be applied to the large fibril MD simulations. As this counters the effect fully, this is probably overcorrecting and, as stated, only used as a reference here. In Fig. 3.9, we compare the breakage points in our collagen model under load: panel b without, and panel c with this hypothetical correction. In the second case, breakages happen in more distinct bonds next to each other, rather than only in a few amino acid types for which the forces are higher than in their neighbours. Whereas there are significant changes on a small spatial scale, already on a mid-size scale this effect seems to be small compared to the effect of the actual force distribution: most breakages still happen in the vicinity of crosslinks. We do note, however, that the ratio of breakages in crosslink vs. backbone bonds also changes (see insets of Fig. 3.9b and c).

Turning towards literature, a previous study found that the force response of polypeptide chains is, in principle, well described by the Amber force field when comparing MD to density functional theory (DFT) calculations [117]. Particularly interesting for our context is their finding that the force-extension curves of different homopeptides depend on the amino acid type. For example, polyglycine gets stretched more into a planar state and, hence, is more extended at the same force level than the other homopeptides. Further, the study corroborates our finding that the interactions with neighbouring amino acids matter for the force-extension curves, by varying length and composition of the polypeptides. This is in line with our analysis above that there are important interactions that lead to deviations from the Hookean spring picture of a polypeptide.

On a bigger picture, the results aiming at the identification of weak regions in collagen are mostly robust against these considerations. We emphasize again that the tested correction factor is only an upper limit, and the actual effect size will be smaller. Besides, this analysis was a nucleation point for future work aiming both to improve KIMMDY and get a better understanding of classical force fields in general. Investigations on the relative force take-up into the different degrees of freedom are ongoing by other group members. They make use of the Judgement of Energy DIstribution (JEDI) method for QM calculations, which decomposes the energy along different coordinate types [118]. The results in this section, while not fully conclusive yet, are likely still relevant in the more general context of understanding and improving atomistic force fields.

### 3.2.3   Ultrasound induced DNA rupture in KIMMDY

In this thesis, we focus on studying ruptures in tensed collagen using our various simulations schemes. However, KIMMDY has been developed with a general applicability in mind and received attention from a total different field: breaking DNA molecules. An ongoing collaboration within the Max Planck School Matter-to-Life, in the course of which we supervised an intern student, aims at identifying propensities of bond ruptures in nicked DNA. For this purpose, DNA models in which one of the two strands has a nick (a missing bond) have been designed and investigated with force-probe MD simulations. This is to guide the experimental design of DNA strands. With the application of ultrasound, a selective bond rupture can be achieved at the nick as predetermined breaking point. Promising preliminary results have been achieved in this project. As this goes outside the scope of this thesis, however, and the main work (aside from guidance and supervision) was done by other people, we mention this only as further evidence for the relevance of advancing reactive simulations at the atomistic level with means like KIMMDY. Moreover, the results from this collaboration might be useful for method validation from our side: Detecting rupture points in DNA is experimentally easier than in collagen. Hence, we could use the obtained rupture distributions to compare our predictions with experimental data more directly, once they are available.

### 3.2.4   Proof of concept simulation of hydrogen atom transfer

Having homolytic bond ruptures in the system, hydrogen atom transfer (HAT) is a natural follow-up reaction as radicals are highly reactive and will migrate through the system. More mechanistically speaking, a hydrogen atom moves to the place of the radical. As the last part of this section about KIMMDY improvements, we show our proof of concept work on how HAT can be implemented in GROMACS. This work was done during the course of this thesis, and subsequently taken up by other group members, with still occasional contributions, mainly through discussions, from our side. It involves many further technical challenges, and is still ongoing.

As a small toy system, we used ALA-DOPA-ALA, into which we induced a bond rupture with KIMMDY and then implemented subsequent hydrogen atom transfers. In principle, these reactions can be incorporated naturally in the Kinetic Monte Carlo step of KIMMDY, given that we can obtain reaction rates — a benefit that we always had in mind when designing KIMMDY. In the Monte Carlo step, we envision the different
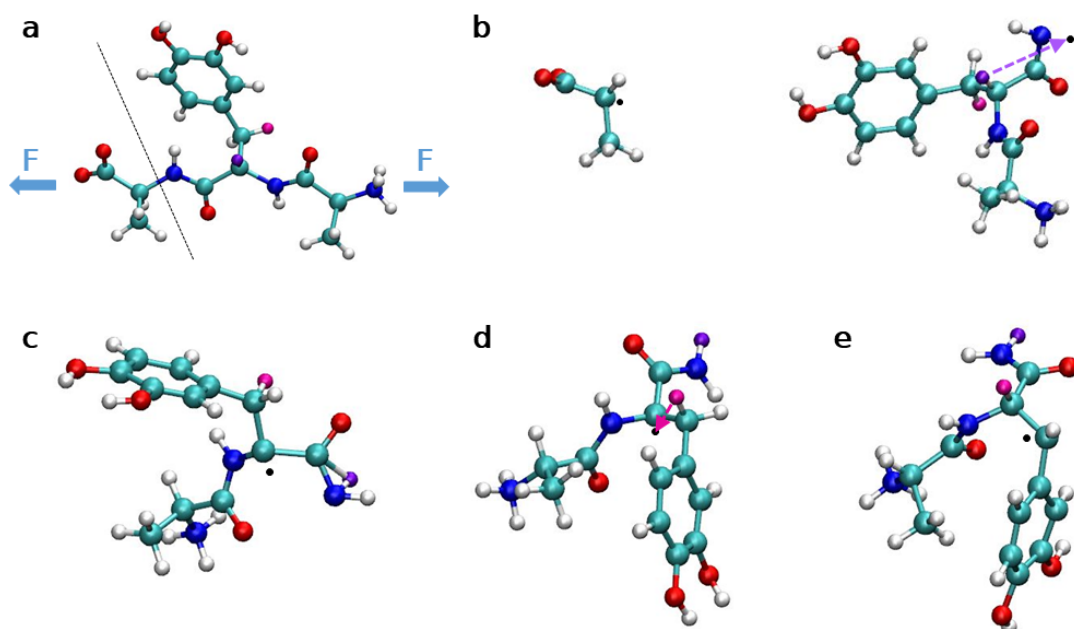
**Figure 3.10: Hydrogen atom transfer (HAT) in a model system**. These frames are snapshots of concatenated all-atom simulations of a ALA-DOPA-ALA model system. We conducted two HATs, but the process could go on afterwards **a** The ALA-DOPA-ALA system is loaded with external force, a bond rupture will happen at the dashed line. **b** After the homolysis, there are two fragments with one radical each (marked with a black dot). In the larger fragment, we identify the purple hydrogen as nearest to the radical site. We will not show the smaller fragment in the following, as it moves away. **c** The purple hydrogen travels, together with this the radical site migrates in the opposite direction. **d** The radical is now at the new position, the magenta hydrogen is identified for the next transfer. **e** After the second transfer, the magenta hydrogen is now at the starting position of the purple hydrogen. The radical took the place of the magenta hydrogen.

reaction types to compete at the same time, with their propensities simply determined by their relative rates. At this stage, the candidates for transfers were just chosen based on their distance to the radicals. However, recent efforts in our group to learn the energy barriers of this reaction have just been published [119] and pave the road towards using Machine Learning models within the flexible KIMMDY scheme. Leaving reaction rates aside, the main technical challenge we started to tackle here was to create the new topology of the system after a hydrogen atom transfer, that is, to get all the updated bonds, angles, and dihedrals written into the input files. Next, we conducted the actual transition using an MD simulation with a smaller step size, before going back to the standard simulation settings. An example of this process can be seen in Fig. 3.10: The bond is broken in Fig. 3.10a and we then observe two hydrogen atom transfers on the larger fragment. First, the violet hydrogen migrates to the original radical site in panels b and c. Then, the magenta atom takes up this place in panels d and e. This could in principle go on until a more stable configuration is reached, for example if the radical stabilizes on the DOPA ring as proposed earlier [8].

## 3.3   Towards ultra-large all-atom collagen simulations

We have seen that collagen has a hierarchical structure spanning multiple length scales, with one individual molecule already reaching a length of 300 nm. At the dawn of exascale computers, the opportunity to simulate a fibril model containing full molecules comes in reach. To this end, we have reached out to the RIKEN institute in Japan, in particular to the authors of the molecular software GENeralized-Ensemble SImulation System (GENESIS) that is optimized for ultra-large systems up to one billion atoms [72] on FUGAKU, currently the second-largest supercomputer in the world. Other MD engines like GROMACS rather optimize for speed to capture larger time scales for standard protein sizes efficiently. In contrast, the implementation of GENESIS is optimized to reduce communication, for example by using a parallel I/O system for each domain, and thereby focusing on scalability to large system sizes [72]. Moreover, the GENESIS code is very much tailored to the specific hardware at FUGAKU [72].

While our collaboration partners provide the computational resources and technical expertise, another PhD student in our group has developed an ultra-large atomistic model (45 million atoms including solvent). As part of this thesis, we iteratively validated and updated the models after preliminary MD simulations on both SuperMUC-NG in Germany, with our computing time grant, and on FUGAKU in Japan, the latter also

during a 3-week research stay in Japan. For instance, we identified that capping the models directly behind the crosslinks resulted in instabilities (as force then concentrates there too much), and we adjusted the systems accordingly. Secondly, as one may expect, already the handling of the arising amount of data is challenging, which is why we proposed a reduced trajectory for on-the-fly analysis that only contains caps and crosslinks. With this, we not only monitor the ends of the strands (and thereby the overall stability), but also the re-arrangements at the border between the gap and overlap regions, changes in D-period, and similar measures, with a minimal amount of data handling.

From our experience with KIMMDY and ColBuilder, we decided for a set of three models: One containing only trivalent crosslinks, one with a mix of divalent and trivalent crosslinks, and lastly a model that is only crosslinked at 70% of all possible crosslink sites. While we chose the connection sites randomly, we constrained the algorithm such that at least one connection per molecules remained to ensure connectivity. As each collagen molecule features four connection points (two at the N- and C-terminal, respectively) that it would then share with a second collagen molecule, a maximum of two crosslinks per (individual) collagen molecule are possible (= 2 mol/mol). Depending on tissue type and age, biologically relevant values of crosslinking density have been reported to be around 1.5 mol/mol (so about 75% crosslinked). Such a system setup is only possible on this large scale, because only then the unconnected parts are still stable inside the network. In our 67 nm long fibril system, in contrast, they would be pulled out in the simulations. For this partially connected collagen system, we expect the force distribution and, therefore, also the rupture propensities to pronouncedly change, and gain insights that would not be possible with our current systems (see especially chapter 5).

Currently, after a number of attempts, we found stable configurations of our three models and constance force MD simulations are being conducted on FUGAKU. To supervise their stability, and also for comparison and validation with existing data, we continuously analyse the end-to-end distance over time, shown in Fig. 3.11 for a recent trajectory status. We reach strains of 21-22%, which is consistent with what we observe for our standard all-atom models (Fig. 4.2). Further, we observe that at the current stage the models are still extending, so more simulation time is needed. Interestingly, differences between the models are already visible: The 'mut' configuration, that is only crosslinked at 70% of the sites, can extend more despite being still a fully connected network (i.e. no strands can be pulled out). This suggests that there is some extra length
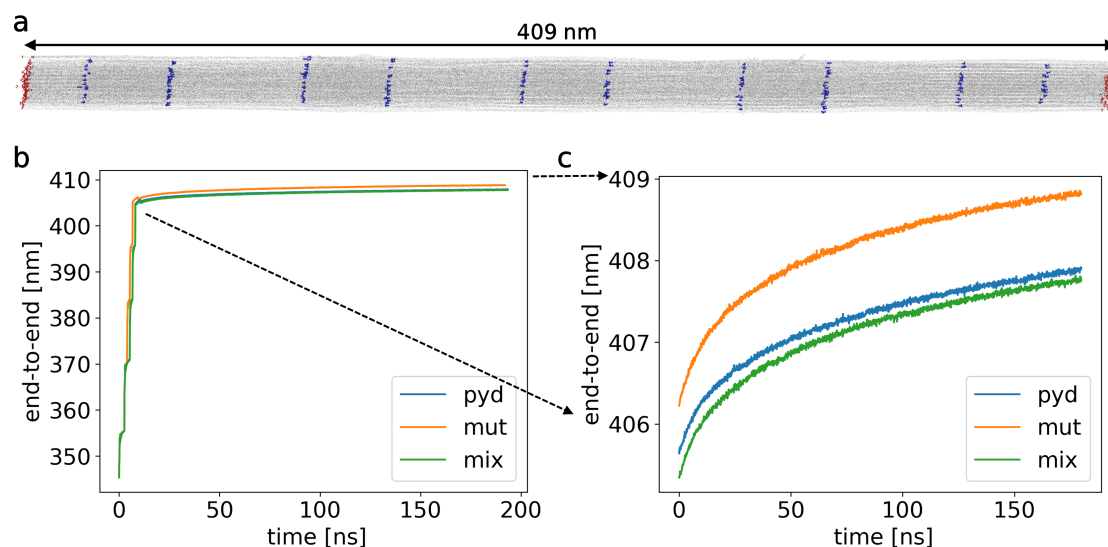
**Figure 3.11: Extensions of our ultra large atomistic models**. We calculate the end-to-end distance over time for our 3 models: pyd with trivalent crosslinks, mix with 50% divalent and 50% trivalent crosslinks, and mut with 70% trivalent crosslinks. **a** The tensed fibril model, here shown for the 70% crosslinked example, reaches about 409 nm length after about 200 ns simulation time. The crosslinks are shown in blue, such that the typical gap/overlap pattern can be seen. **b** Starting at an initial length of about 335 nm, we ramp up the force step-wise before switching to the production run. **c** Production run only (so zoom-in without the first 20 ns ramp-up), showing that the end-to-end distances are not fully converged yet.

in between the crosslinked regions that gets released when there is only one of the two possible connections. We investigate a similar phenomenon in detail in chapter 5. Beside investigating the aforementioned structural features such as gap/overlap region, future plans include the use of KIMMDY to induce bond ruptures in the system. At this size, several breakages distributed throughout the fibril are both more likely, but are also technically easier to achieve while keeping the system stable. Thereafter, continuation runs are planned to assess the impact on the structural integrity, but also on how the force distribution changed and collagen adapts to the new broken topology.

# Collagen Breaks at Weak Sacrificial Bonds Taming its Mechanoradicals

*Figures and text throughout this chapter are reproduced and/or adapted from our publication [38]: Rennekamp et al. "Collagen breaks at weak sacrificial bonds taming its mechanoradicals", Nature Communications 14, 2075 (2023).*

*We provide a more detailed analysis here than in the space-limited article, and combine its Supplementary Information into a continuos text. As stated in the author contribution section of the article, I was the main contributor of the MD and KIMMDY simulations and their analysis, and also analysed the experimental data that was obtained by others. Note especially that the QM part of the study, shortly described in section 4.1 as a prerequisite for the rest of the chapter, was performed by my co-supervised Master student Christoph Karfusehr [120].*

As we have seen in the previous chapters, under elastic deformations, tendon collagen I forms mechanoradicals originating from covalent bond rupture [8], just as in any other synthetic polymer [12, 13]. However, in contrast to synthetic materials, collagen readily stabilizes the primary radicals on dihydroxy-phenylalanines (DOPAs), with these highly efficient radical scavenging residues built into the collagen protein [8, 9]. The microscopic failure mechanisms define macroscopic mechanical properties such as ultimate strain and toughness. Further, the positions of the weakest links within the structure determine where and what types of mechanoradicals are formed and how they will propagate through and potentially damage the system until they get scavenged. Pinning down the rupture process of collagen, as the major substituent of tissues, is a decisive step in addressing tissue degradation and ageing, and for guiding tissue engineering.

In this chapter, we examine where collagen ruptures on a molecular scale and discuss

how the propensities of these ruptures might influence subsequent reactions in the material. We have seen in section 2.3 that modelling a molecular rupture process needs two key factors: First, the initial energy barrier as given by the relative strength of the chemical bonds, which we investigate in section 4.1. The second determinant of rupture is how force distributes through the complex hierarchical structure of collagen and loads its chemical bonds. In collagen, the three-dimensional structure, as introduced in section 2.1, combined with the crosslinking in between the triple helices, results in a molecular force distribution network that critically decides on the rupture propensity of an individual bond. For this reason, we employ scale-bridging hybrid Molecular Dynamics simulations for reactive dynamics (KIMMDY [27]) to assess collagen's mechanochemical rupture. In addition, gel electrophoresis-coupled mass spectrometry (MS) of stressed collagen has been conducted by others and analysed here to validate the computational results of both backbone and crosslink rupture.

## 4.1   Crosslinks harbour the weakest bonds

The first determinant of rupture propensity is the relative thermodynamic strength of the chemical bonds. As force is transmitted via the protein backbone, we first consider the three types of backbone bonds: the C-N peptide bond, the $C_\alpha$-N bond and the $C_\alpha$-C bond. Due to the $\pi$-conjugation across the peptide bond, this bond is an unlikely candidate for homolytic rupture. Which of the two single bonds, $C_\alpha$-N or $C_\alpha$-C, is the weaker link is less obvious, as previous studies suggest bond dissociation energies for these two to be in a similar range [8, 20]. As mentioned before, collagen also features enzymatically derived crosslinks connecting the individual triple helices. All of them are derived from lysines, and as a consequence feature again single bonds of the C-C and C-N type. The thermodynamic differences in homolytic scission of collagen's backbone and crosslink bonds are decisive for collagen mechanics, but were unknown in previous work.

To determine the stability of covalent bonds in collagen I and identify the thermodynamically weakest links, high accuracy QM calculations were performed on the G4(MP2)-6X level of theory, which reaches an accuracy similar to that of G4 for thermochemical properties [121]. Christoph Karfusehr obtained bond dissociation energies (BDEs) of bonds from four major collagen I crosslinks (trivalent: PYD/DPD, divalent: HLKNL and deH-LNL), as well as of backbone bonds of collagen's main amino acids (glycine, proline and alanine). Further, radical stabilization energies (RSEs) of homolytic bond scission
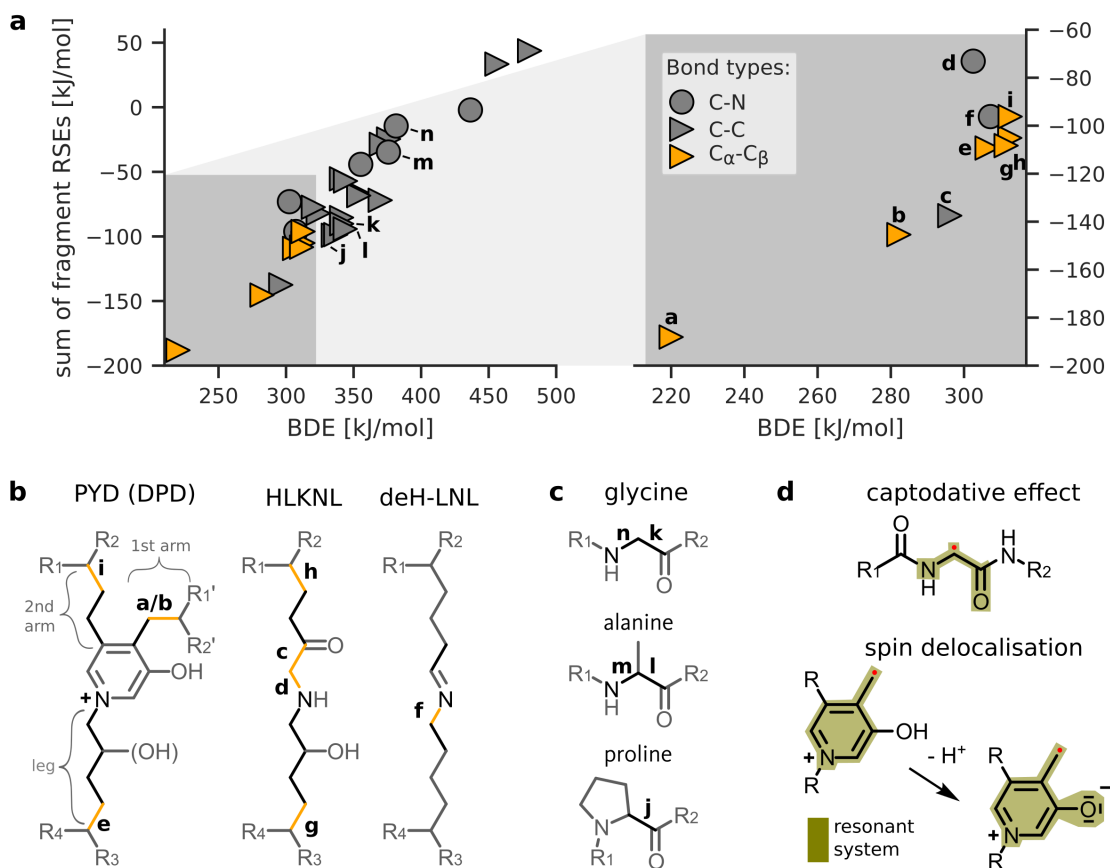
**Figure 4.1: Crosslinks harbour the weakest bonds in collagen type I.** **a** Correlation between the bond dissociation energies (BDEs) and summed radical stabilization energies (RSEs) of homolytically cleaved bonds and formed radicals. All molecular structures used for calculations can be found in the SI of [38]. Bonds with BDEs < 315 kJ/mol are shown in more detail on the right. The small letters close to selected data points indicate respective bonds in the lower panels b and c. **b,c** Molecular structures of (b) investigated collagen crosslinks and (c) prevalent amino acids in collagen. BDEs of bonds coloured in black and orange (BDE < 315 kJ/mol) were explicitly calculated. Residues R1, R1', R2 and R2' connect to the same collagen triple helix. The BDEs of the PYD bond denoted with letter "a" was obtained after deprotonation of the pyridine bound hydroxyl group. **d** Main electronic effects leading to the lowest found BDEs. The captodative effect describes the non-additive stabilization of the amino-carbonyl and carbonyl-amino group present at any $C_\alpha$ peptide radical. The lower part illustrates the increased radical delocalization upon deprotonation of PYD and DPD, lowering the BDE and RSE of bond "b" to that of bond "a". Figure by Christoph Karfusehr, from our publication [38].

products were calculated to examine the origin of the large differences in BDEs, which can exceed 160 kJ/mol across C-C bonds within a single crosslink. Figure 4.1a shows the BDEs of all calculated bonds against the summed RSEs of the two radicals formed by each homolytic bond scission. Overall, the linear correlation between low BDEs and large negative summed RSEs shows that BDE deviations from aliphatic C-C bonds can generally be explained by stabilization of the formed radical centres by proximate chemical groups.

Across collagen crosslinks and backbone bonds, the results suggest that $C_\alpha$-$C_\beta$ bonds are promising bond rupture candidates, having comparatively low BDEs $\leq$ 312 kJ/mol (with other C-C bonds being rather in the range 340-370 kJ/mol). The high stability (low RSE) of the formed $C_\alpha$ centered radical of -87 kJ/mol explains the low $C_\alpha$-$C_\beta$ BDE, as do comparable literature values (-74.1 kJ/mol [122] and -75.5 kJ/mol [20]). We attribute the comparably low RSE to the stabilizing effects of carbonyl-amino and amino-carbonyl substituents present in peptides, which are further increased when combined. Such non-linear donor/acceptor interactions are known as captodative effects (Fig. 4.1, panel d) [123, 124].

Interestingly, the presence of a second strong radical stabilizing system on the $C_\beta$ centered radical further lowers the BDE (Fig. 4.1, bond "b") in the trivalent crosslinks Pyridinoline (PYD) and Deoxypyridinoline (DPD) to just 282 kJ/mol. The stabilizing effect of the conjugated aromatic system becomes even more pronounced upon deprotonation of the pyridine bound hydroxyl group, extending the resonant system and giving an exceptionally low BDE of 220 kJ/mol. Importantly, the pKa value of a reference compound (3-Hydroxypyridine methochloride, pKa = 4.96 [125]) suggests a pKa of about 5, implying that PYD and DPD crosslinks are mostly deprotonated at neutral pH (and even more so their respective radicals). Focusing on divalent crosslinks Hydroxylysino-keto-norleucine (HLKNL) and Dehydro-lysino-norleucine (deH-LNL), we found (additional to $C_\alpha$-$C_\beta$ bonds) two more weak bonds in HLKNL and one more in deH-LNL.

Notably, none of the backbone bonds appear as likely rupture candidates. All calculated BDEs of backbone bonds, both $C_\alpha$-C and $C_\alpha$-N bonds, are considerably higher than those of the weak crosslink bonds ($\Delta$BDE $\geq$ 24 kJ/mol). Within the backbone, $C_\alpha$-C bonds are overall weaker than $C_\alpha$-N bonds and are the primary rupture candidates of the protein backbone, in agreement with previous semi-empirical calculations [126], but

here regardless of the amino acid type.

In essence, the quantum chemical calculations identify bonds within crosslinks as weak bonds within collagen, in particular $C_\alpha$-$C_\beta$ bonds reaching values down to 220 kJ/mol in the trivalent PYD crosslink. These can be backtraced to excellent stabilization of the resulting mechanoradicals.

## 4.2   Ruptures occur dominantly in and around crosslinks, but also unspecific in backbones

To directly predict relative rupture counts within a stretched collagen fibril, we utilized KIMMDY [27] as described in section 2.3. The loaded molecular system rearranges and reaches a new stretched (quasi-)equilibrium in the simulations that mimics the conditions from previous experiments [8], in which homolytic bond ruptures were observed inside macroscopic collagen tendons at a sub-failure force regime (corresponding for example to heavily exercising without macroscopic tendon rupture).

### 4.2.1   Simulation set-up and convergence

For the simulations, we use atomistic collagen type I fibril models taken from Col-Builder [34] as described in section 3.1. The simulated structures comprise one overlap and gap region in length and 41 triple helices in width (in the overlap region).  The different collagen models were solvated in TIP3P water leading to system sizes of approximately 2.5 - 2.9 million atoms, for which periodic boundary conditions were chosen. After energy minimizations, we conducted equilibrations in NVT and NPT ensemble for 10 ns each. The complete set of simulation parameters can be found in the appendix table A.1.

For data production, we subjected the collagen fibrils to a constant force of 1 nN per chain in four different schemes: In the first scheme, we distributed the forces equally leading to 3 nN on each triple helix, pulling from both sides. In schemes two and three, we also pulled equally on all strands, but just from one side while restraining the other side of the molecule. We used this as a control that our results are independent of the time the force needs to propagate through the system and of internal friction. In the last scheme, we inhomogenously distributed forces on the inner triple helices by drawing forces from a Gaussian distribution with the same average of $F_{av} = 1$ nN and a width of

$\sigma = F_{\mathrm{av}}/3$. The ring of outer helices, however, was subject to 1 nN to avoid a sliding of the chains against each other. We refer to this scheme as shear pulling, it mimics more imperfect biological samples. In all cases, we additionally employed torque restraints to prevent unwinding of the capped collagen triple helices [127]. The applied stress level leads to a new, stretched equilibrium of the fibril that corresponds macroscopically to a sub-failure regime with occasional bond rupture, consistent with the scenario of radical generation in stressed tendons from previous work [8].

Overall, we used collagen fibril models from three different species *Rattus Norvegicus*, *Pongo Abelii* and *Loxodonta Africana*, which yet again varied in their crosslink position and also crosslink type (divalent HLKNL or trivalent PYD). To account for the stochastic nature of MD simulations, we simulated 3-4 replica per model type. Including replica, we in total conducted 63 simulations. In each instance, the collagen fibrils were pulled for 100 ns before executing the KIMMDY scheme of bond rupture as discussed below. For 9 simulations with PYD crosslinks, we thereafter conducted a second cycle of 100 ns pulling to investigate secondary breakages. All models and their replica taken together yield a combined total simulation time of more than $7.2\,\mu$s. Along with the publication, the combined results and simulation input files were made publicly available at Hei-Data [128].
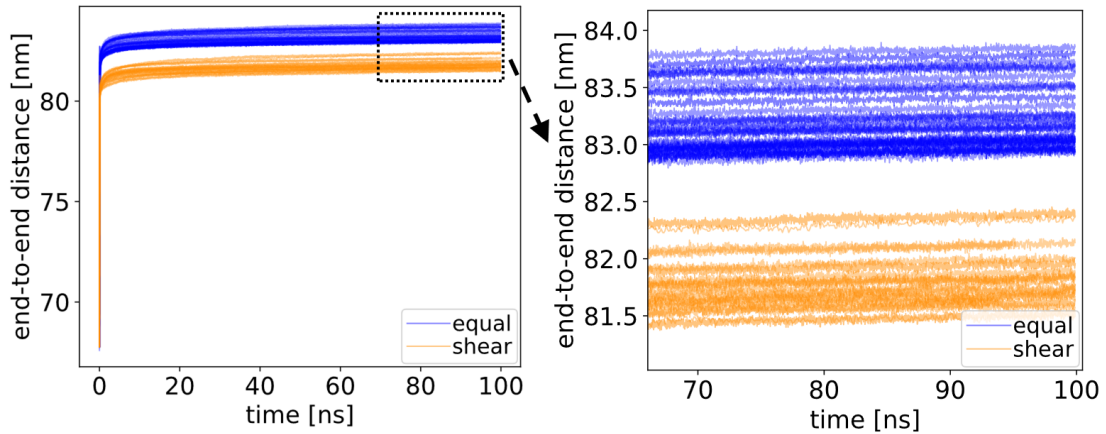


**Figure 4.2: End-to-End distances of the fibrils in the MD simulations**. Data from setups in which all fibrils are pulled with the same force ('equal') are shown in blue, and data from inhomogeneous pulling setups ('shear', as described above) in orange. On the right side, we zoomed in on the last $\approx 30$ ns.

In order to test the convergence of the pulling simulations, we monitored the end-to-end distances of the fibrils as shown in Fig. 4.2. While the two different batches vary slightly, all fibrils get strongly stretched from the starting length of about 67.5 nm in the beginning and then converge to on average $\approx 83.4$ nm or $\approx 81.9$ nm respectively — a behaviour that is expected due to the immediately applied constant force. From approximately 50 ns onwards, the fibrils show only a very minor further increase relative to the whole fibril length, suggesting that the internal force distribution has also largely converged. Of course, even longer simulations times might be desirable to give room for these very last re-arrangements, but in the trade-off (due to limited computational resources) for more replica, we can gain more information having more rather than longer simulations. Another observation is that the main stretching happens on a time scale in the order of 10 ns, such that microscopically a new stretched (quasi-)equilibrium can be reached quickly. Overall, we reach strains in the range 20%-24%. Whereas we try different enzymatic crosslinks, these models do not incorporate so-called advanced glycation endproduct (AGE) crosslinks. They are known to stiffen collagen and lead to a more brittle behaviour [50]. In their absence, the strain we reach appears plausible.
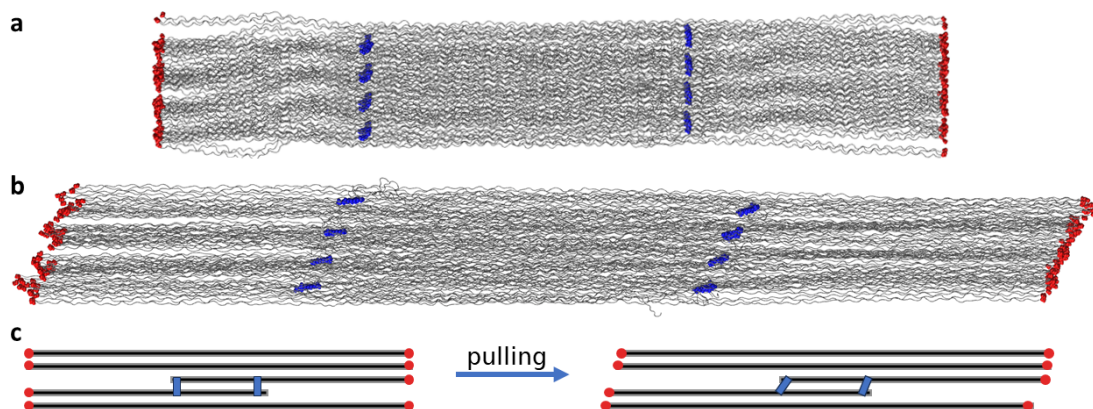


**Figure 4.3: Thinning and tilting of tensed collagen fibrils. a** Our 67 nm long atomistic model after equilibration. We have coloured the caps in red and the crosslinks in blue. **b** After 100 ns of pulling, the fibril is extended. Note especially that the planes of crosslinks and caps are now tilted with respect to the fibril axis. **c** Schematic 2D representation on how the crosslinking in the overlap region can facilitate the tilting.

To get an idea of the conformational changes, we display in Fig. 4.3 exemplary snapshots before and after the pulling simulation. As expected, along with the extension, we can also observe a slight thinning of the fibril. On top of that, we note that the cross-sectional

planes of the caps (red) and crosslinks (blue) have a tilt angle with respect to the fibril axis in Fig. 4.3b. Molecular tilting and kinking has been discussed in collagen literature, with varying reported angles in the range of 5 - 18 degree [129, 130, 21, 5]. In our model this is also a result of the topology following the x-ray fiber diffraction data as described in section 3.1.4. The shift in the overlap region and the stretched crosslinks laying down especially contribute to this, as depicted in a schematic manner in Fig. 4.3c.
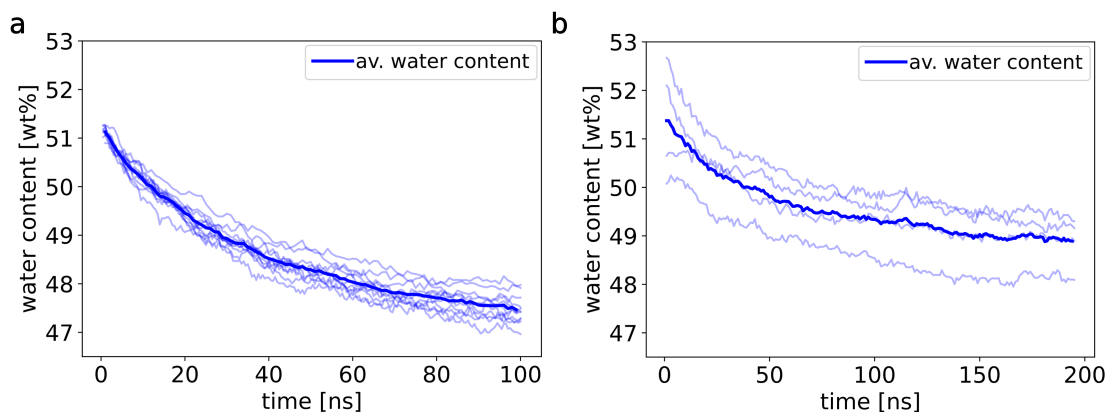


**Figure 4.4: Our collagen models stay highly solvated both in equilibrium and force-probe MD.** We define all waters in and around the fibril within a distance of 0.6 nm to contribute to the relative water content (measured in relative weight). **a** Water content within the ColBuilder validation data set of 12 equilibrium MD simulations, with their average in bold. **b** Water content in steered MD simulations for a sub-set of the described setups for which we obtained 200 ns (instead of 100 ns) long trajectories. In both cases, we see a slight decrease to a constant level of a wet fibril.

Another test of convergence that at the same time also yields interesting results in itself is to look at the water content within the collagen fibril. For this, we defined a cut-off of 0.6 nm around the protein and counted all the water molecules within that cylinder – so mostly inside the fibril and a few hydration layers around it. The results in Fig. 4.4a are comprised of the 12 equilibrium MD simulations that we used also in section 3.1 to check the behaviour of ColBuilder models. We can compare this to panel b, where we show a small sub-set of 4 trajectories from varying species of the aforementioned pulling setup. For these, we conducted 200 ns long steered MD (so twice the length as the remainder of the data set). Overall, in both cases we see that the relative weight of water decreases slightly, but levels-off quickly to values slightly below 50%. The pulling simulations have a larger variation, which can be explained not only due to the smaller

data set, but also due to the fact that the applied force leads to more movement and disturbance of the water than in equilibrium. Still, even in these cases, the fibrils keep their overall level of solvation. While the exact value also depends on the definition of the water shell around the protein (and we here use a rather large cut-off), this analysis confirms the experimental picture of a wet material with water both around and inside the fibril. This implies that subsequent reactions to the radical generation with water, such as the production of hydrogen peroxide, are indeed reasonable.

After the aforementioned 100 ns pulling MD simulations, 5,000 bond elongations per breakage candidate were averaged to generate the bond rupture rates as described previously [27]. These elongations were taken from another 1 ns of simulation now using a PLUMED-patched [131] GROMACS version that can directly output bond distances for a specified set of rupture candidates. In contrast to the original implementation, which used default backbone values for BDEs and did not differentiate between the various crosslink chemistries, we here utilize the BDEs as described in section 4.1. Concerning peptide backbone bonds, we found them to be identical for different amino acids within methodological errors. For subsequent simulations, we used the resulting averages of 379 kJ/mol and 341 kJ/mol for all $C_\alpha$-N and $C_\alpha$-C bonds, respectively. For each of these ensembles of rates, the Kinetic Monte Carlo step was invoked 10,000 times to sample the rupture propensities properly. In the cases where a second rupture was simulated, which we did for 9 simulations with PYD crosslink, we adjusted the topology and relaxed the system with a small 0.2 fs time step before pulling for another 100 ns and repeating the procedure.

## 4.2.2   Identification of primary rupture sites

Notwithstanding the overall higher stability of the backbone bonds according to the quantum chemical calculations, KIMMDY predicts that crosslink and backbone bond ruptures can both occur. The accumulated rupture propensities, from all our simulations for the various models and pulling setups, are shown in Fig. 4.5 in different perspectives. Along the fibril structure (Fig. 4.5a, left), ruptures, albeit at low relative counts, occurred randomly across the whole fiber axis (Fig. 4.5b). Most frequently, however, the crosslinks themselves break (at 73 %). Also, the backbone in their vicinity has a strongly increased rupture propensity due to local stress concentration (12 %). When mapping the rupture statistics onto the collagen I sequence (Fig. 4.5a, right), the prevalence of crosslink ruptures over widely distributed backbone ruptures is again evident (Fig. 4.5c).
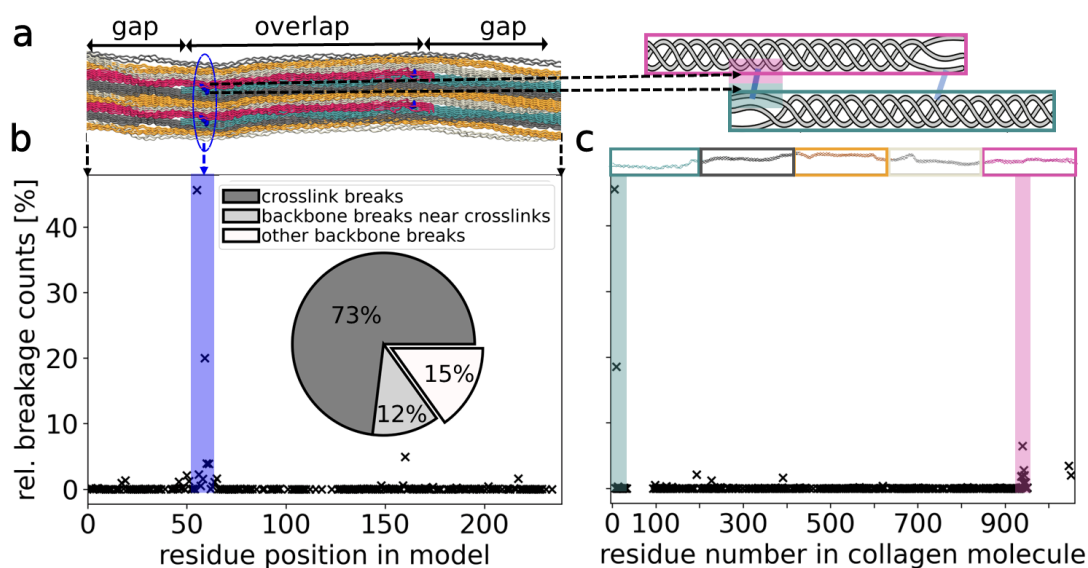
**Figure 4.5: Collagen ruptures primarily in crosslinks, but also unspecific ruptures occur everywhere**. Data was obtained from 63 independent MD simulations including divalent and trivalent crosslink setups, with different collagen sequences and pulling conditions, as described in section 4.2.1. For simulations with trivalent crosslinks, we used the deprotonated BDEs of PYD. In both panels b and c the same data is presented in two different reference systems. **a** Left side: Location of residues and crosslinks in our model along the fiber axis. Triple helices with the same phase are in the same colour, respectively. Upper and lower residues participating in the same crosslinks (blue) appear in the same residue position in the model, see for example the blue encircled area. This view is used in Panel b. Right side: Crosslinks along a whole 300 nm collagen molecule occur at the beginning (cyan area) and at the end of a triple helix (magenta area), shown as cartoon with colouring as in the structure on the left. This reference system, counting along the collagen molecule, is used in Panel c. **b** Propensity of bond breakages in our collagen model. Most ruptures concentrate in the crosslinked area, while there are some scattered backbone ruptures. Inset: Pie chart of summed up ruptures in the crosslinks vs. the backbones in the crosslinked area (up to 5 residues before/behind) vs. elsewhere in the backbone. **c** Propensity of bond breakages within collagen fibrils. The crosslinked areas at the beginning and end of the molecules, as marked exemplary for the N-terminal crosslink in the shaded areas, are most prone to rupture. Note that both in b and c we only show data points where at least one rupture happened.

Only backbone regions where two crosslinked strands are overlapping and thus share the external stress, i.e. at the very beginning and end, are largely void of ruptures (areas without crosses in cyan and magenta boxes). Thus, the external force distributes such that some backbone bonds, in particular those in the vicinity of crosslinks, are sufficiently weakened to compete in terms of bond rupture rates with the intrinsically weaker crosslink bonds. In addition, backbone bonds are statistically still prone to rupture due to their sheer number, with up to ~1,500 amino acids per crosslink in a fully crosslinked collagen I fibril. If crosslink bonds were "only" 10–100 times more prone to rupture than individual backbone bonds, they would still not dominate the distribution.

Inspired by a contemporaneous publication [126], we also analysed the rupture distribution along the GLY-X-Y pattern in collagen to be 0.7%-82.2%-17.0% in backbone bonds (excluding the common crosslink ruptures). As all backbone bonds have the same BDEs, this results from a stronger loading within the bonds, likely due to the h-bonding network that is known to stabilize the collagen triple helix via the GLY-X-Y pattern. Whereas this is consistent with the reported values along this pattern [126], we could not confirm the suggested over-representation of prolines among these X-ruptures, which might be due to methodological differences.
Furthermore, this is another indicator that KIMMDY is able to capture these differences within the collagen triple helix arising from the detailed molecular interactions in MD. As mentioned in section 3.2.2, the preliminary results on the issue of intrinsic differences of amino acids vs. differences due to their position in collagen have culminated in another research project and are, hence, ongoing work.

The results shown in Fig. 4.5 are compiled from a set of KIMMDY simulations varying in crosslinks, collagen sequence (different species) and pulling conditions (see section 4.2.1 above). Importantly, in all the models and settings, we recover a competition between crosslink and backbone bond ruptures, i.e. a concentration of ruptures at and around crosslinks, and the random distribution of the remaining backbone ruptures along the sequence.

As it is shown in Fig. 4.6, this competition is recovered for all species (i.e. different sequences). Moreover, the correspondence between the setups 'left' and 'right' pulling, in which we fixed one side and pulled only from the other side, show that this does not depend on the time of force propagation. Even the heterogenous 'shear' setup shows a qualitatively agreeing rupture concentration at the crosslinks.
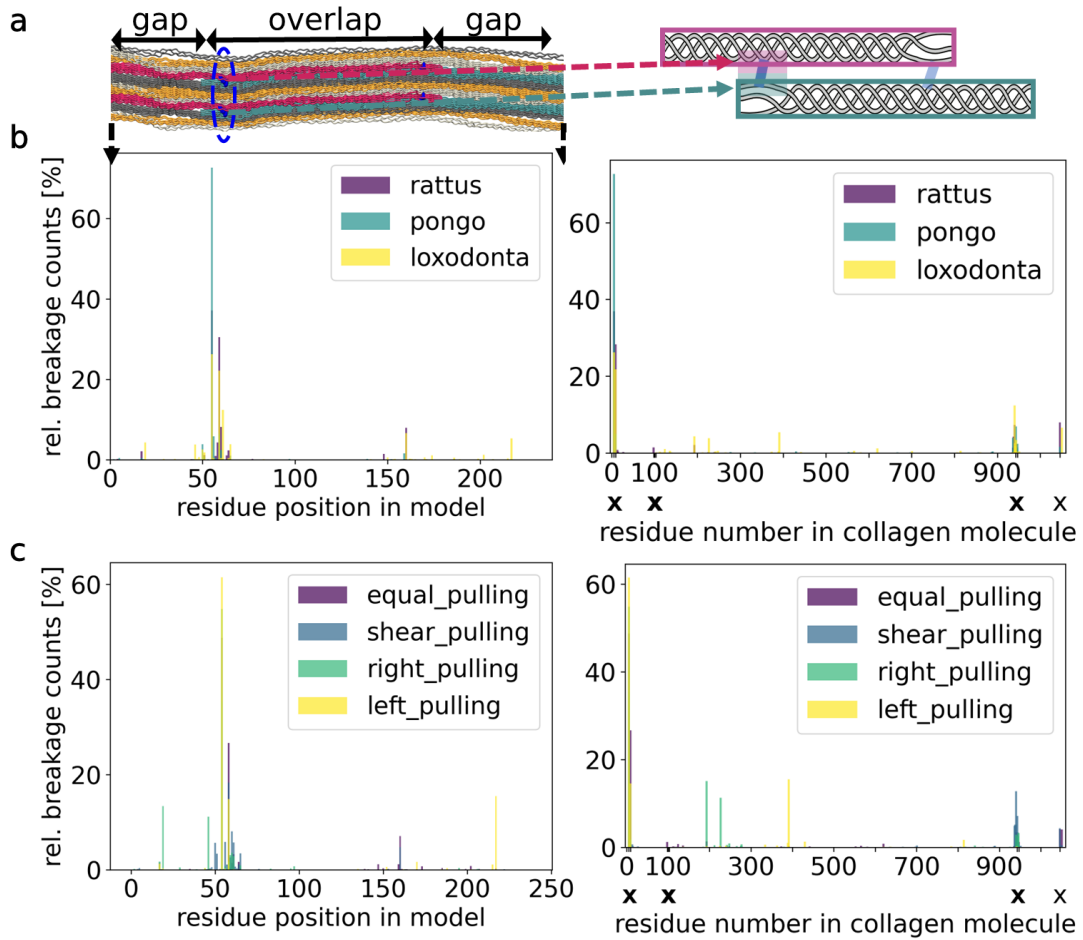
**Figure 4.6: Comparison of breakage sites depending on model species and pulling setup**. This analysis was performed for the full data set, percentages are normalized within each subset. Crosslink positions are marked with an X in the right panels. **a** Location inside the model (left) and along a collagen triple helix (right) as will be used in b and c. **b** Data for the different model species. The subsets consist of 23 simulations for Rattus, 21 for Pongo and 19 for Loxodonta. **c** Data with respect to the pulling setup. The subsets consist of 6 simulations with divalent crosslinks for each left and right pulling, 28 and 23 simulations with mixed crosslinks for equal and shear pulling, respectively.

However, quantitative rupture counts can vary strongly and also depend on model choice and composition. For instance, if we assume acidic pH, that is, 100% protonated PYD, we have a lowest BDE of 282 kJ/mol instead of 220 kJ/mol (Fig. 4.1a, bond b versus a). In this case, the ratio shifts to 28%:41%:31% of breakages in crosslinks to those in adjacent backbone bonds to other unspecific backbone bonds (compare Fig. 4.7). Hence, as crosslinks now feature bonds strengths closer to those of the backbone, ruptures throughout the backbone can compete with those in crosslinks under the dynamic stress concentrations present in the stretched fibril. To conclude, crosslinks indeed represent weak spots, not only in terms of their BDEs but also in the context of the stress distribution across the collagen structure. The weaker the crosslinks are compared to the backbone, the more initial ruptures are funnelled into crosslinks preventing breakages at other unspecific positions in the protein backbone.

## 4.3  Trivalent crosslinks harbour sacrificial bonds

The trivalent crosslinks become more prevalent with age and connect two 'arms', i.e. two modified lysine side chains of different strands from within one triple helix, to one lysine residue ('leg') of another triple helix (see Fig. 4.1b or also Fig. 2.3). It has been speculated that due to this double connection, trivalent crosslinks provide more stability yet lower elasticity. Having established crosslinks as major ruptures sites by our combined calculations, we next asked how the rupture propensity of the trivalent crosslink compares to its divalent counterpart (chemical structures shown in Fig. 4.1b), and how the redundant connectivity of the trivalent crosslink is compromised by rupture.

At first glance, the distribution of ruptures does not depend heavily on the crosslink type, except for an even stronger concentration in the crosslinks for PYD. For the detailed rupture distributions (compare Fig. 4.7b for PYD to Fig. 3.9b for HLKNL). In Fig. 4.8a, however, we summed up all bond rupture rates within one simulation and grouped these total rates per crosslink type used in the model. The models with trivalent (PYD) crosslinks exhibit clearly higher rupture rates. While this appears counter-intuitive at the first glance, since the two arms can share the external mechanical load, the explanation is straightforward: the short arm of the trivalent crosslink has by far the lowest BDE due to the combination of the captodative effect and conjugation with the aromatic ring (see Fig. 4.1d). While this leads to a higher rupture propensity for a first breakage, the overall crosslink remains intact due to the second arm. The low BDE of this bond also
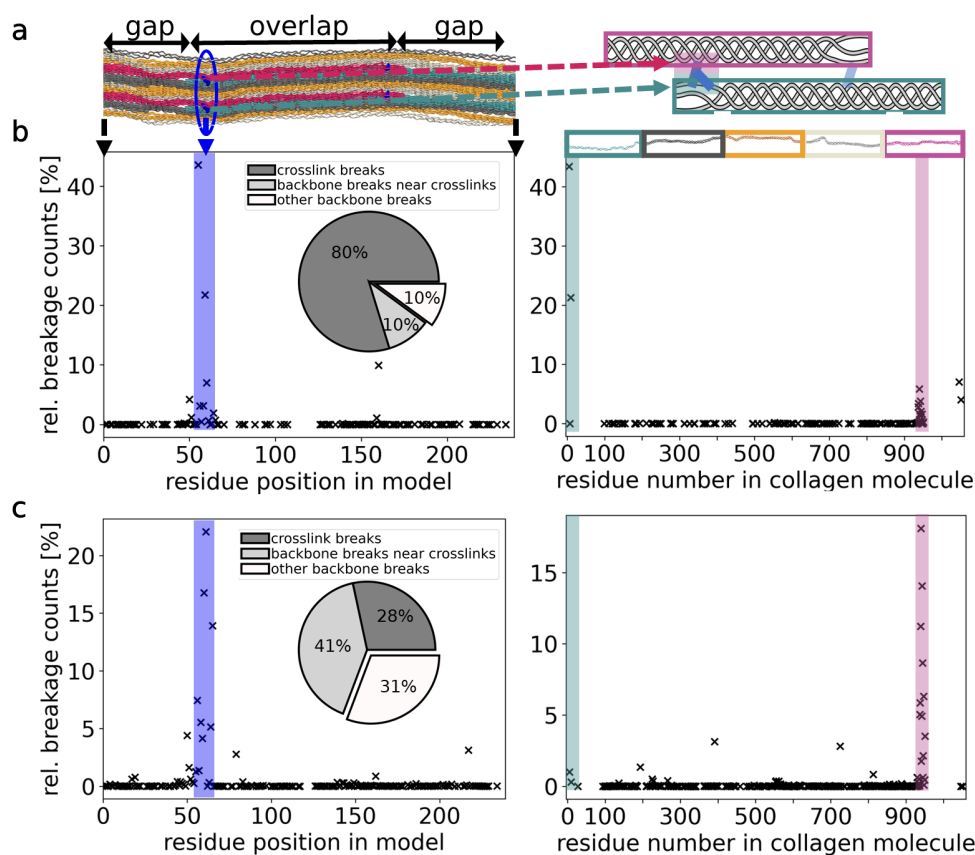
**Figure 4.7: Comparison of breakage sites in PYD using different BDEs depending on protonation states**. This analysis was performed for data of trivalent (PYD) crosslinks only. **a** Location inside the model (left) and along a collagen triple helix (right) as will be used in b and c. **b** Data for (default) BDE of 220 kJ/mol in the $C_\alpha$-$C_\beta$ bond in the shorter arm. N-terminal crosslinks as preferred breakage sites are highlighted in blue in the left panel and, in the right panel, with cyan and magenta depending on the side of the crosslink. **c** Data for BDE of 282 kJ/mol in the $C_\alpha$-$C_\beta$ bond in the shorter arm, corresponding to acidic conditions. The ruptures concentrate overall in the same regions, but are more spread out among different residues (also note the different y-axis). In the representation along the collagen molecule on the right, it can be seen how the breakages shift from one of the two arms to the leg of the trivalent crosslink.
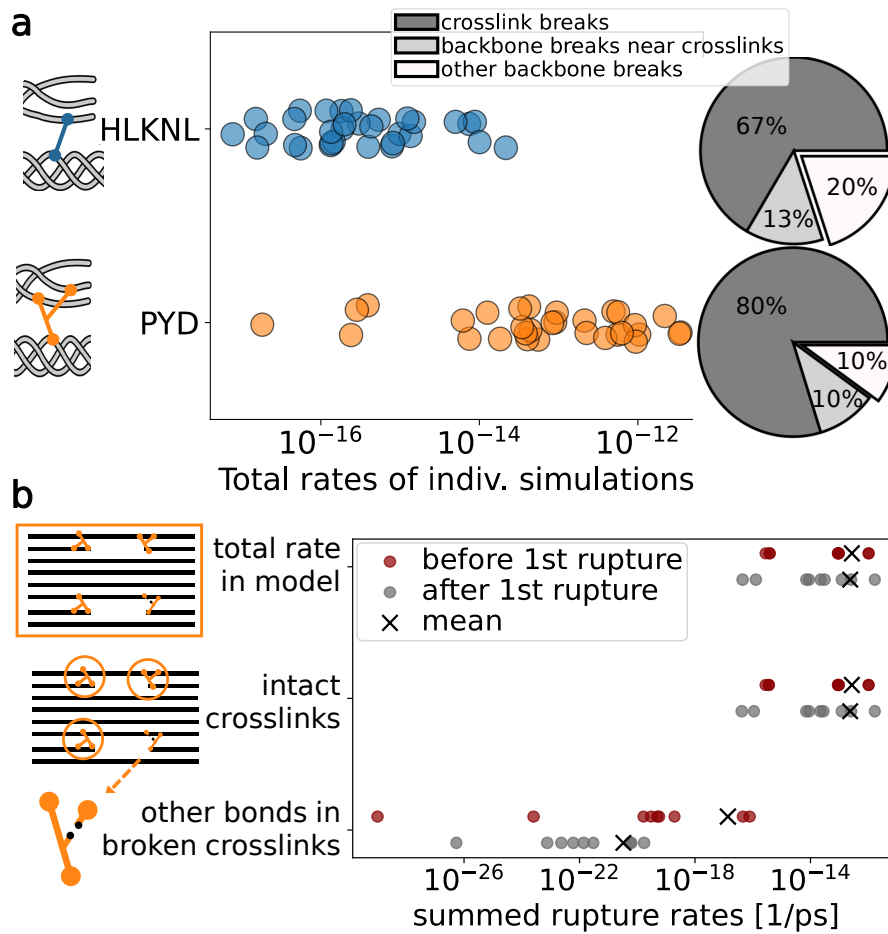
**Figure 4.8: Trivalent crosslinks break faster, but (at first) without loosing connection.**
**a** Comparison of total primary bond rupture rates in simulations with models having divalent (HLKNL) vs. models having trivalent (PYD) crosslinks, which show overall higher rates. The ratio of crosslink ruptures to other breakages confirms that the increased rupture rates are indeed due to higher propensities in the crosslinks. **b** Comparison of secondary bond rupture rates before and after the first arm in one PYD crosslink broke. The rate in the second arm even decreases, being several orders of magnitude lower compared to the sum in the remaining model. Hence, other bonds will break first, and the fibril stays overall intact. Each data point is from one of 9 simulations with trivalent crosslinks, which we continued after a first rupture in the short PYD arm.

renders ruptures across the collagen fibril highly localized to the crosslink, with 80 % ruptures occurring therein (Fig. 4.8a). Unspecific backbone ruptures halved compared to collagen with divalent crosslinks, from 20 % to 10 % of the rupture counts.

Rupture of a trivalent crosslink on one of the two arms leaves the linkage of the two triple helices overall intact. We next examined if subsequent ruptures occur primarily within the same crosslink or rather within other fully intact trivalent crosslinks. To this end, we continued 9 pulling simulation of models with trivalent crosslinks after a rupture inside these arms, examined the new rupture rates and compared to the situation of a fully intact pre-rupture fibril (see Fig. 4.8b). Surprisingly, despite now lacking one force-bearing connection, the total bond rupture rate in the damaged trivalent crosslink decreases by ∼3 orders of magnitude compared to the pre-rupture simulations. As bond strengths are unaffected, this decrease indicates that stress in this crosslink must have relaxed after its rupture. Indeed, we find an average distance increase of the crosslink $C_\alpha$ atoms after rupture by ∼0.1 nm, as the longer arm can now elongate and use the previously hidden length (Fig. 4.9).
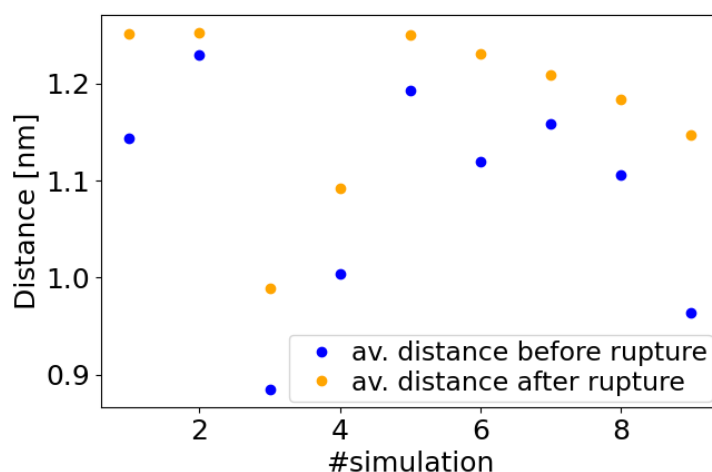


**Figure 4.9: PYD crosslinks extend after rupture of the short arm.** We measure the distance of $C_\alpha$ atoms of the (not broken) second arm to the $C_\alpha$ atoms of the leg of the PYD crosslink before and after rupturing the shorter first arm. In all replica, the distance increases, indicating that the crosslink can now extend by the previously hidden length.

Also, the remaining intact crosslinks, that still each harbour the extraordinarily weak bond, dominate the total of all rates in the simulation. Another rupture in the damaged crosslink is highly unlikely, and this crosslink stays intact. Instead, subsequent ruptures

are predicted to mainly occur in the same type of weak $C_\alpha$-$C_\beta$ bonds in other triple helices. Our combined data in Fig. 4.8 suggests that the PYD and DPD crosslinks are less stable with respect to first ruptures, without compromising fibril integrity.

## 4.4 Experiments confirm covalent backbone rupture and altered crosslinking

We next set out to experimentally validate our predictions of both crosslink and backbone covalent bond ruptures in collagen. Electron paramagnetic resonance (EPR) spectroscopy is a highly sensitive method for detection of mechanoradicals, and has been used for this purpose successfully [8], however it does not allow spatial resolution and in addition rapid migration onto DOPA prohibits the primary rupture site identification. We instead resorted to gel electrophoresis combined with mass spectrometry. More details about the methods that our experimental collaborators used to obtain the raw data, which was then analysed by us during this thesis, can be found in the Appendix A.3 or in the corresponding publication [38].

Fig. 4.10a shows the protein molecular weight distribution of solubilized collagen polypeptides from stressed and unstressed (control) collagen rat tail tendon resolved by an SDS-PAGE gel conducted by our collaborators. We recovered the expected band pattern, including the monomeric $\alpha$-chains ($\alpha$1(I) and $\alpha$2(I)), as well as higher order oligomers (e.g. $\beta$-chains, $\gamma$-chains and $\delta$-chains). For the stressed samples, we additionally observe bands in the mass regime of fragmented chains, indicating backbone rupture of single collagen strands. To quantify the proteolytic breakdown in stressed and control tissues, we obtain the densitometry of stained collagen chains and peptide fragments across samples with ImageJ [132] in Fig. 4.10b. Stressing the samples gives rise to fragments of masses lower than those of full-length collagen chains, across a wide range of masses down to small fragments running at the dye front.

Second, our collaborators identified intact $\alpha$-chains and lower molecular weight polypeptides by mass spectrometry. We took them from regions as indicated in Fig. 4.10a from stressed samples #2 and #4, because peptide fragments from samples #1 and #3 were below the limits of detection. The sequence coverage of the collagen $\alpha$-chains, obtained with Proteome Discoverer, analysed in the course of this thesis, and displayed in Fig. 4.10c, confirms the presence of both $\alpha$1- and $\alpha$2- chain fragments in all these areas, even in the lowest mass fractions of the dye front. Overall, this experimental analysis
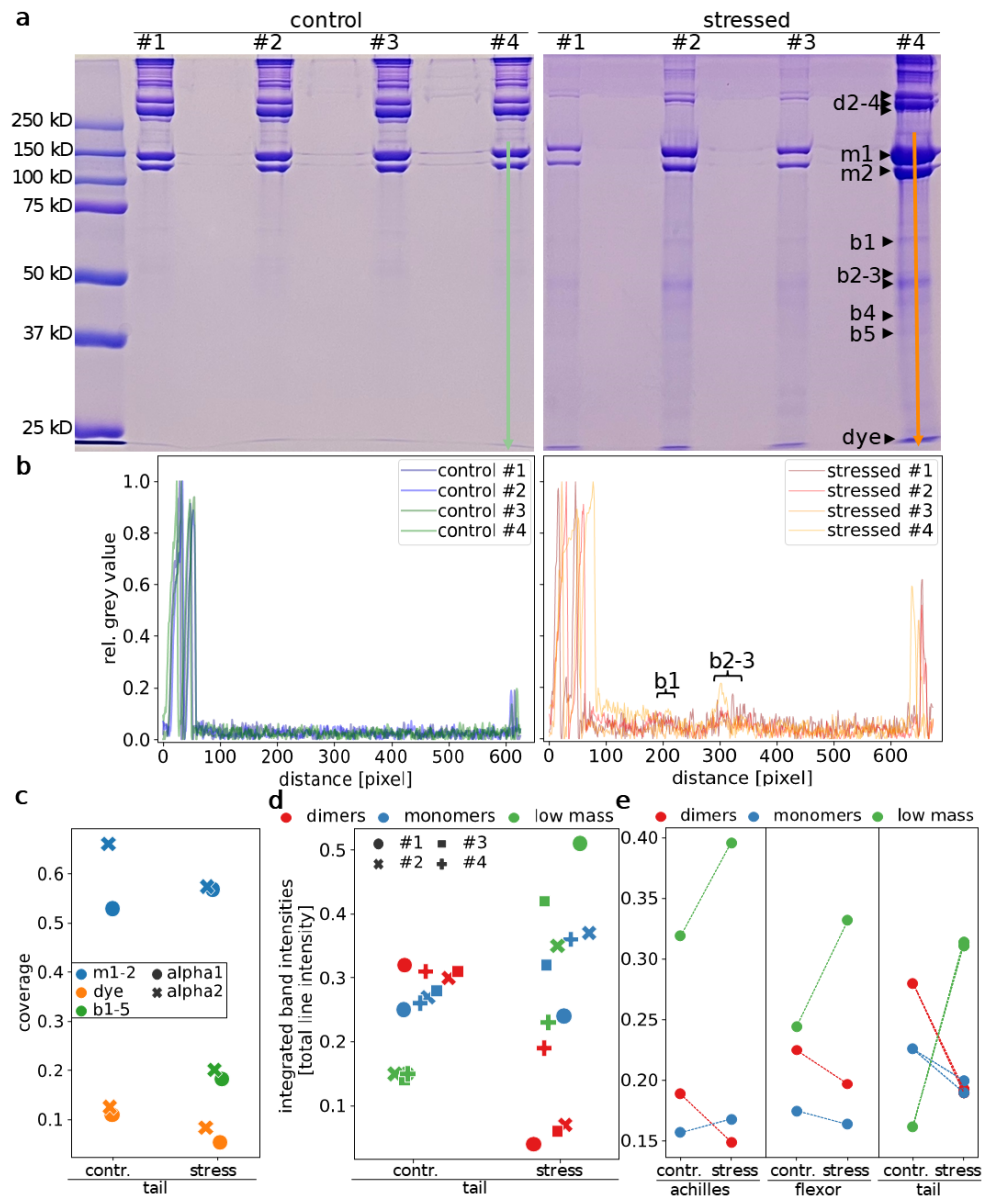
**Figure 4.10: Experiments support the simulations of covalent backbone and crosslink rupture in stressed tendon collagen. a** SDS-PAGE analysis reveals altered patterns of bands and their intensities between stressed and control tendon collagen. The experiment was repeated independently for a second biological sample set, see App. Fig. A.1. We indicate lines used for analysis in panel b in lanes #4. The bands b1-5, cut out for the mass spectrometry, are marked with arrows. **b** Densitometry to quantify stained collagen $\alpha$-chains and degradation products on SDS-PAGE. We show the normalized inverted grey value along the lanes. **c** Percent sequence coverage of collagen $\alpha 1$(I)- and $\alpha 2$(I)-chains from the bands analysed by mass spectrometry. Collagen fragments are present in the mass regime of monomers m1 and m2, low mass bands b1-b5 (averaged) and even the dye front. **d** Integrated intensity in the two monomeric bands (blue) and three dimer (red) bands, as well as the total area below the monomers (incl. dye front, green), normalized for each lane with the total intensity. **e** Same analysis as in panel d for additional gels provided in App. Fig. A.1 from Rat tail, flexor and achilles tendons.

supports the unspecific breakages found in the theoretical analysis (see Fig. 4.5c), as the stressed samples contained multiple $\alpha$-chains peptide fragments indicative of unspecific backbone breakages in the fiber. Given that the intensity of low molecular weight bands increases upon mechanical stress, we attribute this indeed to the mechanical loading.

Lastly, we used variations in the collagen protein mass patterns and their intensities as observed by SDS-PAGE as a qualitative measure of changes in collagen crosslinking. Here, we test our hypothesis of loss of connection between the molecules by rupture of crosslinking upon mechanical stress, as predicted by our simulations (Fig. 3b). Thus, we expect to observe fewer oligomers ($\beta$-, $\gamma$- and $\delta$-chains) and instead more monomers. Indeed, there is precedent for predicting alterations in collagen crosslinking through changes in patterns of bands in SDS-PAGE gels, whether caused by chemical treatment or genetic disorder [133, 134, 135]. It is apparent in Fig. 4.10a that the concentration of oligomers is decreased in the stressed samples. We quantify the integrated intensities of dimers and monomers in each lane in Fig. 4.10d, after normalizing to the overall intensity to account for the different signal strengths in the lanes. In the control, dimers account for slightly more of the total intensity in their lane than monomers. In contrast, after mechanical treatment, dimer intensity decreases, and the monomeric $\alpha$-chains make up a higher fraction in the stressed samples. Hence, stress on the tissue disrupted the crosslinks typically found in these higher molecular weight collagen polypeptides. Again, this is consistent with the crosslink breakages predicted from our simulations.

Together, the experimental evidence of both backbone rupture and altered crosslinking supports the detailed simulation results. The increase in lower molecular weight fractions is indicative of unspecific backbone ruptures, and the shift from oligomeric combinations to monomers can be explained by ruptures of crosslinks that otherwise held those oligomers together. In the simulations, the exact ratio of backbone versus crosslink rupture depended on parameter details as pointed out above, but that competition was always recovered. Similarly, which of the two mechanisms are more prevalent in the experiments remains yet to be clarified for future work and likely depends, among others, on species and tissue age.

# CHAPTER 5

# Collagen's Fibril Structure Buffers Mechanical and Oxidative Stress

Collagen is mostly known for its outstanding mechanical properties, which are attributed to its hierarchical structure. It spans multiple length scales and generates an interplay of molecular mechanism up to macroscopic winding of the fibrils. Experimental studies on this subject are naturally focused on larger length scales, while computational studies like Molecular Dynamics are more focused on collagen molecules or small fibril fragments, as our own work in the previous chapter. Some attempts on mesoscopic models [35, 26], or also on disordered network models [136, 137], have been made. None of these studies, to the best of our knowledge, combined rupture dynamics, as we obtained them from the atomistic simulations, with larger topological fibrillar features. Other (computational) work, for instance, puts more emphasis on the denaturation and unwinding of triple helices as a molecular failure mode in cyclic fatigue experiments [138, 139], or on fibrillogenesis [140].

In this chapter, we explore how breakages on the molecular level influence the subsequent reactions in the fibril. Particularly, we aim to bridge that length gap to the mesoscale mentioned above when it comes to simulations of collagen failure. In section 5.1, we identify that collagen failure modes buffer stress via a sequential rupturing mechanism: The region at the N-terminal crosslink always ruptures first, and never the C-terminal crosslink region. We ascribe this to the peptide length difference between the two otherwise fully equivalent crosslinks, which we find to be a conserved feature of collagen I in section 5.2. This sequential rupturing allows an additional energy dissipation by the release of hidden length, increasing the toughness of collagen for particularly high stress strands.

We present a new mesoscopic model in section 5.3, which allows us to explore the parameters involved in the failure mechanisms widely. We validate our bottom-up parametrization from MD simulations and are able to empirically fit the behaviour from the previous chapters. With this, we explore the properties more widely than any MD simulation would allow in terms of length and time scales, but also in terms of model architecture. We show that there is a trade-off involved when choosing the aforementioned path length differences. On the one hand, force transmission pathways with comparable lengths lead to equal load sharing and, hence, an overall decreased rupture propensity. However, these ruptures are then less controlled. On the other hand, an intermediate difference funnels ruptures into crosslinks, alike the sacrificial bonds in the previous chapter, but here on a larger length scale due to the resulting mechanical loading rather than due to the BDEs in the crosslinks. Even larger length differences, however, lead to an overall decreased survival time. We find our collagen models to be naturally located in a sweet spot of this trade-off, with a high parameter sensitivity enabling collagen to twist these properties to the specific needs of its various locations and functions. Overall, these properties of collagen rupture render the protein as an efficient buffer of mechanical and chemical stress in the stretched material, complementing our atomistic simulations.

## 5.1 Collagen breaks sequentially

We have seen in the previous chapter that collagen breaks at weak sacrificial bonds in crosslinks. The properties of the molecular bonds funnel the arising mechanoradicals into regions with redox-active residues like DOPA that can efficiently deal with the oxidative stress. Intriguingly, we observe another pattern with a similar effect on a larger scale when investigating the location of bond ruptures: Breakages inside the crosslinked region concentrate on one of the two crosslink sites in the overlap region, namely in the vicinity of N-terminal crosslink. In earlier studies, Kwansa et al. [25] already reported N-terminal crosslinks to be "early responders" to mechanical stress in simulations of a small system containing two collagen molecules connected by one crosslink each. This principle is illustrated in Fig 5.1, where the N-terminal crosslink (blue) is stretched more in the fibril direction. Here, we demonstrate that this property extends to being "first breakers" using a full fibril model containing up to 41 molecules in the overlap region, of which 8 are crosslinked on each site. We also find a molecular design principle behind this effect that is conserved among various species containing collagen.
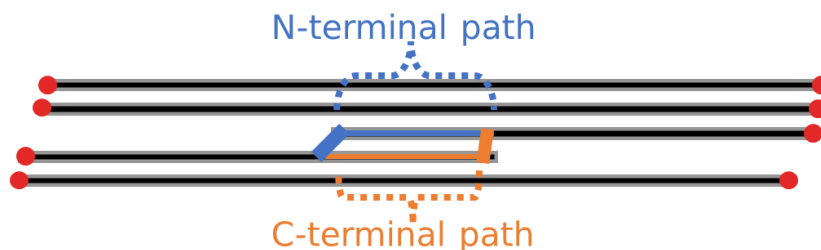
**Figure 5.1: In the overlap region, there are two force pathways due to the two crosslink sites at the N- and C-terminal regions**. In a schematic 2D representation, we show the N-crosslinks that connect the starting molecule in blue, as well as the arising N-terminal path for the force transmission. Their C-terminal counterparts are shown in orange. In red, we symbolically marked the caps on the other collagen molecules that would continue if not cut due to the finite model size.

While this trend is already visible in Fig. 4.5, we have quantified the data more clearly in Fig. 5.2 with respect to whether or not the rupture happens in a crosslink in panel a, and especially whether a crosslink on the N-terminal or C-terminal side broke in panel b. Models with trivalent PYD crosslinks, having the weaker link, show a very strong concentration of bond ruptures in the crosslinked region. This potentially beneficial concentration comes at the cost of higher total rupture rates. Interestingly, as panel b shows, the higher the rate, the more pronounced is the effect that it only breaks at the N-terminal side. For models with divalent HLKNL crosslinks, there are also two main observations. Inspecting panel a, it becomes clear that the funnelling into the crosslinks comes once more at the cost of higher overall breakages rates. As expected, in models where crosslinks have the highest rupture rates, they break first. In contrast to PYD, however, if crosslinks break, then it is in almost all cases exclusively the N-terminal ones.

Overall, the trade-off between stability (lower rupture rates) and specificity (concentration of rupture location) that we identified in the previous chapter looking at the strengths of individual bonds apparently also extends to the microfibril scale. Having the same chemistry and, hence, the same BDEs on the individual bonds on both sides, the differences in rupture propensities between N- and C-terminal crosslinks must arise due to a different load distribution, which we discuss in more detail below. Having crosslinks on one terminal that are more prone to breakages increases the rupture specificity, funnelling them even more into crosslinked regions than in models where the load is apparently more equally distributed.
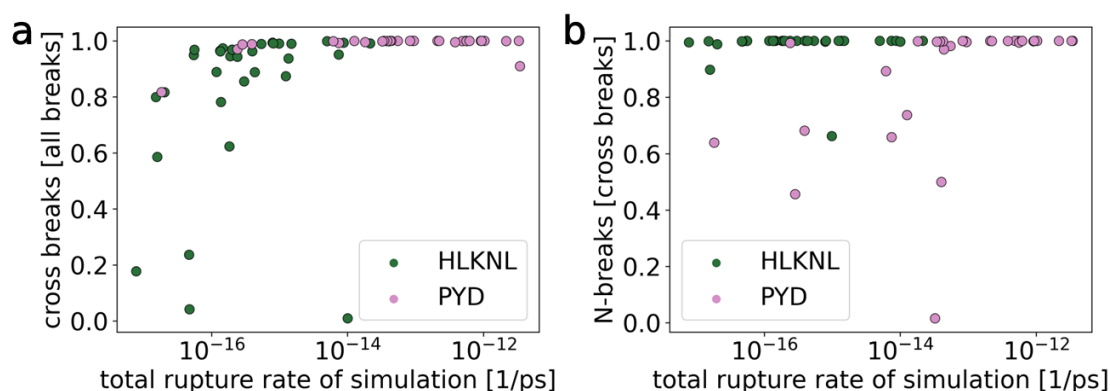
**Figure 5.2: Sequential rupturing: Collagen breaks at N-terminal crosslinks before C-terminal crosslinks**. We analysed again the data set of 63 KIMMDY simulations from the previous chapter, in particular with respect to the ratio of crosslink breakages in **a** and, especially, with respect to the percentage of breakages at N-terminal crosslink sites given the crosslink breaks in panel **b** – which have a much higher rupture propensity than the crosslinks at C-terminal sites.

## 5.2 Hidden length gets released after rupture

Next, we asked what happens after the N-terminal crosslinks have all been broken. To this end, we extended the previous data set with 18 more MD / KIMMDY simulations of collagen models: For each of the same three species as before (*Rattus Norvegicus*, *Pongo Abelii*, and *Loxodonta Africana*), we generated two models with ColBuilder that are crosslinked only at the C-terminal sites, one model with divalent HLKNL and one with trivalent PYD crosslinks. Using these, we conducted three replica each of force-probe MD and KIMMDY simulations with the same setup as described in section 4.2.1.

In Fig. 5.3a we display the aggregated and then normalized KIMMDY rupture distribution of these 18 simulations in the same way as before. Non-surprisingly, the concentration in the vicinity of remaining C-terminal crosslinks is even more pronounced than for the fully crosslinked models, compare Fig. 4.5. This confirms that we deal with a sequential rupturing mechanism, having the C-terminal crosslinks break after the N-crosslinks. Furthermore, we see in panel b that the end-to-end distances of our new models are about 0.75 nm more extended than our fully crosslinked models from before. As they are otherwise identical, this means that there is a release of extra 'hidden' length between the two connections that can be pulled out only when the N-crosslinks are broken. A small but reassuring observation is that the trends between species are
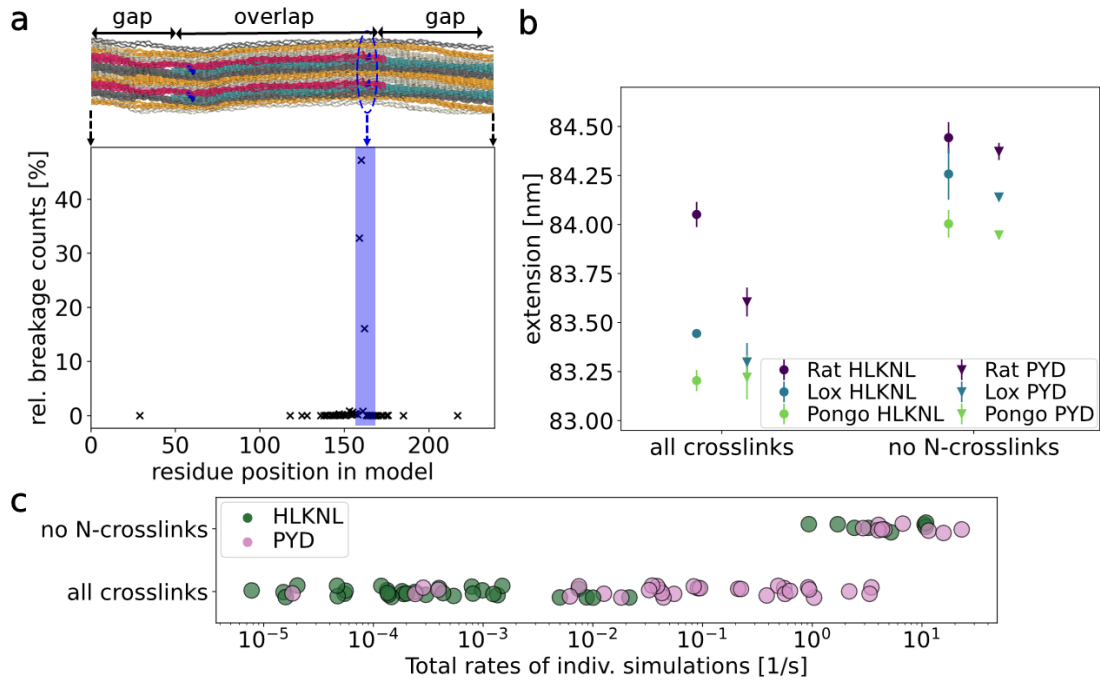
**Figure 5.3: Models without N-terminal crosslinks break at the C-terminal sites, at higher rates, and are more extended.** In a set of 18 pulling simulations with models that are only crosslinked at the C-terminal sites, we modelled the second stage of a sequential rupturing after the N-terminal crosslinks have already been broken. **a** In these models, the relative breakage counts according to KIMMDY strongly concentrate at the crosslinked region (blue) of the C-terminal sites. **b** Extension measured as end-to-end distance, averaged per model and crosslink type. Models not crosslinked at N-terminal sites are, compared to our previous data set ('all crosslinks'), on average more extended. Error bars are standard errors of the mean. **c** Total rupture rates in the individual simulations in models not crosslinked at N-terminal sites are on average higher, compared to our previous data set ('all crosslinks').

very consistent and, interestingly, there is even a difference between crosslinks types, with the more stiff trivalent crosslink leading to slightly smaller extensions. Considering what happens subsequently, we see in Fig. 5.3c that the total rupture rates (obtained by adding up all rates in that model) are orders of magnitudes larger in models without N-terminal crosslinks compared to our previous data set. On average, we obtain total failure rates of $5.7\,\mathrm{s}^{-1}$ and $8.5\,\mathrm{s}^{-1}$ for HLKNL and PYD models without N-terminal crosslinks, respectively. In contrast, in our fully crosslinked models, the average rates are $0.002\,\mathrm{s}^{-1}$ and $0.5\,\mathrm{s}^{-1}$, respectively. These are considerable differences for both crosslink types, while the change is smaller in the PYD models. There, the weak connections in PYD lead to higher overall rates and reduce the influence of the stress distribution on the rate. Nonetheless, these overall substantial differences suggest that the C-crosslinks get stressed stronger in the one-sided models than the N-terminal counterparts in the fully crosslinked models. In other words, in the fully crosslinked models, there has to be some load sharing between the two connections, even if one side is more prone to rupture. In contrast, having only one connection bearing all the force renders collagen less stable.

We shed more light on this idea of different paths and hidden length release in Fig. 5.4. The two overlapping pathways are shown in panel a, from a snapshot of a force-probe MD simulations, where we can already visually see a difference in strain of the crosslinks. The ones shown in blue, which are connected to the starting side of the new collagen molecule, lay down clearly more than the ones that connect the end of the C-terminal sites ending strands, shown in orange.
Quantifying this, we display the z-distance (fibril axis direction) between the $C_\alpha$ atoms of the crosslinks for both pathways for an exemplary simulation in panel b, averaged over the 8 connections in the model. Thus, this is a direct measure of the length of the two brackets indicated in panel a. We see that while the measured lengths get closer to each other under force, the C-terminal paths remain more extended than the counterparts on the other side. Similarly, for panel e, we calculated the z-distance between the beginning and ending of the crosslinks as a measure for how much they are strained and lay down in pulling direction. Clearly, the N-terminal crosslinks stretch both farther and also almost immediately, while the C-terminal crosslinks do not adapt as much to the apparently lower load. Still, as we also concluded from the data in Fig. 5.3c before, a (smaller) share of the load is carried by the latter side as well.

Counting the number of amino acids between the connections at the two sites, we

arrive at the histogram in panel c. For this, we evaluated all models that are available in ColBuilder [34] – currently 20 species (slightly varying in sequence) with different potential crosslink sites (only lysines in certain areas qualify for enzymatic crosslinks as described previously). Clearly, the connections via the N-terminal sites are shorter, with an average length of about 94 instead of 100 residues.

Furthermore, we see in panel d that we always get a non-negative outcome, when exclusively counting the combinatorial possible differences within models of the same species between the two paths. This means that the C-terminal pathways have always at least the same number of amino acids than their counterparts in each model (and are not only on average longer). There is an average difference of 6 to 7 residues as well, with up to 14 extra amino acids in some models. The GLY-X-Y sequence of collagen (with lysine as crosslink precursor only possible on X or Y) leads to the effect that these differences are less likely to occur in every third path.

Overall, Figs. 5.4c and d suggest that, for a length of about 100 amino acids between the two crosslink sites, there is a path difference of on average 6-7%. Note, however, that this way of counting is not always clearly defined. The N-terminal crosslink might connect to the first strand of the triple helix, whereas the C-terminal crosslink might bind to the second or third strand of the triple helix, and both strands can be shifted in relation to each other, as shown in Fig. 5.5. As only the side chains of glycines in the GLY-X-Y sequence pattern point to the inside of the helix, facilitating hydrogen bonding to side chains of other residues at the X- or Y-position, this shift is a well-known feature in collagen, as we also discussed in section 3.1.4 with respect to the chain registration.

Therefore, while the residue counting gives some indication of the potential difference in the connection length, we believe that the comparison of extension under force in Fig. 5.3b yields a more practical definition of the extra extension that is actually available: There is a difference of about 0.75 nm between the two models types. This length corresponds to about two (stretched) amino acids, so more than 2% in this about 100 residue-long section. We obtain a similar result by looking at the extra strain compared to the length of the overlap region between the two crosslinks of about 31 nm (see Fig. 3.2):

$$\Delta\epsilon = \frac{\Delta l}{l} = \frac{0.75\,\mathrm{nm}}{31\,\mathrm{nm}} \approx 2.5\% \tag{5.1}$$

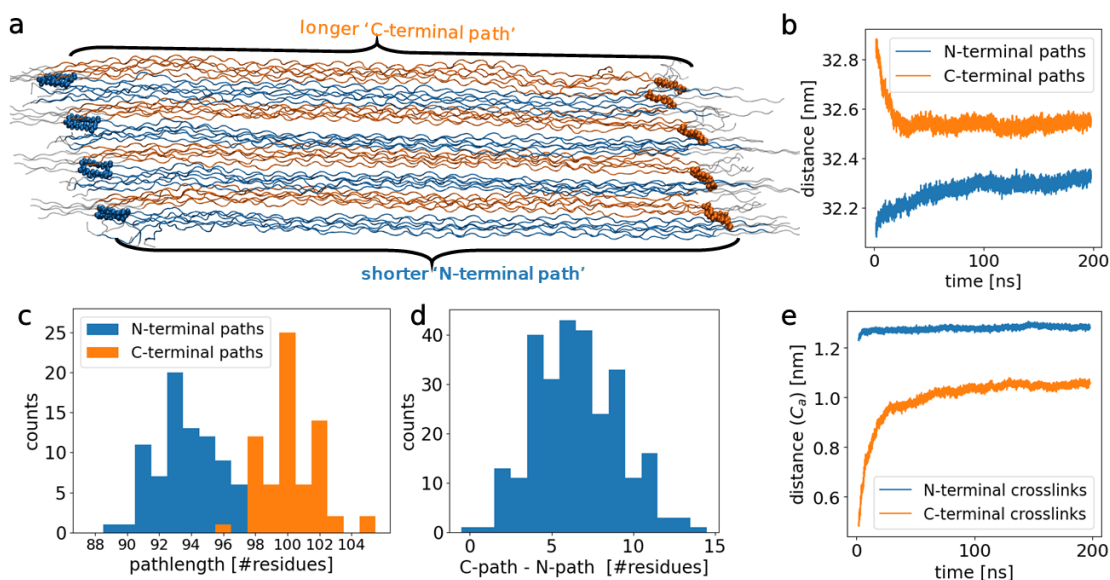We use these 2.5% as our best guess in the parametrization of the mesoscopic model below.

**Figure 5.4: Force pathways via N-terminal crosslink sites are shorter than via C-terminal crosslink sites in our collagen models. a** From a snapshot of a force-probe MD simulation, we show an enlarged view of the overlap region and have coloured all ending strands in orange (coming in from the left in grey) and the new triple helices in blue (continuing in grey to the right). We especially highlight the resulting two force transmission pathways via the crosslinks. Clearly, the N-terminal crosslinks lay down more than their C-terminal counterparts. **b** Length in z-direction (fibril axis) of the two pathways for one simulation, averaged over the 8 connections in that simulation. **c** Obtaining all models from ColBuilder [34] with all possible crosslink connections available, we count the number of residues between the two crosslink sites as a measure of the path lengths. **d** Taking the differences in the number of residues between pathways within all possible models, we obtain the path differences displayed here. **e** Crosslink elongation measured as z-distance of the crosslink $C_\alpha$ atoms in pulling direction, averaged over the 8 crosslinks per site for one simulation.
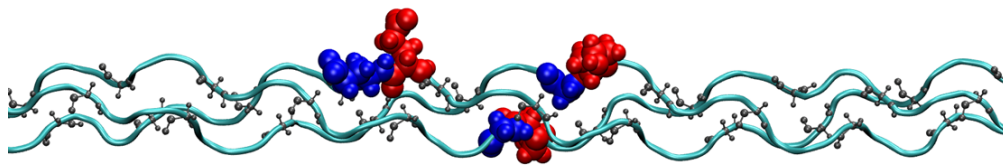
**Figure 5.5: Residues in the triple helix are shifted against each other.** We have coloured two exemplary residues in blue and red with the same residue ID per strand and can see that their position is shifted in relation to each other. This is to be expected due to the GLY-X-Y sequence pattern in collagen, of which only the side chains of glycine (shown in black cpk representation) can face inside the narrow triple helix.

Another important measure for a material like tendon collagen, which has to withstand high stress, is toughness [141]. It is defined as the ability to absorb energy (by deformation) per unit volume, without macroscopic rupturing. The release of hidden length $\Delta l$ is a generally known mechanism that allows energy dissipation $\sim F \cdot \Delta l$ and thereby toughens the material. To get a rough estimate of this effect in our collagen models, we approximate the volume of our fibril model that has 33 triple helices (i.e. 99 pulled strands) in the gap region as a cylinder with a diameter of about $10\,\mathrm{nm}$ (measured in the middle of the model with VMD [142]):

$$\Delta \text{Toughness} = \frac{\Delta E}{V} = \frac{99\mathrm{nN} \cdot 0.75\mathrm{nm}}{\pi \cdot (5\mathrm{nm})^2 \cdot 67\mathrm{nm}} \approx 14\,\mathrm{MPa} \qquad (5.2)$$

We consider this as an upper limit of the effect size, given that we assumed no other processes at work and a constant force. We have to keep in mind that our models are ideal, whereas imperfections and heterogeneities might lead to a stress concentration and earlier ruptures in the real material. In contrast, if we instead take the stress (roughly) estimated in previous macroscopic experiments from our group, in which the authors pulled with $20\,\mathrm{N}$ on collagen tendons with a cross-sectional area of about $1\,\mathrm{mm}^2$ [8], we reach:

$$\Delta \text{Toughness} = \sigma_{\exp} \cdot \Delta\epsilon = \frac{20\mathrm{N}}{1\mathrm{mm}^2} \cdot \frac{0.75\mathrm{nm}}{67\mathrm{nm}} \approx 0.22\,\mathrm{MPa} \qquad (5.3)$$

Nevertheless, given that there are multiple mechanisms throughout the hierarchy of collagen at work, this is still a considerable fraction of stress that can be taken up, which adds to collagen toughness.

## 5.3 ColBreaker: A mesoscale model for collagen failure simulations

Motivated by the above results, we present in this section our mesoscopic collagen model that shall extrapolate our findings to a scale of micrometers, for which to the best of our knowledge no simulations are available that combine the aforementioned topological details with bottom-up parametrized breakage behaviour. Due to the model's simplicity, we can explore more general design principles of collagen fibers by varying its key model parameters. To this end, the model has a particle-based approach utilizing the Smoluchowski equation as defined in eq. 2.31. Collagen molecules are segmented along the gap and overlap pattern, leading to 10 beads per triple helix. They are connected with different variants of worm-like chain potentials. Beads at ending molecules are connected with crosslink potentials to beginning molecules, such that we get a network as shown in Fig. 5.6. Both the crosslinks and the 'backbone' worm-like chain segments have the ability to rupture under force. As we use this model to investigate breakages in collagen, we named it *ColBreaker*, in analogy to our webserver ColBuilder that provides the atomistic models we described in section 3.1. We describe design choices, parametrization, performance evaluation and validation of ColBreaker in this section. We also discuss our ongoing work on a C++ version of the code with the help of software developer Dr. Bernd Doser. Once this is completed, we will make the code publicly available on GitHub.

### 5.3.1 Settings and interaction potentials

ColBreaker is an ultra coarse, particle-based model, in which the most important interactions are defined between the connected beads. We here discuss underlying physical potentials, their parameters, and our implementation of them into ColBreaker. Of course, other settings like length (in units of gap/overlap regions) and width (in units of beads in the hexagonal cross-section) can also be adjusted. A list of all implemented settings can be found in the appendix A.4.

To connect the beads in ColBreaker, we made several choices for the underlying potentials for both the collagen triple helix and the covalent crosslinks. The overall setup with the key properties highlighted can be seen in Fig. 5.6. The fibril is aligned along the z-axis, with a slight tilt angle such that the hexagonal grid in the cross-section and the resulting network can be seen. Furthermore, in order to get a better visualization
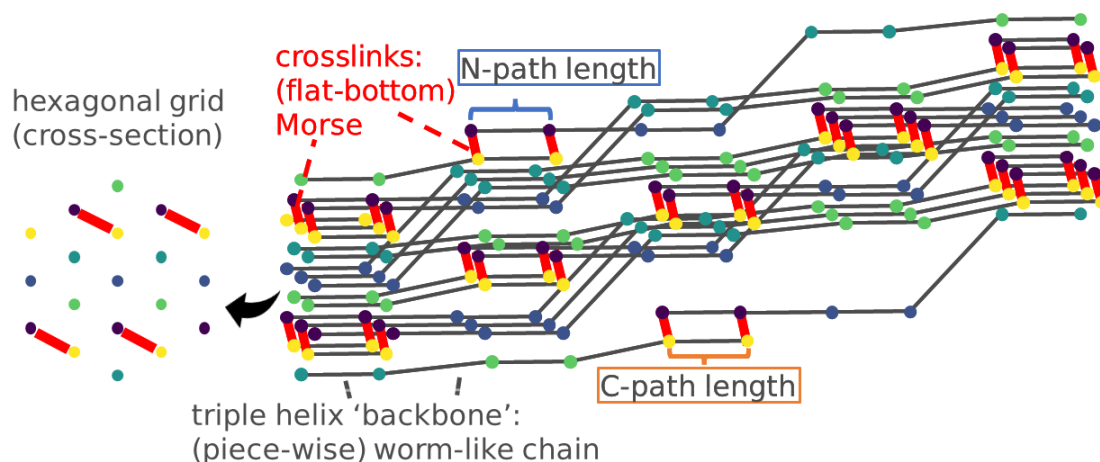
**Figure 5.6: 300 nm ColBreaker section, with main potentials and parameters choices highlighted.** We depict a model that contains at least a few full collagen molecules (which are 300 nm long) such that we have a fully packed hexagonal grid both in the first and in the last layer. The grid of the first layer can be seen in a cross-sectional view on the left. Note that the fibril axis is condensed here compared to the cross-sectional plane, in order to make the characteristics of this otherwise too long fibril visible. The starting beads (dark purple) are connected to the ending beads (yellow) of other molecules via the crosslink potentials (red). The connection between the beads, i.e. the backbone, is modelled with variations of the worm-like chain. We are also able to adjust topological parameters like the length of gap and overlap region and, especially, of the marked contour lengths of the different (N- or C-terminal) pathways between the connected beads in the overlap.

of the gap/overlap features, we display the fibril axis condensed in comparison to the x/y-plane in the figure. On the left, we also show the first layer in a cross-sectional view. In this small model for illustration, we choose a width of three triple helices per side of the hexagon. Beside the two potential types marked in the figure that we discuss in more detail below, we emphasize that the contour length between the overlapping segments can be adjusted. This is to test the effect of varying distances between the crosslinks, marked in the figure as N-path or C-path length, respectively, that can result in differing force transmission.

## 5.3.2  Topology of the ColBreaker fiber

A decisive feature of collagen is the twisting of collagen molecules that we elucidated in section 2.1.1. Much like a rope, the strength of a fibril can be influenced by how the individual triple helices wind around each other. We refer to this here as the topology of

ColBreaker. In MD simulations, the topology, among other things, provides information about which atoms are bonded to each other and, thereby, describes the network of atoms. Here, we extend this notion by the relative position of strands and how they wind around each other with respect to the position on the quasi-hexagonal grid that collagen is known to exhibit in its cross-section [46]. This grid is the building block of ColBreaker, as in each gap and each overlap region we place exactly one bead per triple helix there, see also Fig. 5.6 on the left. The connections between the beads, analogous to bonds, then define the course of the triple helices through theses 'phases' of an individual collagen molecule.
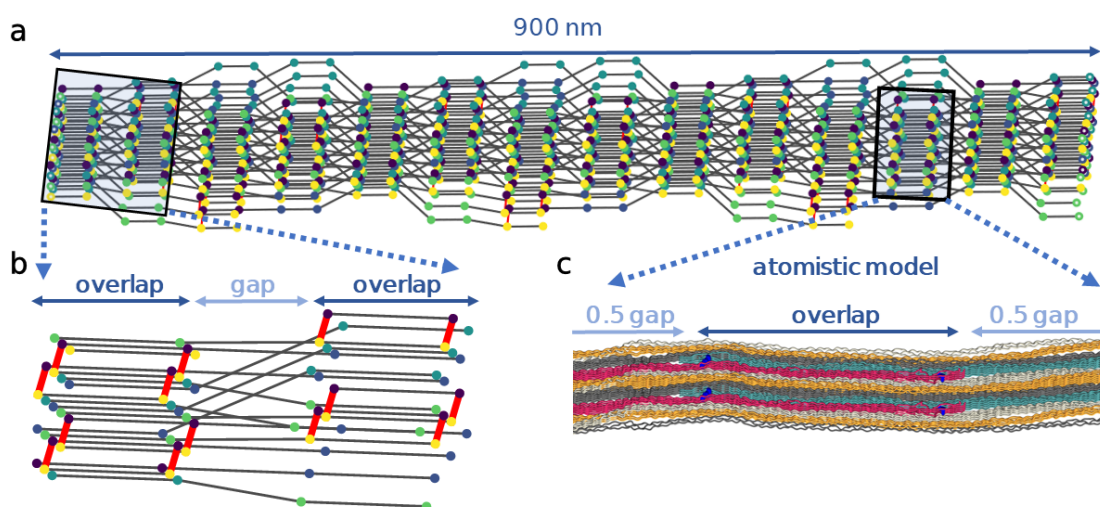


**Figure 5.7: Comparison between the topology of a ColBreaker fibril and our atomistic model. a** We depict a ColBreaker model for a 900 nm long fibril, which is about the length of 3 full individual collagen molecules. Note that the fibril axis is condensed here compared to the x/y plane, in order to make the characteristics of this otherwise too long fibril visible. In each gap and overlap region, there is a layer of beads connected via potentials that mimic the collagen triple helix. In the more dense overlap region, they proceed parallel to the fibril axis. In contrast, the twisting and change of relative positions happens in the less dense gap region. The beads are arranged on a hexagonal grid in the cross-section. **b** The details can be seen on the enlarged (and, for clearer visualization, thinned out) zoom-in onto the first two overlap regions. The red crosslinks connect beads in the last phase (yellow) of a molecule to ones that have just started (violet). **c** For comparison, we show our 67 nm long atomistic model that has one overlap region in the middle and one half gap regions on each side.

The 3D topology of ColBreaker follows the X-ray fiber diffraction data of Orgel et al. [21],

as do our atomistic models, but we also allow different crosslink configurations, as discussed in section 3.1.4. An example of a 900 nm ColBreaker configuration can be seen in Fig. 5.7. To enable a representation that fits a page, we once more had to shrink the fibril axis compared to the cross-sectional plane, such that this view is distorted and exaggerates the characteristic features along the fibril. The immense size difference between this model and our 67 nm atomistic microfibril (also shown in Fig. 5.7c for comparison) underlines the need for such a coarse description. Whereas we hardly incorporate the twisting of triple helices around each other in our atomistic models, we here can easily reach scales that include the 3D network topological effects up to the micrometer scale.

In Fig. 5.7a, the course of the fibril is displayed, starting with the beads on the quasi-hexagonal shape in the overlap region. Once placed, they can move freely in the z-axis (fibril direction) in the simulation, with drift and diffusional forces acting on them. We have coloured the beads according to the 'phases' in the braiding, starting from dark violet over shades of blue and green to yellow – in this way, it is easier to follow the course of one connected triple helix in this convoluted situation. As a result, it is always a yellow bead at the end that is connected to a dark bead at a beginning triple helix via the red crosslinks.

In the overlap, the triple helix 'backbone' connections (black lines) are parallel to the fibril axis. In the gap region, the aforementioned twisting leads to connections to the next layer at different positions – see the zoom-in in Fig. 5.7b (note that it is turned a few degrees to enable a view into the cross-section). For a detailed view of the braiding of individual helices around each other on the molecular level, compare this again to Fig. 2.2. The twisting is also the reason for the perceived widening of the fibril diameter, as the triple helix does not stay within the same position of the hexagonal cross-sectional grid during its course through the fibril. Note that the resulting edge effect of a not fully packed configuration on the outside of the fibril is not relevant here, as we do not include non-bonded interactions between the strands at this point. An alternative approach would be to fill the arising holes with additional fragments of triple helices that would get cut-off when they go out of the boundaries. Hence, they would be shorter than the 300 nm of a full collagen molecule and (at least for now) only serve cosmetic purposes, which is why we leave this for future work if it might be needed. We also like to point out that there is an upward shift of the fibril with respect to the z-axis, if one aligns it parallel to the strands in the overlap region and perpendicular to the hexagonal

cross-section. We have chosen this orientation in our implementation, as it appears to be a natural choice and facilitates the grid-based building of the starting configuration (see Fig. 5.6, but note that the effect appears exaggerated due to the condensed z-axis in the figure). The upward shift arises due to the way the helices are connected and intertwined in the scheme proposed by Orgel et al. [21]. Alternatively aligning the fibril along its course, as done in Fig. 5.7a for better visualization, leads to the fact that the cross-sectional hexagonal planes of the beads are tilted. This is a known feature of collagen that we also observe in our MD simulations when steering the fibril with external force (see Fig. 4.3). Considering the different lengths between the two crosslink connections, this outcome appears even more plausible: force is transmitted via the crosslinks such that, while the upper strands are stretched more on one side, the lower strands stretch more on the other side, enhancing the observed tilting.

## 5.3.3 External pulling forces

Collagen has to withstand large mechanical loads, which we can simulate in ColBreaker by pulling with an external force at the outer beads of those triple helices that would continue if not capped due to the finite size of the model. In other words, we apply force on all beads in the last layer except the ones that belong to ending strands. We have implemented three protocols of steering simulations in ColBreaker. Analogous to what can be done in the atomistic simulations with GROMACS, the first two options are 'constant force' or 'constant velocity' pulling. For constant force, one simply adds a constant external force to each bead (with ColBreaker also having an optional gradual build-up phase to increase the force level more smoothly in the beginning). In the second case of constant velocity pulling, a virtual spring with spring constant $k$ is extended at a constant speed $v_{\mathrm{pull}}$ and drags on the beads.

In addition, we came up with a third protocol that is not available in our MD engine. We term it 'constant extension' or 'constant strain'. It mimics a situation in which tendons get stretched up to a certain extension. This protocol is also implemented as a virtual spring, which gets pulled until it reaches a certain position $x_{\mathrm{max}}$:

$$x_{\mathrm{spring}} = x_0 + \max(v_{\mathrm{pull}} \cdot t, x_{\mathrm{max}}) \tag{5.4}$$

This scheme combines several advantages, making it more faithful to the biological situation: We can keep up a certain force $F_{\mathrm{max}} = k \cdot (x_{\mathrm{spring,max}} - x_{\mathrm{bead}})$ to obtain meaningful rupture rates and then have a relaxation after rupture. After a breakage

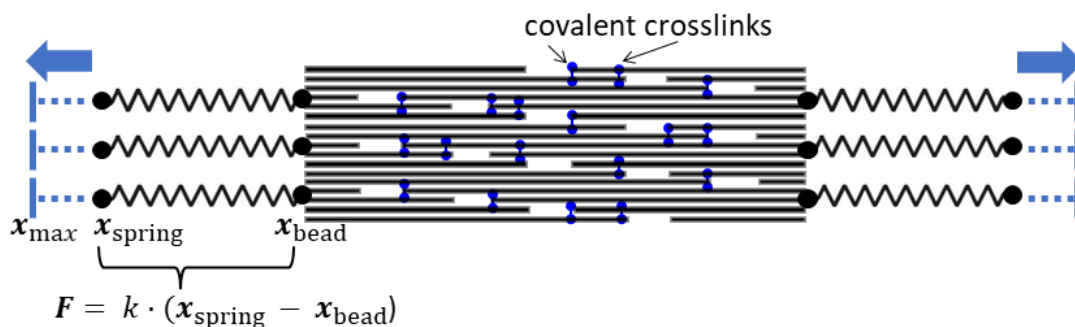$$F = k \cdot (x_{\text{spring}} - x_{\text{bead}})$$

**Figure 5.8: ColBreaker: Constant extension pulling as a steering option.** We have implemented the default constant force and constant velocity pulling options known from GROMACS and, in addition, the here displayed third steering algorithm. In there, a virtual spring at the outer beads gets extended with a certain rate. Having reached a pre-defined maximum, that extension is kept, even if the position of the beads change – leading to varying force levels depending on the extension of the model. This mimics the situation of a tendon being pulled to a certain elongation.

inside one of the two crosslinks, the broken strand can slightly extend due to the release of extra length between the connections, and the resulting force decreases (as the distance to $x_{\text{max}}$ is lower). In contrast, the other setups keep up the forces leading to higher strains after these micro-ruptures and do not lead to the expected relaxation.

## 5.3.4 Crosslinks as flat-bottom extended Morse potentials

For the crosslinks, which are (beside the backbone) one of the two modelled key interactions, we decided to use a Morse potential (see eq. 2.10 for definition), similar to the individual bonds in the atomistic simulations. Under force, the crosslinks first lay down in the MD simulations. To mimic this, we added a flat bottom part of length $r_{\text{fb}} = 1\,\text{nm}$ before the Morse potential. If the actual bond is stretched more than another $r_{\text{cut}} = 0.25\,\text{nm}$, the crosslink breaks. This distance is sufficiently larger than the inflection point of the Morse potential, after which the restoring force monotonically decreases to zero. Overall, the resulting force as a function of the distance between the two beads then reads:

$$f(r) = \begin{cases} 0 & r \leq r_{\text{fb}} \\ 2\beta E_{\text{dis}} e^{-\beta(r-r_{\text{fb}})}(1 - e^{-\beta(r-r_{\text{fb}})}) & r_{\text{fb}} < r \leq (r_{\text{fb}} + r_{\text{cut}}) \\ 0 \text{ (crosslink broken)} & r > (r_{\text{fb}} + r_{\text{cut}}) \end{cases} \quad (5.5)$$

In essence, we coarse-grain the several atomistic crosslink bonds together into one effective potential. For the barrier, we chose the weakest link with the lowest bond dissociation energy $E_{\text{dis}}$. We saw in section 4.1 that typically the weakest bonds in the crosslinks have a much lower barrier than the others, such that we can neglect the influence of other bonds breaking here. We validate the crosslink breakage behaviour together with the backbone against KIMMDY data in section 5.3.6.

## 5.3.5 Collagen backbone as extensible worm-like chains with Morse breakage barrier

Another key question for building a mesoscopic collagen model is how to approximate the triple helices. In the literature, there is a wide range of both experimental and computational studies, and also reviews thereof [97, 98], that compare collagens to worm-like chains (WLCs) or similar models, which we introduced in section 2.4.1. Particularly, we are interested in the force-extension curves of stretched collagen molecules. For a 'bottom-up' comparison, we once more conducted constant force MD simulations, this time of both our full fibril model (67 nm long, containing 41 triple helices) and of an individual 67 nm long triple helix, at various force levels for at least 100 ns each. For low pulling forces, where fluctuations and bending of the elastic rod play a lager role compared to the stretching, longer simulation times were necessary to obtain converged end-to-end distances in pulling direction. Thus, for the single triple helix, a larger range of forces could be explored due to the smaller system size.

The obtained MD force-extension values for the fibril and triple helix are displayed in Fig. 5.9 as the orange and dark red data points, respectively. We show a linear scaling on the upper panel and a logarithmic depiction on the lower one, respectively, to resolve both the more relevant high force regime and the typical WLC characteristics in the lower force regime. A direct comparison of the two MD sets shows good agreement, with the full fibril yielding slightly larger extensions at identical force levels; as expected due to extra length that can be gained by crosslink extension and by a relative shift of helices with respect to each other.

Another observation from Fig. 5.9 is that collagen follows the classical WLC only up to a relative extension of about 5% to 10%, before indeed switching to the linear regime of enthalpic stretching that we discussed together with the theory of WLC and alike models combining both entropic and enthalpic effects in section 2.4.1.
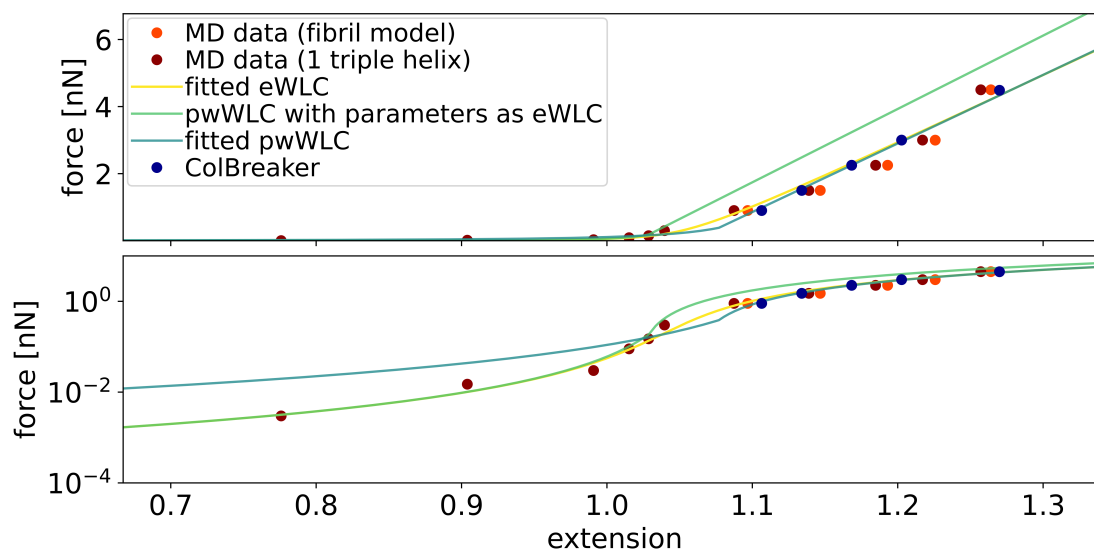
**Figure 5.9: Comparing force-extension curves of worm-like chain models to MD data of collagen under force.** We collect the average extensions, projected in pulling direction, at a given constant force for different simulation setups. Dark red and orange data points are obtained from constant force MD simulations of our fibril model and of an individual 67 nm long triple helix, respectively. To these, we fit the eWLC (yellow), calculate the pwWLC with the parameters obtained from the eWLC (light green), and again fit the pwWLC focussing on the upper force range (green-blue). The parameter set obtained from the latter is utilized for ColBreaker. Thereafter, we compare these to the extensions of ColBreaker at the same force levels (dark blue). The lower panel has a logarithmic scale for the force, showing the same data.

Next, we fit three variants of the discussed WLC-based models to our MD data. First, we use the extensible worm-like chain (eWLC) by Odijk [100] that is most accurate, but does not have an explicit force-extension formulation and can, hence, not be used directly in our code. It reproduces the MD behaviour well over the full force range, as can be seen in both the linear and the logarithmic depiction (yellow line). With this, we obtain a persistence length $l_p$ of 4.39 nm and a contour factor $c_f$ of 1.07 for the worm-like chain regime, as well as a spring constant $K_0$ of 21.88 N/m. This is consistent, though on the very low end, with elsewhere reported collagen persistence lengths [98]. Secondly, using these values as input for the piece-wise defined model (pwWLC, light green line), we regain a good agreement in the lower force regime, but with an offset for high forces. The two curves overlap until an extension of about 1.07, where the kink due to the piece-wise formulation is visible, especially in the logarithmic scaling.

Finally, fitting the pwWLC (to the non-logarithmic values) with a least-squares algorithm naturally focuses more on high forces, which is the most relevant regime for our purposes. With this, we obtain the third parameter set (green-blue line). Whereas these parameters do not accurately reproduce extensions at low forces (and, hence, the fit yields meaningless values of 0.47 nm and 1.16 for persistence length and contour factor, respectively), it is especially the upper linear part after the kink that is decisive in our model. This section is now more accurate, without the previous offset that the second parameter set would cause. This procedure yields, therefore, our final parameter set, in particular the effective spring constant $K_0$ of 20.45 N/m that we use for ColBreaker.

As a consistency check, we then employ ColBreaker and get back a force-extension curve (dark blue points) matching the pwWLC. On the one hand, as an end-to-end test of the code, this validates that our implementation yields back the input. On the other hand, it shows that we can get a good overall agreement with the MD data, despite this very simplified description of collagen.

## 5.3.6   Modelling the breakage behaviour

In chapter 4, we identified a competition between crosslink and backbone rupture in our KIMMDY simulations. Above, we have already introduced the flat-bottom Morse potentials for the crosslinks in ColBreaker, which break once the potential energy barrier is crossed. So the next question arises on how to implement the failure of a worm-like chain mimicking the rupture of a backbone bond in the collagen triple helix. Initially, we tested a simple cut-off of the potential such that the WLC can sustain a maximal force level. Adjusting this cut-off, we are able to reach a reasonable ratio between crosslink and backbone rupture for a fixed force level. However, when varying the stress, the two rupture processes reacted in different ways: The WLC potential is a lot softer than the crosslinks, leading to a large reaction distance $\Delta x$ that the force can act against – the resulting shift in the underlying energy landscape (see the Bell model in Fig. 2.7) is hence much larger and the rupture rate much more force sensitive.

In Fig. 5.10, we compare different backbone breakage implementations in ColBuilder with reference KIMMDY data. For the KIMMDY part, we used single 67 nm (one gap and overlap region) spanning parts of collagen triple helices under varying levels of external force; continuing with KIMMDY the simulations that we used for the MD triple helix data in the force-extension curve in Fig. 5.9. We added up all the individual bond
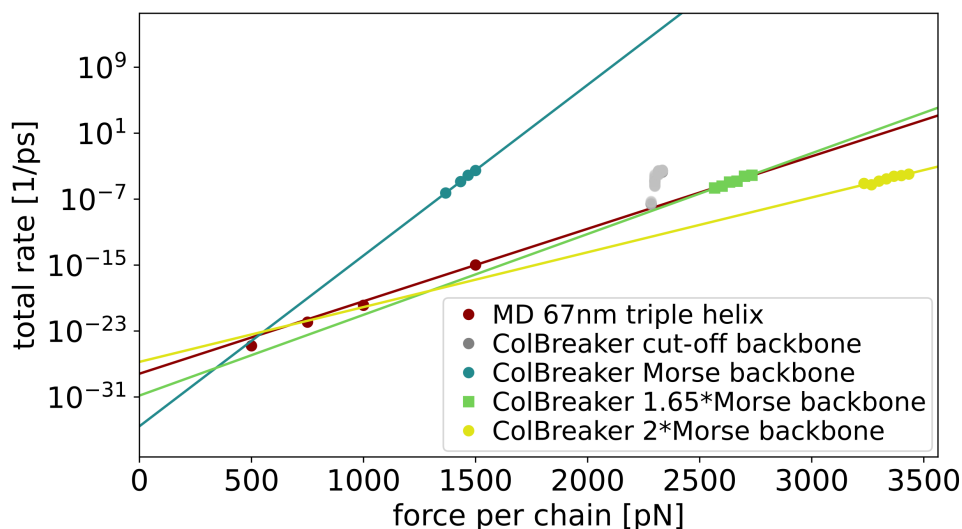
**Figure 5.10: The force-dependency of backbone rupture rates of ColBreaker can be fitted to MD data.** We use the summed-up KIMMDY rupture rates for a 67 nm triple helix and fit a log-linear relation (dark red line) as reference value. To this, we compare different rupture implementations of ColBuilder: A simple cut-off (grey), a Morse potential at the end of the WLC with the strength of an individual covalent bond (blue), with twice that strength (yellow) and, as best fit, with a increased strength of factor 1.65 (green). The ColBreaker rates are obtained as the inverted average failure times of 5 simulations per data point.

rupture rates to obtain a total failure rate of that segment. We compare this to the average breakage times of a ColBreaker set-up with the same length (3 beads connected by two WLCs, so one overlap and gap as well), conducting 5 simulations until failure for each force level. According to the Bell-model described in section 2.3, we expect a linear relation on the logarithmic scale of in this curve, or maybe a slight bending according to the more elaborate Dudko-Hummer-Szabo model.

The aforementioned initial approach in ColBreaker with a simple cut-off yields the data shown in grey in Fig. 5.10, for one exemplary cut-off value. As we can see, the force-dependency does not fit the slope at all, and it is not possible to cover a range of force values due to strong dependency of the rates on the force. The reason behind this is that the lowering of the effective energy landscape is proportional to $F \cdot \Delta x$, and the reaction distance $\Delta x$ is much larger for a soft worm-like chain potential than for an individual covalent bond with a steeper Morse potential.

To mimic the actual breakage process more closely, we updated our backbone potential with a new breakage barrier. In the high force regime, we define another switching point $x_{\text{switch}}$ at a certain force $F_{\text{switch}}$ (per default at $4\,\text{nN}$) where we change to a Morse potential after the pwWLC. This force should be lower than the maximal force given at the inflection point of the Morse potential, and we shift the potentials relative to each other such that the transition is continuos in the force level:

$$F_{\text{pwWLC}}(x_{\text{switch}}) = F_{\text{Morse}}(x_{\text{switch}} + x_{\text{shift}}) \tag{5.6}$$

$$\overset{!}{=} F_{\text{switch}}. \tag{5.7}$$

In this way, the beads do not reach the cut-off via a soft ramp any more, but have to cross a steep bond potential equivalent to what happens in the crosslinks and the MD simulation. Using the parameters of a standard $C_\alpha$-C backbone bond for this last part, we obtain the blue data points in Fig. 5.10. Now, an overall more sensible force vs. rupture rate relationship appears, but at comparatively high rates. This can be explained by the fact that we used the equivalent of a single bond for the whole triple helix, whereas in the MD simulation three strands would share the force. Analogous to having three springs in parallel, we expect the force to be shared and the effective potential to be stronger. Another point worth noting, though working in the opposite direction, is that we also have to consider that in our atomistic system we have many more bonds in a row per backbone that could all break (but have to withstand about the same force level). Thirdly, the idea of springs is idealized and, in practice, the total rupture rate is dominated by individual rates of bonds where force concentrates more due to heterogeneities or simple differences between the amino acids, as discussed in section 3.2.

For these reasons, we implemented an empirical strength scaling factor of the Morse potential. From the above considerations, it should be somewhere in the range between 1 and 3. For comparison, we show in yellow a Morse potential with twice the strength and then, finally, our current best fit in green. For this, we enhanced the barrier by a factor of 1.65. We still see a slight deviation in the fitted slope to the MD data, but also like to note that this is within the given accuracy, as we only cover a small fraction of that graph. Having the faster C++ code at hand in the future, we will strive to improve upon this by covering 2–3 orders of magnitude more in rates. For now, we abstained from spending more resources on this already reasonable reproduction of the KIMMDY data, considering the simplicity of ColBreaker.

Lastly, we combine both backbone and crosslink rupture propensities and compare their relative occurrences. For the crosslinks, we set a BDE of 296 kJ/mol for the barrier height in the Morse potential, which is the lowest value in the divalent HLKNL crosslink obtained from the QM calculations that we have seen in section 4.1 [38]. Using a path length difference of 0% and 3% close to the hidden length release from the previous chapter, we conduct ColBreaker simulations of a 300 nm fibril with the constant strain pulling scheme, to an extension of 35% per side – a force regime that we found to lead to occasional bond rupture inside the computationally accessible time scales. However, we noticed that now the crosslinks are comparatively weak against the backbones, breaking much faster even without a crosslink path difference. The above fitting of the backbone accounted empirically for heterogeneities and imperfections in the backbone, but the crosslinks still have the parameters of an idealized bond. In the MD, in contrast, they would only lay down up to a certain angle, force might distribute into other degrees of freedom such as angles, and also other more complex effects of the fibrillar network can lead to a reduction in force uptake. We did not consider these points in the above fit to a simple collagen triple helix for the backbone. In order to take these factors into account, we iterate once more over the empirical strength of the backbone parameter, to come back to the competitive rupture regime between backbones and crosslinks that we observed with KIMMDY in chapter 4 for a wide range of parameters. We could have adjusted the crosslink potential instead, but opted to work again with the backbone strength in order to keep the number of empirically fitted parameters as low as possible. In the end, an enhancement of the backbone strength by a factor of about 1.45 instead of the above 1.65 turns out to be most reasonable. Again, we verify this in more detail by comparing the results in section 5.4 to the MD data from chapter 4. Whereas this rather iterative empirical procedure shows the difficulties in the bottom-up design of such a model, we also underline that already by only considering the simple triple helices as reference we observe similar rupture characteristics that only needed to be fine-tuned to reproduce the MD data.

### 5.3.7 Performance and time stepping

Having everything parametrized in ColBreaker, we compare its performance to our MD simulations. The first version of ColBreaker is implemented in Python due to the language's simplicity during the development phase. We used just-in-time compilation libraries that convert certain parts of the code to a C-based execution, but are still limited

in speed by this choice of programming language. We reach about 70 ns/day single core performance for a "small" ColBreaker system of a 300 nm fibril. This system has 374 beads with 72 crosslinks. It corresponds roughly in size to the 45 million atom system that our Japanese collaborators currently run on FUGAKU. There, we reach 12 ns/day with an impressive number of 768 nodes having 48 cores each. Hence, we need about 215,000 times more core-h in MD for the same simulation time as in ColBreaker.

Additionally, we have recently started a project together with software developer Dr. Bernd Doser from HITS to port the code into C++. As cross-functional scientific programmer, Dr. Doser takes up such projects at our institute. Whereas he is dealing with the code conversion and technical details, we provide ongoing feedback and testing of the code. Preliminary results indicate a possible speed-up of up to three orders of magnitude (so up to a factor 1000x). Additionally, parallelization will make it easier to utilize several cores on compute clusters. This will enlarge the usage options of ColBreaker significantly, as a much wider parameter space can be explored.

Another reason why we do not reach even faster simulations despite this very coarse setup are our choices for the underlying energy landscape. In other coarse-graining schemes, like Martini [70] and others, a huge part of the speed-up is achieved by increasing the time stepping, e.g. up to 20 fs for Martini, which is possible due to the much smoother underlying energy landscape. However, in order to recover the realistic breakage behaviour, we use the aforementioned Morse potentials comparable to the all-atom MD simulations. This comes at the cost that they are also as steep as in the atomistic force field, and small changes in positions of the beads can lead to large changes in the resulting forces. We saw when experimenting with the time step that too large time steps can lead to (falsely) increased breakage counts, because the beads in this case simply jump over the steep barrier with a large integration step. Hence, small time steps are needed, and we stick to 2 fs, equivalent as in the MD simulations performed in this thesis.

## 5.4 Collagen's fibril structure buffers mechanical stress

With the sequential rupturing of the atomistic microfibril in mind and ColBreaker developed, we wondered how our results scale up to larger lengths scales. We have seen in Fig. 5.4 that the force pathways via N-terminal crosslinks are always shorter than their C-terminal counterparts in our models that follow the experimental X-ray data. A

major advantage of ColBreaker, beside the scalability to larger length scales, is that we can easily explore parameters like this path difference. Further, we can compare this to, for example, homogenous models where equal paths would share the force, which we do not have at hand atomistically.

## 5.4.1  Path length differences funnel ruptures to crosslinks

To explore the influence of the connection lengths on the rupture propensities, we conducted a set of simulations varying the contour factor of the triple helix sections between the two crosslinks in the overlap region, from 0% difference (so both sides equal) to 3% (close to the 2.5% extension after removal of N-crosslinks) to 6% (about the average difference when counting residues as above) for a 900 nm long collagen fibril. To keep proportions, this model is also thicker in the cross-section (5 triple helices on each of the 6 sides of the hexagonal grid), such that we reach 1714 beads with 338 crosslinks in total. This is large enough to include effects of the 3D network, in contrast to the MD system that only has 16 crosslinks (for a comparison of both models, see again Fig. 5.7). We use our strain pulling setup and extend the virtual spring to 20.5% percent on each side such that we reach an initial force level of around 6 nN per triple helix with occasional rupture within our accessible simulation time.

In Fig. 5.11, we see the breakage counts for several independent runs of each of these three models. Each simulation yields one timeline for the N-breaks, C-breaks and backbone breaks, respectively. The total simulation times vary, as we used several workstations for the computations (to avoid being cut by the wall-time limit of our high-performance cluster). When the crosslinks share their load equally, as in panel a, only backbone breakages occur. Introducing a 3% difference in the contour lengths (panel b), we observe a competition with many more crosslink ruptures. As both N- and C-terminal crosslinks in the same strand break almost immediately after each other, the green and blue lines overlay (we consider a different scenario in the next section, however). We have seen in chapter 4, that a typical share of crosslink breakages would be in the range of 65% to 80%, or even 80-90% if we include backbones in the vicinity. This fits very well to the data in panel b, where in each run 6 to 12 crosslink breaks happen for 0 to 2 backbone failures. As we have adjusted our parameters to reach this competition, the agreement, thereby, validates our settings.

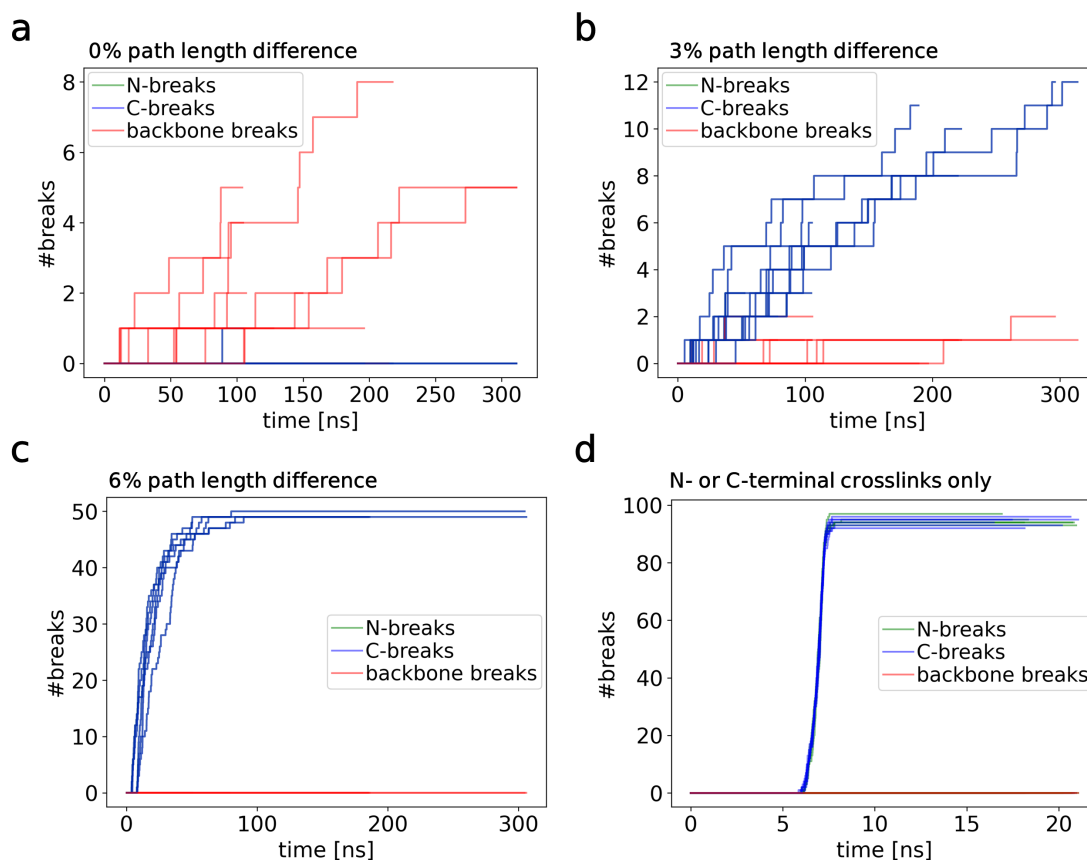When we look at the breakages obtained with the ColBreaker model that is closest to

**Figure 5.11: ColBreaker shows that path length differences funnel ruptures into crosslinks, in the compromise of an increased overall rupture count.** We pull with our constant extension scheme on a 900 nm ColBreaker fibril and monitor breakage events with regard to their location: In the triple helical backbone (red), N- or C-terminal crosslinks (green and blue – note that in cases of overlap, only the blue data is visible). Each line represents an independent simulation with the same starting conditions, simulation time per run varied due to available computational resources. **a** In a model setup with equal path lengths between N- and C-terminal crosslinks, only backbone ruptures occur. **b** In a setup with 3% path difference, close to our atomistic models, we see a competition between less backbone and more crosslink ruptures. **c** With a 6% difference in the contour lengths, only crosslinks break, and also much faster. **d** In models with crosslinks only on one side (either N- or C-terminal), we observe immediate ruptures, after the build-up phase to the same force levels (note the different x-axis).

the MD properties in Fig. 5.11b, and compare it to a hypothetical baseline of an equal path model in panel a, we can see that the total count of breaks is higher in panel b, but the number of breaks in the backbone is greatly reduced. This shows that the path length difference can funnel the breakages away from unspecific failure somewhere in the fiber to a more localized and controllable environment in the crosslinks. We refer to this as breakage specificity in regard to location, as this is critical for the radical scavenging properties of collagen, with the DOPA residues built-in nearby crosslinks. In panel c, we see that when we increase the difference even further to 6%, there are only crosslink ruptures, but at a much higher rate. So this parameter offers a high sensitivity in adjusting the load share, and our simulations suggest that collagen might be in the sweet-spot of gaining control over the rupture location, at an as low as possible expense of increasing the overall rupture count. Lastly, see panel d, we compare this to a situation with crosslinks at only one of the two sites (so either N- or C-terminal) and observe that this would lead to immediate failure, at least at these high force levels.

Comparing our ColBreaker simulations to the MD results (Fig. 5.3), we see that there are very similar mechanisms at play. The atomistic models without N-terminal crosslinks have total failure rates that are several orders of magnitude larger than the fully crosslinked ones (Fig. 5.3c), and we here observe a similar behaviour of almost immediate failures for this case in ColBreaker (Fig. 5.11d). Our atomistic models, while more prone to rupture on one side, therefore must have some (uneven) load-sharing between the two connection sites. This holds true for ColBreaker as well. Additionally, with ColBreaker, we can compare this to the scenario of equally shared burdens on both crosslinks as well (Fig. 5.11a). The latter scenario prevents the crosslink rupture, but instead leads to the unspecific and potentially more harmful backbone breakage.

To conclude, the seemingly small difference of a few percent in contour length offers a strong leverage on adjusting the breakage count and specificity, and we suggest that this could act like a set screw as one mechanism for collagen to regulate its function depending, for instance, on tissue type or on location, such as in low-stress positional versus high-stress energy storing tendons [39].

## 5.4.2   Hidden length release can reduce ruptures

In the exploration phase of this study, we managed to manipulate the ColBreaker parameters such that we observe a more pronounced sequential rupturing mechanism

than in the previous section. In this regime, which we do not reach in the bottom-up parametrization above, the path length differences not only funnel breakages into crosslinks instead of backbones, but we also see an increased survival due to a hidden length release and a resulting relaxation. Subsequent to breaking on the N-terminal sites, the C-terminal crosslinks are still strong enough to survive considerably. This is possible due to a force decrease after rupture, as in the constant strain implementation the extra length that was formerly hidden in the longer connection leads to a smaller distance between pull atoms and their virtual pulling spring and, hence, a lower applied force.

In Fig. 5.12, we see an example of how this can look like. In panel b, we utilized a path length difference of 8% in a 300 nm fibril. The 8% length difference corresponds to the upper average of our simplified counting residues approach. At that stage of data acquisition, we used the simple cut-off for the backbone rupture and adjusted it for that specific force level to match the backbone vs. crosslink rupture ratio from KIMMDY. With this, we observe a sequential rupturing to unfold. First, the more stressed N-terminal crosslinks break (green line), as they experience a higher force due to the shorter contour length that connects them to the next bead. After rupturing, the hidden length, i.e. the length difference to the other longer connection, gets released in the constant strain pulling scheme. In turn, this leads to a small stress reduction in collagen strands with a N-terminal break. Even though the C-terminal crosslinks (blue line) in these have to bear the remaining force subsequently alone, they still survive for a significant time. This can be seen in the difference between the green and blue line of N- and C-breaks in Fig. 5.12b, respectively; contrary to the scenarios in the previous section, where both lines always overlay. Additionally, we again see occasional backbone rupture (red line).

However, these unspecific backbone breakages are more prone in panel a, where we do not have any path length difference. There, both crosslinks share an equal amount of force and do not rupture. Since the force levels in the strands stay high until a backbone rupture occurs, the unspecific ruptures are more likely and their number is comparatively higher. Lastly, in panel c and d, we compare this to setups in which we only have a crosslink on one of the two sides, respectively. In both cases, they break quickly. Overall, we see yet again that there is a sweet spot in the trade-off between strength (panel a) and specificity of location of ruptures (panel b) that would not be possible with a simple one-crosslink setup (panels c and d). Particularly, the hidden length release and resulting reduction in stress can increase the survival time in panel b while maintaining the control on rupture location.
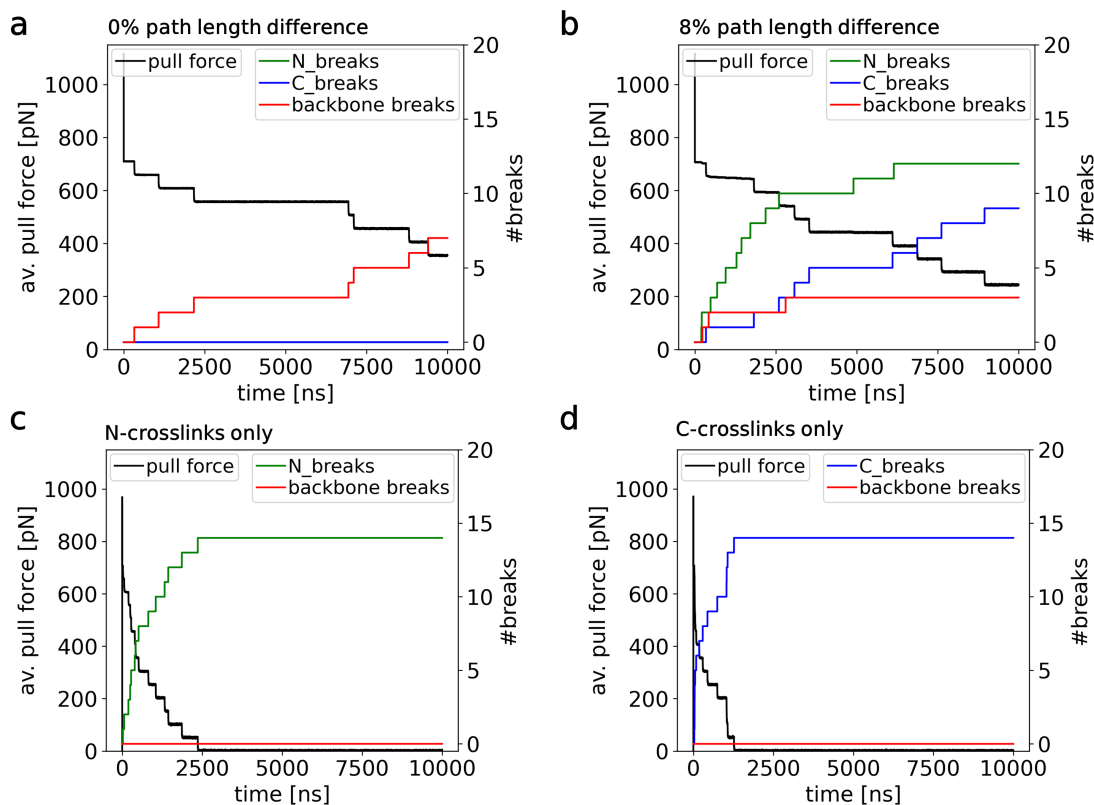
**Figure 5.12: In an exploratory parameter regime, a hidden length release enables a sequential rupturing in ColBreaker that buffers mechanical stress.** We show average pull force (black, left y-axis) and rupture events (right y-axis) depending on their location in N-terminal crosslinks (green), C-terminal crosslinks (blue), or backbone (red), in ColBreaker for a 300 nm model. Here, we used the hard backbone cut-off at a low force level. The force is applied through our constant strain scheme, leading to stepwise relaxations after each rupture. **a** In a model without path difference, we only observe backbone breakages. **b** In this model, the N-terminal path is 8% shorter than the C-terminal one, leading to N-crosslinks breaking more quickly, but also to a reduction in backbone rupture. The C-terminal crosslinks survive longer than their N-terminal counterparts, such that the strands overall rupture later. **c, d** In models with N- or C-terminal crosslinks only, respectively, the survival time is very low.

We speculate that even though we do not reach this regime in the above bottom-up parametrization, it still might be physiologically relevant. There are so many collagen types in various species, as well as within various tissue types, having different crosslinking chemistry, that surely our atomistic models cover a relevant but not exhaustive part of this natural broad spectrum. For example, stronger crosslinks combined with a larger path difference could enable this, which was how we started this exploration. And even if these did not exist in nature, this is still an advantage of ColBreaker, complementing the detailed MD simulations with insights not only into mechanisms of one parameter regime, but also on how manipulating certain factors could influence the results. Thus, ColBreaker furthers our understanding of their contributions and interplay.

# CHAPTER 6

# Discussion

In this thesis, we investigated tensed collagen type I fibrils by means of various scale-bridging simulations and dealt with the identification of breakage sites, the creation of mechanoradicals and their implications on the material due to subsequent reactions. On different scales, we discovered and elucidated mechanisms on how collagen is designed to break in controlled manners that help to deal with the mechanical and arising oxidative stress — adding to the innovative perspective of rich mechanochemistry in collagen upon the established simpler mechanical picture of collagen as a strong rubber band.

As experiments alone cannot provide the necessary level of detail to understand the underlying molecular processes of mechanoradical generation in tensed collagen, we employed scale-bridging hybrid simulations to obtain a comprehensive picture on the rupture propensities in collagen. To achieve this aim, discussed in chapter 3, we first set out to build and improve the necessary (molecular) simulation framework.

A precondition for the MD simulations is an atomistic collagen type I microfibril, as close as possible to native tissue. Note that we limited the computational part of this study to the most common collagen type I, but find decisive features like crosslink sites and DOPA positions to be highly conserved in other minor tendon collagen types III and V [38]. In our group, all-atom models of crosslinked collagen type I fibrils were built and are described here in section 3.1. They are accessible via ColBuilder [34], a web resource of collagen models for 20 different species with a large variety of crosslink types and localization within the fibril, coming along with parameters of the crosslinks for an Amber force field. Since their publication, the models have already been used by other groups [28, 143], proving their usefulness to the community, though to our knowledge nobody else conducted comprehensive failure simulations of the crosslinked fibrils yet.

In this thesis, we especially validated the models with MD simulations and saw that they dynamically explore a conformational space that is close to X-ray fiber diffraction data and that is also in line with other experimental observations (gap/overlap ratio and heavy-atom labelling). However, given the lack of high-resolution data for collagen I and the high disorder of the telopeptide regions, we note that our atomistic models in this region represent only one of the possible conformations that allow for accommodating divalent as well as trivalent crosslinks, and the structure of this region should be interpreted with caution. Apart from validating the atomistic models, we support the current experimental efforts in our group to obtain new high-resolution structural data by cryo-EM with structural comparisons to our MD models. Moreover, we explored possible variations of crosslinking both for future models and for our own ultra coarse model.

Beside the atomistic modelling, we also dealt with improvements of our method toolkit in chapter 3, foremost of our KIMMDY scheme that enables bond ruptures in all-atom MD simulations. We observe the rupture propensity of covalent bonds in collagen to be determined by a competition between how weak a bond intrinsically is, that is, low BDEs, and how much force acts on this bond, that is, the stress distribution in the fibril. We find the latter to depend on computational details of the force calculation in KIMMDY, as we investigated in great detail in section 3.2. Atomistic force fields are suitable to take up the conformational changes of collagen and force distribution throughout the protein, but we also saw the limitations when it comes to the exact split into the different interactions and degrees of freedom of a molecule. This work was important to show that our approach is appropriate for our purpose. Additionally, it enhances our understanding of the applicability (and the limits) of classical force fields to model molecular forces, with ongoing further investigations in our group on the quantum-mechanical level.

The KIMMDY approach is more efficient than hybrid quantum mechanics / molecular mechanics (QM/MM) techniques [144, 145], especially given that the initial rupture sites are unknown and the whole molecule can potentially react. Further, having now detailed BDEs at hand, it is also more accurate than reactive force fields [31], which would also not scale well to these large atomistic models. KIMMDY operates in the (macroscopic) sub-failure regime of collagen with occasional internal bond ruptures, which would otherwise not be computationally accessible. This regime is consistent with previous experiments showing radical generation for time scales of seconds to

hours [8], and can, for instance, be seen as a common scenario of heavy tendon loading without overall fiber rupture.

Albeit more efficient than QM/MM, KIMMDY is still limited to sub-micrometer-scale atomistic systems. Our atomistic simulations thus can not take structural features beyond this length scale into account, such as heterogeneities in crosslinking or flaws that can aggravate stress concentrations [8]. We reach most of the strain already after 10-20 ns and assume a separation of time scales with (un-)loading times beyond the rupture kinetics. We did not consider fast non-equilibrium processes like a catastrophic fiber failure. Further, we limited ourselves to examine the initial micro-ruptures, and KIMMDY is not yet able to deal with bond (re-)formation. Both processes could be a possible extension of KIMMDY for future work. A particular strength of KIMMDY is, however, that it takes structural rearrangements as well as effects of internal friction on the distribution of forces through the structure and thus on the resulting rupture kinetics directly into account. It therefore is well suited to study ruptures in stretched collagen, which has previously been shown to feature conformational changes on various structural levels, from proline ring flips to helix rearrangements and beyond [28, 99, 146, 147].

In addition to bond ruptures in collagen, we paved the ground for a KIMMDY extension with hydrogen atom transfer (HAT) reactions, a natural follow-up after homolytic rupture, by a proof-of-concept simulation presented in section 3.2.4, and by guiding the ongoing developments. It is also exiting to see that machine-learned energy reaction barriers by others in our group are now getting incorporated into KIMMDY [119]. Advancing the limits of computational feasibility, we further saw in section 3.3 that a more than 300 nm long fibril model, containing for the first time full collagen molecules, comes in reach even at the atomistic level. In there, more heterogeneous crosslink set-ups can be probed, and a first analysis of the ongoing simulations shows that we can reproduce structural features from our smaller previous models, but it also hints at intriguing possible differences in the models chosen with varying crosslink densities.

Considering the advances in computational power, modelling, and the ongoing methodological extensions, KIMMDY appears to be a promising method for further future work.

Following the hierarchical structure of collagen, we then focused on the molecular level in chapter 4, nailing down rupture locations in collagen for a large, representative set of collagen models. For this, we varied species and crosslink selections, but also the pulling setups in our simulations. We incorporated a new set of bond dissociation energies into

KIMMDY, yielding rupture statistics, which in turn were validated by experiments.

Notwithstanding the sensitivity of quantitative results with regard to the widely explored parameters, a robust picture emerged: Crosslinks harbour the weakest bonds. In particular the captodative effect stabilizing $C_\alpha$-centered radicals, and, in trivalent crosslinks, the additional resonance with the aromatic ring give rise to exceptionally low BDEs. Hence, they break most often. However, the stressed protein backbone still breaks randomly as well due to stress localization, but also simply due to the vast amount of bonds under force. The lower the difference of backbone and crosslink BDE, the higher the share of backbone ruptures in that competition. Likewise, we could detect both backbone ruptures and altered crosslinking in the SDS-PAGE gels of tensed collagen, compared to untreated samples. The picture emerging from the experiments with different Rat tendons (tail, flexor, achilles) corroborates the simulation results. However, we also note that future experimental work is needed for a more quantitative confirmation of the predictions.

Protein $C_\alpha$ radical stabilization by the captodative effect was found before, e.g., when inducing radical formation via UV light [148], or by mass spectrometry experiments [149]. These studies support our results, and we here extend the decisive role of the captodative effect in proteins to mechanical bond rupture and mechanoradical formation in a stretched protein material. The stable methylpyridine radical in trivalent crosslinks extends the set of aromatic moieties known to stabilize radicals in proteins, namely histidine, tyrosine, tryptophan, and DOPA, as e.g., in ribonucleotide reductase [53, 54, 150], or collagen [8]. A DOPA radical, however, is even more stable, with a BDE for hydrogen abstraction (X-H) as low as 260 kJ/mol for a DOPA anion [151]. We determined a BDE(X-H) of around 337 kJ/mol for the deprotonated PYD [38], hence, in the context of the collagen structure, radical migration to DOPA seems likely from a thermodynamic viewpoint. A recent study from our group experimentally confirmed the presence of DOPA, often nearby crosslinks, in different collagenous tissues and its radical-scavenging properties [9]. DOPA is derived from either tyrosine or phenylalanine, and the accumulation of these amino acids in the proximity to the crosslinks is eye-striking, as shown in Fig. 6.1.

Our data propose in particular several advantages of the unique trivalent crosslinks of collagen. First, its rupture results in exceptionally stable radical species, reducing unspecific reactions such as uncontrolled migration. The stabilizing effect is enhanced by the localization of breakages, as low crosslink BDEs funnel ruptures to sites where
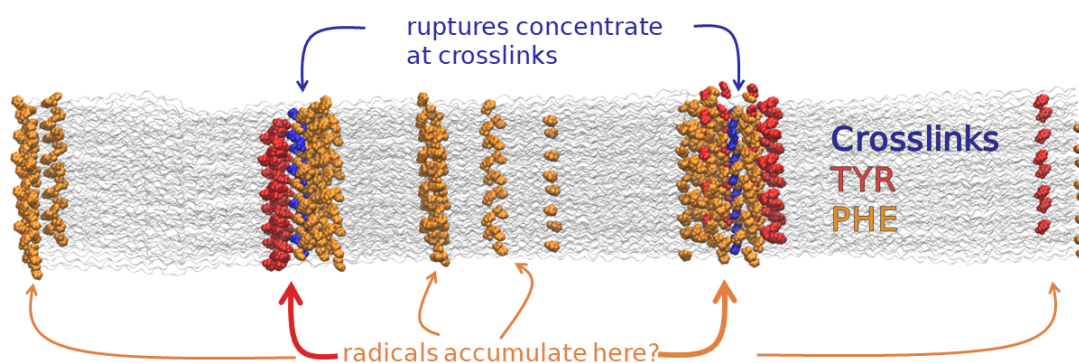
**Figure 6.1: Breakage points and DOPA precursors are in vicinity of each other.** The DOPA pre-cursors tyrosine and phenylalanine are in close proximity to the crosslinks, which have a high rupture propensity, as we saw throughout the chapters in this thesis. This novel picture of an interplay between mechanical and chemical stress indicates that collagen evolved in order to efficiently deal with the arising oxidative stress.

a rapid transfer to nearby DOPAs is likely. Thus, further damage to the material is mitigated. Secondly, we find the mechanical strength of the trivalent crosslink to be pH-sensitive, i.e., BDEs depend on the protonation state of PYD. We speculate that collagen's rupture propensity could adapt to variations in physiological pH, e.g., in muscle after exercising [152, 153] or under inflammatory conditions [154]. Thirdly, we find the weak bonds within the trivalent crosslinks to act as sacrificial covalent bonds. Stress relaxes in the trivalent crosslink after initial rupture, and subsequent ruptures occur in other crosslinks, according to our computational analyses. Mechanically, the sacrificial arm in the trivalent crosslink can increase the toughness of the material due to energy dissipation and relaxation afterwards, similar to what is known from sacrificial bonds in polymers [19]. Finally, radicals formed in crosslinks can migrate rapidly to nearby DOPAs and then in presence of water react with superoxide radicals to create $H_2O_2$ as has previously been shown in pulling experiments on collagen fibers [8]. In this way, the breakage of one of the two arms offers a possibility to already start signalling high loads in collagen, triggering repair mechanisms and other reactions, even before macroscopic failure occurs. Speculating further, a self-healing effect similar to what can be observed in polymers [155] or biological fibers [156] could also come about by either direct DOPA recombination or by the formation of DOPA-Lys and DOPA-His crosslinks [157, 158].

Collagen's remarkable trivalent crosslink chemistry appears perfectly adapted towards

handling mechanical stress and mechanoradicals, and as such can inspire the design of synthetic (biocompatible) material systems with high toughness, self-healing, or self-reporting strategies. Our results presented in chapter 4 propose that collagen might have evolved into a high-performance biomaterial by means of highly specific mechanochemical routes that enable collagen to self-report on rupture while preventing further unspecific damage. If this specific mechanochemistry is similarly at play in other biomaterials such as elastin, and to what extent it plays a role in tissue ageing, remains to be elucidated.

Turning our view from individual molecular bonds to the microfibril structure in chapter 5, we observe a sequential rupturing mechanism in crosslinked collagen fibrils, with the crosslinks at N-terminal sites having the highest rupture propensities. Much like the individual sacrificial bonds, it is here the whole connection that funnels ruptures again into the DOPA-rich regions. In contrast to the molecular bonds, however, it is not the difference in BDEs, but the level of load-sharing between the two crosslinks that is the decisive adjusting screw of the arising competition. We see that while one side is more prone to rupture, a fraction of the load is still carried by the second connection. Without it, the overall failure rate is significantly increased.

As aforementioned, the concept of sacrificial bonds is widely known in synthetic polymer networks. In collagen, only sacrificial ionic bonds have previously been proposed in bones, enabling energy dissipation through the release of hidden length on an intermolecular scale [159, 18, 19]. As discussed, we observe here similar mechanisms for (much stronger) covalent bonds on the intra-molecular scale inside trivalent collagen crosslinks in chapter 4 and extend this in sections 5.1 and 5.2 to the whole crosslink acting as a sacrificial connection between molecules.

Interestingly, it has been reported elsewhere that collagen toughness can increase upon cyclic stretching of reconstituted collagen fibrils, especially at strains above 20% [160]. Whereas the authors speculate that the underlying reasons might be molecular sliding and breakage of hydrogen bonds, our results indicate that sacrificial covalent bonds and, potentially, their reformation after a recovery phase, could also contribute to this strain-induced strengthening, especially at such high loads. We propose that the difference in breakage propensity between the two crosslink sites enables collagen to act as a mechanical buffer. In fibrils with a particular high mechanical load, a bond scission in one connection can release additional hidden length of the other longer pathway,

thus increasing the material's toughness. Given that force probably distributes more heterogeneously in imperfect nature than in our idealized models, this mechanism specifically benefits very highly stressed strands for which elastic deformation is already on its limit and which otherwise might fail completely. In this regard, these strands can relax compared to their neighbours and maintain the overall integrity of the fibril.

To investigate this on larger length and time scales, but also to topologically explore this trade-off further than our atomistic models allow, we developed ColBreaker as presented in section 5.3. Utilizing our knowledge from MD and bottom-up parametrizing this mesoscopic model, we are able to build a description based on piece-wise worm-like chains for the triple helices, connected in the characteristic 3D topology via crosslinks, and with the option to enable both backbone and crosslink breakages. We acknowledge that in order to mimic the rupture process accurately, we resorted to the Morse potentials as in the atomistic simulations leading to a steep energy landscape with the necessity of, unfortunately, the same small integration time steps as in MD simulations. Nonetheless, we can scale our simulations up significantly both in time and, importantly, in many length scales, even without the prospect of a much faster C++ version of the code. We reproduced force-extension curves from MD simulations in the relevant high force regime, as well as regained the competition between backbone and crosslink ruptures that we saw above in the KIMMDY simulations.

Using this new tool, we lastly obtained several complementing results on the fibrillar level (see section 5.4). We see that the contour length of the connection between the two crosslinks is, indeed, a set screw that can adjust the ratio of backbone versus crosslink ruptures with a high sensitivity. Governed by the difference in force transmission pathway lengths, our collagen fibrils can adapt in a trade-off between strength (more equal load sharing) and specificity in breakage (higher difference in forces), with all the advantages of that localization that we discussed above. With ColBreaker, we are able to compare this to hypothetical scenarios that are not accessible in the MD simulations and obtain a more comprehensive picture. Models with fully equal connections only lead to unspecific backbone ruptures, models with a very large length differences fail much faster, and models with crosslinks on one side only cannot even sustain the same force levels. Hence, all of these have obvious disadvantages compared to the natural model setup, which suggests that collagen has found its sweet-spot to best provide its functionality in tensed energy-storing tendons. The obtained rupture simulations are consistent with our atomistic results, where we compared fully crosslinked models to

C-terminal crosslinked variants, the latter breaking much faster. ColBreaker enables us to gain an understanding of the influence of manipulating such properties. We can investigate a wider range of collagen variants or, going further, even explore how nature could potentially have looked like, if evolution did not optimize collagen in the way we uncovered here.

Other experimental work has suggested that crosslink rupture, followed by molecular sliding, might be a mechanism to explain force-extension curves of collagen type I fibrils [161]. Furthermore, it has been speculated that even complete subsections of a fibril might be sacrificial areas or at least more susceptible to mechanical damage, with the origin of material heterogeneity to lie in different crosslink densities [162]. With our work, we provide mechanistic insights on the molecular level to these experimental hypotheses. Even if after N-terminal crosslink rupture also the secondary C-terminal crosslinks broke, and the fibril would fail locally, it would do so in a more controlled manner. This design can help collagen to maintain control with respect to the rupture location and the arising mechanoradicals, having the scavenging DOPA residues nearby the preferred breakage sites.

ColBreaker yields by its nature only a simplified perspective, with rich opportunities for further enhancements. Next future steps, beside a more exhaustive parameter exploration with the faster C++ code, could include to build 'trivalent' crosslinks, e.g. with a double-well energy barrier, that rupture sequentially and thereby mimic the molecular sacrificial bonds we found. Another extension could be to incorporate more interactions between the triple helices, for instance, with an adhesion term to neighbouring strands. In the convoluted fibrillar topology, these non-bonded interactions could contribute to the overall mechanical properties, for example by redistributing force to other strands. However, they would also increase the computational cost.

Moreover, we aim to use the option to switch the crosslinking to alternative connection sites (Fig. 3.4). We have already implemented it, but it requires larger models to avoid molecular sliding of the unconnected parts of the network and, hence, more computational time (or a faster code). With this, we are planning to shed light on a more heterogeneous force transmission network, but also on how the fibrillar arrangement for the model building could be handled outside the limited options that the experimental study by Orgel et al. [21] from 2006 propose. Another desirable extension is to add advanced glycation endproduct (AGE) crosslinks, which are more randomly located, in contrast to the enzymatically derived crosslinks, for which only certain positions

are possible. ColBreaker can enable us to study the arising variety efficiently, whereas this would require much larger, if not prohibitive, efforts in atomistic simulations. The content of AGE crosslinks is known to increase with ageing, and we speculate that the resulting heterogeneities in the force distribution could render collagen more prone to rupture, maybe also in a less controlled manner throughout the fibril than with only enzymatic crosslinks.

## 6.1  Summary

Overall, we show across various length scales that the conjunction of topology and chemical structure of collagen by design funnels bond ruptures towards crosslinks and specific sacrificial bonds therein, very much akin weak sacrificial bonds in synthetic polymers. We contribute in several ways to the available computational toolkit, with respect to both methods specifically tailored towards collagen (such as the development of ColBreaker) and to more general schemes (such as reactive MD with an improved version of KIMMDY). With these, we achieve our initial aims to identify rupture propensities and underlying mechanisms in collagen, and even find several design principles behind. Together with previous findings that potential sites of the radical scavenger DOPA are highly enriched around the crosslinked regions [8], also across different tissue types [9], our combined results put forward collagen as a biomaterial with highly specific mechanochemical routes, a feature which likely prevents further unspecific damage by radicals. Beside their direct mechanochemical implications, these findings might also guide the development of synthetic collagen-like materials or serve as nucleation points for further investigations, e.g. on radical migration and subsequent signalling, onto this newly established route of conversion between mechanical and oxidative stress in collagen.

# Appendix

## A.1 MD simulation parameters

The following table contains the parameters which have been used in the MD simulations that were conducted in the course of this thesis, if not otherwise specified.

| | |
|---|---|
| Simulation time step $\Delta t$ | 2 fs |
| Temperature coupling | v-rescale method, $T_{\text{ref}} = 310$, $\Delta t = 0.1$ ps |
| Pressure coupling | Parinello-Rahman barostat, $\Delta t = 2$ ps |
| Constraint on h-bonds | LINCS constraints |
| Cut-off scheme | Verlet |
| Coulomb cut-off | 1.0 nm |
| Van-der-Waals cut-off | 1.0 nm |
| Long-range electrostatics | Particle Mesh Ewald scheme |
| Fourier grid-spacing | 0.12 nm |

**Table A.1:** Parameters which were (per default) used in the production runs of MD simulations in this thesis.

## A.2 Force field parameters for DOPA and HYP

As described in section 3.1.3, we have updated our Amber force field with parameters for DOPA and our charmm36 force field with parameters for HYP, respectively. The obtained parameters are compiled here.

| atoms | | | bonds | |
|---|---|---|---|---|
| name | type | partial charge | N | H |
| N | N | -0.41570 | N | CA |
| H | H | 0.27190 | CA | HA |
| CA | CT | -0.09192 | CA | CB |
| HA | H1 | 0.12423 | CA | C |
| CB | CT | -0.07334 | CB | HB1 |
| HB1 | HC | 0.07485 | CB | HB2 |
| HB2 | HC | 0.07485 | CB | CG |
| CG | CA | 0.00415 | CG | CD1 |
| CD1 | CA | -0.18682 | CG | CD2 |
| HD1 | HA | 0.19644 | CD1 | HD1 |
| CE1 | CA | -0.34668 | CD1 | CE1 |
| HE1 | HA | 0.18979 | CE1 | HE1 |
| CZ | C | 0.25093 | CE1 | CZ |
| OH1 | OH | -0.66648 | CZ | OH1 |
| HH1 | HO | 0.47708 | CZ | CE2 |
| CE2 | CA | 0.36462 | OH1 | HH1 |
| OH2 | OH | -0.63628 | CE2 | OH2 |
| HH2 | HO | 0.47708 | OH2 | HH2 |
| CD2 | CA | -0.33398 | CE2 | CD2 |
| HD2 | HA | 0.21588 | CD2 | HD2 |
| C | C | 0.59730 | C | O |
| O | O | -0.56790 | -C | N |

**Table A.2:** Amber force field parameters for DOPA: definition of the residue and partial charges (left) and bonds (right).

| improper dihedrals | | | |
|------|------|------|------|
| -C   | CA   | N    | H    |
| CA   | +N   | C    | O    |
| CG   | CE2  | CD2  | HD2  |
| CZ   | CD2  | CE2  | OH2  |
| CD1  | CZ   | CE1  | HE1  |
| CG   | CE1  | CD1  | HD1  |
| CD1  | CD2  | CG   | CB   |
| CE1  | CE2  | CZ   | OH1  |
| N    | CA   | C    | +N   |

| bondtypes | | | | | |
|------|------|--|------|--------|----------|
| CA   | OH   |  | 1    | 0.1364 | 376560.0 |

| angletypes | | | | | |
|------|------|------|--|---------|---------|
| C    | CA   | OH   | 1 | 120.000 | 585.760 |

| dihedraltypes | | | | | | |
|---|----|----|----|---|--------|---------|---|
| C | CA | CA | OH | 4 | 180.00 | 4.60240 | 2 |

**Table A.3:** Amber force field parameters for DOPA: definition of improper dihedrals, and bonded interactions added to the force field.

| atoms | | |
|-------|------|----------------|
| name  | type | partial charge |
| N     | N    | -0 .290        |
| CD2   | CP3  | 0.000          |
| HD1   | HA2  | 0.090          |
| HD2   | HA2  | 0.090          |
| CA    | CP1  | 0.020          |
| HA    | HB1  | 0.090          |
| CB    | CP2  | -0.180         |
| HB1   | HA2  | 0.090          |
| HB2   | HA2  | 0.090          |
| CG    | CP2  | 0.140          |
| HG    | HA   | 0.090          |
| OD1   | OH1  | -0.660         |
| HE    | H    | 0.430          |
| C     | C    | 0.510          |
| O     | O    | -0.510         |

| bonds | |
|------|------|
| C    | CA   |
| C    | +N   |
| N    | CA   |
| CA   | CB   |
| CB   | CG   |
| CG   | CD2  |
| N    | CD2  |
| HA   | CA   |
| HG   | CG   |
| CG   | OD1  |
| OD1  | HE   |
| HD1  | CD2  |
| HD2  | CD2  |
| HB1  | CB   |
| HB2  | CB   |
| O    | C    |

**Table A.4:** Charmm36 force field parameters for HYP: definition of the residue and partial charges (left) and bonds (right).

| improper dihedrals | | | | | | |
|---|---|---|---|---|---|---|
| N | -C | CA | CD2 | | | |
| C | CA | +N | O | | | |
| **cmap** | | | | | | |
| -C | N | CA | C | +N | | |
| **atomtype** | | | | | | |
| CD2 | | | 12.01100 | | | |
| **bondtypes** | | | | | | |
| HA | CP2 | | 1 | 0.11110000 | 258571.20 | |
| **angletypes** | | | | | | |
| HA | CP2 | CP3 | 5 | 110.100000 | 221.752000 | 0.21790000 | 18853.10 |
| HA | CP2 | CP2 | 5 | 110.100000 | 221.752000 | 0.21790000 | 18853.10 |
| HA | CP2 | OH1 | 5 | 111.000000 | 192.045600 | 0.00000000 | 0.0 |
| **dihedraltypes** | | | | | | |
| HA | CP2 | OH1 | H | 9 | 0.000000 | 0.585760 | 0 |

**Table A.5:** Charmm36m force field parameters for HYP: definition of improper dihedrals and cmap, and bonded interactions added to the force field.

## A.3 SDS-PAGE: methods and additional gels

Our experimental collaborators used the following procedures (compare also [38]) to obtain the experimental data that was analysed during this thesis. We here also provide in Fig. A.1 the SDS-PAGE gel of the second experimental set that is not in the main Figure, as it was added in the revisions of the publication.

For the first experimental data set, Rat (*Rattus Norvegicus*) tails were sampled earlier from female rats, 5-6 months old, snap-frozen and stored at -80 °C until the day of preparation. Tendon samples from the tail were excised by pulling out single collagen fibers. The fibers were collected into ice-cold Dulbecco's phosphate-buffered saline (DPBS) w/o calcium and magnesium. Collected samples were then briefly washed by over-head agitation at 4 °C and collected by centrifugation (500 g force, 2 min, 4 °C). This step was repeated twice. After washing, the fibers were equilibrated at atmosphere for 1 h. Dried fibers were crushed in a porcelain mortar, with a manually applied stress level that is likely above the simulation setup. In previous EPR experiments [8], this led to a signal increase without a change in signal type. Next, they cooled and filled with liquid nitrogen, for 5 min. Control samples were left untreated. Both crushed and control samples were collected on dry ice and stored at -80 °C until shipment.
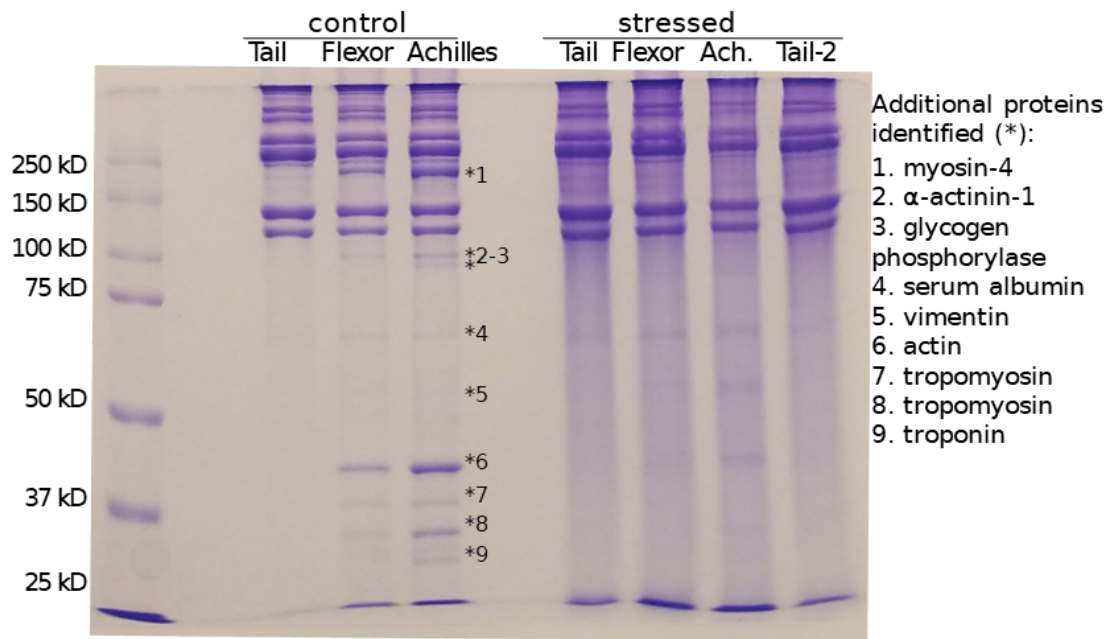
**Figure A.1: SDS-PAGE gels of a second sample set with different tissue types.** We compare control and cryo-milled tail, flexor and achilles rat tendon. In addition to the mass spectrometry analysis for the first gel, we cut out bands deviating from the expected collagen pattern (marked with an asterisk) and find them to be other (muscle) proteins.

For the second experimental set, tail tissue was sampled and prepared as described above. Achilles' and flexor tendons were dissected from hint legs and washed in the same manner. All washed tendon samples were freeze-dried for 48 h. Dried tendons were split randomly into a control and cryo-milling group, then frozen at -80 °C until the day of treatment. We switched from crushing to milling as it is more quantitatively controllable. Cryo-milling in liquid nitrogen flow was conducted on a Retsch Cyromill (Haan, Germany) with a 25 ml zirconium oxide chamber and a 15 mm marble for 2 x 2.5 min at 30 Hz. The control group was left untreated. Control and milled samples were then shipped for further analysis.

After shipment, intact type I collagen was solubilized from the lyophilized rat tissue by SDS-extraction in SDS sample buffer for 5 min at 100 °C. Sample loads were normalized to the dry weight of tendon in the first data set. As sample uptake in stressed material varied, in the second data set loads were normalized manually to the intensity of the $\alpha$-chains. SDS-extracted collagen $\alpha$-chains and collagen peptides were resolved by SDS-PAGE and stained with Coomassie Blue R-250. Collagen peptides and $\alpha$-chains were cut from SDS-PAGE gels and subjected to in-gel trypsin digestion. Electrospray mass spectrometry was carried out on the trypsin-digested peptides using an LTQ XL linear quadrapole ion-trap mass spectrometer equipped with in-line Accela 1250 liquid chromatography and automated sample injection [48, 163]. Thermo Xcalibur software and Proteome Discoverer software (ThermoFisher Scientific) were used for peptide identification and sequence coverage analysis. Tryptic peptides were also identified manually by calculating the possible MS/MS ions and matching these to the actual MS/MS spectrum. Protein sequences used for MS analysis were obtained from the Ensembl genome database.

Densitometry of stained collagen chains (($\alpha$, $\beta$, $\gamma$) from SDS-PAGE was determined using NIH ImageJ software [132]. Briefly, using ImageJ, the densitometry of each SDS-PAGE lane was calculated from a 32bit grey image of the SDS-PAGE after the background (measured as mean intensity of a line between lanes) was subtracted.

# A.4   ColBreaker parameters

There are various options in ColBreaker to choose, both for the topology of the collagen fibril (Tab. A.6) and the more general and physics-related settings of the simulation (Tab. A.7). While we have compiled the complete list here, the most important features are described in section 5.3.

| parameter | values [unit] (default) | description |
| --- | --- | --- |
| layers | integer (5) | number of layers of gap and overlap regions, so length of fibril. Use 1+4*x to get x full triple helices |
| TH_per_side | integer (4) | number of triple helices per side in hexagonal building grid, so thickness of fibril |
| gap_ratio | float (0.54) | length ratio between gap and overlap region |
| periodicity | float [m] (6.7e-8) | length of periodic d-phase (one gap plus one overlap region) |
| spacing | float [m] (1e-8) | lateral spacing between triple helices on hexagonal grid |
| allow_switches | boolean (True) | if crosslink would go outside the model, switch direction to retrieve fully crosslink network |
| connectedness | float [0,1] (1.0) | percentage of crosslinks, if lower 1.0 than randomly chosen sites will not be crosslinked |
| double_crosslink | boolean (False) | use this option to double the amount of crosslinks: Connect both up and downwards |
| low_up_ratio | float [0,1] (1.0) | fraction of crosslinks that take the standard path or switch direction |
| crosslink_sites | [N, C, BOTH] (BOTH) | side of the gap region where crosslinks will be made |
| N_path_difference_factor | float (1.0) | relative length of N-terminal path in crosslinked phase compared to standard contour length |
| C_path_difference_factor | float (1.0) | relative length of C-terminal path in crosslinked phase compared to standard contour length |

**Table A.6:** Options in ColBreaker: Part I (topological parameters)

| parameter | values [unit] (default) | description |
|---|---|---|
| time_total | float [s] (1e-7) | simulation time |
| dt | float [s] (2e-15) | integration time step |
| diffusion_constant | float [m$^2$/s] (4e-10) | diffusion constant used in Smoluchowski eq. that will influence the internal time of the system. |
| pulltype | [VELOCITY, FORCE, STRAIN] | pulling setup |
| force_constant_in_kT | float [multiple of kT] (200) | force constant for velocity or strain pulling |
| max_extension | float [initial length] (0.23) | in strain pulling, maximum extension of the virtual spring |
| v_pull | float [m/s] (100) | speed of virtual spring in constant velocity pulling, or for build-up phase in constant force and constant strain |
| constant_force | float [N] (6e-9) | force level for constant pulling scheme |
| force_constant_pwWLC | float [N/m] (2.05e-8) | force constant for enthalpic stretching in piecewise WLC |
| k_f | float [N/m] (440.5) | width of Morse potentials, 440.5 by harmonic bondtype C-CT from amber99sb* force field |
| persistence_length | float [m] (0.47e-9) | persistence length of entropic part in WLC |
| contour_factor | float (1.17) | contour factor of entropic part in WLC |
| r_fb | float [m] (1e-9) | flat-bottom extension of crosslinks before Morse potential starts |
| Edis_cross | float [multiple of kT] (119) | depth of crosslink Morse potential |
| Edis_bb | float [multiple of kT] (137) | depth of backbone Morse potential |

**Table A.7:** Options in ColBreaker: Part II (simulation and physics parameters)

## A.5 Acronyms

**RNR**        ribonucleotide reductases

**DOPA**        dihydroxy-phenylalanine

**MD**        Molecular Dynamics

**QM**        quantum mechanical

**QM/MM**  quantum mechanics / molecular mechanics

**KIMMDY**  Kinetic Monte Carlo/Molecular Dynamics

**MS**        mass spectrometry

**HLKNL**    Hydroxylysino-keto-norleucine

**PYD**        Pyridinoline

**BDE**        bond dissociation energy

**DPD**        Deoxypyridinoline

**SDS-PAGE**  polyacrylamide gel electrophoresis

**WLC**        worm-like chain

**EPR**        electron paramagnetic resonance

**EVB**        Empirical Valence Bond

**PME**        Particle Mesh Ewald

**AFM**        Atomic Force Microscopy

**PBC**        periodic boundary condition

**kMC**        kinetic Monte Carlo

**eWLC**        extensible worm-like chain

**pwWLC**  piece-wise worm-like chain

**FBM**        Fiber Bundle Model

**HAT**        hydrogen atom transfer

**HYP**        hydroxyproline

**cryo-EM**  cryogenic electron microscopy

**ALA**        alanine

**DFT**        density functional theory

*A Appendix*

**JEDI**      Judgement of Energy DIstribution

**GENESIS**  GENeralized-Ensemble SImulation System

**RSE**       radical stabilization energy

**deH-LNL**  Dehydro-lysino-norleucine

**AGE**       advanced glycation endproduct

# Bibliography

[1] Rennekamp, B. Simulating bond scission in tensed collagen: A hybrid monte carlo / molecular dynamics approach. *Heidelberg University* **Master's Thesis** (2019).

[2] Seifert, C. & Gräter, F. Protein mechanics: How force regulates molecular function (2013).

[3] Chen, Y., Radford, S. E. & Brockwell, D. J. Force-induced remodelling of proteins and their complexes. *Current Opinion in Structural Biology* **30**, 89–99 (2015). URL `https://doi.org/10.1016/j.sbi.2015.02.001`.

[4] Alegre-Cebollada, J. Protein nanomechanics in biological context. *Biophysical Reviews* **13**, 435–454 (2021). URL `https://doi.org/10.1007/s12551-021-00822-9`.

[5] Fratzl, P. e. *Collagen - Structure and Mechanics* (Springer US, 2008).

[6] Komi, P., Fukashiro, S. & Järvinen, M. Biomechanical loading of Achilles tendon during normal locomotion. *Clinics in sports medicine* (1992).

[7] Maffulli, N. Current Concepts Review - Rupture of the Achilles Tendon. *J Bone Joint Surg* (1999).

[8] Zapp, C. *et al.* Mechanoradicals in tensed tendon collagen as a source of oxidative stress. *Nature Communications* **11** (2020). URL `https://dx.doi.org/10.1038/s41467-020-15567-4`.

[9] Kurth, M. *et al.* DOPA residues endow collagen with radical scavenging capacity. *Angewandte Chemie International Edition* (2023). URL `https://doi.org/10.1002/anie.202216610`.

[10] Stubbe, J. & Nocera, D. G. Radicals in biology: Your life is in their hands. *Journal of the American Chemical Society* **143**, 13463–13472 (2021). URL `https://doi.org/10.1021/jacs.1c05952`.

[11] Jackson, M. J., Vasilaki, A. & McArdle, A. Cellular mechanisms underlying oxidative stress in human exercise. *Free Radic Biol Med* **98**, 13–17 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/26912036`.

[12] Staudinger, E., H; Leupold. Über Isopren und Kautschuk, 18. Mitteil.: Viscosittäts-Untersuchungen an Balata. *Ber Dtsch Chem Ges* **63**, 730–733 (1930).

[13] Caruso, M. M. *et al.* Mechanically-induced chemical changes in polymeric materials. *Chem Rev* **109**, 5755–98 (2009). URL `https://www.ncbi.nlm.nih.gov/pubmed/19827748`.

[14] Ducrot, E., Chen, Y., Bulters, M., Sijbesma, R. P. & Creton, C. Toughening elastomers with sacrificial bonds and watching them break. *Science* **344**, 186–9 (2014). URL `https://www.ncbi.nlm.nih.gov/pubmed/24723609`.

[15] Willis-Fox, N., Rognin, E., Aljohani, T. A. & Daly, R. Polymer mechanochemistry: Manufacturing is now a force to be reckoned with. *Chem* **4**, 2499–2537 (2018). URL `https://doi.org/10.1016%2Fj.chempr.2018.08.001`.

[16] Bo, G. D. Polymer mechanochemistry and the emergence of the mechanophore concept. *Macromolecules* **53**, 7615–7617 (2020). URL `https://doi.org/10.1021/acs.macromol.0c01683`.

[17] Chen, Y., Mellot, G., van Luijk, D., Creton, C. & Sijbesma, R. P. Mechanochemical tools for polymer materials. *Chemical Society Reviews* **50**, 4100–4140 (2021). URL `https://doi.org/10.1039/d0cs00940g`.

[18] Fantner, G. E. *et al.* Sacrificial bonds and hidden length dissipate energy as mineralized fibrils separate during bone fracture. *Nature Materials* **4**, 612–616 (2005). URL `https://doi.org/10.1038%2Fnmat1428`.

[19] Fantner, G. E. *et al.* Sacrificial bonds and hidden length: Unraveling molecular mesostructures in tough materials. *Biophysical Journal* **90**, 1411–1418 (2006). URL `https://doi.org/10.1529%2Fbiophysj.105.069344`.

[20] Luo, Y.-R. *Comprehensive handbook of chemical bond energies* (CRC press, 2007).

[21] Orgel, J. P. R. O., Irving, T. C., Miller, A. & Wess, T. J. Microfibrillar structure of type I collagen in situ. *PNAS* (2006).

[22] Bell, G. I. Models for the specific adhesion of cells to cells (1978). `arXiv:1011.1669v3`.

[23] Evans, E. & Ritchie, K. Dynamic strength of molecular adhesion bonds. *Biophysical Journal* (1997).

[24] Dudko, O. K., Hummer, G. & Szabo, A. Intrinsic rates and activation free energies from single-molecule pulling experiments. *Physical Review Letters* (2006).

[25] Kwansa, A. L., De Vita, R. & Freeman, J. W. Tensile mechanical properties of collagen type I and its enzymatic crosslinks. *Biophys Chem* **214-215**, 1–10 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/27160969`.

[26] Depalle, B., Qin, Z., Shefelbine, S. J. & Bühler, M. J. Influence of cross-link structure, density and mechanical properties in the mesoscale deformation mechanisms of collagen fibrils. *J Mech Behav Biomed Mater* **52**, 1–13 (2015). URL `https://www.ncbi.nlm.nih.gov/pubmed/25153614`.

[27] Rennekamp, B., Kutzki, F., Obarska-Kosinska, A., Zapp, C. & Gräter, F. Hybrid kinetic monte carlo/molecular dynamics simulations of bond scissions in proteins. *Journal of Chemical Theory and Computation* **16**, 553–563 (2020). URL `https://dx.doi.org/10.1021/acs.jctc.9b00786https://pubs.acs.org/doi/pdf/10.1021/acs.jctc.9b00786`.

[28] Rowe, J. & Röder, K. Chemical bonds in collagen rupture selectively under tensile stress. *Physical Chemistry Chemical Physics* **25**, 2331–2341 (2023). URL `https://doi.org/10.1039/d2cp05051j`.

[29] Franz, F., Daday, C. & Gräter, F. Advances in molecular simulations of protein mechanical properties and function. *Current Opinion in Structural Biology* **61**, 132–138 (2020). URL `https://doi.org/10.1016/j.sbi.2019.12.015`.

[30] Chenoweth, K., Van Duin, A. C. & Goddard, W. A. ReaxFF reactive force field for molecular dynamics simulations of hydrocarbon oxidation. *Journal of Physical Chemistry A* (2008). `0311084`.

[31] Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *npj Computational Materials* **2** (2016). URL `https://doi.org/10.1038/npjcompumats.2015.11`.

[32] Buehler, M. J. & Keten, S. Colloquium: Failure of molecules, bones, and the Earth itself. *Reviews of Modern Physics* (2010).

[33] Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **8**, 3192–3203 (2017). URL `https://doi.org/10.1039/c6sc05720a`.

[34] Obarska-Kosinska, A., Rennekamp, B., Ünal, A. & Grater, F. Colbuilder: a server to build collagen fibril models. *Biophys J* **120**, 3544–3549 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/34265261`.

[35] Buehler, M. J. Nanomechanics of collagen fibrils under varying cross-link densities: Atomistic and continuum studies. *Journal of the Mechanical Behavior of Biomedical Materials* (2008).

[36] Gautieri, A., Russo, A., Vesentini, S., Redaelli, A. & Buehler, M. J. Coarse-grained model of collagen molecules using an extended MARTINI force field. *Journal of Chemical Theory and Computation* **6**, 1210–1218 (2010). URL `https://doi.org/10.1021/ct100015v`.

[37] Yeo, J. *et al.* Multiscale modeling of keratin, collagen, elastin and related human diseases: Perspectives from atomistic to coarse-grained molecular dynamics simulations. *Extreme Mechanics Letters* **20**, 112–124 (2018). URL `https://doi.org/10.1016/j.eml.2018.01.009`.

[38] Rennekamp, B. *et al.* Collagen breaks at weak sacrificial bonds taming its mechanoradicals. *Nature Communications* **14** (2023). URL `https://doi.org/10.1038/s41467-023-37726-z`.

[39] Gsell, K. Y., Veres, S. P. & Kreplak, L. Single collagen fibrils isolated from high stress and low stress tendons show differing susceptibility to enzymatic degradation by the interstitial collagenase matrix metalloproteinase-1 (MMP-1). *Matrix Biology Plus* **18**, 100129 (2023). URL `https://doi.org/10.1016/j.mbplus.2023.100129`.

[40] Nijhuis, W. H. *et al.* Current concepts in osteogenesis imperfecta: Bone structure, biomechanics and medical management. *Journal of Children's Orthopaedics* **13**, 1–11 (2019). URL `https://doi.org/10.1302/1863-2548.13.180190`.

[41] Di Lullo, G. A., Sweeney, S. M., K??rkk??, J., Ala-Kokko, L. & San Antonio, J. D. Mapping the ligand-binding sites and disease-associated mutations on the most

abundant protein in the human, type I collagen. *Journal of Biological Chemistry* (2002).

[42] Komi, P. V. Relevance of in vivo force measurements to human biomechanics. *Journal of biomechanics* **23**, 27–34 (1990).

[43] Wang, J. H.-C. Mechanobiology of tendon. *Journal of Biomechanics* **39**, 1563–1582 (2006). URL `https://doi.org/10.1016/j.jbiomech.2005.05.011`.

[44] Ricard-Blum, S. The collagen family. *Cold Spring Harbor Perspectives in Biology* **3**, a004978–a004978 (2010). URL `https://doi.org/10.1101/cshperspect.a004978`.

[45] Brodsky, B., Thiagarajan, G., Madhan, B. & Kar, K. Triple-helical peptides: An approach to collagen conformation, stability, and self-association. *Biopolymers* **89**, 345–353 (2008). URL `https://doi.org/10.1002/bip.20958`.

[46] Hulmes, D. J. & Miller, A. Quasi-hexagonal molecular packing in collagen fibrils (1979).

[47] Bailey, A. Molecular mechanisms of ageing in connective tissues. *Mechanisms of Ageing and Development* **122**, 735–755 (2001). URL `https://doi.org/10.1016/s0047-6374(01)00225-1`.

[48] Eyre, D. R., Weis, M. A. & Wu, J. J. Advances in collagen cross-link analysis. *Methods* **45**, 65–74 (2008). URL `https://www.ncbi.nlm.nih.gov/pubmed/18442706`.

[49] Verzijl, N. *et al.* Crosslinking by advanced glycation end products increases the stiffness of the collagen network in human articular cartilage: A possible mechanism through which age is a risk factor for osteoarthritis. *Arthritis Rheumatism* **46**, 114–123 (2002). URL `https://doi.org/10.1002/1529-0131(200201)46:1<114::aid-art10025>3.0.co;2-p`.

[50] Snedeker, J. G. & Gautieri, A. The role of collagen crosslinks in ageing and diabetes - the good, the bad, and the ugly. *Muscles Ligaments Tendons J* **4**, 303–8 (2014). URL `https://www.ncbi.nlm.nih.gov/pubmed/25489547`.

[51] Fitch, K. R. & Goodwin, A. P. Mechanochemical reaction cascade for sensitive detection of covalent bond breakage in hydrogels. *Chemistry of Materials* **26**, 6771–6776 (2014). URL `http://dx.doi.org/10.1021/cm503253n`.

[52] Stratigaki, M. & Göstl, R. Methods for exerting and sensing force in polymer materials using mechanophores. *ChemPlusChem* **85**, 1095–1103 (2020). URL `https://doi.org/10.1002/cplu.201900737`.

[53] Bleifuss, G. *et al.* Tryptophan and tyrosine radicals in ribonucleotide reductase: a comparative high-field EPR study at 94 GHz. *Biochemistry* **40**, 15362–15368 (2001). URL `https://doi.org/10.1021/bi010707d`.

[54] Srinivas, V. *et al.* Metal-free ribonucleotide reduction powered by a dopa radical in mycoplasma pathogens. *Nature* **563**, 416–420 (2018). URL `https://doi.org/10.1038/s41586-018-0653-6`.

[55] Przybytniak, G., Sadło, J., Dąbrowska, M. & Zimek, Z. Radicals initiated by gamma rays in selected amino acids and collagen. *Nukleonika* **64**, 11–17 (2019). URL `https://doi.org/10.2478/nuka-2019-0002`.

[56] Nakagawa, T., Yokozawa, T., Terasawa, K., Shu, S. & Juneja, L. R. Protective activity of green tea against free radical- and glucose-mediated protein damage. *Journal of Agricultural and Food Chemistry* **50**, 2418–2422 (2002). URL `https://doi.org/10.1021/jf011339n`.

[57] Nugala, B., Namasi, A., Emmadi, P. & Krishna, P. Role of green tea as an antioxidant in periodontal disease: The asian paradox. *Journal of Indian Society of Periodontology* **16**, 313 (2012). URL `https://doi.org/10.4103/0972-124x.100902`.

[58] Chandra, H. & Symons, M. C. Sulphur radicals formed by cutting alpha-keratin. *Nature* **328**, 833–4 (1987). URL `https://www.ncbi.nlm.nih.gov/pubmed/2442616`.

[59] Symons, M. C. Radicals generated by bone cutting and fracture. *Free Radical Biology and Medicine* **20**, 831–835 (1996). URL `https://doi.org/10.1016/0891-5849(95)02174-4`.

[60] Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *The Journal of Chemical Physics* **27**, 1208–1209 (1957). URL `https://doi.org/10.1063/1.1743957`.

[61] Alder, B. J. & Wainwright, T. E. Studies in molecular dynamics. I. General method. *The Journal of Chemical Physics* (1959). `1.1730376`.

[62] Rahman, A. Correlations in the motion of atoms in liquid argon. *Physical Review* **136**, A405–A411 (1964). URL https://doi.org/10.1103/physrev.136.a405.

[63] Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* **99**, 1129–1143 (2018). URL https://doi.org/10.1016/j.neuron.2018.08.011.

[64] Steinhauser, M. & Hiermaier, S. A review of computational methods in materials science: Examples from shock-wave and polymer physics. *International Journal of Molecular Sciences* **10**, 5135–5216 (2009). URL https://doi.org/10.3390/ijms10125135.

[65] Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of Molecular Biology* **103**, 227–249 (1976). URL https://doi.org/10.1016/0022-2836(76)90311-9.

[66] Warshel, A. & Karplus, M. Calculation of ground and excited state potential surfaces of conjugated molecules. i. formulation and parametrization. *Journal of the American Chemical Society* **94**, 5612–5625 (1972). URL https://doi.org/10.1021/ja00771a014.

[67] Warshel, A. & Weiss, R. M. An Empirical Valence Bond Approach for Comparing Reactions in Solutions and in Enzymes. *Journal of the American Chemical Society* (1980).

[68] Warshel, A. Multiscale modeling of biological functions: From enzymes to molecular machines (nobel lecture). *Angewandte Chemie - International Edition* (2014). 15334406.

[69] Ingólfsson, H. I. *et al.* The power of coarse graining in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**, 225–248 (2013). URL https://doi.org/10.1002/wcms.1169.

[70] Souza, P. C. T. *et al.* Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods* **18**, 382–388 (2021). URL https://doi.org/10.1038/s41592-021-01098-3.

[71] Bottaro, S. & Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **361**, 355–360 (2018). URL `https://doi.org/10.1126/science.aat4010`.

[72] Jung, J. *et al.* New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems. *Journal of Computational Chemistry* **42**, 231–241 (2020). URL `https://doi.org/10.1002/jcc.26450`.

[73] Shaw, D. E. *et al.* Anton 3. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (ACM, 2021). URL `https://doi.org/10.1145/3458817.3487397`.

[74] Born, M. & Oppenheimer, R. Zur quantentheorie der molekeln. *Annalen der Physik* **389**, 457–484 (1927). URL `https://doi.org/10.1002/andp.19273892002`.

[75] Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

[76] Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function and Bioinformatics* (2010). `arXiv:1309.2828v1`.

[77] Best, R. B. & Hummer, G. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *Journal of Physical Chemistry B* (2009). `NIHMS150003`.

[78] Verlet, L. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review* **159**, 98–103 (1967). URL `https://doi.org/10.1103/physrev.159.98`.

[79] Hess, B. P-lincs: A parallel linear constraint solver for molecular simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2008).

[80] Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).

[81] Abraham, M., van der Spoel, D., Lindahl, E., Hess, B. & the GROMACS development team. *GROMACS User Manual version 2020* (2020).

[82] Darden, T., York, D. & Pedersen, L. Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics* **98**, 10089–10092 (1993).

[83] Sotomayor, M. & Schulten, K. Single-molecule experiments in vitro and in silico (2007).

[84] Grubmüller, H. Force Probe Molecular Dynamics Simulations. *Methods in Molecular Biology; Protein-Ligand interactions* (2005).

[85] Rico, F., Russek, A., González, L., Grubmüller, H. & Scheuring, S. Heterogeneous and rate-dependent streptavidin–biotin unbinding revealed by high-speed force spectroscopy and atomistic simulations. *Proceedings of the National Academy of Sciences* **116**, 6594–6601 (2019). URL https://doi.org/10.1073/pnas.1816909116.

[86] Berendsen, H. J., Postma, J. P., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* (1984). arXiv:1011.1669v3.

[87] Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of chemical physics* **126**, 014101 (2007).

[88] Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* (1981). arXiv:1011.1669v3.

[89] Kramers, H. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* (1940).

[90] Arrhenius, S. Über die Dissociationswärme und den Einfluss der Temperatur auf den Dissociationsgrad der Elektrolyte. *Zeitschrift für Physikalische Chemie* (1889). 1306.4630.

[91] Arrhenius, S. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für Physikalische Chemie* (1889).

[92] Dudko, O. K., Hummer, G. & Szabo, A. Theory, analysis, and interpretation of single-molecule force spectroscopy experiments. *Proceedings of the National Academy of Sciences* (2008).

[93] Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599–1600 (1994). URL https://doi.org/10.1126/science.8079175.

[94] Marko, J. F. & Siggia, E. D. Statistical mechanics of supercoiled DNA. *Physical Review E* **52**, 2912–2938 (1995). URL `https://doi.org/10.1103/physreve.52.2912`.

[95] Smith, S. B., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992). URL `https://doi.org/10.1126/science.1439819`.

[96] Bouchiat, C. *et al.* Estimating the persistence length of a worm-like chain molecule from force-extension measurements. *Biophysical Journal* **76**, 409–413 (1999). URL `https://doi.org/10.1016/s0006-3495%2899%2977207-3`.

[97] Rezaei, N., Lyons, A. & Forde, N. R. Environmentally controlled curvature of single collagen proteins. *Biophysical Journal* **115**, 1457–1469 (2018). URL `https://doi.org/10.1016/j.bpj.2018.09.003`.

[98] Kirkness, M. W., Lehmann, K. & Forde, N. R. Mechanics and structural stability of the collagen triple helix. *Current Opinion in Chemical Biology* **53**, 98–105 (2019). URL `https://doi.org/10.1016/j.cbpa.2019.08.001`.

[99] Buehler, M. J. & Wong, S. Y. Entropic elasticity controls nanomechanics of single tropocollagen molecules. *Biophysical Journal* **93**, 37–43 (2007). URL `https://doi.org/10.1529/biophysj.106.102616`.

[100] Odijk, T. Stiff chains and filaments under tension. *Macromolecules* **28**, 7016–7018 (1995). URL `https://doi.org/10.1021/ma00124a044`.

[101] Hillgärtner, M., Linka, K. & Itskov, M. Worm-like chain model extensions for highly stretched tropocollagen molecules. *Journal of Biomechanics* **80**, 129–135 (2018). URL `https://doi.org/10.1016/j.jbiomech.2018.08.034`.

[102] Pradhan, S., Hansen, A. & Chakrabarti, B. K. Failure processes in elastic fiber bundles. *Reviews of Modern Physics* (2010). `0808.1375`.

[103] Rainey, J. K. & Goh, M. C. An interactive triple-helical collagen builder. *Bioinformatics* **20**, 2458–2459 (2004).

[104] Hodge, A. & Petruska, J. *Recent studies with the electron microscope on ordered aggregates of the tropocollagen molecule*, 289–300 (Academic Press, New York, 1963).

[105] Hulmes, D. J., Miller, A., White, S. W., Timmins, P. A. & Berthet-Colominas, C. Interpretation of the low-angle meridional neutron diffraction patterns from collagen fibres in terms of the amino acid sequence. *International Journal of Biological Macromolecules* **2**, 338–346 (1980). URL https://doi.org/10.1016/0141-8130(80)90015-x.

[106] Bradshaw, J., Miller, A. & Wess, T. Phasing the meridional diffraction pattern of type i collagen using isomorphous derivatives. *Journal of Molecular Biology* **205**, 685–694 (1989). URL https://doi.org/10.1016/0022-2836(89)90314-8.

[107] Sasaki, N. & Odajima, S. Elongation mechanism of collagen fibrils and force-strain relations of tendon at each level of structural hierarchy. *Journal of Biomechanics* **29**, 1131–1136 (1996). URL https://doi.org/10.1016/0021-9290(96)00024-3.

[108] Frisch, M. J. *et al.* Gaussian 09, revision a. 1. *Gaussian Inc. Wallingford CT* **27**, 34 (2009).

[109] Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **25**, 247–260 (2006).

[110] da Silva, A. W. S. & Vranken, W. F. ACPYPE - AnteChamber PYthon parser interfacE. *BMC Research Notes* **5** (2012). URL https://doi.org/10.1186/1756-0500-5-367.

[111] Schroeder, A. Collagen structure and mechanics: Molecular dynamics simulations and network analysis. *Heidelberg University* **Master's Thesis** (2020).

[112] Piez, K. A., Eigner, E. A. & Lewis, M. S. The chromatographic separation and amino acid composition of the subunits of several collagen. *Biochemistry* **2**, 58–66 (1963). URL https://doi.org/10.1021/bi00901a012.

[113] Piez, K. A. & Trus, B. L. Sequence regularities and packing of collagen molecules. *Journal of Molecular Biology* **122**, 419–432 (1978). URL https://doi.org/10.1016/0022-2836(78)90419-9.

[114] Saccà, B., Renner, C. & Moroder, L. The chain register in heterotrimeric collagen peptides affects triple helix stability and folding kinetics. *Journal of Molecular Biology* **324**, 309–318 (2002). URL https://doi.org/10.1016/s0022-2836(02)01065-3.

[115] Brondijk, T. H. C., Bihan, D., Farndale, R. W. & Huizinga, E. G. Implications for collagen i chain registry from the structure of the collagen von willebrand factor a3 domain complex. *Proceedings of the National Academy of Sciences* **109**, 5253–5258 (2012). URL `https://doi.org/10.1073/pnas.1112388109`.

[116] Jalan, A. A. *et al.* Chain alignment of collagen i deciphered using computationally designed heterotrimers. *Nature Chemical Biology* **16**, 423–429 (2020). URL `https://doi.org/10.1038/s41589-019-0435-y`.

[117] Schwarzl, R., Liese, S., Brünig, F. N., Laudisio, F. & Netz, R. R. Force response of polypeptide chains from water-explicit MD simulations. *Macromolecules* **53**, 4618–4629 (2020). URL `https://doi.org/10.1021/acs.macromol.0c00138`.

[118] Stauch, T. & Dreuw, A. On the use of different coordinate systems in mechanochemical force analyses. *The Journal of Chemical Physics* **143** (2015). URL `https://doi.org/10.1063/1.4928973`.

[119] Riedmiller, K. *et al.* Predicting reaction barriers of hydrogen atom transfer in proteins (2023). URL `https://doi.org/10.26434/chemrxiv-2023-7hntk`.

[120] Karfusehr, C. Where does collagen break? a qm, mm and kimmdy study. *Heidelberg University* **Master's Thesis** (2021).

[121] Chan, B., Deng, J. & Radom, L. G4(mp2)-6x: A cost-effective improvement to g4(mp2). *Journal of Chemical Theory and Computation* **7**, 112–120 (2011). URL `https://dx.doi.org/10.1021/ct100542x`.

[122] Johnny Hioe, H. Z. *Radical Stability - Thermochemical Aspects*. Encyclopedia of Radicals in Chemistry, Biology and Materials (2012).

[123] Sustmann, R. & Korth, H.-G. The captodative effect. In *Advances in Physical Organic Chemistry*, 131–178 (Elsevier, 1990). URL `https://doi.org/10.1016/s0065-3160(08)60045-3`.

[124] Peterson, J. P. & Winter, A. H. Solvent effects on the stability and delocalization of aryl dicyanomethyl radicals: The captodative effect revisited. *Journal of the American Chemical Society* **141**, 12901–12906 (2019). URL `https://doi.org/10.1021/jacs.9b06576`.

[125] Albert, A. & Phillips, J. N. 264. ionization constants of heterocyclic substances. part ii. hydroxy-derivatives of nitrogenous six-membered ring-compounds. *J. Chem. Soc.* 1294–1304 (1956). URL `http://dx.doi.org/10.1039/JR9560001294`.

[126] Rowe, J. & Röder, K. Chemical bonds in collagen rupture selectively under tensile stress (2022). URL `https://doi.org/10.1101/2022.09.23.509192`.

[127] Kutzner, C., Czub, J. & Grubmüller, H. Keep it flexible: Driving macromolecular rotary motions in atomistic simulations with GROMACS. *Journal of Chemical Theory and Computation* (2011).

[128] Rennekamp, B. & Gräter, F. Collagen breaks at weak sacrificial bonds taming its mechanoradicals [Data] (2023). URL `https://doi.org/10.11588/data/HJ6SVM`.

[129] Miller, A. & Tocchetti, D. Calculated x-ray diffraction pattern from a quasi-hexagonal model for the molecular arrangement in collagen. *International Journal of Biological Macromolecules* **3**, 9–18 (1981). URL `https://doi.org/10.1016/0141-8130(81)90018-0`.

[130] Ottani, V., Raspanti, M. & Ruggeri, A. Collagen structure and functional implications. *Micron* **32**, 251–260 (2001). URL `https://doi.org/10.1016/s0968-4328(00)00042-1`.

[131] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. Plumed 2: New feathers for an old bird. *Computer Physics Communications* **185**, 604–613 (2014). URL `http://www.sciencedirect.com/science/article/pii/S0010465513003196`.

[132] Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH image to ImageJ: 25 years of image analysis. *Nature Methods* **9**, 671–675 (2012). URL `https://doi.org/10.1038/nmeth.2089`.

[133] Kalamajski, S. *et al.* Increased c-telopeptide cross-linking of tendon type I collagen in fibromodulin-deficient mice. *J. Biol. Chem.* **289**, 18873–18879 (2014).

[134] Hudson, D. M. *et al.* P3h3-null and sc65-null mice phenocopy the collagen lysine under-hydroxylation and cross-linking abnormality of ehlers-danlos syndrome type via. *J Biol Chem* **292**, 3877–3887 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28115524`.

[135] Hudson, D. M., Archer, M., King, K. B. & Eyre, D. R. Glycation of type I collagen selectively targets the same helical domain lysine sites as lysyl oxidase–mediated cross-linking. *Journal of Biological Chemistry* **293**, 15620–15627 (2018).

[136] Ban, E. *et al.* Strong triaxial coupling and anomalous poisson effect in collagen networks. *Proceedings of the National Academy of Sciences* **116**, 6790–6799 (2019). URL https://doi.org/10.1073/pnas.1815659116.

[137] Burla, F. *et al.* Connectivity and plasticity determine collagen network fracture. *Proceedings of the National Academy of Sciences* **117**, 8326–8334 (2020). URL https://doi.org/10.1073/pnas.1920062117.

[138] Zitnay, J. L. *et al.* Molecular level detection and localization of mechanical damage in collagen enabled by collagen hybridizing peptides. *Nat Commun* **8**, 14913 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28327610.

[139] Zitnay, J. L. *et al.* Accumulation of collagen molecular unfolding is the mechanism of cyclic fatigue damage and failure in collagenous tissues. *Science Advances* **6** (2020). URL https://doi.org/10.1126/sciadv.aba2795.

[140] Hafner, A. E., Gyori, N. G., Bench, C. A., Davis, L. K. & Šarić, A. Modeling fibrillogenesis of collagen-mimetic molecules. *Biophysical Journal* **119**, 1791–1799 (2020). URL https://doi.org/10.1016/j.bpj.2020.09.013.

[141] Bühler, M. J. Nature designs tough collagen: explaining the nanostructure of collagen fibrils. *Proc Natl Acad Sci U S A* **103**, 12285–90 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16895989.

[142] Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**, 33–38 (1996).

[143] Rajkumar, D. S., Murugan, G. & Padmanaban, R. Unraveling the interaction of bisphenol a with collagen and its effect on conformational and thermal stability. *Biophysical Chemistry* **298**, 107026 (2023). URL https://doi.org/10.1016/j.bpc.2023.107026.

[144] Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem., Int. Ed.* **48**, 1198–1229 (2009).

[145] Cui, Q. Perspective: Quantum mechanical methods in biochemistry and biophysics. *J. Chem. Phys.* **145** (2016).

[146] Chow, W. Y. *et al.* Proline provides site-specific flexibility for in vivo collagen. *Scientific Reports* **8** (2018). URL https://doi.org/10.1038/s41598-018-31937-x.

[147] Masic, A. *et al.* Osmotic pressure induced tensile forces in tendon collagen. *Nature Communications* **6** (2015). URL `https://doi.org/10.1038/ncomms6942`.

[148] Metreveli, N. *et al.* Identification of free radicals induced by UV irradiation in collagen water solutions. *Journal of Photochemistry and Photobiology B: Biology* **93**, 61–65 (2008). URL `https://doi.org/10.1016/j.jphotobiol.2008.06.003`.

[149] Kjeldsen, F. *et al.* C??c backbone fragmentation dominates in electron detachment dissociation of gas-phase polypeptide polyanions. *Chemistry - A European Journal* **11**, 1803–1812 (2005). URL `https://doi.org/10.1002/chem.200400806`.

[150] Blaesi, E. J. *et al.* Metal-free class Ie ribonucleotide reductase from pathogens initiates catalysis with a tyrosine-derived dihydroxyphenylalanine radical. *Proceedings of the National Academy of Sciences* **115**, 10022–10027 (2018). URL `https://www.pnas.org/content/115/40/10022`. `https://www.pnas.org/content/115/40/10022.full.pdf`.

[151] Treyde, W., Riedmiller, K. & Gräter, F. Bond dissociation energies of x–h bonds in proteins. *RSC Advances* **12**, 34557–34564 (2022). URL `https://doi.org/10.1039/d2ra04002f`.

[152] Street, D., Bangsbo, J. & Juel, C. Interstitial pH in human skeletal muscle during and after dynamic graded exercise. *The Journal of Physiology* **537**, 993–998 (2001). URL `https://doi.org/10.1113/jphysiol.2001.012954`.

[153] Kemp, G., Böning, D., Beneke, R. & Maassen, N. Explaining pH change in exercising muscle: Lactic acid, proton consumption, and buffering vs. strong ion difference. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **291**, R235–R237 (2006). URL `https://doi.org/10.1152/ajpregu.00662.2005`.

[154] Díaz, F. E., Dantas, E. & Geffner, J. Unravelling the interplay between extracellular acidosis and immune cells. *Mediators of Inflammation* **2018**, 1–11 (2018). URL `https://doi.org/10.1155/2018/1218297`.

[155] Imato, K. & Otsuka, H. Self-healing polymers through dynamic covalent chemistry. In *Dynamic Covalent Chemistry*, 359–387 (John Wiley & Sons, Ltd, 2017). URL `https://doi.org/10.1002/9781119075738.ch9`.

[156] Krauss, S., Metzger, T. H., Fratzl, P. & Harrington, M. J. Self-repair of a biological fiber guided by an ordered elastic framework. *Biomacromolecules* **14**, 1520–1528 (2013). URL `https://doi.org/10.1021/bm4001712`.

[157] Usha, R., Rajaram, A. & Ramasami, T. Stability of collagen in the presence of 3, 4-dihydroxyphenylalanine (DOPA). *Journal of Photochemistry and Photobiology B: Biology* **97**, 34–39 (2009). URL `https://doi.org/10.1016/j.jphotobiol.2009.07.009`.

[158] Fuentes-Lemus, E., Hägglund, P., López-Alarcón, C. & Davies, M. J. Oxidative crosslinking of peptides and proteins: Mechanisms of formation, detection, characterization and quantification. *Molecules* **27**, 15 (2021). URL `https://doi.org/10.3390/molecules27010015`.

[159] Thompson, J. B. *et al.* Bone indentation recovery time correlates with bond reforming time. *Nature* **414**, 773–776 (2001). URL `https://doi.org/10.1038/414773a`.

[160] Liu, J. *et al.* Energy dissipation in mammalian collagen fibrils: Cyclic strain-induced damping, toughening, and strengthening. *Acta Biomaterialia* **80**, 217–227 (2018). URL `https://doi.org/10.1016/j.actbio.2018.09.027`.

[161] Gachon, E. & Mesquida, P. Stretching single collagen fibrils reveals nonlinear mechanical behavior. *Biophysical Journal* **118**, 1401–1408 (2020). URL `https://doi.org/10.1016/j.bpj.2020.01.038`.

[162] Baldwin, S. *et al.* A new longitudinal variation in the structure of collagen fibrils and its relationship to locations of mechanical damage susceptibility. *Journal of the Mechanical Behavior of Biomedical Materials* **110**, 103849 (2020). URL `https://doi.org/10.1016/j.jmbbm.2020.103849`.

[163] Eyre, D. Collagen cross-linking amino acids. In *Methods in Enzymology*, 115–139 (Elsevier, 1987). URL `https://doi.org/10.1016/0076-6879(87)44176-1`.

# List of Figures

# Acknowledgments

First and foremost, I would like to thank Prof. Dr. Frauke Gräter for the continuous supervision and the possibility to carry out this thesis in the Molecular Biomechanics group at the Heidelberg Institute for Theoretical Studies. As a German idiom says, people vote with their feet, and I stayed after the Master's thesis for the PhD despite other opportunities. I am grateful to the whole group that supports each other in a collaborative environment. I also thank Prof. Dr. Helmut Grubmüller for contributing to the supervision of this thesis and for the opportunity to engage into the seminars of his department at the MPI, even if Covid-19 moved the planned research exchange online. Furthermore, I would like to thank Prof. Dr. Ulrich Schwarz for fostering my interest in theoretical biophysics and, specifically, for contributing as part of the Thesis Advisory Committee. Moreover, I thank Prof. Dr. Tristan Bereau for reviewing this thesis.

I also acknowledge my overall privileged position that certainly helps on such an endeavour. Lastly, I would like to thank my parents, my partner Cristina, and all my family and friends for ongoing support.

## Statement of Originality (Erklärung):

I certify that this thesis, and the research to which it refers, are the product of my own work. Any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Ich versichere, daß ich diese Arbeit selbständig verfaßt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, October 4, 2023

........................................
(signature)