UNIVERSITÄT LEIPZIG
Faculty of Physics and Earth System Sciences
International Physics Study Program

Influence of Helicity on the Collapse Transition of Polymers

Investigation on Alanine and Glycine Homopolymers and Copolymers

Nicolò De Masi

Bachelor's Thesis

Under the guidance of:
M.Sc. Maximilian Conradi
Prof. Dr. Wolfhard Janke

Bachelor of Science in Physics
Leipzig, 2024

Abstract

The phase in which newly formed proteins self-restructure into well-behaving functional molecules is called protein folding, and is a complex process governed by a multitude of factors. Among these is the formation of secondary structure elements. By means of molecular dynamics (MD) simulations, the collapse transitions of polyalanine, polyglycine, and various heteropolymers made of alanine and glycine are showcased. Trends in the increase of the collapse time τ_c due to the formation of secondary structure helical elements are observed. The influence of α-helices in the folding process is particularly taken into account. The study of τ_c was carried out by analysis of the mean squared radius of gyration $\langle R_g^2 \rangle$, which showed a longer folding process for molecules where alanine was more than 50 % of the total composition. However, no significant increase was observed for the remaining molecules. This is in contrast to the intuitive idea of a linear dependence between the presence of alanine and τ_c, and suggests more complex folding mechanisms due to the exact sequencing of amino acids instead. Qualitative observations and quantitative analysis suggest that the rate of appearance of α-helices follows quite closely the behavior of the extrapolated collapse times and correlates with them. However, further analysis is required to determine the precise extent of such correlations.

Contents

1 Theoretical Background

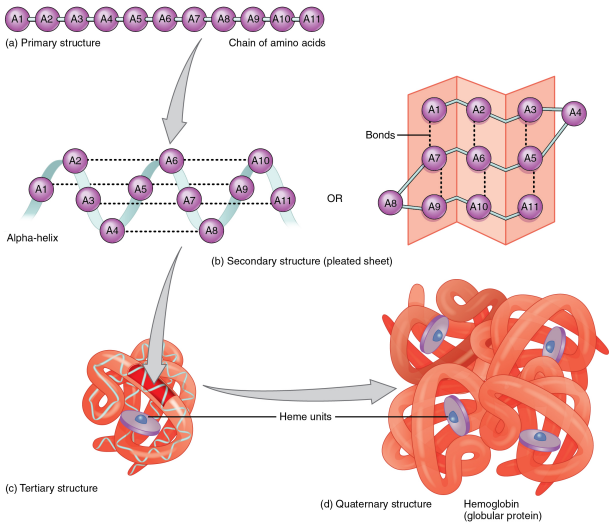


Figure 1: Different levels of the structure of an Hemoglobin protein. [1]

Often referred to as the multitools of biological organisms, proteins serve numerous crucial roles, such as providing structural stability, acting as motors for movement, carrying out metabolic activities, and participating in the expression of genetic information. Recent studies even found polypedtides that remarkably allow the survival of organisms in subzero conditions by reducing the freezing point of water and avoiding the growth of ice crystals [2].

Each of the hundreds of thousands of different proteins is originally synthesized by a ribosome as a long ordered chain of amino acids, with covalent bonds connecting near neighbors tightly. In order to serve their purpose, these must arrange in the specific typical shapes which define their function. This naturally takes place immediately after syntheses, often even initiating while the nascent chain is still attached to the ribosome [3], and is called protein folding. Throughout this folding stage, linear chains of amino acids quickly self-organize into well-behaving and functional three-dimensional structures, due to the interactions between non bonded particles as well as between the molecule and the solvent.

To properly handle the increasing complexity of a protein, different levels of organization exist, as Fig. 1 above illustrates for hemoglobin. In its most fundamental form, the primary structure is delineated, specifying the precise sequence of amino acids. This is followed by three additional levels of structural description that elucidate the self-organizational behavior of these molecules [4]. The secondary structure of an amino acid chain is characterized by its local configurations, manifesting as small units that, while repeating inconsistently, do so in a regular manner. These will be of particular interest in this thesis. Tertiary and quaternary structures, ultimately, delineate the spatial conformation of an individual peptide chain and the assembly of a polypeptide complex composed of multiple chains. The synthesis of such intricate molecules exemplifies a

remarkable biological phenomenon that highlights the asymmetrical complexity inherent in biological systems, continually unveiling novel aspects of them [5].

1.1 Protein Folding

Protein folding has been central to the interests of the scientific community for more than 50 years, ever since the work of Max Perutz and John Kendrew in the 1960s, which took them two decades but gained them a Nobel price. Their research, by means of X-ray crystallography, revealed the 3D structures of hemoglobin and myoglobin, marking them the first solved structures of globular proteins. It demonstrated that these molecules adopt specific three-dimensional conformations essential for their functions, which spurred questions about how and why these structures form. This interest led to the pioneering work of Anfinsen in 1973, which established the now well-known thermodynamic hypothesis (or Anfinsen's dogma) [6]. Anfisen showed that proteins, under certain physiological conditions, spontaneously fold into their functional structures. In his work, he laid down the basic properties of a protein's native structure: (i) uniqueness, (ii) kinetic and thermodynamic stability, and (iii) uniquely determined by its amino acid sequence. Anfisen therefore hypothesized that if the initial composition of the chain is known, its final state should be determinable as well, establishing it as a global minimum of the free-energy landscape.
This came to be known as the protein folding problem.

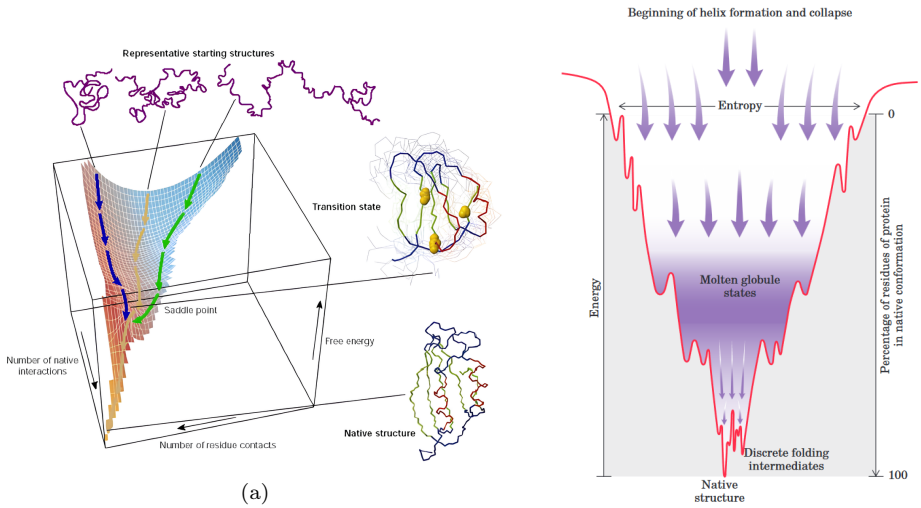


Figure 2: (a) Illustration of the folding process of the same small protein starting from three different random initial configurations. Each configuration is tunneled toward a common saddle point, after which their path coincide. In yellow, the key residues of the structure are highlighted, which formation leads to the contraction of the protein towards the final native fold. (b) Depiction of the funnel-like energy landscape of protein folding. The number of possible states progressively diminishes, as such reducing conformational entropy.

However, the way in which a protein identifies its unique minimum energy configuration is remarkable. Any biological algorithm purely cycling through every possible state is bound to take a very long time, given the huge number of possibilities. This does not reflect the reality of the phenomena; proteins fold extremely rapidly. For example, E. coli cells can produce complete, biologically active protein molecules containing hundreds of amino acid residues in mere seconds at body temperature [7]. The latter is conceptually known as the Levinthal paradox. Fundamentally, it excludes the idea that proteins fold by completely random trial-and-error processes; transitions must move through and be guided by uneven energy landscapes instead. This is represented by a stochastic search for the possible accessible conformations toward a saddle point (Fig. 2 (a)) [8]. From a thermodynamical perspective, unfolded proteins have high conformational entropy, which determines the number of configurations that can be branched out to, and relatively high free energy, hence highly unstable and easily driven by forces and interactions. Both progressively decrease throughout the folding process, with local plateaus representing semi-stable configurations (Fig. 2 (b)).

From its beginning as purely experimentally tackled, the protein folding problem transitioned to computational approaches. However, the questions remained unchanged: How does an amino acid sequence dictate a protein native structure? Is there a folding code a protein follows? What determines different folding mechanisms? Although general approaches to the 'solution' of protein folding have been explored, most recently with the groundbreaking advancements made by AlphaFold in the prediction of millions of native structures by deep learning methods [9], much of the correlation between the initial composition of a chain and the resulting folding pathways remains still unknown. For those answers, it is vital to pull proteins apart, look at single local arrangements, understand the conditions under which they manifest, and the ramification of their presence.

Forces governing protein folding

Although amino acids in a polypeptide are linked by covalent bonding forces, a protein's folded structure is determined by multiple non-bonding interactions, none of which can be considered dominant. However, a series of primary contributors can be identified, each playing a different role, and of great importance given the marginal instability of a protein.

- Hydrophobic interactions: When considering a polar solvent, such as surrounding water molecules, hydrophobic interactions lead the protein to fold into well-packed states with predominantly hydrophobic (H) amino acids at the core, and polar (P) amino acids on the folded protein surface. For this reason, hydrophobic interactions represent a good indicator of the compactness of a protein and a type of solvent-induced forces, as opposed to direct interaction between amino acids [10]. Hydrophobic interaction emerged as particularly dominant in the folding process in the 1980s, with the advent of statistical mechanical modeling, which allowed to simulate solvent induced interactions for the first time and showed how a protein's secondary structure is as much a consequence of the tertiary structure as a cause of it [11].

- Electrostatic interactions(charge-charge, charge-dipole, dipole-dipole): Cou-

lomb forces are the way by which some amino acids attract or repel based on their charge. Early models of protein folding [12] viewed ion bonding interactions as the main engine of the process, but this was soon dispelled [13]: they are concentrated in high-dielectric regions on the protein surface but play a lesser role elsewhere.

- Van der Waals interactions: Amino acids are tightly packed, which means that weak attractive and repulsive forces arise within a folded protein. They are essential to the short-range stability of the chain .

- Hydrogen bonding: Whenever an hydrogen atom is partially shared between a donor and an acceptor with a lone electron pair, a hydrogen bond is formed. In proteins, hydrogen bonds form both between bonding groups within the polymer, as well as between a bonding group and the water molecules surrounding it, which influences the overall stability of the protein [14]. Hydrogen bonds are of particular interests, as they are the main determinant of the differences in secondary structure. This will be covered in extensive detail in the next section.

In the last decades, secondary contributions have been identified as well, such as aromatic rings interactions [15]. These are either (i) weak but abundant interactions in the protein main chain, or (ii) strong but less frequent interactions in side chains.

The Secondary Structure

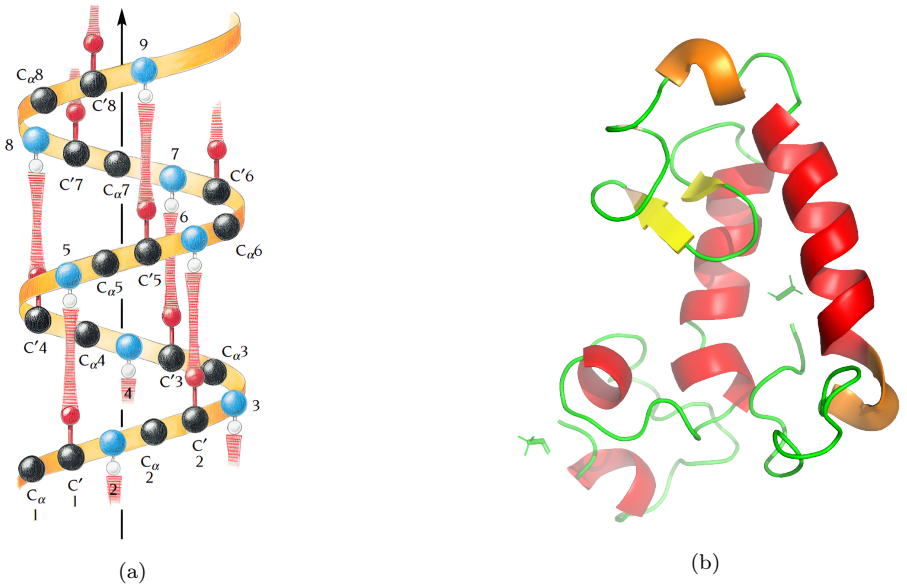
The formation of a secondary structure is the first step in the folding process. As amino acids fold closer and closer, strong hydrogen bonds establish between non consecutive monomers, with residues locally arranging themselves into well known spacial conformations as a result. These repetition units are called secondary structural elements and are of great importance in determining a protein's final native structure.
The most commonly occurring secondary structure, and central to this thesis, are α-helices, ribbon-like helical conformations [16]. The α-helix is the classic element of protein structure. Its presence was first predicted in 1951 by Linus Pauling, working at the California Institute of Technology, on the basis of geometrical calculations on results of crystallographic analyses of small molecules. It was described as a stable and energetically favorable structure and soon received strong experimental support from diffraction patterns that confirmed it. Each right handed turn of the helix holds 3.6 residues (each a 100 degrees turn) and a translation of 1.5 Å along the helical axis (Fig. 3 (a)). This structure element is the result of a hydrogen bond forming between the carbonyl (C=O) group of an amino acid and the amino H (N-H) group of an amino acid four positions down the chain. The prevalence of α-helices is often attributed to the optimal use they make of internal hydrogen bonds [7]: every peptide bond (except those close to the ends of the helix) participates in such bonding. Each successive turn of the helix is held to adjacent ones by three to four hydrogen bonds, which earns the structure considerable stability.

Bonding between residues three positions apart is also frequent, and creates so called 3_{10}-helices. This secondary structure element has often been overshadowed by the considerably more common α-helices, but they have gained more attention in recent years. While α-helices show a natural predilection toward a length of

nine to seventeen residues [17], which maximizes their stability, 3_{10}-helices tend to be shorter, comprised of three to five residues and characterized by smaller bonding angles [18]. They are commonly found either as extensions of an α-helix on their N- or C-terminal (as it is the case in Fig. 3 (b)) [19], or observed as intermediates in the transition between helical and non-helical states [20]. This latter role is justified by the idea that a shorter looped structure is less entropically penalizing for the molecule. The propensity towards 3_{10}-helices, as opposed to α-helices, is noticeably higher for globular proteins, as well as for shorter peptide sequences, due to the lack of side-chains interactions. For longer segments, however, α-helices dominate, and greatly influence their folding dynamic and, possibly, the speed of their collapse.

Another common family of elements, β-sheets, are the result of two (or more) different segments of the peptide chain, singularly called β-strands, held close by a series of at least two hydrogen bonds. The structure will then resemble a crumpled up sheet-like structure. Helices and sheets are connected by turns and loops, although they themselves are made of less frequently observed structural elements.

In the realm of simulated protein chains, secondary structure identification is most often managed by the database of secondary structure assignments DSSP, which calculates the most likely secondary structure by reading the position of the atoms in a protein followed by calculation of the bond energy between them [21, 22]. In the context of helical elements, the most relevant information which characterizes different helices from one another are the backbone torsion angles ϕ and ψ. α-helices present angles ranging from (-90°, -15°) to (-70°, -35°), while 3_{10}-helices are around (-49°, -26°).



(a) (b)

Figure 3: (a) From [16]. Idealized diagram of an α-helix with approximate positions for main-chain atoms and with hydrogen bonds included. The arrow denotes the direction from the N-terminus to the C-terminus. (b) A polyalanine molecule in a folded state. α-helices are red, 3_{10}-helices are orange. Yellow arrows represent $\beta-$sheets.

Homopolymers and Heteropolymers

Protein chains can additionally be grouped based on the different types of amino acids they contain. Molecules with exclusively one type of monomer are called homopolymers:

$$... - A - A - A - A - A - A - A - A - A - A - ...$$

On the other hand, combining different types of monomers in a single chain leads to so-called heteropolymers. The properties of a heteropolymer depend on its composition as well as the exact sequencing of monomers on the chain.

For copolymers, i.e. second degree heteropolymers, the sequencing can lead to the creation of: (i) Alternating copolymers, when the two species alternate evenly. (ii) Statistical copolymers, when the placement of monomers follows well-known statistical distributions (or random copolymers when such distribution is unclear). (iii) Block copolymers, constituted of two (diblock), three (triblock), or more (multiblock) homopolymer sub-units linked by covalent bonds and (iv) periodic copolymers, constructed of sub-units arranged in a repeating sequence. These represent most of the models used in this thesis.

1.2 Structural and Behavioral Differences of Glycine and Alanine

In the following, glycine and alanine are dealt with exclusively. Glycine is the simplest and smallest amino acid, with the amine group and the carboaxylic acid group, characteristic of amino acids, bonding directly to the central carbon atom. However, glycine is considered a poor α-helix former. This is likely due to the flexibility given by the small residue group, made of one single hydrogen atom, which allows for more compact configurations to emerge. On the other hand, protein chains comprised of alanine, which displays a larger $-CH_3$ residue, show dominant presence of helical elements, due to so called large helix-stabilizing propensity [23].

Displaying the preferences of secondary structures for each amino acid is possible by means of Ramachandran plotting. This technique consists in mapping the allowed values of the ϕ and ψ angles for a specific composition and configuration of the chain, based on the avoidance of steric collisions. Ramachandran et al. showed that the physically allowed angle combinations correspond largely to the secondary structures observed in proteins [24]. Recent analysis of the conformational propensities of the two amino acids displayed a slightly denser distribution of points in the Ramachandran map space associated to α-helices for alanine than for glycine [25]. The results were obtained by all-atom molecular dynamics simulations with an explicit solvent of different iterations of Ala-X-Ala and Gly-Gly-X-Gly-Gly molecules, where X represents either Ala or Gly. The backbone conformations differences between Ala and Gly were analyzed (Fig. 4), with the results being that the percentage of residues adopting helical conformations for 20 % or less of the simulation time was around 61 % for glycine and substantially lower for alanine (38 % of residues). Glycine, on the other hand, consistently showed a wider distribution between the two. Due to the reduced size, the steric hindrance will always be lower and as such, more varied angle values will be recorded. Despite the small differences, the authors, as well as other

similar studies [26] highlighted how packing constraints, solvation, and general folding circumstances often overwhelm intrinsic conformational preferences.

While that may be true for more complex molecules, homogeneous polymers appear to conform to the tendency of their building blocks. The behavior of polyglycine and polyalanine has been previously explored and has well-known scaling laws with respect to the length of the chain [27–29]. The predilection of certain secondary structural elements is visible at every step of the folding process and evidently influences the obtained final native structure. Polyglycine collapses to a globular state: a roughly spherical structure with the predominant presence of β-sheets and β-strands, as well as short 3_{10}-helices. On the other hand, polyalanine exhibits a helical native state: a locally bonded structure with a less cohesive central core and the significant presence of α-helices. Different native states entail different folding processes and speed, with a recent study on polyglycine in particular observing a significantly faster transition time than expected from coarse-grained models [30]. As such, the influence of secondary structural elements in the folding process is taken under observation by means of molecular dynamics (MD) simulation and statistical analysis of the results.

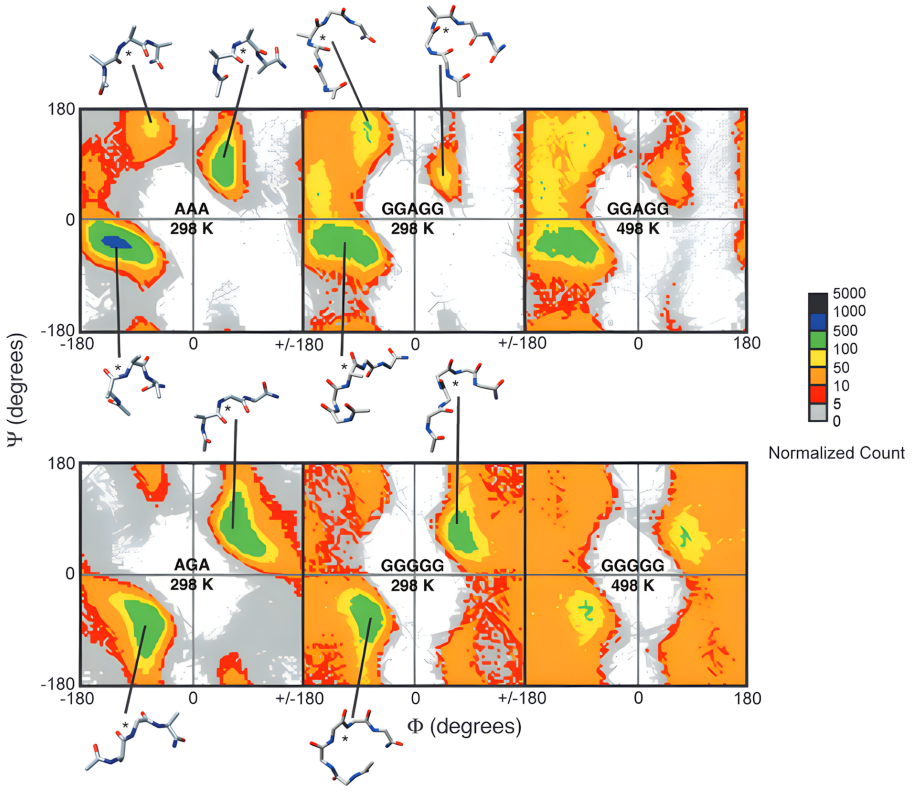


Figure 4: Ramachandran plots for aforementioned models at two different temperatures 298 K and 498 K. The plot is colored according to the population of each bin. The bottom left clustering is indicative of angle values corresponding to α-helix formation and is consistently denser for alanine. Meanwhile, glycine displays more coverage of the overall range of angles. From [25].

2 Models & Method

2.1 Simulation Models

The 11 molecules that will be investigated are the following:

Composition	Sequence	Type of Polymer
100% G	$[G]_{100}$	Homopolymer
90% G - 10% A	$[GGGGAGGGGG]_{10}$	Periodic copolymer
80% G - 20% A	$[GGAGG]_{20}$	Periodic copolymer
70% G - 30% A	$[GGAGGAGGAG]_{10}$	Periodic copolymer
60% G - 40% A	$[GAGAG]_{20}$	Periodic copolymer
50% G - 50% A	$[GA]_{50}$	Alternating copolymer
40% G - 60% A	$[AGAGA]_{20}$	Periodic copolymer
30% G - 70% A	$[AAGAAGAAGA]_{10}$	Periodic copolymer
20% G - 80% A	$[AAGAA]_{20}$	Periodic copolymer
10% G - 90% A	$[AAAAGAAAAA]_{10}$	Periodic copolymer
100% A	$[A]_{100}$	Homopolymer

The molecules, constructed by means of the PyMol software, are exclusively made up of the amino acids alanine (A), and glycine (G) and will be referred to with the shortened notation AnGm, where n,m represent two integers between 0 and 10. For example, the periodic colopymer $[AAAAGAAAAA]_{10}$ will be referred to as A9G1.

All of the models are capped using an amide group (NMA) at the C-terminal and an acetyl group (ACE) at the N terminal. These two terminals are defined by the bonding direction of the peptide chain, which will inevitably end on one side with a free amino group (NH_2), and on the other with a free carboxyl group (COOH). The presence of these caps is necessary in order to avoid strong interactions between the two ends, which would be overly represented due to the simulated chain being so short. The real life peptide chains are usually substantially longer and, as such, termini interaction plays less of a role.

2.2 Molecular Dynamics Simulations

Molecular dynamics (MD) simulations are a powerful computational tool that allows for a detailed study of folding trajectories. Although the process is entirely artificial, MD problem solving approaches are structured very similarly to physical experiments. First, a system is constructed, then, Newton's equations of motion for this modeled system are repeatedly solved throughout the non-equilibrium phase, and lastly, the necessary statistical analysis is performed on the results.

The software used for the MD simulations is the OpenMM package [32]. Here the PDB (Protein-Data-Bank) files, containing the topological information of all the atoms that make up each molecule, are acted upon by a force field. The force field used is Amber14 (Assisted Model Building with Energy Refinement), defined in terms of functional form of the following potential energy [33]:

$$
\begin{aligned}
V_{\text{total}} = & \sum_{\text{i} \in \text{ bonds}} k_{r,i}(r_i - r_i^0)^2 + \sum_{\text{i} \in \text{ angles}} k_{\theta,i}(\theta_i - \theta_i^0)^2 + \\
& + \sum_{\text{i} \in \text{ torsions}} \sum_{n} \frac{V_i^n}{2} \left[1 + \cos(n\phi_i - \gamma_i)\right] + \\
& + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} f_{ij} \left[\varepsilon_{ij} \left(\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right]
\end{aligned}
\tag{1}
$$

The first term represents the energy between covalently bonded atoms ($k_{r,i}$ are the bonds force constants), while the second term represents the energy due to the geometry of electron orbitals involved in covalent bonding ($k_{\theta,i}$ are the angle force constants). The third and fourth terms represent, in turn, the energy due to the twisting of bonds, and, the last one, the non-bonded energy sum of Van der Waals and electrostatic interactions between all atom pairs, respectively the Lennard-Jones (LJ) and Coulomb potentials.

When creating a force field, some explicit options can be specified. In this case, a cutoff bonding interaction distance of 1 nm is established, at which the reaction field method is used to eliminate all interactions beyond it [34]. This truncation procedure assumes a finite-radius spherical cavity around each particle, within which the electrostatic interactions are calculated explicitly, while treating the region outside of the sphere as a dielectric continuum (illustrated in Fig. 5). This helps neglect parts of the long-range interactions to focus more on local structures. Both Coulomb and LJ potentials have an infinite range, but since they quickly decrease with distance, their negligible contribution from far away atoms are computationally redundant. This approach is not suited to every circumstance however, and may introduce significant simulation errors if mishandled.

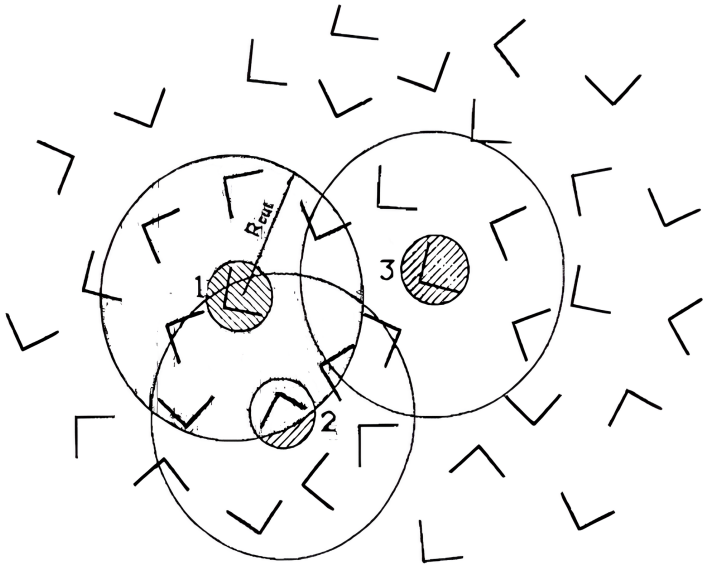


Figure 5: A schematic picture of a system of water molecules illustrating the basic idea behind the reaction field method. From [34].

Langevin Dynamics

Once the system has been established, the approach to solving it repeatedly needs to be implemented as well. This is done by defining an integrator function. An integrator is used to advance the equations of motion in time. At every iteration, the force field, as well as the current molecular topology information contained in the system, will be subject to the explicit calculations determined by the chosen integrator. Different integrators vary with respect to the algorithmic approach they take to solve the equations of motion. The one used in these pages is the Langevin integrator, which models itself on the basis of Langevin dynamics.

The assumption behind Langevin dynamics is that since molecular systems are present within a gaseous or liquid solvent instead than a vacuum, collisions with other particles need to always be accounted for. When these interactions are prevalent, Brownian motion takes place, with a particle's seemingly random trajectory guided by a fluctuating component and balanced out by the frictional forces due to the solvent's viscosity. At its simplest, the Langevin equation at the core of this theory can be described by:

$$m_i \frac{\mathrm{d}^2 \mathbf{x}_i}{\mathrm{d}t^2} = -\gamma m_i \frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t} + \mathbf{R}_i(t) \tag{2}$$

In the above, for each i_{th} particle, γ represents the friction coefficient, given by Stokes Law, while $\mathbf{R}_i$ is a noise term that originates from the collisions between particles. $\mathbf{R}_i$'s intensity changes infinitesimally quick, as it is randomly determined; at any instant any numbers of collision could take place, or no collision at all. Thus the effects of the fluctuating force can be summarized by characterizing its first and second moments (mean and auto-correlation), as time averages over infinitesimal time intervals:

$$\langle \delta\mathbf{R}(t) \rangle = 0, \qquad \langle \delta\mathbf{R}(t)\delta\mathbf{R}(t') \rangle = 2B\,\delta(t-t') \tag{3}$$

B represents the magnitude of the fluctuation. The delta function indicates the independence between impacts at distinct time intervals. That, as well as the zero mean, indicates a Gaussian distributed fluctuating force.

The OpenMM Langevin Integator tool introduces a heat bath in communication to the system, with the equation of motion:

$$m_i \frac{\mathrm{d}\mathbf{v}_i}{\mathrm{d}t} = -\gamma m_i \mathbf{v}_i + \mathbf{f}_i + \mathbf{R}_i \tag{4}$$

Compared to eq. 2, OpenMM adds a generic term $\mathbf{f}_i$ representing all the forces from the force field adding upon the particle at any give time, while single R force contributions for $\mathbf{R}_i$ are accounted as random numbers generated as part of a normal distribution with mean zero and unit variance.

The integration is done using the Langevin leap-frog method, with each step updating the i_{th} particle's position and velocity in the following way (with $\alpha = \exp[-\gamma\Delta t]$) [35]:

$$\begin{aligned} \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) &= \alpha \mathbf{v}_i\left(t-\frac{\Delta t}{2}\right) + \mathbf{f}_i(t)\frac{1-\alpha}{\gamma m_i} + \sqrt{\frac{k_B T(1-\alpha^2)}{m_i R}} \\ \mathbf{x}_i\left(t+\Delta t\right) &= \mathbf{x}_i(t) + \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right)\Delta t \end{aligned} \tag{5}$$

For this analysis the values of the friction coefficient and of the time step size are $\gamma = 1$ ps^{-1} and $\Delta t = 0.001$ ps. ($\alpha \approx 0.999$) The temperature is initially set very high, at approximately $T_{\text{high}} = 2000K$.

2.3 Collapse Time and Radius of Gyration

To monitor the kinetics of a protein collapse process, and subsequently extrapolate the collapse time τ_c, one aims to observe the decay of the squared radius of gyration with respect to time.

R_g^2 is defined as the average square distance between monomers ($\mathbf{R}_i$) in a given conformation and the polymer's center of mass ($\mathbf{R}_{cm}$) [31]:

$$R_g^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{R}_i - \mathbf{R}_{cm})^2. \tag{6}$$

With the center of mass position vector defined as the average over all monomers position vectors:

$$\mathbf{R}_{cm} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{R}_j \tag{7}$$

Accounting for monomers with different masses, eq. 7 generalizes to:

$$\mathbf{R}_{cm} = \frac{\sum_{j=1}^{N} M_j\mathbf{R}_j}{\sum_{j=1}^{N} M_j} \tag{8}$$

The radius of gyration represents the mean range of motion of all the object in a system, around a single well-defined center. As such, it is expected to decrease in a folding process, where the protein collapses onto itself. The general decay of R_g^2 consistently behaves according to the following fit [30]:

$$R_g^2(t) = b_0 + b_1 \exp\left(-\left(\frac{t}{\tau_c}\right)^{\beta}\right) \tag{9}$$

b_0 represents $R_g^2(t \to \infty)$, i.e. the squared radius of gyration in the collapsed state. b_1 and β are associated nontrivial fitting parameters, while τ_c denotes the collapse time that is of interest.

An additional, more straightforward way to estimate decay times is to extrapolate the time when $R_g^2(t)$ has decayed to a certain collapse percentage p % of itself, i.e., $\Delta(R_g^2) = R_g^2(t=0) - Rg^2(t \to \infty)$ [29]. The exact collapse percentage p % chosen is not as important as the comparison between different polymers. In the literature, 50 % as well as 80 % have been both used before. $p_1 = 50\%$ and $p_2 = 80\%$ hence correspond to the collapse times: τ_{50} and τ_{80}.

2.4 Simulation Details

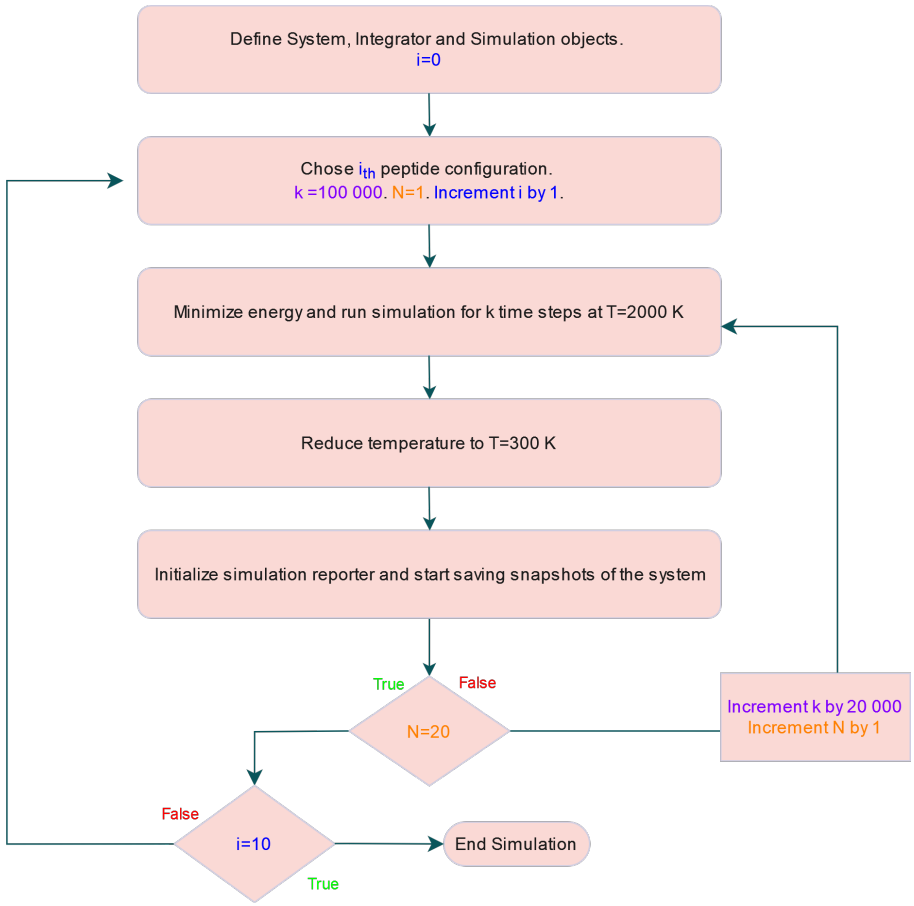


Figure 6: Diagram of the simulation. Each of the eleven molecule models goes through twenty simulations, for a total of 220 simulation cycles.

Fig. 6 depicts the different phases of the simulation. The first step is to initialize a simulation function, which collects the protein chain's topological information, the force field and the integrator. After performing a local energy minimization - which is usually a good idea at the start of a simulation to avoid producing too large forces - it is left running at T_{high} for k time steps. At $T >> T_{\text{crit}}$, every further iteration in time increases the disorder of the previous configuration, until a state is reached when the amino acid chain is unfolded in a seemingly random structure. This is often a non reversible process in nature, with the proteins unfolding to the point of denaturation, like an egg white turning grey when overly boiled. For mesophilic proteins, i.e. proteins that function in moderate temperature environments such as the human body, this critical functional temperature is between around 280 to 330 degrees Kelvin. [36] A minimum of k=100 000 steps for the high temperature phase was chosen, which corresponds to 0.1 nanoseconds of simulation time the protein will spend at 2000 K.

For each model 20 independent runs are performed. It is essential to avoid

auto-correlation errors and to make sure the initial unfolded states are independent from one another. To that end, for each new iteration of the same molecule, the k steps parameter is incremented by 20 000 (0.02 ns), which entails a maximum time spent at T_{high} of 0.48 ns (k=480 000). The high temperature phase is followed by a quenching, where the temperature is lowered to $T_{\text{quench}} = 300\,K \leq T_{\text{crit}}$. Arashiro et al. looked at the non-equilibrium evolution processes of an Ala_{40} chain with short-time Monte Carlo simulations and found a critical temperature $T_{\text{crit}} = 470K$ [37]. Although specific information on polyglycine are not as readily available as for polyalanine, Majumder et al. mention a rough collapse transition temperature of $T = 310$ K for glycine when dealing with chains of $[\text{Gly}]_N$, with N ranging from 20 to 200 residues [30]. Similar protein follow similar folding pathways, which suggest similar T_{crit} values as well. As every model used here is a combination of the same two amino acids, all the specific T_{crit} are likely to be found between 310 K and 470 K, well above T_{quench}. At such temperature, the proteins will surely undergo the folding process. The low temperature phase needs to be long enough for the protein to reach a stable configuration. For short chains and in the absence of solvent interaction, the folded state will be quickly reached, with 10^7 (10 nanoseconds) of simulation time being sufficient.

2.5 Data Analysis and Jackknife Resampling

The final results of the simulation are 20 PDB files for each of the 11 molecules. Each file contains the complete atomic position of 300 different frames of one iteration of a specific molecule. At each frame, by means of Eq. 6, the square radius of gyration is computed.

Since this observable is not a direct measurement of the simulation, variance estimation is handled by resampling. The method of choice, illustrated in Fig. 7, is jackknife resampling, developed by Maurice Quenouille [38, 39] and refined by John Tukey [40].

Consider a data set of N different measurements $(x_1, x_2, x_3 ... x_N)$. Jackknifing then consists in the construction of N data sets, each systematically lacking one single measurement, and as such, of size $(N-1)$. The required observable is then estimated for each individual jackknife set and subsequently aggregated. For the purpose of this investigation, the jackknifed data are different independent measurements of R_g^2, for which, at every given frame, $N = 20$ distinct values exist. As such the only relevant observables are the mean and associated variance. For each subset, the i_{th} jackknife mean is:

$$\overline{x}^i_{\text{jack}} = \frac{1}{N-1} \sum_{j \in [N], j \neq i}^{N} x_j \tag{10}$$

From which the total mean is:

$$\overline{x}_{\text{jack}} = \frac{1}{N} \sum_{i}^{N} \overline{x}^i_{\text{jack}} \tag{11}$$

The variance will be given by:

$$\sigma_{\overline{x}}^2 = \frac{N-1}{N}\sum_{i=1}^{N}(\overline{x}_{\rm jack}^i - \overline{x}_{\rm jack})^2 \equiv \epsilon_{\overline{x}}^2 \tag{12}$$

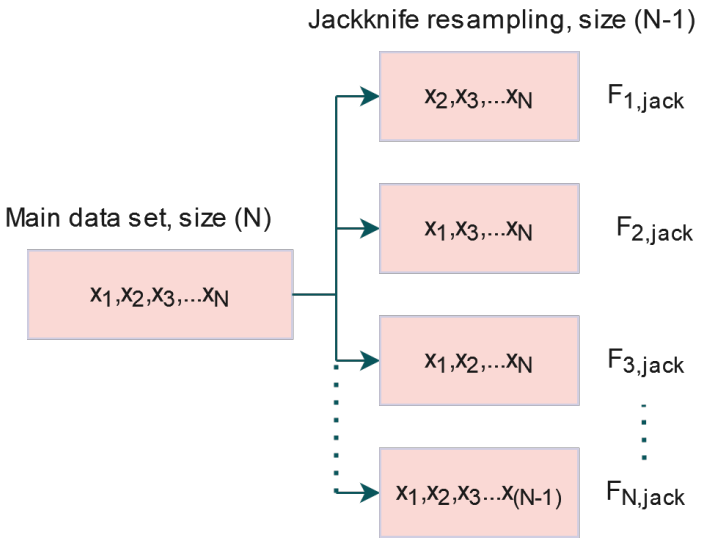
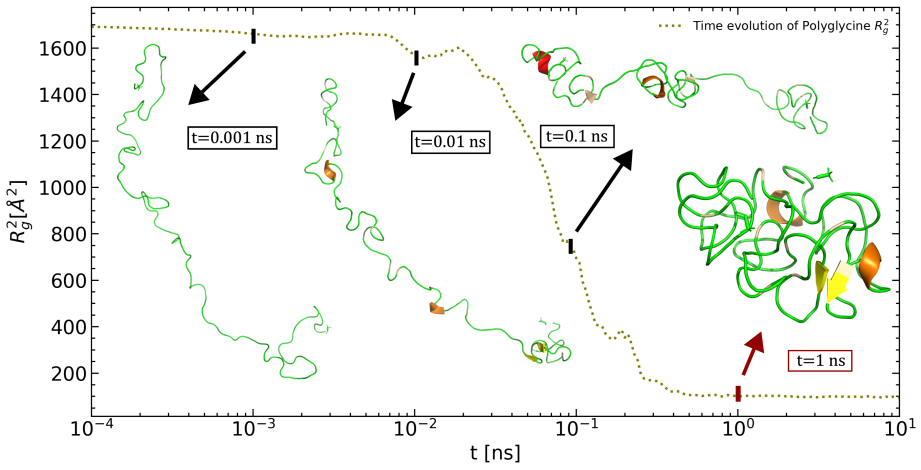


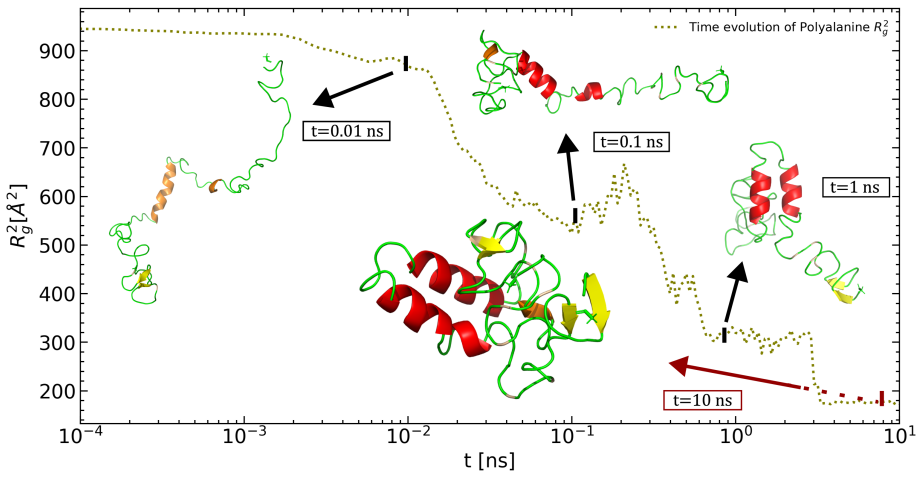
Figure 7: Illustration of the Jackknife resampling method.

3 Results & Discussion

3.1 Phenomenological Behavior of Polyalanine and Polyglycine



(a)



(b)

Figure 8: Time evolution of the square radius of gyration for a single molecule of (a) polyglycine and (b) polyalanine, with associated visual representation of the state of each molecule at a given time. 3_{10}-helices are colored orange, α-helices red, β-sheets are yellow, while single β-strands are colored white. The black arrows represent mid-folding intermediate stages, while the red arrows indicate the folded structures corresponding to an energy minimum, which usually remains consistent for the rest of the simulated frames.

The quench is almost instantaneous, with the simulation reaching the target temperature of 300 K in less then 0.01 ns, coinciding with the first sharp drop of R_g^2 in Fig. 8. The structural evolution of the protein is equally as rapid, with the initial phase of the transition requiring significantly more resolution to observe

then the middle and end. Equally spaced snapshots of the molecule would fail to convey its exact behavioral evolution, as such, logarithmic spacing is preferable. Fifty snapshots are taken for each decade of measurement (fifty between t=10 and t=100, fifty between t=100 and t=1000, and so on...).

After spending 0.2 ns in the high temperature phase, each molecule goes trough the quench. As a result of it, it folds closer and closer, until a native structure is reached. Fig. 8 showcases this folding process for two single iterations of polyglycine and polyalanine as a way to visually associate changes in R_g^2 with respect to shifts in the configurations of a molecule.

The initial stages of the folding process present shorter secondary elements, with the prevalent presence of 3_{10}-helices. Polyglycine does not showcase the transition to stable and long α-helices, with only a single short four residue element that quickly unravels (in about 0.01 nanoseconds). This happens repeatedly throughout the simulation. The collapse of the protein itself is very quick, with a general globular state being reached in around 0.3 nanoseconds and being completely established at the 1 nanosecond mark. Short β-sheets and strands, as well as 3_{10}-helices remain prevalent.

Polyalanine behavior is quite different, showcasing a collapse speed that is visibly lower than polyglycine, with multiple plateaus alternated by sharp drops of R_g^2 instead than a continuous decrease. The most energetically favorable state appears as a helical U-like structure and is achieved and maintained in around 3 nanoseconds of simulation time. The presence and reciprocal influence of two long α-helices might be to account for that. Their formation also had 3_{10}-helices as intermediates, but contrary to the previous molecule, the long helices -8 and 11 residues by the end of the simulation- do not fall apart. The significance of these more substantial secondary element is well-showcased by the behavior of R_g^2, which does not have the same steep linear decrease for polyalanine as it does for polyglycine.

R_g^2's behavior for the homopolymers is meant to represent a sort of boundary within which the remaining molecules are expected to be observed. As such, insights into it are precious for the observation of the copolymers, which are next to be dealt with.

3.2 Analysis of $\langle R_g^2 \rangle$ for all models

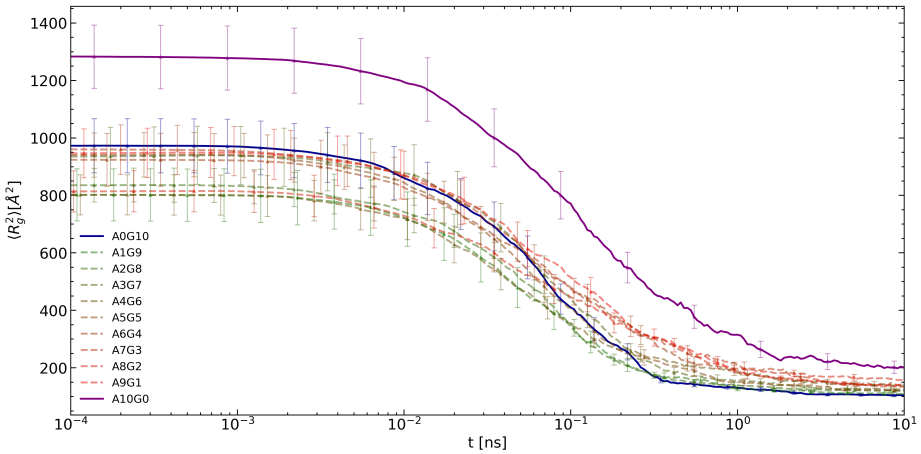


Figure 9: Behavior of $\langle R_g^2 \rangle$ throughout the folding process of all 11 molecules. Each curve is an average of 20 different independent simulations. Red indicates alanine content while green glycine, with the hue of each line is meant to represent its copolymers composition. The error is given by Jackknife resampling according to section 2.5.

Fig. 9 shows the result of the entire simulation. The behavior of the mean square radius of gyration is taken under observation for each molecule. At the start of the quench glycine and alanine content does not seem to play a significant role, although the overlap to the error bars suggests that a lot more simulation iterations would be necessary. However, considering the random configuration each chain reaches after the high temperature phase, such a degree of disorder may be a good indicator of the initial states being highly uncorrelated. The behavior of polyalanine is the exception, with the homopolymer's mean square radius of gyration remaining 10 (at the start of the simulation) to 100 % (at the end) higher than all others. This is possibly due to the structure of the amino acid. Alanine's side chain, consisting of a methyl group, introduces considerable steric hindrance and restricts the flexibility of the polymer chain. This leads to a more extended conformation, increasing the radius of gyration. In contrast, glycine has a hydrogen atom as its side chain, which is much smaller and allows for greater flexibility, a more compact structure and a lower R_g. Even minimal amounts of glycine lead to an increase in compactness that is not necessarily representative of the composition of the molecule.

From around $t = 0.1$ nanoseconds of simulation time, the different curves slowly start to arrange themselves accordingly to their composition. Polyglycine (in blue), as well as other molecules with high glycine content, represented by greener lines, show a steeper decrease than redder lines, as they go though a faster collapse. In the tail end of the graph the collapse state is calmly reached by the redder lines as well, which also start decaying more rapidly. In the Flory mean-field approximation, R_g^2 has well known scaling approximations that are random coil-like $R_g^2 \sim N^{\nu_F}$ with $\nu_F = \frac{3}{5}$ in a good solvent, and $R_g^2 \sim N^{\frac{2}{3}}$ for globular states in a poor solvent [29]. N, the degree of polymerization, or simply the number of individual polymers, is constant at 100, giving values of $R_{g,c}^2 = 251$

and $R^2_{g,c} = 21.54$. Polyalanine's last frame has a mean R^2_g of 199.8 Å^2. For polyglycine $\langle R^2_g \rangle = 104.05$ Å^2, with all other molecules coming in-between those two values and as such being approximately within the set bounds, with an acceptable difference due to the absence of a solvent.

3.3 Extrapolation and analysis of τ_c

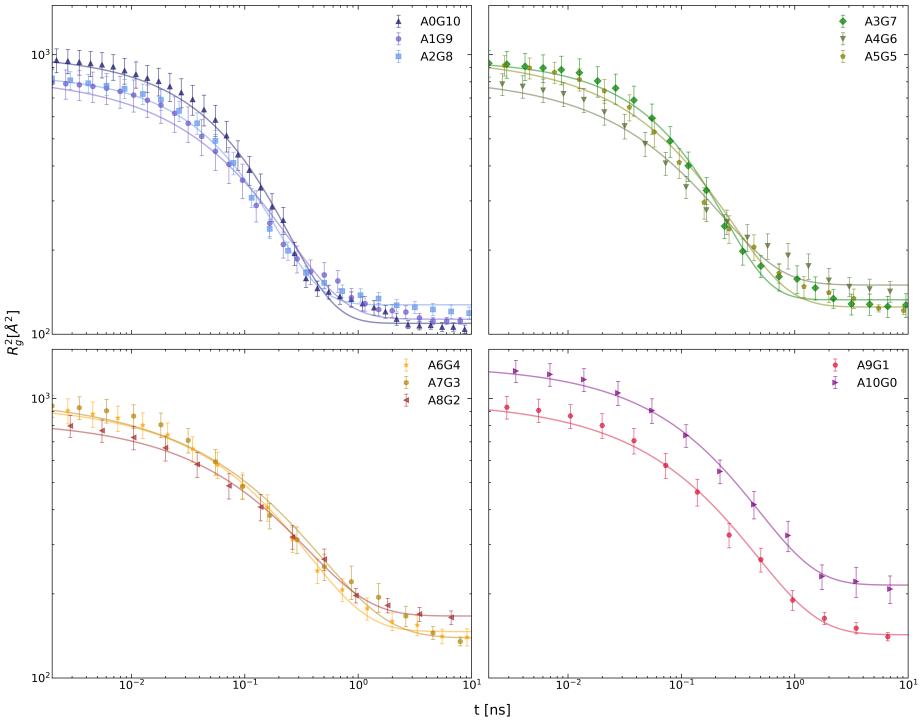


Figure 10: $\langle R^2_g \rangle$ for the 11 polymers in a log-log plot to better showcase the different behaviors. The whole lines represent the fits according to eq. 9. For better visibility, less data points are visualized.

Fig. 10 above describes the fitting of $\langle R^2_g \rangle$ for each model. The list of the obtained fitting parameters can be found in Appendix A, while τ_c is analyzed in Fig. 11 at the end of the section. A chi-squared test was carried out to quantify the 'goodness' of each fit, with the χ^2 parameter, also found in Appendix A, ranging between 0.1 and 0.5. Slightly lower than 1 values are expected due to the relatively large error bars, which could be improved upon by increasing the amount of measurements. When also considering the visual consistency of the curves, however, the fits are deemed adequate and are in line with similar analysis [30]. Despite the difference in decay, each fitting equation falls within error bars, satisfactorily justifying τ_c values. From Fig. 11(a), as well as the values in the table in Appendix A, it is clear to see that the collapse time for polyalanine is almost exactly twice as for polyglycine, with all the other molecules falling close to or within the two. Although not central to this thesis, a comparison of the collapse time τ_c with respect to the degree of polymerization

N (the number of monomers), is also looked at. Non-biological polymers have a firmly established scaling law $\tau_c \sim N^z$, with a debatable dynamic exponent which has been found to range from values $z \approx 1$ for MD simulations, all the way to $z \approx 2$ (and occasionally closer to 1) when Monte Carlo simulations were run instead, which do not account for hydrodynamics [41]. An extensive summary of the different values for z which have been reported can be found in [42]. Since N is constant at 100, the single value for z is easily calculated instead:

$$z = \frac{\ln \tau_c}{\ln N} \tag{13}$$

The values of the dynamic exponent can be found in Fig. 11(b), with z ranging from 0.96 for molecules with high glycine content, to 1.13 for polyalanine, which is more akin to the predicted value for MD simulation, as one would have expected, although higher than [30], which found, for pure polyglicine, a hyper fast folding time with z as low as $z \approx 0.5$.

τ_c is compared with τ_{50} and τ_{80} as depicted in Fig. 11. To facilitate a consistent comparison of the behaviors associated with these varying collapse times, all three parameters underwent min-max normalization:

$$\tau' = \frac{\tau - \tau_{\min}}{\tau_{\max} - \tau_{\min}} \tag{14}$$

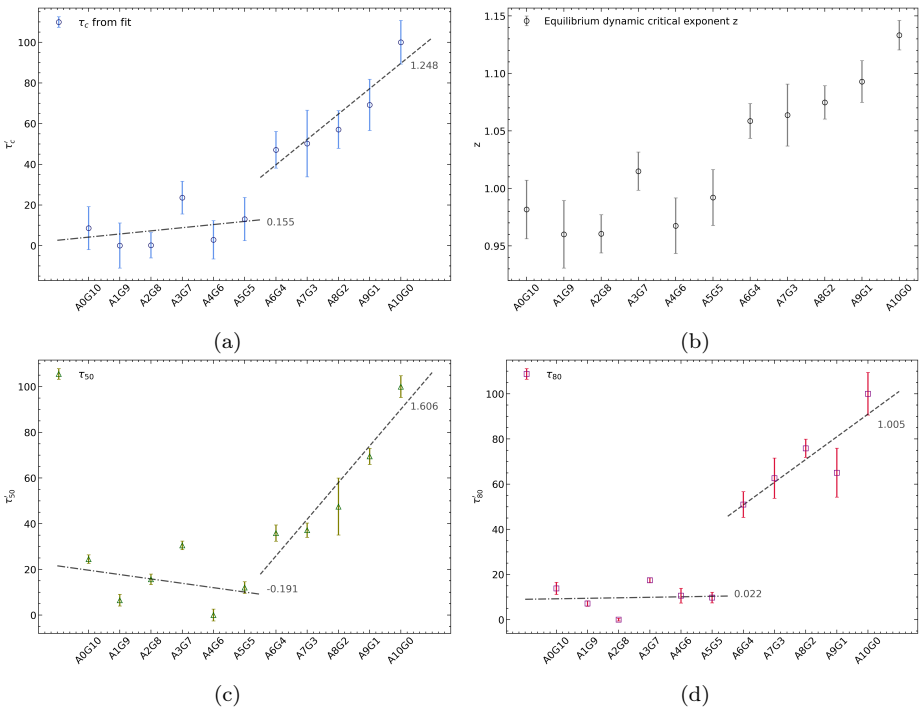


Figure 11: Behavior of the normalized collapse time with respect to the composition of the Polymer.(a) showcases the behavior of τ_c as obtained from the fitting equation, with (b) depicting the extrapolated dynamic critical at N=100 for all different molecules. (b) and (c) showcase τ_{50} and τ_{80}, corresponding to the respective relaxation times and obtained by interpolation of sections of the curves in Fig. 10.

Visualization of the different collapse times shows a clear behavioral difference in the speed of decay of the copolymers based on their composition. Significant increase in the length of collapse is evident only for polymers where the alanine content is higher than glycine. Let $\mathcal{A} \in [0, 100]$ and $\mathcal{G} \in [0, 100]$ be respectively the percentage alanine and glycine content such that $\mathcal{A} + \mathcal{G} = 100$. The behavior of the dynamic collapse scaling parameter γ_c can be defined such that: $\tau' = \gamma_c \mathcal{A}$. For $\mathcal{G} \geq \mathcal{A}$, by averaging the above values:

$$\gamma_c(\mathcal{G} \geq \mathcal{A}) = -0.005 \pm 0.101 \tag{15}$$

The dimensionless parameter's error is two orders of magnitude larger than the real value, as such the behavior of the collapse time in this range is predominantly guided by the fluctuation in the measurement due to the low sample size. In the absence of further analysis, it is perfectly valid to say that the speed of collapse is unaffected by the relative presence of alanine and glycine.

On the other hand, molecules where alanine is present in greater quantities than glycine have a time of collapse that increases proportionally to the amount of alanine: for $\mathcal{A} > \mathcal{G}$ the scaling parameter will amount to:

$$\gamma_c(\mathcal{A} > \mathcal{G}) = 1.286 \pm 0.174 \tag{16}$$

To draw conclusion on such a steep behavioral difference, a more detailed visual analysis follows.

3.4 Qualitative Analysis

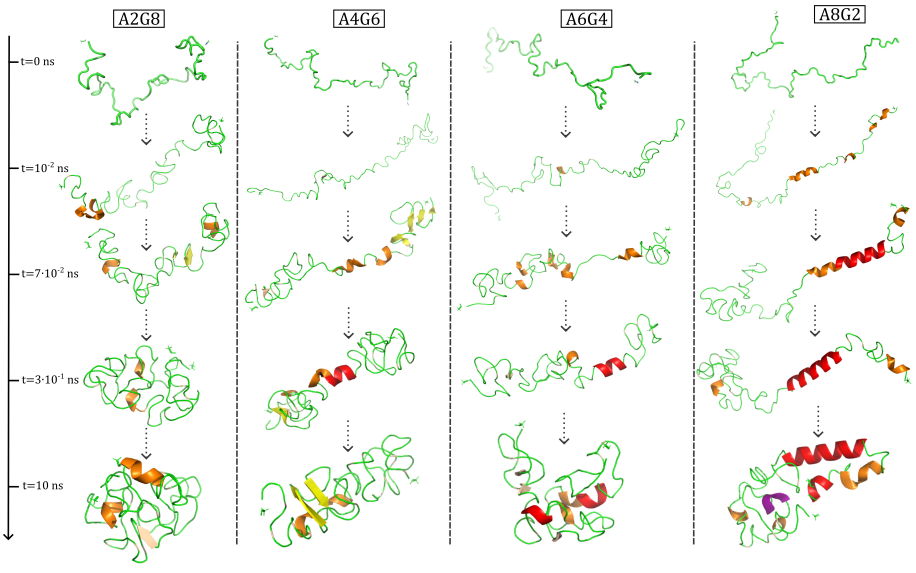


Figure 12: Visual analysis of four models from the start to the end of the simulated quench. Red represents α-helices, orange 3_{10}-helices, purple rare $\pi-$helices, while yellow and white β-sheets and strands.

A visual analysis of the helical secondary structure elements was conducted by taking snapshots of a single model for each molecule. Fig. 12 depicts A2G8, A4G6, A6G4 and A8G2. The remaining copolymers can be found in the Appendix B. The molecules are showcased at $t = 0$ of simulation time, right at the start of the quench, as well as at the start of the folding process $t = 0.01$ ns, at $t = 0.07$ ns, as the average τ_{50} value, at $t = 0.3$ ns, as the average τ_{80} value, and lastly, in the last simulation frame at $t = 10$ ns.

A qualitative analysis, despite its limits, is still insightful. A2G8 does not showcase any α-helix formation. A4G6 does manage to construct a short strand, although it unravels before a fully collapsed state is achieved. This is indicative of the correlation between the stability of α-helices and their length. Every single structure analyzed in this thesis consistently produced α-helical structures, but while shorter strands almost always quickly collapse and hardly ever appear in the final snapshots, longer ones are maintained throughout the folding once established. This is also evident from the main helix in A8G2 (Fig. 12). It is preceded by two 3_{10}-helices that come together and lead the molecule to bend towards the bifurcating helical structure of polyalanine (Fig. 8). in this case, however, the second helix is incomplete. A8G2 additionally showcases a rare short π-helix. Often associated to functional sites of the protein [43], π-helices are almost exclusively found in conjunction to destabilized α-helices [44]. With no such elements around it and since it appears exclusively in the last two frames, it is most likely a simulation artifact rather than a biologically relevant feature.

The sequence of snapshots for glycine rich molecules, to an extent, demonstrates the same process first seen with polyglycine: the initial local ordering of the residues along the chain with an emergence of local structures, such as short loops, and later forming a single central cluster. This is known as a pearl-necklace collapse and has been commonly reported when dealing with polyglicine both in the presence and absence of a solvent [29, 30, 42]. Contrary to the homopolymer, however, the behavior of the heteropolymers is not as predictable, with two local clusters emerging at the two ends of a slightly elongated molecule.

Taking a look back to τ_{80}, in reference to the table in Appendix A, the steep difference between A5G5 and A6G4 is particularly highlighted. For the first, $\tau_{80} = 175.44$ ps, while the latter almost doubles it. Given the assumption that the higher collapse time is the result of an increase in helical formation, this is likely due to the sequencing change of the amino amino acids. A6G4, with $[AGAGA]_{20}$ is the first molecule to present two connected alanine amino acids. These alternate with a $[GAG]$ block, which is considerably less active, with the common short α-helices, centering around an $[AA]$ block instead. Similar observations on sequencing can be done for the remaining molecules. A7G3 sequencing is such that three alternating alanine amino acids are bound together, separate by a $[GAAGAAG]$ block, while A8G2 sequencing leads to four bonded alanine monomers. For both A7G3 and A8G2, a consistent local folding pathway is the initial formation of 3_{10}-helices on the two alanine blocks, later occasionally followed by an expansion and connection into a long α-helix by the incorporation of the middle heterogeneous block. By A9G1 the homogeneous blocks of alanine are nine residues long and comprise the almost totality of helical element. The single glycine amino acid never appears in short helices and only rarely in α-helices that are long enough to join two alanine blocks. These folding pathways are akin to the two-step process described by Alves and Hansmann, which looked at the helicity of a peptide with two long strains of alanine joined by

a central glycine block ($Ala_{15} - Gly_5 - Ala_{15}$) [45]. The first step consist of the formation of two long α-helices, which afterwards lead the molecule into a U-like arrangement. Fig. 12 as well as Fig. 13 in Appendix B show that while polyalanine perfectly exemplifies this two-step process, A9G1, A8G2, A7G3, A6G4 increasingly struggle with the formation of both α-helices, but still show tendency towards the same behavior. The α-helices formation is more fragmented due to localized areas with high glycine working as a potential perimeter for the expansion of helices.

The final collapsed state achieved by each molecule is akin to our expectations. A2G8 is very similar to the globular polyglycine, and A8G2 resembles polyalanine, which suggests that the dominant presence of a specific amino acid indeed plays a major role in the folding pathways. The lack of clear influences for A4G6 and A6G4, however, does indicate more complex folding propensities in intermediate compositions. The difference in speed of the structural change of the different molecules is also quite evident. A2G8 and A4G6 achieve and maintain the final shape by the fourth snapshot, while the other two molecules undergo further structural changes between the fourth and the fifth. It is hard to determine how much of the increase in collapse time is due to the secondary structure elements, as accounting for the presence and nature of every helix is not trivial and goes beyond the scope of a qualitative analysis. An attempt at a quantitative approach was done by means of simply counting the amounts of secondary structural element of interest at each molecule's collapse time τ_c as estimated from the fit. The results are found in Appendix C. A table displays the count of each secondary helical element. Accounting for the size of α-helices was deemed relevant, as there is a significant difference between shorter helices, which tend to behave more akin to 3_{10}-helices, quickly disappearing after a few frames, and longer α-helices. The general count of α-helices is showcased in Fig. 14 and it displays a remarkable similarity to that of τ_c, although accounting for all helical elements shows the limitations of this method.

4 Conclusions & Outlook

From the results of the analysis, the presence of α-helices has been reconfirmed to be associated with longer collapse dynamics, however, the extent of that relationship is still unclear, especially in light a cutoff composition dictating both the production of long α-helices and the increase in folding time. A linear increase of the collapse time τ_c with a proportionality constant $\gamma_c = 1.286 \pm 0.174$ is found only in models where alanine is more then 50 % of the total, with little to no changes for lower percentages. Equivalently, α-helices were displayed with higher prominence in these same molecules. The behavior of those molecules resembled the two-step process with the local formation of twin helices followed by the re-arrangement of the molecule in a U-like bifurcated configuration. While polyalanine displays this behavior repeatedly and consistently, all other polymers rich in alanine fail to form both helices at full length and fold into hybrid cluster-like elongated shapes, with one -or more- major central helix instead. This is indicative of the limiting effect of glycine for the expansion of an helix. One single residue, placed between conjoining block of alanine, is sufficient to stunt its folding pathway.

Molecules rich in glycine where found to have behaviors similar to that of polyglycine, which folds by mean of a pearl necklace-like evolution, although with the alanine content increase, the final globular structure is often elongated. In terms of direct correlation between helicity and collapse time, similar linear scaling was found between τ_c values and amounts of helical elements at specific key frames of the molecules cycle, which reinforces the key role α-helices have in the slowing down of the collapse.

Although the approach used was sufficient here, further studies on the subject would need to invest time into defining a more rigorous method to verify the relationship between collapse time and presence of α-helices, such as analysis of average helicity $\langle n_H \rangle$ and end-to-end distance $\langle d_{e-e} \rangle$ with respect to temperature for each model, as well as modeling the shifting thermal characteristics of the systems: internal energy, free energy and heat capacity. Additionally, an analysis of the exact composition of every single helical element, that is, how much glycine and alanine is actually of each secondary structure element with respect to the total amount, as well as Ramachandran plotting of key residues, may reveal useful information on folding mechanics.
As illustrated by this thesis, however, there remains a strong foundation for further exploration of this subject.

Acknowledgments

My full acknowledgments go to my supervisor M.Sc. Maximilian Conradi, for his constant guidance and never-ending patience in these last months.
I also thank Prof. Dr. Wolfhard Janke for the opportunity to work on this topic, as well as for his lectures, which accompanied me in my last semester.
Additional thanks to M.Sc. Fabio Müller for the contributions to the writing of the logarithmic spacing snapshots code.

References

[1] Abdulsalam A.; Nafaa A.; Amer A.: *Intracellular Protein Biosynthesis: A Review* Asian J. Biochem. **2**, 10-18, (2020).

[2] Eskandari A.; Leow T. C.; Rahman M. B. A.; Oslan S. N.: *Antifreeze Proteins and Their Practical Utilization in Industry, Medicine, and Agriculture*, Biomol. **10(12)**, 1649, (2020).

[3] Hardesty B.; Kramer G.; *Folding of a nascent peptide on the ribosome*, Prog. Nucleic Acid Res. Mol. Biol. **66**, 41–66, (2001).

[4] Khan R. H.; Siddiqi M. K.; Salahuddin P.: *Protein structure and function*, Basic Biochem. **5**, 1-39, (2017).

[5] Radford S. E.; Dobson C. M.: *From Computer Simulations to Human Disease: Emerging Themes in Protein Folding*, Cell **97(3)**, 291–298, (1999).

[6] Anfinsen C.B.: *Principles that govern the folding of protein chains*, Science **181(4096)**, 223-230, (1973).

[7] Lehninger A.; Nelson D. L.; Cox M. M.: *Lehninger principles of biochemistry (5th ed.)*, Chapter 4, 116-156, (2017)

[8] Dobson C. M. : *Protein folding and misfolding*, Nature **426**, 884–890, (2003).

[9] Jumper J. et al.: *Highly accurate protein structure prediction with AlphaFold*, Nature, **596**, 583–589, (2021).

[10] Durell S. R.; Ben-Naim A.: *Hydrophobic-hydrophilic forces in protein folding*, Biopolymers **107(8)**, (2017).

[11] Dill K. A.; Ozkan S. B.; Shell M. S.; Weikl T. R.: *The Protein Folding Problem*. Annu. Rev. Biophys. **37**, 289–316, (2008).

[12] Mirsky A. E.; Pauling L.: *On the Structure of Native, Denatured, and Coagulated Proteins*, PNAS **22(7)**, 439-447, (1936)

[13] Jacoesen C.; Linderstrom-Lang K.: *Salt Linkages in Proteins*. Nature **164**, 411–412, (1949).

[14] Murphy K. P.: *Protein Structure, Stability, and Folding*, MIMB **168**, 5-13, (2001).

[15] Newberry R. W.; Raines R. T.: *Secondary Forces in Protein Folding*, ACS Chem. Biol. **14(8)**, 1677-1686, (2019).

[16] Branden C.; Tooze J.: *Introduction to Protein Structure (2nd ed.)*, Chapter 2, 12-20, (1999).

[17] Qin Z.; Fabre A.; Buehler M. J.: *Structure and mechanism of maximum stability of isolated alpha-helical protein domains at a critical length scale*, Eur. Phys. J. E. Soft Matter **36(5)**, 53, (2013).

[18] Richardson J.S.; Richardson D.C.: *Amino acid preferences for specific locations at the ends of α-helices*, Science **240**, 1648–1652, (1988).

[19] Millhauser G.L.: *Views of helical peptides: A proposal for the position of 3_{10}-helix along the thermodynamic folding pathway*, Biochemistry **34**, 3873–3877, (1995).

[20] Daggett V.; Levitt M.: *Molecular dynamics simulations of helix denaturation*, J. Mol. Biol. **233**, 1121–1138, (1992).

[21] Touw G. W.; Baakman C.; Black J.; Beek T. A.; Krieger E.; Joosten R. P.; Vriend G.:*A series of PDB related databases for everyday needs*, Nucleic Acids Res. **43**, (2015).

[22] Kabsch W.; Sander C *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features*, Biopolymers *22*, 2577-2637, (1983).

[23] Rohl C. A.; Fiori W.; Baldwin R. L.: *Alanine is helix stabilizing in both template-nucleated and standard peptide helices*, PNAS **96**, 3682–3687, (1999).

[24] Ramachandran G. N.; Sasisekharan V.: *Conformation of polypeptides and proteins*, Adv. Protein Chem. **23**, 283-438, (1968).

[25] Scott K.; Alonso D.; Sato S.; Fersht A.; Daggett V.: *Conformational entropy of alanine versus glycine in protein denatured states*, PNAS U.S.A. **104**, 2661-2666, (2007).

[26] Beck D.; Alonso D.; Inoyama D.; Daggett V.: *The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins*, Proc. Natl. Acad. Sci. U.S.A. **105**, 12259-12264, (2008).

[27] Byrne A.; Kiernan P.; Green D.; Dawson K.A.: *Kinetics of homopolymer collapse*, J. Chem. Phys. **102**, 573–577, (1995).

[28] Christiansen H.; Majumder S.; Janke W.: *Coarsening and aging of lattice polymers: Influence of bond fluctuations*, J. Chem. Phys. **147(9)**, (2017).

[29] Majumder S.; Zierenberg J.; Janke W.: *Kinetics of polymer collapse: Effect of temperature on cluster growth and aging*, Soft Matter **13**, 1276–1290, (2017).

[30] Majumder S.; Hansmann U. H. E.; Janke W.: *Pearl-Necklace-Like Local Ordering Drives Polypeptide Collapse*, Macromolecules **52**, 5491–5498, (2019).

[31] Rubinstein M.; Colby R. H.: *Polymer Physics*, Chapter 2.4, 60-66, (2003).

[32] Eastman P.; Swails J.; Chodera J. D.; McGibbon R. T.; Zhao Y.; Beauchamp K. A.; Wang L.-P.; Simmonett A. C.; Harrigan M. P.; Stern C. D.; Wiewiora R. P.; Brooks B. R.; Pande V. S.: *OpenMM 7: Rapid development of high performance algorithms for molecular dynamics*, PLOS Comp. Biol. **13(7)**,(2017).

[33] Cornell W. D.; Cieplak P.; Bayly C. I.; Gould I. R.; Merz K.M. Jr.; Ferguson D.M.; Spellmeyer D. C.; Fox T.; Caldwell J.W.; Kollman P.A.: *A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules*, J. Am. Chem. Soc. **117(19)**, 5179–5197, (1995).

[34] Lee F. S.; Warshela A.: *A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations*, J. Chem. Phys. **97**, 3100–3107, (1992).

[35] Izaguirre J. A.; Sweet C. R.; Pande V. S.: *Multiscale dynamics of macromolecules using Normal Mode Langevin*. PSB **15**, 240–251, (2010).

[36] Taylor T. J.; Vaisman I. I.: *Discrimination of thermophilic and mesophilic proteins*, BMC Struct. Biol. **10(1)**, (2010).

[37] Arashiroa E.; Drugowich J. R.; Hansmann U. H. E.: *Short-time dynamics of polypeptides*, Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **73(4)**, (2006).

[38] Quenouille M. H.: *Problems in Plane Sampling*, Ann. Math Stat. **20(3)**, 355–375, (1949).

[39] Quenouille M. H.: *Notes on Bias in Estimation*, Biometrika **43(3–4)**, 353–360,(1956).

[40] Tukey J. W.: *Bias and confidence in not quite large samples (abstract)*, Ann. Math Stat. **29(2)**, 614, (1958).

[41] Hasenbusch M.: *Dynamic critical exponent z of the three-dimensional Ising universality class: Monte Carlo simulations of the improved Blume-Capel model*, Phys. Rev. E. **101(2)**, (2020).

[42] Majumder S.; Christiansen H.; Janke W: . *Understanding nonequilibrium scaling laws governing collapse of a polymer*, Eur. Phys. J. **93(142)**, (2020).

[43] Weaver T. M.: *The pi-helix translates structure into function*, Protein Science **9(1)**, 201–206, (2000).

[44] Cooley R. B.; Arp D. J.; Karplus P.A.: *Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality*, J. Mol. Biol.**404(2)**, 232–246,(2010).

[45] Alves N. A.; Hansmann U. H. E.; *Helix Formation and Folding in an Artificial Peptide*. J. Chem. Phys. **117**, 2337–2343 (2002).

Appendix

A Fit and Interpolation Parameters

	b_0[Å^2]	b_1[Å^2]	τ_c[ps]	β	χ^2
A0G10	109.34 ± 1.83	888.51 ± 28.84	91.90 ± 10.76	0.71 ± 0.06	0.546
A1G9	113.16 ± 1.53	723.44 ± 25.46	83.19 ± 11.26	0.60 ± 0.05	0.363
A2G8	127.41 ± 2.31	724.01 ± 15.91	83.37 ± 6.36	0.76 ± 0.05	0.410
A3G7	132.77 ± 3.05	822.85 ± 19.32	107.13 ± 8.18	0.77 ± 0.06	0.214
A4G6	149.94 ± 4.96	682.40 ± 21.43	86.11 ± 9.57	0.59 ± 0.05	0.343
A5G5	125.12 ± 2.25	850.19 ± 23.36	96.40 ± 10.77	0.61 ± 0.05	0.411
A6G4	146.58 ± 2.59	805.55 ± 15.53	130.98 ± 9.15	0.59 ± 0.03	0.150
A7G3	139.06 ± 3.36	867.98 ± 27.18	134.11 ± 16.62	0.50 ± 0.04	0.351
A8G2	166.63 ± 2.43	669.99 ± 11.64	141.10 ± 9.43	0.57 ± 0.03	0.109
A9G1	142.80 ± 2.06	838.26 ± 17.96	153.38 ± 12.80	0.56 ± 0.03	0.195
A10G0	214.99 ± 6.11	1097.63 ± 18.25	184.68 ± 10.86	0.61 ± 0.03	0.110

Table 1: Fitting parameters and χ^2 values.

	τ_{50}[ps]	τ_{80}[ps]
A0G10	64.28 ± 1.09	188.17 ± 8.48
A1G9	53.92 ± 1.49	167.08 ± 3.76
A2G8	59.20 ± 1.33	144.59 ± 1.90
A3G7	67.76 ± 1.08	199.60 ± 3.33
A4G6	50.16 ± 1.50	178.09 ± 10.31
A5G5	57.10 ± 1.47	175.44 ± 7.41
A6G4	70.85 ± 2.05	305.38 ± 17.93
A7G3	71.59 ± 1.82	342.22 ± 28.09
A8G2	77.56 ± 7.20	383.99 ± 12.71
A9G1	90.23 ± 2.03	349.84 ± 34.22
A10G0	107.84 ± 2.75	460.14 ± 29.66

Table 2: Interpolation parameters

B Visual representation of Remaining Molecules

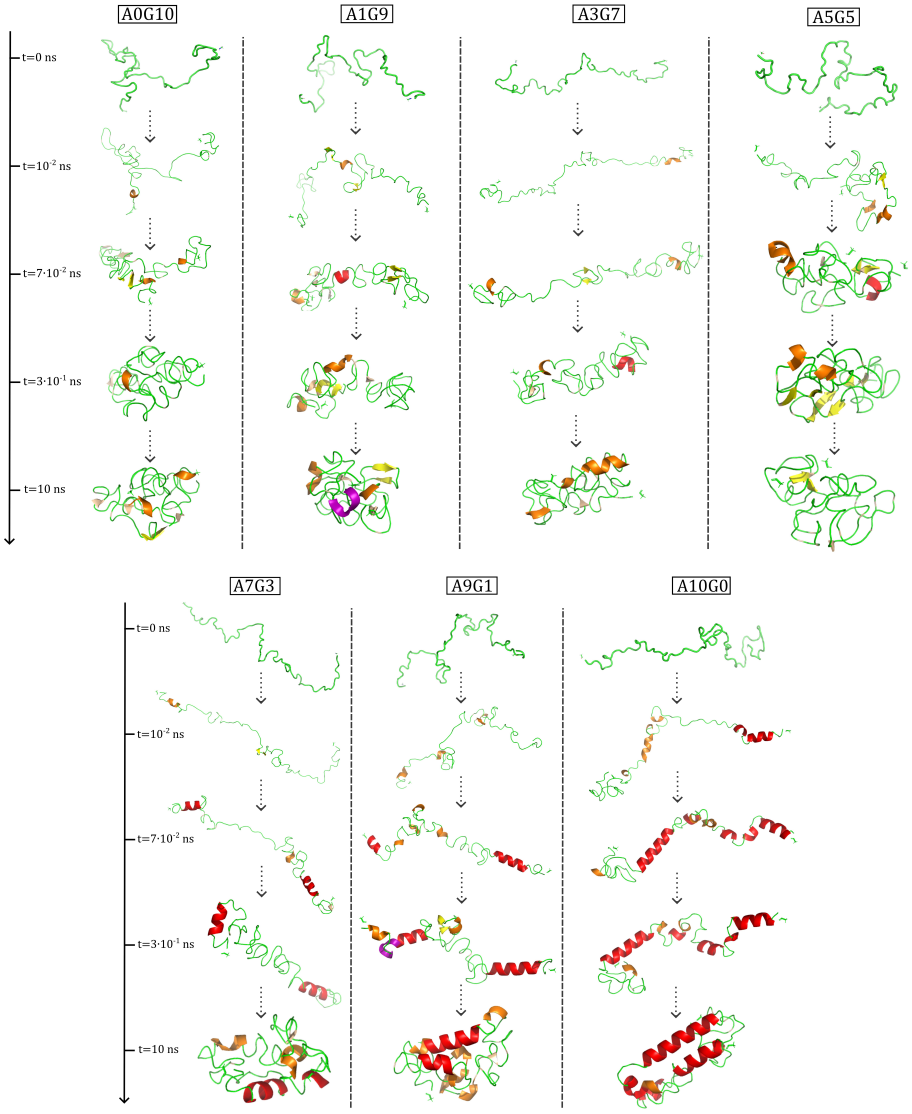


Figure 13: Visual representation of one random configuration for each of the remaining molecules. The sudden increase in α-helices is clearly visible from the molecules above to those below. Red once again indicates α-helices, orange 3_{10}-helices, yellow $\beta-$sheets and purple $\pi-$helices.

C Count of Helical Elements at τ_c

	τ_c frame	α-helices (<6 residues)	α-helices (≥ 6 residues)	3_{10}-helices
A0G10	198	15	2	21
A1G9	198	9	2	36
A2G8	197	5	1	39
A3G7	199	16	4	50
A4G6	198	10	3	45
A5G5	200	8	4	49
A6G4	211	11	8	53
A7G3	212	18	16	49
A8G2	213	15	13	52
A9G1	215	18	20	70
A10G0	218	25	34	58

Table 3: Summary of the secondary structure elements for each polymer.

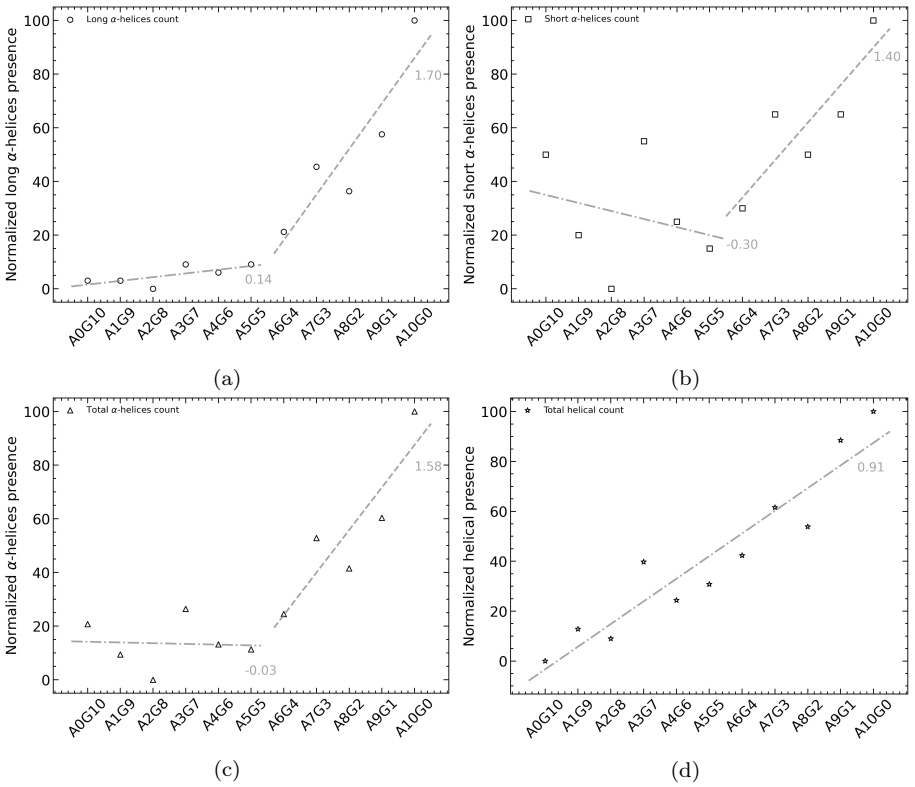


Figure 14: Plotting of the count of helices at collapse time.(a) showcases the count of long α-helices made up of 6 or more residues, meanwhile (b) showcases short α-helices, (c) is the total count of α-helices, (d) is the count including 3_{10}-helices as well.