University College London

Fully Atomistic Modelling of Collagen Cross-linking

Thesis submitted for the degree of Doctor of Philosophy (PhD) by

Thomas Collier

Supervisors:

Professor Nora H. de Leeuw

And

Professor Helen L. Birch

University College London

Department of Chemistry

September 16

Declaration

I, Thomas Andrew Collier confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis

Thomas Collier

28 September 2016

Abstract

The extracellular matrix (ECM) undergoes progressive age-related stiffening and loss of proteolytic digestibility due to an increase in concentration of advanced glycation end products (AGEs). Detrimental collagen stiffening properties are believed to play a significant role in several age-related diseases such as osteoporosis and cardiovascular disease. Currently little is known of the potential location of covalently cross-linked AGEs formation within collagen molecules; neither are there reports on how the respective cross-link sites affect the physical and biochemical properties of collagen. Using fully atomistic molecular dynamics simulations (MD) we have identified preferential sites for exothermic formation of two lysine-arginine derived AGEs, glucosepane and DOGDIC. Identification of these favourable sites enables us to align collagen cross-linking with experimentally observed changes to the ECM. For example, formation of both AGEs were found to be energetically favourable within close proximity of the Matrix Metalloproteinase-1 (MMP1) binding site, which could potentially disrupt collagen degradation. With the aid of a number of dynamic analysis techniques we have provided an explanation for the site specificity of the two AGE cross-links. The mechanical properties of collagen were also investigated through the use of steered MD to determine the effect of the cross-links presence. Additionally the effect of the sequence on the collagen mechanical properties was also investigated, owing to the heterogeneous response of collagen to an applied load.

A homology model for the *Homo sapiens* sequence was developed from the crystal structure of the *Rattus norvegicus* structure that was shown to produce stable simulations. Through the use of the homology model and implementation of a novel simulation technique we attempted to ascertain the orientations of the collagen molecules within a fibril, that is currently below the resolution limit of experimental techniques.

Table of Contents

Table of Figures

Table of Tables

List of Abbreviations

AFM – Atomic Force Microscopy

AGE – Advanced Glycation End Products

CLP - Collagen-Like Peptide

DFT – Density Functional Theory

ECM – Extracellular Matrix

HF – Hartree Fock

HSP-47 – Heat Shock Protein 47

MD – Molecular Dynamics

MMP1 – Matrix Metalloproteinase 1

NMR – Nuclear Magnetic Resonance

PBC – Periodic Boundary Conditions

QM – Quantum Mechanical

RESP – Restrained Electrostatic Potential

RMSD – Root Mean Squared Deviation

SMD – Steered Molecular Dynamics

SPE – Single Point Energy

YM – Young's Modulus

Acknowledgements

Firstly, I would like to express my sincere thanks to my supervisors, Professor Nora de Leeuw and Professor Helen Birch for their support and guidance throughout my PhD studies.

This work would not have been possible without the generous financial support of the EPSRC funded Molecular Modelling and Materials Science Centre for Doctoral Studies, who have provided valuable funding, tutelage and guidance for the duration of my studies. In addition I would like to thank Research Computing at UCL and ARCHER, through our membership of the Materials Chemistry Consortium, for access to high performance computing facilities.

For four years I have been part of two amazing groups; firstly the de Leeuw group, who have provided me with both fruitful intellectual discussion but also many happy memories I will keep with me for life. Secondly, the collaborative collagen research group, who have shared so much of their vast and diverse expertise with me, and have inspired me to think differently in my approach to scientific problems. Special thanks to Dr. Anthony Nash who has been instrumental in providing support and guidance during my studies.

Not forgetting thank you to my family who have been amazingly supportive and encouraging throughout all of my studies, despite not understanding what I am talking about most of the time.

Finally, I cannot end without thanking my wonderful wife, Ella, who has been a patient, supportive and loving partner throughout.

List of Publications

The work described in this thesis has been published in the following papers:

T.A. Collier, A. Nash, H.L. Birch, N.H. de Leeuw, Intra-molecular Lysine-Arginine derived Advanced Glycation End-product cross-linking in Type I collagen: A molecular dynamics simulation study, Biophysical Chemistry, **218**, (2016) 42-46. doi:10.1016/j.bpc.2016.09.003.

T.A. Collier, A. Nash, H.L. Birch, N.H. de Leeuw, Preferential sites for intramolecular glucosepane cross-link formation in type I collagen: A thermodynamic study., *Matrix Biology*, **48** (2015) 78–88. doi:10.1016/j.matbio.2015.06.001.

List of Selected Presentations

The work described in this thesis has been presented at the following conferences:

"Identification of preferential sites for intra-molecular lysine-arginine derived advanced glycation end products formation in fibrillar type I collagen and their effect on the function of collagenous tissues: an all atom molecular dynamics approach" T.A. Collier, A. Nash, L. Bozec, H.L. Birch, N.H. de Leeuw, **talk,** 10th World Biomaterials Congress, Montreal, Canada, 17th-22nd May 2016. Proceedings: Front. Bioeng. Biotechnol. Conference Abstract: 10th World Biomaterials Congress. doi: 10.3389/conf.FBIOE.2016.01.02048

"Identification of Preferential Sites for Glucosepane Cross-link Formation in Fibrillar Type I Collagen and Their Effect on the Function of Collagenous Tissues: An All Atom Molecular Dynamics Approach", T.A. Collier, A. Nash, L. Bozec, N.H. de Leeuw, H.L. Birch, **talk and poster**, The Collagen Gordon Research Conference and Seminar, New London, NH, USA, 11th-17th July 2015

"The Thermodynamic Identification of Glucosepane Cross-linking in Type I Collagen – An All Atom Molecular Dynamics Study", T.A. Collier, A. Nash, A.F. Lopez-Clavijo, L. Bozec, N.H. de Leeuw, H.L. Birch, **poster**, Joint meeting of the BSMB and BSDB - The musculoskeletal system from development to disease, Norwich, UK, 1st – 3rd September 2014. Proceedings: Int. J. Exp. Pathol. **96**(2), 2015, doi: 10.1111/iep.12125

Chapter 1 Introduction

1.1 Collagen

The name “Collagen” is used as a generic term for a family of proteins which typically have functions in tissue assembly or maintenance and form the characteristic triple helix of three polypeptide chains. The family of proteins, currently contains 28 different human collagen types (1), with a number of closely related structures considered to be part of the collagen superfamily, such as, acetyl cholinesterase, adiponectin and C1q. Distinguishing between “collagen” and “collagen-like” proteins is non-trivial as there are proteins with triple helical domains that are not considered to be collagen. The collagen family can be subdivided further into a number of groups based on structure and role, these include; fibril forming collagens, FACIT (Fibril Associated Collagens with Interrupted Triple Helices), network forming collagens, transmembrane collagens, basement membrane collagens and unclassified collagens. The most abundant subgroup of collagens are the fibril forming collagens, which constitute approximately 90% of total collagen types (2). The majority of collagen proteins are homotrimers, containing three identical polypeptide chains, for example collagen II. However there are also collagens that are heterotrimeric containing different polypeptide chains, for example type I collagen, which contains two identical α1 chains and one α2 chain.

1.1.1 Biosynthesis of Collagen

Collagen biosynthesis is a complex multistep process, from gene transcription within the nucleus to aggregation of collagen heterotrimers into large fibrils (2, 3). It had long been believed that the organisation of the collagen was a “self-

assembly" process; secreted collagen molecules are ejected into the inter-cellular space to self assemble. However this is no longer fully supported (4, 5).

It is believed that genes are transcribed in the nucleus of the eukaryotic cells, before the translation of the ribosome-bound mRNA into pre-procollagen molecules which protrude into the lumen of the rough endoplasmic reticulum. After removal of a signal peptide by a signal peptidase, the pro-collagen then undergoes a number of post-translational modifications. The main post-translational modifications are the hydroxylation of the proline and lysine residues by the enzymes prolyl 3-hydroxylase, prolyl 4-hydroxylase and lysyl hydroxylase. The presence of 4-hydroxyproline is essential for the formation of stabilising intramolecular hydrogen bonds whilst hydoxylysine residues are required to form mature enzymatic intermolecular cross-linking of the collagen molecules into the fibrils (6). The extent of lysine hydroxylation varies depending on the organism and the tissue type (7). The next stage is the assembly of the three-polypeptide chains into the triple helical structure. The formation of the triple helical region from the C- to the N- terminus begins with the alignment of the C-terminal domains of the three-polypeptide chains.

For efficient folding and formation of the pro-collagen to occur, the presence of several enzymes is required, for example peptidyl-prolyl cis-trans-isomerase (8) and collagen specific chaperones like Heat Shock Protein – 47 (9). Globular propeptides are essential during this process as they ensure association between monomeric pro-collagen chains and provide stability through disulphide bonds in the C-terminal pro-peptides. After forming the triple helical pro-collagen structure and post-translational modifications, the pro-collagen is transported via the Golgi complex; they are packaged into secretory vesicles,

called Golgi to plasma membrane carriers and finally passed into the extracellular space. Once secreted into the extracellular space the pro-collagen is processed, with cleavage of the C- and N-propeptides by Zn^{2+} dependent metalloproteinases procollagen C-proteinase and pro-collagen N-proteinase, respectively (10). Some argue that some processing may be done in the carrier during secretion, as the N- and C-proteinases have been identified in the Golgi network (11). However it is still believed that the majority of procollagen processing is performed in the extracellular space. Upon completion of the pro-collagen to collagen fibril formation, collagens, including type I collagen, spontaneously aggregate into ordered fibrillar structures *in vivo* to form long thin fibrils in a process known as fibrillogenesis (12), (although the exact mechanism is still not known). These fibrils then bundle further, in various orientations depending on the tissue type. For example, in tendon collagen fibrils align parallel to each other and form fibres whereas in the skin, the orientation is more random, forming a complex network of interlaced fibrils. Fibril diameter can be modulated by the binding of small leucine rich proteoglycans such as decorin, which inhibits lateral growth of the fibril once bound (13, 14). The molecular arrangement of collagen molecules is further stabilised by enzymatic covalent cross-links, which we will discuss in more detail in the section 1.1.2.3. The fibrils then bundle together further to form tissues, for example in tendon, the fibrils bundle to form fibres, which then further associate to form a fascile, which then group to form the tendon.

1.1.2 Type I Collagen

Type I collagen is the most abundant protein in animals, playing an important structural role in the extracellular matrix (ECM) of all vertebrates and accounting for over a quarter of the dry mass of the human body (15, 16). The fibril-forming type I collagen dominates in organs and tissues that require tensile strength, such as tendon, ligament, bone and the dermis. The precise way in which type I collagen molecules are organized and chemically linked into collagen fibrils provides these tissues with the specific mechanical properties required for efficient biological function (17).

1.1.2.1 Structure

The tropocollagen (a collagen molecule) is the basic structural unit of all collagen types; it is a rope-like macromolecule comprised of three polypeptides strands twisted into a continuous triple helix, with a right-handed supercoil of the polypeptide chains with a one residue stagger between the chains (18). It is approximately 300 nm in length and 1.5 nm in diameter, and is flanked at both ends by a non-helical telopeptides domain. These collagen molecules are secreted by human cells into the extracellular matrix, whereupon they spontaneously bundle tightly together to form hydrated collagen fibrils which typically have diameters of 20-500 nm (12). This process is called fibrillogenesis. Collagen microfibrils are made up of the high-aspect ratio collagen molecules arranged in a staggered configuration. According to the Hodge-Pertruska model, the fibrils are deposited side-by-side and parallel but also staggered with respect to each other (19), in a quasi-hexagonal packing (20). This structure creates an observable periodicity known as the D-band, where D=67 nm (21). This is made up of an overlap “region” which has a higher

protein density, in addition to a gap between two consecutive collagen molecules known as the “gap region”, which measures 0.54 D or 36 nm, as seen in Figure 1 (19).

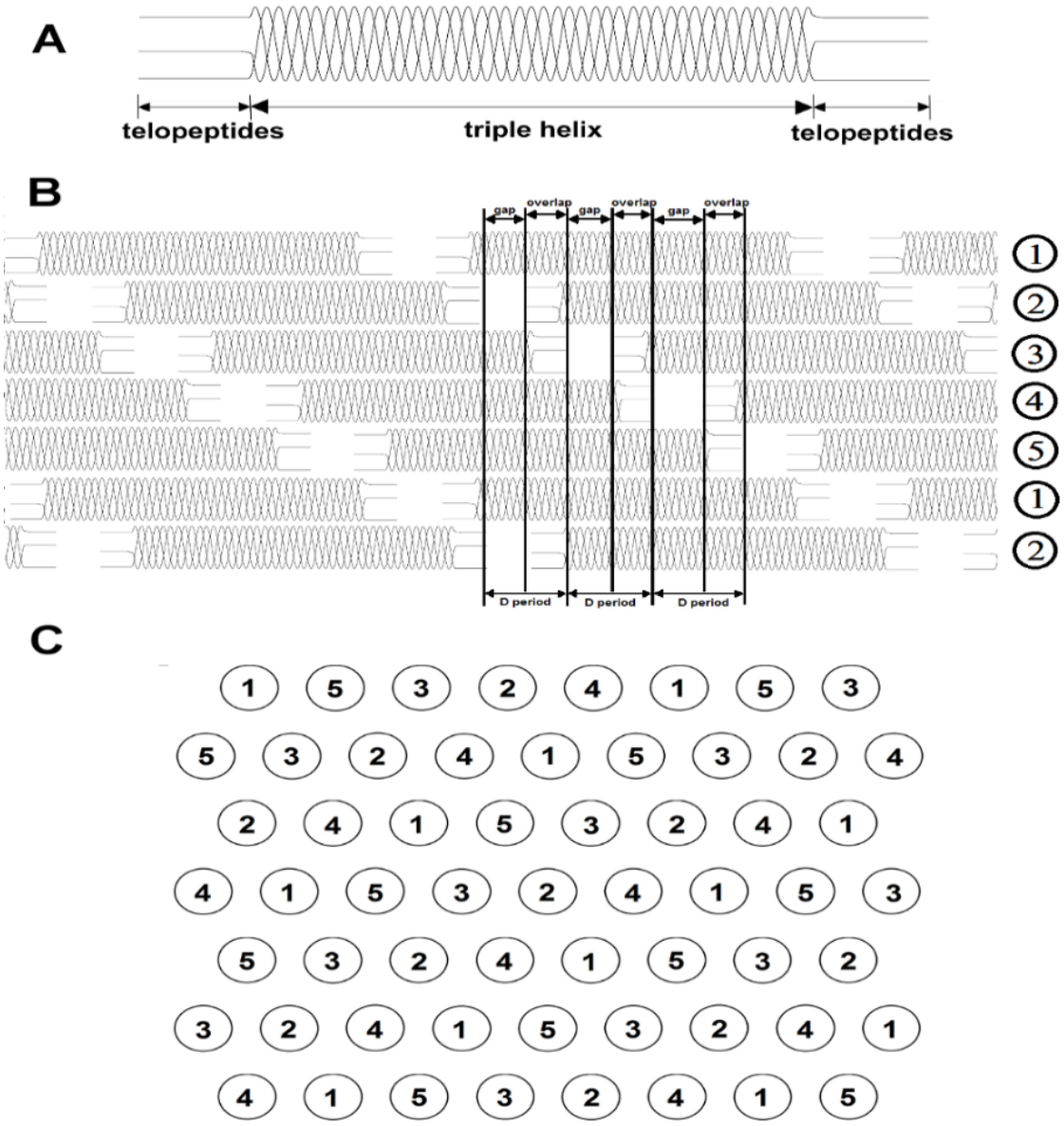


Figure 1: Schematic of (A) a single collagen protein; (B) and (C) are schematics showing the supramolecular arrangement of the collagen proteins in a collagen fibril. Specifically (B) shows the staggered axial alignment in the fibril, with each collagen molecule represented as a straight rod. (C) Cross section through a fibril in the overlap region, showing the quasi-hexagonal packing, with each collagen molecule represented as a circle. In both (B) and (C) the number represent the five possible axial alignments of the proteins.

Collagen fibrils are formed through the bundling of several microfibrils, these vary in length and diameter dependent on the organism and location of the tissue. Collagen fibrils within tendon typically have diameters ranging between 20 and 150 nm and a length on the millimetre scale. For example, in human Achilles tendon, average fibril diameters of 50 – 90 nm have been measured. In the flexors and extensors of the fingers diameters are 20 - 60 nm (22, 23).

Ramachandaran initially proposed the concept of the triple helical structure for collagen over 60 years ago, by employing fibre diffraction theory with stereo-chemical consideration (24). Collagen-like peptides have backbone torsional angles φ and ψ which fall in the region of -76°, 127° in the Ramachandaran plot (25). Since then a significant number of studies have been conducted to better understand the structure of type I collagen. However, given the large size, insolubility, complex hierarchical structure and repetitive sequence most conventional biochemical analysis techniques are unable to obtain atomic level structural information.

These challenges have resulted in a large number of studies that utilise collagen-like peptides (26–31). For example Berman *et al.*, confirmed the existence of inter-chain hydrogen bonds between N-H(Gly)---O=C(Xxx) and Cα-H---O=C(Xxx/Gly) through the use of high resolution crystallography of a collagen-like peptide (26, 32). In addition to these inter-strand hydrogen bonding, there is also water-mediated hydrogen bridging, where a water molecule simultaneously forms hydrogen-bonding interactions with two residues on different molecules or strands, as seen in Figure 2. Experimental studies have previously hypothesised that these water mediated hydrogen bridges

could be the driving force for fibrillogenesis and a major component to collagen fibril stability (33–35).

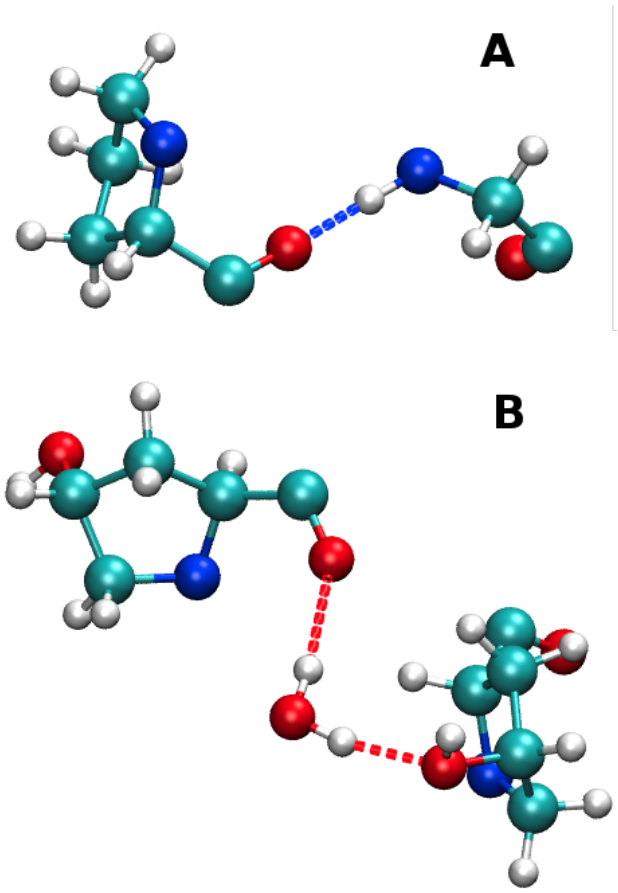


Figure 2: Image showing two different types of hydrogen bonding interactions (A) A direct inter-protein hydrogen bond (B) A water mediated hydrogen bonding interaction, using the bridging water molecule.

The collagen molecules are initially only stabilised by the non-covalent interactions immediately after fibrillogenesis. As the fibrils mature, the tissue is stabilised further by an enzyme-mediated cross-link forming within the telopeptide region of the molecules (6). Collagen arranged and cross-linked in this way results in tissues with high tensile strength (36).

In addition to a mechanical contribution, the precise arrangement of collagen molecules within the fibril governs important interactions with other matrix

macromolecules and cellular component of the tissue. For example, decorin, a small leucine-rich proteoglycan (SLRP), binds to fibrillar collagen at specific sites, where sufficient space is available to accommodate the protein core (13). The binding of decorin plays a role in regulating the collagen fibril diameter by inhibiting lateral growth of the fibril (14). Collagen also contains cell interaction domains, which enable binding to integrins on the cell surface. This cell matrix interaction is important for mechanotransduction and other cell signaling events (37).

1.1.2.2 Sequence dependencies

The primary sequence is vitally important in the generation of the collagen structure. The presence of a glycine residue in the third position of the polypeptide chain is essential for formation of the triple helix (2). The amino acid triplet sequences occurring in type I collagen were characterized by Heidemen and Roth, where typically Gly-X-Hyp and Gly-Pro-Y appear approximately 25% of the time each in the primary sequence and the other 50% is typically Gly-X-Y in native type I collagen (38). The presence of glycine in the third position enables the polypeptide chains to arrange to form a helical structure, with the glycine residues typically on the inside of the helix, owing to its small size, and the bulkier residues being located on the outer positions, maximizing the creation of Van der Waals interactions. Furthermore, glycine plays a significant role in the formation of inter-strand hydrogen bonds, adding further to the structure's stability. The results of an amino acid single point substitution of glycine to bulkier residues are considered to be some of the most damaging mutations to collagen genes, and have been found to play a role in the

pathogenicity of a number of genetic disorders such as *Osteogenesis Imperfecta* (or brittle bone disease) (39). Bella *et al.*, have shown that a simple glycine to alanine substitution in a collagen-like peptide resulted in a triple helical molecule with an partially untwisted or “bulge” region at the site of the substitution (26). This was a result of a disruption to both the packing and the hydrogen bonding network.

Proline and its post-translational modified variant, hydroxyproline are also abundant in type I collagen, accounting for 28% and 38% of the total number of amino acids respectively (40). Raines *et al* suggested that this is due to their significant contribution to the structural stability of collagen from the inductive effects of the pyrrolidine ring (41–43). Proline residues in the Yyy position of the triplet are modified enzymatically by polyl 4-hydroxylase, which hydroxylates the 4(R) position of the proline ring to form a hydroxyproline residue (7, 44). It is well known that electron withdrawing substituents at 4(R) position on the proline ring, such as is present in hydroxyproline, stabilize the helix at the Yyy position, as it reduces the conformational freedom of the proline ring and constrains the dihedral angles of the backbone (45–47). The same was not found when the hydroxyl group was added in the 4S configuration (48), or for hydroxyprolines present in the Xxx position (49). Hydroxyproline is also considered to be a significant contributor to the water mediated hydrogen bridges between collagen molecules acting, as the water bridge donor for 21.2% of the inter-protein interactions, owing to its abundance and high polarity (35).

When the Xxx and Yyy positions are not occupied by proline or hydroxyproline respectively, it could potentially be populated by any amino acid, giving rise to over 441 possible triplet sequences. However some amino acids are found to

destabilise the triple helical structure. Brodsky *et al.*, conducted an exhaustive study using a collagen-like protein with the sequence $(ProHypGly)_5(XxxYyyGly)(ProHypGly)_5$, where they varied Xxx and Yyy with all possible variations of the 20 standard amino acids (40, 50). Their study showed a strong correlation between the residue's contribution to triple helical stability as a function of its propensity to adopt a left handed polyproline II-type helical conformation. Some residues confer greater stability. For example, an arginine or a lysine residue in the Yyy position contributes greatly to triple helical stability (51, 52). Conversely the aromatic amino acid residues, tryptophan, phenylalanine and tyrosine are strongly destabilising.

Given the strong influence of the primary sequence on the stability of the triple helix, variations to a wild type collagen sequence can lead to disastrous phenotypic consequences such as *Osteogenesis Imperfecta,* as previously mentioned. However, even more subtle point mutations, which do not alter the size of the amino acid significantly, can yield detrimental effects with an $arginine_{(Xxx)}$ to cysteine mutation, causing Ehlers-Danlos syndrome through the generation of disulphide bonds (53). More recently abnormal activity of collagen hydroxyproline has been linked to lung cancer (54, 55).

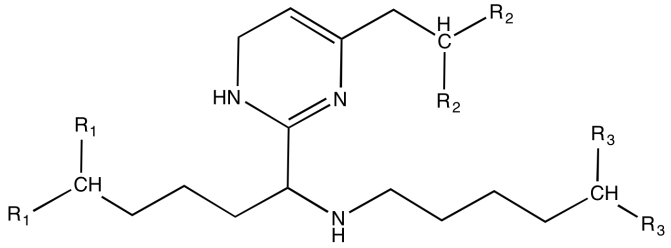
1.1.2.3 Enzymatic Mature Cross-links

After fibrillogenesis a number of "maturing" processes occur, with the molecules only initially being aggregated through the inter-molecular interactions. However this is relatively weak and, to provide the mechanical properties required of collagenous tissues, covalent cross-linking is necessary. The formation of the enzymatic cross-links depends on the presence of specific enzymes, amino

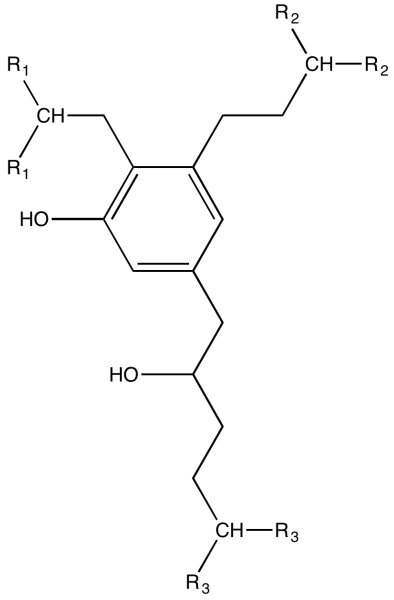
acid sequences and quaternary structure. The first stage in the generation of the enzymatic cross-links is aldehyde formation from lysine and hydroxylysine residues in the telopeptide regions, catalysed by the copper-dependent lysyl oxidase. The aldehydes then go on to spontaneously react further to form intermediate cross-links, with the lysyl-aldehydes reacting with lysines on neighbouring residues to form intermediate cross-links called Schiff bases. Hydroxylysyl-aldehydes react to form ketoimine bonds, generating hydroxylysino-keto-norleucine (HLKNL). Further maturation converts the intermediate cross-links into the non-reducible mature products; the Schiff Bases are converted to non-reducible histidin adducts, predominately histidino-hydroxylysinonor-leucine (HHL) (6). Maturation of the hydoxylysine residue product HLKNL with a hydroxylysine-aldehyde or a second ketoimine can form a pyridinium cross-link (56). Alternatively HLKNL can react with a lysyl-aldehyde to form the hydroxylysyl-pyrrole (57). The structures of these cross-links can be seen in Figure 3.

The main determinant of the type of cross-link formed is whether the post-translational enzymatic hydroxylation of the lysine residues by lysyl hydroxylase has taken place (58). The proportion of hydroxylysine varies depending on the collagen type and location. For example, in the skin, the levels of hydroxylysine are much lower than in the bone, which accounts partly for the difference in the biomechanical properties of the tissue (58). The mature enzymatic cross-links are predominantly located in the overlap region of the fibril, with the cross-links occurring between the telopeptide regions of neighbouring molecules (36). In cases of copper deficiency, lysyl-oxidase activity is reduced and thus the number of cross-links is found to be much lower, with the tissues exhibiting

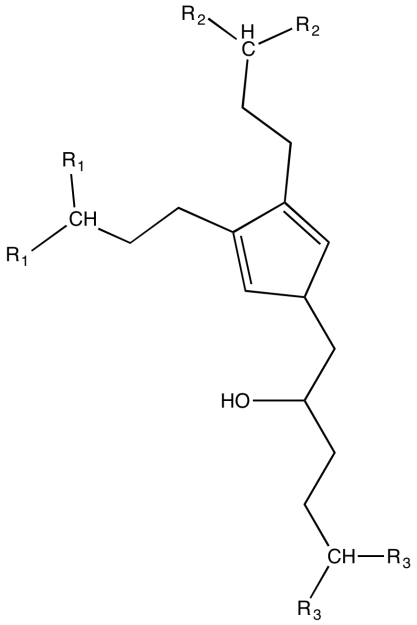
greater fragility, further highlighting the role of cross-links in the biomechanics of collagenous tissues.



Histidino-hydroxylysinonorleucine (HHL)



Hydroxylysyl-Pyridinoline



Hydroxylysyl-Pyrrole

Figure 3: Schematic representation of the three main enzymatic cross-links, histidino-hydroxylysinonor-leucine, hydroxylysyl-pyridinoline and hydroxylylsyl-pyrrole.

1.1.3 Mechanical Properties of Type I Collagen

With collagen making up a significant proportion of connective tissues, approximately 90% in some cases, its biomechanical and energy storage properties are of utmost importance. The mechanical functions of the

supramolecular structure in collagenous tissues are optimised for the direction and magnitude of load. Tendons have unidirectional tensile strength, a consequence of fibre alignment of thick bundles parallel to the long axis of the tendon (59). In skin, the fibres form an isotropic network capable of managing multidirectional forces (15). Forces experienced by the collagenous tissues vary greatly in magnitude and direction. Applied force can be sporadic, sustained or repetitive. For example, a runner's Achilles tendon can experience peak forces of 11.4 times their body weight (60), experiencing over 2000 cyclic loading events during a 5 km run (61).

The mechanical properties of collagenous tissues can be broken down into a number of different scales; the molecular scale, the response of the collagen molecule to strain; the fibrillar level with the response of fibrils to an applied load; moving into the microscale, which incorporates the response of a collagen fibre; and finally the macroscale, in which the mechanics of the whole collagenous tissue are considered. At the lowest scale, a number of atomistic and coarse grained molecular dynamic simulations have been conducted to the response of the molecule to an applied load (62–64), with a small number of experimental studies even claiming to be probing single molecule responses to a load (65, 66). At the microfibril and fibril level the amount of experimentally determined data increases, with experiments using a wide variety of techniques now possible (67–71). As the hierarchical scale increases, it has been observed that the Young's modulus decreases significantly with the molecular level ranging from 2-9 Gpa. On the tissue scale the modulus varies between 0.001-1 Gpa depending on the tissue type (72). The most likely reason for this is the inter-fibrillar sliding that occurs on a macroscopic scale; in addition to the straightening and reorientation of the fibrils/fibres (73).

1.1.4 Computational Modelling of Collagen

Since the proposal of the helical structure for collagen by Ramachandran and Kartha in 1955 (74), a huge number of computational studies have been conducted to elucidate the structure further, from the early simulation of the 1970's, using simple techniques on very small peptide sequences (75, 76), to the more complex and sophisticated simulations of the full collagen molecule developed over the past decade (77, 78). Despite huge advances being made in the software development and the capability of hardware, simulations of a solvated full collagen fibre are still beyond current capabilities, with simulations of the collagen molecule pushing the limits of what is realistically achievable. It is for this reason that a lot of research has gone into developing novel techniques to be able to probe the structural properties of fibrillar collagen molecules, all with the intention of reducing the computational expense of the calculations by reducing either the timescale (79), number of atoms (77, 78), complexity of the system (80), the variety of properties able to be studied (81, 82), or a combination of these. What follows is a short summary of some successful computational studies used to probe a variety of structural, electronic or dynamic properties of fibrillar collagen molecules.

Some dynamic events, particularly of large molecules, are often not observable by the use of conventional MD techniques, such as the folding of proteins, owing to the timescale at which these events occur. Folding of coiled-coil structures normally occurs on millisecond to second timescales (83, 84), and collagen is thought to fold on a timescale of minutes to hours (85). Both are beyond current computational capabilities. To overcome long time scale events, Stultz developed a method to promote folding within a shorter period of time, enabling the observation of a folding event during a short MD trajectory (79). In

this approach, a gentle bias is introduced during MD simulations to favour the formation of the pre-specified conformation. This is done by 1) selecting movements in the simulation which take the system closer to the target, 2) a small energetic penalty is applied to movements that take the system away from the target (86–88). The trajectory starts with three unfolded polypeptides and the target is the triple helical conformation.

After application of the bias approach, it was observed that most of the progress in folding occurs within the first 3.7 ns of the simulation. Despite the mechanism observed being consistent with previous experimental folding studies by Boudko *et al* (89), there are still two limitations to this approach. Firstly, the target structure must be known. Secondly, artifacts can be introduced into the trajectory as a result of the biasing potential. However this approach remains a useful technique to probe slow dynamics on computationally achievable timescales.

Another approach to increase the timescale achievable through simulation is to increase the simulation integration time step used. One study that utilized this approach is the work of Gautieri *et al.*, with their coarse grained (CG) model of collagen using the Martini force-field (77). Coarse grained models allow the study of larger systems, up to micrometre dimension and millisecond duration, which would allow access to many materials phenomena, such as tissue deformation and failure, which require large samples and occur on long timescales (90). CG models work by reducing the number of degrees of freedom through grouping of atoms into pseudo-atoms referred to as beads (91–93). The MARTINI force-field model assigns atoms into individual beads, retaining information about the amino acid sequence. For example, a particular

single bead type will represent amine or methyl groups, or saturated carbon atoms (94). This significantly reduces the number of atoms, and therefore the degrees of freedom. The number of beads used to represent an amino acid is dependent on the size of the side chain, with small residues like glycine being described by just one bead, whilst large amino acids such as tyrosine are modelled with up to five beads (95). The MARTINI model also takes into account the polarity of each bead described by a letter (P, polar; C, apolar; N, non-polar; and Q, charged) and a number (from 1, low polarity, to 5, high polarity), as well as characterizing the beads' hydrogen bonding capabilities. Validation of a short 8 nm collagen-like peptide showed a good correlation in root-mean-square deviation (RMSD) relative to an all atom simulation, which allowed an expansion to a full -length CG molecule in solution. This approach has the benefit of having significantly lower computational costs from fewer degrees of freedom, relative to a corresponding fully atomistic simulation, allowing much large molecules to be constructed. However this comes at the cost of a reduction in structural detail, as CG models cannot undergo a change to their secondary structure. Recent shape based CG simulation studies have been able to access collagen fibrils in the micrometre scales, yielding macroscopic structural properties of collagen fibrils (96).

The exploitation of periodic boundary conditions can be used to reduce the number of explicitly defined protein and water molecules. This replicates the regularly repeating densely packed collagen protein and solvent. The first such study to exploit the D-periodicity of collagen in MD simulations was Streeter and de Leeuw (78). The dense fibrillar environment of collagen was replicated whilst using only one fifth of the amount of water compared to the fully solvated system, in addition to creating a more realistic model of collagen within a fibril

(35). Validation of the method was conducted by comparison of alpha carbon positions to those of an experimental crystal structure (21), with agreement to within 2.48 Å. This study is the most realistic simulation of the fibrillar collagen at present to utilize atomic resolution detail, whilst minimizing the computational cost. However this approach does have the limitation that, due to the period boundary conditions, it is less amenable to studying low concentration point mutations within its primary sequence.

Whilst MD simulations are excellent for studying structural properties of collagen, their basis in Newtonian physics means the electronic structure is not considered beyond the coulombic interaction of point charges. Hence to probe the electronic properties of collagenous systems density functional theory (DFT) techniques are needed, although due to the significantly greater computational cost of such methods the size of the systems studied is much smaller. The results identify trends, which can be applied to the wider structure. One such example is the use of *ab initio* and DFT techniques to study the role of various collagen triplets on the stability of collagen (80). Through the use of these simulations to calculate chemical hardness and solvation free energies, Madhan *et al.*, were able to probe the stability and solvation effect of amino acids within various triplets present in collagen. The models were run using a relatively small amount of atoms, approximately 50 atoms, using the widely distributed Gaussian98 package. However this study does suffer from the drawback that only triplets were studied and, so making statements about the entire collagen triple helix is not possible, although it does allow us to comment on the local-helix stability.

Machine learning algorithms are a branch of artificial intelligence that allows us to hypothesize structure and relationships based on work that is currently known experimentally or theoretically from conventional methods. It has a huge potential in bioinformatic projects based on the complex nature of modelling such large structures. Z. Yang applied machine learning techniques to predict collagen hydroxyproline sites using support vector machines (SVM) (81). This requires a well-defined proper kernel function, dealing with the sequence homology alignment. There are two simple metrics to score the similarity or distance between two sequences; 1) the Needleman-Wunsch score, and 2) Dayhoff score as well as its variants (97–99). The Needleman-Wunsch score is binary, while the Dayhoff score is based on a probability estimation. This enables two different kernels to be derived using SVM. An odd size-sliding window is used to scan the whole collagen sequence to generate peptides with a proline in the middle residue, because hydroxyproline always requires a proline to form, the middle residue is therefore removed for computational efficiency. The generated peptides are based on thirty-seven hydroxyproline-annotated collagen-like sequences collected from the NCBI (National Center for Biotechnology Information). From the generated peptides the training set is created by mixing 80% of the annotated peptides with 80% of the non-annotated peptides. The remaining 20% from each forms the testing set. This process is repeated five times for fivefold cross-validation. Despite the method being able to predict, on small amounts of computational resources, the high frequency of glycine sites within the model peptides at high sensitivity (85%). A large number of misclassifications for the less abundant hydroxyproline sites was observed, making this method unreliable especially for prediction of less abundant amino acids. The use of both kernel SVM models ensured no

hydroxyproline sites were completely missed, with no sites missed by both kernels. Yang *et al.* concluded that the study could be improved had they used a greater number of training sets instead of restricting to thirty-seven collagen-like protein sequences.

Whilst no single methodology is currently available to fully model every structural property of collagen, the adoption of a number of techniques can yield a better picture of the structure, function and dynamics of collagen. Currently MD provides the most promising methodology for modelling full size collagen peptides, although the results are limited by the quality of the force-field. DFT methods, however, are capable of probing electronic effects and bond breaking or forming, but are limited to smaller peptides due to the associated computational cost. Machine learning algorithms show potential for the prediction in trends of local stability despite their limitations probing physical perturbations to the system. However these techniques require large data training sets to give reliable results.

1.2 Advanced Glycation End Products

Whilst enzyme-mediated cross-linking described previously is physiological and provides functionality to tissues, glycation-mediated cross-linking is considered to be pathological and is thought to jeopardize the functionality of the musculoskeletal system (100, 101). Advanced Glycation End products (AGEs) have additionally been linked to the pathogenesis of a number of chronic diseases from neurodegenerative diseases to cancer metasis (102, 103). AGEs are formed by a series of successive chemical reactions between a reducing sugar, such as glucose (an aldose) or fructose (a ketose), and a protein or lipid.

When no enzyme is present to catalyse the reaction, the non-enzymatic glycation is called the Maillard reaction (104). Formation of AGEs and cross-links are site-specific processes influenced by steric constraints and the amino acid side chains (105). The protein side chain functional groups exert a strong influence on the reactivity of amino groups with glucose and also on the kinetics and products of subsequent Maillard reactions (106). Amino acids on the surface of the protein will also be easily glycated, owing to the availability of their exposed amino acid side chains to glycation.

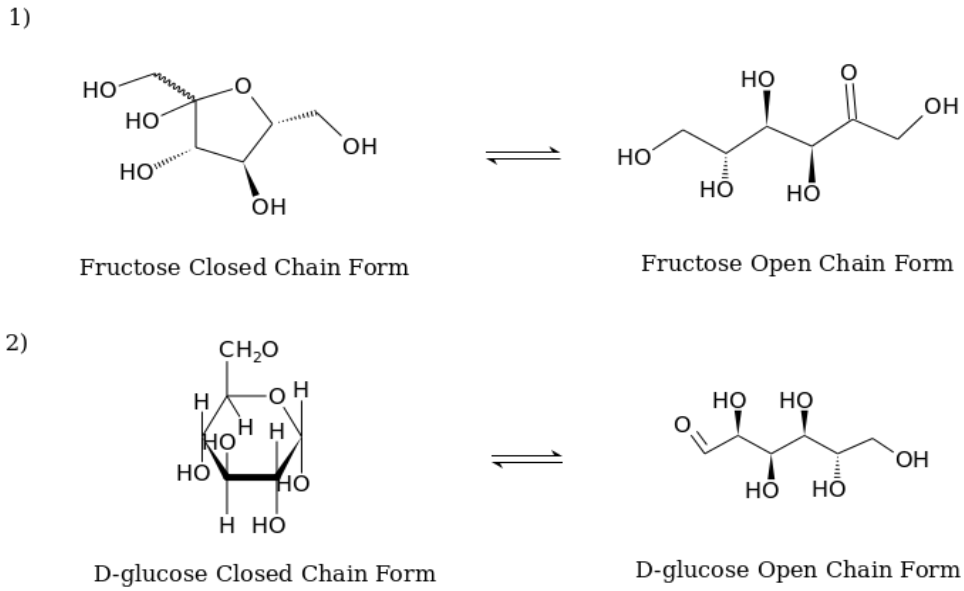


Figure 4: Open and closed cyclic structural forms of 1) fructose 2) D-glucose.

The main external influence on the rate of AGEs formation is the equilibrium between the sugars open-chain and cyclic form. This is because only an open-chain sugar molecule can react with an amino acid residue, as the cyclic form does not contain a reactive aldehyde or ketone group (107), as seen in Figure 4. Aldoses are more reactive than ketoses, owing to the fact that the terminal aldehyde group is more accessible and electrophilic than the ketone group

(107). However, for most sugars, the cyclic form is thermodynamically favoured, owing to the bond enthalpy of two C-O bonds in the cyclic form being lower in energy than the C=O bond enthalpy present in the linear form. Hence for glucose only 0.002% of molecules exist in the reactive open-chain form *in vivo*, with galactose and fructose having 0.02% and 0.7% in the open-chain form respectively. Despite this strong preference for glucose to be in the cyclic form its role in protein cross-linking is significant, owing to the fact that glucose is the most abundant sugar *in vivo*, with a concentration in the blood plasma of 5 mmol/L. Fructose on the other hand has a concentration of only 35 µmol/L (108). However in some tissues glucose can be converted to fructose by the polyol pathways, although the activity of this pathway is low in healthy patients. Conversely in diabetic patients this pathway is very active and leads to levels of fructose in lenses and nerves that exceed that of glucose (101).

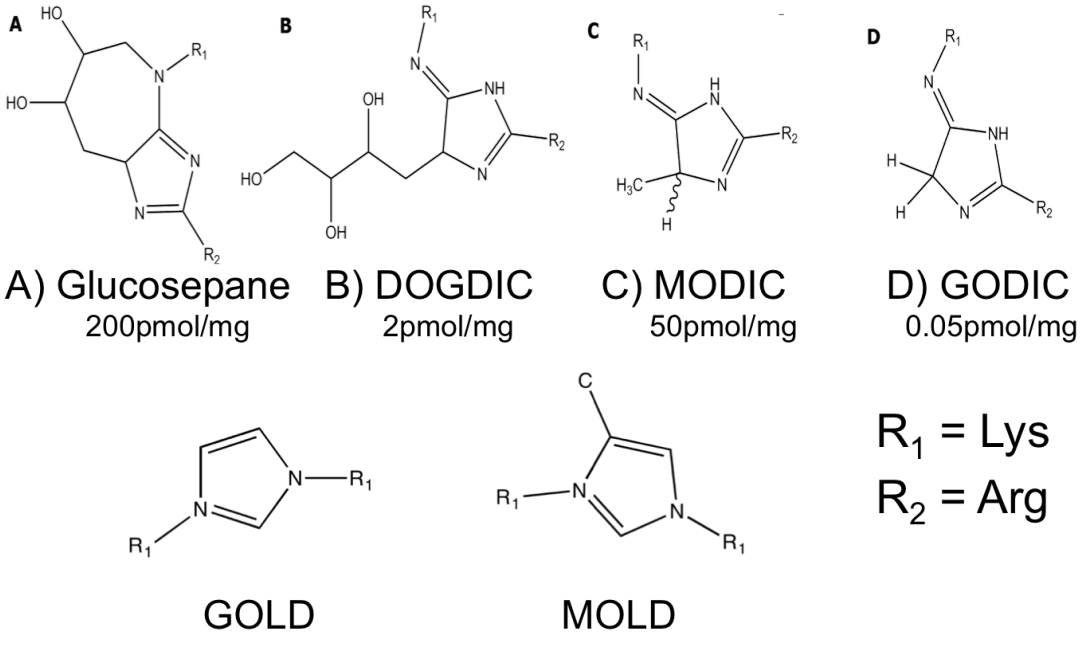


Figure 5: Schematic Image of the common AGE cross-links, R1=Lysine R2=Arginine

Only a few cross-linking AGEs have so far been detected *in vivo*. There are two main sets of AGEs, distinguished depending on the bound amino acids; lysine-lysine AGEs and lysine-arginine AGEs. The lysine-lysine AGES are MOLD and GOLD, two imidazolium compounds which form from methyl glyoxal and glyoxal (101). There are four main lysine-arginine AGEs; glucosepane, DOGDIC, MODIC and GODIC, seen in Figure 5. MODIC and GODIC also form via reaction of the lysine and arginine residues with methyl glyoxal and glyoxal. However DOGDIC and glucosepane form by different mechanisms via the open-chain sugar. A 2002 study by Biemel *et al* quantified the levels of these AGEs in human lens protein and found concentrations of 132.3-241.7 pmol/mg of protein, 1.3-8.0 pmol/mg of protein, 40.7-97.2 pmol/mg of protein and concentrations below the quantifiable level of the instrument respectively (101).

Glucosepane was first identified, by Lederer *et al.*, through model reactions of protected lysine and arginine residues with D-glucose (109). The structure of glucosepane consists of a seven membered ring made from glucose, with a molecular weight of 647 Da. A number of chemical properties have made glucosepane difficult to study via common experimental techniques. For example the cross-link only absorbs UV light at very short wavelengths, making HPLC with UV detection not possible. Also glucosepane is labile to acid hydrolysis, which means it can only be quantified after enzymatic digestion and not in intact tissue (110).

The first stage in the formation of glucosepane is the binding of an open chained D-glucose molecule to an amino group of a lysine residue to form a Schiff base. However, this structure is unstable and will undergo a rearrangement to form a more stable fructosamine called the Amadori product.

Through a series of carbonyl shifts that takes several weeks, the Amadori product produces a dideoxyosone, sometimes referred to as Lederer's glucosone (111, 112). The dideoxyosone then undergoes cyclisation to form a cyclic aldimine that reacts with the guanidine group of the arginine residue to form glucosepane. This simplified reaction mechanism can be seen in Figure 6. Reihl *et al.*, have also shown that the additional hexose sugars, such as, D-galactose and D-mannose can generate the dideoxyosone (112). One theory to explain the abundance of glucosepane is that the final carbonyl rearrangement from the Amadori product undergoes a non-reversible dehydration step that ultimately leads to an accumulation of glucosepane, whereas other AGEs are formed reversibly (113).

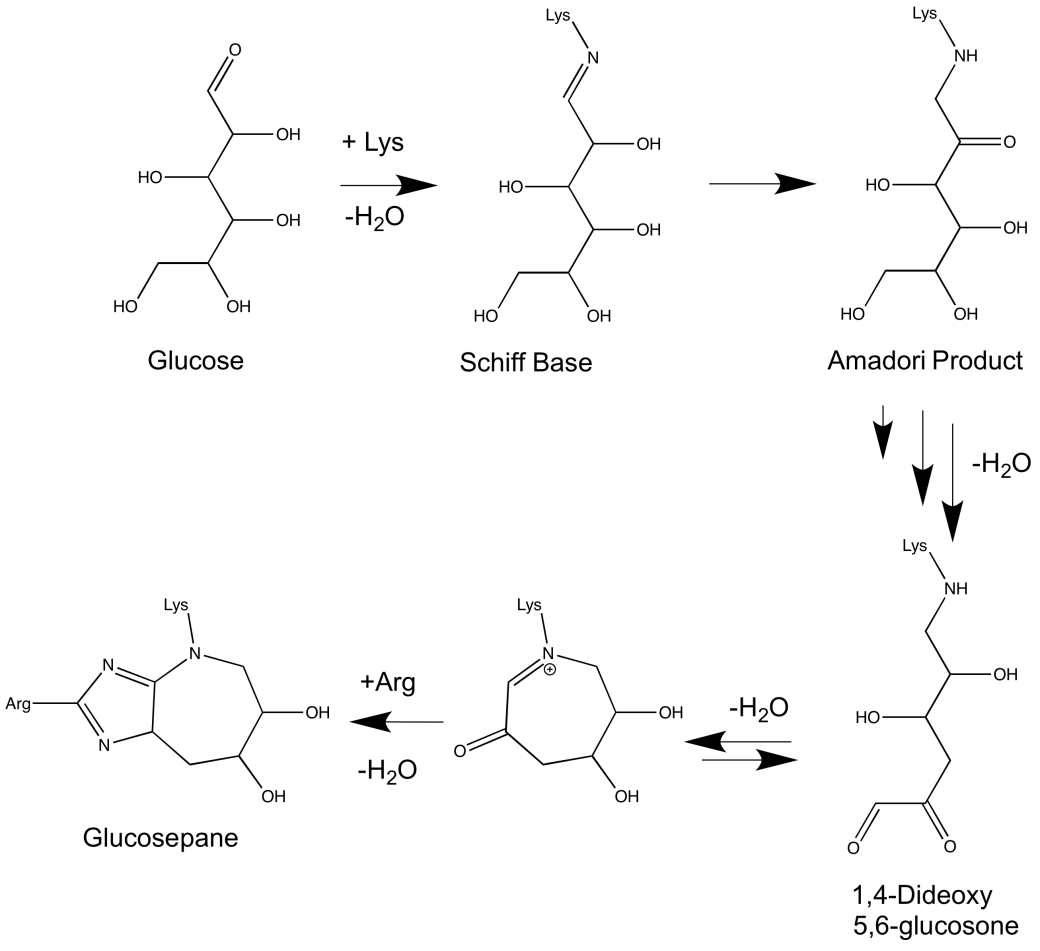


Figure 6: Schematic representation of the abbreviated glucosepane formation mechanism from glucose

Less is currently known about DOGDIC owing to its lower abundance within tissues, which has resulted in studies focusing more on the abundant glucosepane cross-link. However its significance should not be underestimated, as the position of the cross-linking sites may have biological significance, or its presence may impede other lysine-arginine AGEs from forming. DOGDIC, standing for deoxyglucosone-derived imidazoline cross-link, consists of a five membered ring cross-linking a lysine and arginine residue with a 4-carbon hydroxylated aliphatic chain extending from the ring, as seen in Figure 5, with a molecular weight of 664 Da. Currently the exact mechanism of formation is not known. However it has been shown that it too proceeds via the Amadori product (114), as opposed to the glucose degradation products.

1.2.1 Impact of Advanced Glycation End Products

As mentioned previously in 1.2, it has been shown that the polyol pathway is more active in patients with *diabetes mellitus* (108). Combined with the lack of insulin, human diabetic patients can develop an increase of glucosepane cross-link concentration, in skin reaching levels of 4500 pmol/mg (100), resulting in 1 in 2 collagen molecules being cross-linked. In healthy human skin glucosepane levels reach concentrations of 2000 pmol/mg, with on average 20% of the collagen molecules being cross-linked by glucosepane. This is significantly less than the 50% cross-linked collagen molecules found for diabetic patients, which is not surprising considering that serum glucose concentration levels are significantly higher than in non-diabetic patients (100). A 2012 audit by Diabetes UK found 2.9 million people were diagnosed with *diabetes mellitus* within the UK (115). This value is predicted to double by 2025, such that 10% of the

population will be diagnosed with diabetes. For this reason an understanding of the complications affecting diabetes sufferers is essential.

Over the past 30 years there has been a widespread adoption of high fructose corn syrup (HFCS) as a replacement for sucrose in the food industry. As a result, we will see blood fructose concentration increase and insulin sensitivity decrease in consumers of large quantities of products containing HFCS (116, 117). Both the increase in patients diagnosed with *diabetes mellitus,* as well as the widespread adoption of HFCS in the food industry, means that now an understanding of AGEs formation and effect on the body is essential.

AGEs have been linked to the pathogenesis of several chronic diseases from neurodegenerative diseases to cancer (102, 103). Stiffening of the lung ECM from the build up of AGEs is likely to contribute to age associated changes in lung function, such as loss of elastic recoil and subsequently reduced lung capacity (118). So far only pentosidines presence has been proven, although the study measured cross-links using fluorescence measurements which will not show the presence of other AGEs within the lung tissue, most other AGEs not being fluorescence active (118). Several studies have proposed that AGEs, through collagen cross-linking, play a role in increasing myocardial and vascular stiffness, accumulating in heart and vascular tissue, as well as promoting the development of cardiac hypertrophy (119–122).

It is also widely known that AGEs are responsible for age-related increase in stiffness of many collagen-rich tissues (100). The functionality of the musculoskeletal system is believed to be strongly reduced by the build-up of glucosepane cross-links and other AGEs within its tissues; a long-lived protein collagen is particularly vulnerable to AGE cross-linking (123). For example, in

tendon a half-life of up to 200 years was calculated for the collagenous component based on aspartic acid racemization (124). Components that make up the musculoskeletal system tend to have longer half-lives compared to other collagenous tissues such as skin (123). Accumulation of reducible sugars in bone tissue leads to increased levels of cross-linking of collagen, which has been cited as an important contributing factor to the age-related deterioration of bone density, potentially leading to osteoporosis (125). A previous study by Reddy *et al*., found that *in vitro* incubation of rabbit Achilles tendon in ribose increased levels of the AGE pentosidine and increases Young's modulus by 159% from 24.89 ± 1.52 MPa to 65.087 ± 14.41 MPa. This suggests that the presence of AGE cross-links increased stiffness of soft tissue (126). This decrease in functionality of the musculoskeletal tissues predominantly affects the elderly, causing infirmity, and reducing mobility in these generations. By gaining a better understanding of the processes that occur within the tissue it may be possible to achieve a more active life for the older members of the population. This could in turn have huge socio-economic implications. By 2025 it is predicted that half of the population of the UK will be over the age of 50. A better understanding of this aging mechanism could help reduce the pressure on public healthcare and the benefits and pension costs that the UK will face.

In recent years there has been a surge in the use of artificial tissues; created by combining cultured cells and a polymer scaffold. Collagen is an ideal scaffold, primarily due to it being permissive to host remodelling, as well as the fact it is a natural cell substrate conducive to the critical events of cell migration (spreading), and its ability to bind to a large number of ECM components (127–129). It is via this application of collagen as a cell scaffold in tissue engineering that we see an advantageous benefit of AGEs presence. The main limitation of

the use of collagen in tissue equivalents, particularly in load bearing applications, has been the insufficient stiffness and poor strength of the implant, which can result in terrible shape retention. A new approach to overcome this was the use of cross-linking. However all conventional cross-linking methods use chemical and thermal treatment that makes them unsuitable for tissue engineering applications, given that such treatments kill cells (130–133). The targeted use of non-enzymatic glycation during the fabrication phase of tissue constructs, could be utilised to stiffen and strengthen the tissue equivalents without damaging the cells within (134, 135).

1.2.2 Treatment of Advanced Glycation End Products

Due to how little is presently known about cross-linking by AGEs within the body there are currently no clinically licensed treatments. Focus has been placed on finding methods to slow down, stop or even reverse formation of AGEs. Dietary control of glucose consumption alone is not sufficient to prevent cross-links from forming, as the serum glucose levels necessary for inhibition would result in severe hypoglycaemia and malnutrition (136–139). However, having lower serum glucose concentrations through dietary control, would slow down the rate of formation.

The second approach is to find ways of preventing binding to amino acid residues. Aspirin has been proven to acetylate the μ-amino residues in many proteins including collagen. Pyridoxal-5'-phosphate is also able to protect the amino acid by forming a Schiff base with them (140–142). However aspirin and Pyridoxal-5'-phosphate's clinical importance is only minor, owing to the abundance of lysine and arginine residues within the body, and thus it would be

impossible to block every single amino acid residue within the human body. Additionally, modification through binding to the amino acid residue will likely alter the secondary structure of the protein and may itself introduce detrimental effects on function (143).

Another potential method is to break down the intermediates before the cross-links form. The human body already produces some enzymes that can remove through conversion certain intermediates such as glyoxal, methylglyoxal, 3-deoxyglucosone and Amadori products (144–146), to less reactive species. One example of such an enzyme is Fructosamine-3-kinase, which can convert the intermediate Amadori product back to 3-deoxyglucosone and the unglycated protein (147). Current research focuses on identifying all of the intermediates, to allow further development of treatments by mimicking the role of these natural enzymes within the body.

Cross-link breakers are the most promising field of treatment currently being researched; AGEs are very different to any functional chemical structure within the body reducing the effect of complications (148). However development of cross-link breakers is very complicated, owing to the need to cleave between two to four covalent bonds, depending on the AGE. The most promising example of these cross-link breakers is alagebrium chloride (ALT-711), developed by the Alteon Corporation, which went into phase II clinical trials in 2005 (149, 150). Unfortunately the promising results seen in the animal model were not replicated within the human body, highlighting the need to fully understand the processes surrounding AGE formation, its preference for binding within the collagen molecule and the role of surrounding molecules (151).

1.2.3 Previous Studies on Advanced Glycation End Products

To date few studies have been conducted on AGEs owing to their complexity in study both experimentally and computationally. One of the main barriers to experimental studies is the long time scale over which these cross-links form, with collagen gel integrity decreasing over time; a faster synthesis approach is necessary to isolate the effects of cross-linking from those of gel degradation. One method adopted was to use more reactive species to generate non-specific AGE cross-links for investigation of the mechanical properties. The species used were typically glyoxal or ribose based. However for studies of the impact of glucosepane cross-links specifically this time factor was still a problem until recently, when work by Spiegel *et al.* produced a new rapid one-pot approach for glucosepane synthesis (152). It is hoped that, with the development of this new synthesis technique the number of experimental studies probing glucosepane will increase.

Computational studies on advanced glycation end products are also limited, owing to the large size of the collagen molecule, with the main focus on the formation mechanisms of the cross-links (153, 154). Even the mechanistic studies previously conducted have had to make significant compromises in their approach. For example, Nasiri *et al.* conducted a QM study on the mechanism of formation of glucosepane, excluding the effect of water and instead using a polarisable continuum model, as well as reducing the lysine and arginine residues down to be represented by just methyl guanidine and methyl amine, thus removing the effect of the remainder of the side-chains on the energetics (153).

1.3 Hypothesis

This work will aim to address the general hypothesis that the accumulation of AGE cross-links within the collagenous matrix of tendon, ligament and bone is detrimental to the tissue's mechanical and biological function. This hypothesis can be broken down further into three main hypotheses:

1. AGE cross-linking is site specific.
2. The presence of glucosepane cross-link has a detrimental effect on the biological function of the tissue, specifically on the tissues susceptibility to enzymatic degradation.
3. Determine to what extent the position and quantity of glucosepane cross-links affects the elastic properties of collagenous tissues.

Chapter 2 Methodology

Molecular modelling is the general term used to simulate the behaviour and features of molecular systems and encompasses many different theoretical methods and computational techniques. Modelling techniques can be applied across a wide range of scales, from small chemical compounds, to large biomolecules and across many different fields, including physics, chemistry, biology and materials science. The level of detail also varies depending on the techniques applied. For example, in quantum mechanical simulations, electrons are considered explicitly, whereas in coarse-grained molecular dynamics simulations, whole amino acids are the smallest level of detail. The current chapter aims to introduce principal components of the simulation techniques employed in this thesis. A broader overview of computational techniques can be found in many accessible resources (155–158).

2.1 Interaction Potentials

In force-field methods the lowest unit treated explicitly is atoms meaning electrons are not treated explicitly and thus the bonding information must be provided rather then it being the result of solving the electronic Schrödinger's equation. A force-field is composed of the bonded and non-bonded interaction potentials along with some associated parameters, such as polarisability. The sum of the bonded and non-bonded interactions for a single particle yields the total energy, with the negative differential of that total energy with respect to particle position returning the force on that particle.

2.1.1 Bonded Terms

The bonded portion of a typical interaction Hamiltonian can be seen in Eq.1. The first term $U_b(r_{ij})$ gives the potential energy for the stretching of the bond between two atoms (i and j). The second term the potential energy function $U_a(\theta_{ijk})$ of the angle bending between three atoms (i, j and k). The third term is the potential energy function $U_{id}(\phi_{ijkl})$ for keeping the chirality and planar structure of carbon rings through improper dihedrals and the last the potential energy $U_a(\theta_{ijkl})$ associated with bond rotation.

$$U_{bonded,i} = U_b(r_{ij}) + U_a(\theta_{ijk}) + U_{id}(\phi_{ijkl}) + U_d(\theta_{ijkl}) \qquad \text{(Eq. 1)}$$

Each of the four contributing terms to the total potential energy of the molecule are explained in more detail below.

2.1.1.1 Two Body Harmonic Bond Stretching Potential

A simple harmonic oscillator approximation can be applied when considering the bond between two particles, such that the potential energy of bond stretching is given by Eq.2

$$U_b(r) = \frac{1}{2} k_b (r - r_0)^2 \qquad \text{(Eq. 2)}$$

where k_b is a spring constant between the two particles and where r is the bond length and r_0 is an ideal bond length such that $(r\text{-}r_0)^2$ is the squared distance of bond length from the ideal value. The contribution to the force from particle i is given by Eq.3 and takes the form:

$$F_i(r) = -k_b (r - r_0) \qquad \text{(Eq. 3)}$$

The simple harmonic bond stretching potential is the method implemented within the AMBER force-field, owing to its simplicity and thus greater efficiency. However a number of different bond stretching potentials exist; they typically take the form of anharmonic bond stretching potentials such as the cubic bond stretching potential (159) and the Morse potential (160). With the later being capable of explicitly including the effects of bond breaking, such as the existence of unbound states, at the cost of a greater complexity and reduced computational efficiency.

2.1.1.2 Three Body Harmonic Angle Potential

The three body angular bond potential, represented by Eq. 4, describes the angular vibrational motion occurring between three atoms (i,j,k), and is implemented much like the harmonic bond stretching potential.

$$U_a(\theta) = \frac{1}{2} k_a(\theta - \theta_0)^2 \qquad \text{(Eq. 4)}$$

where k_a is the harmonic spring constant, the current angle θ and θ_0 is an idealised angle such that potential energy increases as the angle deviates greater from the ideal value. The contribution to the force on each particle within the angle can be seen from the three equations below (Eq. 5 - Eq.7):

$$F_i = -\frac{\delta U_a(\theta_{ijk})}{\delta r_i} \qquad \text{(Eq. 5)}$$

$$F_k = -\frac{\delta U_a(\theta_{ijk})}{\delta r_k} \qquad \text{(Eq. 6)}$$

$$F_j = -F_i - F_k \qquad \text{(Eq. 7)}$$

This simple potential can be extended to the Urey-Bradley potential (161) (Eq. 8), as implemented in the CHARMM force field by inclusion of a second term, which is used to describe a covalent spring between the outer atoms (i and k), where r_{ik} gives the distance between the two atoms, and where r_{UB} is the equilibrium distance, k_{UB} is a constant which activates the Urey-Bradley term when it is non zero.

$$U_a(\theta) = \frac{1}{2}\sum_{angles} k_a(\theta - \theta_0)^2 + k_{UB}(r_{ik} - r_{UB})^2 \qquad \text{(Eq. 8)}$$

However its use has not been adopted in AMBER, owing to its additional parameter that needs fitting when determining the optimal parameters for a simulation, also resulting in a reduction in the transferability. Additionally the gains from inclusion of this term only produces relatively minor beneficial subtleties in the vibrational spectra. For this purpose it is not necessary for inclusion in the studies we conduct.

2.1.1.3 Four Body Improper Dihedral Angle Potential

Improper dihedral angles are designed to ensure that planar groups such as carbon rings remain planar and the ring structures do not pucker or flip. The improper dihedral angle is defined by four atoms not bonded successively to one another, for example; the carboxylate carbon in aspartic acid where the improper dihedral is defined by CG-CB-OD1-OD2. The improper dihedral terms are implemented typically using a harmonic potential, which takes the form shown in Eq. 9, where ϕ is the dihedral angle and ϕ_0 is the equilibrium dihedral angle between four atoms (i,j,k and l).

$$U_{id}(\phi_{ijkl}) = \frac{1}{2}k_\phi(\phi - \phi_0)^2 \qquad \text{(Eq. 9)}$$

2.1.1.4 Four Body Proper Dihedral Angle Potential

As the harmonic description is valid only for small deformations, some force fields account for the anharmonic effects by adding higher order terms in the potential function. For the torsional potential (1,4 interactions) a periodic function provides a better description, which can be shown mathematically by a Fourier expansion as shown in Eq. 10

$$U_d(\phi_{ijkl}) = \sum \frac{1}{2} k_\phi [1 + \cos(n\phi - \delta)] \qquad \text{(Eq. 10)}$$

where k_ϕ is a force constant proportional to the barrier to rotation, n is the periodicity, indicating the number of minima in the function and δ is a phase angle that determines which torsional angle ϕ that corresponds to an minima (optimum value).

2.1.2 Non-Bonded Terms

The interactions discussed in 2.1.1 refer to the bonded interactions, defined by the connectivity of the molecule. Conversely the non-bonded terms are not defined by the connectivity and are instead distance-dependent, calculated as the sum over all atoms with a 4 atom or greater separation. These interactions can be considered to consist of two main factors, firstly the Van der Waals interactions and secondly the electrostatic or Coulombic interactions.

2.1.2.1 Van der Waals Interactions

The Van der Waals interactions consist of a repulsion and a attraction term, which can be described by a simple Lennard-Jones (162), Buckingham (163) or Born-Mayer (164) potential, to name just a few. Despite many functional forms

existing, the relatively simple 6-12 Lennard Jones (L-J) potential is most frequently employed. Its relative simplicity is due to the absence of the need for calculating large number of square roots and exponentials, as is the case in more sophisticated potentials such as the Buckingham potential. The general form of the L-J potential is given by Eq. 11 below:

$$U_{LJ}(r_{ij}) = \varepsilon \left[\left(\frac{r_{ij}^{eq}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^{eq}}{r_{ij}} \right)^{6} \right] \qquad \text{(Eq. 11)}$$

where ε is the energy well depth and r_{ij}^{eq} is the inter-atomic separation for the which the energy is a minimum. The contributing force, with respect to the distance between particle i and j, can be obtained by differentiation of the above equation, such that the force is given by Eq. 12, where C is a constant.

$$F_i(r_{ij}) = \left[12 \frac{C_{ij}^{12}}{r_{ij}^{13}} - 6 \frac{C_{ij}^{6}}{r_{ij}^{7}} \right] \qquad \text{(Eq. 12)}$$

In the Lennard-Jones potential the short-range repulsions are accounted for by the r^{-12} term, whereas the London dispersion-attraction terms are mediated by the r^{-6} term, hence at short distances the repulsive term dominates. The L-J potential goes to zero as r_{ij} increases, so typically cut-off distances are used to truncate the potential to zero more rapidly, increasing the computational efficiency.

2.1.2.2 Electrostatic Interactions

The electrostatic interactions are typically calculated using partial charges at the atom centres with the energy being calculated by using Coulomb's law.

$$U_{el} = \sum \frac{q_i q_j}{r_{ij}} \qquad \text{(Eq. 13)}$$

where q_i and q_j are the charges separated by a distance r_{ij}. This electrostatic term becomes less accurate for highly polarizable groups or ions, where polarisable force-fields need be employed. However, for organic systems such as in proteins, the Coulombic approach is sufficient.

2.1.3 Amber Force-Field

The basic force-field equation implemented in AMBER has the form presented below in Eq. 14. This is the simplest functional form that preserves the essential nature of molecular interactions in a condensed phase at an optimum efficiency.

$$U(r) = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} \left(\frac{V_n}{2}\right)(1 + \cos[n\emptyset - \delta]) + \sum_{nonb} \left(\varepsilon \left[\left(\frac{r_{ij}^{eq}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{eq}}{r_{ij}}\right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right) \quad \text{(Eq. 14)}$$

2.2 Molecular Dynamics Theory

Molecular dynamics simulation is a computational technique originally developed to simulate liquids by modelling them as hard spheres (165). It was later adapted in 1977 to be capable of modelling biological molecules, with McCammon simulating the bovine pancreatic trypsin inhibitor (157). Albeit a very small system of ~500 atoms in vacuum for only 9.2 ps, this simulation was fundamental in showing the dynamic nature of proteins. It was not until 1978 that the first microsecond simulation of a protein in explicit water was conducted (166). Since then advances in both methodology and technology have allowed MD to grow from a complementary tool to a field in its own right (167). With increasing advances in technologies, the capabilities of MD simulations will

continue to grow, for example the use of graphical processing units (GPUs) developed in the past 10 years has resulted in a 3600% speed-up compared to an equivalent number of CPUs (168).

The basic underlying principles of molecular dynamics are fairly simple and are the same regardless of the software package used. Commonly used molecular dynamics software packages such as CHARMM (169), AMBER (170), NAMD (171), DLPOLY and GROMACS (172, 173). All employ the same central physical principle; that the nuclei are heavy enough that the Newton's second law of motion (Eq. 15), which relates the force F experienced by a mass m in motion to its acceleration (rate of change in momentum) applies. With the vector r containing the coordinates of the atoms in Cartesian coordinates such that it is a vector of length $3N_{atom}$.

$$F = ma = m\frac{\delta^2 r}{\delta t^2} \quad \text{(Eq. 15)}$$

Newton's equations of motion relate the force (F) to the changes in atomic positions as a function of time. The evolution of the atomic coordinates can be calculated by integrating Newton's equation of motions simultaneously in small time steps. The force is considered constant during that time step and equal to the negative derivatives of a potential energy function $U(r_1,r_2,\ldots,r_N)$ (Eq. 16).

$$F_i = -\frac{\delta U}{\delta r_i} \quad \text{(Eq. 16)}$$

For i = 1,2,…..N, and where the potential U is the sum of the contributions to the potential energy from both bonded and non-bonded interactions. Through combining Eq. 15 and Eq. 16 we obtain Eq. 17, which relates the change in position of the atom with the derivative of the potential energy.

$$-\frac{\delta U}{\delta r_i} = m_i \frac{d^2 r_i}{dt^2} \quad \text{(Eq. 17)}$$

Given the initial coordinates and velocities, the forces on the atoms determine the new positions and velocities at the subsequent time step, through integration of Newton's 2nd equation of motion.

$$v = \int_t^{t+\Delta t} \frac{d^2 r}{dt^2} = \int_t^{t+\Delta t} \frac{1}{m} \frac{\delta U}{\delta r} \delta r \quad \text{(Eq. 18)}$$

As the force on the atom depends on the position of all of the other atoms in the system, an analytical solution even for the smallest system is not possible. Instead, a solution must be found through the use of a time integration algorithms.

2.2.1 Time Integration Algorithm

A number of numerical time integration algorithms have been developed for integrating the equations of motion. There are three main elements that need to be considered when choosing which integration algorithm to use; the algorithm should be computationally efficient; as well as preserving the energy and momentum of the system; and allow the adoption of adequately long time steps for integration. It is important that the use of the time integration algorithm gives a true trajectory. A true trajectory is theoretically possible, however, due to the implementation of the mathematics a number of sources of error are introduced. For example, round off errors as a result of the finite size of the floating-point arithmetic on current computers and the errors introduced from the truncation of Taylor expansions. Through careful selection of a time step, the error can be minimised such that the trajectory approximates the true trajectory. Although as the aim of MD is to obtain average behaviour as opposed to absolute system

configurations, this approximate true trajectory following is sufficient as long as a good initial configuration is given.

All the integration algorithms assume the positions; velocities and accelerations can be approximated by a Taylor series expansion. The most commonly used time integration algorithm is the Verlet algorithm (174), of which there are now several variations (175, 176). The basic idea is to write two third order Taylor expansions for the positions r(t), one forward (Eq. 19) and one backward in time (Eq. 20).

$$r(t+\Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 + \frac{1}{6}b(t)\Delta t^3 + \mathcal{O}(\Delta t^4) \quad \text{(Eq. 19)}$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + \frac{1}{2}a(t)\Delta t^2 - \frac{1}{6}b(t)\Delta t^3 + \mathcal{O}(\Delta t^4) \quad \text{(Eq. 20)}$$

Where v is velocities, a the accelerations and b the third derivatives of r with respect to t. Adding the two expressions (Eq. 19 and Eq. 20) gives:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + a(t)\Delta t^2 + \mathcal{O}(\Delta t^4) \quad \text{(Eq. 21)}$$

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} \quad \text{(Eq. 22)}$$

this is the simplest form of the Verlet algorithm, which uses the positions from time (t – Δt) to calculate the new positions at (t + Δt). One problem with this form is that the velocities are not directly generated. Although the velocity is not needed for the time evolution, it is necessary for computing the kinetic and total energy of the system, and hence it has to be computed via a different equation (Eq. 22). More recent iterations of the time integration algorithms reproduce the same trajectory as this method, but their computation of the velocities is more straightforward.

The most widely adopted time integration algorithm and the one of most interest to us, owing to its implementation within Amber12, is the leap frog algorithm. In this method the velocities are first calculated at time (t + 1/2Δt). This is then used to calculate the positions at time (t + Δt); hence the velocities leap over the positions and then the positions leap over the velocities resulting in the name the leap frog algorithm. The velocities are calculated at half integer time steps:

$$v\left(t - \frac{\Delta t}{2}\right) = \frac{r(t) - r(t - \Delta t)}{\Delta t} \qquad \text{(Eq. 23)}$$

where r(t– Δt) is defined as

$$r(t - \Delta t) = r(t) - v(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 - \dddot{r}(\Delta t^3) + \mathcal{O}(\Delta t^4) \qquad \text{(Eq. 24)}$$

and

$$v\left(t + \frac{\Delta t}{2}\right) = \frac{r(t + \Delta t) - r(t)}{\Delta t} \qquad \text{(Eq. 25)}$$

where r(t + Δt) is defined as

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 + \dddot{r}(\Delta t^3) + \mathcal{O}(\Delta t^4) \qquad \text{(Eq. 26)}$$

From these equations we are able to obtain an expression for the new position at (t + Δt) based on the old position and the velocities:

$$r(t + \Delta t) = r(t) + v\left(t + \frac{1}{2}\Delta t\right)\Delta t \qquad \text{(Eq. 27)}$$

With the update in the velocities occurring at half integer time steps given by:

$$v\left(t + \frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + \Delta t \, \frac{F(t)}{m} \qquad \text{(Eq. 28)}$$

The leap frog algorithm will produce a trajectory identical to that produced by the above Verlet scheme. However as the velocities are calculated at half

integer time points, and not at the same time as position, then the kinetic and potential energies are not calculated at the same time, making it impossible to obtain the total energy of the system at the current time step. Instead the total energy comes from an average value over a given time interval.

2.2.2 Canonical Ensemble

In the canonical ensemble, more commonly referred to as the NVT ensemble, the amount of substance N, volume V, and temperature T are conserved. In the NVT ensemble the temperature T can be calculated by Eq. 29, where N_f equals the number of degrees of freedom and k_b is the Boltzmann constant. The energy of endothermic or exothermic processes are exchanged with a thermostat, which adds or removes energy to the system to maintain the temperature around an average. There have been a number of thermostats developed; velocity scaling (177), the Andersen thermostat (178), the Nosé-Hoover thermostat (179), the Langevin thermostat (180) and the coloured-noise Langevin thermostat (181). Below we discuss in greater detail the Langevin thermostat and the Berendsen thermostat, which were used in our studies.

$$T(\Delta t) = \sum_{i=1}^{N} \frac{m_i v_i^2(\Delta t)}{k_b N_f} \qquad \text{Eq. 29}$$

2.2.2.1 Berendsen Thermostat

The Berendsen thermostat maintains the temperature of the system by coupling the system to a thermal bath, which gradually scales the velocities proportionally to the differences between the system temperature and that of

the thermal bath. Mathematically the change in temperature is illustrated in Eq. 30.

$$\frac{\delta T(t)}{\delta t} = \frac{1}{\tau}(T_0 - T(t)) \quad \text{(Eq. 30)}$$

where τ is the coupling parameter between the heat bath and the system, T(t) is the actual temperature at time t, and T_0 is the desired temperature. At the limit where τ equals infinity, the system replicates exactly the isokinetic canonical ensemble although, as τ never reaches infinity, the system only approximates the canonical ensemble. However its relative stability, seen by the absence of large oscillations during its function, makes it an ideal thermostat for early equilibration steps, as the thermostat exponentially decays to equilibrium.

2.2.2.2 Langevin Thermostat

The Langevin is based on the principle that the motion of a large molecule through a continuum of smaller particles, such as a protein through a water environment, will be altered. The water alters the dynamics of the protein, via random collisions and by imposing a frictional drag force on the protein.

For each time step, Δt, the Langevin thermostat changes the equation of motion by introduction of a dampening factor, so that the change in momentum is given by:

$$\Delta p_i = F_i - \gamma_i p_i + f_i(\Delta t) \quad \text{(Eq. 31)}$$

where γ_i is a friction coefficient, with γp_i damping the momenta and f_i is a random force with dispersion σ_i related to the friction coefficient γ_i via (Eq. 32) :

$$\sigma^2 = 2m_i\gamma_i k_b T/\Delta t \quad \text{(Eq. 32)}$$

where m_i is the mass, k_b the Boltzman constant, T temperature and Δt being the time step used in the time integration algorithm. This random fluctuating force represents the thermal fluctuations from the small particles. The random force and the friction coefficient combine to create the correct canonical ensemble.

2.2.3 Isothermal–Isobaric Ensemble

The isothermal-isobaric (NPT) ensemble is a constant pressure extension to the canonical ensemble. In this ensemble, temperature, pressure and number of particles are all kept as a constant. This is the most commonly employed ensemble, as chemical reactions are typically carried out under constant pressure conditions. The ensemble is typically maintained by one of two methods; a weak coupling to an external pressure bath Berendsen barostat (182); or via addition of an extra degree of freedom to the Hamiltonian, Parinello-Rahman barostat (183).

2.2.3.1 Berendsen Barostat

The Berendsen barostat (182) maintains the pressure of the system by a weak coupling of the system to an external pressure bath using the principle of least local perturbation. The coordinates and box vectors are rescaled by a coordinate rescaling factor μ at each MD time point using the relationship shown in Eq. 33:

$$\mu = (1 - \frac{\beta \Delta t}{\tau_p}(P_0 - P))^{\frac{1}{3}} \quad \text{(Eq. 33)}$$

where P_0 is the applied external pressure, Δt is the size of the MD time increment, β is isothermal compressibility, τ_p is a time constant and P is the

internal system pressure. Where the instantaneous internal system pressure P is given by Eq. 34, where V is the system volume and $f(r_{ij})$ is the force exerted on particle i by j. The change in pressure over time is proportional to a diminishing approximation of the pressure relative to a reference pressure P_0, given by Eq. 35.

$$P = \rho T + \frac{\sum_{i>j} f(r_{ij}) \cdot r_{ij}}{3V} \quad \text{(Eq. 34)}$$

$$\frac{\delta P}{\delta t} = \frac{P_0 - P}{\tau_p} \quad \text{(Eq. 35)}$$

This relationship results in the phenomenon that the average pressure of the system is correct, however as τ_p never goes to infinity, the approach will never yield an exact isothermal-isobaric ensemble (183).

2.2.3.2 Parrinello-Rahman Barostat

The Parrinello-Rahman barostat (183) conversely does theoretically yield an exact isothermal isobaric ensemble, by allowing the volume and the shape of the cell to fluctuate, as well as through the extension of the Hamiltonian. The Hamiltonian (Eq. 36) is extended by inclusion of a thermal reservoir term s and a friction parameter γ:

$$H = K + U + K_s + U_s \quad \text{(Eq. 36)}$$

Where K is the kinetic energy and U the potential energy terms, with the equation of motion (Eq. 37) showing that the acceleration of an particle i is reduced by a factor given by $\gamma \frac{\delta r}{dt}$.

$$\frac{\delta^2 r_i}{\delta t^2} = \frac{F}{m_i} - \gamma \frac{\delta r}{dt} \quad \text{(Eq. 37)}$$

2.2.4 Constraint Algorithm

In an ideal scenario the time step used in the molecular dynamics simulations should be small enough to capture all of the intramolecular atomic interactions, including the higher frequency bond vibrations, especially those to the light elements. However, even for the simplest of molecular systems, this would be an excessive drain on computer resources.

Often, in order to use longer time steps, the highest frequency motions, such as the bonds to light elements, need to be eliminated. The most common way to do this is to constrain the bond lengths to those elements, typically hydrogen, through the use of a number of different algorithms such as SHAKE (184), or linear constraint solver (LINCS) algorithm (185). The algorithms work by making a correction to the set of new atomistic positions for all atoms connected by the constrained bonds. Throughout our studies, all hydrogen – heavy atom bonds are constrained using the AMBER12 implementation of SHAKE. SHAKE works by making a modification to the Leapfrog time integration algorithm, the velocities are first generated for the unconstrained system, then are modified iteratively until the constraint is satisfied, with each iteration the velocities and positions are updated by adding a contribution due to the restoring forces.

2.2.5 Periodic Boundary Conditions

Another approach to minimise the computational cost of simulations of bulk periodic structures is the use of periodic boundary conditions. Periodic boundary conditions are applied to a finite particle system to mimic an infinite system. The simulation box is replicated though space to form an infinite lattice by rigid translation in all three Cartesian directions, completely filling space.

Such that each of the surrounding boxes are exact copies in every detail of the simulation box, including the velocities and concentrations of impurities. A consequence of this is that a particle moving out of the unit cell is replaced by an identical atom at the opposite face of the cell. The use of periodic boundary conditions has two additional effects; firstly the number of atoms within the cell is conserved; secondly that no atom feels surface forces, as the surface of the cells are removed by the presence of the periodic boundary conditions. During a molecular dynamics study an atom may interact with an atom in a neighbouring cell (which is an image of an atom within the simulation cell), as it is within the cut-off radius. It will then ignore the equivalent atom in the simulation cell, as it will be too far away, thus the interaction is always calculated with the closest image. This is known as the minimum image convention. For this reason, when choosing the cut-off radius for the simulation, it must be less than half the lowest dimension of the simulation cell.

2.2.6 Solvent Models

In computational modelling of biomolecules, a realistic representation of the local environment surrounding biomolecules is necessary, crucially an accurate model for the solvent-molecule interactions. There are two main approaches, which can be employed; explicit and implicit solvent models.

In implicit solvent models, solvent effects are simulated as a perturbation to the gas-phase behaviour of the system (186). The introduction of additional equations to model the mean field effect of the solvent whilst simultaneously reducing the number of degrees of freedom that need to be simulated, resulting in greater computational efficiency. A variety of implicit solvent models have

been developed over the years, all with slight variations on the same key principles (187–191). One of the major drawbacks of implicit solvent models is their inability to reproduce the microscopic features of the solvent environment. Additionally, in the context of collagen modelling, we have seen in Chapter 1, the importance of solvent mediated hydrogen bonding in the stabilisation of the triple helical structure; it is for this reason that all of the simulations are conducted using an explicit solvent approach.

Explicit solvent approaches include water molecules physically within the simulation cell. This allows for the solvent-solute interactions to be considered explicitly. Despite water being a small molecule, its inclusion within molecular dynamics simulations is not simple owing to its complex behaviour; it is for this reason that a number of water molecules have been developed over the years. The most commonly adopted water models include; Berendsen's single point charge model (192); and Jorgensen's TIP3P (193), TIP4P (193) and TIP5P (194) models. The TIP3P water model assumes a rigid geometry with three atom-centred partial charges, which are exactly balanced between the positive charge on the hydrogen atoms, and the negative charges on the oxygen atom. Additionally the TIP3P model contains only one Van der Waals interaction site, localised on the oxygen atom. The TIP4P model builds on this idea by still maintaining a rigid geometry but, instead of having the negative charge centred on the oxygen atom, it is instead at a point along the bisector of the HOH angle, closer to the hydrogen atoms. The TIP5P model localises the negative charges to the lone pairs on the oxygen atoms, resulting in 4 point charges and one Van der Waals interaction site being considered within the model, hence it's name TIP5P. Explicit water molecules within the system increases the number of degrees of freedom and thus leads to a slower statistical convergence of

molecular properties. It was for this reason that we have chosen to use TIP3P water throughout our simulations owing to its simplicity, and its greater computational efficiency that makes it amenable to studies of larger systems.

2.3 Optimisation Algorithms

Optimisation is the general term used to describe the process of finding a stationary point of a function, which in most cases is a minimum with a first derivative of zero and values for the second derivative being positive. Potential energy surfaces are not always a simple function, but instead are multi-dimensional functions, which may contain many points at which the first derivative is zero. Therefore the potential energy landscape can be made up of multiple minima. The lowest value is the global minimum with all the others referred to as local minima.

The simplest reasonable approach to minimising a function is the simplex method, which uses function values to construct an irregular polyhedron in parameter space and then moving this polyhedron towards the minimum whilst allowing the size to fluctuate to improve the convergence (195). However it becomes too slow for multi-dimensional functions, so is not used for minimisation of the potential energy surface in MD studies, although it has been implemented in the refinement of force-field parameters in a number of studies (196, 197).

Owing to the multivariate nature of modelling techniques, most methods assume that at least the first derivative of the function, with respect to all the other variables, the gradient g, can be calculated directly. Additionally the function and derivatives are calculated with a finite precision, which depends on

the computational implementation. Consequently, a stationary point cannot be located exactly but the gradient can be reduced to certain limits within the cut-off value. Hence if the gradient is reduced to within certain limits, or if the difference in the values between two steps is less than a threshold tolerance, then it can be deemed that the optimisation has converged. First order methods use the energy gradient (steepest descent and conjugate gradient), while second order methods use the second derivatives (Newton Raphson).

2.3.1 Steepest Descent Algorithm

The steepest descent method is a first order method and uses the first derivative of the function to determine the direction towards the minimum. It is named steepest descent owing to the fact that the direction of the first minimisation is in the direction opposite to which the gradient vector g is largest. This can be defined as d = -g, where d is the direction vector for the line search. Once the function begins to increase an approximate minimum can be determined by interpolation between the two points. At this new point a new gradient is determined and the step repeated, with the d being orthogonal to the previous direction.

The method is fast, easy to apply and, if the minimum exists, then the method is guaranteed, given an infinite number of iterations to find it. However there are a couple of drawbacks to the method. Firstly, as the line searches are always perpendicular to one another, if there is a gradient component along the previous search direction, which could further lower the function in that direction, it is ignored, hence the algorithm has a tendency for each iteration to partly spoil the function lowering from the previous step. Additionally the

algorithm begins with a reasonable convergence, but as the minimum is approached the rate of convergence slows down, with the line searches crawling toward it with an ever-decreasing speed. Therefore it is most commonly adopted as an initial method to relax a poor starting point, owing to the fact it is one of the few methods guaranteed to lower the function. However it is nearly always used in conjunction with another method, which can converge to the minimum in close proximity quicker.

2.3.2 Conjugate Gradient Algorithm

The conjugate gradient method is another first order method which tries to overcome some of the limitations present in the steepest descent method, namely the partial undoing of the previous step. The first step is performed in a the same manner as the steepest descent step. However the next stage incorporates a small portion of the previous direction in the next search to prevent the oscillating back and forth that can sometimes be present in the steepest descent approach. The direction vector of the conjugate gradient search vector can be defined by Eq. 38, where the value for β can be derived by a variety of methods. β defines the degree of weighting of the previous directions placed on the current direction.

$$d_i = -g_i + \beta_i d_{i-1} \qquad \text{(Eq. 38)}$$

This method allows a more rapid movement towards the minimum once in close proximity, giving greater convergence properties. Structures far away from the minimum the convergence is much slower, hence typically a combination is used in such circumstances.

2.4 Steered Molecular Dynamics

Steered molecular dynamics (SMD) simulations apply an external force to the system, to a particular atom or group of atoms, to probe their mechanical properties or to accelerate processes that are otherwise too slow to model via conventional MD simulations, owing to the timescales on which they occur. SMD bears similarities to umbrella sampling techniques (198–200), which also aim to guide the system through an event that occurs on a long timescale. However SMD does not require equilibrium simulations. Instead perturbation of the system via a constant force result in simulations deviating far from equilibrium, and hence they must be analysed as such (201). The main principle behind SMD is to apply an external factor to guide the system from one state into another. There are two types of SMD; constant force pulling and constant velocity pulling. During constant force SMD, the selected atoms are subject to a fixed constant force in addition to the force generated from the force-field potential. Constant velocity pulling is of particular interest as it mimics the implementation of experimental techniques such as atomic force microscopy (AFM) and optical tweezers, giving complementary atomistic scale detail on the response of the system to an a mechanical load. It is for this reason that it is this implementation that will be used throughout our investigations.

In constant velocity SMD the defined atom, or the centre of mass of a group of defined atoms of the protein, is harmonically restrained to a dummy atom (a point in space) via a virtual spring, with spring constant k, which is then moved along a given vector ($\vec{n}$) at a defined constant velocity (v)(171). The movement of the dummy atom results in the defined pulling group of atoms experiencing a resultant force that depends linearly on the distance between the dummy atom and pulling group. This may result in the pulling group following along the same

vector. The force can then be extracted from the output of simulations to be able to probe the mechanical properties with its relationship to the potential energy shown below in Eq. 40.

$$\vec{F}_{SMD}(\vec{r}) = -\nabla U_{SMD}(\vec{r}) \quad \text{(Eq. 39)}$$

$$U_{SMD}(\vec{r}) = \frac{1}{2}k[vt - (\vec{r} - \vec{r}_0)\vec{n}]^2 \quad \text{(Eq. 40)}$$

Analysis of the energetics of non equilibrium structures, is also possible using the Jarzynski relationship (202). This is based on the assumption that, when some external parameters of a system are changed infinitely slowly, then the total work done on a system is equal to the free energy difference between the initial and final states.

2.5 Electronic Structure Methods

Unlike the force-field approaches discussed above, electronic structure methods apply the laws of quantum mechanics to consider the electrons explicitly. The electronic and structural properties of a system, with M nuclei and N electrons, can be calculated by solving the Schrödinger's equation. The time independent Schrödinger's equation in its barest form is:

$$\widehat{H}\Psi = \mathrm{E}\Psi \quad \text{(Eq. 41)}$$

where $\widehat{H}$ is the Hamiltonian, a differential operator which represents the total energy of the system, Ψ the wave function and E is the energy of the system. The non-relativistic Hamiltonian operator $\widehat{H}$ consists of the sum of the kinetic and potential energy operators of all of the particles within the system, both

electrons and nuclei. This is defined by Eq. 42:

(Eq. 42)

$$\left[-\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2-\sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2-\sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}}+\sum_{i=1}^{N}\sum_{j>1}^{N}\frac{1}{r_{ij}}+\sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_AZ_B}{R_{AB}}\right]\Psi=E\Psi$$

where the indices i and j are electrons which run from 1 to N, while the indexes A and B, are nuclei enumerated from 1 to M. M and Z are the mass and the atomic number of the nucleus A and r_{ia}, r_{ij} and R_{ab}, are distances which define the electron-nucleus, electron-electron and nucleus-nucleus separations respectively. ∇^2 represents the Laplacian operator, a differential operator, with ∇_i^2 and ∇_A^2 involving the second derivative with respect to the coordinates of the i^{th} electron and the A^{th} nuclei. The equation shows that the Hamiltonian is the sum of all of the contributions to the total energy including; the kinetic energy of all the electrons and nuclei (term 1 and 2); the Coulomb attraction between electrons and nuclei (term 3); the Coulomb repulsion between all unique electrons (term 4); and all unique pairs of nuclei (term 5).

The Schrödinger's equation can be solved exactly for hydrogen atoms and other one-electron systems. However for systems of two or more electrons approximations are required. The forces on both nuclei and electrons due to the electric charge are of the same order of magnitude, thus the change in momenta as a result of this force must also be equal. Hence it can be assumed that the momenta of both electrons and nuclei are of a similar magnitude, yet the mass of the nuclei is significantly larger than the mass of an electron (for hydrogen the mass ratio is larger than 1800), hence the nuclear motion must be much slower to be of similar momenta. Thus when solving the time independent Schrödinger equation, the nuclei can be considered as stationary and that the

electrons will relax to the instantaneous ground state. This approximation makes it possible to separate the electronic and nuclear coordinates in the many-body wavefunction, which reduces the problem to the solution of the dynamics of the electrons in some frozen configuration in the nuclei (the Born-Oppenheimer approximation) (203). When the approximation is included, the wavefunction can be obtained by solving the electronic Schrödinger's equation (Eq. 44), the kinetic energy of the nuclei and the coulomb repulsion term for the nuclei are now constant and hence are omitted from the Hamiltonian.

$$\widehat{H}_{elec} = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \qquad \text{(Eq. 43)}$$

The solution of the Schrödinger's equation involving the electronic Hamiltonian then becomes:

$$\widehat{H}_{elec}\Psi_{elec} = \mathrm{E}_{elec}\Psi_{elec} \qquad \text{(Eq. 44)}$$

where Ψ_{elec} is the electronic wavefunction:

$$\Psi_{elec} = \Psi_{elec}\left(\{\vec{X}_i\};\{\vec{R}_A\}\right) \qquad \text{(Eq. 45)}$$

and E_{elec} is the electronic energy, which depends parametrically on the nucleus coordinates:

$$\mathrm{E}_{elec} = \mathrm{E}_{elec}\left(\{\vec{R}_A\}\right) \qquad \text{(Eq. 46)}$$

From solution of the electronic Schrödinger's equation in a fixed nuclear configuration $\{\vec{R}_A\}$, we can obtain the potential generated by the electrons on the nuclei, which allows calculation of the forces acting on the nuclei. This force represents the basis of geometry optimisation and *ab initio* molecular dynamics. The surface defined by Eq. 46, is the Born Oppenheimer surface at which at

optimal geometry of a system the surface is at a minimum. Even with the use of the Born-Oppenhemier approximation, solving the electronic Schrödinger's equation is still a difficult task, even for simple systems.

The main difficulty is the presence of the interaction term $\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}}$. Reformulation of the multi-electron problem is possible, into a series of one electron problems, which define for each ith electron of the system an effective potential $\hat{h}_i$ associated with the presence of the other electrons this allows the N-electron Hamiltonian $\hat{H}_{elec}$, to be rewritten as the sum of N single-electron Hamiltonians $\hat{h}_i$ (Eq. 47)

$$\hat{H}_{elec} = \sum_{i=1}^{N}\hat{h}_i \qquad \text{(Eq. 47)}$$

There are two common approaches to reformulating the multi-electron problem, either Hartree Fock theory and Density Functional Theory

2.5.1 Hartree Fock Theory

The Hartree Fock (HF) theory uses the original multi-electron wavefunction for an atom as the product of one-electron orbitals $\Psi_i(x)$ in the following way:

$$\Psi\,(x_i, \dots, x_N) = \Psi_i(x_1)\Psi_j(x_2)\dots\Psi_k(x_N) \qquad \text{(Eq. 48)}$$

The wave function of the system is also shared in one-electron functions $\Psi_i(N)$ called spin orbital, which is the product of a spatial orbital and a spin function. This function is called the Hartree product $\Psi^{HF}\,(x_i, \dots, x_N)$. Electronic wavefunctions are antisymmetric with respect to the exchange of labels, but the Hartree product does not satisfy this anti-symmetry principle. These defects were corrected by Slater (204), in the self-consistent-field-approach, in which

determinant functions are used to introduce the asymmetry of the wave function.

$$\Psi^{HF}(x_i, \dots, x_N) = \frac{1}{\sqrt{N!}} \det \begin{pmatrix} \Psi_1(x_1) & \cdots & \Psi_N(x_1) \\ \vdots & \ddots & \vdots \\ \Psi_1(x_N) & \cdots & \Psi_N(x_N) \end{pmatrix} \qquad \text{(Eq. 49)}$$

This is the so called Slater determinant in Eq. 49, where $x \equiv \{r, \sigma\}$ is the set of all spatial and spin coordinates of one electron. This form is not arbitrary, it is the simplest form that holds the antisymmetry principle for the wavefunction: when the coordinates of two electrons are exchanged in the wavefunction, the result equals the original wavefunction but with the opposite sign. This principle is just a mathematical implementation of Pauli's exclusion principle, which states that two electrons in a system cannot have identical quantum numbers.

The corresponding energy of the wavefunction can then be calculated by:

$$E = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} = \frac{\int \Psi^* \hat{H} \Psi \, d\tau}{\int \Psi^* \Psi \, d\tau} \qquad \text{(Eq. 50)}$$

where Ψ^* is the complex conjugate of Ψ, and the integration is with respect to the three spatial coordinates and the one spin coordinate for each electron. Additionally the variational principle states that the energy calculated from the above equation (Eq. 50) must be greater than or equal to the true ground state energy E_0 (Eq. 51).

$$E = \frac{\langle \Psi | \hat{H} | \Psi \rangle}{\langle \Psi | \Psi \rangle} \geq E_0 = \frac{\int \Psi_0 | \hat{H} | \Psi_0}{\int \Psi_0 | \Psi_0} \qquad \text{(Eq. 51)}$$

ψ can be chosen to be expressed using an orthonormal set owing to the fact that the value of the determinant is unchanged by any non-singular linear transformations. A Lagrange multiplier ε_i normalises the ψ.

$$\frac{\delta E}{\delta \psi_i} = 0 \qquad \text{(Eq. 52)}$$

This reduces to a set of one-electron equations of the form:

$$\hat{F}\psi_i = \varepsilon_i \psi_i \qquad \text{(Eq. 53)}$$

which is called the Hartree Fock equation, where the Fock operator $\hat{F}$ is defined as:

$$\hat{F} = -\frac{1}{2}\nabla^2 - \sum_{A=1}^{M} \frac{Z_A}{|r-R_A|} + \hat{v}^{HF} \qquad \text{(Eq. 54)}$$

the Fock operator $\hat{F}$, describes the effect of all particles, including the electrons, on the solution of a one electron from the system. The Fock operator consists of three key terms; the first term is the kinetic energy of the electron; the second is the interaction of the electron with the nuclei; the third is the exact-exchange operator $\hat{v}^{HF}$ representing the average potential experienced by the ith electron due to the presence of the other electrons. The exact-exchange operator is due to the interaction of the electron with the cloud of the rest of the electrons and the repulsion between electrons of like spin, which is a consequence of the antisymmetry principle.

The Hartree Fock approximation describes the motion of one electron in the average field created by all other electrons of the system. The difference between the exact total energy and the total energy given by the HF method is generally called the correlation energy. Although the magnitude of the correlation is not always significant and can sometimes be neglected, in other systems it can play an important role and different post-HF methods have been developed to deal with this discrepancy, such as DFT.

2.5.2 Basis Sets

Basis sets are used as an approximation to the true orbitals with mathematical functions. Larger basis sets approximate more accurately the orbitals by imposing fewer restrictions on the locations of electrons in space. With basis sets, assigning basis functions to each atom within a molecule to approximate its orbitals, with these basis functions themselves being composed of a linear combination of Gaussian functions. These basis functions are called contracted functions, with their component Gaussian functions referred to as primitives. These localised orbitals have a maximum concentrated on a nucleus and decay to zero at infinite distance. The mathematical form of the Gaussian type orbital is shown below in Eq. 55, where N is a normalisation constant, r,x,y,z are the Cartesian coordinates, and ξ is a parameter which determines the height of the exponential function.

$$\chi^{GTO}(x,y,z) = Nx^{l_x}y^{l_y}z^{l_z}e^{-\xi r^2} \qquad \text{(Eq. 55)}$$

A variety of different basis set types exist; minimal basis sets, split valence basis sets, polarised basis sets and basis sets with diffuse functions. Minimal basis sets contain the minimum number of basis functions that are needed for each atom. Split valence basis set uses more basis functions per atom. They can commonly be double zeta valence basis sets from the linear combination of two sets of functions per atomic valence orbital, or triple split valence basis sets which uses three sets of contracted function per valence orbital type. Polarised basis sets improve split valence basis sets by adding orbitals (of different shapes), with angular momentum greater than is required for a proper description of the ground state, such as adding d-functions to carbons atoms and p-functions to hydrogen atoms. Diffuse functions are added to basis sets

occupy a larger region of space, which is important for systems where electrons are far from the nucleus.

Chapter 3 Parameterisation of a Force Field for Glucosepane and DOGDIC

3.1 Introduction

The fundamental basis for force field methods is atoms, with electrons not considered as individual elements. This means that the bonding information must be provided explicitly in the form of a force field, rather than being the result of solving the electronic Schrödinger equation. It has long been recognised that the accuracy of the force field is fundamental to successful application of the computational methods.

There are two main methods to derive the force field; either by fitting to quantum mechanical data or fitting to experimentally obtained data. If fitting to experimental data care has to be taken to fit the data to experiments, which reflect the kind of applications the force field will be used in. For example data on geometries of small molecules is best taken from gas phase structural studies, such as microwave or electron diffraction. Solution-based data, should be avoided owing to the unknown influence of the environment. Other sources of experimental data, such as spectroscopic data, can be used to provide information on rotational barriers and vibrational frequencies. Thermodynamic data, such as enthalpy of formation, can also be used although care must be taken to determine the group contributions. This approach requires having a suitable amount of experimental data to generate a reliable force field, which may not be possible for reactive/unstable compounds.

Data from *ab initio* calculations can also provide data for fitting a force field to, which is useful when experimental data is inadequate or lacking. The obvious

limitation to this approach is the quality of the DFT method and basis set used in the calculations. However with significant amounts of computational resources this becomes less of a problem. Force constants can be calculated from diagonalisation of the quantum mechanical hessian obtained from frequency calculations on the optimized structure. If the intra-molecular portion of the force field is expressed in terms of bond, angle or dihedral (molecular internal coordinates) force constants, then a significant problem arises in ensuring invariance of the individual force constants with respect to the given internal coordinate, as a molecule can be described by a variety of sets of internal coordinates, all of which may give different values for the force constant. Two main methods have been developed to overcome this problem in the parameterisation; Seminario developed a method in 1996 that is fully invariant to the internal coordinates used (205) and hence allows full parameterisation direct from the results of the quantum mechanical calculation; Ayers *et al.* (206) developed a iterative program which uses a Molecular Mechanics Matrix based on the gaff force field (207) as well as the quantum mechanical hessian to derive intra-molecular force constants independent of the internal coordinate. In addition to the force constants and equilibrium term values, point charges are also required for a complete force field. To date there is still not a unified approach between the different published force fields. However the basic approach is to derive the charges from fitting to an electrostatic potential surrounding the molecule (208).

3.1.1 Amber Force Field

The Amber force-fields are among the most widely used for biomolecular simulations; with the original 1984 article currently being the 10th most-cited in the history of the Journal of the American Chemical Society (209). The early Amber force field was initially derived for simulating the structures, conformational and interaction energies of proteins, nucleic acids and a small number of related organic molecules. Over proceeding years it has been extended to include force field terms for carbohydrates (GLYCAM force field) (210), and lipids (Lipid14) (211). The GLYCAM force field focussed, not only on development of the bonded terms, but optimisation of the terms for the glycosidic torsion angles as well as the non-bonded interactions. The initial motivation for the development of the Amber force field was to accurately describe structures in the condensed phase with a simple, transferable and general model. The Amber force fields all work by defining atom types and parameterising these for atoms that are both chemically and physically alike.

The initial Amber force field, developed by Weiner *et al.*, generated its bonded terms using microwave and X-ray data on compounds which corresponded to fragments of the amino acids they were being developed to model (209). The initial equilibrium bond lengths and angles were taken directly from this data and the initial force constants were generated through a linear interpolation algorithm between the pure single bond and pure double bond. These initial equilibrium lengths, angles and force constants were then adjusted as necessary to reproduce experimental normal mode frequencies. The electrostatic potential derived charges were computed at the HF/STO-3G level of theory, where STO-3g is a minimal basis set with 3 primitive Gaussian functions fitted to one slater type orbital. A decade on the Weiner *et al.* force

field was replaced by a second generation force field developed by Cornell *et al.*, ff94 (208). The major change in this new, and subsequent iterations of the fixed charge force fields, were the restrained electrostatic potential (RESP) method for charge derivation at HF/6-31G* level of theory (212, 213). The use of the 6-31G* basis set alone (ESP-fit) leads to a uniformly overestimate of molecular polarity, however it gives excellent reproduction of condensed phase inter-molecular properties (214). The use of a 6-31G* ESP-fit does suffer from two drawbacks; there is considerable variation in charges depending on the conformation of the molecule, and less than ideal charges for “buried” atoms are given (208). Given these deficiencies, the RESP approach was implemented, via a two-stage process, using a least-squares fit of the charges to the molecular electrostatic potential (MEP), with the addition of hyperbolic restraints on non-hydrogen atoms, thus reducing the charge on interior atoms of the molecules which can be reduced to more reasonable values, such as buried carbons. The second stage is to fit methyl groups, which require equivalent charges on hydrogen atoms which are not symmetrically equivalent (212, 213). R.E.D. is a web-based server which conducts an automated implementation of the methodology for RESP charge derivation, as was conducted in the Cornell force field production, from a PDB structure (215, 216).

As very few studies have been conducted on AGEs cross-links previously, especially computational investigations, there is a real need to create reliable parameters for these structures. Implementation within Amber should be efficient, owing to the organic nature of the cross-links. However we will opt to parameterise them using a similar approach to the ffXX force fields of Amber. This is owing to the fact the use of GAFF force field (207) for the AGE, would result in covalent bonds forming between the collagen and AGE which would be

described by two different force fields, which were generated using different philosophies, and hence may introduce inconsistencies into the system. The ff99SB force field was chosen owing to its reduced computational expense when compared to the polarisable Duan *et al.*, ff03 force field (217, 218). We intend to use the collagen specific additional terms for hydroxyproline in our collagen simulations which are only implemented in ff99SB force field (219). Through the use of Gaussian09 (220) and R.E.D. tools (215, 216) we will generate reliable parameter files for the most abundant AGE glucosepane and another lysine-arginine derived cross-link DOGDIC.

3.2 Methodology

To generate the structural data for the creation of the parameters for the two AGE cross-links, we begin by loading a lysine and arginine residue, with their Cα atoms separated by 10 Å, into the molecular modelling and building software Avogadro (221). This is the starting point for the building of both of our AGE structures. The cross-links are then drawn between the two residues in Avogadro before a universal force field minimisation (UFF) is conducted to optimise the geometry of the drawn AGE, reducing the bond lengths to within a reasonable range (222). An acetyl group is added onto the backbone N-terminus and a methyl amide group is added to the C-terminus of the amino acid residues. This is for generation of more accurate RESP charges for the backbone atoms, taking into account the peptide backbone.

The two structures are then used in HF calculations, conducted using Gaussian09, to further optimise the geometries, as well as for computing the MEP around the optimised geometry.

3.2.1 Gaussian Methodology

A tight convergence geometry optimisation of the two AGE models was performed, using the HF method with the 6-31G* basis set, implemented in Gaussian09 (223). The 6-31G* basis set is a split valence double zeta polarized basis set, which adds 6d functions to the first row elements, including orbitals with angular momentums greater than is required for a proper description of the ground state. The calculation was run with a charge equal to 0 and a multiplicity equal to one. Confirmation that a minimum for the energy of the optimised geometries had been reached was proven via the absence of imaginary modes in frequency calculations of the structure.

3.2.2 PyRED Server

The QM output files, in addition to the optimised geometry from the Gaussian calculations, are then uploaded to the R.E.D. Server in the form of a PDB file and P2N file (215, 216). This implements mode 2 of R.E.D. server in which no re-optimisation is performed and only the MEP computation and charge fitting steps are performed using the output from the previous Gaussian simulation. In the System.config file we specified for RESP-A1 charge fitting, which corresponds to the RESP fitting algorithm implemented by Cornell *et al.*, as previously described in 3.1.1, with a weighting factor of 0.0005 and 0.0010 on non-hydrogen atoms. These settings are the same as were used for generation of all of the amber FF fixed charge force fields (208, 224–226), hence an additional term (FFPARM= AMBERFF99SB) is necessary to specify for output of the ff99SB atom types. R.E.D. server will then output the RESP charges, with the atom names being the same as those specified in the input PDB file. In

addition PyRED will make an assignment of the atom types based on the optimised structure before generating two frcmod files, one containing known parameters and one containing unknown.

3.2.3 Integration into AMBER

Manual verification of the PyRed assigned atom types is conducted to ensure the most suitable assignment has been made. If changes are made to the assigned atom types, then the parmchk script part of the AmberTools12 package can be used to extract the minimum number of unknown suitable parameters (where connectivity between atom types is not present in the standard Amber libraries) in the ff99SB force field, for our systems. The bond and angle terms still missing from the force field are manually inserted into the frcmod file, using the equilibrium values for the bond or angle from the QM optimised geometry and a force constant, based on an analogous structure already present in, ideally, the ff99SB force field. However, if necessary, the force constant can be used from another fixed-point charge force field within the Amber group. After the frcmod, containing all of the additional bond, angle and dihedral data, is complete a library file is generated for the new residue. The library file is generated by loading in both the frcmod file and the mol2 file of the optimised geometry, which contains the atom types and RESP charges, into tleap. The library file contains a template of the new residue, including the geometry and connectivity data as well as the amber atom types and RESP charges.

After identifying suitable values for the missing parameters for our models, these parameters were added into a force field modification file. The file was

then loaded into tleap to confirm that each of the systems could be successfully used to generate a corresponding input and topology file.

3.3 Parameters Developed

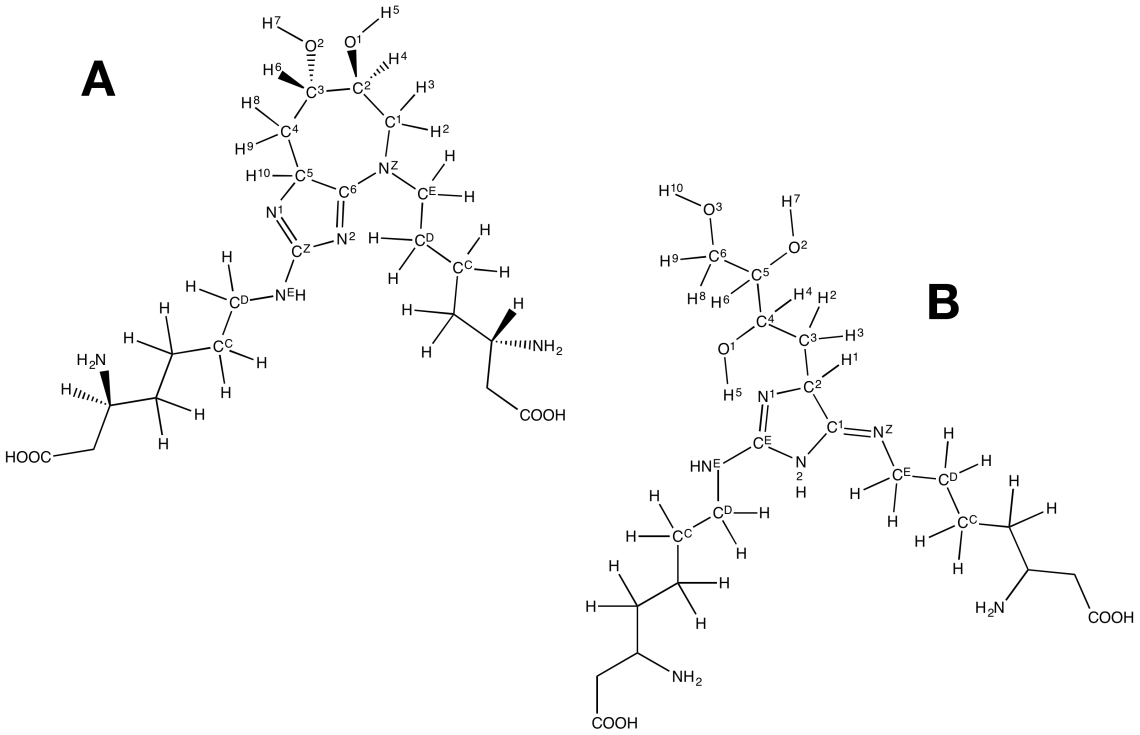


Figure 7: Schematic representation of (A) glucosepane and (B) DOGDIC with parameterised atoms labelled with atom names.

When implementing the derivation of the library files, it was decided, for ease of use, that the AGE cross-linked system would be split across three residues; the cross-linked lysine, the cross-linked arginine and the sugar derived cross-link. For the glucosepane cross-linked system, the system is separated into three residues; Arc – cross-linked arginine; LYC – cross-linked lysine and ORG – Glucosepane, for DOGDIC the three residues are named ARD, LYD and DOG respectively. This would require the use of the cross-link command in LeaP, during the parameterisation, to ensure the cross-link is inputted correctly.

However it will lead to more logical implementation of these for inter-molecular cross-link studies, where the backbone-backbone separation might vary. Figure 7 shows the structure of the two AGEs being parameterised, with the atom names shown in superscript. These names correspond to the atom entries in the library file, as shown in Figure 8. The full library files and frcmod files are available in the Appendix. What can be seen from the portion of the library file, shown in Figure 8, is the successful implementation of the RESP fitting, with the two asymmetrical hydrogen atoms HD2 and HD3 in the ARC residue exhibiting the same charge. Additionally the net charge on the cross-linked three residues sum to an integer value of 0.000, which is a requirement of an Amber force field.

```
!!index array str
 "ORG"
!entry.ORG.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "C1" "CC" 0 1 131072 1 6 0.64350
 "C2" "CT" 0 1 131072 2 6 0.078300
 "H2" "H1" 0 1 131072 3 1 0.038300
 "H3" "H1" 0 1 131072 4 1 0.038200
 "C3" "CT" 0 1 131072 5 6 0.080800
 "C4" "CT" 0 1 131072 6 6 -0.024700
 "C5" "CT" 0 1 131072 7 6  0.8421
 "H8" "HC" 0 1 131072 8 1 0.029100
 "H9" "HC" 0 1 131072 9 1 0.029100
 "C6" "CT" 0 1 131072 10 6  0.89940
 "H10" "H1" 0 1 131072 11 1 0.052600
 "O1" "OH" 0 1 131072 12 8 -0.389500
 "H4" "H1" 0 1 131072 13 1 0.061600
 "H5" "HO" 0 1 131072 14 1 0.209900
 "O2" "OH" 0 1 131072 15 8 -0.389300
 "H6" "H1" 0 1 131072 16 1 0.061900
 "H7" "HO" 0 1 131072 17 1 0.209900

!entry.ARC.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "CD" "CT" 0 1 131072 11 6 0.048600
 "HD2" "H1" 0 1 131072 12 1 0.068700
 "HD3" "H1" 0 1 131072 13 1 0.068700
 "NE" "N2" 0 1 131072 14 7 -0.529500
 "HE" "H" 0 1 131072 15 1 0.345600
 "CZ" "CA" 0 1 131072 16 6 0.77990
 "NH1" "N2" 0 1 131072 17 7 -0.862700
 "NH2" "N2" 0 1 131072 18 7 -0.862700


!entry.LYC.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "CD" "CT" 0 1 131072 11 6 -0.047900
 "HD2" "HC" 0 1 131072 12 1 0.062100
 "HD3" "HC" 0 1 131072 13 1 0.062100
 "CE" "CT" 0 1 131072 14 6  0.154300
 "HE2" "H1" 0 1 131072 15 1 0.113500
 "HE3" "H1" 0 1 131072 16 1 0.113500
 "NZ" "NT" 0 1 131072 17 7 -0.185400
```

Figure 8: Figure depicting relevant section of the new library file generated for the glucosepane cross-link, which is split into three residues; ORG = Glucosepane, ARC= cross-linked arginine and LYC - cross-linked lysine. The columns from left to right contain; atom name, amber atom type. The next two columns are; unused, residue number, atom number in residue template, element, RESP charge.

```
!!index array str
 "DOG"
!entry.DOG.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "C1" "CC" 0 1 131072 1 6 0.596400
 "C2" "CT" 0 1 131072 2 6 0.467300
 "C3" "CT" 0 1 131072 3 6 0.045500
 "C4" "CT" 0 1 131072 4 6 0.070620
 "H2" "HC" 0 1 131072 5 1 0.033200
 "H3" "HC" 0 1 131072 6 1 0.033200
 "C5" "CT" 0 1 131072 7 6 0.106100
 "O1" "OH" 0 1 131072 8 8 -0.387700
 "H4" "H1" 0 1 131072 9 1 0.062100
 "C6" "CT" 0 1 131072 10 6 0.115600
 "O2" "OH" 0 1 131072 11 8 -0.387000
 "H6" "H1" 0 1 131072 12 1 0.064200
 "H8" "H1" 0 1 131072 13 1 0.058800
 "O3" "OH" 0 1 131072 14 8 -0.389400
 "H9" "H1" 0 1 131072 15 1 0.058800
 "H1" "H1" 0 1 131072 16 1 0.085900
 "H5" "HO" 0 1 131072 17 1 0.210400
 "H7" "HO" 0 1 131072 18 1 0.210800
 "H10" "HO" 0 1 131072 19 1 0.21000


!entry.ARD.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "CD" "CT" 0 1 131072 11 6 0.048600
 "HD2" "H1" 0 1 131072 12 1 0.068700
 "HD3" "H1" 0 1 131072 13 1 0.068700
 "NE" "N2" 0 1 131072 14 7 -0.729500
 "HE" "H" 0 1 131072 15 1 0.345600
 "CZ" "CA" 0 1 131072 16 6 0.779900
 "NH1" "N2" 0 1 131072 17 7 -0.69610
 "NH2" "N2" 0 1 131072 18 7 -0.69610
 "HH2" "H" 0 1 131072 19 1 0.457800

!entry.LYD.unit.atoms table  str name  str type  int typex  int resx
int flags  int seq  int elmnt  dbl chg
 "CD" "CT" 0 1 131072 11 6 -0.047900
 "HD2" "HC" 0 1 131072 12 1 0.062100
 "HD3" "HC" 0 1 131072 13 1 0.062100
 "CE" "CT" 0 1 131072 14 6  0.154300
 "HE2" "H1" 0 1 131072 15 1 0.113500
 "HE3" "H1" 0 1 131072 16 1 0.113500
 "NZ" "N2" 0 1 131072 17 7 -0.585400
```

Figure 9: Figure depicting relevant section of the new library file generated for the DOGDIC cross-link, which is split into three residues; DOG = DOGDIC, ARD= cross-linked arginine and LYD - cross-linked lysine. The columns from left to right contain; atom name, amber atom type, the next two columns are unused, residue number, atom number in residue template, element, RESP charge.

Figure 9 gives a similar portion of the created library file for the DOGDIC cross-link. It can be seen from comparing these figures, the geometry, and connectivity of an atom can have a significant effect on the calculated RESP charges. If we look at NZ in the LYC and LYD residues respectively, we see that the charge difference between the two is -0.4, which results from the difference in the bonding order of the atom. In DOGDIC, NZ is sp^2 hybridised forming a double bond to C^1. In glucosepane NZ is sp^3 hybridised forming two covalent bonds to C^1 and C^6. This illustrates the need to have well optimised structures to ensure accurate charge derivation. It was also found that the C^c atoms on both lysine and arginine, for both AGES, had values very similar;

differing by less than 2%, to those present in their respective non-cross-linked residues. The C^b atoms had almost identical values to their respective atomic positions in the non-cross-linked residues, which will also ensure that there will be no artificial disruption to backbone formation.

After assigning all of the atoms types, as seen in Figure 8 and Figure 9, parmchk was run to ensure there were no missing parameters still present. This resulted in a number of missing parameters arising for the N2, which is due to the fact that this atom type was designed for sp^2 nitrogen atoms within the arginine residue. One of the bonding parameters that were missing is the N2-CC. CC is the atom type for a sp^2 carbon in a five membered ring, with one substituent next to a nitrogen. To find a force constant for such a bonding term we looked at the analogous, CA-N2 bonding term, where CA describes an sp^2 carbon atom in a 6 membered ring with one substituent, giving us a force constant of 481 kcal/(mol $Å^2$). This analogy was chosen as it maintained the N2 atom type, which is a fairly unique atom type, as well as combining it with a carbon atom with the correct hybridisation. A number of bond angles were also missing which tended to involve the CC and N2 atom types. One such example was the CC-CT-H1 bond angle. In this case we used a simple analogy with the non-polar HC atom type, such that the bond angle force constant of 50 kcal/(mol rad^2), is for the analogous CC-CT-HC term.

3.4 Discussion

The approach used to generate the bonded terms of the force field was analogous; a suitable force constant was chosen, from assigning an Amber atom type to the new atoms based on their bonded environment, and then using

the force constant data for bonds between similar atom types. This is the approach many force fields implemented within Amber have employed. Since initially parameterising the cross-links in 2013, Nash *et al.*, have produced a purely quantum mechanically derived force-field for a range of AGEs at HF/6-31G* level theory (227) using new software which employs the Seminario *et al.*, method for force constant derivation (205).

We decided to compare our force field to their new force field, to see if there were significant differences in the force constant values derived, which may indicate un-reliability in either of our methods. The C^D upward portions of the side-chains were used for ease of comparison, between the two force fields. As can be seen in Table 1 below, the majority of the force-constants are within 10% with only 35% of bonding force constants and 8% of bond angle force constants deviating by more than 5%. The results indicate that the analogous approach employed in the production of our force field is reliable. In addition any differences will be relatively minor, when considering the concentration of cross-links compared to the standard amber residues in collagen, 1 in 3000.

	% of FCs 0 - 2% Different	% of FCs 2 - 5% Different	% of FCs 5 - 10% Different
Bonds	40	15	35
Angles	72	12	8

Table 1: Percentage difference in the force constant (FCs) values for the bond and angle terms of the force field for glucosepane developed using the analogy method compared to those derived for glucosepane purely quantum mechanically by Nash *et al.* (227)

Despite the force field being largely reproducible by QM to 10% accuracy, we were interested to see where the large deviation in force constant terms was arising from. To do this we looked at the bonds with the largest deviation in the force constants, these were predominately the carbon-hydrogen bonds. What was apparent from our simulation is that the hydrogens bonded to the same carbon were assigned the same atom type, whereas looking at the QM data from Nash *et al.*, this was frequently not the case. For example the CD-HD1 and CD-HD2 bonds in the QM force field have force constants of 345 kcal/(mol $Å^2$) and 258 kcal/(mol $Å^2$) respectively, compared to the value of 340 kcal/(mol $Å^2$) used for both bonds as in our force field. The likely cause for this is the asymmetry of the hydrogen atoms around the carbon atom, which results in different stretching frequencies and subsequently slightly different force constant values. This asymmetry is neglected in our designation, owing to the same atom type being assigned to both of the hydrogen atoms. However the effect of this variation is minor overall, as the correct C-H bond lengths are maintained throughout MD simulation, using our force field.

3.5 Summary

Through the use of R.E.D. server, the web implementation of R.E.D. tools for RESP charge fitting, we have been able to develop two sets of force-fields for implementation in studies on AGE cross-linking within collagen. Missing force constant terms were derived by using the analogy approach. If the force constant is missing, a value is taken from a similar structure, and the distance/angle value is taken as the equilibrium value from the QM geometry optimisation. The set for glucosepane consists of three residues; ARC the

cross-linked arginine; LYC the cross-linked lysine; and finally, ORG the sugar derived portion of the glucosepane cross-link. For implementation in LeaP, the cross-link command should be used to join the residues. The DOGDIC set was produced with the same philosophy and contains three residues named; LYD, ARD and DOG. The force field generated in this approach gave good agreement with a recently developed wholly quantum mechanically derived force field for glucosepane. The development of a reliable force field for the two lysine arginine derived AGEs will enable a wide variety of computational studies to take place to probe a variety of properties.

Chapter 4 Intra-Molecular Lysine-Arginine AGE cross-linking of Type I Collagen

4.1 Introduction

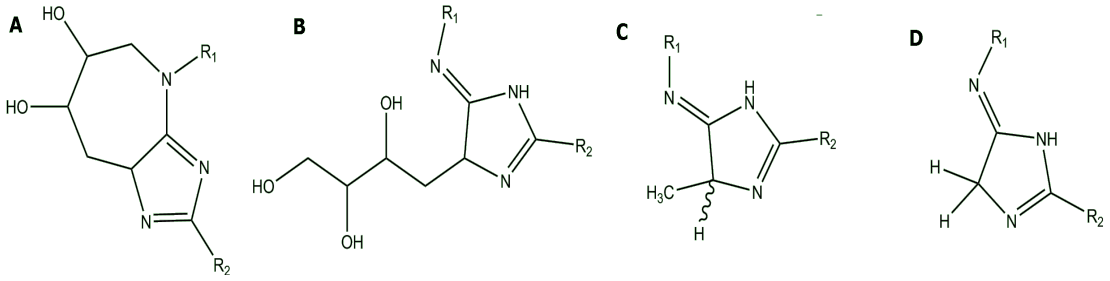


Figure 10: Schematic image of Lysine (R1)-Arginine (R2) cross-linking AGES, A) Glucosepane B) DOGDIC, C) MODIC and D) GODIC.

Although a series of AGE cross-links is possible, the lysine-arginine-glucose AGE cross-link glucosepane is the most abundant in collagen, with levels 100 to 1000 times higher than all other currently known cross-links (228). The concentration of AGE cross-links are found to vary depending on the tissue and AGE type, with glucosepane found in concentrations of 250 pmol/mg in human lens protein whereas MODIC and DOGDIC are found in much lower quantities, < 75 pmol/mg and < 5 pmol/mg, respectively (100). One theory to explain the abundance of glucosepane, is that the final carbonyl rearrangement from the Amadori product undergoes a non-reversible dehydration step, which ultimately leads to an accumulation of glucosepane. However AGEs formed by other glycation agents are formed reversibly (113). In this chapter we will focus on AGEs which form between lysine and arginine residues, more specifically the AGEs which form primarily from D-glucose products via the Schiff base, glucosepane and DOGDIC, as opposed to those that form via other methods including, glucose degradation products methylglyoxal and glyoxal.

Although the absolute levels of AGE cross-links are important, the exact position within the collagen molecules is critical in determining the impact that they have on the collagen properties. In this study we have used a fully atomistic model of an entire collagen molecule in a fibrillar environment to identify, based on energetics, the residues responsible for forming intra-molecular glucosepane and DOGDIC cross-links. We discuss how the identified potential sites of glucosepane formation might cause disruption of the biological function of collagen, and give a structural rationale for their preference for particular sites within the collagen molecule.

4.2 Methodology

An all-atom model of a *Rattus norvegicus* type I collagen molecule, exploiting periodic boundary conditions to replicate the fibrillar environment, was used to study the energetics of AGE cross-link formation.

4.2.1 Building the Model

Our model based on the previous model of Streeter *et al.* (78), uses the amino acid sequence for *Rattus norvegicus* owing to the availability of the crystal structure and similarity to the *Homo sapiens* sequence, making it suitable for this study. A straight-chained structure of a collagen molecule, with the correct helical propensity, was generated using the Triple Helical Building Script (THeBuScr) (229). The primary sequences of the collagen peptide chains α1 and α2, translated from the genes COL1A1 (P02454) and COL1A2 (P02466) (230), were the inputs. The primary sequences used included the post-translational modified residues such as hydroxyproline and hydroxylysine, which

were present in the UniProt entries. A custom script was used to combine the output from THeBuScr and the fibrillar arrangement taken from Protein Data Bank entry 1Y0F (229, 231). The supramolecular structure in 1Y0F contains the Cα atomic coordinates of each amino acid as determined by low-resolution X-ray diffraction experiments (21). The combined model had the helical propensity from the THeBuScr output and the supramolecular positions from the crystal structure. The linear telopeptides and side chain atoms were finally added using LeaP, part of the AMBER12 software package (232).

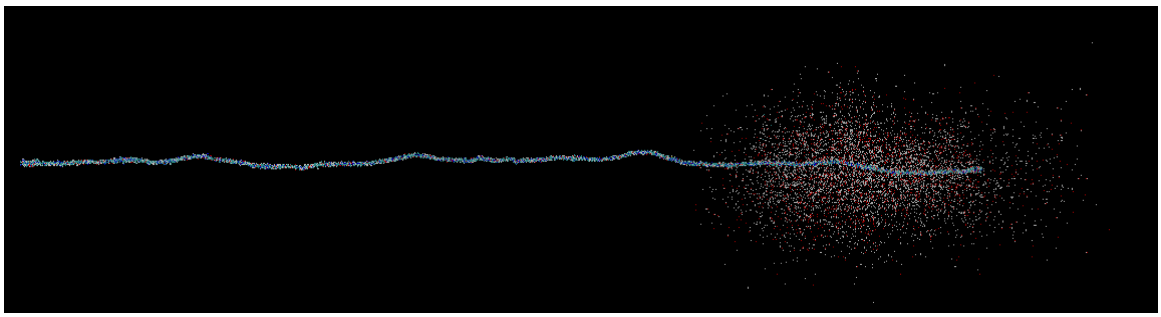


Figure 11: Image depicting the collagen model used, with the water molecules highlighting the unit cell dimensions employed to model the D-period.

The triclinic unit cell dimensions came from the low resolution X-ray diffraction experiment of the collagen molecule (21). The triclinic unit cell has dimensions 39.97 Å, 26.95 Å and 677.90 Å for edges *a*, *b*, and *c* respectively, and 89.24°, 94.59° and 105.58° for angles α, β and γ, respectively. The unit cell is long and thin with the value of the c lattice parameter representing the fibril's D period. The length of the collagen molecule is approximately 300 nm in length, meaning the unit cell which describes the system's periodicity is over four times smaller than the collagen molecule itself, this can be seen in Figure 11, by the presence of the water molecule highlighting the unit cell, with the whole image depicting the system explicitly studied. The system was solvated in LeaP by the addition

of 11,980 explicit water molecules into the interstitial gaps between neighbouring collagen molecules. This value is equal to that derived by Streeter *et al*., with the aim of preserving the crystallographic dimensions of the fibril during an isothermal-isobaric MD simulation (78), whilst remaining in good agreement with experimental values for the intra-fibrillar water content of 0.75 g water g^{-1} of collagen (233). All amino acids were assumed to be in their standard protonation states for physiological pH, resulting in 268 cationic sites from the amino acids with acidic side-chains and 235 anionic sites from the amino acids with basic side-chains. The remaining +33 net charge was neutralised by 33 chloride ions per collagen molecule, resulting in an effective chloride concentration of 0.14 M, which is in general agreement with the experimentally observed concentration of 0.1 M sodium chloride (234, 235). Through adoption of this model in which we model a full-length single collagen molecule we are able to reproduce an infinite fibril of type I collagen, which has been proven to give good agreement with experimental structures (236). The model reproduces the D-periodicity experienced within the fibril, seen here in Figure 12, with each colour of the collagen molecule depicting a single unit cell of our collagen system.

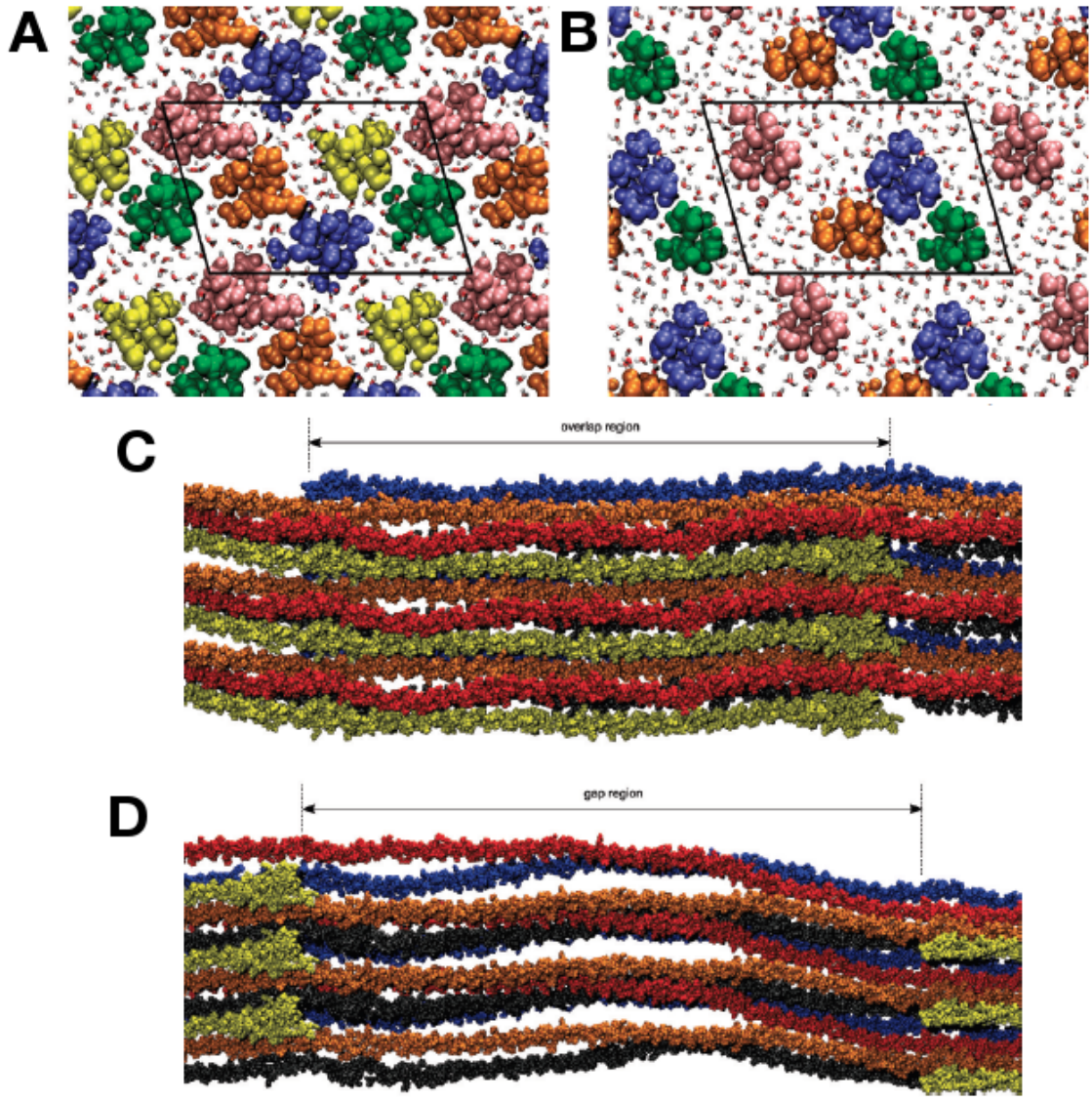


Figure 12: Cross-sectional images of the collagen fibril taken from a molecular dynamics simulations, employing the same unit cell dimensions and MD set-up as our present study. A and B show all atoms within a 5 Å thick slice, including proteins (pink, orange, blue, yellow, and green) and water molecules (red and white). The collagen proteins lie perpendicular to the cross-sectional plane and therefore appear as small clusters of atoms. (a) “Overlap” region of the fibril in which five different collagen proteins pass through the cross section of the unit cell (white quadrangle). (b) “Gap” region of the fibril in which only four collagen proteins pass through the cross section. Image C (“Overlap” region) and D (“Gap” region) shows the longitudinal cross-section, with each image depicting three adjacent unit cells; each protein that passes through the unit cell has a different colour, with the water omitted for clarity. Reprinted adapted with permission from (78). Copyright 2010 American Chemical Society.

A single cross-link was inserted into the collagen molecule between the residues identified during the distance-based criterion search (see 4.2.3),

totaling 24 unique models of collagen molecules with single cross-links for each cross-link type. A native collagen model without cross-links but with an unbound D-glucose; minus four water molecules and minus three water molecules for glucosepane and DOGDIC respectively; were also created to act as a reference for thermodynamic comparison to our covalently cross-linked collagen models. The reference models were generated such that the number of atoms within the system is conserved; this had the added benefit that the enthalpy contribution from forming the covalent bonds is kept very close to zero. This is necessary owing to the implementation of the bond stretching potential within the Amber force field, in which the bonding energy is only given by the energy related to strain away from an equilibrium bond length, such that at the equilibrium bond length the bonding energy contribution is zero. From calculation of the total change in the energetic contribution of the covalent bonds from bond dissociation enthalpy data (237) we were able to determine that the change in enthalpy as a result of the change in covalent bonding is – 2.39 kcal/mol and – 0.72 kcal/mol for glucosepane and DOGDIC respectively.

4.2.2 Modifications to the Amber12 Source Code

A small modification was required to the AMBER12 source code, otherwise the MD algorithm will not progress beyond the first time step within the NPT ensemble, as detailed in the work by Streeter *et al.* (78). Such a problem did not occur in the NVT ensemble. The problem occurred during the resizing of the unit cell and the rescaling of the atomic coordinates, part of the Berendsen barostat algorithm for maintaining constant pressure (182). The problem is

believed to be the result of the unusually long molecule within a periodic unit cell, which is more than four times smaller.

In the standard AMBER12 MD code for isotropic external pressure, the coordinate rescaling factor μ, is calculated after each time step as shown in equation 1. Where P_0 is the applied external pressure, P is the internal system pressure, Δt is the size of the MD time increment, β is isothermal compressibility and τ_p is a time constant.

$$\mu = (1 - \frac{\beta \Delta t}{\tau_p}(P_0 - P))^{\frac{1}{3}} \quad \text{(Eq. 1)}$$

The unit cell is normally rescaled after each MD time step so that a, b and c become μa, μb and μc; correspondingly each atom is repositioned from a position r to a position μr. To overcome the problem the code is modified so that the unit cell parameters a and b are rescaled by μ. C is not rescaled essentially making the calculation a constant pressure simulation, with respect to the x- and y-coordinates, and a constant volume simulation, with respect to the z-coordinates, the periodic box angles not being allowed to deviate from the crystallographic values.

As a result the internal system pressure also had to be calculated slightly differently. The instantaneous pressure in the system is a tensor, P, and it is calculated in the AMBER12 code from the kinetic energy and the forces acting upon each atom, shown in equation 2

$$P = \frac{1}{V}\left(\sum_i m_i v_i v_i^T + \sum_{i<j} r_{ij} F_{ij}^T\right) \quad \text{(Eq. 2)}$$

$$P = \frac{P_{xx} + P_{yy} + P_{zz}}{3} \quad \text{(Eq. 3)}$$

Where V is the volume of the periodic box, m is an atomic mass, v is an atomic velocity vector and r is the vector between two atoms, and F is the force between two atoms. The superscript T represents the transpose of a column vector to a row vector, and the sum is over all the atoms in the unit cell. When the external pressure is isotropic, the internal pressure can also be approximated to be isotropic, and the scalar value for P can be estimated from the elements of the matrix P. In the unmodified AMBER12 code the internal system pressure is given by Equation 3, where the subscripts refer to the matrix elements of matrix P. Hence it can be said that, in the unmodified code, the scalar pressure can be described by the average of the main diagonal elements of the pressure tensor. However for my system, owing to the fact that the atomic coordinates are only rescaled in the x and y direction, the AMBER12 code is modified such that the scalar pressure can be described by the average of the P_{xx} and P_{yy} elements, as shown below in Equation 4:

$$P = \frac{P_{xx} + P_{yy}}{2} \qquad \text{(Eq. 4)}$$

4.2.3 Distance Based Criterion Search

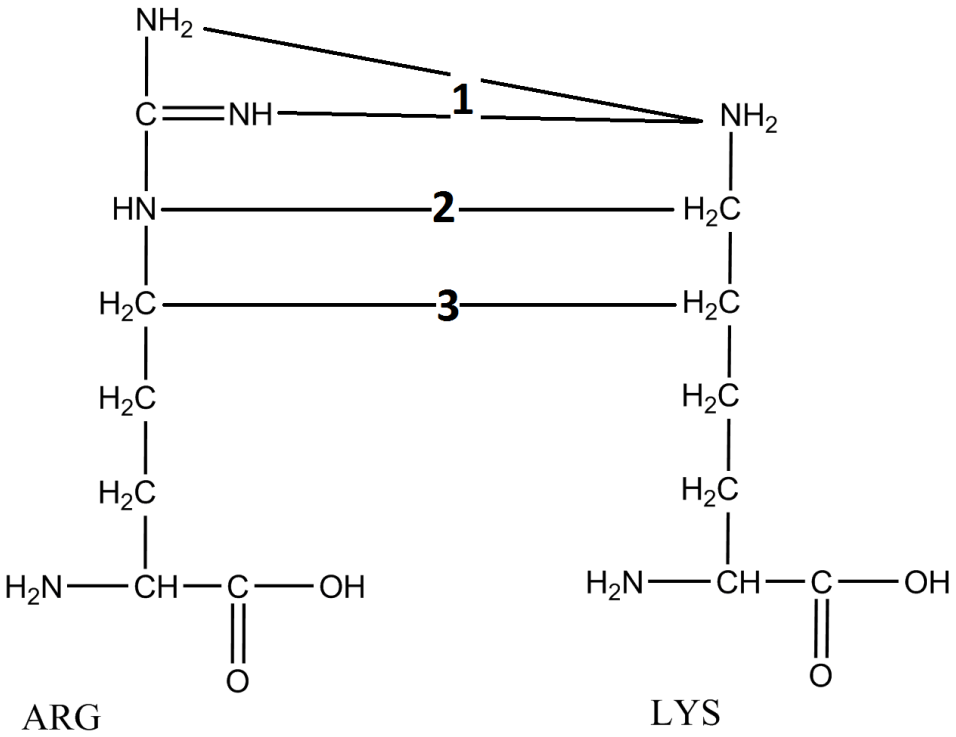


Figure 13: A schematic representation of the three points at which the distance was measured between the lysine and arginine during the distance based criterion search. Measurements are between: 1) arginine N^{η} and lysine N^{ζ}, 2) arginine N^{ε} and lysine C^{ε}, and 3) arginine C^{δ} and lysine C^{δ}.

A custom script was used to scan the triple helical portion of a low energy conformer of a collagen molecule for lysine and arginine residues, within a 5 Å cut-off across separate polypeptide chains, within the collagen molecule. The distance-based criterion of 5 Å was chosen based on two main factors. Firstly, previous studies have suggested that the two residues within proximity of 5 Å are a strong indicator for preferential glycation (238). Secondly, doubling the distance between the nitrogen atoms within glucosepane (approximately 2.5 Å and 3.7 Å) could reveal a reasonable number of potential sites where cross-links are likely to form (238). The distance was calculated at three separate points along the residues' side chains, as shown in Figure 13, namely between the three terminal nitrogen atoms lysine N^{ζ} and arginine N^{η}; the lysine C^{ε} and

the arginine N^{ε}; and lysine C^{δ} and arginine C^{δ}. Any site where at least one distance criterion was met was considered for cross-linking.

4.2.4 Molecular Dynamics Simulation Detail

MD simulations were performed on all models using SANDER, part of the AMBER12 software package (232). Periodic boundary conditions were applied to the unit cell in order to simulate the densely packed fibrillar environment. The ff99SB force field was used for the parameterisation of the collagen molecule with additional terms based on published values for hydroxyproline (225). Water molecules were represented using the TIP3P model (219). The ff99SB force field was parameterised specifically for biological molecules and describes the non-bonded interactions by pairwise additive Lennard-Jones 6-12 potentials and pairwise additive coulombic potentials. Coulombic potentials were calculated using the Particle Mesh Ewald summation with a cut-off radius of 8.0 Å (239).

A time step of 2 fs was adopted for all MD simulations and hydrogen-bond lengths were constrained using the SHAKE algorithm (184). This time step was chosen, owing to it's widespread usage in MD studies of collagen (68, 78). Constant temperature and pressure was maintained with the Berendsen algorithm (182), using a barostat time constant of 5.0 ps atm^{-1} and a thermostat time constant of 1.0 ps. As the periodic unit cell has a *c* lattice parameter much larger than *a* and *b*, it is more appropriate to use anisotropic coordinate rescaling than isotropic rescaling, for maintaining constant pressure. This was achieved by making a small modification to the AMBER code, the details of which are discussed in 4.2.2. To reduce instabilities in the starting structure, the

models underwent 5000 steps of steepest decent energy minimisation followed by a further 5000 steps of conjugate gradient minimisation. The system was heated to 310 K for 120 ps using the NVT ensemble followed by a further 320 ps using the NPT ensemble. Production simulations ran for 60 ns at 310 K using the NPT ensemble. Analysis was performed over the final 25 ns of the production simulations, as this was found in our previous work to give stable energies (78) and a fibrillar arrangement, which is in good agreement with experimental values (21). The system density and the potential energy were monitored up to 35 ns at which point they were shown to converge.

4.3 Results

The distance-based criterion search identified 24 potential lysine-arginine cross-link sites distributed along the length of the triple helical portion of the protein.

4.3.1 Glucosepane Cross-linking

The average binding enthalpies of the glucosepane cross-links were calculated using the total energy over the last 25 ns from each cross-linked collagen simulation, where the average total energy of the native reference collagen model was subtracted from the total energy over the last 25 ns for the cross-linked collagen. The binding enthalpies are reported in Table 2. Residue numbering originates from the original Uniprot sequence data entries. The statistical error of the formation enthalpy was calculated using the standard error of the mean, which was found to be approximately 0.7 kcal mol^{-1} for all of the cross-linked simulations. Sites 1 and 23 were found to have strong steric clashes with neighbouring images of the collagen molecule. It was therefore

decided early on that these cross-links would not form within the fibrillar environment, and hence the simulations were not continued.

Table 2: The difference in enthalpy formation of glucosepane for all 24 identified cross-link sites. The six energetically favourable sites were aligned to ECM binding sites of the human collagen type I sequence. Column 1 gives the site number, columns two to four highlight the cross-linked residue pair between two of the three polypeptide chains (labeled using the Uniprot residue number with the helical residue number shown in brackets) and the fifth column lists the change in enthalpy (kcal/mol).

Cross-link	Chain α1 (a)	Chain α1 (b)	Chain α2	ΔEnthalpy
1	^{229}ARG(62)	^{226}LYS(59)		-
2		**^{257}ARG(90)**	**^{183}LYS(87)**	**-13.572**
3		^{419}LYS(252)	^{348}ARG(252)	+38.54
4	^{458}ARG(291)		^{386}LYS(290)	+7.883
5		^{494}LYS(327)	^{419}ARG(323)	+39.176
6	^{509}LYS(342)		^{438}ARG(342)	+4.357
7	**^{527}LYS(360)**		**^{456}ARG(360)**	**-2.304**
8	^{587}ARG(420)		^{516}LYS(420)	+43.326
9	^{620}ARG(453)		^{549}LYS(453)	+76.636
10		^{646}LYS(479)	^{579}ARG(483)	+4.076
11	^{734}ARG(567)	^{731}LYS(564)		+23.157
12	^{740}LYS(573)		^{669}ARG(573)	+19.280
13$^{\alpha}$	**^{748}LYS(581)**		**^{677}ARG(581)**	**-23.968**
14		^{770}LYS(603)	^{699}ARG(603)	+73.645
15	^{854}ARG(687)	^{851}LYS(684)		+92.728
16	^{896}LYS(729)		^{825}ARG(729)	+55.401
17	**^{958}LYS(791)**	**^{956}ARG(789)**		**-2.315**
18	^{958}LYS(791)		^{884}ARG(788)	+65.516
19	^{1025}ARG(858)	^{1022}LYS(855)		+16.130
20	**^{1055}ARG(888)**		**^{980}LYS(884)**	**-34.501**
21	^{1085}LYS(918)	^{1082}ARG(915)		+21.912
22	**^{1094}ARG(927)**		**^{1020}LYS(924)**	**-36.130**
23	^{1100}ARG(933)		^{1029}LYS(933)	-
24	^{1141}LYS(974)		^{1073}ARG(977)	+90.852

Six potential sites yielded an exothermic enthalpy change on formation of the glucosepane cross-link between the two identified amino acids, all of which occur within regions which are thought to be of biological significance. The details of the biomolecule binding site overlaps for each favourable cross-link site are presented in Table 4. Specifically, site 2 occurs at a position local to the interaction sites of heparan sulfate (HS), α1α1 and α2β1 integrins, and an enzymate mediated cross-link (240–243). Site 7 is within the binding sites of heat shock protein 47 (HSP-47) chaperone and the proteoglycan decorin (241, 244). Site 13 was found within the binding region of phosphoryn, a protein of dentine which plays a role in bone mineralization (245). Site 17 occurs within the binding sites of α2β1 integrin, a HSP-47 chaperone, a fibrillogenesis inhibitor, and is also within close proximity of the binding site of the collagenase Matrix Metalloproteinase-1 (MMP-1) (244, 246). Site 20 is within the binding site of dermatan sulfate (DS) proteoglycan and the secreted protein factor interleukin 2 (IL-2) (241, 247, 248). Finally, site 22 is also within the IL-2 binding domain as well as the binding location of the anticoagulant heparin (243, 248). The remaining 18 sites were found to have energetically unfavourable changes in enthalpy upon cross-link formation.

4.3.2 DOGDIC Cross-linking

Using the same approach for DOGDIC as we used for glucosepane we have been able to determine six sites where DOGDIC formation is an energetically favourable process, the binding enthalpies for which are recorded in Table 3. The same residue numbering system is used, and the statistical error worked out using the standard error of the mean, is comparable at approximately 0.7 kcal mol^{-1} for all of the cross-linked simulations.

Table 3: Difference in the enthalpy of formation for DOGDIC at all 24 identified cross-link sites. The six energetically favourable sites were aligned to ECM binding sites of the human collagen type I sequence. Column 1 gives the site number, columns two to four highlight the cross-linked residue pair between two of the three polypeptide chains (labelled using the UniProt residue number and the triple helical residue number shown in brackets) and the fifth column lists the change in enthalpy (kcal/mol).

Cross-link	Chain α1 (a)	Chain α1 (b)	Chain α2	ΔEnthalpy
1	^{229}ARG(62)	^{226}LYS(59)		+85.78
2		^{257}ARG(90)	^{183}LYS(87)	+109.66
3		^{419}LYS(252)	^{348}ARG(252)	+50.08
4	**^{458}ARG(291)**		**^{386}LYS(290)**	**-8.68**
5		^{494}LYS(327)	^{419}ARG(323)	+20.21
6	^{509}LYS(342)		^{438}ARG(342)	+25.38
7	^{527}LYS(360)		^{456}ARG(360)	+11.61
8	^{587}ARG(420)		^{516}LYS(420)	+72.09
9	^{620}ARG(453)		^{549}LYS(453)	+55.07
10		^{646}LYS(479)	^{579}ARG(483)	+58.68
11	**^{734}ARG(567)**	**^{731}LYS(564)**		**-14.33**
12	^{740}LYS(573)		^{669}ARG(573)	+1.03
13$^{\alpha}$	^{748}LYS(581)		^{677}ARG(581)	+9.03
14		^{770}LYS(603)	^{699}ARG(603)	+30.74
15	^{854}ARG(687)	^{851}LYS(684)		+83.03
16	^{896}LYS(729)		^{825}ARG(729)	+23.38
17	^{958}LYS(791)	^{956}ARG(789)		+53.69
18	**^{958}LYS(791)**		**^{884}ARG(788)**	**-20.38**
19	**^{1025}ARG(858)**	**^{1022}LYS(855)**		**-61.58**
20	**^{1055}ARG(888)**		**^{980}LYS(884)**	**-4.85**
21	**^{1085}LYS(918)**	**^{1082}ARG(915)**		**-1.62**
22	^{1094}ARG(927)		^{1020}LYS(924)	+28.15
23	^{1100}ARG(933)		^{1029}LYS(933)	+3.63
24	^{1141}LYS(974)		^{1073}ARG(977)	+32.15

By transposing the energetically favourable formation sites onto the candidate cell and matrix interaction domain map, a number of overlaps with regions of biological significance were identified. The details of the biomolecule binding site overlaps for each favourable cross-link site are presented in Table 5. Site 4

Intra-Molecular Lysine-Arginine AGE cross-linking of Type I Collagen occurs at the interaction sites of α2β1 integrin, IL-2 and an HSP-47 chaperone (242, 244, 248). Site 11 is local to the binding sites of HSP-47 chaperone, phosphoryn and the proteoglycan keratan sulfate (KS) (241, 244, 245, 247). Site 18 is within the binding sites of α2β1 integrin, a HSP-47 chaperone, which is a fibrillogenesis inhibitor, and is also within close proximity of the binding site of the collagenase MMP-1 (5, 7, 9). Site 19 occurs within the binding site for IL-2, HSP-47 chaperone and the proteoglycan DS (244, 247, 248). Site 20 is within the binding site of DS proteoglycan and IL-2 (247, 248). Finally, site 21 is within the IL-2 binding domain, as well as the binding location of the HSP-47 chaperone (244, 248).

The number of sites identified as being potential sites for DOGDIC formation equaled those identified for glucosepane formation, but only one site (site D20) overlapped, although at this site glucosepane formation is more exothermic. Additionally, despite the force field approach not taking into account covalent bond formation and dissociation explicitly, the use of a carefully selected reference system minimised the net contribution. With the contributions for both glucosepane and DOGDIC being exothermic at -2.39 kcal/mol and -0.72 kcal/mol respectively. Hence the contribution from covalent bond formation will not alter the sign of any of the above identified formation sites. Additionally the lowest energy values for glucosepane formation were in general agreements with the enthalpy change of formation of glucosepane determined in the QM study conducted in the absence of a protein environment by Nasiri *et al.*, for which they report values of – 33.3 kcal/mol (153).

4.4 Biological Implications

It has long been assumed that the presence of AGEs in and between collagen molecules alters their physical properties, as well as increasing the lifetime of collagen molecules within the body. The presence of either a DOGDIC or a glucosepane cross-link at any of the sites identified in this study could result in an alteration in the physical properties of the collagen molecule. In addition, alteration to the collagen could potentially impede essential ECM function from the glycation of side chains in, or close to, binding sites for ECM biological molecules, such as heparin, proteoglycans or collagenases. Sweeney *et al*., created a descriptive map of human type I collagen with marked ECM interaction domains based on an earlier map and database (241, 249).

The amino acid sequence and structural information used in this study was that of *Rattus norvegicus,* owing to the availability of the experimental X-ray diffraction data and the strong biosimilarity between rat and human sequences, (92% and 91% similarity for α1 and α2 chains respectively). Sweeney *et al*., used the same crystallographic data to obtain a descriptive map and then localised functional domains of the human type I collagen onto the rat type I microfibril (241). By using a similar approach we have been able to directly map favourable and unfavourable cross-link sites onto the human sequence and compile a list of ECM interaction sites that may be impeded by cross-linking of collagen. Table 4 and Table 5 summarise the biomolecule binding sites that overlap with the location of the energetically favourable cross-linking sites for glucosepane and DOGDIC respectively.

Table 4: Biomolecule binding sites that overlap with the energetically favorable glucosepane cross-linking sites

Cross-link	Aligned ECM Binding Sites	Enthalpy (kcal/mol)
2	Heat Shock Protein 47 (244), Heparan Sulfate (243), α1β1 integrin (242), α2β1 integrin (242), Enzyme Mediated mature cross-link(240)	-13.572
7	Heat Shock Protein 47 (244), Guanidine extracted decorin(241)	-2.304
13	Phosphophoryn (245)	-23.968
17	α2β1 integrin (242), Heat Shock Protein 47 (244), Matrix Metalloproteinase 1 (246)	-2.315
20	Dermatan Sulfate (247), Interleukin-2 (248)	-34.501
22	Interleukin-2 (248), heparin (243), Amyloid precursor protein (241)	-36.130

Table 5: Biomolecule binding sites that overlap with the energetically favourable DOGDIC cross-linking sites.

Cross-link	Aligned ECM Binding Sites	Enthalpy (kcal/mol)
4	Heat Shock Protein 47 (244), Interleukin-2 (248), α2β1 integrin (242)	-8.68
11	Heat Shock Protein 47 (244), Phosphophoryn (245), Keratan Sulfate PG (241, 250)	-14.33
18	Heat Shock Protein 47 (244), α2β1 integrin (242), Matrix Metalloproteinase 1 (246),	-20.38
19	α2β1 integrin (242), Heat Shock Protein 47 (244), Amyloid Precursor Protein, Interleukin-2 (248), Dermatan Sulfate (247)	-61.58
20	Dermatan Sulfate (247), Interleukin-2 (248), Amyloid Precursor Protein	-4.85
21	Interleukin-2 (248), heparin (243), Amyloid precursor protein (241)	-1.62

An important part of forming mechanically competent collagen fibrils is the formation of chemical cross-links between collagen molecules. These cross-links occur during fibrillogenesis and are mediated by the enzyme Lysyl oxidase. This enzymatic mediated cross-linking is thought to occur from a young age, between α1-^{254}Lys (228, 240), and a lysine residue from a neighbouring molecule. The α1-^{254}Lys residues already involved in enzymatic cross-links are therefore unlikely to be available for glucosepane cross-linking. However the α1-^{254}Lys not involved in the enzymatic cross-link is available for glycation. Another active enzyme in the ECM is lysyl hydroxylase, which hydroxylates lysine residues, producing hydroxylysine during post-translational modification of the collagen molecule. The hydroxylysine residues within our model were included in the distance based criterion search, owing to the availability of their amine group, which is needed to form the glucosepane cross-link (251). Our modelling approach presented in this chapter adopts a homotypic microfibril of type I collagen. However healthy tissue is heterotypic with tendon fibrils, although predominantly type I collagen, it still contains small amounts of other minor collagen types such as type III and type V. Given that our study focuses on intra-molecular cross-links, a heterotypic microfibril composition would have little effect if any on the predicted cross-links locations.

A number of biomolecular interactions identified would remain unaffected by cross-linking, owing to the fact that the process in which they are involved occurs prior to secretion of the procollagen into the ECM. One particular example of this is HSP-47, an intracellular collagen-stress binding protein local to the endoplasmic reticulum and responsible for maturation of a number of types of collagen. HSP-47 binding site overlaps with potential cross-linking sites G2, G7, G17, D4, D11, D17 D19 and D21. However, as it acts within the cell

before glycation can occur, formation at these sites will have no impact on HSP-47 binding.

The energetically favoured cross-link site G13 and D11 may result in structural variation of the binding site of phosphophoryn. The binding of type I collagen to phosphophoryn, the major non-collagenous dentin protein, is believed to play an important part in the nucleation of the mineral phase within the dentin matrix (245). Phosphophoryn has a large number of Asp-Ser-Ser repeats interspersed throughout the molecule and Ser-Asp domains towards the C-terminal, both of which are readily phosphorylated (252). In cases where most of the available Asp-Ser-Ser and Ser-Asp motifs are phosphorylated, the molecule will have a strong negative charge and thus acts as a sink for binding calcium ions, potentially directing the location and moderating the speed of mineralisation within the dentin matrix (253). The effect of cross-link formation on phosphophoryn binding, and thus dentin mineralisation, is dependent on which of the two processes will occur first. The exact point in time at which mineralisation occurs is still unknown, although it is often suggested that biomineralisation will occur shortly after fibrillogenesis (254). This, combined with the fact that rate of formation of AGEs is in the order of weeks, we would expect mineralization to occur before AGEs could form. If this were the case we would not expect cross-links to form at these sites, as AGEs are unable to form in mineralized tissue. Also it has been hypothesized that, due to the much greater remodelling rate and larger abundance of mature enzymatic cross-links in bone, the biomechanical impact of the non-enzymatic AGEs will not be significant (58).

One potential area where the presence of AGEs may be significant is in pre-glycated tissue engineered constructs for bone regeneration, where the presence of such cross-links like glucosepane and DOGDIC, might significantly reduce the extent of mineralization owing to their energetically favourable binding locations G13 and D11 within the phosphophoryn binding region.

Glycosaminoglycans constitute a considerable fraction of the glycoconjugates found in the ECM of virtually all mammalian tissues, where they play a significant role in the biological function of the tissue. Heparan sulphate (HS) and DS serve as: key biological response modifiers by acting as stabilizers; cofactors, and co-receptors for growth factors and cytokines. HS and DS also act as regulators of enzyme activity; signalling molecules in response to cellular damage, such as wounding, infection, and carcinogenesis; and targets for bacterial, viral, and parasitic virulence factors for attachment and invasion (255–257). KS acts as a hydration agent due to its distinct water binding properties, although its anti-adhesive character has also been suggested to play a role in cell migration.

Reigle *et al*. had observed that KS proteoglycans and heparin experience a reduced affinity for glycated collagen, whereas DS proteoglycans exhibit no change in affinity (250). Co-electrophoresis analysis of heparin's affinity for collagen revealed that unglycated collagen had an appreciably stronger heparin binding (K_d of 100 nM) compared to glycated collagen (K_d of 250 nM). The same study found, through blocking fibrillogenesis by casting in agarose gel, that the defective heparin binding found in the glycated collagen is independent of the supramolecular state of the collagen. It is possible that *intra*-molecular glucosepane cross-link formation at G22 could reduce the concentration of

bound heparin. HS is a structural analogue of heparin which is therefore likely to have similar binding mechanism to the collagen. Therefore we could also expect disruption to HS binding at G2, by the presence of a glucosepane cross-link. This is also supported by the work of Reigle *et al.,* which showed that the HSPG also had a reduced affinity for glycated collagen. In addition they proposed that glycation of collagen weakens the KSPG-Collagen interactions *in vivo,* which could potentially be explained by the presence of a DOGDIC cross-link within the KS binding region at site D11. The blocking of heparin or KS from binding due to the presence of AGEs, could happen in two ways; firstly, through steric blocking of the binding site; secondly, by altering the electrostatic potential after the occupation of the lysine side chain during cross-linking. Reigle *et al*. found no reduction in affinity between DS and glycated collagens, which they attributed to heparin and KS proteoglycans having significant electrostatic contributions towards binding, unlike other proteoglycans (250). A combination of a reduced electrostatic contribution to binding, in addition to only the arginine of the pair involved in the cross-link situated within the DS binding region at G20, D19 and D20, suggests that the presence of an AGE in this region would have little effect on the binding of DS. However additional investigation would be necessary to develop this suggestion further, for example through explicit modelling of the molecules and interactions. The presence of a DOGDIC cross-link within the KS binding region at site D11, may also help to explain the experimental observations of Reigle *et al.*, that glycation of collagen may weaken the KSPG –Collagen interactions *in vivo*.

The two major membrane-bound integrins α1β1 and α2β1 are in part responsible for eukaryotic cell-collagen (type I) interaction within the ECM. No favourable cross-linking sites are present in the key collagen-cell interaction

Intra-Molecular Lysine-Arginine AGE cross-linking of Type I Collagen domain. However sites G2, G17, D4 and D18 occur at positions aligned with the cell interaction domain of a neighbouring collagen molecule. Integrins bind using their A-like domain, which contains a trench centred onto a metal ion-dependent adhesion site (MIDAS), in the presence of Mg^{2+} or Mn^{2+}. Upon binding, a glutamine residue becomes attracted to the metal ion in the MIDAS site (258, 259). Due to the size of the A-like domain there are some additional interaction between the integrin and residues in the neighbouring collagen molecules. The introduction of a cross-link into the collagen cross-link sites identified would therefore alter the polarity and structure of the additional interaction sites, potentially leading to a lower binding affinity. If that were the case, one might expect to observe a drop in integrin interactions.

Figure 14: Crystal structure of the immunoregulatory cytokine InterLeukin-2

Type I collagen acts as an extracellular store for bioactive interleukin 2 (IL-2) through reversible binding, thereby increasing the bioavailability of IL-2 in a spatial pattern dictated by the organisation of the collagen molecules (248). The binding of IL-2 has been shown to be very site-specific, with a K_d of

approximately 10^{-2} – 10^{-8} M and molar ratios of four to six IL-2 to one collagen molecule. IL-2 is an important stimulator and modulator of T-cell activation, adopting a key role in the pathophysiology of various immune-mediated diseases such as rheumatoid arthritis, multiple sclerosis and transplant rejection. Somasundaram *et al.,* found that the binding of IL-2 was very sensitive to the collagen sequence and structure of the binding site, although the exact mechanism of attachment is unknown (248). The presence of a cross-link at any of the four favourable formation sites G20, D20, D21 and G22 within an IL-2 binding region may therefore have a significant effect on IL-2 binding, potentially leading to disruption in the attachment of IL-2 to the collagen, which could decrease T-cell response time.

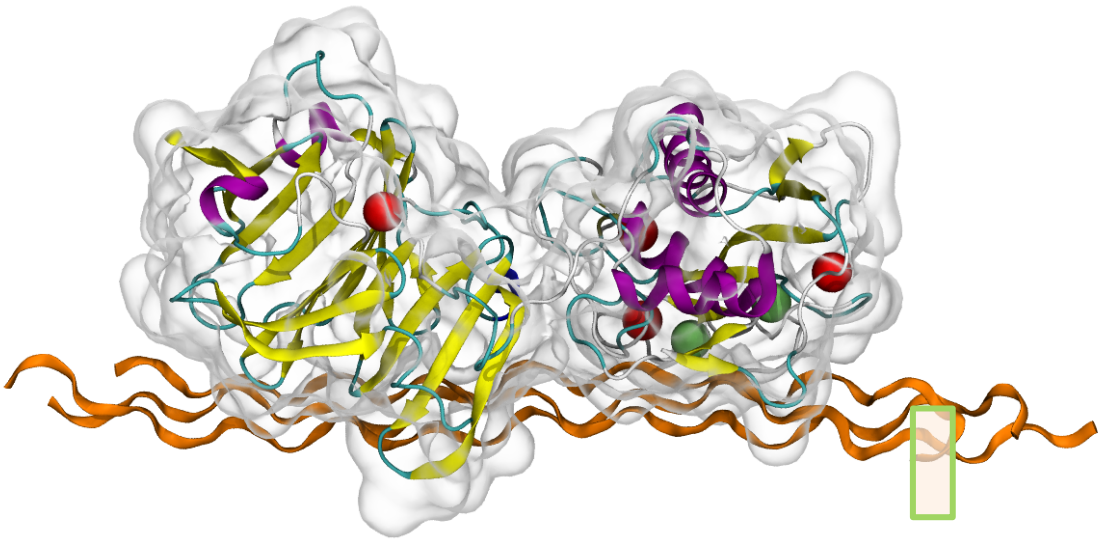


Figure 15: Matrix Metalloproteinase 1 bound to type I collagen, the location of the cross-linking sites in relation to the active site is illustrated by the green box. Hameopexin domain on left closest to N-terminus, catalytic domain to right closest to C-terminus, Zinc ions green and Calcium ions in red.

Another important implication of the findings presented here is the fact that favourable cross-link sites G17 and D18 involve a lysine residue that is only two

Intra-Molecular Lysine-Arginine AGE cross-linking of Type I Collagen residues down from the proposed MMP1 binding site (241, 246), Figure 15. Inhibition of the MMP1 collagenase via collagen cross-linking may significantly affect the digestibility of the cross-linked tissue (260). Although the exact mechanism of action of the MMP1 enzyme has so far not been determined in significant detail, it is known that it operates by first uncoiling the three polypeptide chains before cleaving the peptide bond. Cross-links between polypeptide chains could potentially hinder this uncoiling process (246). By conducting a mutation (^{200}Glu → ^{200}Ala) of the ^{200}Glu residue, which is essential for peptide hydrolysis, Chung *et al*., were able to observe the "unwinding" action of the MMP1 without bond cleavage. This allowed them to determine that "unwinder" MMP1 (E200A) preferentially interacts with the α2 (I) chain, whilst the α1 (I) chains were more exposed and susceptible to a cutter "proteinase" such as serine proteases. This may suggest that the presence of the cross-link would not completely hinder the uncoiling of the polypeptide chains in G17 but would reduce the turnover rate. However for D18 the cross-link forms between ^{958}Lys (α1a) and ^{885}Arg (α2), thus potentially preventing MMP1 from uncoiling, thus having a direct impact on the ability of the enzyme to cleave the peptide bonds. However due to the low concentrations of DOGDIC cross-links within the body, the overall effect of the turnover rate of collagen would be small. In the same study by Chung *et al.*, they showed, through the use of an active site-directed synthetic MMP inhibitor GM6001X, that an unoccupied active site is necessary for unwinding to occur, which they assumed was due to the need to accommodate the α2(I) chain. This could suggest that the binding of MMP1 to collagen could be sensitive to the variation in the local structure at the binding site and thus the presence of the cross-link at either G17 or D18 could have a significant effect on the binding of MMP1 to collagen. Yet without an identified

mechanism for the binding of MMP1 and the uncoiling of the collagen substrate it is difficult to know to what extent the presence of a cross-link affects proteolysis. To determine the exact implications of the cross-linking at site G17 and D18 further explicit modelling of the MMP1 with a glycated collagen would be necessary.

4.5 Structural Implications

A wide variety of structural analysis was conducted in an attempt to elucidate the reasoning why specific sites formed energetically favourable cross-links. Before discussing these in detail, we will discuss the new interactions introduced upon formation of the cross-link.

The six favourable cross-link sites are presented in Figure 16, along with their neighbouring amino acids. It can be inferred that the major contribution to the decrease in enthalpy is an increase in the number of favourable side-chain to side-chain interactions or side-chain to backbone interactions, in addition to the contribution of glucosepane to protein interactions. Here we give an example using the average bond distances of these additional interactions, found after cross-link formation. A side-chain to side-chain interaction found within the immediate vicinity of position 20, is a potential hydrogen bonding interaction between ^{983}Asn (α2) HD22 and ^{1055}Arg (α1a) HE with an average bond length of 2.16 Å, over the last 25ns of the simulation. However most of the additional interactions occurred between side chains and the backbone of the polypeptide. For example, after cross-linking at position 17, an additional hydrogen bonding interaction is formed with an average length of 2.00 Å between ^{956}Arg (α1b) HE and the backbone carbonyl of ^{957}Gly (α1b) O. The formation of a cross-link

Intra-Molecular Lysine-Arginine AGE cross-linking of Type I Collagen results in four additional carbon containing covalent bonds being present in the system, thus lowering the energy of the system. However, the overall enthalpy change between the bound and native models is a combination of the energy released by the formation of the covalent bonds, plus the additional energy contributions from the increased number of interactions, both favourable and unfavourable. In addition, cross-links potentially contribute additional electrostatic interactions with the surrounding residues. One example is the formation of two long-range electrostatic interactions with an average distance of 3.05 Å and 3.22 Å, between the cross-link hydroxyl groups and the carbonyl oxygen of ^{679}Gly (α2) O at position 13.

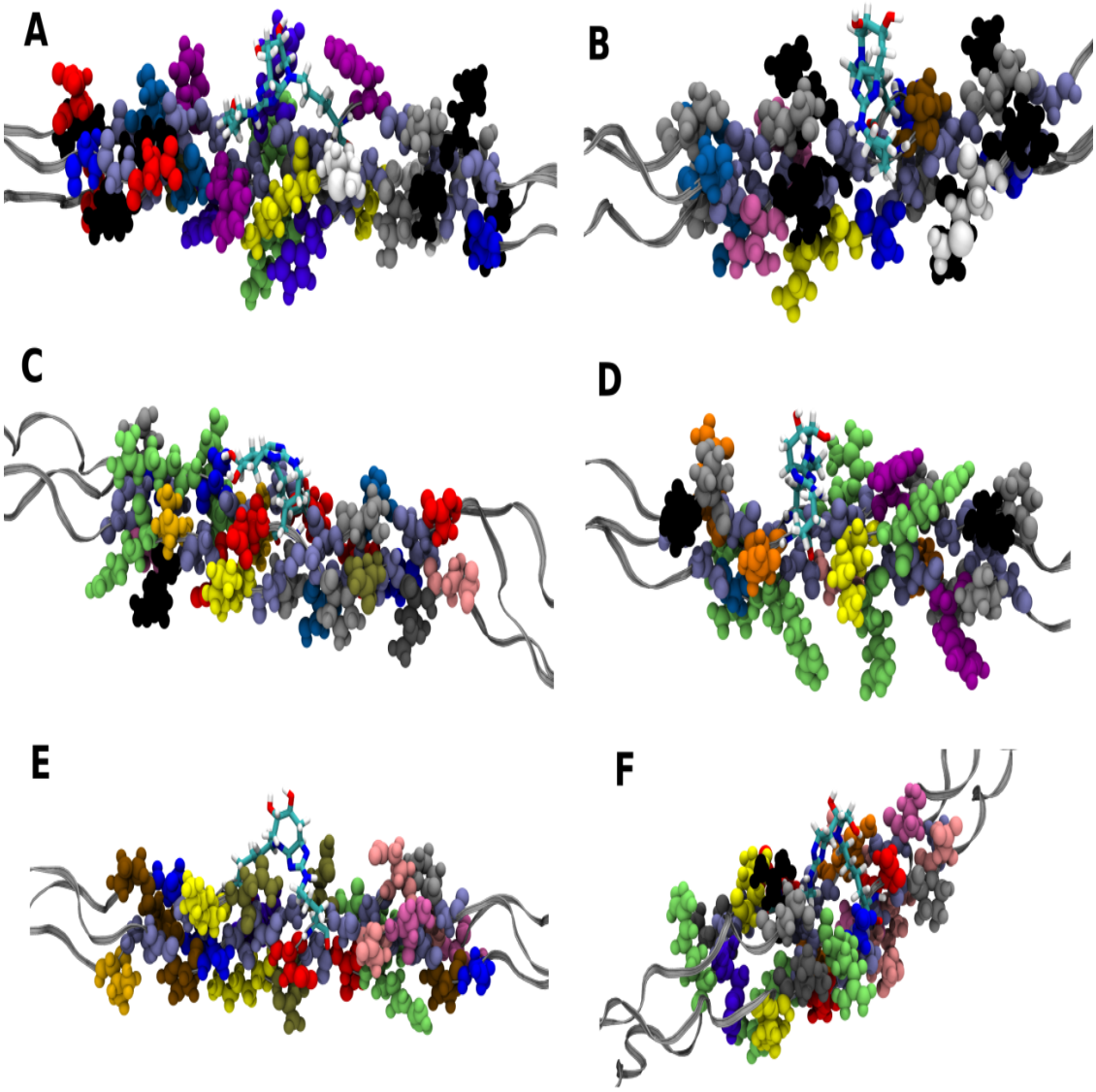


Figure 16: Local environment around the favourable glucosepane cross-link sites a) Position 2, b) Position 7, c) Position 13, d) Position 17, e) Position 20 and f) Position 22. (Residue colours: Ala – Blue; Asn - Tan; Asp – Red; Arg – Lime; Gln – Orange; Glu – Pink; Gly – Ice Blue; His – Violet; Hyp – Silver; Ile – Gray; Leu – Black; Lys – Yellow; Lyz - Yellow; Met – White; Phe – Purple; Pro – Ochre; Ser – Light Blue; Thr – Mauve; Tyr – Magenta; Val – Gold; glucosepane cross-link shown as sticks.

The six favorable cross-link sites for DOGDIC are presented in Figure 17, along with their neighboring amino acids. It can be inferred that, like glucosepane, the major contribution to the decrease in enthalpy is an increase in the number of non-bonded interactions upon cross-link formation and rearrangement of the local environment. Noticeable non-bonded interactions are formed during the simulations, for example in between a backbone carboxyl group and a side-

chain of the cross-linked residue at position 11, where a hydrogen-bonding interaction between ^{734}Arg (α1a) HE and ^{733}Gly (α1a) O with an average bond length of 2.07 Å is formed in the last 25 ns of the simulation. There are also side-chain to side-chain interactions, e.g. a potential hydrogen-bonding interaction at site 21, between ^{1084}Asp (α1a) OD2 and ^{1082}Arg (α1b) HH2, with an average bond distance of 1.75 Å.

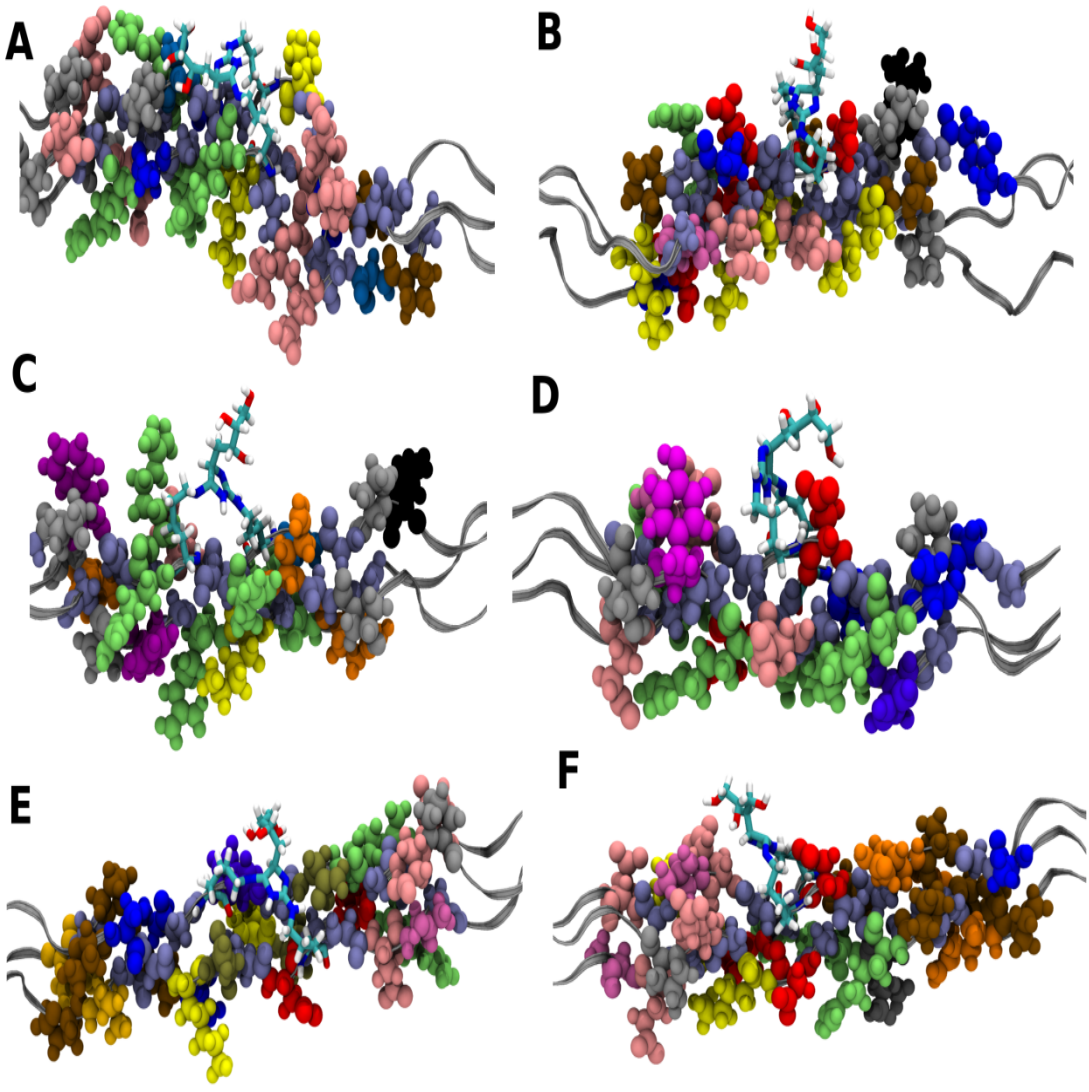


Figure 17: Local environment around the favourable DOGDIC cross-link sites a) Position 4, b) Position 11, c) Position 18, d) Position 19, e) Position 20 and f) Position 21. (Residue colours: Ala – Blue; Asn - Tan; Asp – Red; Arg – Lime; Gln – Orange; Glu – Pink; Gly – Ice Blue; His – Violet; Hyp – Silver; Ile – Gray; Leu – Black; Lys – Yellow; Lyz - Yellow; Met – White; Phe – Purple; Pro – Ochre; Ser – Light Blue; Thr – Mauve; Tyr – Magenta; Val – Gold; DOGDIC cross-link shown as sticks

The interactions observed for DOGDIC differ slightly from those observed for glucosepane in two ways. Firstly, there are non-bonded interactions at the DOGDIC sites, which are not observed at the sites for glucosepane, and occur between the side-chains of two cross-linking residues. For example at site 21 where we see a potential hydrogen-bond between ^{1085}Lys (α1a) O and ^{1082}Arg (α1b), with an average separation of 2.16 Å. Secondly, there are noticeably fewer cross-link to side-chain or cross-link to backbone interactions. This can potentially be explained by two main differences between the two different cross-links; the first is the greater flexibility of the DOGDIC cross-link, in particular the four carbon aliphatic chain; the second is the greater polarity of the DOGDIC cross-link itself, which means that it is more likely to be involved in water-mediated hydrogen-bonding if not involved in direct hydrogen bonding.

Eighteen of the potential 24 cross-link sites, for both glucosepane and DOGDIC, were found to be energetically unfavourable. There are three main reasons for the unfavourable formation enthalpies of some of these binding sites; the local structure of the collagen at the site; the configuration of the binding side-chains; and the presence of steric clashes and close contacts. Close contact can occur between residues within the same collagen molecule, as is the case at G19 where the arginine N^{η} is within 3.5-5.0 Å of the O^{δ} of the neighbouring ^{1024}Asp residue, or between residues on neighbouring collagen molecules within the fibril, e.g. at site G15 where there is a large number of close contacts of 1.5 - 4.0 Å between several positions on ^{851}Lys and ^{1085}Lys on a neighbouring molecule. When a cross-link forms, a rotation of both side chains around their α-carbon may be necessary, which causes the residues to adopt configurations not experienced in the native state. Upon cross-linking the ^{851}Lys's movement is severely restricted and thus the presence of the nearby ^{1085}Lys residue, which

also has its movements limited by nearby residues from neighbouring molecules, causes an increase in energy as the two residues come into close contact. Smaller contributions to the energy arise from the local configuration of the linking side-chains, which arises when, upon cross-link formation, the number of degrees of freedom for the bound residues decreases. For example, in G8 we see a rotation of 180° around the $C^{\beta} - C^{\gamma}$, thus adopting an eclipsed conformation with both the two larger groups located on the same position on their respective carbon atoms. The local structure of the collagen around the binding site dictates the extent to which the two previously described contributors to the energy occur, as the proximity and size of residues near the site can have a large impact on the ability of the residue to form a glucosepane cross-link.

Gautieri *et al.* (261), identified potential sites on a homology model of the human sequence, based upon the amount of time the two potential cross-linking side chains were within 5 Å of one another. Our study builds on this work by conducting fully atomistic MD simulations of the actual cross-linked molecules to ascertain whether the structural rearrangement around the binding site is energetically favorable or not. Direct comparison between the sites identified in our study and those of the previous work is not possible. As the local structure will vary due to the different methods used to incorporate the human sequence and the fact that the explicit cross-links in our study influence the local environment.

What we have found is that not a single analysis technique is able to fully explain the impact of these cross-links. However through the use of a variety of techniques we are able to build a more complete reasoning. Before going

further I want to discuss two concepts I will be using throughout my analysis; the first is the Root-Mean-Square-Deviation (RMSD), which measures the average displacement of the atomic positions relative to a reference, the reference typically being the position of a superimposed protein, or an initial structure. It can be used selectively to probe certain regions of the protein by applying the analysis to only certain atom types, for example, to analyze only the backbone atoms.

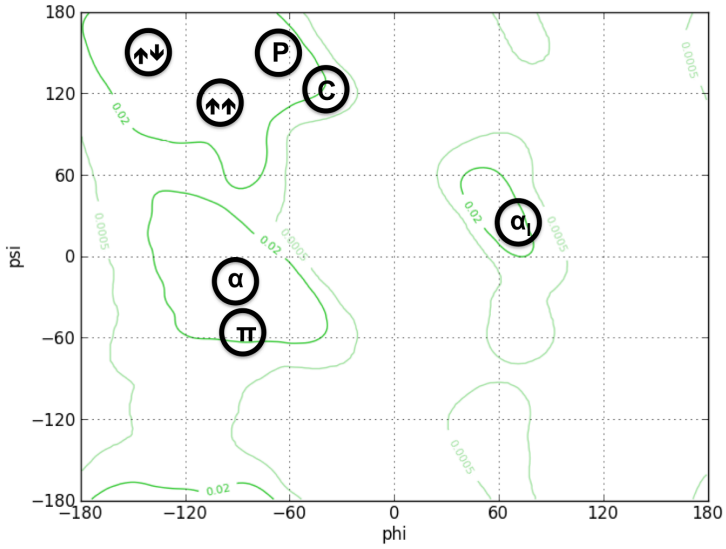


Figure 18: The canonical Ramachandran plot from Ramachandran and Sasisekharan original work with outlines defining the core allowed (dark green lines), and extreme-limit allowed (light green lines) regions for an Ala dipeptide. The widely accepted locations of linear groups, are also shown for the α-helix (α), π-helix (π), left-handed α-helix (α_l), polyproline-II (P), collagen (C), parallel β-sheet (↑↑), and anti-parallel β-sheet (↑↓).

The second method is the Ramachandran plot which is a fundamental tool in structural biology research. Within a polypeptide the main chain N-C_α and C_α-C bonds are relatively free to rotate, these rotations are represented by the dihedral (torsional) angles ϕ and ψ. A Ramachandran plot graphically plots the dihedral angle ψ against ϕ of all the amino acid residues in protein structure. In doing so Ramachandran found areas of the plot where amino acid residues would frequently be located, dependent on their conformation. Figure 18b

outlines these, with the regions of most interest to us highlighted by the C and α symbols on the figure. Ramachandran identified these areas by treating the atoms as hard spheres with a radius equal to the Van der Waals radii. The core areas highlighted by the solid line are where the conformations cause no steric clashes, the dashed lines are where the radius of the sphere is allowed to be slightly less than the Van der Waals radii, and in doing so suffered no clashes (classical region). Areas outside these lines are where clashes occurred in both cases (262). The only exception would be glycine, which is frequently found in areas of the graph owing to its small side chain (R=H). Thus glycine is often found in the turns of proteins where other proteins would normally not be found due to steric clashes of their side chains.

Proteins undergo dynamic fluctuations in structures, mostly from variations in side chain positions. To better understand the overall structure of the protein we also need to analyse the system dynamically. To do this specifically with the Ramachandran plots, we monitor the two dihedral angles ϕ and ψ over the whole analysis trajectory for the regions of interest and then plot these angles for both dihedral bonds in a frequency histogram, which then allows us to use the modal angle in further analysis. The use of frequency histograms also allows us to plot the dihedrals of two systems, i.e. cross-linked and native system, on the same plots, which allows us to see shifts in the dihedrals as a result of the cross-links introduction.

Commonly the largest increases in energies for structural changes of proteins, specifically collagens, are from changes to the conformation of the backbone. The universal approach adopted is measuring the Root-Mean-Square-Deviation (RMSD) of the positions of the atoms, which constitute the protein’s backbone.

As can be seen in Figure 19, the RMSD for both the cross-link and the native system remain fairly equal compared to one another. This is especially clear in Figure 19B, where you can see that the presence of the cross-link does not significantly alter the position of the backbone or its ability to undergo fluctuations in atomic positions, with the RMSD observed occurring in a similar range for the cross-linked (Average RMSD – 1.68 Å) and native collagen models (average RMSD – 1.63 Å). What these RMSD plots clearly show is the dynamic nature of the proteins and hence the need to analyse them accordingly. This is seen by the fact that, for both the native and the cross-linked systems, the RMSD for the whole backbone fluctuates by about 1.6 Å. These fluctuations in positions are just part of natural protein dynamics, as proteins are not static systems.

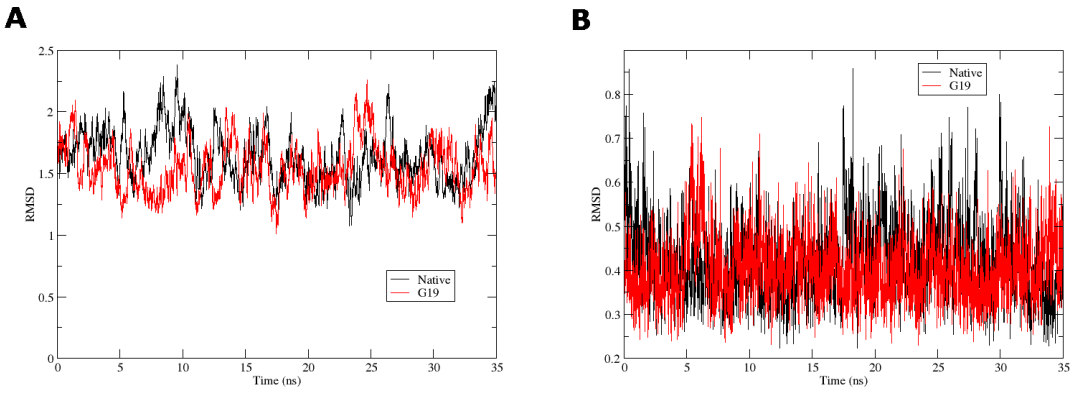


Figure 19: RMSD of the protein backbones relative to their average structure for the native model (black) and the system with a glucosepane cross-link present at site 19 (red) for (A) the whole protein and (B) for 4 residues either side of the cross-linking site

As we saw above, there was no significant deviation in the dynamics of the backbone relative to their dynamics of a native collagen molecule. One idea I believed warranted further investigation was that there could be a torsional tension introduced into the backbone upon introduction of a glucosepane cross-

link into the system. To investigate this further I decided to select two potential cross-linking sites identified in the distance-based criterion search, one which had an unfavourable formation enthalpy G9, and one which had a favourable formation enthalpy G20, for cross-linking. At the end of the 60 ns simulation, we set-up two further simulations for each site. In one we ran the system as is for a further 4 ns and in the other we removed the cross-link from the system, running the simulation for a further 4 ns. What we would expect to observe if there was a build-up of torsional tension in the helix, is that upon removing the cross-link, there would be a significant shift in the positions of the atoms in the backbone. This would be shown as an initial large increase in the uncoiled RMSD, which gives a maximal difference in values between the cross-linked and uncross-linked system. The reference frame used for the RMSD is that of the last frame in the 60 ns trajectory, i.e. the point in time at which the cross-link is removed. The results of which are presented in Figure 20.

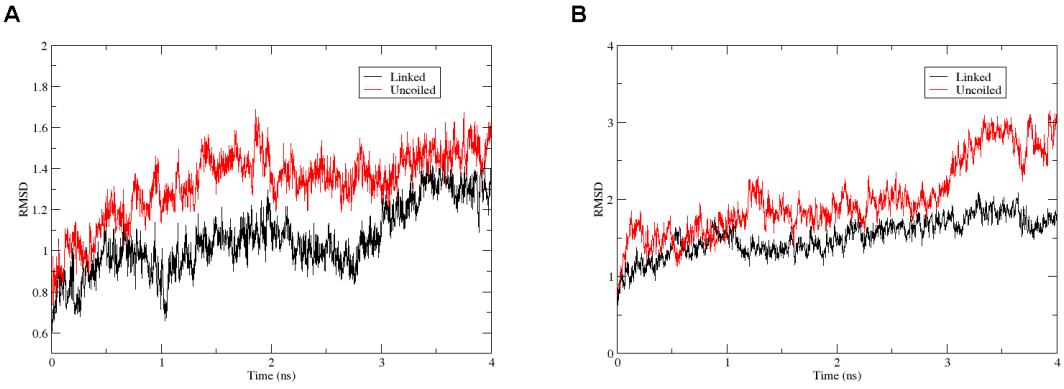


Figure 20: RMSD of the uncoiling simulations for A) Site G9 b) Site G20, the red line showing the RMSD of the system with the cross-link removed relative to the same initial frame and the black line showing the RMSD of the system with the cross-link still present.

What our results show in Figure 20 is that in site G9 we see a large increase in the RMSD upon removing the cross-link, suggesting the backbone is

undergoing a structural relaxation, before beginning to equilibrate. This can be seen by looking at the two systems at the 1ns time point, then as time increases we observe the difference in the RMSD between the two systems decreases. For the favourable G20 site we see little difference between the RMSD of the uncross-linked and cross-linked systems for the first 3 ns of the trajectory. We then see a difference in the RMSD for the two systems begin to develop. However the lines follow roughly the same pattern but with a varying magnitude. These results suggest that at an unfavourable cross-link site, like G9, we have an initial rapid relaxation from a strained system upon cross-link removal. This suggests a larger amount of torsional tension than in the favourable sites, like G20, where there is no initial rapid relaxation of the system. Upon looking at the corresponding dynamic dihedral plots (Appendix 2) for the residues around the cross-link site, for the uncross-linked and cross-linked systems on the same plots, we see no significant difference in the dihedrals for the two systems. At G9 we see a slight shift in the modal value for the ϕ for ^{623}Gln (α1a), three residues down from the cross-linking ^{620}Arg (α1a) residue. At G20 the only significant shift in the modal dihedral values is for the cross-linking ^{1055}Arg (α1a) residue and ^{980}Lys (α2) residue. This can easily be explained by the fact that the side-chains rotate around to form the cross-link, which upon removal of the cross-link, will adopt a state much similar to that in the native system.

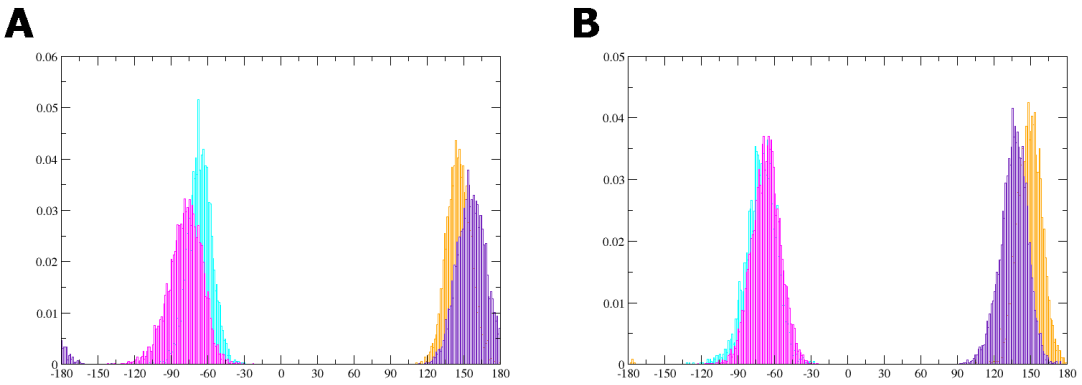


Figure 21: Frequency histograms for the two dihedrals angles ϕ and ψ in the cross-linking residues A) Arginine and B) Lysine of site D20, for the native and the cross-linked systems. (Colours: ϕ_{Nat} – Pink; ϕ_{Cross} – Blue; ψ_{Nat} – Purple; ψ_{cross} – Yellow)

We have already looked at the effects of cross-links on the properties of the collagen's backbone. Here we use the dynamic dihedral data for a number of different cross-link sites, G9, D9, G20, D20, G22 & D22, to compare the variance and shift in the modal dihedral values for cross-linked molecules, compared to the same positions in the native model, (Appendix 3). Unfortunately, from this data, no conclusive trends could be seen from the effect on the sites neighbouring the cross-link that would allow us to explain why some sites were exothermic and others endothermic for AGE formation. What this technique did allow us to see was the effect cross-linking had on the residues directly involved. We see two main possible changes; the first, as illustrated in Figure 21, is the shift in the values for the angle. This is a result of the need for the side-chains to rotate into a position at which they are able to form the cross-link. The other change we see is a reduction in the distribution and a rise in the value for the highest frequency, seen by a narrowing and rise of the cross-link peak compared to the native model seen in Figure 21. This is a consequence of the cross-link creating a more constrained local environment where there is restricted movement of the cross-linked chain. This reduction in the number of

possible angles able to be adopted is based on those energetically favoured, for example rotation outside of these regions could mean that the side-chain atoms needed to form the cross-link are further than is possible for them to covalently bond. Alternatively further rotation may bring the side-chains into close contact with neighbouring residues, resulting in the creation of energetically unfavourable interactions.

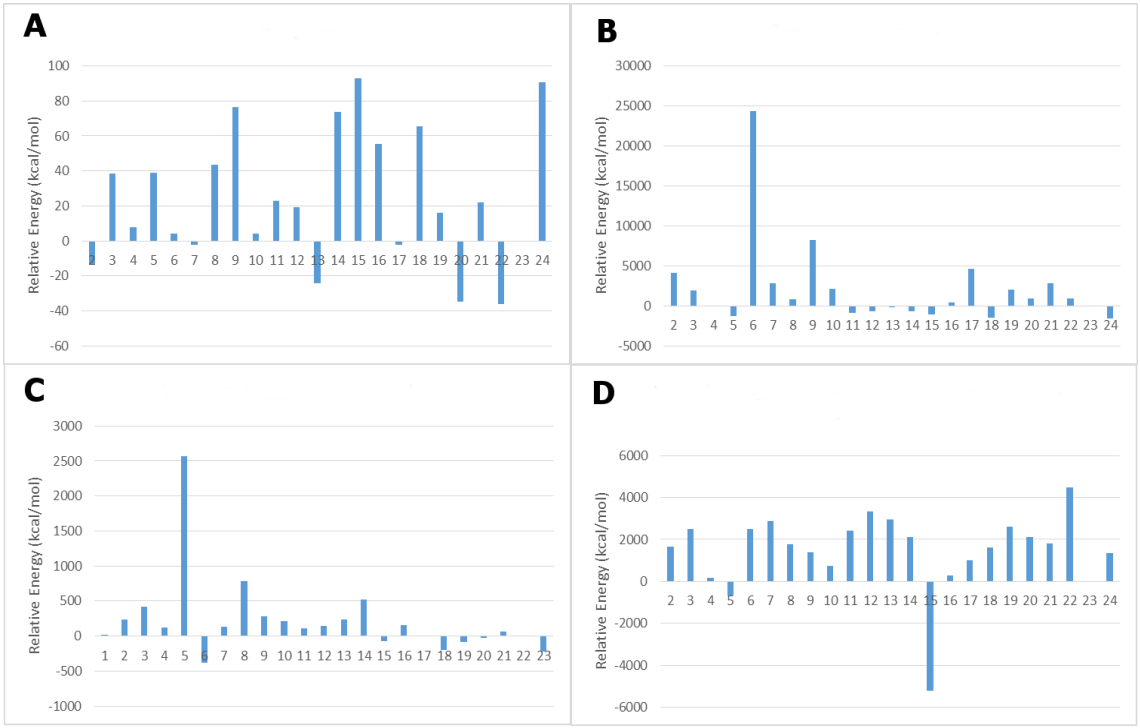


Figure 22: Graphs of the relative energies of the A) Glucosepane cross-linked collagen molecules at site numbers given by the x-axis: then the same systems with the B) Cross-link removed C) Cross-link and solvent removed D) Solvent, cross-link and periodic boundary conditions removed using tloop.

Another approach we adopted was to write an in-house script ("*tloop*"), which could be used to probe a number of properties over the whole duration of the trajectory. The in-house script worked by changing a variable, such as removing water or the periodic boundary conditions (PBC), then running a single point energy (SPE) on every single image of the trajectory without the variable present. We then calculated the mean total energy of the system with the varied

property from all the single snapshots. This approach does still have two limitations; the first is that the number of snapshots analysed is dependent on the printout frequency in the initial simulation trajectory, for this work this was every 1000 steps. Secondly, if studying the interactions introduced by the presence of the cross-link, hence removal of the cross-link and re-introduction of the hydrogens to the lysine and arginine side-chains, there are sometimes close contacts between these protons. Thus energies which are 200% higher than the median energy values are omitted when calculating the mean energies. Despite these limitations we were able to obtain a better understanding of a number of sites.

We conducted further work on the glucosepane sites to ascertain the impact of three different aspects of the simulation; the PBC, the solvent and the physical presence of the cross-link, Figure 22. Upon removing the cross-link Figure 22B), there are no quantitative conclusive trends in the energies, however qualitatively we can see a decrease in energies for the unfavourable formation sites upon removal of the cross-link, which suggests that there may be close contacts occurring between the cross-link and their neighbouring environment. The energy scale is distorted, particularly by the large peak at site G6 as a result of a large number of proton-proton clashes, which did not pass the threshold but result in this large increase in energy. Removal of solvent and the cross-link results in a large destabilisation of all the systems, as this removes the water mediated hydrogen bonding from the system. The biggest conclusions we took away from the use of the tloop script (Figure 22D) is that, upon removing the PBC, the trend in the relative energies (relative to the native system) of the glucosepane cross-linked system changed completely. This suggests that, in regions of higher protein density, the higher energy is likely the

result of close contacts with neighbouring collagen molecules. Site G5 and G15 show the most significant changes in their relative energies when the PBC are removed, illustrated by their negative values compared to the native system under the same conditions. Upon further investigation we found that for G15 the arginine residue has to rotate into a position where it is capable of forming the cross-link, and in doing so the guanidinium group (and hence the five membered ring portion of glucosepane) comes into very close contact with a neighbouring collagen molecule, thus raising the energy significantly. Upon removing just the cross-link we see a small drop in energy as the carbons of the sugar moiety bound to the collagen are removed reducing the number of close contacts, a larger drop in energy is then seen on removing the PBC to completely remove the close contacts.

A comparison between the number of favourable DOGDIC and glucosepane formation sites (105), reveals an equal propensity for each AGE cross-link formation although, of the six DOGDIC sites, only site 20 was identified in both studies. The most likely reason for this variation is the difference in the separation between the three terminal nitrogen atoms lysine N^{ζ} and arginine N^{η}. In glucosepane these separations are 2.6 Å and 3.8 Å respectively. In DOGDIC the separation is smaller at 2.5 Å and 3.5 Å, which is potentially the determining factor as to whether cross-link formation will be favourable or not. Upon rearrangement of the side-chains the resulting configuration may impose close contacts of those side-chains with their neighbouring residues. The difference between the two sets of distances suggests that the structure of the residues at the same site for DOGDIC and glucosepane is not the same. In one structure there may be an unfavourable close-contact introduced or a favourable electrostatic interaction may be missing, depending on the cross-link formed.

Another possible explanation is the difference in the degree of polarity between the two cross-links, with glucosepane having two hydroxyl groups whilst DOGDIC has three. However, there is no net change in the polarity, leading us to believe that this is unlikely to be a significant contributor. The one common site, site 20, is seen in Figure 16 and Figure 17, to have fewer bulky amino acid residues surrounding the cross-link. This means that it is likely to have a greater flexibility in its movement and thus the ability to form the cross-link without the introduction of close-contacts or unfavourable interactions with the neighbouring residues. It is therefore found to be energetically favourable for both of the studied AGEs.

Despite there being an equal number of favourable glucosepane and DOGDIC sites within the collagen molecule, other factors may affect the formation within the tissue, which will account for the difference in the reported *ex vivo* concentrations. The simulations conducted in this study do not take into account activation barriers for cross-link formation, nor do they take into account the kinetics of the reaction. It has previously been reported that the dehydration step in the glucosepane formation is non-reversible, whereas DOGDIC formation is reportedly reversible, potentially accounting for the difference in relative abundance *ex vivo* (101).

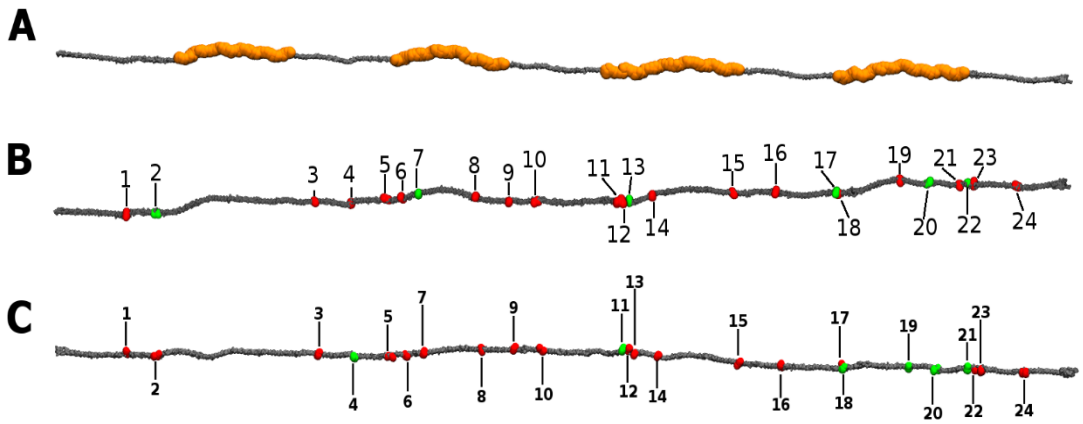


Figure 23: Image depicting a single collagen molecule with A) the gap regions illustrated by the orange regions B) Favourable glucosepane sites highlighted by the green regions and the unfavourable regions highlighted by the red regions C) Favourable DOGDIC sites highlighted by the green regions and the unfavourable regions highlighted by the red regions. Labels denote the number of the site highlighted

The location of the cross-links within the collagen molecule, as shown in Figure 23, may play a role in whether the cross-link will be energetically favourable to form or not. It can be seen that the sites identified by the distance based criteria search are spread along the whole length of the collagen molecule. However there is a large region of the molecule between site 2 and 3 where no potential sites exist. One striking observation is that all but one of the cross-links for DOGDIC are located within the gap region of the collagen fibril, where the gap region is defined as the lower protein density region produced as a result of the D-periodicity of the Hodge Petruska packing model. Glucosepane shows a similar affinity for cross-link formation in the gap regions of the collagen (site G2 and D4 are both in very close proximity to the gap region). There are two potential explanations for this observation. First, both AGEs are polar and hence capable of forming hydrogen-bonds to the intra-fibrillar water molecules; the number of water molecules per unit volume in the fibril is 20% higher in the gap region than in the overlap region (78, 263). Second, the overlap region has

a higher protein density (volume), which results in an increased likelihood of an unfavourable interaction occurring between the newly formed AGEs and the neighbouring collagen molecules. The influence of both of these factors can be seen in the conformations of the DOGDIC cross-link shown in Figure 17. The one cross-link not located in the gap region shows a configuration with the hydroxyl chain of the DOGDIC cross-link running flat (or parallel) to the backbone. Those cross-links in the gap region have the DOGDIC hydroxyl-chains perpendicular to the backbone, thus maximizing the amount of hydrogen-bonds to the intra-fibrillar water molecules.

4.6 Oxidized-DOGDIC

The have been some reports of the literature in that DOGDIC can undergo further oxidation to form Ox-DOGDIC, changing the aliphatic chain into a six membered ring, although the exact mechanism of this is currently not known. To date few studies have been conducted into Ox-DOGDIC on its structure, function or impact on the body. This is most likely due to its low concentration within the body. Concentrations of Ox-DOGDIC are only 0.23% that of glucosepane in skin, with glucosepane having levels as high as 2000 pmol/mg compared to 7 pmol/mg for Ox-DOGDIC (100). From the few reports that have been constructed it has also been shown that the concentration of Ox-DOGDIC increases with age whilst DOGDIC showed a decrease in concentration with age, suggesting an enhanced oxidative process occurring in the skin, with direct conversion between the two AGEs (100).

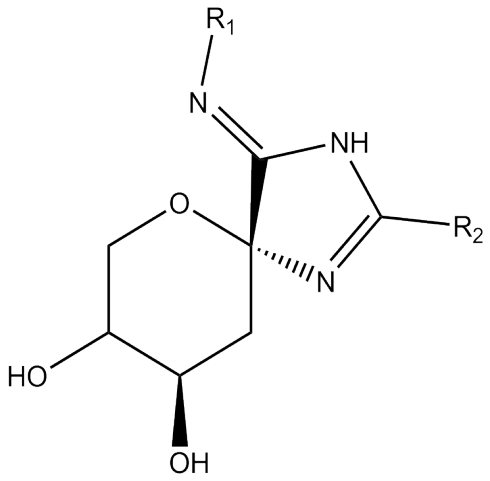


Figure 24: Schematic diagram of Ox-DOGDIC, where R_1= Lysine and R_2 = Arginine.

As DOGDIC is a precursor to Ox-DOGDIC, we used only the six identified in the previous investigation to be exothermic for DOGDIC formation, instead of the 24 sites located in the distance based-criterion search to reduce the computational expense of the study. The DOGDIC cross-links were converted to Ox-DOGDIC at the six favourable sites, before running for a further 60 ns of simulation. The average relative formation energies for the Ox-DOGDIC cross-links were then calculated relative to their native collagen system, with the data being collected again from the stable final 25 ns of simulation, with the results shown in Table 6.

Table 6: The Difference in enthalpy formation of Ox-DOGDIC for all the 6 favourable DOGDIC cross-link sites, identified in the previous study.

Cross-linking Site	**Relative Formation Enthalpy (kcal/mol)**
4	+60.30
11	+3.96
18	+21.96
19	-
20	+12.15
21	-4.28

As can be seen in Table 6 the formation of Ox-DOGDIC is only energetically favourable at one of the six identified cross-link sites. Unfortunately site 19 clashed (C5 specifically) with a neighbouring collagen molecule backbone upon forming the cross-link, which meant the energy was significantly higher than the others and certainly would not occur within the body in this arrangement. For all of the sites investigated we see an increase in their energy as a result of this further oxidation. This is most likely due to the fact that the structure is placed into a more constrained configuration upon formation of the six membered ring, Figure 24, and due to the fact that the separation between the three terminal nitrogen atoms is reduced even further to 2.35 Å and 3.53 Å in Ox-DOGDIC, from 2.5 Å and 3.5 Å in DOGDIC. Both of these changes mean that there will be increased steric constraints imposed and hence more likely to be close contacts between the rest of the side-chains in the cross-linking residues, or between the cross-linking residues and neighbouring residues. This increased constrained configuration is illustrated by the fact that the standard deviation (spread) in the values for the dihedral angles of the cross-linking reduces is decreased by 10-20%, as shown in Figure 25. Figure 25 also shows a significant rotation around the alpha-carbon in the oxidation process, with the mean value of the phi angle in arginine changing from 156.61° to 8.88°, thus further illustrating the big differences in the structures of the two cross-links.

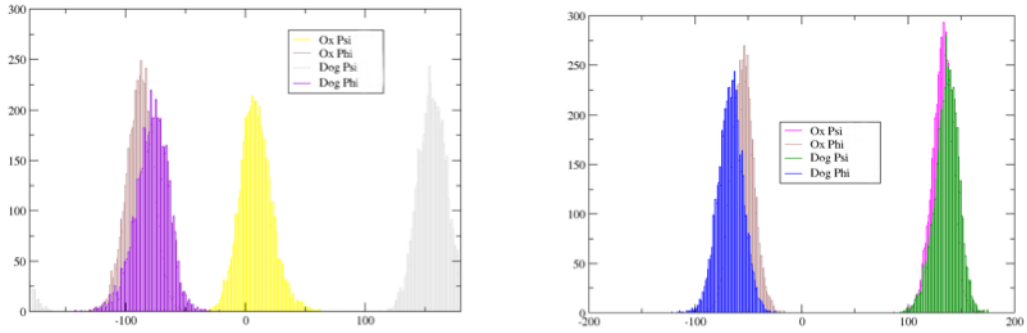


Figure 25: Dihederal angles for the cross-linked arginine (left) and lysine (right) residues for Ox-DOGDIC and DOGDIC at Site 20

Site 21 gives a relative formation enthalpy (compared to the native model) that is slightly exothermic, - 4.28 kcal/mol, suggesting it may possibly form within the body. However the magnitude of the enthalpy, and the fact that there is only one potential intra-molecular formation site where Ox-DOGDIC is energetically favourable, may explain why the concentration of Ox-DOGDIC in the body is so low.

4.7 Summary

In conclusion, we have identified six sites each for DOGDIC and glucosepane where the intra-molecular formation of the cross-links in type I collagen is energetically favourable, with only one site being equivalent. We have then shown that the reduced N-N intra-distance in DOGDIC means that there is little competition for lysine arginine sites with glucosepane, as they form exothermically at different sites. Our results suggest that lower levels of DOGDIC in human lens tissue is most likely as a result of the differences in the availability of carbonyl metabolite, or the non-reversibility of the glucosepane formation mechanism. The positions of these cross-links are likely to have a significant impact on collagen properties, with some overlapping with key

collagen-biomolecule binding sites. Both AGEs studied show a preference to form energetically in the gap region, owing to its lower protein density and higher intra-fibrillar water content. This is owing to the fact that the local environment around a formation site is a key factor in the energetics of the process, with little deviation observed in backbone atom positions upon cross-link formation. Additionally a single site for Ox-DOGDIC formation from DOGDIC under oxidative conditions was identified, although its significance within the body is likely to be small.

Chapter 5 Mechanical Properties of Collagen and the Impact of Cross-linking

5.1 Introduction

As was shown in 1.1.3 the mechanical properties of collagen are vital to its function as a material. Therefore a large number of experimental studies have focussed on investigating these properties (63, 65, 71, 264–269). However due to the size and complexity of collagen molecules, to date, very few computational investigations have been conducted. The first computational study to probe the mechanical properties of the collagen molecule was conducted by Lorenzo *et al.*, in 2005, when they conducted a steered molecular dynamics approach, testing the molecular response of short 29-30 amino acid collagen like peptide, based on a two springs in series model (270). However this study only used a polarisable continuum model, which was later proven not to be sufficient by a 2007 study by Zhang *et al.*, in which they looked at the extent to which structural water participates in carrying load (271). During this study they observed that water behaved as a weak lubricant during in plane axial stretching, but as a resisting factor during microfibril bending, suggesting that the absence of water would lead to inconsistent results with experiment. A couple of other studies have been conducted since looking at several elements such as how helical hierarchy controls collagen deformation (264), the role of the mature enzymatic cross-links (272) and even the viscoelastic (creep) behaviour of collagen micro fibrils (273).

A number of previous studies have tried to combine the results of AFM single molecule pulling experiments with those obtained through fully atomistic steered

molecular dynamics. However limits in capabilities of both techniques, primarily the different time scales of the two techniques, prevent direct comparison between the two being possible (274, 275). Despite this a number of papers have been published using a combination of the two (276, 277) as, although direct comparison of absolute values of the two techniques is not possible, it is possible to generate qualitative results. These qualitative results can be used to correlate with trends in the other techniques to explain some of the trends observed.

The major problem is that no consensus has come out of any of the studies to date for the absolute value of the Young's modulus of collagen. Table 7 gives results from a number of both experimental and computational investigations using a variety of techniques and their respective values. There are a number of possible reasons for these inconsistencies from; the accumulation of uncertainty in the experiments; or probing the mechanics at different states of hydration (70).

Authors	Methodology	Value for YM (GPa)	Reference
Harley *et al.*,	Brillouin light scattering	9.0	(278)
Cusack *et al.*,	Brillouin light scattering	5.1	(279)
Sasaki *et al.*,	X-ray Diffraction	2.9 ± 0.1	(280)
Sun *et al.,*	Estimate from Persistence Length	0.35 – 12	(66)
Lorenzo *et al.,*	Computational – SMD	4.8 ± 1.0	(270)
Zhang *et al.*,	Computational – SMD	6.5 ± 0.5	(271)
Gautieri *et al.,*	Computational – Coarse Graining	4	(77)

Table 7: Values for the Young's modulus of molecular type I collagen derived from a variety of techniques, illustrating the inconsistencies of the values produced from different methods.

If we focus solely on the computational techniques, we have identified three key reasons that may contribute to the inconsistent values for the Young's modulus. Firstly there are a wide variety of values reported for the constant of the cross-sectional area from 167 – 214 $Å^2$. This of course leads to a wide variety of values being reported for the Young's Modulus (270, 271, 281). Secondly there is no consistent value used for the pulling velocity, which may result in greater uncertainty in reported values, which may thus introduce variance in the reported absolute values. Finally the difference in the primary sequences of the collagen like peptides used in the studies, which it has been suggested, may alter the mechanical response of the collagen to an applied load. It is this last variable which we will investigate further in our study whilst trying to keep the influence of the other two factors to a minimum through careful experiment design.

5.2 Steered Molecular Dynamics Methodology

All of the steered molecular dynamics (SMD) calculations were conducted using NAMD version 2.11, owing to its excellent scalability and compatibility with the Amber force-field, such that the same force-field is able to be used, removing the need to re-parameterise the cross-links for a new force-field implementation.

5.2.1 Tensile Modulus

The terminal nitrogen atoms of each of the three strands are fixed at a position in space. The three carbon atoms at the other end of the three strands are defined as the SMD pulling group. The centre of mass of the three pulling atoms is attached to a dummy atom via a virtual spring of force constant 7 kcal/(mol

$Å^2$). The dummy atom is then moved at a constant velocity of 0.1 Å /ps, along a vector defined between the centre of mass of the pulling and the fixed atoms, the resultant reactant force is printed out every 50 steps. The simulations are conducted for 80 ps at 310 K using the same ff99SB force field from our previous investigations. The resultant force displacement curve can then be used to obtain information on the stress strain relationship and even calculation of the Young's modulus, as detailed in 5.2.3.

5.2.2 Lateral Modulus

The terminal carbon atoms, in addition to the terminal nitrogen atoms, of each of the three strands are fixed at a position in space. The SMD pulling group for the lateral investigation is defined as the alpha carbon atom of the central α1 residue of a cross-linked chain (in the case of the cross-link forming between the two α1 chains the α1a chain is used). The centre of mass of the pulling atoms is attached to a dummy atom via a virtual spring of force constant 7 kcal/(mol $Å^2$). The dummy atom is moved at a constant velocity of 0.1 Å /ps, with the resultant reactant force again being printed out every 50 steps. The pulling vector in this case runs perpendicular to the principal axis and is defined by the centre of mass of the pulling atom and the alpha carbon atom of the residue below in the structure. The simulations are conducted for 50 ps at 310 K using the same ff99SB force field from our previous investigations.

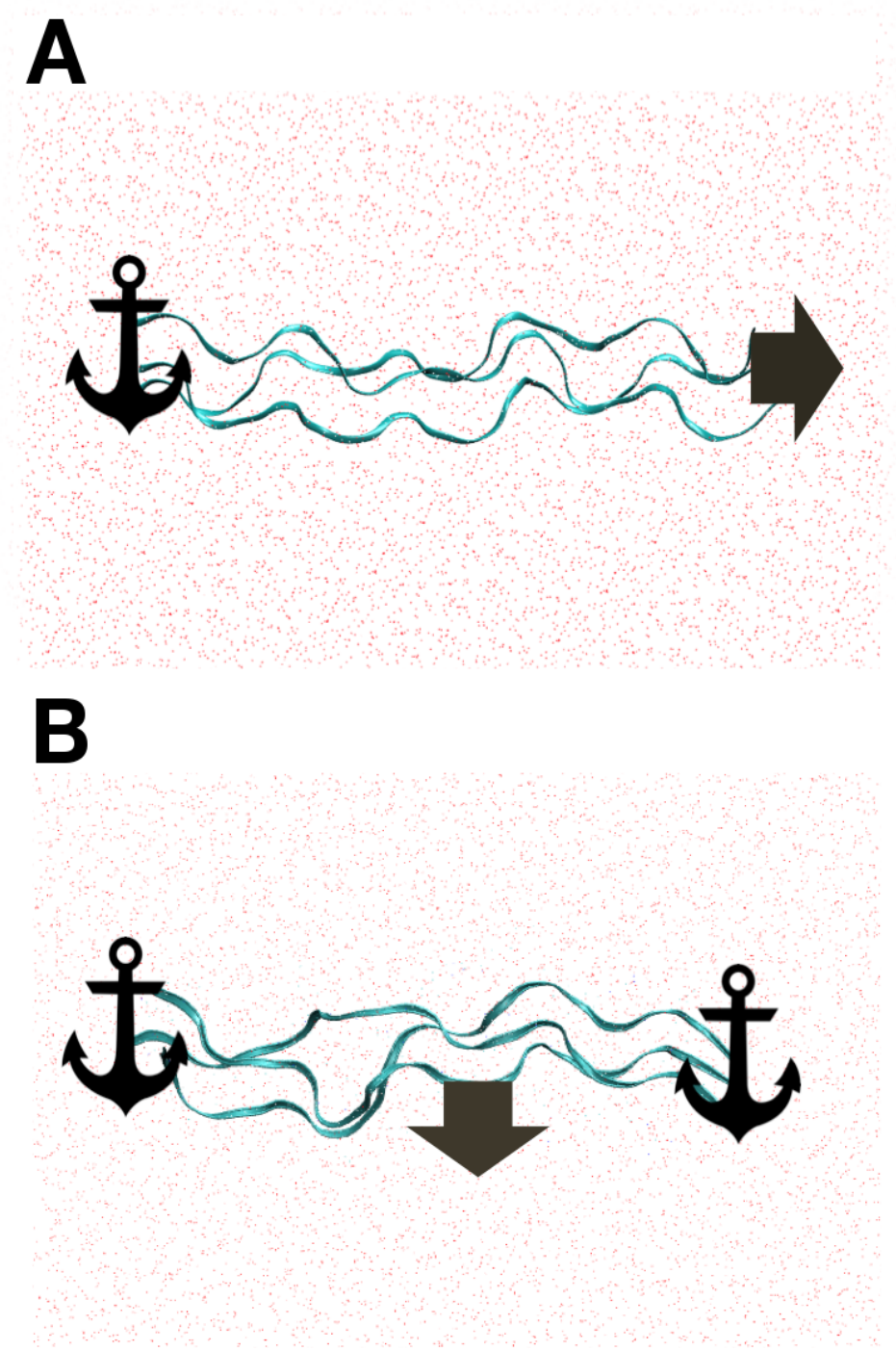


Figure 26: Graphical representations of the two SMD protocols implemented for probing the mechanical properties of collagen, A) Tensile modulus and (B) Lateral Modulus

5.2.3 Analysis of the Nano-mechanical Properties

The force displacement data obtained from the SMD calculations can be converted into stress strain curves using the relationships given below.

The stress applied to a material is the force per unit area applied to the material. Hence the stress inside the protein σ, can be calculated by the force exerted on the molecule F, divided by the cross-sectional area A.

$$\sigma = \frac{F}{A} \quad \text{(Eq. 1)}$$

The strain ε of the protein is the unit less ratio of extension (L-L_0) to original length L_0.

$$\varepsilon = \frac{L - L_0}{L_0} \quad \text{(Eq. 2)}$$

The Young's modulus E, which gives a measure of the stiffness of the material, can then be calculated, under the elastic approximation, as the stress divided by the strain (gradient of a stress strain plot).

$$E(\varepsilon) = \frac{\sigma}{\varepsilon} \quad \text{(Eq. 3)}$$

5.3 Heterogeneous Response to Strain – Triplet Variance

All but one of the molecular dynamics or AFM studies conducted to date have excluded looking at the effects of the amino acid sequence. The studies conducted on different types of collagen, or collagen like peptides, all exhibited a wide variety of moduli, suggesting that the variance in the sequence used could result in alteration to the mechanical behaviour. This difference in behaviour could be due to the differences in the way the side-chains move with

respect to one another; the cause could be influenced by a number of biochemical parameters such as polarity, charge and side-chain volumes.

Buehler *et al.,* previously conducted the only investigation into the effect of the sequence of a collagen molecule on its response to an applied load (281). However, Buehler *et al,* only investigated four different sequences, which they considered to be the most abundant sequences in collagen, $(GPO)_3$ (10.5%), $(GAO)_3$ (3.4%), $(GPA)_3$ (3.4%) and $(GEQ)_3$(GEK) (2.0%) listed in decreasing abundance, with abundance in type I collagen in brackets. Although abundant, the four triplet sequences still only accounts for fewer than 20% of the whole collagen sequence (50), and thus a study using a larger number of the abundant sequences is required to gain a better understanding of the mechanical behaviour of the collagen molecule. What Buehler *et al.'s* study showed was that the nano-mechanical response varied for different sequences up until 25% strain, with the Young's modulus varying by up to 50% between the four sequences. After 25% strain the backbone was stretched which they deemed to be independent of sequence, as moduli of the four sequences converged after this point. Additionally they showed the importance of hydrogen bonding in the response to an applied load, with the stiffness being related to the rate of hydrogen bond breaking.

In our investigation we aim to quantify the effect of varying just a single residue in the central triplet of a homotrimeric collagen like peptide on its response to an applied load, whilst simultaneously investigating a wider variety of the triplets encountered within the full collagen molecule. To test this we ran an investigation on the triplets that make up just 28.1% of the sequence triplets of

the collagen molecule, which are the GlyProYyy triplets, by testing over 20 different triplet compositions, based on varying the identity of the Yyy residue.

To conduct this study we use a small collagen like peptide, which is shown to be capable of forming quasi-staggered packing in the crystal structure. First synthesised by Kramer *et al.,* (28) in 2000, the peptide was initially used to investigate the effect of charged pairs in the centre of the triplet, we will use this central triplet to study the sequence dependent effect on the mechanical properties of collagen like polypeptides. The CLP we use in this study is homotrimeric with the sequence $(ProHypGly)_4YyyProGly(ProHypGly)_4$, the three polypeptide chains are staggered with a leading, middle and tailing chain. The presence of the repeating GlyProHyp triplets ensures that a triple helical structure is adopted. This was theorised by Brodsky *et al.,* (32) who showed that the GlyProHyp triplet is the most stable for the triple helical conformation with substitutions of the Pro or Hyp, in the Xxx and Yyy, decreasing the melting temperature by destabilising the triple helix. Arginine in the Yyy is the only substitution that does not lower the melting temperature (282). However where Hyp in the Yyy promoted triple helix formation, arginine destabilised triple helix formation (51).

5.3.1 Methodology – Building the Model

The CLP we use in this study is homotrimeric, with the sequence $(ProHypGly)_4YyyProGly(ProHypGly)_4$, taken from the protein databank accession number: 1QSU (28). An in-house script was created to vary the identity of Yyy for each of the 20 naturally occurring amino acids plus hydroxyproline independently in the central GlyProYyy triplet.

Once the sequences have been defined they are transposed onto a PDB, which contains the coordinates of the backbone atoms from the crystallographic study. The PDB is then loaded into LeaP which adds in the missing side-chain atom coordinates, Cl^- or Na^+ ions to negate any overall charge from the charged amino acid residues and we finally solvate the CLP with a buffer of 30 Å of TIP3P water molecules in all directions. The models are then run for 5000 steps of minimisations (500 steepest descent followed by 4500 conjugate gradient), 120 ps of heating before finally being run for 2 ns of NPT equilibration. 2 ns was chosen as we believe this is more than sufficient for a stable conformation to be formed. This was confirmed by checking the RMSD of ten random models all of which had an RMSD less than 2.5 Å. The pulling experiments were conducted using the above tensile SMD methodology outlined in 5.2.1, for five repeats taken at 0.1 ns increments for the final 0.5 ns of the simulation. The tensile Young's modulus was calculated by Eq. 3 to determine whether the sequence has an effect on the mechanical properties of triple helical collagen-like peptides.

5.3.2 Results

Models were constructed and MD simulations were run for the 20 different sequences generated by varying the Yyy residue in the central triplet. Unfortunately the methionine based sequence continuously uncoiled upon running the simulations. The restraints necessary to force a triple helical confirmation were too high to be of relevance; therefore it was decided to omit methionine substitution from our study. The remaining 19 models then underwent the tensile modulus protocol outlined in 5.2.1, with the modulus

being calculated in the low strain domain. The choice of the low strain region, of up to 25% strain, was taken based on the observations of Buehler *et al.,* who observed that changes to the structure in the higher strain regions were independent of sequence (281). The averaged results of the SMD simulations were then normalised against the most frequently occurring GlyProHyp triplet, so that they any resulting sequence dependence differences are easier to observe, the results of which are presented in Figure 27. The standard errors of the mean for these average values were no greater than 0.4%.

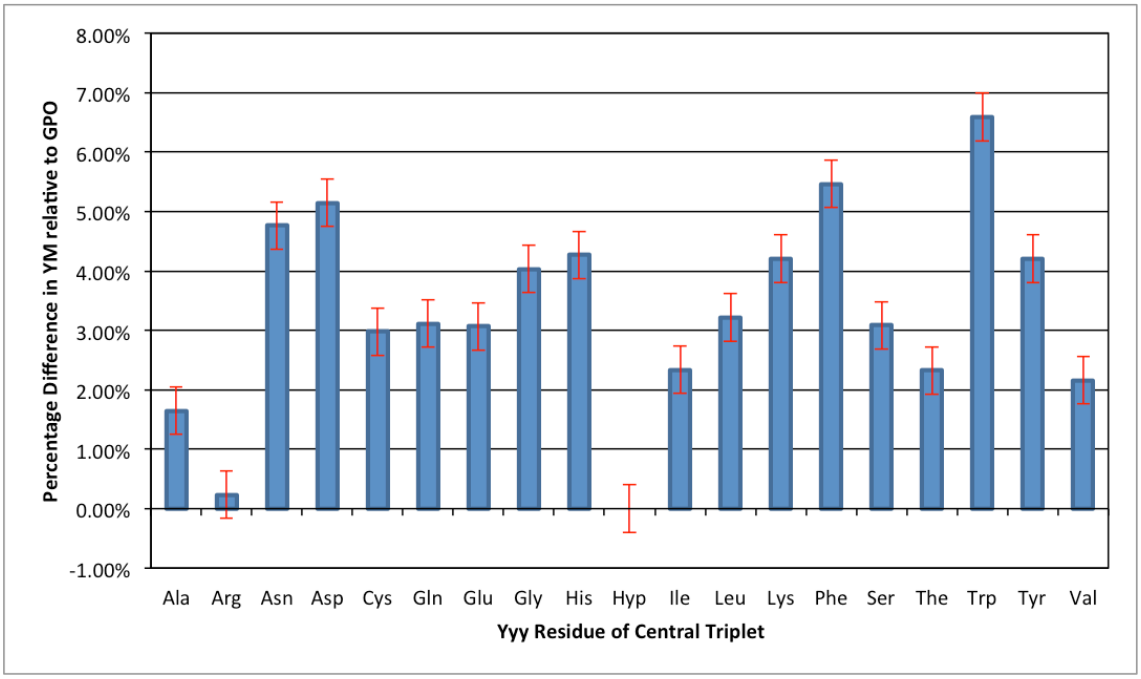


Figure 27: Plot showing the change in the value of the Young's modulus on varying the Yyy residue in the sequence $(ProHypGly)_4YyyProGly(ProHypGly)_4$, relative to the value of the most frequently occurring GlyProHyp triplet. Uncertainty in above values no greater than ±0.4%.

What is immediately apparent from Figure 27, is not only that the hydroxyproline residue in the Yyy position is the most frequent, but it also yields the most compliant molecule of those tested, with all the other residues resulting in an increase in value for the Young's modulus. The absolute value for the Young's modulus, calculated for a hydroxyproline residue in the Yyy position, was 7.0

GPa which is in the upper range of values reported in Table 7, and close to the value reported by Zhang *et al.* (271), supporting the reliability of the results presented herein.

As has been mentioned previously in 1.1.2, hydroxyproline plays a significant role in the stabilisation of the triple helix. Hydroxyproline does this in one of two ways; through water mediated hydrogen bridging interactions between collagen molecules or strands; or through the entropically favoured imino acid specific constraint of the backbone angles to optimum collagen values. If the imino specific constraint of the backbone angles were the reason for the higher elasticity then we would expect a GlyProPro triplet to exhibit a similar phenomenon. However proline in the Yyy position results in a value for the Young's modulus which is 0.84% greater than that of hydroxyproline, yet significantly lower than for most other residues. So instead, if the hydrogen bonding properties played the dominant role, then we would expect similar values for other polar residues to be reported. Apart from arginine, which is a potential hydrogen-bonding source, this was not observed. Therefore it is likely that a combination of both effects, in addition to the inductive effects introduced by the hydroxyl group, are the reason for this elastic behaviour.

Interestingly arginine also exhibits a value for the Young's modulus that is much lower than for any other residue (excluding hydroxyproline). It has been shown previously in the work of Yang *et al.*, that arginine in the Yyy position can exhibit a similar stability as hydroxyproline in the same position (51). Its stability was not thought to be a result of its charged group, something which we concur with given that the Young's modulus for the lysine residue in the Yyy position is significantly greater than that of the arginine derived model. Yang *et al.* instead

hypothesised that it was the role of the guanidine group of arginine and its direct hydrogen bonding interactions with the protein backbone or water network that stabilised the structure, and its constrained position in the Yyy position that provided the optimum interaction network (283, 284).

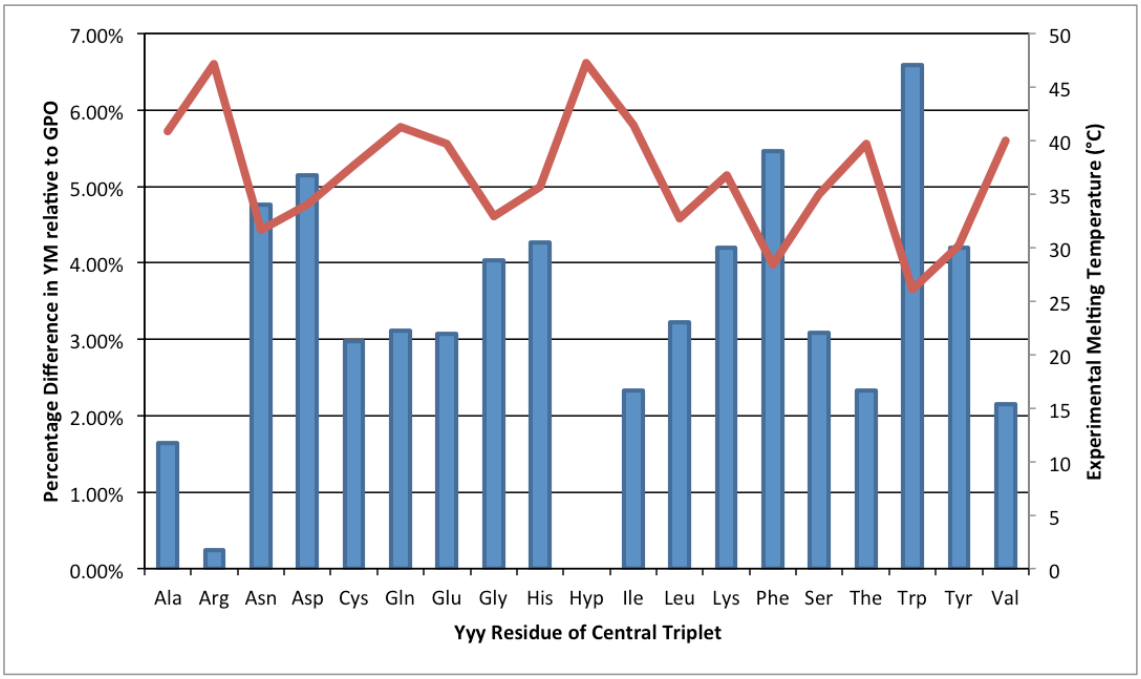


Figure 28: Bar plot showing the relative difference in Young's modulus on varying the Yyy residue from hydroxyproline. With the red line plot showing the experimental melting temperature for each triplet, reported in the work of Brodsky *et al.*, (30). Uncertainty in above calculated YM values no greater than ±0.4%.

Given that both hydroxyproline and arginine exhibit similar stability in the Yyy positions of model peptides, we plotted in Figure 28 the relative Young's modulus values, with an overlaid line plot of the experimentally derived melting temperature for the corresponding model sequences (30). What this shows is that if the model peptide is stabilised by inclusion of a particular residue, illustrated by a higher melting temperature, then the Young's modulus will also decrease. Through looking at the classifications of amino acids, we can also see that the aromatic amino acids have significantly higher values for the

Young's modulus then non-aromatic. Additionally, the charged residues, with the exception of arginine, will result in a 3% or greater increase in the Young's modulus by their inclusion at the Yyy position.

5.3.3 Discussion

The sequence dependency of the mechanical properties will not only affect comparison of mechanical responses of different collagen or collagen like molecule, but will impact the deformation patterns within a single collagen molecule. The sequence of collagen, although having a common repeating pattern of GlyXxxYyy triplets, is still extremely heterogeneous and thus, as has been shown in this work, will respond in a non-uniform manner to an applied load. This may lead to localised strain and stress concentrations within the collagen molecule, owing to the different response of the varying sequences to a mechanical load. Sections of sequence which have been shown to be highly compliant may deform at a greater rate, potentially inducing micro unfolding at that region or inter-molecular sliding of the supramolecular structure.

The impact of varying just one residue within the collagen molecule on the calculated value for the elastic modulus is significant, in some cases the Young's modulus increasing by as much as 6.6%. For this reason it is no surprise that the numbers reported from different experimental and computational investigations differ so much, where sequences used can vary significantly. However we do not believe that it will be a simple additive effect, as the interaction between the different possible Xxx and Yyy residues, and hence their stability and mechanical properties, will vary depending on their identity. Instead a more comprehensive study would require testing all possible

permutations of Xxx and Yyy to get a complete picture on the sequence dependency of the mechanical properties. What our study does highlight however is that, when comparing two sequences, the stability in the form of melting temperature can be a good indicator as to how they will respond to a mechanical load. Additionally this can be used to build up an estimate of the mechanical response of GlyProYyy triplet only containing triple helical peptides. In future work we aim to calculate the mechanical properties for the GlyXxxHyp triplets so that over 50% of the sequences of collagen have been calculated, thus allowing a better understanding of the elasticity and mechanical properties of collagen molecules. Ultimately we hope to obtain an understanding of the mechanical and energy storage properties of connective tissues such as tendons and ligaments. We hope the results presented here may form the basis of a data library, which could in future be added to such that the mechanical properties of a collagen molecule could be predicted based purely on sequence information.

5.4 Impact of Intra-molecular AGEs Cross-linking on Mechanical Properties of a Collagen Molecule

In this portion of our work we aimed to test whether the presence of AGEs within a collagen molecule will alter the mechanical properties of the collagen molecule. To test this theory we use constant velocity steered molecular dynamics simulations on short collagen sections extracted from the full collagen molecule.

5.4.1 Methodology

5.4.1.1 – Building the Model

Sections of the collagen were created by taking 5 or 7 residues either side of the cross-linking residues depending on the curvature of the molecule at that point. Care had to be taken to ensure that the ends of the collagen sections created were as straight as possible, so that the pulling and fixed groups were close to linear and that the principal axis of the collagen, for tensile modulus, was being probed. The sections were created for several different time points, 37 ns, 41 ns, 45 ns, 49 ns, 53 ns and 57 ns, from our previous studies in which we identified the cross-linking sites, allowing a number of different starting configurations to be employed. As has been shown in 5.3 above, the heterogeneous nature of collagen means the absolute values will alter depending on the type of amino acid residues neighbouring the cross-linking site. It was decided to provide relative (percentage difference) values calculated between the same sites in native collagen and the cross-linked collagen, thus removing the effects of the sequence on the mechanical properties.

Once the smaller collagen sections had been created they were loaded into LeaP, where the Cl^- and Na^+ ions were added to negate the overall charge from the anionic and cationic amino acid residues. The models were then solvated with TIP3P water model with a buffer of 30 Å. Once the solvated models had been created they underwent 2500 steps of minimisations (500 steepest descent followed by 2000 conjugate gradient), 120 ps of heating to 310 K before finally being run for 1 ns of NPT equilibration. During the minimisation and heating stages, in addition to the first 500 ps, the whole protein was kept restrained. In the second 250 ps just the backbone was restrained. In the final

250 ps only the terminal residue's backbone atoms were restrained, using a force constant of 75 kcal/(mol Å^2).

The equilibrated structures then underwent SMD calculations using the outlined SMD protocol (5.2) in both the tensile and lateral directions to probe the effect of the cross-link on the mechanical properties of the short collagen sections.

5.4.1.2 Full Collagen Mechanics

A structure for the collagen molecule is extracted from our previous simulation of the wild type collagen. The water molecules are stripped out of the structure using Ptraj, owing to the huge computational cost that would exist to model a fully solvated collagen molecule (approximately 2 million water molecules with a buffer value of 20 Å). To generate the cross-linked models, the cross-linked regions of the collagen molecules are inserted into the native model; such that two models are generated, one with all six favourable glucosepane cross-link sites and another with six favourable DOGDIC cross-link sites inserted. This is repeated for all 6 time points. Once the two models had been created they underwent 2500 steps of minimisations (500 steepest descent followed by 2000 conjugate gradient) and 120 ps of heating before finally running a production simulation for 1 ns of NPT equilibration. During the minimisation and heating stages the backbone was restrained using a force constant of 75 kcal/(mol Å^2).

5.4.2 Results and Discussion

As mentioned in 5.1, a major source for the variation in the absolute values in the Young's modulus is the different pulling velocities used in the simulations.

We tested three different velocities for three of the cross-link sites and from the results illustrated in Figure 29, we can see that increasing the velocity decreases the calculated values for the Young's modulus. This is owing to the larger perturbation of the structure per unit of time. However as we are using the relative differences between wild type and glycated peptides this will not change the results we are interested in. However, through the use of a consistent pulling velocity for all of the sequences, we are able to limit the uncertainty introduced owing to this.

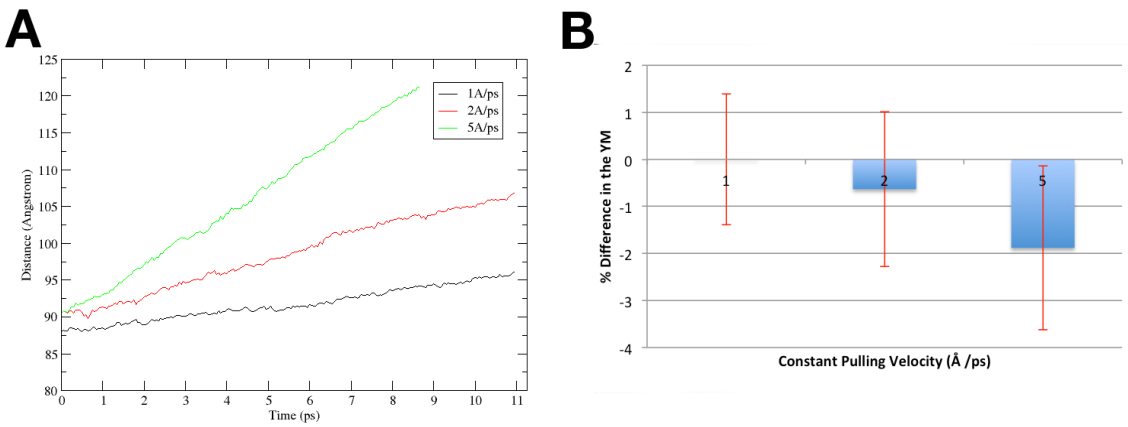


Figure 29: Figure showing the mechanical response of a collagen-like peptide (collagen region 4) to strain applied at varying velocities. A) Illustrates the length vs time plot and B) The effect of the velocity on the Young's modulus calculated from N=6 repeats.

Another source of variation for the absolute values for the Young's modulus is the values used for the area in Eq. 1. It was for this reason, and due to the different sequences being investigated, that it was decided to measure the diameter at approximately 30-40 positions along each of the short collagen snippets, such that the cross-sectional area can be calculated for each of the snippets to give a more accurate value for their Young's modulus. Even with this approach there will be some uncertainty introduced, owing to the assumption that the cross-section of the collagen is circular which, although a

good approximation, is still not entirely accurate. Given the six time points for each of the cross-link positions it was possible to reduce the uncertainty from values reported in previous simulations, ± 7.5 – 20 % (270, 271). However, at times this uncertainty could still be relatively large, up to ± 2.25 % for the tensile modulus and up to ± 4.1 % in the lateral modulus.

Despite this uncertainty we were able to obtain results for the change in the mechanical properties. The resultant force displacement curve and stress-strain from our tensile modulus simulations, Figure 30, resembles those obtained in the work of Zhang *et al.* (271). The general overview of the curve shows an initial “toe-shaped” region that changes to a linear region of greatly increased gradient that would continue up until failure (285). The toe shaped region has been described previously for a tendon under axial stretching (265), where it was considered to be the low strain region necessary to remove small “crimps” in the collagen molecular structure. Hence the “toe-shaped” region in our curve is likely the low strain region where the removal of curvature from the collagen fragments occurs. In the low strain region there is also a maintained number of overall hydrogen bonded interactions between the chains, but as the structure becomes too strained the separation between the polypeptide chains increases as the chains straighten, and the total number of hydrogen bonded interactions decreases.

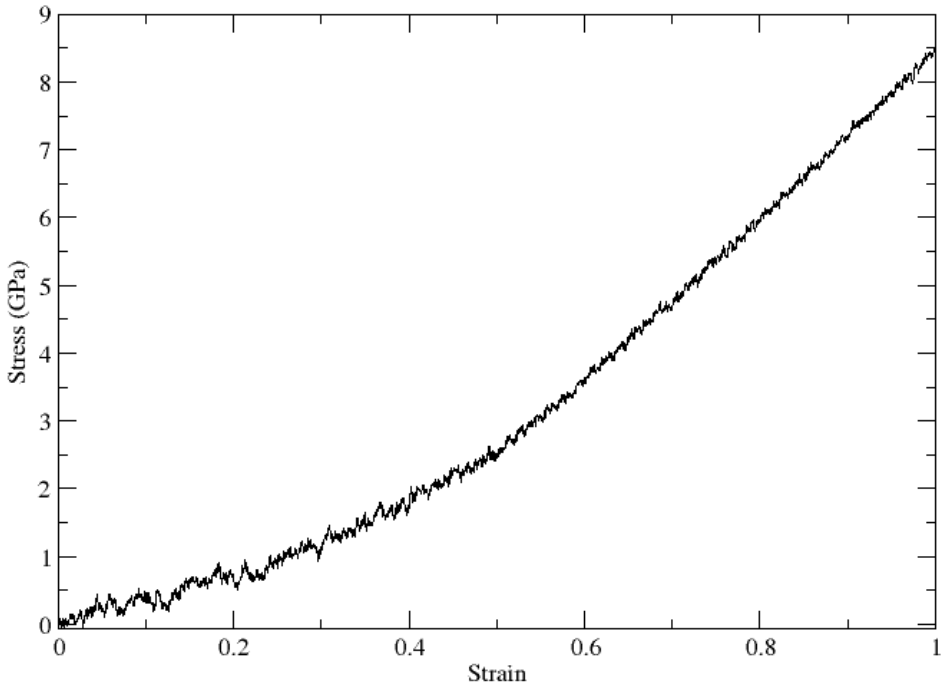


Figure 30: Illustrative stress vs. strain plot for a collagen like peptide (collagen region 4), showing the “toe shaped” curve region at low strain, followed by an linear region which would continue until fracture (fracture not possible with MD technique employed, instead simulation was run to 100% extension).

For calculation of the Young’s modulus it was decided to use the linear medium strain region of the stress strain curve (strain 25 - 50%), for both the tensile and lateral modulus simulations, to obtain an optimum line fit. Absolute values calculated for the tensile Young’s modulus varied from 8-14 GPa, making them at the upper range of the previously published values. However, as we are looking at the relative differences between the cross-linked and wild type collagen peptides, owing to the different primary sequences and the heterogeneity of collagen, this slightly high value will have no influence.

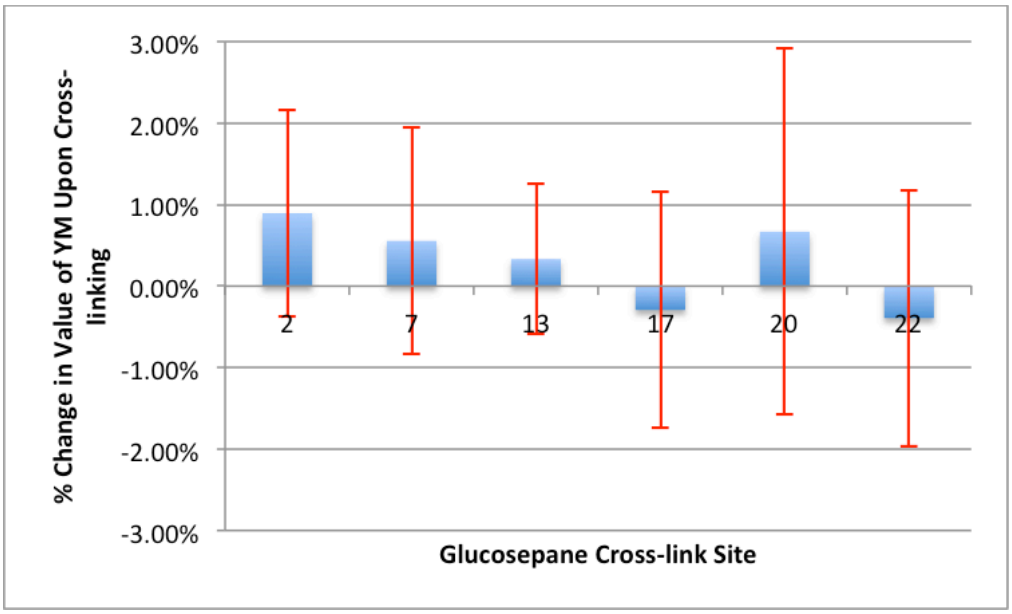


Figure 31: Bar chart showing the percentage change in the tensile Young's modulus upon the formation of a Glucosepane cross-link relative to the wild type collagen. The uncertainty in the calculated values is illustrated by the red error bars.

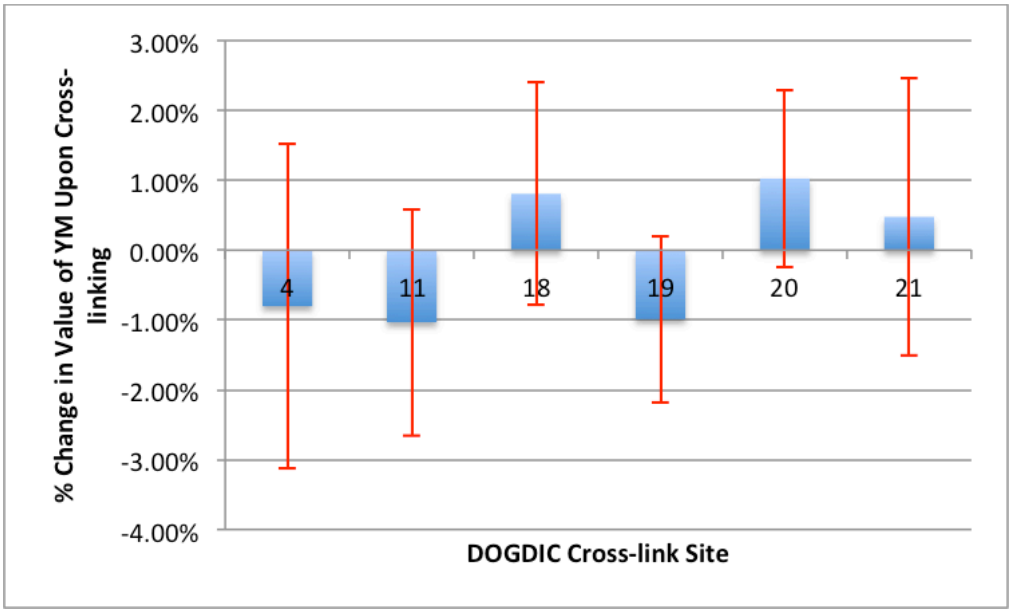


Figure 32: Bar chart showing the percentage change in the tensile Young's modulus upon the formation of a DOGDIC cross-link relative to the wild type collagen, the uncertainty in the calculated values is illustrated by the red error bars.

As can be seen from Figure 31 and Figure 32, there is no statistically significant increase in the tensile Young's modulus on introduction of a glucosepane or

DOGDIC cross-link respectively. Decreases in the values for the Young's modulus is likely an artefact of inaccuracies in the values used for the cross-sectional area of the peptide. As a result of this, no conclusions can be drawn from the data other than the fact that intra-molecular AGEs cross-linking has little effect on the tensile modulus of the collagen molecule, within the uncertainty limits of the method.

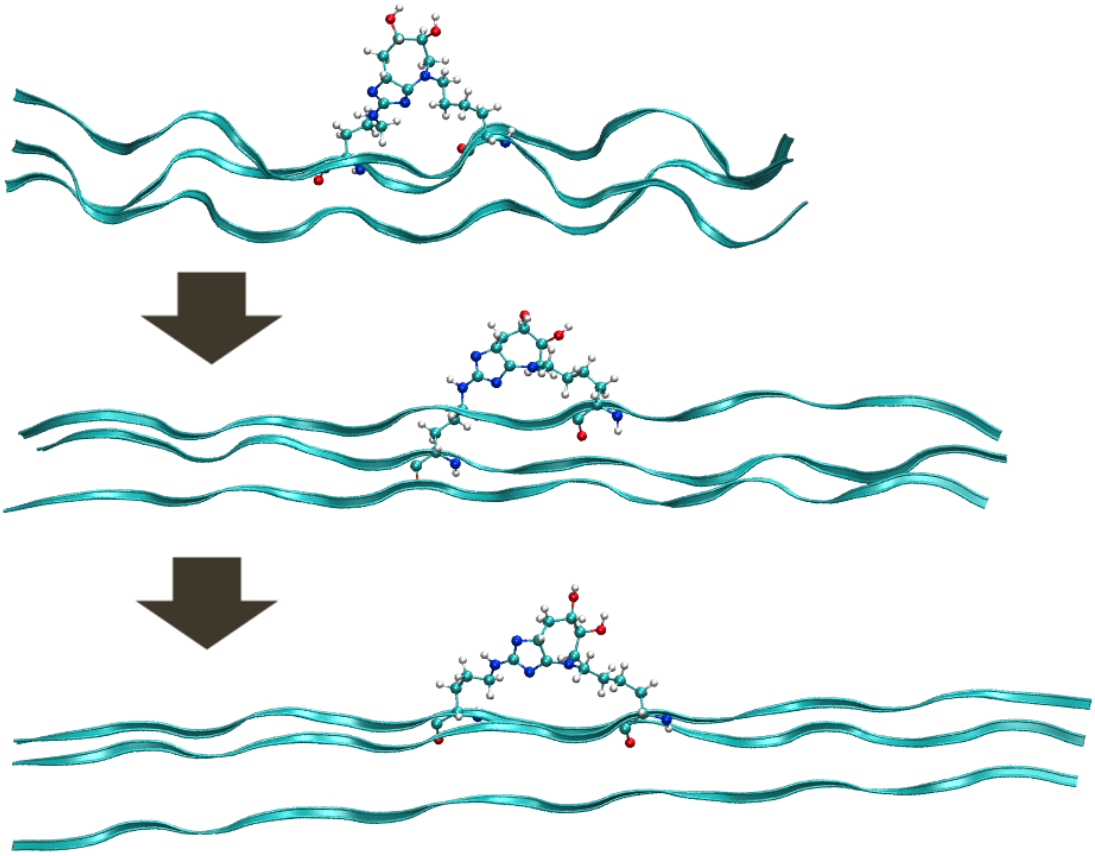


Figure 33: Series of three images showing the glucosepane cross-linked at site 20; the top image illustrating the starting structure, middle image depicting the structure at 20% strain and the final image showing the final structure at 50% strain.

The most likely reason for no increase being observed is that in our simulation we assume the load is applied uniformly to all three chains through the dummy

atom attached to the centre of mass of the three-polypeptide chains. This results in an extension of the chains at an almost equal rate, as seen in Figure 33, such that the separation of the two cross-linked amino acids occurs at a gradual rate such that it does not significantly increase beyond typical values until high strains, near the fracture region. Figure 34, shows this graphically, with the straining of the cross-link not occurring until above 25 Å extension, as illustrated by the 15% difference in Cα-Cα separation. Even in this region of increasing Cα-Cα separation, the N1-NZ and N2-NZ separation does not increase dramatically. Based on the data presented in Figure 34 and the final structure in Figure 33 it is my belief that region where the Young's modulus would increase, as a result of intra-molecular AGE cross-linking, would not be accessible physiologically in a fully solvated state, as the strain required would likely result in protein fracture.

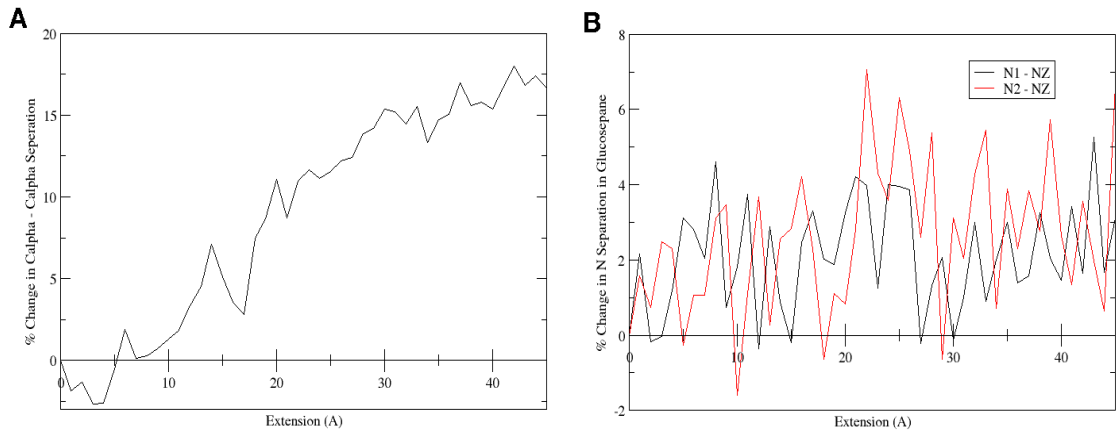


Figure 34: Plots showing the percentage increase in the separation between (A) alpha carbon atoms in the backbone of the cross-linked lysine and arginine residues (B) the nitrogen atoms within glucosepane (N1 and N2 from arginine and NZ from the lysine residue).

We then set out to calculate the lateral force-displacement ratio of the same collagen peptides to test whether this different loading mode is affected by the presence of the AGE cross-links. A force-displacement ratio was used instead

of the YM owing to difficulties in defining the cross-sectional area for lateral pulling, however as we are using relative differences the results will be equivalent. The results are presented in Figure 35 and Figure 36; again as with the tensile modulus there is no statistically significant effect of cross-linking on the mechanical properties of the collagen peptides. Figure 37 shows again, that the cross-links themselves are not strained heavily until the strain applied to the collagen molecule is above 50%, as seen in the final image.

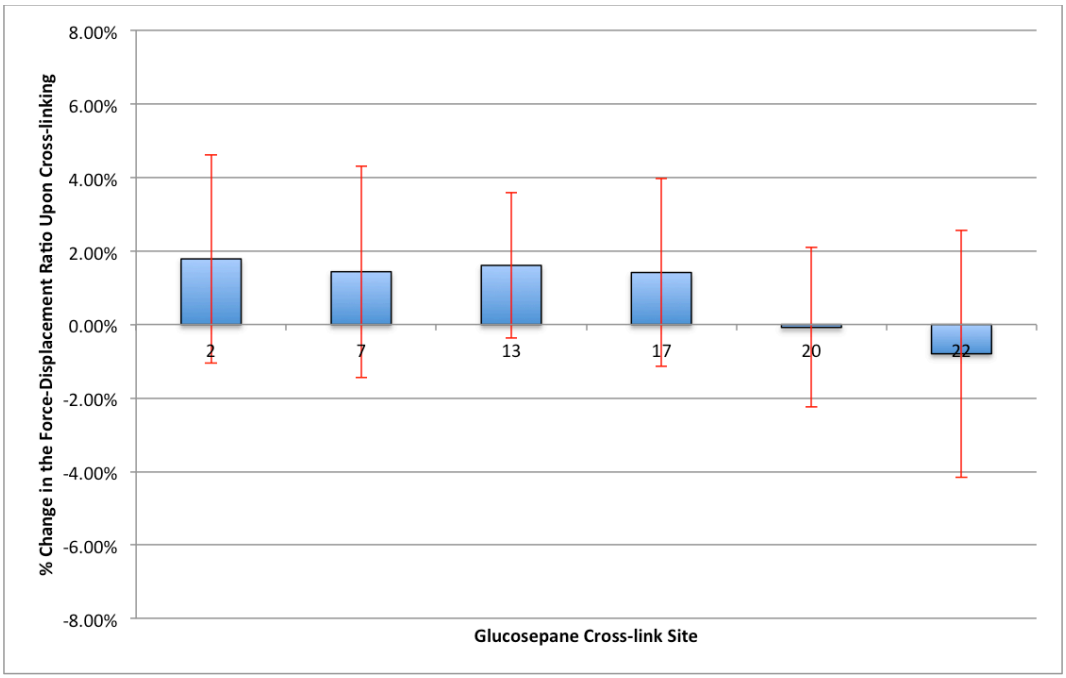


Figure 35: Bar chart showing the percentage change in the lateral force-displacement ratio upon the formation of a glucosepane cross-link relative to the wild type collagen, the uncertainty in the calculated values is illustrated by the red error bars.

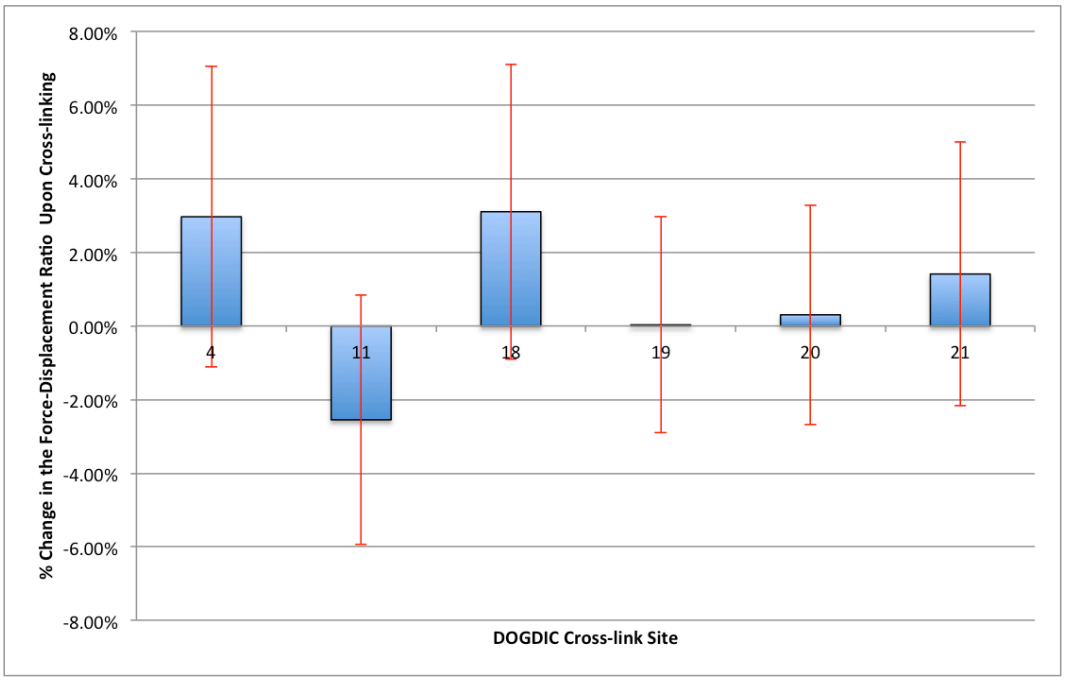


Figure 36: Bar chart showing the percentage change in the lateral force-displacement ratio upon the formation of a DOGDIC cross-link relative to the wild type collagen, the uncertainty in the calculated values is illustrated by the red error bars.

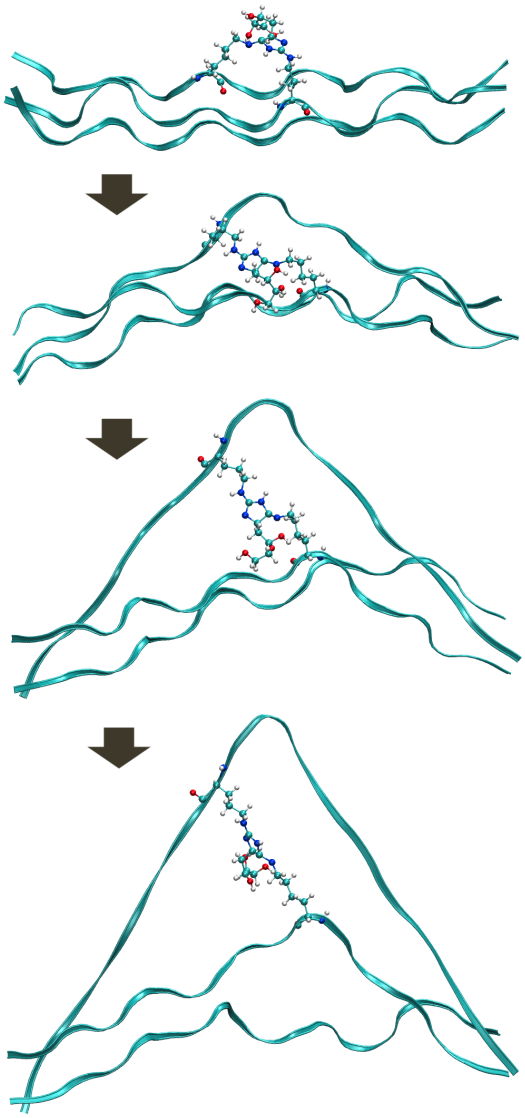


Figure 37: Series of four images showing the lateral pulling of DOGDIC cross-linked at site 20; the top image illustrating the starting structure, second image depicting the structure at 10%, the third at 35% strain and the final image showing the final structure at 50% strain.

It was decided to verify these results, to ensure that the null result is correct and not an artefact of the pulling methodology. Two further approaches were implemented to verify the effect of intra-molecular cross-linking; the first verification was conducted by comparing the results of a tensile SMD approach on the full collagen molecule that contained all 6 DOGDIC cross-links or all 6 glucosepane cross-links or no cross-links. The second approach was to make a slight alteration in the SMD tensile methodology, assigning the pulling group to

be only the terminal C atom of one chain, thus only one polypeptide chain will be explicitly pulled in the simulations. The peptide generated at site 20 was used so that both cross-links could be considered.

The whole collagen molecule models generated as described in 5.4.1.2 were subjected to the tensile modulus methodology to obtain results for the effect of cross-linking on the collagen molecules. The resulting force displacement and stress-strain curves obtained were significantly more complex than those for the solvated short collagen like peptides. The stress-strain curve, in Figure 38, exhibits the same initial “toe region” as shown for the collagen peptide in Figure 30. However, instead of being followed by a linear region to fracture, what we observe for the full collagen molecule is that there is a short linear region up to the 18% strain region, followed by a region 18 – 35% strain region where the stress undulates with increasing strain. This is most likely the result of the removal of the macroscopic kinks within the collagen molecule, resulting in an temporary decrease in the stress of the system. We also calculated values for the tensile Young’s modulus of the three models, from the linear region of the stress-strain plot with the results shown in Figure 38. As can be seen the changes again are not statistically significant, supporting the previous results suggesting that intra-molecular AGE cross-linking has no effect on the Young’s modulus of molecular collagen.

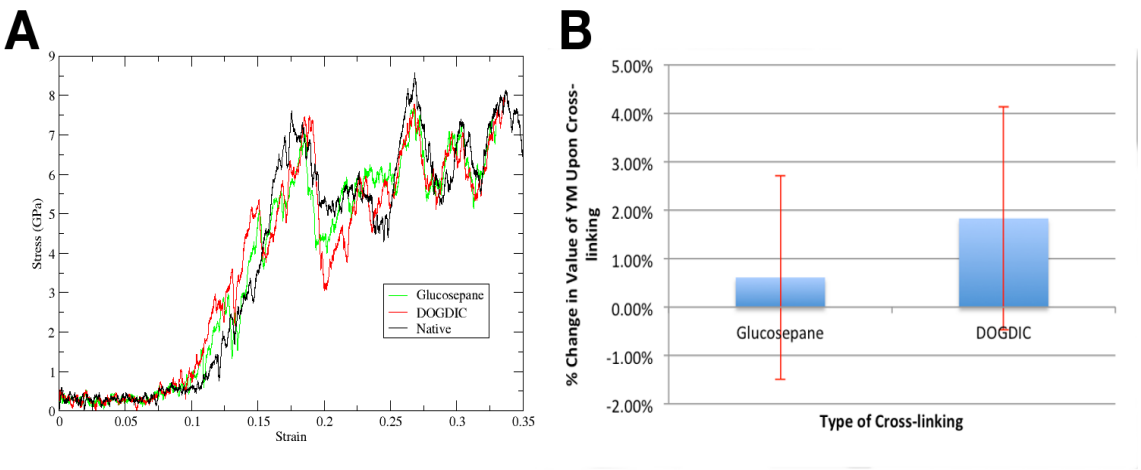


Figure 38: Figure showing the tensile mechanical response of a whole collagen with either DOGDIC or glucosepane cross-links present at all of the favourable binding sites to an applied load. A) Illustrates the stress vs. strain plot and B) The percentage change in the tensile Young's modulus upon the formation of all 6 of the glucosepane or all 6 of the DOGDIC cross-link relative to the wild type collagen, the uncertainty in the calculated values is illustrated by the red error bars.

To confirm this negative result we decided to conduct one more type of calculation using the peptide from site 20, allowing for testing of both cross-links effect. The calculations were repeated for all time points of the peptide model, so that a variety of different starting structures could be investigated. The single strand pulled, the α2 chain, was kept the same across the variety of cross-links and initial time-point simulations. The C-terminus end of the two other polypeptide chains were allowed to behave normally with no restraints present, whilst the N-terminus end of all three chains were fixed. Figure 39 shows the development of the simulations with time. Immediately from the force-displacement plot, Figure 40, it was apparent that there was an effect from the cross-links presence, as seen by the separate lower stress-strain line of the wild type collagen, with the two AGEs showing similar stress-strain curves. In the low strain region all three models exhibit similar behaviour, with a "toe-like" region. However at intermediate strains (~20% strain) we see the curves

beginning to diverge. This is a result of the strain on the cross-link increasing, as can be seen in the middle structure in Figure 39 by the linearization of the cross-link structure when compared to the top image. As the strain increases further (>30% strain) we see the other cross-linked polypeptide chain experience a localised high strain in the region between the cross-link and the fixed N-terminus, as can be seen clearly by the kink in the polypeptide chain in Figure 39. Through fitting to the linear regions of the stress-strain graph (at intermediate strain rates) we were able to calculate the tensile Young's modulus for this single chain pulling. Here we see a statistically significant increase in the Young's modulus from the wild type to the AGE cross-linked, as seen in Figure 40. This 8.8% increase in the Young's modulus suggests that, if a load is not applied uniformly, i.e. all chains strained by the same amount, then this may result in an increase in the Young's modulus. This increase from the single strand pulling also suggests that inter-molecular cross-linking will likely have a significant effect on the mechanical properties, unless all of the molecules respond uniformly to an applied load which, as shown from 5.3, is not the case.

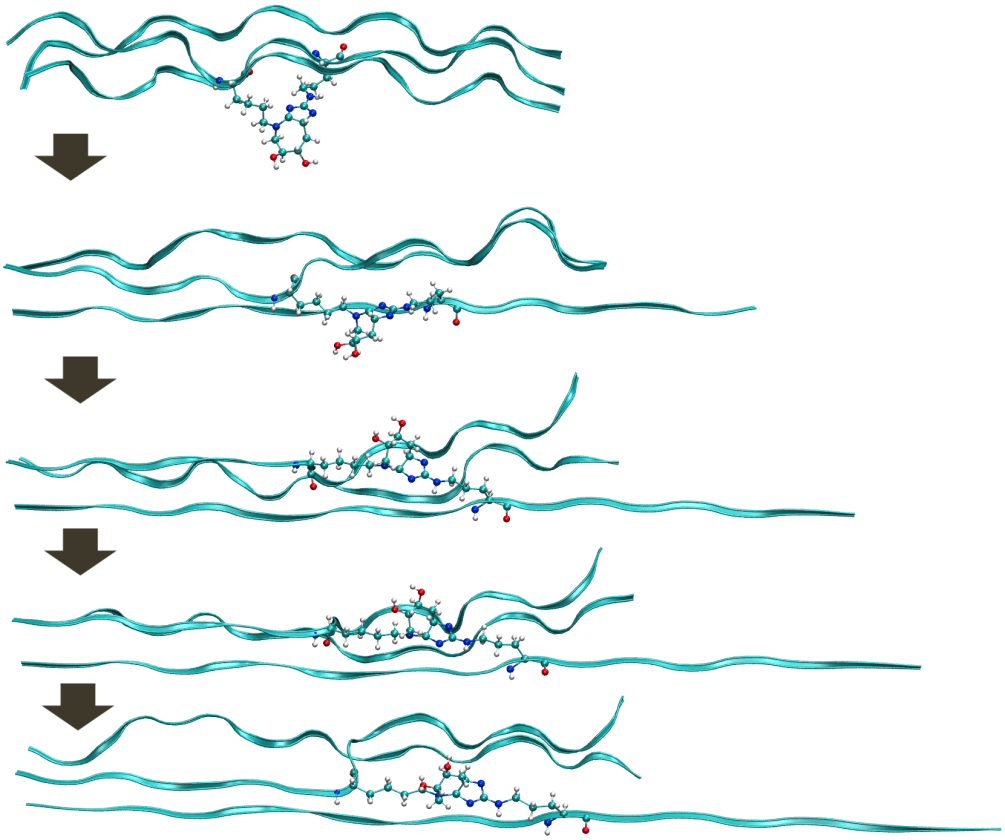


Figure 39: Series of five images showing the tensile pulling of a single polypeptide chain of a glucosepane cross-linked polypeptide (Collagen region 20); the top image illustrating the starting structure, second image depicting the structure at 10%, the third at 20% strain, the fourth at 30% strain and the final image showing the final structure at 50% strain.

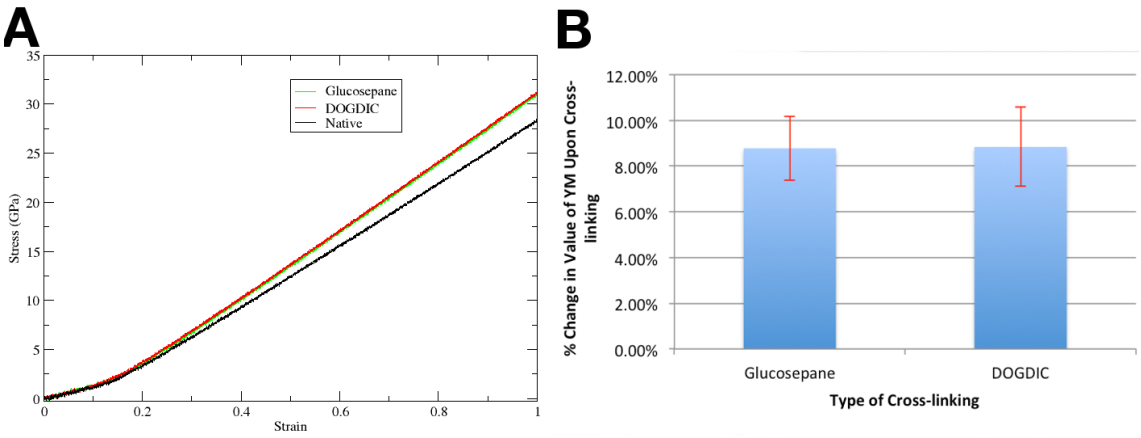


Figure 40: Figure showing the mechanical response of a tensile pulling of a single polypeptide chain of an AGEs cross-linked polypeptide (Collagen region 20). A) Illustrates the resultant stress-strain plot and B) Bar chart showing the percentage change in the tensile Young's modulus upon the formation of a DOGDIC cross-link relative to the wild type collagen, the uncertainty in the calculated values is illustrated by the red error bars, N=6.

This appears to be in contradiction to the macroscopic mechanics study conducted by Reddy *et al.*, in which they measure a 158% increase in the Young's modulus of ribosome cross-linked rabbits Achilles tendon (126). A possible reason for the discrepancy between their study and ours is the absence of the molecular environment in our study, with all of our calculation being conducted in a fully solvated system. Our decision was made to reduce the computational cost of the simulations. However in doing so we have removed the influence of neighbouring molecules on the mechanics of the pulled molecule, through intermolecular interactions. Another important factor that may account for the difference between our reported values and those from the macroscopic study of Reddy et al., is that the concentration of cross-links between our two samples may differ, with our study only employing a single cross-link whereas the experimental study was conducted with an abundance of ribose present, enabling multiple cross-links to form.

5.5 Summary

As was shown in 5.3, the mechanical properties of the collagen molecule, or collagen like peptide, exhibit a sequence dependency, with a change in a single residue resulting in up to a 6.6% change in the Young's modulus. The more stable triplets, higher melting temperature triplets, exhibit a lower value for their elastic modulus. This has two main consequences; the first is that care must be taken in experiment design, as sequence variance will make experiments utilising different sequences incomparable. Secondly, the heterogeneity of the collagen sequence will likely lead to localised regions of high stress or strain, which could lead to micro-unfolding or axial sliding. We presented the

mechanical influence of 28% of the possible triplets present in collagen, which can be used to quantitatively predict mechanical properties.

In the second part of this study, we aimed to probe the change in the mechanical properties upon AGEs cross-link formation, specifically glucosepane and DOGDIC, testing both the tensile and lateral modulus. After a comprehensive study of all 12 of the favourable cross-linking sites compared to the wild type, using six repeats, it was found that no significant changes occurred within the errors of the technique. Further testing utilising all of the cross-linking sites within a full-length collagen molecule equally resulted in no significant changes being observed. This was thought to be due to one of two things; either the absence of the fibrillar environment during the simulations or that only inter-molecular AGE cross-linking had an effect on the mechanical properties, as has been observed in macroscopic experimental studies previously (126).

Chapter 6 Constructing a Realistic *Homo sapiens* Homology Model

6.1 Introduction

Proteins are polymeric biomolecules built up of the 20 naturally occurring amino acid monomers coded for in DNA, in addition to the non-naturally occurring post-translational modified amino acids, such as hydroxyproline. The primary sequence alone gives no implication of the conformational structure of the protein; it is for this reason that, for successful computational studies of biomolecules, an experimentally derived structure for the protein is necessary, most commonly from X-ray diffraction studies. Despite advances in technology, the timescale for obtaining crystal structure data for a protein is still on the scale of a few months to a year, and therefore there are a huge number of identified proteins for which an experimental structure is still missing. Some proteins are too large for NMR analysis and cannot be crystallised for X-ray diffraction and thus experimental structures may never be known. It is for this purpose that computational modelling has a part to play, utilizing information from known structures and applying it to make valid sequence based predictions of the structure for unknown proteins.

Homology modelling is based on two main principles; first that knowing the amino acid sequence should be sufficient to predict the structure, owing to the structure being uniquely determined by its amino acid sequence. The second is that the overall structure of a protein alters much slower during evolution than the associated primary sequences. Therefore proteins with a similar sequence are going to adopt an almost identical structure, and vaguely similar sequences

will adopt structures with similar conformation. With an abundance of structural and sequence information present in open access databases, such as the Protein Data Bank (286) and wwPDB (287), in the late 90s, Burkhard Rost, was able to determine two zones of sequence alignments. One zone was called the "safe" zone, where the sequence is almost guaranteed to fold into the same structure. The second is known as "twilight" zone, where the likelihood is completely random (288). The regions are defined based on the relationship between the sequence similarity and the number of aligned residues.

Homology modelling for two sequences of the same length normally takes the form of a five-stage process, with the initial stage being the identification of a suitable template sequence from a database of known protein structures. This stage is followed by alignment of the sequences, based on the overlap of similar residues, initially using fast programs such as BLAST (99) or FASTA (289). Large dissimilar regions are filled in, with segments of other template proteins in a more refined alignment.

Once the aligned sequence is determined, the backbone of the target sequence is built based on the coordinates of the template structure. If the template residue is the same as the target sequence, then the whole side-chain is added in at this stage. If not, the next stage is to add in the side-chains for the non-overlapping residues. Two possible approaches are used; either add in the C_α and C_β atoms, though this only works for high sequence identity structures, or a combinatorial approach is used, which places rotamers of the side-chains into the structure, based on libraries of common rotamers obtained from experiment, with an energy scoring function used to determine the best placement (290).

The fourth stage is model optimisation, through the use of predominately MD (occasionally DFT level theory) calculations, to minimise the structure to its low energy form. With sufficiently long simulations it is hoped that the trajectory will result in the complete folding to the true structure.

The final stage is the validation of the model. With large sequence similarity structures, < 90% similarity, the results of the optimisation can be used to compare to the crystal structure of the template structure. Alternative methods of validation are; by monitoring the energies of the structure, RMSD, bond lengths, radial distribution functions and distributions of polar and apolar residues over the duration of the simulation. Sometimes inconsistencies in the model, such as misfolding, may not invalidate the model for its intended use, for example, if the inconsistency occurs at a position far away from the site of interest, whether it is an active site of the enzyme or a biomolecule-binding site.

For disease modelling and determination of pathology and possible treatments it is the proteins present in the human body which are of most interest to researchers. However, so far, none of the structures for the 28 members of the collagen family have been determined for the *Homo sapiens* species. The largest portion of the native *Homo sapiens* structure determined to date is a 20 triplet region of the type III collagen molecule (291). Determination of collagen structure is complicated owing to a number of factors, for example its large size prevents determination by NMR spectroscopy. Additionally, unlike crystals, X-ray diffraction determination of biological fibres is complicated by the anisotropic resolution of the resultant electron density map, with a lower resolution in the axis perpendicular to the fibre axis making it hard to distinguish supramolecular structural features (236). Therefore a homology model of the fibrillar type I

Homo sapiens structure is vital for a better understanding of the structural related changes and responses in collagen disease pathology.

6.2 Methodology

6.2.1 Identifying Template Structures - BLASTp

The web portal version of Standard Protein BLAST, part of the blastp suite from the US National Center for Biotechnology Information, is used for database searching for the template sequence. The accession numbers for the human target sequence used are CO1A1_human (P02452) and CO2A1 (P08123), which includes all the hydroxyproline and hydroxylysine residues, as designated in the post-translational modification section of the entries. A number of databases were used for the search including the Protein Data Bank (231, 286, 287), UniProt (230), SwissProt (292) and NCBI own libraries (293). For gene sequence data only, a manual search is conducted to identify if an available crystal or experimentally derived PDB file is obtainable.

6.2.2 Transposing the system – BLAST

Once a suitable reference model PDB has been identified, the next stage is alignment of the template and target sequences to look at the relative differences in residue types at the local positions within the primary sequence. Aligning the sequences, using BLAST, beginning from a point of strongest sequence similarity at the C-terminus, such that if the target and reference sequences are a similar length then they will align, with maximum coverage.

Once the sequences are aligned the positional co-ordinates of the side-chain atoms of the reference model are removed, leaving just the backbone intact. The sequence data of the target molecule is then transposed onto the positions of the reference protein's backbone, giving a protein backbone with the sequence of the target molecule. The side chain positions are added to the backbone atoms through the use of Leap, which will add the atoms in based on its template structure. Chloride ions will then be added to negate the charge and the same periodic boundary conditions used in the initial model will be applied to this simulation.

An initial MD simulation of 20 ns is run using the same MD conditions described previously for the simulations of the cross-linking within the Rattus *norvegicus* model (as described in detail in section 4.2). However initial restraints of 100 kcal/(mol Angstrom^2), were added during the minimisation. Firstly, all protein atoms were restrained for the first 5000 minimisation steps of steepest descent minimisation, to remove any overlap of the water or chloride ions. Then the restraints were removed from the light atoms, so that the heavy carbon, nitrogen and oxygen atoms were restrained for the next 5000 minimisation steps. The restraints were then lifted such that only the backbone was restrained for the following 5000 steps, before finally all of the restraints were removed for the final 5000 steps of minimisation. The restraints were briefly reintroduced onto the backbone atoms during the initial heating from 0 K to 100 K at a heating rate of 0.5 K/ps. Finally unrestrained heating in the NVT ensemble was finished, taking the temperature of the simulation up from 100 K to 298 K, at a heating rate of 1 K/ps. A production run was then conducted in the NPT ensemble for a simulation time of 19.6 ns.

6.2.3 Determining the D-band periodicity

Due to the difference in sequence between the target and reference sequence there may be increased or decreased diameter of the collagen triple helix, with the replacement of small residues such as glycine with bulkier residues such as tyrosine or vice versa. This introduction of the bulkier residues in the structure may result in an alteration of the packing of the collagen molecules within the fibril. To test if varying the packing will alter the stability, and thus the energetics of the new homology model, we have tried varying the dimensions of the periodic box by – 1.5% to + 1.5%, to expand or decrease the separation of the molecules within the fibrillar structure.

The input for testing whether the packing may alter with the different sequence, is the final structure from the previous 20 ns simulation. In this investigation the dimensions of the periodic box is varied from – 1.5 % to + 1.5% in 0.5% increments, whilst maintaining the density of water in the simulation cell by adding or removing water based on the difference in volume. The seven models with varying dimensions are then run for 500 steps steepest descent minimisation, followed by 4500 steps of conjugate gradient minimisation, with heating under the NVT ensemble for 450 ps, followed by 30 ns of production at 310 K. The RMSD and energy of the protein in the model is monitored over the final 10 ns of the simulation to determine the relative stability of the model to find the preferred d-period dimensions.

6.3 Results and Discussion

Accession Number	Description	Max Score	Query Coverage	Sequence Similarity	PDB Available
NP_001003090.1	Collagen1 α1 - *Canis lupus familiaris*	2573	100%	97%	NO
NP_001029211.1	Collagen1 α1 - *Bos Taurus*	2570	100%	97%	NO
NP_001310708	Collagen1 α1 - *Equus Asinus*	2559	100%	97%	NO
P02454	Collagen1 α1 - *Rattus Norvegicus*	1683	100%	91%	YES
3HQV_A	Low resolution Molecular Envelope	939	76%	84%	YES

Table 8: Five highest scoring, reference sequences from the blastp search, of the alpha1 chain of the *Homo sapiens* collagen type I sequence. Column one gives the accession number for the corresponding database entry, column two describes where the sequence is from, column three is the BLAST max score, column four the overlap of two sequences, column five the sequence identity similarity based on the two sequences and finally the sixth column shows whether an experimentally derived structure is available for the reference sequence.

Accession Number	Description	Max Score	Query Coverage	Sequence Similarity	PDB Available
P02466	Collagen1 α2 - *Rattus Norvegicus*	1984	100%	91%	YES
Q01149.2	Collagen1 α 2 – *Mus*	1927	100%	90%	NO
O46392	Collagen1 α2 - *Canis lupus familiaris*	1906	100%	94%	NO
P02465	Collagen1 α2 - *Bos Taurus*	1769	100%	92%	NO
3HQV_B	Low resolution Molecular Envelope	986	78%	85%	YES

Table 9: Five highest scoring, reference sequences from the blastp search, of the alpha2 chain of the *Homo sapiens* collagen type I sequence. Column one gives the accession number for the corresponding database entry, column two describes where the sequence is from, column three is the BLAST max score, column four the overlap of two sequences, column five the sequence identity similarity based on the two sequences and finally the sixth column shows whether an experimentally derived structure is available for the reference sequence.

Typically a scoring function program is used to determine the accuracy of the sequence template relative to the model sequence. This was conducted in this investigation through the use of the Blastp software suite. Table 8 and Table 9 display the five highest scoring hits for both the α1 and α2 chain respectively, along with the respective percentage coverage and similarity to the target sequence.

From the results of the blastp database search it is clear to see that the highest scoring sequence, for which an experimentally determined structure is known, is that of collagen α1 and α2 for the *Rattus norvegicus,* which had scores 1.8 and 2 times larger than the next sequence with an experimentally determined structure 3HQV. It was decided to attempt to use the structure of the *Rattus norvegicus* sequence, which we had been using in our previous simulations, as the reference structure to generate the model structure for the *Homo sapiens* sequence. We hope this approach will generate a reliable method, owing to the strong sequence identity similarity of 91% between the two sequences, thus far exceeding the “safe” region threshold described by Rost *et al.* (288).

Upon aligning the two sequences it was possible to see that they had a nearly equal number of residues; the α2 chains contained the same number of residues with the α1 chain containing one extra residue in the human sequence. However the additional residue is located in the non-helical telopeptide regions, such that the helical portions of both sequences consisted of exactly the same number of residues. For this reason we aligned the sequences from the beginning of the triple helical portion so that this had less of an influence. The side-chains of the *Rattus norvegicus* structure are removed in ptraj part of AmberTools14, leaving just the backbone atoms. The human sequence is then

transposed onto the backbone of the *Rattus norvegicus* sequence, with one amino acid added in the telopeptide region, before being loaded into LeaP to add in the missing side-chains of the amino acids. The structure produced had a number of local clashes between side-chains, owing to the different volume of the side-chains between the two sequences, and the fact that LeaP uses a template based approach to add in the missing side-chains, which does not take into account the neighbouring amino acids. These overlaps were overcome when the backbone-restrained minimisation was conducted. This produced a single collagen molecule with no local steric clashes. However it may not contain the optimum structure at this point, as this calculation was conducted assuming the same packing of the collagen molecules within the fibril, which may not be the case. Hence further MD runs to find the lower energy structures are required, which take into account the local environment around the collagen molecule within the fibril, as this environment will influence the conformations that the side-chains will adopt.

Owing to the small variations in the sequences between the two species, there may be local regions where structural properties are altered. For example in the human α1 chains there are a number of proline substitutions in the 310-360 region, which results in a tighter coiling of the three polypeptide chains. There are a number of small sequence variations between the two different organisms. However these predominately occur in the telopeptide regions of the collagen molecules and therefore will not have a significant effect on the structure. Focussing on the triple helical region of the collagen molecule; there are 34 amino acid differences in the α1 and 79 differences in the α2 chains, with 17 and 27 of these differences changing the polarity of the residue from hydrophobic to polar or vice versa. Polarity is a big driver in protein folding, with

the hydrophobic residues clustering on the inside of the protein and the polar side-chains either on the outside, interacting with the water, or on the inside forming intra-strand hydrogen bonding. Hence these local differences in polarity between the two sequences may result in a different conformation of the collagen triple helix forming, with the potential for a slightly greater separation of the strands to accommodate the hydrophobic side-chain. The most significant difference between the two sequences is found in the α2 strand, with a ^{645}Glycine ➔ ^{639}Ser difference between the *Rattus norvegicus* and *Homo sapiens* sequences respectively which creates a slight “bulge” in the human structure compared to the rat structure. This phenomenon is well described by the previous studies of Bella *et al*, who reported a partial untwisting at the site of a glycine to alanine substitution (26). However the extent of separation is smaller in this case, owing to the fact that the glycine substitution in the rat to human sequence is an Yyy position glycine, as opposed to the Zzz position glycine that was subject of the Bella *et al.*, study. Due to these small differences in the local structure, the packing of the collagen molecules may also be slightly altered, and it was decided that we would also investigate small variations to the crystallographic unit cell dimensions of the *Rattus norvegicus* values to obtain the model crystallographic unit cell dimensions for our *Homo sapiens* homology model.

Seven different crystallographic unit cell dimensions were investigated for the homology modelling, with values used varying from -1.5 % to +1.5 % of the dimensions for the unit cell of the *Rattus norvegicus.* The water density was maintained in all of the different models by the subtraction or addition of TIP3P water molecules. The plot of the average energies of the models over the last 10 ns of simulation, in Figure 41, shows a very clear U-shaped curve with a

clear minimum at the 0.0% crystallographic dimensions. This suggests that the use of the same dimensions, as have previously been described for the *Rattus norvegicus* sequence, would give on average the lowest energy structures, implying that the packing of the *Homo sapiens* collagen is very similar to the *Rattus norvegicus* sequence. The U-shaped curve of the graph in Figure 41 shows a slight slant in its arms with a steeper gradient towards more negative values of variation. This would suggest that the homology model collagen is already close packed, and that further reduction in the separation of collagen molecules will lead to close contacts being introduced between molecules. This close packing may be the result of a wider average diameter of the *Homo sapiens* collagen sequence, owing to the above mentioned sequence differences.

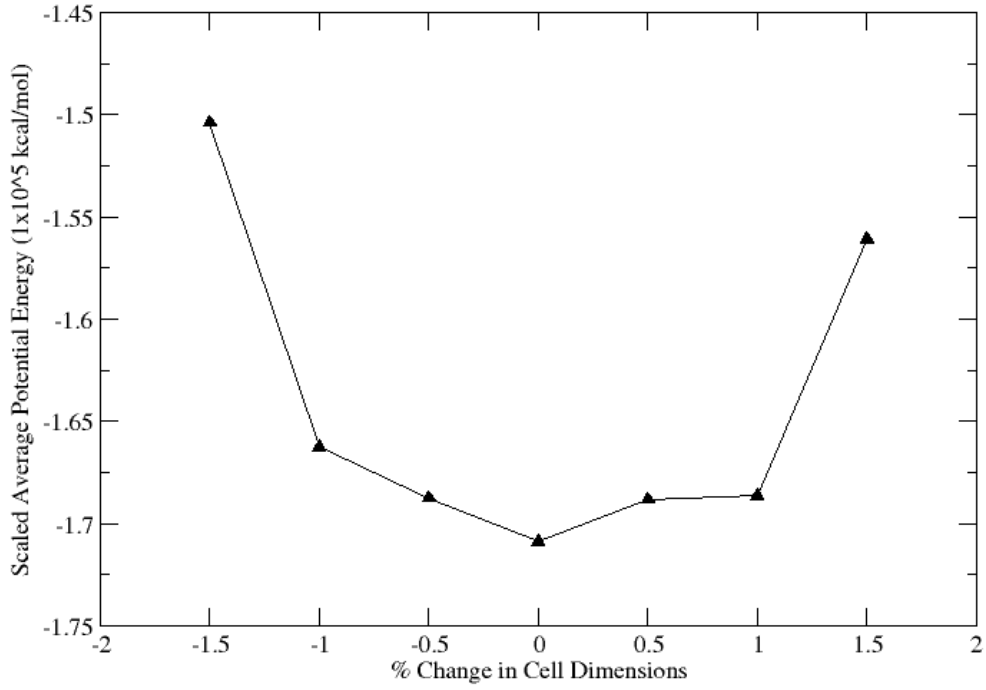


Figure 41: Plot showing the average scaled energies of the seven *Homo sapiens* homology models with varying cell dimensions over the last 10 ns of simulation. Energies are scaled to take into account the differing water content of the seven models.

In addition to looking at the average energies with varying cell dimensions, we also monitored the RMSD, as large fluctuations in the structure, after sufficient time for equilibration, would suggest instabilities introduced from sub-optimal cell parameters. A plot of the average RMSD of the backbone atoms is shown in Figure 42. The plot shows an M like shape, with a minimum RMSD value located at 0%. With increasing cell dimensions there is an initial increase, as the collagen molecule remains within the cut-off distance of the simulation. However, upon increasing the dimensions by 1.0% and above, the separation of the collagen molecules is large and the RMSD decreases as the molecule becomes more like an isolated molecule. With decreases in the cell dimensions we see an initial increase in the RMSD, as the protein undergoes large structural fluctuations to minimise the close contacts with neighbouring molecules. However, at the smallest values of the cell dimensions, there is a large number of close contacts with neighbouring molecules, which restricts the degree of movement of the molecules and thus resulting in a reduced average RMSD.

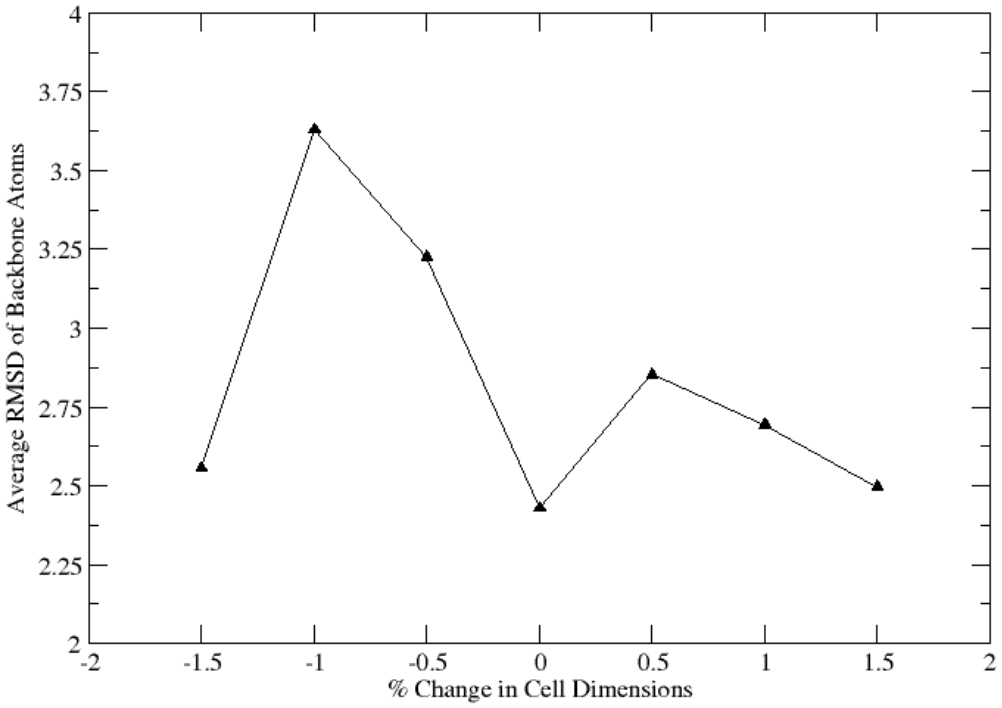


Figure 42: Plot showing the average root mean squared deviations of backbone atom positions for the seven *Homo sapiens* homology models with varying cell dimensions over the last 10 ns of simulation.

It was also noted that, on moving from the NVT to the NTP ensemble, there were minor variations in the volume of the cells, both compressive and expansive. This was necessary to maintain the 1 atmosphere pressure within the system. These occurred in all but one of the models. However the variations were only very minor, less than 0.03%. For example the 0% cell dimensions encounter a 0.03% increase in the x and y dimensions that would further suggest that the Homo sapiens sequence results in a slight expansion of the average diameter of the molecule compared to the *Rattus norvegicus* structure.

Given the results of the simulations into the investigation of the packing of our homology *Homo sapiens* molecule, relative to the template *Rattus norvegicus* structure, as summarised in Figure 41 and Figure 42, it was decided to use the values for the cell dimensions developed from the end of the template structure

simulation, as this gives the most stable simulation. Hence the unit cell dimensions used are 39.982 Å, 26.96 Å and 677.90 Å for edges *a*, *b*, and *c* respectively, and 89.24°, 94.59° and 105.58° for angles α, β and γ, respectively. The average structure for our homology model, using the chosen cell dimensions, is generated from the last 10 ns of the simulations. This will be the structure that will be used as a starting configuration for any future investigations upon fibrillar *Homo sapiens* type I collagen molecules.

6.4 Validation

To test that the homology model structure generated is reliable for use in future MD simulations, we conducted a comparison of a number of observables generated between two short MD simulations, one of our homology model and one of our well established model of fibrillar *Rattus norvegicus* type I collagen molecule as described in 4.2.1. A portion of 2 ns simulation of the equilibrated structures, tested by convergence of the potential energies, was used for the comparisons. Four main observables were monitored over this period; the density, system volume, temperature and the RMSD of the backbone atoms. The values for each of these observables, as a function of time, can be seen in Figure 43, with the red line representing the *Rattus norvegicus* model and the black line our *Homo sapiens* homology model. What can be seen in Figure 43A-C is that the values are not equal at the same time, nor would we expect them to be. However the range of values for each of the three observables is equal for both the *Homo sapiens* and *Rattus norvegicus* model simulations. This equivalent range in the observables would suggest a similarity in the stability of the two simulations, meaning that our homology model could be considered to

give results with a reliability that is equivalent to those obtained from the simulation of the widely used model for the *Rattus norvegicus* structure (35, 68, 78). The plot of the backbone RMSD in Figure 43D supports the conclusion of a reliable structure; with fluctuations being in a range consistent with the simulations from the *Rattus norvegicus* model simulations.

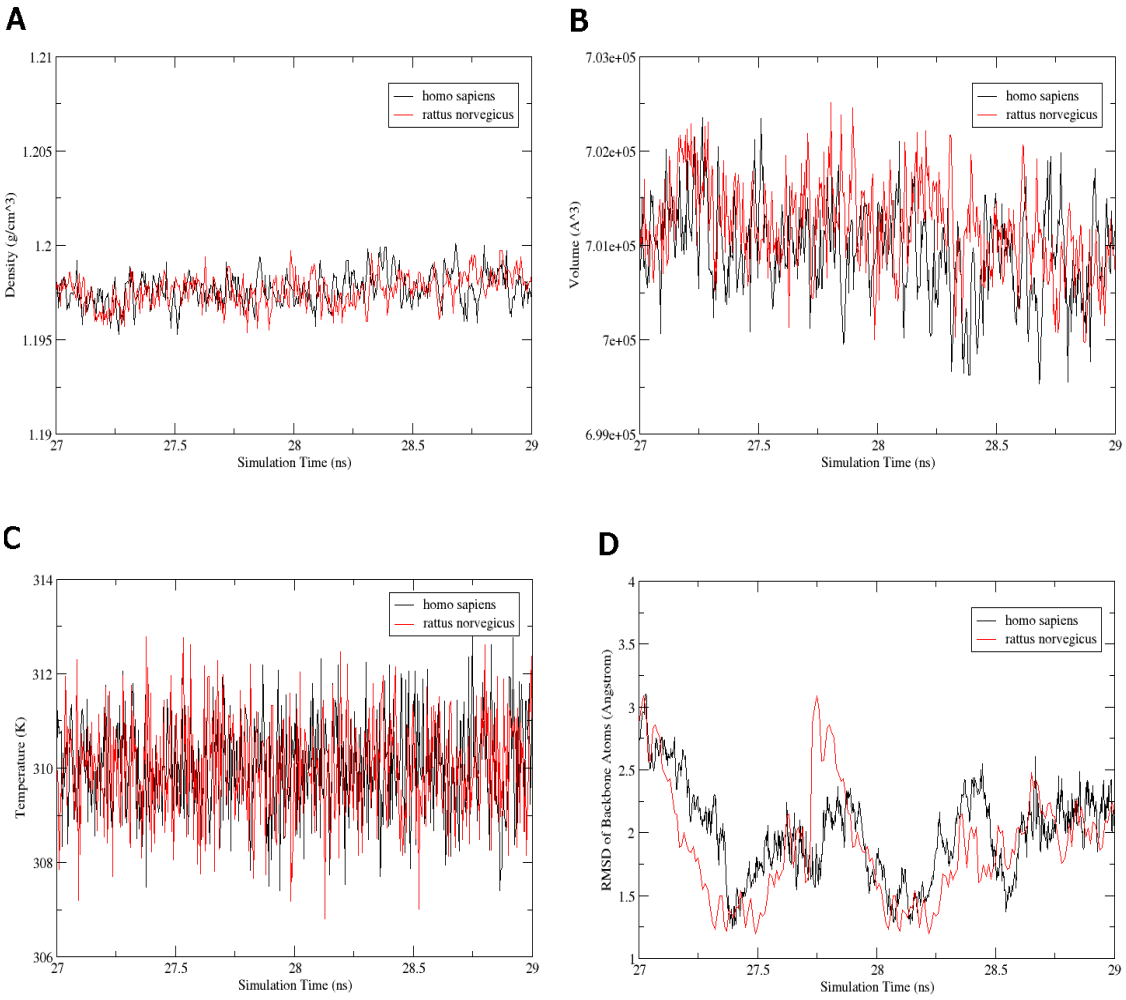


Figure 43: Comparison of system observables of a 2 ns simulation of the *Rattus norvegicus* model (red line) against our newly developed homology model for the *Homo sapiens* sequence (black line). The observables plotted are A) System density, B) System volume, C) Temperature and D) Root mean squared deviation of the positions of the backbone atoms relative to the average structure for the respective model.

6.5 Summary

Given the absence of a crystal structure for the structure of fibrillar *Homo sapiens* type I collagen, we developed a homology model using the crystal structure of the *Rattus norvegicus* sequence as a template structure. Owing to a number of variations in the sequence between the target and template, local variations in the structure were observed. An investigation varying the dimensions of the cell was conducted to explore whether the packing of the collagen molecules within the fibril varied. Through this investigation it was determined that no significant variation occurred to the packing, although the *Homo sapiens* structure did have a slightly larger average diameter which resulted in a 0.03% increase in the a and b dimensions. A number of observables were compared from a short 2 ns simulation of the equilibrated homology model to those of the established *Rattus norvegicus* model. The observables were found to be in a similar range and it was therefore inferred that the homology model for the *Homo sapiens* sequence produced a stable simulation.

Chapter 7 Relative Orientation of Collagen Molecules in a Fibril

7.1 Introduction

The packing in collagen networks has huge implications on how the tissue responds to a mechanical load. For example, in the skin the fibres form an anisotropic network to respond effectively to multidirectional forces, whereas in tendons the fibres align in one direction to maximise their effectiveness to respond to a uniaxial load. A lot of attention has previously been paid to investigating the way in which the collagen fibrils align and orientate within collagen fibres, as well as how collagen fibres align in fasicles. A variety of techniques have been employed to do this; scanning electron microscopy (SEM) (294), small angle X-ray scattering (295, 296), polarized light microscopy (297), infrared and polarized Raman spectroscopy (298–301). A major drawback of some of these techniques is the complex sample preparation. For example, in the SEM work conducted by Pannarale *et al.,* the samples were dehydrated prior to scanning thus changing their structural properties (294). The use of small angle X-ray scattering allowed the sampling of larger hydrated histological sections of 1.5mm thick. However this technique is still unable to sample *in vivo* or below the fibril scale (295).

Spectroscopic techniques offer the best chance of sampling *in vivo* collagen orientation, owing to their non-destructive, relatively simple, sample preparation. Non-invasive Raman Spectroscopy is even being used as a diagnostic tool to identify Osteogenesis Imperfecta (OI) type abnormalities in bone composition of patients (302). Spectroscopic techniques are possible owing to the IR and Raman active amide I (~1620-1700 cm^{-1}), amide II (1600-1500 cm^{-1}) and amide

III (1215-1300 cm^{-1}) bond stretch frequencies. However the use of spectroscopic studies to determine the orientation of the collagen fibrils is only possible through the use of a polarised light source, relying on the anisotropic response of the amide bonds to the incidence beam. When the incidence of the light source is parallel “out of plane” to the principal axis of the molecule, an isotropic response is obtained, whereas a sinusoidal anisotropic response is obtained when the incidence of light is perpendicular “in plane” (299). In the anisotropic response for “out of plane” incidence of radiation, a minimum intensity response is obtained when the polarisation angle of the incident light is parallel to the plane of the amide bond, and is maximal when perpendicular. Fourier transform infrared imaging spectroscopy (FT-IRIS) on highly orientated tendon collagen was used by Bi *et al.*, to generate spectral parameters based on the ratio of integrated areas of the amide I and amide II absorbance peaks, which they then used to determine collagen fibril orientation in various regions and types of cartilage (301). However FT-IRIS is limited to imaging the exposed surface fibres, owing to the strong adsorption of IR radiation by water, leading to low light penetration of the IR radiation in hydrated biological tissues (303). However the Raman spectrum of water is weaker and unobtrusive and thus Raman spectroscopy is often favoured for studying hydrated tissues. A number of studies have been conducted using Polarisable Raman Spectroscopic approaches to map fibril orientation within osteonal lamellae (298), tendon (299) and in the Haversian bone structure (304).

It is worth noting that, although the spectroscopic techniques (Raman and FT-IRIS) offer the closest to atomistic scale detail at present, the response of the amide I band is made up of contributions from all of the amide I scattering centres present in the structure, thus it will consist of multiple collagen

molecule's responses to the incidence light. For this reason there is still a need to develop more advanced techniques and methodologies to be able to sample the orientation of the individual collagen molecules within the fibril.

As previously demonstrated, the orientation and macroscopic structure of collagen fibrils plays a vital role in the function of the collagenous tissues. If we dive deeper into the structure of the fibril and begin to think about the role of the collagen molecule, we can see that it too has a significant role to play in the function. The alignment of the collagen molecules has been well studied, with the D-banding periodicity being the subject of countless research articles (15, 21, 305–307). However, to the best of our knowledge, no study to date has looked at the orientation of the individual collagen molecules within the fibril. The orientation of the collagen molecules about their principal axis will determine mechanical properties owing to the different possible intermolecular forces, biological interactions owing to the accessibility of the biomolecule binding sites and finally, in the context of cross-linking, the different residues which will be available to form inter-molecular AGE cross-links.

As mentioned previously, initially the collagen molecules aggregate based on the intermolecular forces, before later forming the covalent interactions via the mature enzymatic cross-link. This could therefore mean that the driver for the determination of the orientation of the collagen molecules will be to maximise the number of favourable inter-molecular interactions to form a low energy fibril. To investigate the lowest energy orientations of the collagen molecules we will use a novel two stage modelling approach. This new modelling approach takes inspiration from Adams *et al.,* 1995 work on computational method development for the determination of conformation and rotation angles of the pentameric

transmembrane domain of phospholamban (308). Our approach begins by conducting a comprehensive single point energy search of all of the possible orientations at small rotation intervals of 6°. The results of the single point energy search are then used to conduct short molecular dynamics searches of the lowest 150 orientations, for further sampling of the potential energy landscape, to find the lowest energy orientations.

7.2 Methodology

The study can be divided into two sections, both using the same model system; firstly a rapid single point energy scan of the potential energy landscape, followed by a more in depth investigation of the low energy orientations.

7.2.1 Building the Model

The model was constructed using the amino acid sequence for *Homo sapiens*, as this is the species of most interest in collagen research. A straight-chained structure of a collagen molecule, with the correct helical propensity, was generated using the Triple Helical Building Script (THeBuScr) (229). The primary sequences of the collagen peptide chains α1 and α2, translated from the genes COL1A1_human (P02452) and COL1A2_human (P08123) (230), were the inputs. Once again proline residues present in the Yyy position of the triplets were considered to be hydroxyproline in the study. Additionally hydroxyl-lysine residues stated in the modified residues of the UniProt entry were also included in the sequence, giving the same primary sequence as utilized in Chapter 6. To apply the rotation for the collagen molecule accurately, the assumption is made that the collagen molecule must be considered a straight

rod/cylinder, thus the straight molecule from the THeBuScr program was used directly. The next stage in the preparation was to align the principal axis (c-axis) of the collagen molecule to a Cartesian origin axis using the Orient script in VMD (309). In our case we aligned to the x-axis such that the backbone atoms had almost zero displacement in the y and z components.

From this aligned straight collagen molecule two different strands are created. First the molecule is replicated along the x-axis, preceded by a 36 nm gap region. The strands are then generated by taking a box of length 360 nm and positioning the box at the beginning of the collagen molecule, so that the strand contains a full collagen molecule, a gap region and a short 110 residue triple helical and telopeptide section. The second strand is generated by placing the end of the box at the end of the collagen molecule, such that this strand also includes a short triple helical region, telopeptide, a gap region and a full collagen molecule as seen in Figure 44. Thus, we have two models aligned to the x-axis, one a collagen molecule followed by a short collagen snippet and the second a short collagen like snippet followed by a collagen molecule.

Additionally I developed a Perl script that is capable of rotating a selection or all of a protein by a set number of degrees about a chosen axis; the script is also capable of translating a portion or the whole of a protein by a set number of angstroms along a chosen axis. Through an input of a PDB file the script reads the correct rows and columns for the selected atom’s coordinates, and will either multiply them by a rotation matrix, or will add the selected translation to the coordinates and generate the modified PDB file. Rotation must be conducted before translation, as the rotation occurs around the axis of the Cartesian system and not the principal axis of the molecule. However, as the

two strands principal axes are aligned with the Cartesian axis, then rotation is conducted accurately. It is from these two principally aligned strands that all of the subsequent models are generated from. Through translation we can build our complete model, which is two staggered collagen molecules, including the gap and overlap region at varying degrees of rotation. The reference model for our simulations is the 0° - 0°, two strands not rotated and then translated by 17 Å. This orientation is the linear version of the orientation from the 2006 Orgel crystal structure in which the glycine of the alpha1 chain is above the first residues of the other two chains that lay almost in a horizontal plane parallel to the z axis.

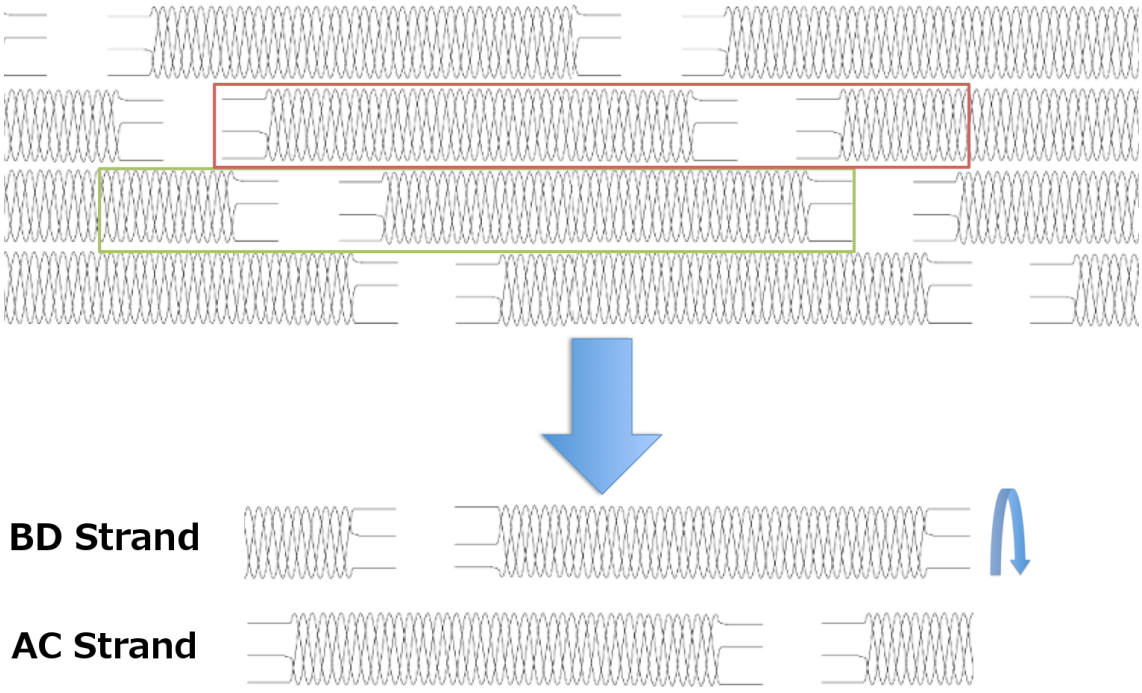


Figure 44: Schematic of the fibril (Top), with the red box (AC) and green box (BD), illustrating the regions of the collagen fibril used in the orientation study. After generation of the two strands, alignment to the x-axis, rotation about the x-axis, followed by translation, we obtain the model illustrated at the bottom of this figure, with the AC strand on the bottom and the BD strand above.

7.2.2 Single Point Energy

A bash script is used to generate 3600 different models for the two strands, orientated independently at 6° increments. The PDB files for all of the models are then fed through LeaP part of the AmberTools14 to generate the input files, during which the models are solvated using TIP3P water with a buffer of 8.0 Å and the charge of the system is negated by the addition of chloride ions. The models then undergo a very short 1000 step conjugate gradient minimization, during which all of the protein atoms are restrained using a force constant of 1000 kcal/(mol Angstrom^2). This is necessary to remove any high energy fluctuations caused by close contacts with the recently added water and Cl^- ions. A further one step minimization was conducted to get a single point total energy for the system, which is then used to direct the search in the second stage of this investigation.

7.2.3 Short MD runs

A three dimensional plot of the total energy against the rotation angle in the two different strands is generated to show areas of low energy, which are strong indicators of possible favourable orientations. From the 3600 different models the 150 lowest energy structures are identified from the results of the single point energy determination. Owing to the large size of the models (~1.6 million atoms), and hence long computational times involved in the calculations, it was not possible, with the resources, to investigate any more than the lowest 4%. The lowest 150 structures then undergo short molecular dynamics simulations using a cut-off of 8.0 Å; this is to allow the models to relax further into their most favorable orientations. The models initially undergo 500 steps steepest descent and 2500 steps of conjugate gradient minimization; before a two stage heating

simulation of 20 ps from 0 K to 100 K and 30 ps for 100 K to 310 K with 200 kcal/(mol Angstrom^2) restraints applied up until this point on all the protein atoms being removed. Finally the model undergoes a further 100 ps simulation in the NPT ensemble at 1.0 atm pressure. The results of the 100 ps NPT simulation are used to determine the low energy orientations of the collagen molecules within the fibril. This is done in two ways, firstly through comparing the energies of the respective orientations and secondly through the use of another script, which calculates the orientations from the trajectories of the simulation, allowing us to monitor and compare the abundance of certain orientations.

The second Perl script uses ptraj, part of the AmberTools14 package, to remove the water ions and side-chain atoms. Then using 2D vector based mathematics about the x-axis we determine the relative orientations of each of the heavy atoms within the molecule relative to the 0°-0° model. More specifically the script calculates the angle of rotation from the dot product of the vector defined by the new position to a point on x-axis, relative to the position of the same atom in the 0°-0° configuration to a point on the x-axis. The results are then averaged over all the atoms to get the relative orientation of the whole molecule, and this is repeated for each of the time points within the trajectory. The scripts functionality and accuracy were tested on a test collection (N=30) of known rotated models of the same system, with results reported to within 0.345% accuracy.

7.3 Results

7.3.1 Single Point Energy

3600 individual single point energy calculations were conducted rotating the two strands independently by 6° increments from 0° - 354°. The energies of these systems were then plotted, showing the energy as a function of its orientation in the AC and BD strands, which can be seen in Figure 45. For a small proportion, less than 2%, of the simulations there were close contacts or steric clashes between large side-chains on the collagen molecules, which resulted in a huge increase in energies, orders of magnitudes larger than the average energy. Due to the relatively low proportion of these high energy structures it was decided to omit these from the results of the simulations. These can be see by the bright yellow squares in Figure 45.

What can be seen from Figure 45 is that there are several areas of higher energy illustrated by the red regions, and a large region of lower energy illustrated by the blue regions. Due to the scale of the energy, further differentiation within the blue region was not possible from Figure 45. However through changing the range plotted on the graph from -6.5E+6 - -7.0E+6 kcal/mol to -6.85+6 - -6.91E+6 kcal/mol, it was possible to obtain good differentiation, as seen in Figure 46. This reduction in range plotted means that the red regions are now off of the scale and hence are represented by white (absent) regions. From Figure 46 we can see there is a wide distribution of energies throughout, with no significant clustering of low energy regions. However what can be seen in Figure 46, is that there are a large number of smaller regions of low energy configurations, illustrated by the dark blue regions. Through comparison of the energies of these regions we are able to identify the lowest 150 orientations from the single energy point scan.

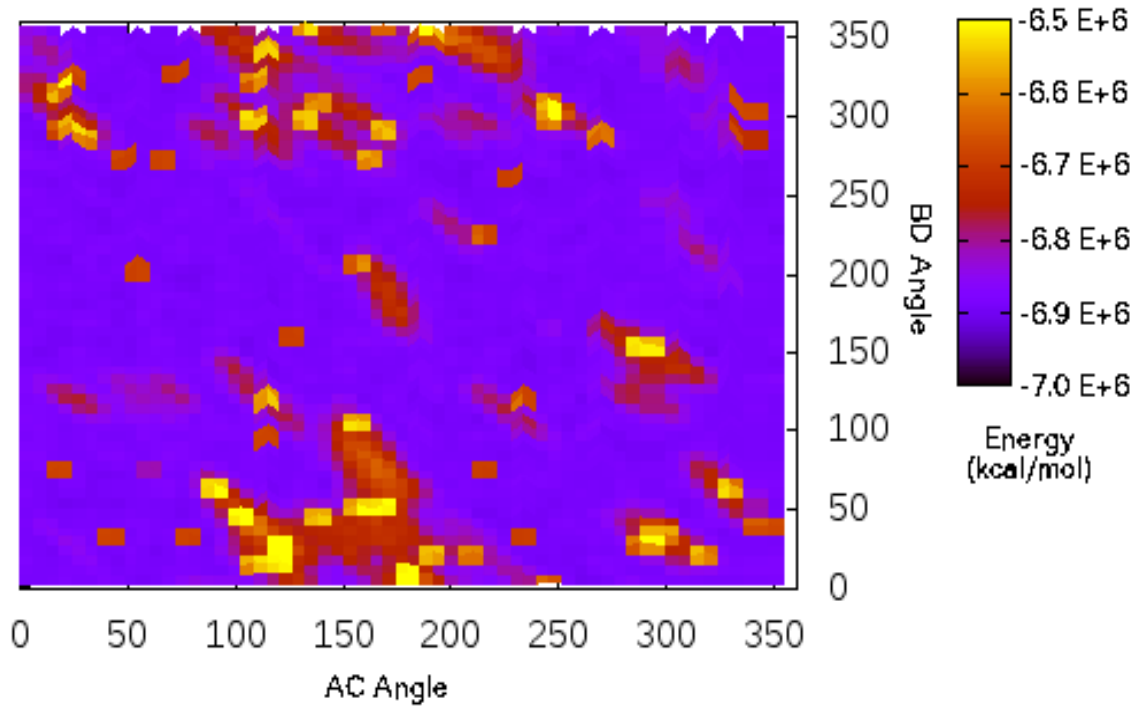


Figure 45: Plot of the potential energy as a function of the orientation angle of the AC strand and the orientation of its corresponding BD strand. Potential energy is defined by the colour, on a sliding scale yellow - high energy to blue – low energy.

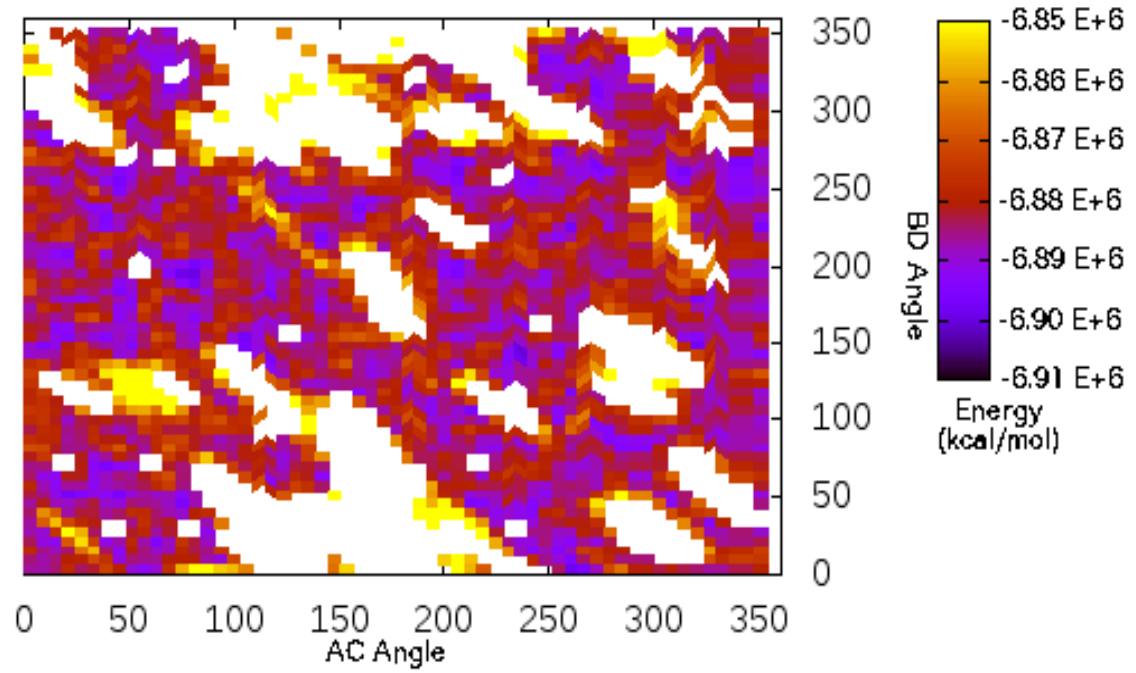


Figure 46: A Plot of the potential energy as a function of the orientation angle of the AC strand and the orientation of its corresponding BD strand, with the scale plotted reduced, for increased resolution. Potential energy is defined by the colour, on a sliding scale yellow - high energy to blue – low energy, with white representing values significantly off of the scale.

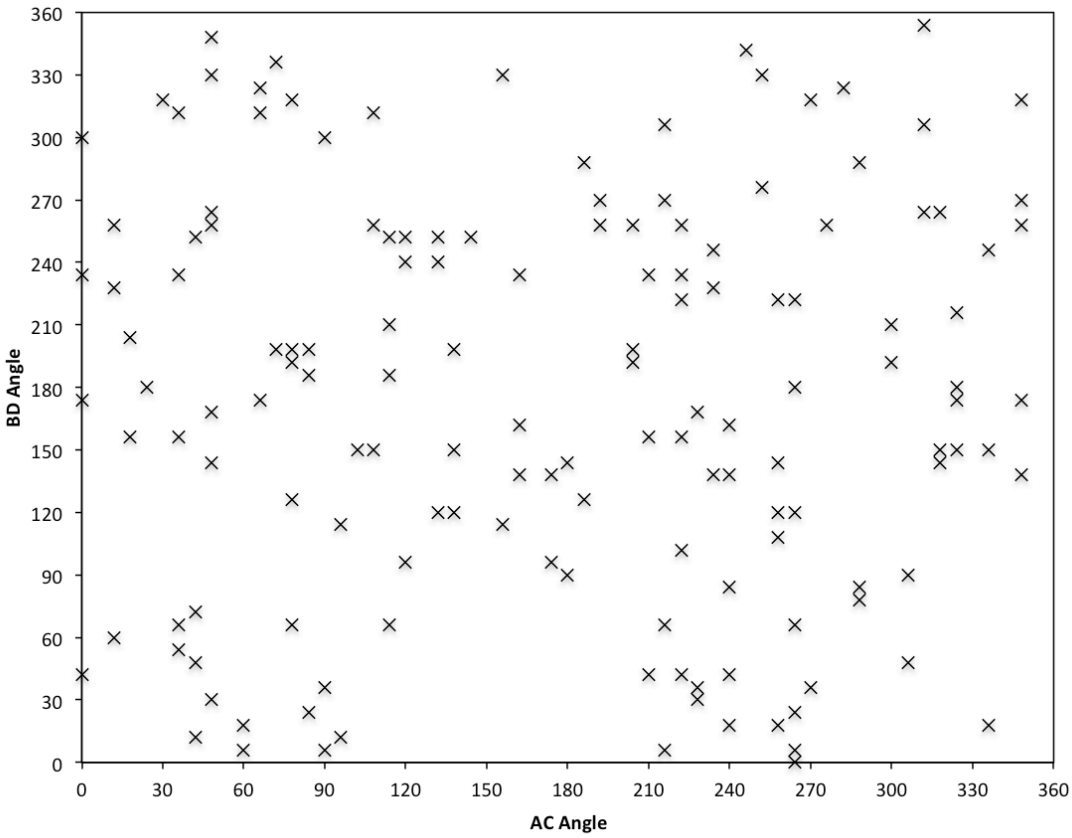


Figure 47: Distribution of the 150 lowest energy orientations determined from the single point energy rotation search.

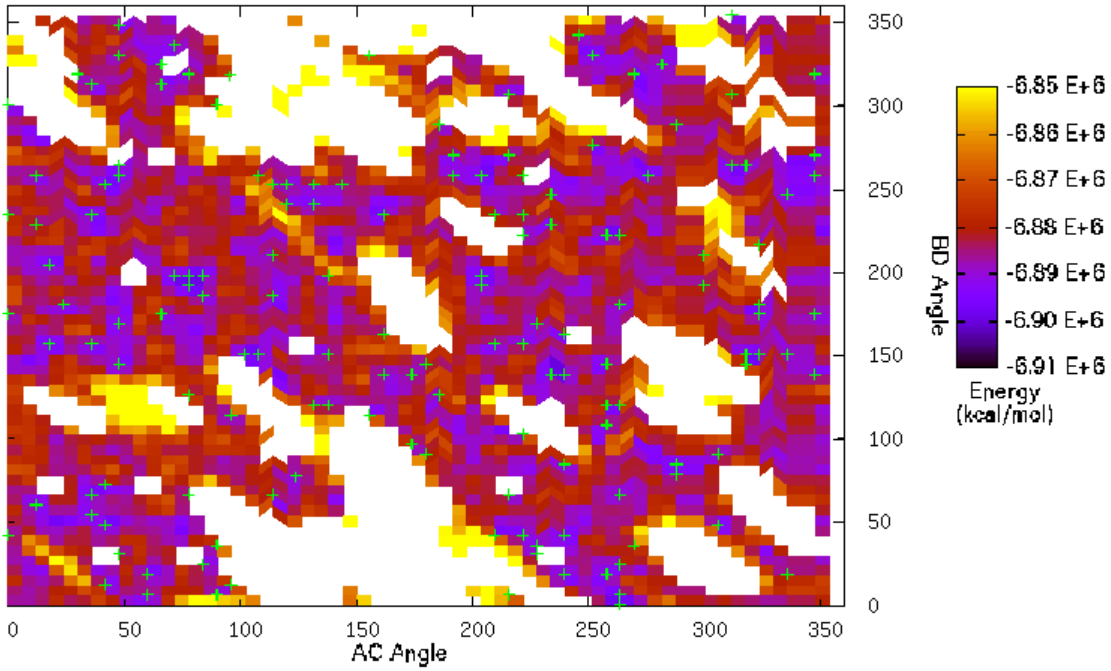


Figure 48: Overlaid image of the 150 lowest energy orientations from the single point energy rotation search, green cross, on the potential energy plot.

The lowest 150 energy orientations have energies in the range of -6899382 to -6909418 kcal/mol.; the values for these orientations can be seen in Figure 47. By overlaying Figure 47 on Figure 46 to produce Figure 48, we can begin to get an idea of the values that may be adopted by the collagen molecules in nature. The majority of the lowest energy orientations lie on dark blue, purple and a few on maroon regions of the energy profile. However, as the plots were generated using a 6° rotation increment, the resolution is not sufficient to be able to distinguish the most optimum orientations, as upon relaxation, a structure may find a more stable confirmation by a small rotation, for example by moving from 0°-0° to 0°-6° there are 2160 integer orientations. However the resolution is sufficient to act as a guide for a more in depth investigation of the areas surrounding the lower energy orientations, hence short MD simulations were conducted according to the protocol outlined in 7.2.3.

7.3.2 Short MD runs

Beginning from the wide range of orientations defined by the lowest 150 structures identified in the single point energy searching, short MD simulations were run to identify both the most abundant orientations in addition to the lowest energy orientations. During thermostated molecular dynamics simulations the free energy of a system tends to a minimum and hence the lower energy states are more probable, although random thermal fluctuations will introduce occasional higher energy states. This has two consequences for our investigation; the first is that a sub-optimal orientation will tend towards a thermodynamic equilibrium, therefore rotating into a state of optimal interaction with the second molecule. The second is that we can use the frequency of

orientations as a measure of the stability of that particular orientation. Therefore, as the simulations proceed, higher energy structures will move towards a lower energy state. This was observed as illustrated by an example for the 240°-84° initial orientation in Figure 49, where we see an change in the orientation observed until a relatively steady state is observed for the AC strand, in this case. The BD strand, in Figure 49, begins at an initial good approximation to the lower energy orientation, hence it remains at steady state. Hence after an initial relaxation period, which we exclude from our results, we are able to monitor the average orientation, to determine the most frequent orientations present.

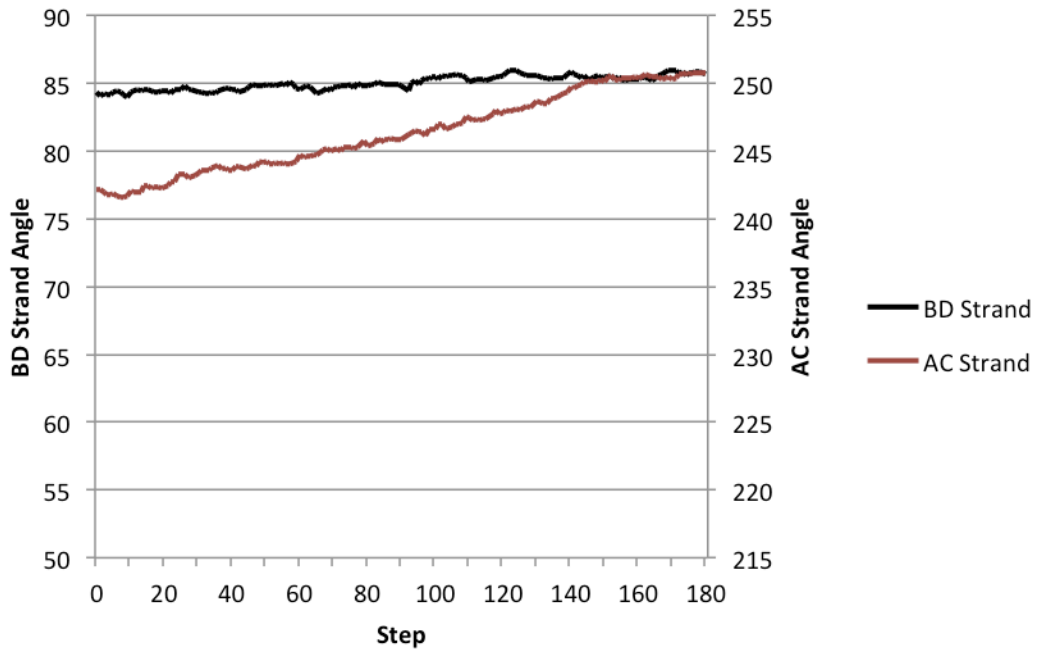


Figure 49: Figure depicting equilibration of the two collagen molecules from the initial 240°-84° orientation, AC strand in red moving to equilibration, and the BD strand in black, beginning at equilibrium.

Through monitoring the frequency of particular orientations within the 100 ps MD simulations, we were able to collate orientation frequency data as shown in graphical form in Figure 50. The data was collated into bins such that only

integer values were used for the remainder of the study; this reduces the number of possible configurations to 129,600 possible discrete orientations, making the data more manageable and reliable within the limits of the calculated ±0.30° standard error of the mean. What can immediately be seen from Figure 50 is that a larger proportion of the possible orientations remain unpopulated. Instead there is a clustering of frequently populated orientations and an almost complete exclusion of states outside of these. Of particular note is the four key exclusion regions from the 340°-0° to 20°-359°, 160°-0° to 200°-359°, 0°-340°- to 359°-20° and 0°-160° to 359°-200°, which leads to a cross shaped region through the plot. This observation supports the idea presented in Figure 49, that despite a large number of potential orientations being identified in the SPE scan, within that region, equilibration of the system results in the molecules rotating into a lower energy orientation, resulting in the frequency of orientations reducing to zero during the MD simulations.

Within these high occurrence orientation clusters observed in Figure 50, we see a small number of very high occurrence orientations, or two very closely related orientations. To gain a better understanding of the favourable orientations that collagen molecules like to adopt, we identified the 30 most frequent orientations, which are presented in Figure 51, along with their accompanying frequency as a percentage of the total number of calculated orientations. Although the frequencies reported look relatively low, with the largest frequency being 0.21%, when you consider that nearly 42,000 orientations were calculated from the MD simulations in which a possible 129,600 orientations are possible the significance of these values becomes apparent.

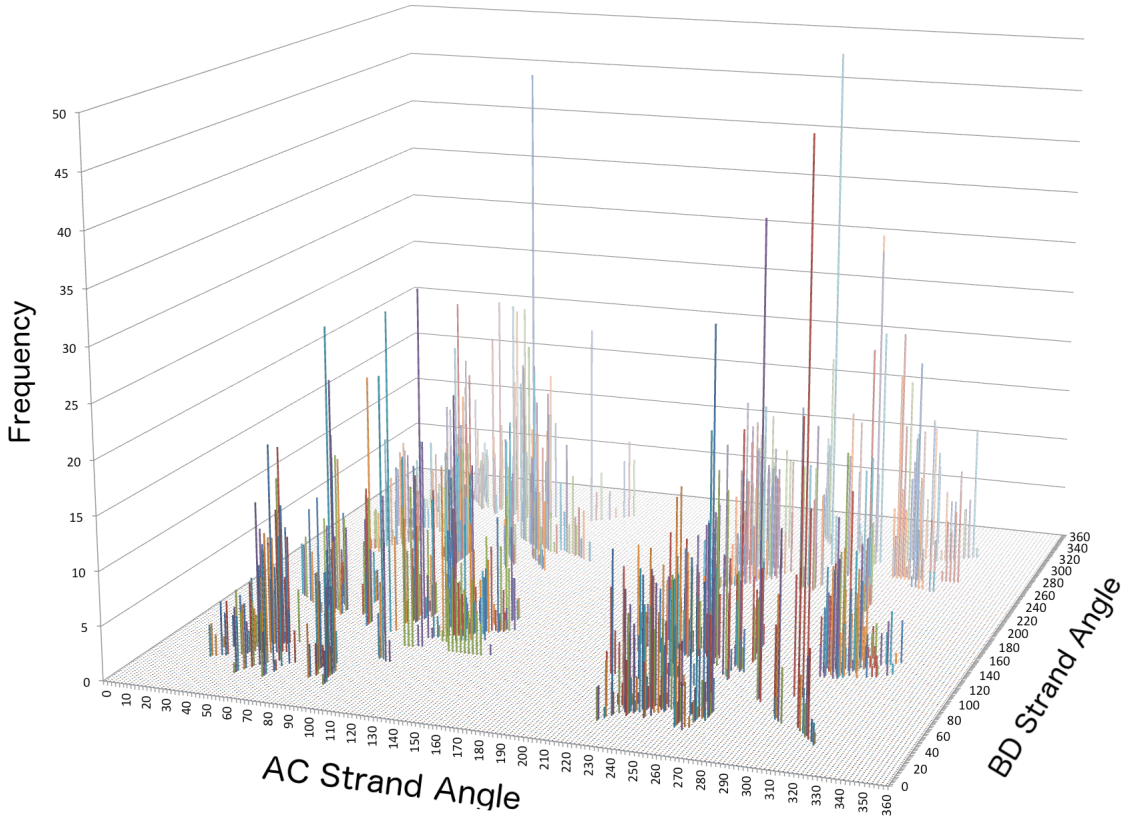


Figure 50: 3D frequency histogram plot of the relative orientations of the two collagen model strands, with angle of the AC strand on the x axis, BD strand angle on the y axis and the frequency of the orientation on the z axis.

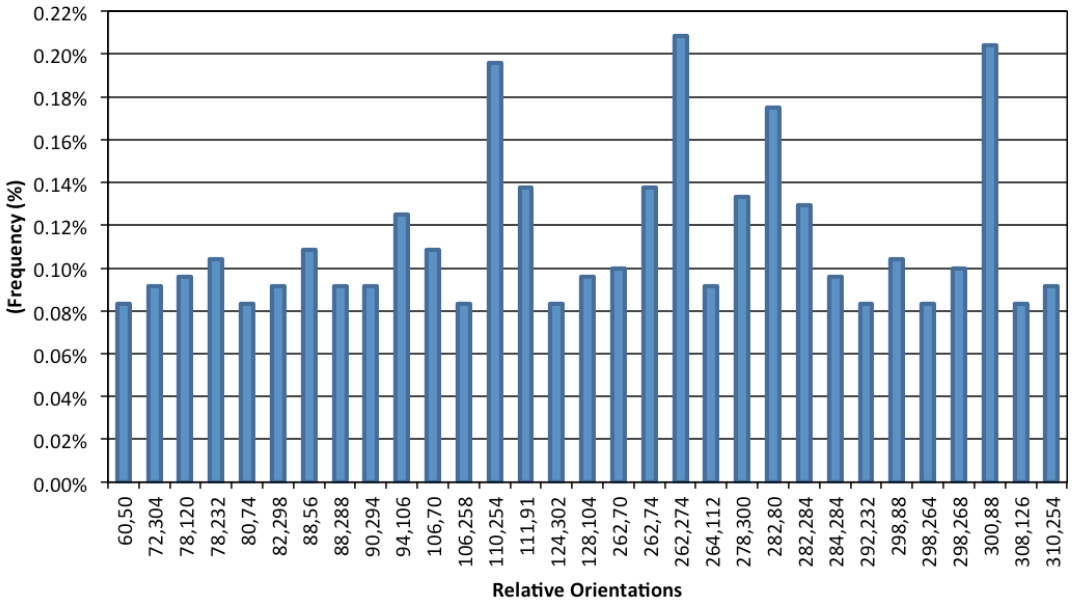


Figure 51: Thirty most frequent orientations identified from the molecular dynamics simulations accompanied with their frequency as a percentage of the total calculated orientations.

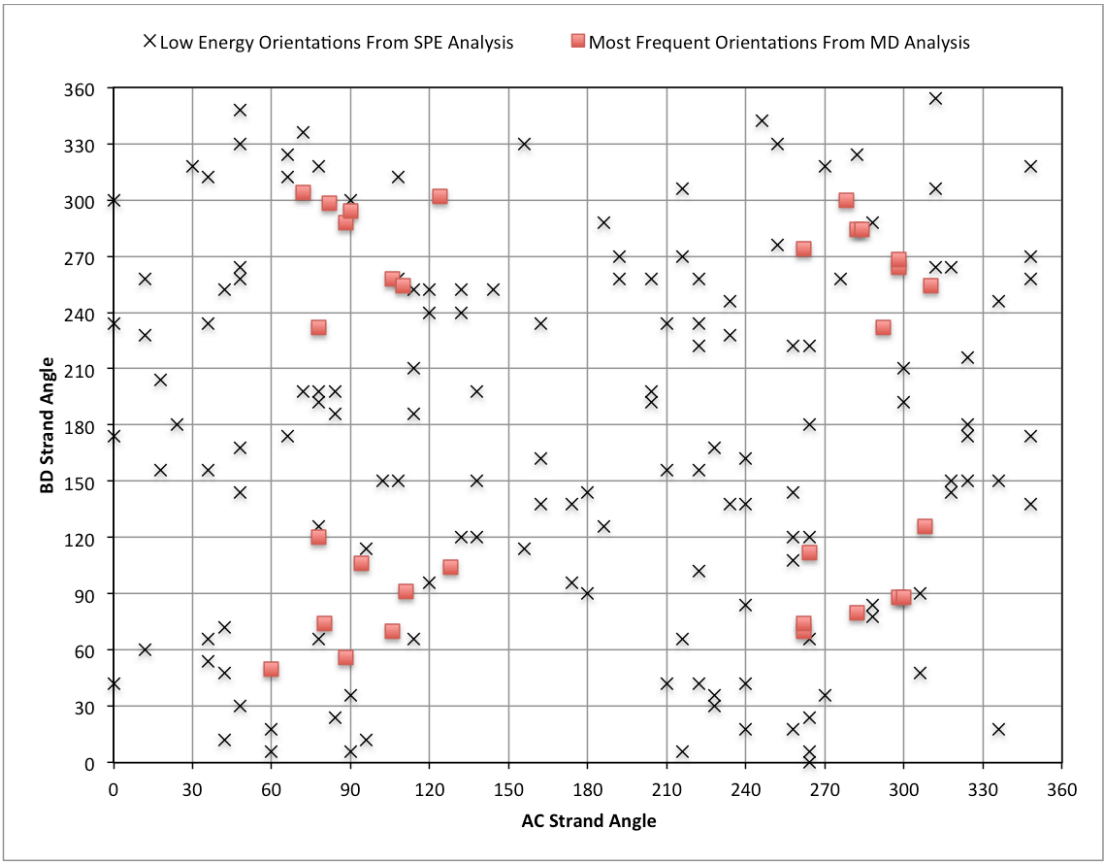


Figure 52: Plot illustrating the angles of the AC and BD collagen strands for the thirty most frequent orientations identified from the molecular dynamics simulations, red squares and the 150 lowest energy orientations determined from the single point energy rotation search

Upon extracting the most frequently occurring 30 orientations, we first wanted to look at the distribution of these orientations. The distribution of the thirty values can be seen in Figure 52, as red squares. Additionally we overlaid the position of the lowest 150 orientations output from the single point energy searches. What is apparent of the MD derived orientations is that they are located in four distinct regions of the orientation plot, with the same four exclusion regions as for the SPE identified orientations, slightly extended. This is of great interest as it indicates that the interaction of either the top and bottom surface (320° – 40° and 130° -240° regions) of the collagen molecule likely results in unfavourable interactions. Another feature of significance is that, although there are large

clusters relatively nearby to the thirty highest frequency orientations, the SPE investigation owing to its reduced resolution would not have found these lowest energy orientations.

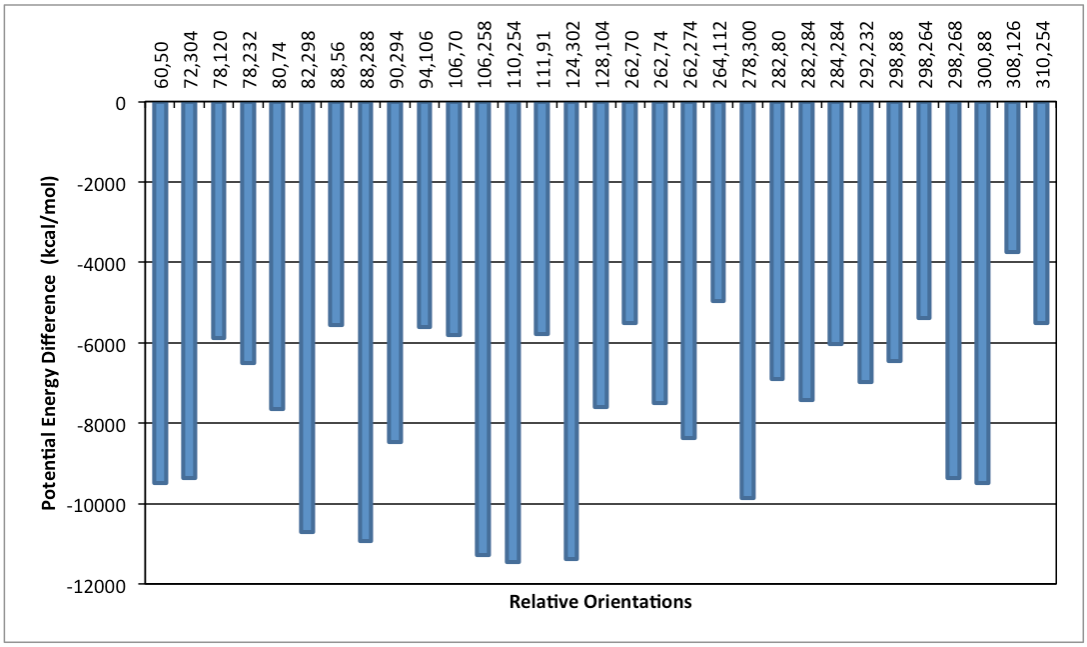


Figure 53: Figure showing the average potential energy difference of the thirty most frequent orientations identified from the molecular dynamics simulations, relative to the average potential energy of the 0°-0° orientation system.

The final stage in identifying the preferential orientations of the collagen molecule for packing in a microfibril is to investigate its effect on the energetics of the system. To do this we calculated the potential energy for the thirty most frequently occurring orientations, for the duration of the 100 ps simulations relative to the 0°- 0° model energy, the results of which are presented in Figure 53. What is immediately apparent is that all thirty of the most frequent orientations have lower energies than those calculated for the 0°-0° interaction model. Within these 30 frequent orientations we have three orientations; 106°-258°; 110°-254°; 124°-302°. These have values 50% lower than the average values for the other 27 orientations, making them the optimum interaction surfaces. It is seen that the orientations with higher frequencies tend to have

lower energy, with a couple of exceptions. This is because a neighbouring orientation is also abundant, for example the 298°-268° orientations reports a very low energy but a relatively small frequency. However the frequency is low, owing to the 298°-264° orientation also exhibiting a high frequency, allowing fluctuations between the two structures. Considering this data, we can use the clusters of the thirty highest frequency sites to guide the likely most favourable interactions sites. If we transpose the clusters, consisting of the 30 most frequent orientations, onto a representative collagen molecule we see a bow shape of “favourable” values in Figure 54a. If we then do the same for the 15 lowest energy orientations we see a narrowing of the left hand bow as seen Figure 54b. The red-regions being unfavourable orientations for interaction with any orientation of the neighbouring collagen molecules.

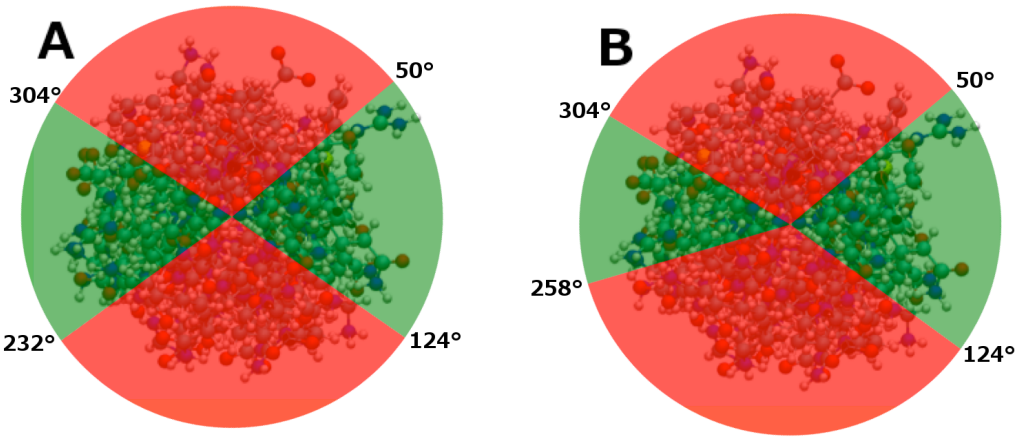


Figure 54: Figures showing the calculated favourable interaction regions shown in green and unfavourable shown in red, based on A) frequency data and B) energetics data.

To obtain the lowest energy packing of the collagen molecules, the system needs to consist of interactions between the collagen molecules based on the optimum orientations. As previously mentioned in 1.1.2.1 collagen molecules within a microfibril can pack in a quasi-hexagonal manner (20) as was seen in

Figure 1. The current implementation in collagen models, taken from the crystal structure for the *Rattus norvegicus* sequence (310), for the packing, results in 12 interactions within the hexagonal arrangement; one 90°-70°, three 270°-90°, four 150°-330° and four 30°-210°. Four of these lie within the high frequency cluster range, whilst the remaining eight lie outside of this area. To obtain the optimum packing for our *Homo sapiens* sequence, rotation of the explicit collagen molecule needs to be conducted, so that the number of favourable interactions is optimised. Owing to the periodic nature of the arrangement of collagen molecules within the fibril, the rotation can only be applied to one molecule so that the periodicity is conserved. Rotation of the explicit collagen by 26° in a clockwise direction results in a 50% reduction in the number of unfavourable orientation interactions and a subsequent 100% increase in the favourable orientation interactions. The angles of interaction now being one 64°-244°, three 244°-64°; four 124°-304°; and four 4°-184° interactions within the hexagonal unit after the 26° rotation of the collagen molecules. The impact of such a rotation can be seen in Figure 55, by a doubling in the number of favourable (green) interactions, within the hexagonal close packed unit. In addition to increasing the number of favourable interactions by this 26° rotation, the rotation also had the added effect that the second lowest energy orientation could be adopted 124°-304°, thus having a significant stabilising effect on the fibril.

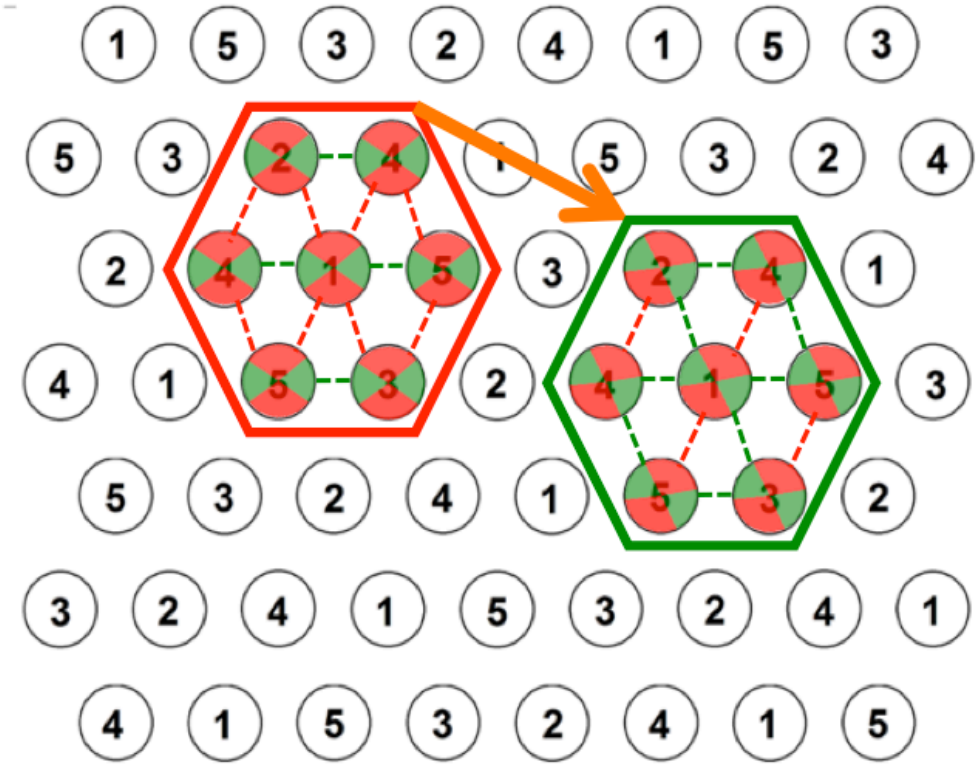


Figure 55: Image depicting the impact of a 26° clockwise rotation of the collagen molecules within a hexagonal closed packed unit. The red hexagonal unit showing the orientation present within the Rattus norvegicus unit cell and the green hexagonal unit showing the configuration after the clockwise rotation. Green and red areas on the collagen molecules illustrating the favourable and unfavourable interaction surfaces respectively, as previously described in Figure 54, with the dashed line similarly coloured showing the interaction orientation of the collagen molecules.

7.4 Implications

As was seen in 7.3, the current model used to simulate collagenous systems may not be the most accurate, given that we have shown that a mere 26° clockwise rotation of the explicit collagen molecule will result in a reduction in the energy. However there are a number of other factors which will also dictate the possible orientations within the fibril. For example, the mature enzymatic cross-links will likely reduce the number of possible orientations further. Taking this into account, what we are actually showing from this current study is the

possible orientations of the collagen molecules within a collagen gel or artificial construct, in the absence of mature cross-links. Further studies would be required, using a three strand model with explicit cross-links, to see how this would influence the possible orientations. This is beyond the capabilities of current computational resources. In addition, collagenous tissues are rarely homogeneous. Even tendons which consist of 65-80% (dry weight) collagen, which is predominately type I collagen, contain other ECM proteins such as 2% elastin which may alter the alignment of the collagen molecules (22, 311). Interactions of the type I collagen molecules with these other proteins will be different to those of the type I collagen - type I collagen interactions. However, if we consider the widespread used of collagen gels as tissue engineering scaffolds, then we can confidently say that the orientations exhibited in these samples, which have much lower concentrations of other proteins and absence of enzymatic cross-links, are likely to be those identified in this study.

In conducting this study two assumptions were made about the model used. The first was that the collagen can be described as a straight rod. The second is that the rotation will occur uniformly over the whole length of the collagen. The assumption that collagen molecules is a straight rod is valid when considering the time point being probed here, which is shortly after secretion into the ECM and processing from the procollagen to collagen; as the collagen molecules are relatively straight until shortly after fibrillogenesis has occurred. Once this stage has past, crimping of the collagen molecule will occur, with “waves” in the molecule being introduced by translation into the lower protein density gap region. Once this has been done then the orientation of the system is fixed. Hence if MD simulation is conducted using the straight molecule then the optimum orientations of the molecule will be encountered given sufficient time.

However the crimping process occurs on longer time scales and hence it would be more beneficial for simulations to begin from the crimped structure to probe physiologically relevant phenomena. Hence, for this purpose, the simulations should begin from the optimum orientation.

Also the second Perl script, that conducts the vector based mathematics to yield the orientation values from the MD trajectory of the model, calculates the angular displacement of the backbone atom's position, relative to their respective position in the 0°-0° reference model, for each outputted time-step. In doing so it averages the orientation of each constituent atom, to report a single value for the orientation of the entire collagen molecule. However it is possible that a portion of the collagen molecule may rotate to a greater extent than the rest of the molecule, resulting in an inaccurate value being recorded for the orientation. To overcome this issue we verified using a random selection (N=10) of the orientations from the short MD runs, to calculate the standard error of the mean for the molecular orientation from its constituent atomic angular displacements. It was found that the standard error of the mean had an average value of ±0.30°, and therefore twisting of the molecule was not occurring to a great enough extent so as to influence the molecular orientation values reported.

7.5 Summary

The orientation of collagen molecules packed within a collagen fibril has significant implications on the fibrils mechanical and biological properties, as well as the accessible surface that will be capable of undergoing AGE cross-linking. Through the use of a comprehensive single point energy scan of the

potential energy landscape, based on 6° rotation increment, we identified the 150 lowest energy orientations to undergo a more comprehensive study. The 150 orientations were then used as inputs for short MD simulations in which the molecules rotated towards a low energy state. There were four distinct clusters of highly frequent structures, with the 30 most frequent within these being used to predict the likely favourable interaction orientations. This was confirmed by comparing the average potential energy for these orientations relative to the 0°-0° model and finding all of them had much lower energies. Using both the favourable orientation angles from the frequency and energetics data, we identified that the current model of 0° is not the most stable based, on inter-molecular interactions. A 26° clockwise rotation, of the collagen molecule in the current model used (or crystal structure for *Rattus norvegicus* (310)), would half the number of unfavourable orientation angles and increase the number of favourable interactions by 100%. However this study did exclude some factors, such as tissue heterogeneity and enzymatic cross-link, which might result in a different result being obtained, although further study would be required to confirm this. However the 26° rotation would hold true for collagen tissue scaffolds used in tissue engineering applications, which may allow a better understanding of host-scaffold interaction and processes to be obtained.

Chapter 8 Conclusion

8.1 Main Conclusions

In Chapter 1 we gave a general introduction to all aspects of collagen, from biosynthesis to its mechanical properties, before introducing advanced glycation end products, their structure, their impact and their treatments. This was followed by a review of the computational techniques employed throughout this work, from molecular dynamics to Hartree-Fock electronic structure methods.

Our modelling approach began with the development of reliable force fields to describe the AGEs of interest, for implementation in the wider AMBER ff99SB force field. Derivation was conducted using frequency calculations at HF/6-31G* level theory using Gaussian09 (220) and R.E.D. server, for RESP charge fitting (216). The terms developed were shown to be comparable to a later purely quantum mechanically derived force field (227), supporting the validity of results obtained by its use.

In Chapter 4, we used these newly derived force field terms to identify preferential glucosepane and DOGDIC formation sites within the collagen molecule. A fully atomistic approach exploiting the D-banding periodicity was used to replicate the fibrillar environment of the collagen molecule under pseudo-physiological conditions. Six sites were found to be energetically favourable for the formation of each of the AGEs, with only one of these sites being equivalent. The formation of the AGEs at different sites was thought to be due to the reduced nitrogen atom separation in the different types of cross-link. We also showed a strong preference for the AGE cross-links to form within the gap region of the protein, owing to its reduced protein density. Finally by overlaying these preferential formation sites on known collagen-biomolecule

binding sites, we have identified a number of overlapping sites, which may impede the biological function of glycated collagenous tissues.

Chapter 5 consists of two related steered molecular dynamics studies into the mechanical properties of collagen and the effect of collagen cross-linking on these. The first part of the study was to probe the heterogeneous response of a collagen molecule to an applied load, to identify if this is a sequence dependent consequence. We varied a single residue at the Yyy position of a homotrimeric collagen peptide $(ProHypGly)_4YyyProGly(ProHypGly)_4$ and measured the resulting change in the Young's modulus of the molecule. The values obtained by changing just a single residue varied by as much as 6.6%. A relationship between stability and elasticity was also observed: peptides with higher melting temperatures reported lower values for the Young's modulus. The second stage was to probe the impact on the mechanical properties of the collagen molecule due to the presence of an AGE cross-link. Using relative differences for the cross-linked model compared to a wild type model it was possible to mitigate the heterogeneous response, and to probe both the lateral and tensile moduli. It was observed that the presence of these favourable glucosepane cross-links and DOGDIC cross-links had no significant impact, within the uncertainty of the techniques, on the values of the tensile and lateral Young's modulus. Even when all favourable cross-links were inserted into a full-length collagen molecule and the tensile modulus was calculated, again no significant difference was reported.

So far, owing to the size of the collagen molecule, only the *Rattus norvegicus* structure has been crystallised and this is the sequence used in chapter four and five. However, for disease modelling and pathology studies the human

sequence is of greatest interest. It is for this reason that in Chapter 6, we derived a homology model of the *Homo sapiens* sequence of type I collagen. After using the blastP suite to score potential template sequences, we identified the *Rattus norvegicus* to be the highest scoring template sequence for which a crystal structure was available. After undergoing a standard homology modelling protocol we decided to consider whether the D-banding periodicity would increase as a result of the different primary sequence. Upon completion of a comparison of 7 alternative cell dimensions, it was found that no significant change (+0.03% change in a and b dimensions) occurred to the packing pattern of the *Homo sapiens* sequence. Through monitoring a number of system observables during two short (2 ns) MD simulations, one using our homology model and one of the well-established model for *Rattus norvegicus*, it was found that they remained within the same range, thus proving the stability of our homology model.

The orientation of collagen molecules packed within a collagen fibril potentially has huge implications on the fibril's mechanical and biological properties. However their determination remains below the resolution of current experimental techniques. In Chapter 7, we use a single point energy scan of 6° rotation increments of two staggered collagen strands (a full collagen molecule, a gap region and a short collagen peptide) to identify the lower energy interaction surfaces. The lowest energy orientations identified from this single point energy scan are then used as starting configurations for short MD simulations. The frequency of orientations and energies are computed over these MD simulations to determine the most favourable orientations. A clustering of low energy and high-frequency orientations were observed, such that the interactions were optimum within two small windows of orientation,

between 50° - 124° and 232° - 304°, with respect to the orientation given of the collagen molecule in the Orgel crystal structure (21). Given the hexagonal close packing of collagen molecules, we identified that a 26° clockwise rotation of the explicit collagen molecule in current models would result in an increased number of favourable interactions, including the second lowest energy orientation, and produce the likely orientation of collagen molecules within a human fibril.

8.2 Limitations

Owing to the large size of a collagen molecule, over 3000 amino acids residues, balancing the size of the system modelled and computational resources has been a challenge. However, through the use of periodic boundary conditions, this impact has been minimised. Despite this over 270 CPU years have been utilised to produce the results contained within this thesis. One potential study in which alteration of the model to reduce computational expense may have influenced the results, is in the mechanical properties studies, where fully solvated models of short collagen peptides were used instead of a fibrillar environment. Therefore further study is required to ascertain whether the environment has a significant effect on the results reported.

Our modelling approach throughout adopts a homotypic model of type I collagen. Typically, healthy tissue, however, is heterotypic, with tendon fibrils, for example containing other minor collagen types and elastin. Although, as our investigations have predominantly focussed on intra-molecular interactions, the composition should have little effect if any on the results reported. The one study which did consider inter-molecular interactions was the orientations

investigation in Chapter 7, at which point we were probing the inter-molecular interactions between the type I collagen molecules so the heterogeneity will have no influence.

8.3 Future Work

Having developed a reliable homology model for the *Homo sapiens* sequence, it would be of interest to verify that the favourable glucosepane and DOGDIC cross-links form at the same preferential sites as they had for the *Rattus norvegicus* structure. After verifying the intra-molecular positions it would be extremely interesting to identify the location of favourable inter-molecular sites between two *Homo sapiens* collagen molecules.

Another key topic we would like to expand upon is to probe the mechanical influence of varying a single residue within a wide variety of different triplets. Initially, we would like to extend the triplet mechanical properties library to include the GlyXxxHyp triplets, thus extending the known influences to over 50% of the triplets present within collagen, making possible a quantitative approach to predict the mechanical properties from the biochemical sequence data.

The last topic we would like to explore further is the key biomolecule collagen interactions disrupted by the presence of AGEs cross-links, as identified in Chapter 4. To do this we would like to conduct explicit modelling of the biomolecule–collagen interaction, with and without AGE cross-links present to determine the thermodynamic effect of their presence and the mechanism of interference.

Chapter 9 Bibliography

1. Kadler, K. E., Baldock, C., Bella, J., and Boot-Handford, R. P. (2007) Collagens at a glance. *J. Cell Sci.* **120**, 1955–1958

2. Gelse, K. (2003) Collagens—structure, function, and biosynthesis. *Adv. Drug Deliv. Rev.* **55**, 1531–1546

3. Canty, E. G., and Kadler, K. E. (2005) Procollagen trafficking, processing and fibrillogenesis. *J. Cell Sci.* **118**, 1341–1353

4. Gelman, R. A., Williams, B. R., and Piez, K. A. (1979) Collagen fibril formation. Evidence for a multistep process. *J. Biol. Chem.* **254**, 180–186

5. Leslie, M. (2006) Making tendons. *J. Cell Biol.* **172**, 167

6. Reiser, K., McCormick, R. J., and Rucker, R. B. (1992) Enzymatic and nonenzymatic cross-linking of collagen and elastin. *FASEB J.* **6**, 2439–2449

7. Kivirikko, K. I., Ryhanen, L., Anttinen, H., Bornstein, P., and Prockop, D. J. (1973) Hydroxylation of lysyl residues in collagen by protocollagen lysyl hydroxylase in vitro. *Biochemistry*. **12**, 4966–4971

8. Lang, K., Schmid, F. X., and Fischer, G. (1987) Catalysis of protein folding by prolyl isomerase. *Nature*. **329**, 268–270

9. Nagata, K. (1998) Expression and function of heat shock protein 47: A collagen-specific molecular chaperone in the endoplasmic reticulum. *Matrix Biol.* **16**, 379–386

10. Prockop, D. J., Sieron, A. L., and Li, S.-W. (1998) Procollagen N-proteinase and procollagen C-proteinase. Two unusual metalloproteinases that are essential for procollagen processing probably have important roles in development and cell signaling. *Matrix Biol.* **16**, 399–408

11. Leighton, M., and Kadler, K. E. (2003) Paired basic/furin-like proprotein convertase cleavage of pro-BMP-1 in the trans-golgi network. *J. Biol. Chem.* **278**, 18478–18484

12. Parry, D. A. D., and Craig, A. S. (1979) Electron microscope evidence for an 80 A unit in collagen fibrils. *Nature*. **282**, 213–215

13. Scott, J. E. (1996) Proteodermatan and proteokeratan sulfate (decorin, lumican/fibromodulin) proteins are horseshoe shaped. Implications for their interactions with collagen. *Biochemistry*. **35**, 8795–8799

14. Birk, D. E., Nurminskaya, M. V, and Zycband, E. I. (1995) Collagen fibrillogenesis in situ: fibril segments undergo post-depositional modifications resulting in linear and lateral growth during matrix development. *Dev. Dyn.* **202**, 229–243

15. Ottani, V., Raspanti, M., and Ruggeri, A. (2001) Collagen structure and functional implications. *Micron*. **32**, 251–260

16. Bhattacharjee, A., and Bansal, M. (2005) Collagen structure: the Madras triple helix and the current scenario. *IUBMB Life*. **57**, 161–172

17. Shoulders, M. D., and Raines, R. T. (2009) Collagen structure and stability. *Annu. Rev. Biochem.* **78**, 929–958

18. Brodsky, B., and Persikov, A. V (2005) Molecular structure of the collagen triple helix. *Adv. Protein Chem.* **70**, 301–339

19. Petruska, J. A., and Hodge, A. J. (1964) A subunit model for the tropocollagen macromolecule. *Proc. Natl. Acad. Sci.* **51**, 871–876

20. Hulmes, D. J. S., and Miller, A. (1979) Quasi-hexagonal molecular packing in collagen fibrils. *Nature*. **282**, 878–880

21. Orgel, J. P. R. O., Irving, T. C., Miller, A., and Wess, T. J. (2006) Microfibrillar structure of type I collagen in situ. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9001–9005

22. Kannus, P. (2000) Structure of the tendon connective tissue. *Scand. J. Med. Sci. Sports*. **10**, 312–320

23. Moeller, H. D., Bosch, U., and Decker, B. (1995) Collagen fibril diameter distribution in patellar tendon autografts after posterior cruciate ligament reconstruction in sheep: changes over time. *J. Anat.* **187**, 161–167

24. Ramachandran, G. N. (1988) Stereochemistry of collagen. *Int. J. Pept. Protein Res.* **31**, 1–16

25. Ramachandran, G. N., and Sasisekharan, V. (1968) Conformation of Polypeptides and Proteins. in *Advances in Protein Chemistry* (Anfinsen, C. B., Anson, M. L., T., E. J., and Richards, F. M. eds), pp. 283–437, Academic Press, **23**, 283–437

26. Bella, J., and Berman, H. M. (1996) Crystallographic evidence for C alpha-H...O=C hydrogen bonds in a collagen triple helix. *J. Mol. Biol.* **264**, 734–42

27. Kramer, R. Z., and Berman, H. M. (1998) Patterns of Hydration In Crystalline Collagen Peptides. *J. Biomol. Struct. Dyn.* **16**, 367–380

28. Kramer, R. Z., Venugopal, M. G., Bella, J., Mayville, P., Brodsky, B., and Berman, H. M. (2000) Staggered molecular packing in crystals of a collagen-like peptide with a single charged pair. *J. Mol. Biol.* **301**, 1191–205

29. Persikov, A. V, Ramshaw, J. A. M., Kirkpatrick, A., and Brodsky, B. (2002) Peptide investigations of pairwise interactions in the collagen triple-helix. *J. Mol. Biol.* **316**, 385–94

30. Persikov, A. V, Ramshaw, J. A. M., and Brodsky, B. (2005) Prediction of

collagen stability from amino acid sequence. *J. Biol. Chem.* **280**, 19343–19349

31. Persikov, A. V, and Brodsky, B. (2002) Unstable molecules form stable tissues. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 1101–3

32. Bella, J., Eaton, M., Brodsky, B., and Berman, H. M. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 A resolution. *Science*. **266**, 75–81

33. Leikin, S., Rau, D. C., and Parsegian, V. A. (1995) Temperature-favoured assembly of collagen is driven by hydrophilic not hydrophobic interactions. *Nat. Struct. Biol.* **2**, 205–210

34. Kuznetsova, N., Rau, D. C., Parsegian, V. A., and Leikin, S. (1997) Solvent hydrogen-bond network in protein self-assembly: solvation of collagen triple helices in nonaqueous solvents. *Biophys. J.* **72**, 353–62

35. Streeter, I., and de Leeuw, N. H. (2011) A molecular dynamics study of the interprotein interactions in collagen fibrils. *Soft Matter*. **7**, 3373–3382

36. Avery, N. C., and Bailey, A. J. (2005) Enzymic and non-enzymic cross-linking mechanisms in relation to turnover of collagen: relevance to aging and exercise. *Scand. J. Med. Sci. Sports*. **15**, 231–40

37. Jokinen, J., Dadu, E., Nykvist, P., Käpylä, J., White, D. J., Ivaska, J., Vehviläinen, P., Reunanen, H., Larjava, H., Häkkinen, L., and Heino, J. (2004) Integrin-mediated cell adhesion to type I collagen fibrils. *J. Biol. Chem.* **279**, 31956–63

38. Heidemann, E., and Roth, W. (1982) Synthesis and investigation of collagen model peptides. in *Advances in Polymer Science*, pp. 143–203, Advances in Polymer Science, Springer Berlin Heidelberg, **43**, 143–203

39. Rauch, F., and Glorieux, F. H. (2004) Osteogenesis imperfecta. *Lancet*. **363**, 1377–1385

40. Persikov, A. V, Ramshaw, J. A. M., Kirkpatrick, A., and Brodsky, B. (2000) Amino acid propensities for the collagen triple-helix. *Biochemistry*. **39**, 14960–14967

41. DeRider, M. L., Wilkens, S. J., Waddell, M. J., Bretscher, L. E., Weinhold, F., Raines, R. T., and Markley, J. L. (2002) Collagen stability: Insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations. *J. Am. Chem. Soc.* **124**, 2497–2505

42. Jenkins, C. L., and Raines, R. T. (2002) Insights on the conformational stability of collagen. *Nat. Prod. Rep.* **19**, 49–59

43. Kersteen, E. A., and Raines, R. T. (2001) Contribution of tertiary amides to the conformational stability of collagen triple helices. *Biopolymers*. **59**, 24–8

44. Kivirikko, K. I., Myllylä, R., and Pihlajaniemi, T. (1989) Protein hydroxylation: prolyl 4-hydroxylase, an enzyme with four cosubstrates and a multifunctional subunit. *FASEB J.* **3**, 1609–1617

45. Improta, R., Mele, F., Crescenzi, O., Benzi, C., and Barone, V. (2002) Understanding the role of stereoelectronic effects in determining collagen stability. 2. A quantum mechanical/molecular mechanical study of (Proline-Proline-Glycine)(n) polypeptides. *J. Am. Chem. Soc.* **124**, 7857–65

46. Improta, R., Benzi, C., and Barone, V. (2001) Understanding the role of stereoelectronic effects in determining collagen stability. 1. A quantum mechanical study of proline, hydroxyproline, and fluoroproline dipeptide analogues in aqueous solution. *J. Am. Chem. Soc.* **123**, 12568–77

47. Berg, R. A., and Prockop, D. J. (1973) The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen. *Biochem. Biophys. Res. Commun.* **52**, 115–120

48. Jiravanichanun, N., Nishino, N., Okuyama, K., and Biro, A. (2006) Conformation of alloHyp in the Y position in the host–guest peptide with the pro–pro–gly sequence: Implication of the destabilization of (Pro–alloHyp–Gly)10. *Biopolymers*. **81**, 225–233

49. Inouye, K., Kobayashi, Y., Kyogoku, Y., Kishida, Y., Sakakibara, S., and Prockop, D. J. (1982) Synthesis and physical properties of (hydroxyproline-proline-glycine)10: Hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix. *Arch. Biochem. Biophys.* **219**, 198–203

50. Ramshaw, J. A., Shah, N. K., and Brodsky, B. (1998) Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides. *J. Struct. Biol.* **122**, 86–91

51. Yang, W., Chan, V. C., Kirkpatrick, A., Ramshaw, J. A. M., and Brodsky, B. (1997) Gly-Pro-Arg confers stability similar to Gly-Pro-Hyp in the collagen triple-helix of host-guest peptides. *J. Biol. Chem.* **272**, 28837–28840

52. Persikov, A. V, Ramshaw, J. A. M., Kirkpatrick, A., and Brodsky, B. (2005) Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability. *Biochemistry*. **44**, 1414–22

53. Nuytinck, L., Freund, M., Lagae, L., Pierard, G. E., Hermanns-Le, T., and De Paepe, A. (2000) Classical Ehlers-Danlos syndrome caused by a mutation in type I collagen. *Am. J. Hum. Genet.* **66**, 1398–402

54. Guruvayoorappan, C., and Kuttan, G. (2008) Anti-metastatic effect of Biophytum sensitivum is exerted through its cytokine and immunomodulatory activity and its regulatory effect on the activation and nuclear translocation of transcription factors in B16F-10 melanoma cells. *J. Exp. Ther. Oncol.* **7**, 49–63

55. Sunila, E. S., and Kuttan, G. (2006) A preliminary study on antimetastatic activity of Thuja occidentalis L. in mice model. *Immunopharmacol. Immunotoxicol.* **28**, 269–80

56. Robins, S. P., and Duncan, A. (1987) Pyridinium crosslinks of bone collagen and their location in peptides isolated from rat femur. *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* **914**, 233–239

57. Bailey, A. J., Paul, R. G., and Knott, L. (1998) Mechanisms of maturation and ageing of collagen. *Mech. Ageing Dev.* **106**, 1–56

58. Knott, L., and Bailey, A. J. J. (1998) Collagen cross-links in mineralizing tissues: a review of their chemistry, function, and clinical relevance. *Bone*. **22**, 181–7

59. Silver, F. H., Freeman, J. W., and Seehra, G. P. (2003) Collagen self-assembly and the development of tendon mechanical properties. *J. Biomech.* **36**, 1529–1553

60. Dixon, S. J., and Kerwin, D. G. (2002) Variations in Achilles tendon loading with heel lift intervention in heel-toe runners. *J. Appl. Biomech.* **18**, 321–331

61. Lichtwark, G. a, Cresswell, A. G., and Newsham-West, R. J. (2013) Effects of running on human Achilles tendon length-tension properties in the free and gastrocnemius components. *J. Exp. Biol.* **216**, 4388–94

62. Gautieri, A., Buehler, M. J., and Redaelli, A. (2009) Deformation rate controls elasticity and unfolding pathway of single tropocollagen molecules. *J. Mech. Behav. Biomed. Mater.* **2**, 130–7

63. Bhowmik, R., Katti, K. S., and Katti, D. R. (2007) Mechanics of molecular collagen is influenced by hydroxyapatite in natural bone. *J. Mater. Sci.* **42**, 8795–8803

64. Buehler, M. J. (2006) Atomistic and continuum modeling of mechanical properties of collagen : Elasticity , fracture , and self-assembly. *J. Mater. Res.* **21**, 1947–1961

65. Bozec, L., and Horton, M. (2005) Topography and mechanical properties of single molecules of type I collagen using atomic force microscopy. *Biophys. J.* **88**, 4223–31

66. Sun, Y. L., Luo, Z. P., Fertala, A., and An, K. N. (2002) Direct quantification of the flexibility of type I collagen monomer. *Biochem. Biophys. Res. Commun.* **295**, 382–386

67. Eppell, S. J., Smith, B. N., Kahn, H., and Ballarini, R. (2006) Nano measurements with micro-devices: mechanical properties of hydrated collagen fibrils. *J. R. Soc. Interface*. **3**, 117–121

68. Gautieri, A., Vesentini, S., Redaelli, A., and Buehler, M. J. (2011) Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. *Nano Lett.* **11**, 757–766

69. Buehler, M. J. (2008) Nanomechanics of collagen fibrils under varying cross-link densities: atomistic and continuum studies. *J. Mech. Behav. Biomed. Mater.* **1**, 59–67

70. Van Der Rijt, J. A. J., Van Der Werf, K. O., Bennink, M. L., Dijkstra, P. J., and Feijen, J. (2006) Micromechanical testing of individual collagen fibrils. *Macromol. Biosci.* **6**, 699–702

71. Wenger, M. P. E., Bozec, L., Horton, M. A., and Mesquida, P. (2007) Mechanical properties of collagen fibrils. *Biophys. J.* **93**, 1255–63

72. Sherman, V. R., Yang, W., and Meyers, M. A. (2015) The materials science of collagen. *J. Mech. Behav. Biomed. Mater.* **52**, 22–50

73. Parry, D. A. D. (1988) The molecular fibrillar structure of collagen and its relationship to the mechanical properties of connective tissue. *Biophys. Chem.* **29**, 195–209

74. Ramachandran, G. N., and Kartha, G. (1954) Structure of collagen. *Nature*. **176**, 593–595

75. Tumanyan, V. G., and Iaxyan, T. (1970) Investigation of fibrous structures. I. Computations for collagen. *Biopolymers*. **9**, 955–963

76. Hobza, P., Hurych, J., and Zahradník, R. (1973) Quantum chemical study of the mechanism of collagen proline hydroxylation. *Biochim. Biophys. Acta - Gen. Subj.* **304**, 466–472

77. Gautieri, A., Russo, A., Vesentini, S., Redaelli, A., and Buehler, M. J. (2010) Coarse-grained model of collagen molecules using an extended MARTINI force field. *J. Chem. Theory Comput.* **6**, 1210–1218

78. Streeter, I., and de Leeuw, N. H. (2010) Atomistic modeling of collagen proteins in their fibrillar environment. *J. Phys. Chem. B*. **114**, 13263–13270

79. Stultz, C. M. (2006) The folding mechanism of collagen-like model peptides explored through detailed molecular simulations. *Protein Sci.* **15**, 2166–77

80. Madhan, B., Subramanian, V., Ramasami, T., and Parthasarathi, R. (2003) Ab initio and density functional theory based studies on collagen triplets. *Theor. Chem. Acc.* **110**, 19–27

81. Yang, Z. R. (2009) Predict collagen hydroxyproline sites using support vector machines. *J. Comput. Biol.* **16**, 691–702

82. Bolboaca, S. D., and Jäntschi, L. (2008) A structural informatics study on collagen. *Chem. Biol. Drug Des.* **71**, 173–179

83. Zitzewitz, J. A., Bilsel, O., Luo, J., Jones, B. E., and Matthews, C. R. (1995) Probing the folding mechanism of a leucine zipper peptide by stopped-flow circular dichroism spectroscopy. *Biochemistry*. **34**, 12812–12819

84. Wendt, H., Berger, C., Baici, A., Thomas, R. M., and Bosshard, H. R. (1995) Kinetics of folding of leucine zipper domains. *Biochemistry*. **34**, 4097–4107

85. Baum, J., and Brodsky, B. (1997) Real-time NMR investigations of triple-helix folding and collagen folding diseases. *Fold. Des.* **2**, R53-60

86. Harvey, S. C., and Gabb, H. A. (1993) Conformational transitions using molecular dynamics with minimum biasing. *Biopolymers*. **33**, 1167–1172

87. Marchi, M., and Ballone, P. (1999) Adiabatic bias molecular dynamics: A method to navigate the conformational space of complex molecular systems. *J. Chem. Phys.* **110**, 3697–3702

88. Paci, E., and Karplus, M. (1999) Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J. Mol. Biol.* **288**, 441–459

89. Boudko, S., Frank, S., Kammerer, R. R. A., Stetefeld, J., Schulthess, T., Landwehr, R., Lustig, A., Bächinger, H. P., and Engel, J. (2002) Nucleation and propagation of the collagen triple helix in single-chain and trimerized peptides: transition from third to first order kinetics. *J. Mol. Biol.* **317**, 459–470

90. Wang, Y., Noid, W. G., Liu, P., and Voth, G. A. (2009) Effective force coarse-graining. *Phys. Chem. Chem. Phys.* **11**, 2002–2015

91. Periole, X., and Marrink, S.-J. (2013) The Martini Coarse-Grained Force Field. in *Biomolecular Simulations: Methods and Protocols* (Monticelli, L., and Salonen, E. eds), pp. 533–565, Methods in Molecular Biology, Humana Press, Totowa, NJ, **924**, 533–565

92. Ayton, G. S., Noid, W. G., and Voth, G. A. (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **17**, 192–198

93. Noid, W. G., Chu, J.-W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A., and Andersen, H. C. (2008) The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244111–244114

94. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., and de Vries, A. H. (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B*. **111**, 7812–24

95. Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. (2008) The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **4**, 819–834

96. Mlyniec, A., Mazur, L., Tomaszewski A, K., and Uhl, T. (2015) Viscoelasticity and failure of collagen nanofibrils: 3D Coarse-Grained simulation studies. *Soft Mater.* **13**, 47–58

97. Needleman, S. B., and Wunsch, C. D. (1970) A general method

applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453

98. Dayhoff, M. O. O. (1978) Observed frequencies of amino acid replacements between closely related proteins. *Atlas Protein Seq. Struct.* **5**, Supl. 3

99. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410

100. Sell, D. R., Biemel, K. M., Reihl, O., Lederer, M. O., Strauch, C. M., and Monnier, V. M. (2005) Glucosepane is a major protein cross-link of the senescent human extracellular matrix. Relationship with diabetes. *J. Biol. Chem.* **280**, 12310–12315

101. Biemel, K. M., Friedl, D. A., and Lederer, M. O. (2002) Identification and quantification of major maillard cross-links in human serum albumin and lens protein : evidence for glucosepane as the dominant compound. *J. Biol. Chem.* **277**, 24907–24915

102. Takeuchi, M., Kikuchi, S., Sasaki, N., Suzuki, T., Watai, T., Iwaki, M., Bucala, R., and Yamagishi, S. (2004) Involvement of Advanced Glycation End-products (AGEs) in Alzheimers Disease. *Curr. Alzheimer Res.* **1**, 39–46

103. Bucciarelli, L. G., Wendt, T., Rong, L., Lalla, E., Hofmann, M. A., Goova, M. T., Taguchi, A., Yan, S. F., Yan, S. D., Stern, M. D., and Schmidt, A. M. (2002) RAGE is a multiligand receptor of the immunoglobulin superfamily : Implications for homeostasis and chronic disease. *Cell. Mol. Life Sci.* **59**, 1117–1128

104. Maillard, L. C. (1913) Action des acides amines sur les sucres: formation des melanoidines par voie methodique. *C.R.Acad.Sci.Ser.2*. **154**, 66–68

105. Collier, T. A., Nash, A., Birch, H. L., and de Leeuw, N. H. (2015) Preferential sites for intramolecular glucosepane cross-link formation in type I collagen: A thermodynamic study. *Matrix Biol.* **48**, 78–88

106. Walton, D. J., and Shilton, B. H. (1991) Site specificity of protein glycation. *Amino Acids*. **1**, 199–203

107. Laroque, D., Inisan, C., Berger, C., Vouland, É., Dufossé, L., and Guérard, F. (2008) Kinetic study on the Maillard reaction. Consideration of sugar reactivity. *Food Chem.* **111**, 1032–1042

108. Schalkwijk, C. G., Stehouwer, C. D. a, and van Hinsbergh, V. W. M. (2004) Fructose-mediated non-enzymatic glycation: sweet coupling or bad modification. *Diabetes. Metab. Res. Rev.* **20**, 369–82

109. Lederer, M. O., Bühler, H. P., and Bu, H. P. (1999) Cross-linking of proteins by maillard processes—characterization and detection of a lysine-arginine cross-link derived from d-glucose. *Bioorg. Med. Chem.* **7**, 1081–1088

110. Monnier, V. M., Sun, W., Sell, D. R., Fan, X., Nemet, I., and Genuth, S. (2014) Glucosepane: a poorly understood advanced glycation end product of growing importance for diabetes and its complications. *Clin. Chem. Lab. Med.* **52**, 21–32

111. Peyroux, J., and Sternberg, M. (2006) Advanced glycation endproducts (AGEs): pharmacological inhibition in diabetes. *Pathol. Biol.* **54**, 405–419

112. Reihl, O., Rothenbacher, T. M., Lederer, M. O., and Schwack, W. (2004) Carbohydrate carbonyl mobility—the key process in the formation of α-dicarbonyl intermediates. *Carbohydr. Res.* **339**, 1609–1618

113. Sjöberg, J. S., and Bulterijs, S. (2009) Characteristics, formation, and pathophysiology of glucosepane: a major protein cross-link. *Rejuvenation Res.* **12**, 137–148

114. Biemel, K. M., Reihl, O., Conrad, J., and Lederer, M. O. (2001) Formation pathways for lysine-arginine cross-links derived from hexoses and pentoses by Maillard Processes. *J. Biol. Chem.* **276**, 23405–23412

115. Diabetes UK (2012) *Diabetes in the UK 2012 - Key statistics on diabetes*

116. Johnson, R. J., Segal, M. S., Sautin, Y., Nakagawa, T., Feig, D. I., Kang, D.-H., Gersch, M. S., Benner, S., and Sánchez-Lozada, L. G. (2007) Potential role of sugar (fructose) in the epidemic of hypertension, obesity and the metabolic syndrome, diabetes, kidney disease, and cardiovascular disease. *Am. J. Clin. Nutr.* . **86**, 899–906

117. Sánchez-Lozada, L. G., Le, M., Segal, M., Johnson, R. J., and Sa, L. G. (2008) How safe is fructose for persons with or without diabetes? *Am. J. Clin. Nutr.* **88**, 1189–1190

118. Bellmunt, M. J., Portero, M., Pamplona, R., Cosso, L., Odetti, P., and Prat, J. (1995) Evidence for the Maillard reaction in rat lung collagen and its relationship with solubility and age. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1272**, 53–60

119. Susic, D., Varagic, J., Ahn, J., and Frohlich, E. D. (2004) Crosslink breakers: a new approach to cardiovascular therapy. *Curr. Opin. Cardiol.* **19**, 336–40

120. Zieman, S. J., and Kass, D. A. (2004) Advanced Glycation End Product Cross-Linking: Pathophysiologic role and therapeutic target in cardiovascular disease. *Congest. Hear. Fail.* **10**, 144–151

121. Corman, B., Duriez, M., Poitevin, P., Heudes, D., Bruneval, P., Tedgui, A., and Levy, B. I. (1998) Aminoguanidine prevents age-related arterial stiffening and cardiac hypertrophy. *Proc. Natl. Acad. Sci.* . **95**, 1301–1306

122. Kass, D. A., Shapiro, E. P., Kawaguchi, M., Capriotti, A. R., Scuteri, A., Robert, C., and Lakatta, E. G. (2001) Clinical investigation and reports improved arterial compliance by a novel advanced. *Circulation*. **104**, 1464–1470

123. Verzijl, N., DeGroot, J., Thorpe, S. R., Bank, R. A., Shaw, J. N., Lyons, T. J., Bijlsma, J. W., Lafeber, F. P., Baynes, J. W., and TeKoppele, J. M. (2000) Effect of collagen turnover on the accumulation of advanced glycation end products. *J. Biol. Chem.* **275**, 39027–39031

124. Thorpe, C. T., Streeter, I., Pinchbeck, G. L., Goodship, A. E., Clegg, P. D., and Birch, H. L. (2010) Aspartic acid racemization and collagen degradation markers reveal an accumulation of damage in tendon collagen that is enhanced with aging. *J. Biol. Chem.* **285**, 15674–15681

125. Wang, X., Shen, X., Li, X., and Agrawal, C. M. (2002) Age-related changes in the collagen network and toughness of bone. *Bone*. **31**, 1–7

126. Reddy, G. K. (2004) Cross-linking in collagen by nonenzymatic glycation increases the matrix stiffness in rabbit achilles tendon. *Exp. Diabesity Res.* **5**, 143–153

127. Wallace, D. (2003) Collagen gel systems for sustained delivery and tissue engineering. *Adv. Drug Deliv. Rev.* **55**, 1631–1649

128. Huang, D., Chang, T. R., Aggarwal, A., Lee, R. C., and Ehrlich, H. P. (1993) Mechanisms and dynamics of mechanical strengthening in ligament-equivalent fibroblast-populated collagen matrices. *Ann. Biomed. Eng.* **21**, 289–305

129. Barocas, V. H., and Tranquillo, R. T. (1997) An anisotropic biphasic theory of tissue-equivalent mechanics: The interplay among cell traction, fibrillar network deformation, fibril alignment, and cell contact guidance. *J. Biomech. Eng.* **119**, 137

130. Weadock, K. S., Miller, E. J., Bellincampi, L. D., Zawadsky, J. P., and Dunn, M. G. (1995) Physical crosslinking of collagen fibers: comparison of ultraviolet irradiation and dehydrothermal treatment. *J. Biomed. Mater. Res.* **29**, 1373–9

131. Weadock, K. S., Miller, E. J., Keuffel, E. L., and Dunn, M. G. (1996) Effect of physical crosslinking methods on collagen-fiber durability in proteolytic solutions. *J. Biomed. Mater. Res.* **32**, 221–6

132. Speer, D. P., Chvapil, M., Eskelson, C. D., and Ulreich, J. (1980) Biological effects of residual glutaraldehyde in glutaraldehyde-tanned collagen biomaterials. *J. Biomed. Mater. Res.* **14**, 753–64

133. Nishi, C., Nakajima, N., and Ikada, Y. (1995) In vitro evaluation of cytotoxicity of diepoxy compounds used for biomaterial modification. *J. Biomed. Mater. Res.* **29**, 829–34

134. Girton, T. S., Oegema, T. R., and Tranquillo, R. T. (1999) Exploiting glycation to stiffen and strengthen tissue equivalents for tissue engineering. *J. Biomed. Mater. Res.* **46**, 87–92

135. Roy, R., Boskey, A., and Bonassar, L. J. (2010) Processing of type I collagen gels using nonenzymatic glycation. *J. Biomed. Mater. Res. A*.

93, 843–51

136. Walford, R. L., Mock, D., Verdery, R., and MacCallum, T. (2002) Calorie restriction in biosphere 2: Alterations in physiologic, hematologic, hormonal, and biochemical parameters in humans restricted for a 2-year period. *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.* **57**, B211–B224

137. Cefalu, W. T., Bell-Farrow, A. D., Wang, Z. Q., Sonntag, W. E., Fu, M.-X., Baynes, J. W., and Thorpe, S. R. (1995) Caloric restriction decreases age-dependent accumulation of the glycoxidation products, N∈-(Carboxymethyl)lysine and pentosidine, in rat skin collagen. *Journals Gerontol. Ser. A Biol. Sci. Med. Sci.* **50A**, B337–B341

138. Sell, D. R., Lane, M. a, Obrenovich, M. E., Mattison, J. a, Handy, A., Ingram, D. K., Cutler, R. G., Roth, G. S., and Monnier, V. M. (2003) The effect of caloric restriction on glycation and glycoxidation in skin collagen of nonhuman primates. *J. Gerontol. A. Biol. Sci. Med. Sci.* **58**, 508–16

139. Teillet, L., Verbeke, P., Gouraud, S., Bakala, H., Borot-Laloi, C., Heudes, D., Bruneval, P., and Corman, B. (2000) Food restriction prevents Advanced Glycation End Product accumulation and retards kidney aging in lean rats. *J. Am. Soc. Nephrol.* **11**, 1488–1497

140. Swamy, M. S., and Abraham, E. C. (1989) Inhibition of lens crystallin glycation and high molecular weight aggregate formation by aspirin in vitro and in vivo. *Invest. Ophthalmol. Vis. Sci.* **30**, 1120–6

141. Yue, D. K., McLennan, S., Handelsman, D. J., Delbridge, L., Reeve, T., and Turtle, J. R. (1984) The effect of salicylates on nonenzymatic glycosylation and thermal stability of collagen in diabetic rats. *Diabetes*. **33**, 745–751

142. Khatami, M., Suldan, Z., David, I., Li, W., and Rockey, J. H. (1988) Inhibitory effects of pyridoxal phosphate, ascorbate and aminoguanidine on nonenzymatic glycosylation. *Life Sci.* **43**, 1725–1731

143. Khalifah, R. G., Baynes, J. W., and Hudson, B. G. (1999) Amadorins: novel post-Amadori inhibitors of advanced glycation reactions. *Biochem. Biophys. Res. Commun.* **257**, 251–258

144. Vander Jagt, D. L., and Hunsaker, L. a (2003) Methylglyoxal metabolism and diabetic complications: roles of aldose reductase, glyoxalase-I, betaine aldehyde dehydrogenase and 2-oxoaldehyde dehydrogenase. *Chem. Biol. Interact.* **143**–**144**, 341–51

145. Delpierre, G., Collard, F., Fortpied, J., Van Schaftingen, E., Schaftingen, E. V. A. N., and Collard, S. (2002) Fructosamine 3-kinase is involved in an intracellular deglycation pathway in human erythrocytes. *Biochem. J.* **365**, 801–808

146. Delplanque, J., Delpierre, G., Opperdoes, F. R., and Van Schaftingen, E. (2004) Tissue Distribution and Evolution of Fructosamine 3-Kinase and Fructosamine 3-Kinase-related Protein. *J. Biol. Chem.* **279**, 46606–46613

147. Delpierre, G., Collard, F., Fortpied, J., and Schaftingen, E. Van (2002) Fructosamine 3-kinase is involved in an intracellular deglycation pathway in human erythrocytes. *Biochem. J.* **365**, 801–808

148. de Grey, A. D. (2006) Foreseeable pharmaceutical repair of age-related extracellular damage. *Curr. Drug Targets*. **7**, 1469–1477

149. Kass, D. A., Shapiro, E. P., Kawaguchi, M., Capriotti, A. R., Scuteri, A., Robert, C., Lakatta, E. G., and DeGroof, R. C. (2001) Improved arterial compliance by a novel advanced glycation end-product crosslink breaker. *Circulation*. **104**, 1464–1470

150. Vaitkevicius, P. V, Lane, M., Spurgeon, H., Ingram, D. K., Roth, G. S., Egan, J. J., Vasan, S., Wagle, D. R., Ulrich, P., Brines, M., Wuerth, J. P., Cerami, A., and Lakatta, E. G. (2001) A cross-link breaker has sustained effects on arterial and ventricular properties in older rhesus monkeys. *Proc. Natl. Acad. Sci.* . **98**, 1171–1175

151. Vasan, S., Foiles, P., and Founds, H. (2003) Therapeutic potential of breakers of advanced glycation end product–protein crosslinks. *Arch. Biochem. Biophys.* **419**, 89–96

152. Draghici, C., Wang, T., and Spiegel, D. A. (2015) Concise total synthesis of glucosepane. *Science*. **350**, 294–298

153. Nasiri, R., Zahedi, M., Jamet, H., and Moosavi-Movahedi, A. A. (2012) Theoretical studies on models of lysine-arginine cross-links derived from α-oxoaldehydes: a new mechanism for glucosepane formation. *J. Mol. Model.* **18**, 1645–59

154. Nasiri, R., Field, M. J., Zahedi, M., and Moosavi-Movahedi, A. A. (2011) Cross-linking mechanisms of arginine and lysine with α,β-dicarbonyl compounds in aqueous solution. *J. Phys. Chem. A*. **115**, 13542–55

155. Young, D. C. (2001) *Computational Chemistry*, John Wiley & Sons, Inc., New York, USA, 10.1002/0471220655

156. Jensen, F. (2007) *Introduction to Computational Chemistry*, Second, John Wiley & Sons, Ltd.

157. McCammon, J. A., Gelin, B. R., and Karplus, M. (1977) Dynamics of folded proteins. *Nature*. **267**, 585–590

158. Hayward, D. O. (2002) *Quantum mechanics for chemists*, Tutorial chemistry texts, Royal Society of Chemistry

159. Ferguson, D. M. (1995) Parametrization and evaluation of a flexible water model. *J. Comput. Chem.* **16**, 501–511

160. Morse, P. (1929) Diatomic molecules according to the wave mechanics. II. Vibrational levels. *Phys. Rev.* **34**, 57–64

161. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983) CHARMM: A program for

macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217

162. Jones, J. E. (1924) On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. London A*. **106**, 463–477

163. Buckingham, R. A. (1938) The classical equation of state of gaseous helium, neon and argon. *Proc. R. Soc. London A*. **168**, 264–283

164. Gaydaenko, V. I., and Nikulin, V. K. (1970) Born-Mayer interatomic potential for atoms with z=2 to z=36. *Chem. Phys. Lett.* **7**, 360–362

165. Rahman, A. (1964) Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **136**, A405–A411

166. Duan, Y., and Kollman, P. A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*. **282**, 740–744

167. Schlick, T., Collepardo-Guevara, R., Halvorsen, L. A., Jung, S., and Xiao, X. (2011) Biomolecular modeling and simulation: a field coming of age. *Q. Rev. Biophys.* **44**, 191–228

168. Anderson, J. A., Lorenz, C. D., and Travesset, A. (2008) General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **227**, 5342–5359

169. Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., Karplus, M., Gumbart, J., Aksimentiev, A., Tajkhorshid, E., Wang, Y., and Schulten, K. (2009) CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–1614

170. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., and Woods, R. J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688

171. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802

172. Hess, B., Kutzner, C., van der Spoel, D., and Lindahl, E. (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447

173. Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–18

174. Verlet, L. (1967) Computer “experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98–103

175. Van Gunsteren, W. F., and Berendsen, H. J. C. (1988) A Leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1**, 173–185

176. Nicos S. Martys, R. D. M. (1999) Velocity Verlet algorithm for dissipative-particle-dynamics-based models for suspensions. *Phys. Rev. E*. **59**, 3733–3736

177. Bussi, G., Donadio, D., and Parrinello, M. (2007) Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 1–8

178. Andersen, H. C. (1980) Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **72**, 2384

179. Evans, D. J., and Holian, B. L. (1985) The Nose–Hoover thermostat. *J. Chem. Phys.* **83**, 4069

180. Grest, G. S., and Kremer, K. (1986) Molecular dynamics simulation for polymers in the presence of a heat bath. *Phys. Rev. A*. **33**, 3628–3631

181. Ceriotti, M., Bussi, G., and Parrinello, M. (2010) Colored-noise thermostats àla Carte. *J. Chem. Theory Comput.* **6**, 1170–1180

182. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A., and Haak, J. R. (1984) Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690

183. Parrinello, M., and Rahman, A. (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190

184. Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. C. (1977) Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341

185. Hess, B., Bekker, H., Berendsen, H. J. C., and Fraaije, J. G. E. M. (1997) LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472

186. Roux, B., and Simonson, T. (1999) Implicit solvent models. *Biophys. Chem.* **78**, 1–20

187. Tsui, V., and Case, D. A. (2000) Theory and applications of the Generalized Born solvation model in macromolecular simulations. *Biopolymers*. **56**, 275–291

188. Hassan, S. A., and Mehler, E. L. (2002) A critical analysis of continuum electrostatics: The screened Coulomb potential-implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. *Proteins Struct. Funct. Genet.* **47**, 45–61

189. Im, W., Lee, M. S., and Brooks, C. L. (2003) Generalized born model with a simple smoothing function. *J. Comput. Chem.* **24**, 1691–1702

190. Chen, J., Brooks, C. L., and Khandogin, J. (2008) Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr. Opin. Struct. Biol.* **18**, 140–8

191. Simonson, T. (2001) Macromolecular electrostatics: continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **11**, 243–252

192. Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987) The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271

193. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926

194. Mahoney, M. W., and Jorgensen, W. L. (2000) A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.* **112**, 8910

195. Nelder, J. A. A., and Mead, R. (1965) A simplex method for function minimization. *Comput. J.* **7**, 308–313

196. Faller, R., Schmitz, H., Biermann, O., and Müller-Plathe, F. (1999) Automatic parameterization of force fields for liquids by simplex optimization. *J. Comput. Chem.* **20**, 16

197. Meyer, H., Biermann, O., Faller, R., Reith, D., and Müller-Plathe, F. (2000) Coarse graining of nonbonded inter-particle potentials using automatic simplex optimization to fit structural properties. *J. Chem. Phys.* **113**, 6264–6275

198. Torrie, G. M., and Valleau, J. P. (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199

199. Roux, B. (1995) The calculation of the potential of mean force using computer simulations. *Comput. Phys. Commun.* **91**, 275–282

200. Kästner, J. (2011) Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 932–942

201. Gullingsrud, J. R., Braun, R., and Schulten, K. (1999) Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations. *J. Comput. Phys.* **151**, 190–211

202. Jarzynski, C. (1997) Nonequilibrium equality for free energy differences, Phys. *Rev. Lett.* **78**, 2690

203. Born, M., and Oppenheimer, R. (1927) Zur Quantentheorie der Molekeln. *Ann. Phys.* **389**, 457–484

204. Slater, J. C. (1930) Note on Hartree's Method. *Phys. Rev.* **35**, 210–211

205. Seminario, J. M. (1996) Calculation of intramolecular force fields from second-derivative tensors. *Int. J. Quantum Chem.* **60**, 1271–1277

206. Burger, S., Lacasse, M., Verstraelen, T., Drewry, J., Gunning, P., and Ayers, P. W. (2012) Automated parameterization of AMBER force field terms from vibrational analysis with a focus on functionalizing dinuclear zinc (II) scaffolds. *J. Chem. Theory Comput.* **8**, 554–562

207. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004) Development and testing of a general Amber force field. *J. Comput. Chem.* **25**, 1157–1174

208. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., and Kollman, P. A. (1995) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197

209. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **106**, 765–784

210. Kirschner, K. N., Yongye, A. B., Tschampel, S. M., Gonza, J., Daniels, C. R., Foley, B. L., Woods, R. J., and González-Outeiriño, J. (2008) GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29**, 622–55

211. Dickson, C. J., Madej, B. D., Skjevik, Å. A., Betz, R. M., Teigen, K., Gould, I. R., and Walker, R. C. (2014) Lipid14: The Amber lipid force field. *J. Chem. Theory Comput.* **10**, 865–879

212. Cornell, W. D., Cieplak, P., Bayly, C. I., and Kollman, P. A. (1993) Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.* **115**, 9620–9631

213. Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97**, 10269–10280

214. Kuyper, L. F., Hunter, R. N., Ashton, D., Merz, K. M., and Kollman, P. A. (1991) Free energy calculations on the relative solvation free energies of benzene, anisole, and 1,2,3-trimethoxybenzene: Theoretical and experimental analysis of aromatic methoxy solvation. *J. Phys. Chem.* **95**, 6661–6666

215. Dupradeau, F.-Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., and Cieplak, P. (2010) The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **12**, 7821

216. Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F. Y. (2011) R.E.D. Server: A web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **39**, 511–517

217. Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P. (2003) A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012

218. Lee, M. C., and Duan, Y. (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized Born solvent model. *Proteins Struct. Funct. Genet.* **55**, 620–634

219. Park, S., Radmer, R. J., Klein, T. E., and Pande, V. S. (2005) A new set of molecular mechanics parameters for hydroxyproline and its use in molecular dynamics simulations of collagen-like peptides. *J. Comput. Chem.* **26**, 1612–1616

220. Frisch, M. J., and al, et. (2009) Gaussian09 Revision A.1. *Gaussian 09, Revis. A.02*

221. Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeerschd, T., Zurek, E., and Hutchison, G. R. (2012) Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **4**, 1–17

222. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard III, W. A., and Skiff, W. M. (1992) UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035

223. Rassolov, V. A., Pople, J. A., Ratner, M. A., and Windus, T. L. (1998) 6-31G* basis set for atoms K through Zn. *J. Chem. Phys.* **109**, 1223–1229

224. Wang, J., Cieplak, P., and Kollman, P. A. (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **21**, 1049–1074

225. Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006) Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins Struct. Funct. Bioinforma.* **65**, 712–725

226. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015) ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713

227. Nash, A., Collier, T. A., Birch, H. L., and de Leeuw, N. H. (2016) A toolset for the mechanical parameterization of Advanced Glycation End Products. *Unpubl. Work*

228. Monnier, V. M., Mustata, G. T., Biemel, K. L., Reihl, O., Lederer, M. O., Zhenyu, D., and Sell, D. R. (2005) Cross-linking of the extracellular matrix by the maillard reaction in aging and diabetes: an update on “a puzzle nearing resolution”. *Ann. N. Y. Acad. Sci.* **1043**, 533–544

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

230. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42**, D191–D198

231. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr., E. E., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542

232. Case, D. A., Darden, T. A., Cheatham, T. E., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Götz, A. W., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wolf, R. M., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. (2012) AMBER 12

233. Sivan, S., Merkher, Y., Wachtel, E., Ehrlich, S., and Maroudas, A. (2006) Correlation of swelling pressure and intrafibrillar water in young and aged human intervertebral discs. *J. Orthop. Res.* **24**, 1292–1298

234. Linn, F. C., and Sokoloff, L. (1965) Movement and composition of interstitial fluid of cartilage. *Arthritis Rheum.* **8**, 481–494

235. Wuthier, R. E. (1977) Electrolytes of isolated epiphyseal chondrocytes, matrix vesicles, and extracellular fluid. *Calcif. Tissue Res.* **23**, 125–133

236. Orgel, J. P. R. ., Miller, A., Irving, T. C., Fischetti, R. F., Hammersley, A. P., and Wess, T. J. (2001) The in situ supermolecular structure of type I collagen. *Structure*. **9**, 1061–1069

237. Mallard, P. J., and Linstrom, W. G. (eds.) (2016) *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, National Institute of Standards and Technology, Gaithersburg MD, 20899

238. Acosta, J., Hettinga, J., Flückiger, R., Krumrei, N., Goldfine, A., Angarita, L., and Halperin, J. (2000) Molecular basis for a link between complement and the vascular complications of diabetes. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 5450–5455

239. Darden, T., York, D., and Pedersen, L. (1993) Particle mesh Ewald: An

N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089

240. Eyre, D. R., Weis, M. A., and Wu, J. (2009) Advances in dollagen cross-link analysis. *Methods*. **45**, 65–74

241. Sweeney, S. M., Orgel, J. P., Fertala, A., McAuliffe, J. D., Turner, K. R., Di Lullo, G. A., Chen, S., Antipova, O., Perumal, S., Ala-Kokko, L., Forlino, A., Cabral, W. A., Barnes, A. M., Marini, J. C., and San Antonio, J. D. (2008) Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates. *J. Biol. Chem.* **283**, 21187–21197

242. Knight, C. G., Morton, L. F., Peachey, a. R., Tuckwell, D. S., Farndale, R. W., and Barnes, M. J. (2000) The collagen-binding A-domains of integrins α1β 1 and α2β1 recognize the same specific amino acid sequence, GFOGER, in native (triple-helical) collagens. *J. Biol. Chem.* **275**, 35–40

243. San Antonio, J. D., Lander, A. D., Karnovsky, M. J., and Slayter, H. S. (1994) Mapping the heparin-binding sites on type I collagen monomers and fibrils. *J. Cell Biol.* **125**, 1179–1188

244. Razzaque, M. S., Nazneen, A., and Taguchi, T. (1998) Immunolocalization of collagen and collagen-binding heat shock protein 47 in fibrotic lung diseases. *Mod. Pathol.* **11**, 1183–1188

245. Veis, A., and Perry, A. (1967) The Phosphoprotein of the Dentin Matrix. *Biochemistry*. **6**, 2409–2416

246. Chung, L., Dinakarpandian, D., Yoshida, N., Lauer-Fields, J. L., Fields, G. B., Visse, R., and Nagase, H. (2004) Collagenase unwinds triple-helical collagen prior to peptide bond hydrolysis. *EMBO J.* **23**, 3020–3030

247. Scott, J. E., and Haigh, M. (1988) Identification of specific binding sites for keratan sulphate proteoglycans and chondroitin-dermatan sulphate proteoglycans on collagen fibrils in cornea by the use of Cupromeronic Blue in ' critical-electrolyte-concentration ' techniques. *Biochem. J.* **253**, 607–610

248. Somasundaram, R., Ruehl, M., Tiling, N., Ackermann, R., Schmid, M., Riecken, E. O., and Schuppan, D. (2000) Collagens serve as an extracellular store of bioactive interleukin 2. *J. Biol. Chem.* **275**, 38170–38175

249. Di Lullo, G. A., Sweeney, S. M., Korkko, J., Ala-Kokko, L., and San Antonio, J. D. (2002) Mapping the ligand-binding sites and disease-associated mutations on the most abundant protein in the human, type I collagen. *J. Biol. Chem.* **277**, 4223–4231

250. Reigle, K. L., Lullo, G. Di, Turner, K. R., Last, J. A., Chervoneva, I., Birk, D. E., Funderburgh, J. L., Elrod, E., Germann, M. W., Sanderson, R. D., and San Antonio, J. D. (2009) Non-enzymatic glycation of type I collagen diminishes collagen-proteoglycan binding and weakens cell adhesion. *J.*

Cell Biochem. **104**, 1684–1698

251. Reiser, K. M., Amigable, M., and Last, J. A. (1992) Nonenzymatic glycation of type I collagen. *J. Biol. Chem.* **4**, 24207–24216

252. George, A., Bannon, L., Sabsay, B., Dillon, J. W., Malone, J., Veis, A., Jenkins, N. A., Gilbert, D. J., and Copeland, N. G. (1996) The Carboxyl-terminal domain of phosphophoryn contains unique extended triplet amino acid repeat sequences forming ordered carboxyl-phosphate interaction ridges that may be essential in the biomineralization process. *J. Biol. Chem.* **271**, 32869–32873

253. Veis, A., Sfeir, C., and Wu, C. B. (1997) Phosphorylation of the proteins of the extracellular matrix of mineralized tissues by casein kinase-like activity. *Crit. Rev. Oral Biol. Med.* **8**, 360–379

254. Nanci, A. (2012) *Ten Cate's Oral Histology: development, structure, and function.*, 8th Ed., Elsevier Mosby, St. Louis

255. Ruoslahti, E. (1988) Structure and biology of proteoglycans. *Annu. Rev. Cell Biol.* **4**, 229–55

256. Sarrazin, S., Lamanna, W. C., and Esko, J. D. (2011) Heparan sulfate proteoglycans. *Cold Spring Harb. Perspect. Biol.* **3**, 1–33

257. Funderburgh, J. L. (2000) MINI REVIEW Keratan sulfate: structure, biosynthesis, and function. *Glycobiology*. **10**, 951–958

258. Emsley, J., King, S. L., Bergelson, J. M., and Liddington, R. C. (1997) Crystal Structure of the I Domain from Integrin α2β1. *J. Biol. Chem.* **272**, 28512–28517

259. Rich, R. L., Deivanayagam, C. C. S., Owens, R. T., Carson, M., Hook, A., Moore, D., Yang, V. W.-C., Narayana, S. V. L., and Hook, M. (1999) Trench-shaped binding sites promote multiple classes of interactions between collagen and the adherence receptors, α1β1 integrin and Staphylococcus aureus Cna MSCRAMM. *J. Biol. Chem.* **274**, 24906–24913

260. Schnider, S. L., and Kohn, R. R. (1981) Effects of age and diabetes mellitus on the solubility and nonenzymatic glucosylation of human skin collagen. *J. Clin. Invest.* **67**, 1630–1635

261. Gautieri, A., Redaelli, A., Buehler, M. J., and Vesentini, S. (2014) Age- and diabetes-related nonenzymatic crosslinks in collagen fibrils: Candidate amino acids involved in advanced glycation end-products. *Matrix Biol.* **34**, 89–95

262. Hollingsworth, S. A., and Karplus, P. A. (2010) A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomol. Concepts*. **1**, 271–283

263. Minary-Jolandan, M., and Yu, M.-F. (2009) Nanomechanical heterogeneity in the gap and overlap regions of type I collagen fibrils with

implications for bone heterogeneity. *Biomacromolecules*. **10**, 2565–70

264. Pradhan, S. M., Katti, K. S., and Katti, D. R. (2012) Structural hierarchy controls deformation behavior of collagen. *Biomacromolecules*. **13**, 2562–2569

265. Abrahams, M. (1967) Mechanical behaviour of tendon In vitro. *Med. Biol. Eng.* **5**, 433–443

266. Misof, K., Rapp, G., and Fratzl, P. (1997) A new molecular model for collagen elasticity based on synchrotron X-ray scattering evidence. *Biophys. J.* **72**, 1376–81

267. Fratzl, P., Misof, K., Zizak, I., Rapp, G., Amenitsch, H., and Bernstorff, S. (1998) Fibrillar structure and mechanical properties of collagen. *J. Struct. Biol.* **122**, 119–22

268. Freed, A. D., and Doehring, T. C. (2005) Elastic model for crimped collagen fibrils. *J. Biomech. Eng.* **127**, 587–593

269. Gutsmann, T., Fantner, G. E., Kindt, J. H., Venturoni, M., Danielsen, S., and Hansma, P. K. (2004) Force spectroscopy of collagen fibers to investigate their mechanical properties and structural organization. *Biophys. J.* **86**, 3186–3193

270. Lorenzo, A. C., and Caffarena, E. R. (2005) Elastic properties, Young's modulus determination and structural stability of the tropocollagen molecule: A computational study by steered molecular dynamics. *J. Biomech.* **38**, 1527–1533

271. Zhang, D., Chippada, U., and Jordan, K. (2007) Effect of the structural water on the mechanical properties of collagen-like microfibrils: a molecular dynamics study. *Ann. Biomed. Eng.* **35**, 1216–30

272. Kwansa, A. L., De Vita, R., and Freeman, J. W. (2014) Mechanical recruitment of N- and C-crosslinks in collagen type I. *Matrix Biol.* **34**, 161–9

273. Ghodsi, H., and Darvish, K. (2016) Characterization of the viscoelastic behavior of a simplified collagen micro-fibril based on molecular dynamics simulations. *J. Mech. Behav. Biomed. Mater.* **63**, 26–34

274. Su, T., and Purohit, P. K. (2009) Mechanics of forced unfolding of proteins. *Acta Biomater.* **5**, 1855–1863

275. Sotomayor, M., and Schulten, K. (2007) Single-molecule experiments in vitro and in silico. *Science*. **316**, 1144–1148

276. Mikulska, K., Strzelecki, J., and Nowak, W. (2014) Nanomechanics of β-rich proteins related to neuronal disorders studied by AFM, all-atom and coarse-grained MD methods. *J. Mol. Model.* **20**, 2144

277. Sharma, D., Perisic, O., Peng, Q., Cao, Y., Lam, C., Lu, H., and Li, H. (2007) Single-molecule force spectroscopy reveals a mechanically stable

protein fold and the rational tuning of its mechanical stability. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 9278–83

278. Harley, R., James, D., Miller, A., and White, J. W. (1977) Phonons and the elastic moduli of collagen and muscle. *Nature*. **267**, 285–287

279. Cusack, S., and Miller, A. (1979) Determination of the elastic constants of collagen by Brillouin light scattering. *J. Mol. Biol.* **135**, 39–51

280. Sasaki, N., and Odajima, S. (1996) Stress-strain curve and Young's modulus of a collagen molecule as determined by the X-ray diffraction technique. *J. Biomech.* **29**, 655–658

281. Uzel, S. G. M., and Buehler, M. J. (2009) Nanomechanical sequencing of collagen: tropocollagen features heterogeneous elastic properties at the nanoscale. *Integr. Biol.* **1**, 452

282. Shah, N. K., Ramshaw, J. A. M., Kirkpatrick, A., Shah, C., and Brodsky, B. (1996) A host-guest set of triple-helical peptides: Stability of Gly-X-Y triplets containing common nonpolar residues. *Biochemistry*. **35**, 10262–10268

283. Vitagliano, L., Némethy, G., Zagari, A., and Scheraga, H. A. (1993) Stabilization of the triple-helical structure of natural collagen by side-chain interactions. *Biochemistry*. **32**, 7354–9

284. De Simone, A., Vitagliano, L., and Berisio, R. (2008) Role of hydration in collagen triple helix stabilization. *Biochem. Biophys. Res. Commun.* **372**, 121–5

285. Roeder, B. A., Kokini, K., Sturgis, J. E., Robinson, J. P., and Voytik-Harbin, S. L. (2002) Tensile mechanical properties of three-dimensional type I collagen extracellular matrices with varied microstructure. *J. Biomech. Eng.* **124**, 214–222

286. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242

287. Berman, H., Henrick, K., and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980

288. Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94

289. Lipman, D. J., and Pearson, W. R. (1985) Rapid and sensitive protein similarity searches. *Science*. **227**, 1435–1441

290. Dunbrack, R. (2002) Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **12**, 431–440

291. Boudko, S. P., Engel, J., Okuyama, K., Mizuno, K., Bachinger, H. P., and Schumacher, M. A. (2008) Crystal structure of human type III collagen Gly991-Gly 1032 cystine knot-containing peptide shows both 7/2 and 10/3

triple helical symmetries. *J. Biol. Chem.* **283**, 32580–32589

292. Bairoch, A., and Apweiler, R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48

293. Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, 501–504

294. Pannarale, L., Braidotti, P., D'Alba, L., and Gaudio, E. (1994) Scanning electron microscopy of collagen fiber orientation in the bone Lamellar system in non-decalcified human samples. *Cells Tissues Organs*. **151**, 36–42

295. Moger, C. J., Barrett, R., Bleuet, P., Bradley, D. A., Ellis, R. E., Green, E. M., Knapp, K. M., Muthuvelu, P., and Winlove, C. P. (2007) Regional variations of collagen orientation in normal and diseased articular cartilage and subchondral bone determined using small angle X-ray scattering (SAXS). *Osteoarthritis Cartilage*. **15**, 682–7

296. Liao, J., Yang, L., Grashow, J., and Sacks, M. S. (2005) Molecular orientation of collagen in intact planar connective tissues under biaxial stretch. *Acta Biomater.* **1**, 45–54

297. Ugryumova, N., Jacobs, J., Bonesi, M., and Matcher, S. J. (2009) Novel optical imaging technique to determine the 3-D orientation of collagen fibers in cartilage: variable-incidence angle polarization-sensitive optical coherence tomography. *Osteoarthritis Cartilage*. **17**, 33–42

298. Schrof, S., Varga, P., Galvis, L., Raum, K., and Masic, A. (2014) 3D Raman mapping of the collagen fibril orientation in human osteonal lamellae. *J. Struct. Biol.* **187**, 266–275

299. Galvis, L., Dunlop, J. W. C., Duda, G., Fratzl, P., and Masic, A. (2013) Polarized Raman anisotropic response of collagen in tendon: Towards 3D orientation mapping of collagen in tissues. *PLoS One*. **8**, e63518

300. Masic, A., Bertinetti, L., Schuetz, R., Galvis, L., Timofeeva, N., Dunlop, J. W. C., Seto, J., Hartmann, M. a., and Fratzl, P. (2011) Observations of multiscale, stress-induced changes of collagen orientation in tendon by polarized Raman spectroscopy. *Biomacromolecules*. **12**, 3989–3996

301. Bi, X., Li, G., Doty, S. B., and Camacho, N. P. (2005) A novel method for determination of collagen orientation in cartilage by Fourier transform infrared imaging spectroscopy (FT-IRIS). *Osteoarthritis Cartilage*. **13**, 1050–8

302. Buckley, K., Kerns, J. G., Gikas, P. D., Birch, H. L., Vinton, J., Keen, R., Parker, A. W., Matousek, P., and Goodship, A. E. (2014) Measurement of abnormal bone composition in vivo using noninvasive Raman spectroscopy. *IBMS Bonekey*. **11**, 1–3

303. Krafft, C., and Sergo, V. (2006) Biomedical applications of Raman and infrared spectroscopy to diagnose tissues. *Spectroscopy*. **20**, 195–218

304. Falgayrac, G., Facq, S., Leroy, G., Cortet, B., and Penel, G. (2010) New method for Raman investigation of the orientation of collagen fibrils and crystallites in the Haversian system of bone. *Appl. Spectrosc.* **64**, 775–780

305. Fraser, R. D. B., MacRae, T. P., Miller, A., and Suzuki, E. (1983) Molecular conformation and packing in collagen fibrils. *J. Mol. Biol.* **167**, 497–521

306. Kukreti, U., and Belko, S. M. (2000) Collagen fibril D-period may change as a function of strain and location in ligament. *J. Biomech.* **33**, 1569–1574

307. Cameron, G. J., Cairns, D. E., and Wess, T. J. (2007) The variability in type I collagen helical pitch is reflected in the D periodic fibrillar structure. *J. Mol. Biol.* **372**, 1097–107

308. Adams, P. D., Arkin, I. T., Engelman, D. M., and Brünger, A. T. (1995) Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nat Struct Biol*. **2**, 154–162

309. Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD -Visual Molecular Dynamics. *J. Mol. Graph.* **14**, 33–38

310. Orgel, J. ., Irving, T. ., Miller, A., and Wess, T. J. (2006) Microfibrillar Structure of Type I Collagen in Situ. *Proc.Natl.Acad.Sci.USA*. **103**, 9001–9005

311. Franchi, M., Trirè, A., Quaranta, M., Orsini, E., and Ottani, V. (2007) Collagen structure of tendon relates to function. *ScientificWorldJournal.* **7**, 404–420

Appendix 1

9.1 Glucosepane Parameters

```
!!index array str
 "ARC"
!entry.ARC.unit.atoms table  str name  str type  int typex  int resx  int flags  int seq  int elmnt  dbl chg
 "N" "N" 0 1 131072 1 7 -0.347900
 "H" "H" 0 1 131072 2 1 0.274700
 "CA" "CT" 0 1 131072 3 6 -0.263700
 "HA" "H1" 0 1 131072 4 1 0.156000
 "CB" "CT" 0 1 131072 5 6 -0.000700
 "HB2" "HC" 0 1 131072 6 1 0.032700
 "HB3" "HC" 0 1 131072 7 1 0.032700
 "CG" "CT" 0 1 131072 8 6 0.039000
 "HG2" "HC" 0 1 131072 9 1 0.028500
 "HG3" "HC" 0 1 131072 10 1 0.028500
 "CD" "CT" 0 1 131072 11 6 0.048600
 "HD2" "H1" 0 1 131072 12 1 0.068700
 "HD3" "H1" 0 1 131072 13 1 0.068700
 "NE" "N2" 0 1 131072 14 7 -0.529500
 "HE" "H" 0 1 131072 15 1 0.345600
 "CZ" "CA" 0 1 131072 16 6 0.77990
 "NH1" "N2" 0 1 131072 17 7 -0.862700
 "NH2" "N2" 0 1 131072 18 7 -0.862700
 "C" "C" 0 1 131072 19 6 0.734100
 "O" "O" 0 1 131072 20 8 -0.589400
!entry.ARC.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "N" "N" 0 -1 0.0
 "H" "H" 0 -1 0.0
 "CA" "CT" 0 -1 0.0
 "HA" "H1" 0 -1 0.0
 "CB" "CT" 0 -1 0.0
 "HB2" "HC" 0 -1 0.0
 "HB3" "HC" 0 -1 0.0
 "CG" "CT" 0 -1 0.0
 "HG2" "HC" 0 -1 0.0
 "HG3" "HC" 0 -1 0.0
 "CD" "CT" 0 -1 0.0
 "HD2" "H1" 0 -1 0.0
 "HD3" "H1" 0 -1 0.0
 "NE" "N2" 0 -1 0.0
 "HE" "H" 0 -1 0.0
 "CZ" "CA" 0 -1 0.0
 "NH1" "N2" 0 -1 0.0
 "NH2" "N2" 0 -1 0.0
 "C" "C" 0 -1 0.0
 "O" "O" 0 -1 0.0
!entry.ARC.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.ARC.unit.childsequence single int
 2
!entry.ARC.unit.connect array int
 1
 19
```

```
 17
 18
!entry.ARC.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 3 4 1
 3 5 1
 3 19 1
 5 6 1
 5 7 1
 5 8 1
 8 9 1
 8 10 1
 8 11 1
 11 12 1
 11 13 1
 11 14 1
 14 15 1
 14 16 1
 16 17 2
 16 18 1
 19 20 1
!entry.ARC.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
 "R" 1 "A" 17
 "R" 1 "A" 18
 "R" 1 "A" 19
 "R" 1 "A" 20
!entry.ARC.unit.name single str
 "ARC"
!entry.ARC.unit.positions table  dbl x  dbl y  dbl z
 3.325770 1.547909 -1.607204E-06
 3.909407 0.723611 -2.739882E-06
 3.970048 2.845795 -1.311163E-07
 3.671663 3.400129 -0.889820
 3.576965 3.653838 1.232143
 2.496995 3.801075 1.241379
 3.877484 3.115795 2.131197
 4.274186 5.009602 1.194577
 5.354271 4.863178 1.185788
 3.973781 5.548460 0.295972
 3.881105 5.817645 2.426721
 2.801135 5.964881 2.435959
 4.181626 5.279602 3.325774
 4.540320 7.142723 2.424483
 5.151805 7.375492 1.655065
```

 4.364284 8.040989 3.389382
 3.575026 7.807606 4.434133
 5.006254 9.201287 3.286991
 5.485541 2.705207 -4.398755E-06
 6.008824 1.593175 -8.449768E-06
!entry.ARC.unit.residueconnect table int c1x int c2x int c3x int c4x int c5x int c6x
 1 19 17 18 0 0
!entry.ARC.unit.residues table str name int seq int childseq int startatomx str restype int imagingx
 "ARC" 1 21 1 "p" 0
!entry.ARC.unit.residuesPdbSequenceNumber array int
 0
!entry.ARC.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.ARC.unit.velocities table dbl x dbl y dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0

!!index array str

 "LYC"
!entry.LYC.unit.atoms table str name str type int typex int resx int flags int seq int elmnt dbl chg
 "N" "N" 0 1 131072 1 7 -0.347900
 "H" "H" 0 1 131072 2 1 0.274700
 "CA" "CT" 0 1 131072 3 6 -0.240000
 "HA" "H1" 0 1 131072 4 1 0.142600
 "CB" "CT" 0 1 131072 5 6 -0.009400
 "HB2" "HC" 0 1 131072 6 1 0.036200
 "HB3" "HC" 0 1 131072 7 1 0.036200
 "CG" "CT" 0 1 131072 8 6 0.018700
 "HG2" "HC" 0 1 131072 9 1 0.010300
 "HG3" "HC" 0 1 131072 10 1 0.010300
 "CD" "CT" 0 1 131072 11 6 -0.047900
 "HD2" "HC" 0 1 131072 12 1 0.062100

```
 "HD3" "HC" 0 1 131072 13 1 0.062100
 "CE" "CT" 0 1 131072 14 6  0.154300
 "HE2" "H1" 0 1 131072 15 1 0.113500
 "HE3" "H1" 0 1 131072 16 1 0.113500
 "NZ" "NT" 0 1 131072 17 7 -0.185400
 "C" "C" 0 1 131072 18 6 0.734100
 "O" "O" 0 1 131072 19 8 -0.589400
!entry.LYC.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "N" "N" 0 -1 0.0
 "H" "H" 0 -1 0.0
 "CA" "CT" 0 -1 0.0
 "HA" "H1" 0 -1 0.0
 "CB" "CT" 0 -1 0.0
 "HB2" "HC" 0 -1 0.0
 "HB3" "HC" 0 -1 0.0
 "CG" "CT" 0 -1 0.0
 "HG2" "HC" 0 -1 0.0
 "HG3" "HC" 0 -1 0.0
 "CD" "CT" 0 -1 0.0
 "HD2" "HC" 0 -1 0.0
 "HD3" "HC" 0 -1 0.0
 "CE" "CT" 0 -1 0.0
 "HE2" "H1" 0 -1 0.0
 "HE3" "H1" 0 -1 0.0
 "NZ" "NT" 0 -1 0.0
 "C" "C" 0 -1 0.0
 "O" "O" 0 -1 0.0
!entry.LYC.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.LYC.unit.childsequence single int
 2
!entry.LYC.unit.connect array int
 1
 18
 17
!entry.LYC.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 3 4 1
 3 5 1
 3 18 1
 5 6 1
 5 7 1
 5 8 1
 8 9 1
 8 10 1
 8 11 1
 11 12 1
 11 13 1
 11 14 1
 14 15 1
 14 16 1
 14 17 1
 18 19 1
!entry.LYC.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
```

```
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
 "R" 1 "A" 17
 "R" 1 "A" 18
 "R" 1 "A" 19
!entry.LYC.unit.name single str
 "LYC"
!entry.LYC.unit.positions table  dbl x  dbl y  dbl z
 3.325770 1.547909 -1.607204E-06
 3.909407 0.723611 -2.739882E-06
 3.970048 2.845795 -1.311163E-07
 3.671663 3.400129 -0.889820
 3.576965 3.653838 1.232143
 2.496995 3.801075 1.241379
 3.877484 3.115795 2.131197
 4.274186 5.009602 1.194577
 5.354271 4.863178 1.185788
 3.973781 5.548460 0.295972
 3.881105 5.817645 2.426721
 2.801135 5.964881 2.435959
 4.181626 5.279602 3.325774
 4.578325 7.173410 2.389153
 5.658410 7.026987 2.380363
 4.277917 7.712267 1.490550
 4.199422 7.952309 3.576860
 5.485541 2.705207 -4.398755E-06
 6.008824 1.593175 -8.449768E-06
!entry.LYC.unit.residueconnect table  int c1x  int c2x  int c3x  int c4x  int c5x  int c6x
 1 18 17 0 0 0
!entry.LYC.unit.residues table  str name  int seq  int childseq  int startatomx  str restype  int imagingx
 "LYC" 1 20 1 "p" 0
!entry.LYC.unit.residuesPdbSequenceNumber array int
 0
!entry.LYC.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.LYC.unit.velocities table  dbl x  dbl y  dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
```

```
!!index array str
 "ORG"
!entry.ORG.unit.atoms table  str name  str type  int typex  int resx  int flags  int seq  int elmnt  dbl chg
 "C1" "CC" 0 1 131072 1 6 0.64350
 "C2" "CT" 0 1 131072 2 6 0.078300
 "H2" "H1" 0 1 131072 3 1 0.038300
 "H3" "H1" 0 1 131072 4 1 0.038200
 "C3" "CT" 0 1 131072 5 6 0.080800
 "C4" "CT" 0 1 131072 6 6 -0.024700
 "C5" "CT" 0 1 131072 7 6  0.8421
 "H8" "HC" 0 1 131072 8 1 0.029100
 "H9" "HC" 0 1 131072 9 1 0.029100
 "C6" "CT" 0 1 131072 10 6  0.89940
 "H10" "H1" 0 1 131072 11 1 0.052600
 "O1" "OH" 0 1 131072 12 8 -0.389500
 "H4" "H1" 0 1 131072 13 1 0.061600
 "H5" "HO" 0 1 131072 14 1 0.209900
 "O2" "OH" 0 1 131072 15 8 -0.389300
 "H6" "H1" 0 1 131072 16 1 0.061900
 "H7" "HO" 0 1 131072 17 1 0.209900
!entry.ORG.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "C1" "CC" 0 -1 0.0
 "C2" "CT" 0 -1 0.0
 "H2" "H1" 0 -1 0.0
 "H3" "H1" 0 -1 0.0
 "C3" "CT" 0 -1 0.0
 "C4" "CT" 0 -1 0.0
 "C5" "CT" 0 -1 0.0
 "H8" "HC" 0 -1 0.0
 "H9" "HC" 0 -1 0.0
 "C6" "CA" 0 -1 0.0
 "H10" "H1" 0 -1 0.0
 "O1" "OH" 0 -1 0.0
 "H4" "H1" 0 -1 0.0
 "H5" "HO" 0 -1 0.0
 "O2" "OH" 0 -1 0.0
 "H6" "H1" 0 -1 0.0
 "H7" "HO" 0 -1 0.0
!entry.ORG.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.ORG.unit.childsequence single int
 2
!entry.ORG.unit.connect array int
 10
 7
 1
!entry.ORG.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 1 4 1
 2 5 1
 2 13 1
 2 12 1
 5 6 1
 5 16 1
 5 15 1
 6 8 1
 6 7 1
 6 9 1
 7 10 1
 7 11 1
 12 14 1
 15 17 1
!entry.ORG.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
```

```
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
 "R" 1 "A" 17
!entry.ORG.unit.name single str
 "ORG"
!entry.ORG.unit.positions table  dbl x  dbl y  dbl z
 170.712000 33.080000 1.510203E+03
 170.801000 32.169000 1.511438E+03
 170.476000 32.378000 1.509387E+03
 169.804000 33.689000 1.510307E+03
 171.239000 32.851000 1.512734E+03
 172.722000 33.208000 1.512706E+03
 172.904000 34.572000 1.512056E+03
 173.303000 32.423000 1.512207E+03
 173.105000 33.237000 1.513736E+03
 172.405000 34.829000 1.510658E+03
 172.391000 35.289000 1.512715E+03
 169.468000 31.623000 1.511691E+03
 171.435000 31.298000 1.511238E+03
 168.853000 32.378000 1.511661E+03
 170.966000 31.950000 1.513829E+03
 170.614000 33.732000 1.512928E+03
 170.149000 31.484000 1.513560E+03
!entry.ORG.unit.residueconnect table  int c1x  int c2x  int c3x  int c4x  int c5x  int c6x
 10 7 1 0 0 0
!entry.ORG.unit.residues table  str name  int seq  int childseq  int startatomx  str restype  int imagingx
 "ORG" 1 18 1 "p" 0
!entry.ORG.unit.residuesPdbSequenceNumber array int
 0
!entry.ORG.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.ORG.unit.velocities table  dbl x  dbl y  dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
```

9.2 DOGDIC Parameters

```
!!index array str
 "ARD"
!entry.ARD.unit.atoms table  str name  str type  int typex  int resx  int flags  int seq  int elmnt  dbl chg
 "N" "N" 0 1 131072 1 7 -0.347900
 "H" "H" 0 1 131072 2 1 0.274700
 "CA" "CT" 0 1 131072 3 6 -0.263700
 "HA" "H1" 0 1 131072 4 1 0.156000
 "CB" "CT" 0 1 131072 5 6 -0.000700
 "HB2" "HC" 0 1 131072 6 1 0.032700
 "HB3" "HC" 0 1 131072 7 1 0.032700
 "CG" "CT" 0 1 131072 8 6 0.039000
 "HG2" "HC" 0 1 131072 9 1 0.028500
 "HG3" "HC" 0 1 131072 10 1 0.028500
 "CD" "CT" 0 1 131072 11 6 0.048600
 "HD2" "H1" 0 1 131072 12 1 0.068700
 "HD3" "H1" 0 1 131072 13 1 0.068700
 "NE" "N2" 0 1 131072 14 7 -0.729500
 "HE" "H" 0 1 131072 15 1 0.345600
 "CZ" "CA" 0 1 131072 16 6 0.779900
 "NH1" "N2" 0 1 131072 17 7 -0.69610
 "NH2" "N2" 0 1 131072 18 7 -0.69610
 "HH2" "H" 0 1 131072 19 1 0.457800
 "C" "C" 0 1 131072 20 6 0.734100
 "O" "O" 0 1 131072 21 8 -0.589400
!entry.ARD.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "N" "N" 0 -1 0.0
 "H" "H" 0 -1 0.0
 "CA" "CT" 0 -1 0.0
 "HA" "H1" 0 -1 0.0
 "CB" "CT" 0 -1 0.0
 "HB2" "HC" 0 -1 0.0
 "HB3" "HC" 0 -1 0.0
 "CG" "CT" 0 -1 0.0
 "HG2" "HC" 0 -1 0.0
 "HG3" "HC" 0 -1 0.0
 "CD" "CT" 0 -1 0.0
 "HD2" "H1" 0 -1 0.0
 "HD3" "H1" 0 -1 0.0
 "NE" "N2" 0 -1 0.0
 "HE" "H" 0 -1 0.0
 "CZ" "CA" 0 -1 0.0
 "NH1" "N2" 0 -1 0.0
 "NH2" "N2" 0 -1 0.0
 "HH2" "H" 0 -1 0.0
 "C" "C" 0 -1 0.0
 "O" "O" 0 -1 0.0
!entry.ARD.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.ARD.unit.childsequence single int
 2
!entry.ARD.unit.connect array int
 1
 20
 17
 18
!entry.ARD.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 3 4 1
 3 5 1
 3 20 1
 5 6 1
```

```
 5 7 1
 5 8 1
 8 9 1
 8 10 1
 8 11 1
 11 12 1
 11 13 1
 11 14 1
 14 15 1
 14 16 1
 16 17 1
 16 18 1
 18 19 1
 20 21 2
!entry.ARD.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
 "R" 1 "A" 17
 "R" 1 "A" 18
 "R" 1 "A" 19
 "R" 1 "A" 20
 "R" 1 "A" 21
!entry.ARD.unit.name single str
 "ARD"
!entry.ARD.unit.positions table  dbl x  dbl y  dbl z
 175.243000 44.272000 1.630719E+03
 175.388700 44.212300 1.629711E+03
 176.383000 44.700000 1.631316E+03
 176.631000 45.665000 1.630926E+03
 177.747400 44.004900 1.631152E+03
 178.141600 43.715300 1.632135E+03
 178.446000 44.782300 1.630798E+03
 177.889500 42.825400 1.630196E+03
 177.439800 41.939000 1.630656E+03
 177.346500 43.018700 1.629264E+03
 179.388000 42.542200 1.629917E+03
 179.616700 41.478000 1.630054E+03
 180.000700 43.056600 1.630670E+03
 179.825500 43.023500 1.628614E+03
 180.798100 43.315700 1.628607E+03
 179.675300 42.084900 1.627561E+03
 178.617700 41.135000 1.627494E+03
 180.613300 42.085500 1.626625E+03
 181.263200 42.850100 1.626828E+03
 176.276000 45.358000 1.632576E+03
 176.211000 45.813000 1.633679E+03
!entry.ARD.unit.residueconnect table  int c1x  int c2x  int c3x  int c4x  int c5x  int c6x
 1 20 17 18 0 0
!entry.ARD.unit.residues table  str name  int seq  int childseq  int startatomx  str restype  int imagingx
 "ARD" 1 22 1 "?" 0
!entry.ARD.unit.residuesPdbSequenceNumber array int
 0
!entry.ARD.unit.solventcap array dbl
 -1.000000
```

```
 0.0
 0.0
 0.0
 0.0
!entry.ARD.unit.velocities table  dbl x  dbl y  dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
```

```
!!index array str
 "LYD"
!entry.LYD.unit.atoms table  str name  str type  int typex  int resx  int flags  int seq  int elmnt  dbl chg
 "N" "N" 0 1 131072 1 7 -0.347900
 "H" "H" 0 1 131072 2 1 0.274700
 "CA" "CT" 0 1 131072 3 6 -0.240000
 "HA" "H1" 0 1 131072 4 1 0.142600
 "CB" "CT" 0 1 131072 5 6 -0.009400
 "HB2" "HC" 0 1 131072 6 1 0.036200
 "HB3" "HC" 0 1 131072 7 1 0.036200
 "CG" "CT" 0 1 131072 8 6 0.018700
 "HG2" "HC" 0 1 131072 9 1 0.010300
 "HG3" "HC" 0 1 131072 10 1 0.010300
 "CD" "CT" 0 1 131072 11 6 -0.047900
 "HD2" "HC" 0 1 131072 12 1 0.062100
 "HD3" "HC" 0 1 131072 13 1 0.062100
 "CE" "CT" 0 1 131072 14 6  0.154300
 "HE2" "H1" 0 1 131072 15 1 0.113500
 "HE3" "H1" 0 1 131072 16 1 0.113500
 "NZ" "N2" 0 1 131072 17 7 -0.585400
 "C" "C" 0 1 131072 18 6 0.734100
 "O" "O" 0 1 131072 19 8 -0.589400
!entry.LYD.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "N" "N" 0 -1 0.0
 "H" "H" 0 -1 0.0
 "CA" "CT" 0 -1 0.0
 "HA" "H1" 0 -1 0.0
 "CB" "CT" 0 -1 0.0
 "HB2" "HC" 0 -1 0.0
 "HB3" "HC" 0 -1 0.0
 "CG" "CT" 0 -1 0.0
 "HG2" "HC" 0 -1 0.0
 "HG3" "HC" 0 -1 0.0
 "CD" "CT" 0 -1 0.0
 "HD2" "HC" 0 -1 0.0
 "HD3" "HC" 0 -1 0.0
 "CE" "CT" 0 -1 0.0
```

```
 "HE2" "H1" 0 -1 0.0
 "HE3" "H1" 0 -1 0.0
 "NZ" "N2" 0 -1 0.0
 "C" "C" 0 -1 0.0
 "O" "O" 0 -1 0.0
!entry.LYD.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.LYD.unit.childsequence single int
 2
!entry.LYD.unit.connect array int
 1
 18
 17
!entry.LYD.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 1 3 1
 3 4 1
 3 5 1
 3 18 1
 5 6 1
 5 7 1
 5 8 1
 8 9 1
 8 10 1
 8 11 1
 11 12 1
 11 13 1
 11 14 1
 14 15 1
 14 16 1
 14 17 1
 18 19 1
!entry.LYD.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
 "R" 1 "A" 17
 "R" 1 "A" 18
 "R" 1 "A" 19
!entry.LYD.unit.name single str
 "LYD"
!entry.LYD.unit.positions table  dbl x  dbl y  dbl z
 3.325770 1.547909 -1.607204E-06
 3.909407 0.723611 -2.739882E-06
 3.970048 2.845795 -1.311163E-07
 3.671663 3.400129 -0.889820
 3.576965 3.653838 1.232143
 2.496995 3.801075 1.241379
 3.877484 3.115795 2.131197
 4.274186 5.009602 1.194577
 5.354271 4.863178 1.185788
```

```
 3.973781 5.548460 0.295972
 3.881105 5.817645 2.426721
 2.801135 5.964881 2.435959
 4.181626 5.279602 3.325774
 4.578325 7.173410 2.389153
 5.658410 7.026987 2.380363
 4.277917 7.712267 1.490550
 4.199422 7.952309 3.576860
 5.485541 2.705207 -4.398755E-06
 6.008824 1.593175 -8.449768E-06
!entry.LYD.unit.residueconnect table  int c1x  int c2x  int c3x  int c4x  int c5x  int c6x
 1 18 17 0 0 0
!entry.LYD.unit.residues table  str name  int seq  int childseq  int startatomx  str restype  int imagingx
 "LYD" 1 20 1 "p" 0
!entry.LYD.unit.residuesPdbSequenceNumber array int
 0
!entry.LYD.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.LYD.unit.velocities table  dbl x  dbl y  dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0


!!index array str
 "DOG"
!entry.DOG.unit.atoms table  str name  str type  int typex  int resx  int flags  int seq  int elmnt  dbl chg
 "C1" "CC” 0 1 131072 1 6 0.596400
 "C2" "CT" 0 1 131072 2 6 0.467300
 "C3" "CT" 0 1 131072 3 6 0.045500
 "C4" "CT" 0 1 131072 4 6 0.070620
 "H2" "HC" 0 1 131072 5 1 0.033200
 "H3" "HC" 0 1 131072 6 1 0.033200
 "C5" "CT" 0 1 131072 7 6 0.106100
 "O1" "OH" 0 1 131072 8 8 -0.387700
 "H4" "H1" 0 1 131072 9 1 0.062100
 "C6" "CT" 0 1 131072 10 6 0.115600
 "O2" "OH" 0 1 131072 11 8 -0.387000
 "H6" "H1" 0 1 131072 12 1 0.064200
 "H8" "H1" 0 1 131072 13 1 0.058800
 "O3" "OH" 0 1 131072 14 8 -0.389400
 "H9" "H1" 0 1 131072 15 1 0.058800
 "H1" "H1" 0 1 131072 16 1 0.085900
 "H5" "HO" 0 1 131072 17 1 0.210400
 "H7" "HO" 0 1 131072 18 1 0.210800
```

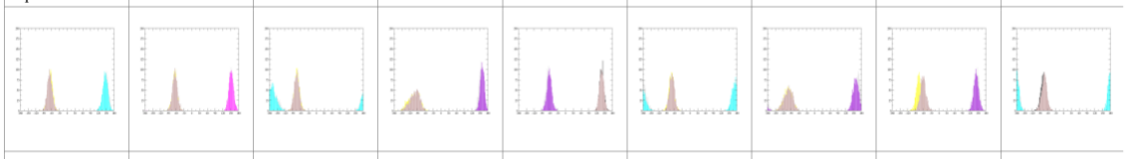
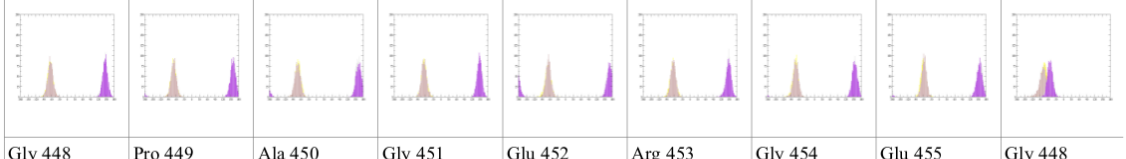
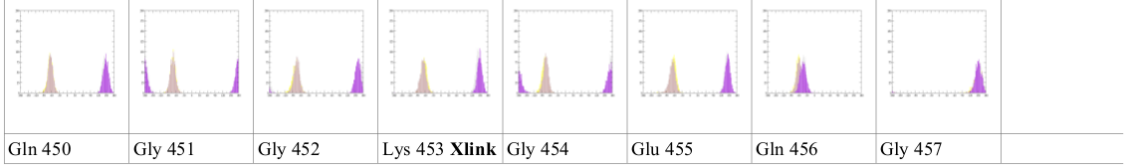
```
 "H10" "HO" 0 1 131072 19 1 0.21000
!entry.DOG.unit.atomspertinfo table  str pname  str ptype  int ptypex  int pelmnt  dbl pchg
 "C1" "CC" 0 -1 0.0
 "C2" "CT" 0 -1 0.0
 "C3" "CT" 0 -1 0.0
 "C4" "CT" 0 -1 0.0
 "H2" "HC" 0 -1 0.0
 "H3" "HC" 0 -1 0.0
 "C5" "CT" 0 -1 0.0
 "O1" "OH" 0 -1 0.0
 "H4" "H1" 0 -1 0.0
 "C6" "CT" 0 -1 0.0
 "O2" "OH" 0 -1 0.0
 "H6" "H1" 0 -1 0.0
 "H8" "H1" 0 -1 0.0
 "O3" "OH" 0 -1 0.0
 "H9" "H1" 0 -1 0.0
 "H1" "H1" 0 -1 0.0
 "H5" "HO" 0 -1 0.0
 "H7" "HO" 0 -1 0.0
 "H10" "HO" 0 -1 0.0
!entry.DOG.unit.boundbox array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.DOG.unit.childsequence single int
 2
!entry.DOG.unit.connect array int
 1
 2
!entry.DOG.unit.connectivity table  int atom1x  int atom2x  int flags
 1 2 1
 2 3 1
 2 16 1
 3 4 1
 3 5 1
 3 6 1
 4 7 1
 4 8 1
 4 9 1
 7 10 1
 7 11 1
 7 12 1
 8 17 1
 10 13 1
 10 14 1
 10 15 1
 11 18 1
 14 19 1
!entry.DOG.unit.hierarchy table  str abovetype  int abovex  str belowtype  int belowx
 "U" 0 "R" 1
 "R" 1 "A" 1
 "R" 1 "A" 2
 "R" 1 "A" 3
 "R" 1 "A" 4
 "R" 1 "A" 5
 "R" 1 "A" 6
 "R" 1 "A" 7
 "R" 1 "A" 8
 "R" 1 "A" 9
 "R" 1 "A" 10
 "R" 1 "A" 11
 "R" 1 "A" 12
 "R" 1 "A" 13
 "R" 1 "A" 14
 "R" 1 "A" 15
 "R" 1 "A" 16
```

```
 "R" 1 "A" 17
 "R" 1 "A" 18
 "R" 1 "A" 19
!entry.DOG.unit.name single str
 "DOG"
!entry.DOG.unit.positions table  dbl x  dbl y  dbl z
 -1.849500 -1.936100 0.221000
 -1.057100 -0.683500 0.630000
 0.316200 -0.621600 -0.033900
 1.165800 0.524600 0.516000
 0.210200 -0.518300 -1.121000
 0.827800 -1.572800 0.151000
 2.560200 0.625100 -0.128100
 1.339800 0.346300 1.929700
 0.643100 1.479500 0.382400
 3.435700 -0.617300 0.084200
 3.243000 1.740800 0.482600
 2.479200 0.849600 -1.197600
 3.533500 -0.858700 1.147300
 4.750400 -0.331500 -0.408400
 3.046200 -1.484300 -0.455900
 -1.004500 -0.648000 1.723400
 1.905800 1.100000 2.197500
 4.174100 1.639200 0.193800
 5.297700 -1.110700 -0.202100
!entry.DOG.unit.residueconnect table  int c1x  int c2x  int c3x  int c4x  int c5x  int c6x
 1 2 0 0 0 0
!entry.DOG.unit.residues table  str name  int seq  int childseq  int startatomx  str restype  int imagingx
 "DOG" 1 20 1 "?" 0
!entry.DOG.unit.residuesPdbSequenceNumber array int
 0
!entry.DOG.unit.solventcap array dbl
 -1.000000
 0.0
 0.0
 0.0
 0.0
!entry.DOG.unit.velocities table  dbl x  dbl y  dbl z
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
 0.0 0.0 0.0
```

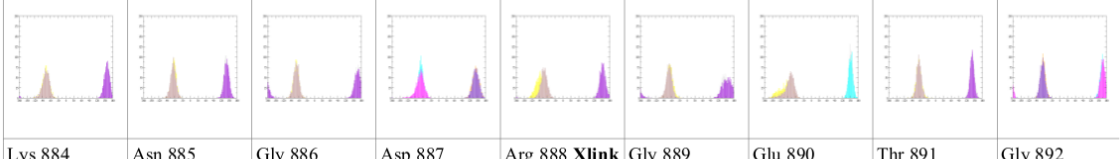
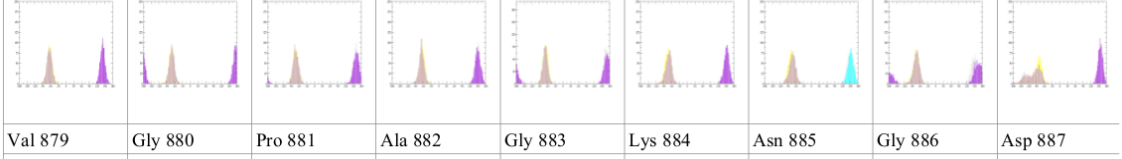
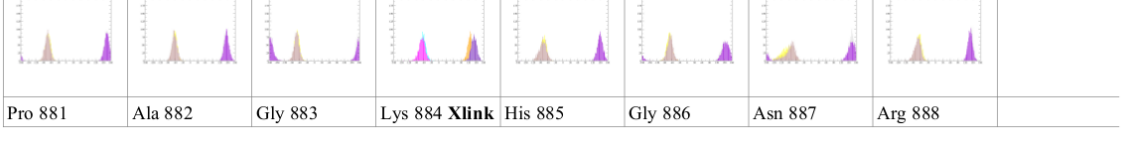
Appendix 2

Frequency histograms for the two dihedrals angles ϕ and ψ in the cross-linking and surrounding residues, for the cross-link removed and the glucosepane cross-linked systems. (Colours: $\phi_{Removed}$ – Pink; ϕ_{Cross} – Blue; $\psi_{Removed}$ – Purple; ψ_{cross} – Yellow)

Site 9 – Unfavourable +76.636kcal/mol

Alpha1 a								
								
Pro 449	Ala 450	Gly 451	Gly 452	Arg 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457
Alpha 1 b								
								
Gly 448	Pro 449	Ala 450	Gly 451	Glu 452	Arg 453	Gly 454	Glu 455	Gly 448
Alpha2								
								
Gln 450	Gly 451	Gly 452	Lys 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457	

Site 20 – Favourable -34.501kcal/mol

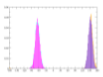
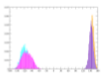
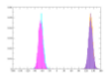
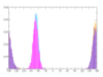
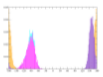
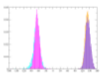
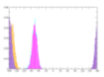
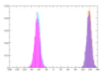
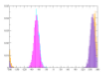
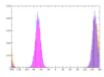
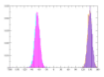
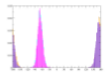
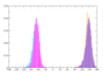
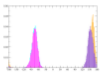
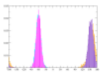
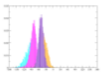
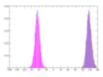
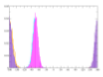
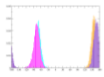
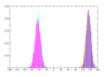
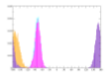
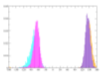
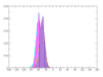
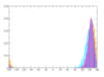
Alpha 1a								
								
Lys 884	Asn 885	Gly 886	Asp 887	Arg 888 **Xlink**	Gly 889	Glu 890	Thr 891	Gly 892
Alpha 1b								
								
Val 879	Gly 880	Pro 881	Ala 882	Gly 883	Lys 884	Asn 885	Gly 886	Asp 887
Alpha 2								
								
Pro 881	Ala 882	Gly 883	Lys 884 **Xlink**	His 885	Gly 886	Asn 887	Arg 888	

Appendix 3

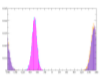
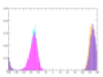
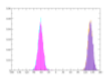
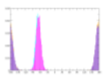
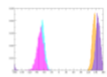
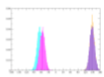
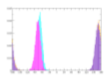
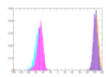
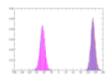
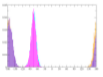
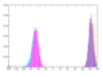
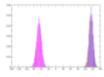
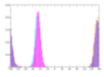
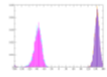
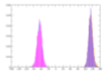
9.3 Glucosepane Dihedral Plot

Frequency histograms for the two dihedrals angles ϕ and ψ in the cross-linking and surrounding residues, for the native and the glucosepane cross-linked systems. (Colours: ϕ_{Nat} – Pink; ϕ_{Cross} – Blue; ψ_{Nat} – Purple; ψ_{cross} – Yellow)

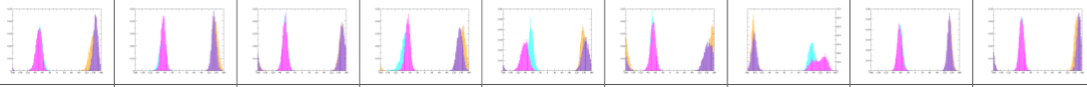
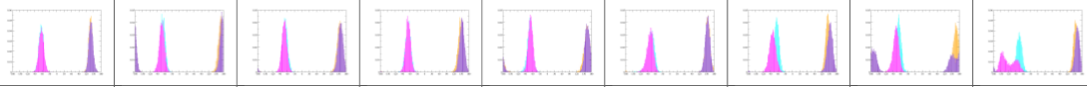
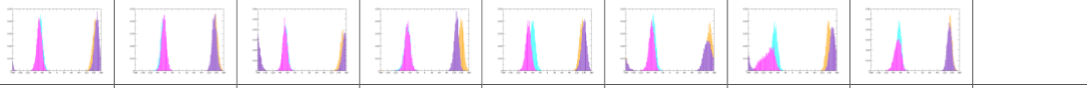
Site 9 – Unfavourable +76.636kcal/mol

Alpha 1a								
								
Pro 449	Ala 450	Gly 451	Gly 452	Arg 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457
Alpha 1b								
									
Gly 448	Pro 449	Ala 450	Gly 451	Glu 452	Arg 453	Gly 454	Glu 455	Gly 448
Alpha 2								
									
Gln 450	Gly 451	Gly 452	Lys 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457	

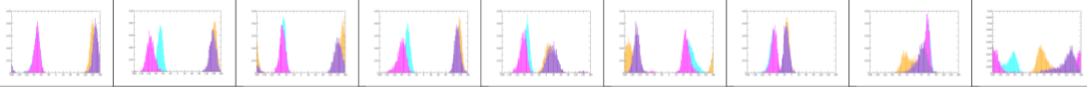
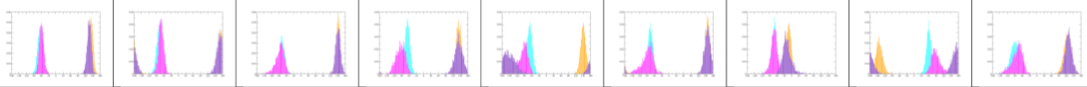
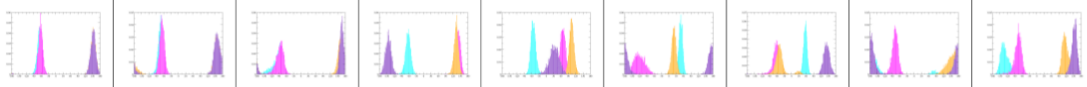
Site 19 – Unfavourable +16.130kcal/mol

Alpha 1a								
									
Gly 853	Ala 854	Lys 855	Gly 856	Asp 857	Arg 858 **Xlink**	Gly 859	Glu 860	Thr 861
Alpha 1b								
									
Gly 853	Ala 854	Lys 855 **Xlink**	Gly 856	Asp 857	Arg 858	Gly 859	Glu 860	Thr 861
Alpha 2								
Gly 853	His 854	Lys 855	Gly 856	Glu 857	Arg 858	Gly 859	Tyr 860	Hyp 861

Site 20 – Favourable -34.501kcal/mol

Alpha 1a								
								
Lys 884	Asn 885	Gly 886	Asp 887	Arg 888 **Xlink**	Gly 889	Glu 890	Thr 891	Gly 892
Alpha 1b								
								
Val 879	Gly 880	Pro 881	Ala 882	Gly 883	Lys 884	Asn 885	Gly 886	Asp 887
Alpha 2								
								
Pro 881	Ala 882	Gly 883	Lys 884 **Xlink**	His 885	Gly 886	Asn 887	Arg 888	

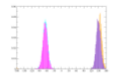
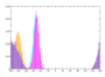
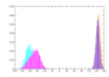
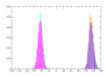
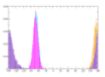
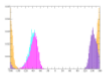
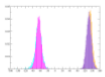
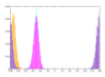
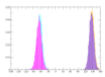
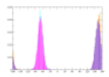
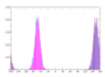
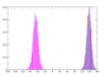
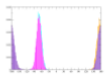
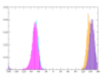
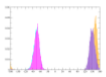
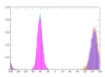
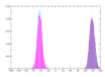
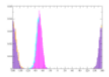
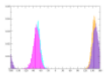
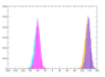
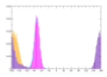
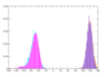
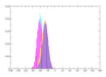
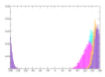
Site 22 – Favourable -36.130kcal/mol

Alpha 1a								
								
Glu 923	Gln 924	Gly 925	Asp 926	Arg 927 **Xlink**	Gly 928	Ile 929	Lys 930	Gly 931
Alpha 1b								
								
Thr 921	Gly 922	Glu 923	Gln 924	Gly 925	Asp 926	Arg 927	Gly 928	Ile 929
Alpha 2								
								
Hyp 921	Gly 922	Asp 923	Lys 924 **Xlink**	Gly 925	Ala 926	Arg 927	Gly 928	Leu 929

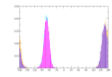
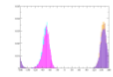
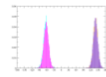
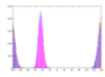
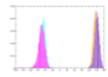
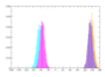
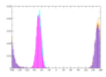
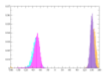
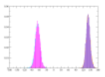
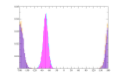
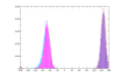
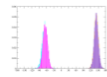
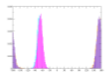
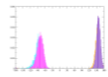
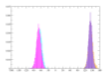
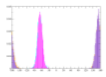
9.4 DOGDIC Dihedral Plots

Frequency histograms for the two dihedrals angles ϕ and ψ in the cross-linking and surrounding residues, for the native and the DOGDIC cross-linked systems. (Colours: ϕ_{Nat} – Pink; ϕ_{Cross} – Blue; ψ_{Nat} – Purple; ψ_{cross} – Yellow)

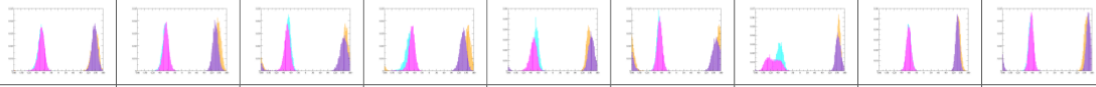
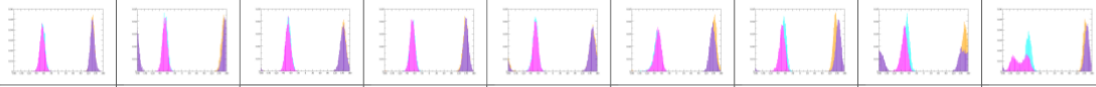
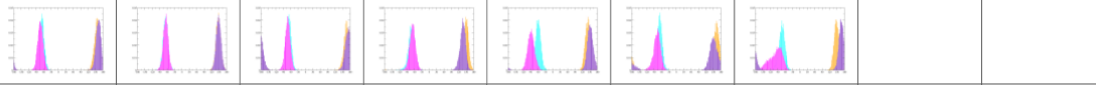
Site 9 – Unfavourable +55.07kcal/mol

Alpha 1a								
								
Pro 465	Ala 450	Gly 451	Gly 452	Arg 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457
Alpha 1b								
								
Ala 447	Gly 448	Pro 449	Ala 450	Gly 451	Glu 452	Arg 453	Gly 454	Glu 455
Alpha 2								
								
Gln 450	Gly 451	Gly 452	Lys 453 **Xlink**	Gly 454	Glu 455	Gln 456	Gly 457	

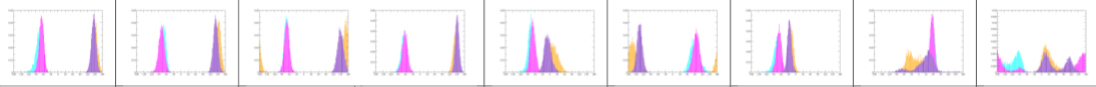
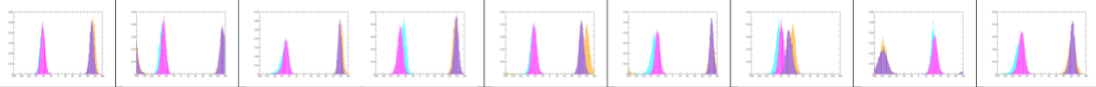
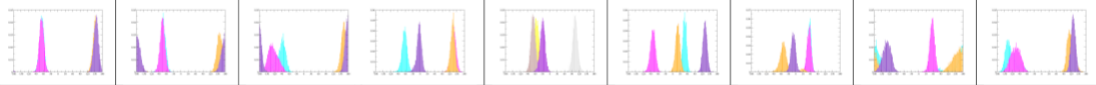
Site 19 – Favourable -61.5804kcal/mol

Alpha 1a								
									
Gly 853	Ala 854	Lys 855	Gly 856	Asp 857	Arg 858 **Xlink**	Gly 859	Glu 860	Thr 861
Alpha 1b								
									
Gly 853	Ala 854	Lys 855 **Xlink**	Gly 856	Asp 857	Arg 858	Gly 859	Glu 860	Thr 861
Alpha 2								
Gly 853	His 854	Lys 855	Gly 856	Glu 857	Arg 858	Gly 859	Tyr 860	Hyp 861

Site 20 – Favourable -4.853kcal/mol

Alpha 1a								
								
Lys 884	Asn 885	Gly 886	Asp 887	Arg 888 **Xlink**	Gly 889	Glu 890	Thr 891	Gly 892
Alpha 1b								
								
Val 879	Gly 880	Pro 881	Ala 882	Gly 883	Lys 884	Asn 885	Gly 886	Asp 887
Alpha 2								
								
Pro 881	Ala 882	Gly 883	Lys 884 **Xlink**	His 885	Gly 886	Asn 887	Arg 888	

Site 22 – Unfavourable + 28.1477kcal/mol

Alpha 1a								
								
Glu 923	Gln 924	Gly 925	Asp 926	Arg 927 **Xlink**	Gly 928	Ile 929	Lys 930	Gly 931
Alpha 1b								
								
Thr 921	Gly 922	Glu 923	Gln 924	Gly 925	Asp 926	Arg 927	Gly 928	Ile 929
Alpha 2								
								
Hyp 921	Gly 922	Asp 923	Lys 924 **Xlink**	Gly 925	Ala 926	Arg 927	Gly 928	Leu 929