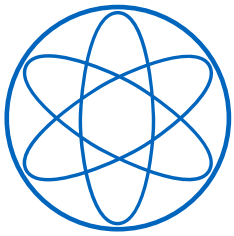
Studying Mutations and Perturbations in Biomolecules using Molecular Dynamics Simulations

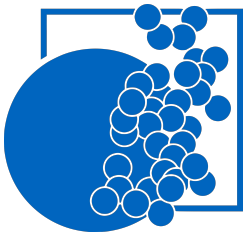
DISSERTATION

Julian Myrddin Pascal Hartmann

TECHNISCHE UNIVERSITÄT MÜNCHEN

FAKULTÄT FÜR PHYSIK





TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM SCHOOL OF NATURAL SCIENCES

Studying Mutations and Perturbations in Biomolecules using Molecular Dynamics Simulations

Julian Myrddin Pascal Hartmann

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genemigten Dissertation.

Vorsitz:	Prof. Dr. Julia Herzen
Prüfer der Dissertation:	1. Prof. Dr. Martin Zacharias 2. Prof. Dr. Ulrich Gerland

Die Dissertation wurde am 12.7.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 21.9.2023 angenommen.

Abstract

Within the framework of this thesis, the Molecular Dynamics (MD) simulation method was used to investigate three different biological systems of biomolecules. Each biomolecule was considered in terms of the effect of an mutation or perturbation in the sequence and/or structure of the molecule. The folding of a protein was statistically simulated using collagen peptides. Collagen fibres consist of three collagen strands that fold into a triple helix. In a given environment, the amino acid sequence of the individual strands and the repeated stabilization of the helix by heat shock protein 47 (HSP47) are decisive. It was shown that insufficient stabilization of the initial nucleus makes subsequent folding more difficult. Furthermore, it was shown that even a point mutation of glycine (Gly), which occurs periodically in the protein sequence, to alanine (Ala), impedes folding and bends the folded helix.
The transport protein Transthyretin (TTR) is normally present in our blood as a tetramer. However, it can also appear in the form of pathological amyloid fibrils. These deposits, which are hardly degradable, are a common cause of cardiovascular diseases in old age. However, several mutations are known which lead to similar clinical pictures at a young age. It is assumed that some mutations destabilize the tetrameric protein and thus tend towards increased amyloid formation. MD simulations combined with free energy calculations were used to investigate 36 known mutations with regard to their tendency to stabilize or destabilize. In good agreement with experimental data, a rapid method was developed to estimate the effect of single point mutations. The developed method delivered better results compared to the FoldX package.
Finally, the dynamics of an extra nucleotide within a double-stranded Ribonucleic Acid (dsRNA) were studied. For this purpose, an artificially generated Ribonucleic Acid (RNA) sequence containing a bulge loop was simulated. The behaviour of the bulge loop was statistically evaluated with regard to the position of the bulge along the helix and to its neighbouring residues. Clear differences were observed between an adenine (A) bulge and a uracil (U) bulge. While the A-bulge mostly stays within the helical structure, the U-bulge often loops out of the helix, which fixes it at that moment, while otherwise it moves more flexibly along the helix.

Zusammenfassung

Im Rahmen dieser Arbeit wurde die Methode der Molecular Dynamics (MD) Simulation genutzt, um drei biologische Systeme von Biomolekülen zu untersuchen. Jedes Biomolekül wurde in Hinsicht auf Auswirkung eines Fehlers oder einer Störung in der Sequenz und / oder der Struktur des Moleküls betrachtet. Anhand von Kollagenpeptiden wurde die Faltung eines Proteins statistisch simuliert. Das Kollagenmolekül besteht aus drei Kollagensträngen, die sich zu einer Dreifachhelix falten. Entscheidend dabei sind unter anderem die Aminosäuresequenz der einzelnen Stränge und die wiederholte Stabilisierung der Helix durch heat shock protein 47 (HSP47). Es wurde gezeigt, dass eine zu schwache Stabilisierung des Anfangsnukleus die folgende Faltung erschwert. Des Weiteren wurde gezeigt, dass bereits eine Punktmutation des periodisch in der Proteinsequenz auftretenden Glycin Gly zu Alanin Ala, die Faltung behindert und die gefaltete Helix krümmt.
Das Transportprotein Transthyretin (TTR) liegt normalerweise als Tetramer in unserem Blut vor. Allerdings kann es auch in Form von pathologischen Amyloidfibrillen auftreten. Diese schwer abbaubaren Ablagerungen sind eine häufige Ursache für Herzkreislauferkrankungen im Alter. Allerdings sind mehrere Mutationen bekannt, welche schon in jungem Alter zu ähnlichen Krankheitsbildern führen. Es wird vermutet, dass einige Mutationen das Tetramer-Protein destabilisieren und so zur verstärkten Amyloidbildung neigen. MD Simulationen kombiniert mit freie Energie Berechnungen wurden genutzt, um 36 bekannte Mutationen hinsichtlich ihrer Neigung zur Stabilisierung oder Destabilisierung zu untersuchen. Bei guter Übereinstimmung mit experimentellen Daten wurde eine schnelle Methode entwickelt, um den Effekt einzelner Punktmutationen abzuschätzen. Die gewählte Methode lieferte im Vergleich zum FoldX-Paket bessere Ergebnisse.
Schließlich wurde die Dynamik eines zusätzlichen Nukleotids in einer double-stranded Ribonucleic Acid (dsRNA) beobachtet. Hierfür wurde eine künstlich erzeugte Ribonucleic Acid (RNA) Sequenz, die einen Bulge Loop enthält, simuliert. Das Verhalten des Bulge Loops wurde sowohl bzgl. der Position des Bulges entlang der Helix, als auch bzgl. der Stellung zu seinen Nachbarresiduen statistisch ausgewertet. Hierbei wurden deutliche Unterschiede zwischen einem Adenin(A)-Bulge und einem Uracil(U)-Bulge beobachtet. Während der A-Bulge meist innerhalb der Helixstruktur bleibt, klappt der U-Bulge häufig aus der Helix heraus, was ihn in diesem Moment fixiert, während er ansonsten beweglicher entlang der Helix wandert.

Contents

Chapter 1

Introduction and Overview

If you look at the molecular building blocks of biological life on earth, the available kit appears very clear and simple at first glance. Starting with the information-storing "libraries" Deoxyribonucleic Acid (DNA) and Ribonucleic Acid (RNA), which enable inheritance and thus development and evolution, there are only four nucleic acids each that form the code that makes every living being unique. This code contains the blueprints for the "machines" of our cells, the proteins. Again, there are only 21 basic building blocks – amino acids – that make up each protein. How can the incredible diversity of life emerge from this small number of building blocks? If one imagines a toy box with building bricks that contains so few different types of bricks, one quickly realises what it takes to have fun with them (assuming you loved playing with such bricks as a child and never quite grew up...). Firstly, a huge number of bricks is needed and secondly, it must be possible to combine them as freely as possible in order to be able to build a wide variety of shapes. Then it is possible to decorate entire amusement parks with it. With regard to DNA and RNA, the first point is particularly true. With as many nucleotides as we have fingers on our hands, already over a million different combinations can theoretically be generated. The human genome contains 3.2 billion building blocks! [1] By the way even with just two "blocks" – 0 and 1 – many things are possible, otherwise this thesis would not have been possible.
The second point does not apply to biomolecules (DNA, RNA and proteins) at first glance. Each building block has only two docking points and, in principle, only long spaghetti molecules can be built with them. However, these molecules are very flexible. Therefore, the structures folded from it are not only diverse but also variable, which further increases the variety.
Furthermore, if we take a closer look, those building blocks of biomolecules have more than two docking points of different strengths. An essential aspect of life and evolution is reproduction. Due to the enormous number of reproductive steps at the molecular level, it is inevitable that mutations occur. These mutations or perturbations can have positive, negative, or no effect on the resulting molecule. It depends on the perspective. In the complex interaction of all biomolecules,

we are usually not able to recognize all the effects of a mutation. In most cases, a single faulty protein or an incorrectly transcribed RNA is quickly degraded by the cell, and a mutation in the DNA is corrected again by its repair mechanisms. However a single mistake in the wrong or right place can have major consequences. For example, a mutation in the DNA can lead to repeatedly incorrectly built proteins, which can cause serious diseases. A mutated RNA can suddenly make a virus much more contagious or enable it to infect entirely new hosts. A modified protein can adopt an unfavourable structure, thereby influencing other proteins, which can ultimately lead to completely different complexes. On the other hand, a mutation can also mean that a certain protein is more stable in its natural conformation, or can perform its task in a better or quicker way. This way mutations are fundamental parts of evolution.

Nowadays there are a variety of measuring instruments to study such small structures as biomolecules. In most cases, however, the focus is always on spatial or temporal resolution. If particularly fast processes are to be measured, it is often only possible to trace few points within the system. If a molecule's complete structure should be mapped, this usually happens so slowly that fast processes cannot be tracked. In order to be able to observe both at the same time, it would be practical if one could slow down time. Alternatively, it would be useful if one knew the rules governing the processes at the molecular level so well that one could re-enact them at an acceptable speed. Nowadays both the knowledge of these rules and the possible game speed (computer power) are large enough to make this possible. The instrument for this is called Molecular Dynamics (MD) simulations.
As part of this thesis, the theoretical foundations of MD simulations are explained first. Then three examples of mutations in biomolecules and their effects are discussed. The first example is the protein collagen, which is the most abundant protein in our body. It forms long fibres that are a fundamental part of the structure of a variety of tissue types from skin to tendons to bone. First, the process of folding three collagen peptides into a triple helix is examined. Individual mutations are then introduced to analyse how they affect the folding process. Furthermore perturbations of the starting nucleus are simulated to investigate their influence on the protein structure..
The second example is the tetrameric protein Transthyretin (TTR), which is a transport protein for various molecules in our body. Unfolding and faulty refolding of this protein can lead to so-called amyloid formation. These are long fibrils made of incorrectly folded proteins, which are difficult to degrade and can therefore lead to pathological deposits in the body. Several mutations of TTR are known which either promote or prevent this process. In order to get a quick method to estimate the effect of a mutation, already folded tetramers are simulated over time to obtain an ensemble of conformations, which is then used to calculate free energy differences.
Finally, the third example deals with a RNA double helix containing an excess nucleobase (bulge loop) which disrupts the normal double helix structure. By creating an artificially repetitive RNA sequence, it is possible for this perturbation

to propagate along the helix. The behaviour and dynamics of the perturbation are observed and analysed using long MD simulations.

Chapter 2

Theory of Molecular Dynamics Simulations

As mentioned at the beginning, Molecular Dynamics (MD) Simulations have proven to be an appropriate tool to describe the behaviour of biomolecules like proteins, DNA or RNA. It is a completely theoretical *in silico* method, using formerly gained knowledge of molecular principles and laws to obtain better insights into biomolecular processes. The basic principles of MD Simulations are described in the following section.

2.1 Discretization of continuous Processes

In short MD is a method to numerically iterate the positions of given set of atoms of a molecular structure through time using time steps small enough to preserve the system's realistic behaviour and at the same time big enough to obtain satisfying progress of the system's behaviour. Obviously the adherence of the first boundary is crucial for the latter goal and can not be evaded. Experience has shown, that the limit is around few femtoseconds, determined by the fastest movements within the system, the hydrogen atoms.[2] Fortunately steadily rising processing power of modern computer chips allows to simultaneously reach the second goal by increasing the number of iterations.
To describe the atoms' dynamics Newton's equation of motion is applicable, determining the dynamic of a particle i with mass m_i and location $r_i(t)$ caused by a force F_i.

$$F_i = m_i \frac{d^2\mathbf{r_i}(t)}{dt^2} \tag{2.1}$$

The force F_i results from the rest of the particles in the system and possible external influences and can be calculated at any time by the negative gradient

of a potential energy function U, called *force field*.

$$F_i = \frac{\partial U(r_1, ..., r_N)}{\partial r_i} \tag{2.2}$$

As biomolecular systems usually consist of many ($N \gg 2$) atoms, there is no analytical solution of this many-body problem. On the contrary many algorithms exist to solve it numerically. A simple method to integrate the equation of motion is to just perform a Taylor expansion and truncate it after the second order. Further simplification is possible by applying the Verlet algorithm, based on the Leapfrog integration, which omits calculating the atoms' velocities after each step by performing a forward and backward Taylor expansion in time and summing both:

$$\begin{aligned} \mathbf{r_i}(t + \Delta t) &= \mathbf{r_i}(t) + \frac{d\mathbf{r_i}(t)}{dt}\Delta t + \frac{1}{2}\frac{d^2\mathbf{r_i}(t)}{dt^2}\Delta t^2 + \frac{1}{6}\frac{d^3\mathbf{r_i}(t)}{dt^3}\Delta t^3 + \mathcal{O}(\Delta t^4) \\ \mathbf{r_i}(t - \Delta t) &= \mathbf{r_i}(t) - \frac{d\mathbf{r_i}(t)}{dt}\Delta t + \frac{1}{2}\frac{d^2\mathbf{r_i}(t)}{dt^2}\Delta t^2 - \frac{1}{6}\frac{d^3\mathbf{r_i}(t)}{dt^3}\Delta t^3 + \mathcal{O}(\Delta t^4) \end{aligned} \tag{2.3}$$

resulting in the new location $r_i(t + \Delta t)$, only depending on the current and prior location and the force acting on the atom.

$$\mathbf{r_i}(t + \Delta t) = 2\mathbf{r_i}(t) - \mathbf{r_i}(t - \Delta t) + \frac{d^2\mathbf{r_i}(t)}{dt^2}\Delta t^2 \tag{2.4}$$

Other common integrators are the velocity Verlet method [3] or the Beeman algorithm [4].
Since atoms consist of nuclei and electrons, a quantum mechanical consideration ought to be necessary. The Born-Oppenheimer approximation allows us to treat atoms as classical particles regardless of the electrons' quantum mechanical character. It assumes that electrons move much faster compared to their nuclei and thus follow them instantaneously. In MD there are no distinct electrons but only classically treated nuclei which are equal to the atoms. The effect of the electrons is described by a global energy landscape acting on the atoms.

2.2 Force Field

While optimizing the integration of the equation of motion 2.1 leads to an acceleration of the calculations, the accuracy of the physical behaviour of the system is mainly determined on the force field from equation 2.2. It describes all interactions between the system's atoms. In this study the force field from the AMBER software package [5] was used. Only a basic description can be presented here.

The system's energy is determined by five additive contributions.

$$\begin{aligned} U(\mathbf{r_1},...,\mathbf{r_N}) &= \sum_{bonded} k_d(d-d_0)^2 \\ &+ \sum_{angles} k_\theta(\theta-\theta_0)^2 \\ &+ \sum_{dihedrals} V_n[1+cos(n\Phi-\gamma)] \\ &+ \sum_{i<j} 4\upsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \\ &+ \sum_{i<j}\left[\frac{q_i q_j}{\epsilon r_{ij}}\right] \end{aligned} \tag{2.5}$$

The first three terms mimic the behaviour of all covalent bonds within the system, with two simple harmonic potentials around a ground state, representing the oscillations of the bond length d and the angle θ of neighbouring (bound) atoms around their equilibrium state d_0 and θ_0. k_d and k_θ are spring constants. The third term describes the periodic behaviour of the dihedral angles Φ represented by a cosine function with scaling factor V_n, multiplicity n and phase γ. (figure 2.1)

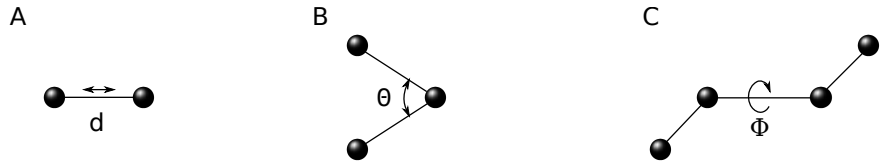


Figure 2.1: **Basic covalent interaction parameters.** Variation of distance (A), angle (B) and dihedral angle (C) between neighbouring atoms.

Unbound interactions are represented by Lennard-Jones potentials with parameters υ_{ij} and σ_{ij} and the distance r_{ij} between interacting partners describing the van der Waals (vdW) force and Coulomb potentials of interacting charges q_i and q_j with the the effective dielectric constant ϵ. (figure 2.2)

The additivity allows to easily expand the force field by new correcting terms or artificial forces. It also necessitate the assumption that all contributions are independent from each other.
The parameters of the force field are either gained by empirical data from experiments or by quantum mechanical calculations of very small sample systems. Since these force fields are not build bottom-up, just following basic physical principles, there are different force fields for different types of biomolecules, like proteins, DNA, RNA, membranes or solvent particles.[6]

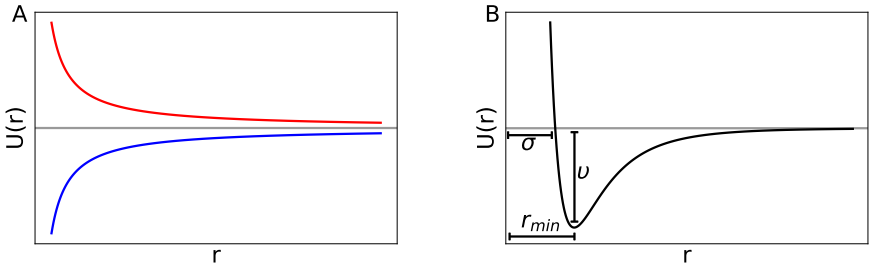


Figure 2.2: **Unbound interactions.** Schematic presentation of attractive and repulsive Coulomb potential (A) and Lennard-Jones potential (B).

2.3 From microcanonical ensemble to isobaric-isothermal ensemble

Since all interactions described so far arise from intra-system forces, the total energy is conserved and we remain in a microcanonical NVE ensemble with constant particle number N, volume V and energy E. This is an unrealistic description of a biological microsystem like a cell or parts of it, which normally sits in an aqueous environment and is therefore coupled to an macroscopic heat bath. This means that energy transfer in and out of the simulated system should be possible. Hence the temperature in our simulation should be regulated to a biologically reasonable value. This is called a canonical ensemble or NVT ensemble.

2.3.1 Temperature Regulation

There are several methods to regulate the temperature towards a desired value. A simple weak-coupling algorithm was described by Berendsen [7]. It rescales the temperature depending on the deviation to a desired temperature T_0 and a time constant τ_T:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \tag{2.6}$$

The temperature is described by the mean kinetic energy of an ensemble's particles.

$$T(t) = \sum_{i=1}^{N} \frac{m_i v_i^2}{N_f k_B} \tag{2.7}$$

Where m_i, v_i and N_f are the particles' mass, velocity and degree of freedom and k_B is the Boltzmann constant. Since the masses and degrees of freedom are not meaningfully adjustable, the velocities have to be manipulated to adjust the Temperature to a certain value. However with this method all velocities are changed the same way, only guaranteeing, that the total kinetic energy and therefore the global temperature are correct. This does not prevent local

temperature divergence.

Another thermostat, which was used for this study, is the Langevin thermostat [8] [9] In Langevin dynamics the Newton equation 2.1 is expanded by two additional terms.

$$m_i \frac{d^2\mathbf{r_i}(t)}{dt^2} \;=\; -\frac{dU(\mathbf{r_i})}{dt} \;-\; \gamma m_i \frac{d\mathbf{r_i}(t)}{dt} \;+\; \eta(t) \tag{2.8}$$

The force is replaced by the time derivative of the potential function $U(\mathbf{r})$. The first addition depicts a solvent drag force, describing the friction with the solvent by collisions using a collision frequency γ. The second addition introduces a random force $\eta(t)$ connected to friction by the fluctuation-dissipation theorem

$$\langle \eta_i(t), \eta_j(t') \rangle = 2m\gamma k_B T \delta_{ij} \delta(t - t') \tag{2.9}$$

The average $\langle\rangle$ is an average over time and k_B is the Boltzmann constant. How strong the particles are coupled to the heat bath is regulated by the collision frequency γ. The delta function δ_{ij} and $\delta(t - t')$ implies that the random force component at one moment t on one particle i is completely uncorrelated to the random force at another moment t' or on another particle j. In reality this is not valid for the time interval of the collision, unless relevant time scales of the simulation are much larger.

Other common thermostats are the Andersen thermostat [10], which describes the energy exchange of the system with the environment by introducing a heat bath which collides stochastically with the atoms, or the Nosé-Hoover thermostat [11], which introduces one virtual particle instead.

2.3.2 Pressure Regulation

For most experiments additionally a constant pressure is preferred. Keeping the number of particles N constant, leads to an NPT ensemble, also called isobaric-isothermal ensemble.
Similar to temperature regulation equation 2.6 can be written for pressure P to describe an easy coupling to an external pressure bath of P_0.[7]

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P} \tag{2.10}$$

Similar to 2.6 a time constant τ_P determines the inertness of the barostat. The pressure P is given by

$$P = \frac{2}{3V} \left(E_{kin} + \frac{1}{2} \sum_{i<j} \mathbf{r_{ij}} \cdot \mathbf{F_{ij}} \right) \tag{2.11}$$

with the total kinetic energy E_{kin} and distances r_{ij} and forces F_{ij} between particle i and j.
Now to regulate the pressure to a desired value, distances between particles r_{ij} and the system's volume V can be manipulated.
Note since all following analysed systems in this study did not undergo major fluctuations of volume, the Berendsen barostat was used.

2.4 Environment

An important part of a MD simulation plays the environment of the analysed biomolecule that is the solvent. In biological systems the solvent mainly consists of water molecules and few ions like sodium and chloride.

2.4.1 Explicit Solvent

The intuitively best method to model the solvent is to calculate its molecules individually. Just as for the solute, simplifications are needed to make the calculations feasible. Since these simplifications are especially required for the smallest and fastest hydrogen atoms, they particularly affect water molecules.
To ensure a reasonable volume of solvent around the solute, the required number of water molecules claim most of the simulation's calculation time. Therefore particular force fields for the solvent were developed. A common water model is the Transferable Intermolecular Potential 3-Point (TIP3P) model [12], [13], which restricts the bond lengths and angle of water molecules to certain values. These parameters together with the near-tetrahedral atomic partial charge distribution and the Lennard-Jones parameters are optimized to describe the properties of liquid water as good as possible. [14]
The water model used in this study was the Optimal Point Charge (OPC) model, which uses four points to optimally reproduce the lowest order moments of water molecules. [14]

2.4.2 Periodic Boundary Conditions

The implementation of explicit water molecules arises the question how many of them are needed and how are they contained, or with other words, what happens at the borders of the simulated volume. An artificial kind of wall would lead to the next question of how particles inside the volume interact with this border and would probably cause undesirable artefacts like particles sticking to the wall [15].
Making the simulation box large enough, that boundary artefacts can be neglected, would not be feasible due to calculation power of today's CPUs / GPUs. A better solution is to introduce periodic boundary conditions (PBC). This means, that molecules cannot interact with the walls, instead if they pass an

"invisible" boundary, they enter the box on the opposite side again. So any particle inside the box does not see any border or exterior of the box.
Another way to understand the concept is, to imagine a lattice of copies of the simulated box volume aligned to each other leading to an infinite volume. (see figure 2.3)

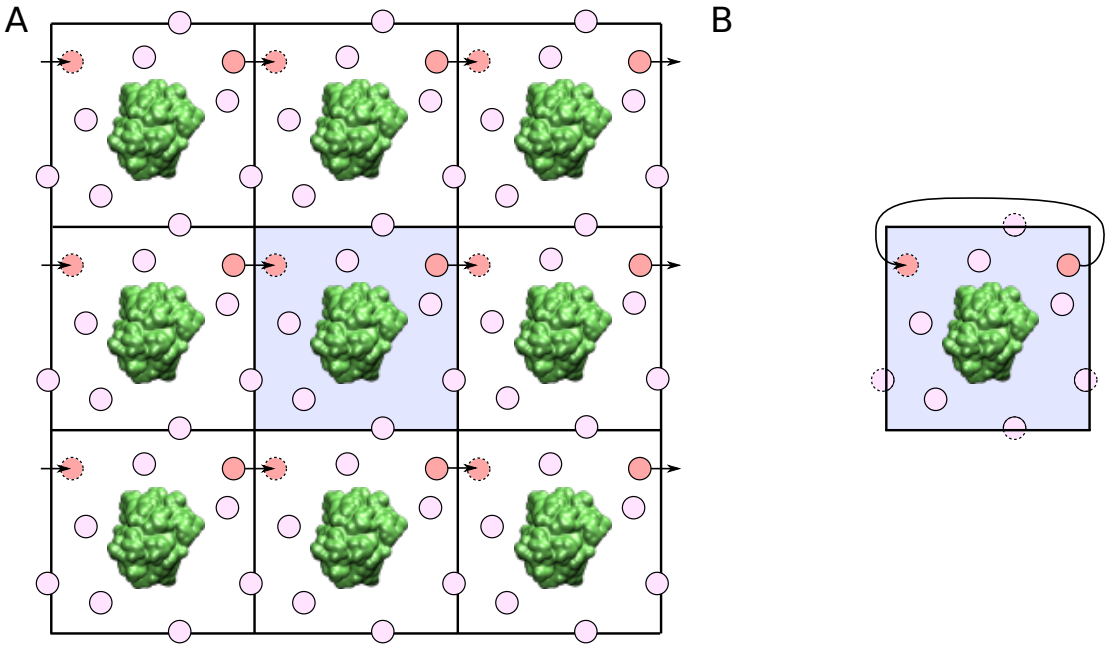


Figure 2.3: **Periodic boundary conditions.** (A) PBC visualized as a grid of copies of the simulation box. (B) PBC visualized as handled in calculations.

This solution avoids undesired boundary artefacts but can cause other problems in practice. To prevent as much computing time as possible one may be tempted to define the box as small as possible. This can lead to undesirable interactions between the big analysed molecule inside with itself or in other words with its own copy, which would probably interfere with its natural behaviour. The risk of this to happen is particularly high, if the molecule drastically changes its shape, for instance by unfolding itself or if a longish molecule rotates in a longish box fitted to its original position. For example if the folding process of a molecule should be simulated, it is possible that parts of the not yet completely folded molecule interact with parts of its copy which perform as a different molecule in this situation. This can either just falsify the observed time interval, the molecule needs to fold, if metastable conformations are formed or make it completely impossible for the molecule to form its biologically relevant conformation, if stable complexes are formed with the copy, leading to an infinite protein complex.

Other problem arises by the theoretically infinite range of non-covalent forces like the vdW force and the Coulomb force.

2.4.3 Cut-Off and Ewald Summation

The most time-consuming processes during MD simulations is the evaluation of energies and forces, especially those resulting from non-bonded Lennard-Jones and Coulomb terms [16]. So every interaction between every molecule within our system has to be considered. In case of a periodic system the number of particles alone is infinite. The calculation of these energies and forces requires $\mathcal{O}(N^2)$ operations, which quickly becomes unhandy or impossible within reasonable time scales with rising particle numbers N, especially if explicit solvent molecules are used.
In case of Lennard-Jones interactions the problem could be solved by cutting them at a certain distance for each molecule, since they decay by $(1/r)^6$. For long ranged Coulomb interactions this method cannot be applied without artefacts at the cutting edge. Instead of simply truncating them, we can artificially divide these interactions in a short ranged and a long ranged part. The short ranged part is used to calculate interactions of a particle with all its neighbours within a cut-off radius in real space. For long ranged interactions the so called Ewald summation is used, which sums the long ranged interactions in reciprocal space. [17] Since summation in reciprocal space over a periodic lattice converges quickly, it also can be truncated at a certain cut-off length. The Ewald summation is a special case of the Poisson summation, which says that a summation in real space can be replaced by a summation in Fourier space. [18], [19]
Using the Fourier transformation for long ranged interactions implicitly assumes a periodic system. The periodic boundary condition (see section 2.4.2) perfectly fits to this approach. The combination of real space particle interactions and long range interactions calculated by a mesh of periodic unit cells by Ewald summation is also called particle mesh Ewald (PME) method and is widely used in MD simulations. [20], [21]

2.4.4 Implicit Solvent

An alternative way to mimic the biomolecules environment would be to just consider one copy of this molecule without periodic repetitions of this unit cell. As mentioned above this again leads to the question, how to confine the molecule and the solvent around and how to treat particles, which cross or touch any artificial border around the system. A very quick method is to put the molecule in an infinite space and give up all solvent molecules. [22] In this way we have only one biomolecule and no borders at all (assuming the biomolecule doesn't consist of multiple subdomains, or these subdomains don't separate significantly). Regardless the effect of the solvent can't be neglected. A way to take the solvent back into account is, to treat it not as discrete molecules but as a continuous mean-field. This approximation ignores all distinct interactions between single

solvent molecules and the solute molecule, but reduces the enormous effort of calculating all solvent molecules. [23]
What is basically left to calculate is the solvation free energy contribution to the total energy of the simulated molecule.

$$E_{tot} = E_{vac} + \Delta G_{solv} \tag{2.12}$$

E_{vac} is the molecule's potential energy in vacuum and determined by the force field. Its calculation is comparably uncomplicated. [24] The energy to put the molecule into the solvent ΔG_{solv} can be split into two parts:

$$\Delta G_{solv} = \Delta G_{nonpolar} + \Delta G_{el} \tag{2.13}$$

The non-polar part contains attractive van der Waals interactions between solute and solvent molecules on the one hand and hydrophobic repulsion by disturbing the structure of the surrounding solvent on the other hand. [23] Both contributions can be linearly approximated by the solvent-accessible surface area (SASA), which can be calculated by 'rolling a ball' (corresponding to a solvent molecule) over the molecules surface. [25]

$$\Delta G_{nonpolar} = \sigma \cdot SASA \tag{2.14}$$

The proportionality constant σ was derived by experimental solvation energies of small non-polar molecules. [23]
This leaves the electrostatic contribution ΔG_{el} to the solvation free energy, which is long-ranged and most time consuming to calculate. An appropriate description of the electrostatic potential $\Phi(\mathbf{r})$, which is generated by a molecular charge distribution $\rho(\mathbf{r})$, delivers the Poisson-Boltzmann (PB) equation, which is derived via mean-field approximation [23], [26]:

$$\nabla[\epsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) + \kappa^2\epsilon(\mathbf{r})\Phi(\mathbf{r}) \tag{2.15}$$

where $\epsilon(\mathbf{r})$ describes the position-dependent dielectric constant. The additional term with the Debye-Hückel parameter $\kappa \sim \sqrt{[salt]}$ models electrostatic interactions of salt ions in the solution. [22]
With the numerical solution of this equation, the electrostatic part ΔG_{el} of the solvation energy can be calculated. The combination of MD, PB and SASA to calculate free (binding) energies is called the Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) method.
Since this relatively expensive method with regard to computational effort has to be done every step during an MD simulation, the even more simplified Generalized Born (GB) model can be used [27]:

$$\Delta G_{el} \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_w}\right)\sum_{i,j}\frac{q_i q_j}{f_{GB}(r_{ij})} \tag{2.16}$$

$$f_{GB}(r_{ij}) = \sqrt{r_{ij}^2 + R_i R_j e^{-\frac{r_{ij}^2}{4R_i R_j}}} \tag{2.17}$$

The dielectric constant inside the molecule ϵ_{in} is normally around 1, whereas the dielectric constant of the surrounding water ϵ_w is high. So the first term can be further approximated to $(1 - 1/\epsilon_w)$. [23] The sum over atoms i and j contains partial charges q_i and q_j, as well as the distance r_{ij} between these atoms and their effective Born radii R_i and R_j. The latter are defined as the radius the whole molecule without any partial charge but the one of the considered atom would have if it was a spherical ion with the same ΔG_{el}. In other words it describes, how deep an atom is placed inside the solute. In practice the effective Born radii are calculated by integrating over the interior of the solute excluding a sphere around the atom of interest. [28]
The function $f_{GB}(r_{ij})$ is equal to the effective Born radius R_i if $i = j$. For $i \neq j$ it becomes the effective interaction distance which approximates to the real distance r_{ij} between atoms i and j.
Like the PB equation it can be extended by a correction for screening effects by salt ions. [29]
The GB equation can be understood as an interpolation between analytic solutions for a single sphere on the one hand and for broadly separated spheres. From computational point of view the GB equation is much faster than the PB equation. It is a widely used compromise between approximating reality and computational effort. One should keep in mind, that the GB model is not as accurate as explicit water models. For instance all effects involving tightly bound water molecules, which may be important for a biomolecule's stability, are neglected. Furthermore all solvent is treated the same way within this approach. If a molecule's shape is strongly curved or has deep binding pockets, solvent inside this regions may behave different in reality than bulk solvent.
The combination of MD, GB and SASA is called Molecular Mechanics Generalized Born Surface Area (MMGBSA) method.
In summary, the total free energy difference of different conformations can be calculated:

$$\Delta G_{tot}^{conf} = \Delta G_{forcefield}^{vac} + \Delta G_{nonpolar}^{solv} + \Delta G_{el/GB}^{solv} \tag{2.18}$$

$\Delta G_{forcefield}^{vac}$ contains all bonded (distance, angle, dihedral) and non-bonded (electrostatics, Lennard-Jones) contributions from the protein force field (see equation 2.5).

Chapter 3

Folding Process of Collagen Peptides

Each protein is characterized by its sequence of amino acids. This sequence, which is encoded in a cell's Deoxyribonucleic Acid (DNA) is similar to a protein's fingerprint. But whereas we understand a fingerprint to be a unique attribute of a single entity, the sequence of a protein is only an attribute for one certain species of proteins. Nevertheless, it is the first necessary piece of information needed to build a specific protein.
After all the amino acids have been joined, the resulting chain must be brought into its correct form. This is essential for proteins' stability and functions [30], [31]. Correctly arranging such a chain of up to thousands of elements appears much more complex than putting the chain itself together. It is all the more astonishing that the mere specification of a certain sequence always leads to the same respective structure. The process of this so-called folding of a protein is illustrated in this chapter using collagen peptides as an example. Furthermore the effect of small changes in the sequence is analysed.
The content of this chapter has been previously published in a similar form [32].[1]

3.1 Importance of Collagen and its structure

With one third of human proteins, collagen is the most abundant protein of our body. It appears in many different tissue types like skin, cartilage, bone or hair and plays an essential role for the stability of the extracellular matrix and whole body. [33] The various types of collagens allow them to serve different functions ranging from stiff structures like bones to elastic tissue like skin or cartilage.
The characteristic feature of collagens is a parallel right-handed triple helical structure which was already assumed by Ramachandran and Kartha [34], Rich and Crick [35], and Cowan and co-workers [36]. It consists of three polypeptide

[1]

chains, that form a left-handed poly-proline II-type helical coil. For this conformation it is important that every third residue is a glycine (Gly) [37] resulting in the characteristic Gly-X-Y repeating unit with X and Y mostly representing proline (Pro) or hydroxyproline (Hyp).
The structures of various triple-helical structures have been determined by X-ray crystallography [33]. Sequence variants are known in different collagens that are associated with several human diseases like Osteogenesis Imperfecta [38], Ehlers-Danlos syndromes [39], Alport syndrome [40], [41] and many others. Mutations in collagen can result in a destabilization, misfolding or delayed folding of the collagen fibre. Understanding such defects and design of treatments requires a comprehensive understanding of the structure formation processes.
Kinetic measurements of folding rates for collagen-like peptides suggest that folding consists of several steps, which include a first formation of a triple-helical nucleus followed by rapid propagation of the triple-helical structure from the C-terminus to the N-terminus in a "zipper-like" mechanism [42]–[45]. The rate determining folding steps correspond to the formation of a sufficiently long and stable initial triple-helical segment and the cis-trans isomerisation of Gly-Pro bonds that can interrupt or prolong the subsequent propagation step [42], [45]. It has been possible to determine the rate of nucleus formation for model peptide chains and to conclude that the nucleus formation is associated with a purely entropic barrier (determined by the speed with which the three strands diffuse together to form a nucleus consisting of $\sim$3.3 tripeptide units per strand). The subsequent propagation step is too fast to be measured by the fastest available kinetic mixing experiments [44], hence, propagation (with all-trans Pro) must happen in a time regime significantly below 1 ms.

In vivo, collagens are formed in the Endoplasmic reticulum (ER) and involve several chaperone molecules including the collagen specific heat shock protein 47 (HSP47). Initiation of the collagen folding process starts at the C-terminus and involves disulphide bonds in the non-helical procollagen part, which is cut away at the final stage [46]. Since collagens can include more than 1000 residues, any incorrect propagation can prevent folding completion.
The HSP47 chaperone can bind to already folded parts of collagen [47] and stabilizes the folded part allowing further propagation to continue towards the N-terminus.
Experimental biophysical kinetic measurements are well suited to study the initial nucleation of triple-helix formation on model systems [44], [45]. However, the time resolution is insufficient to study the propagation steps that are strongly affected by known mutations and give rise to various diseases. Molecular Dynamics (MD) simulations are well suited to study structure formation processes at atomic resolution. Several MD simulation studies have already been performed on triple-helical collagen model peptides [48]–[53], however typically starting from the already folded triple helix to investigate the local dynamics and the effect of substitutions [51]–[53]. The triple-helix folding process has also been studied but was guided by a force to gradually move the unfolded system towards the known folded structure [52]. However such added external driving forces may

artificially bias the structure formation processes.
In the present study, we use multiple MD simulations starting from an already formed folding nucleus to directly follow triple-helix propagation.
The diffusive formation of an initial nucleus occurs on time scales beyond current MD simulations [44]. However, for the natively folded triple helix formation of long collagen fibrils the propagation steps are of central importance.
The simulations allow us to give an atomic detail picture of the structure formation process. We find that the folding follows a sequential process with a first transient formation of two chains forming a short 3 residue folding template. The third chain then propagates by using this segment as template for folding to complete one folding propagation cycle. We also study how an unstable nucleus affects the folding process and find that it dramatically lowers the number of successful folding simulations. However, if the process successfully starts, folding follows the same mechanism as observed for the simulations starting from a stable nucleus. Finally, we also characterize misfolding events and study systematically the substitution of a central Gly by alanine (Ala) or threonine (Thr) on the folding propagation.

3.2 Simulation Setup

The experimental structure of a collagen-like triple helix (PDB 3b0s [54]) served as starting and reference structure. This structure consists of 27 residues for each strand of the triple helix, resulting in 81 residues in total. However, in order to allow for multiple extensive folding and unfolding simulations we reduced the system to a folded triple helix consisting of three strands with 15 residues per strand with the sequence $(\text{Gly-Pro-Pro})_5$ still including 5 repeating tripeptide units per strand.
Simulations on mutations included the in silico replacement of the central Gly residue in one strand or all three strands by either Ala or Thr.
For all simulations the AMBER14 [55] package in combination with the parmff14SB [56] force field for the peptides was used.
To each system explicit water molecules (Optimal Point Charge (OPC) model [14] see 2.4.1) were added in a cuboid box with 15Å distance to the molecule for peptides without any or only a single-strand mutation and 35Å distance to the molecule for peptides with mutations in all three chains. Sodium and chloride ions were added ($\sim$0.1 M) to neutralize the system and to create a physiological salt concentration.
The masses repartitioning option for hydrogens and heavy atoms was used to allow for a large time step of 4 fs during simulations [57]. Without masses repartitioning time steps of 2 fs are usually made. The repartitioning "shifts" mass from heavy atoms to hydrogen atoms, resulting in slower oscillating hydrogen atoms.
Long-range electrostatic interactions were included using the particle mesh Ewald method [20], [21] (see 2.4.3) in combination with a 9Å real space cut-off.
After energy minimization (1000 steps steepest gradient method followed by 1500

steps conjugate gradient method) the systems were heated to 300 K for several nanoseconds simulation time. To generate unfolded triple helical structures the systems were heated to 700 K for 10 ns and cooled down to 310K again. (see figure 3.1) To avoid trans-cis isomerisation at Pro residues during this phase a dihedral restraining potential to keep a trans configuration was added.

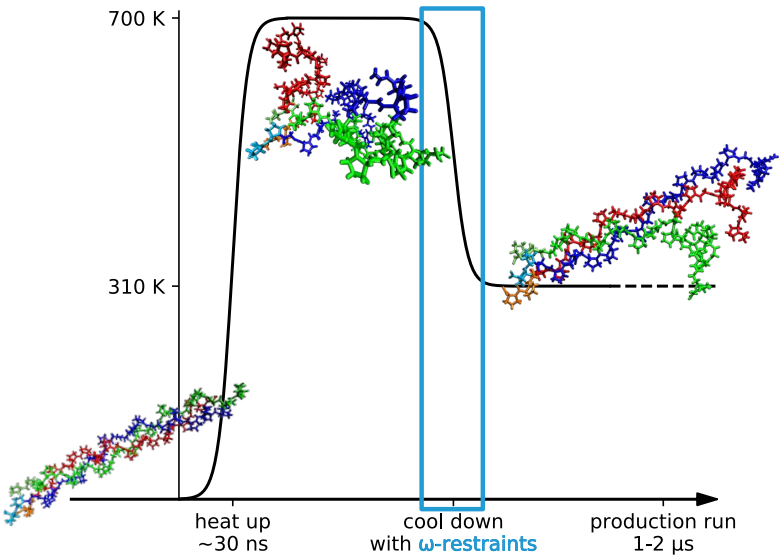


Figure 3.1: **Schematic time schedule of the simulation process.**
To unfold the triple helix, collagen peptides were heated up to 700 K for several nanoseconds. During this hot phase some ω-angles can flip to undesired cis conformation. Therefore these angles were restrained to trans conformation during the cooldown phase. Positionally restrained residues during the whole simulation are coloured in light blue, orange and lime.

During unfolding and the folding propagation run the C_α atoms of the two first residues of each strand were harmonically restrained to the reference structure at the C-terminus in order to mimic a defined (but still flexible) nucleus of the triple helical folding process corresponding to an already folded triple helical segment.
The force constant for this restraint was optimized at 0.5 kcal·mol^{-1}·Å^{-2} to optimally represent the flexibility of an already formed triple helical segment (see A.2). Furthermore, simulations with positional restraints at a reduced force constant of 0.05 kcal·mol^{-1}·Å^{-2} were also performed to mimic a badly formed

previous segment of the helix.
Each folding simulation started from a different unfolded structure and was extended to 1 µs after the cooldown process.
In addition to simulations starting from unfolded systems, simulations starting from the folded triple helix with and without residue substitutions were also performed to analyse the stability of an already folded helix.
For each setup 10 simulations were performed. Trajectory snapshots were taken every 50 ps. Bonds involving hydrogen were constrained by SHAKE [58]. The temperature was controlled by a Langevin thermostat. (see 2.3.1) This thermostat was chosen to create more variance between multiple simulations of the same system setup.
To determine triple helix folding times the root-mean-square deviation (RMSD) of the third residue triplet (meaning residue 4 of each chain starting from the N-terminus) was recorded. The open end of an fully folded helix opens from time to time along few residues but also closes again. The third residue triplet was preferred to the first one to avoid opening and closing fluctuations of the latter. A window of 50 frames (2.5 ns) was shifted through the trajectory and the first moment the mean value of the window was below a threshold of 1 Å was defined as folding time.
All non-hydrogen contacts of the crystal structure with maximum distance of 4.0 Å were counted as native contacts.
Energies were calculated using the Molecular Mechanics Generalized Born Surface Area (MMGBSA) method as implemented in the Amber package. [59] (Amber input options: PBRadii = mbondi3, igb = 8, salt concentration = 0.1M). To flatten the energy time course, a Gaussian filter ($\sigma = 25$ frames) was used. The energy differences between the beginning and the end of a simulation are the differences of the mean energies of the first respectively last 20 frames of the simulation.

3.3 Results and Discussion

The simulation model system consists of three chains $(\text{Gly-Pro-Pro})_5$ forming a collagen like triple helix (figures 3.2 and A.1). In order to specifically study the propagation during MD simulations we added an artificial harmonic restraint to the two residues at the C-terminus of each strand (only acting on the backbone C_α coordinates) to keep this nucleus on average reasonably close to the structure in a folded reference triple helix. It mimics an already formed triple helical nucleus structure at the C-terminus still allowing fluctuations of the restrained segment. Care was taken to adjust the flexibility to a level comparable to the fluctuations of a regularly folded triple-helical structure (figure A.2).

Multiple folding propagation simulations of the $(\text{Gly-Pro-Pro})_5$ structure were performed using in each case different unfolded starting structures (see table 3.1). Out of 10 simulations 9 successfully reached the completely folded triple helix within <1 µs simulation time. The finally reached state is indistinguishable

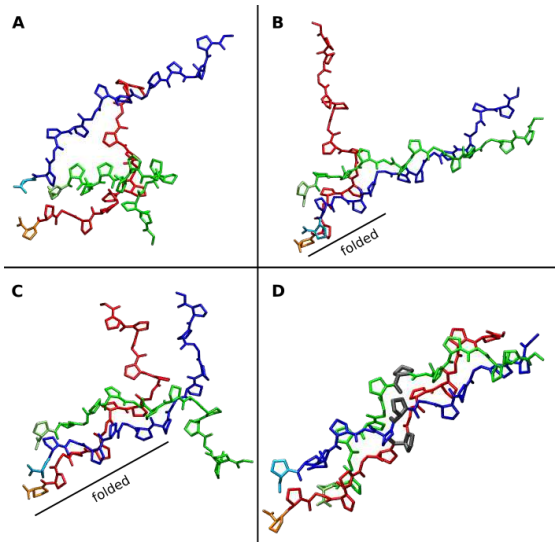


Figure 3.2: **Representative snapshots from a collagen folding propagation simulation.** (A) Initial random arrangement of the three $(Gly\text{-}Pro\text{-}Pro)_5$ chains (coloured red, green and blue). (B) Folding nucleus formed at the C-terminus due to positional restraints on C_α atoms of the last two residues of each strand and partial formation of a near native structure of two chains (blue and green). (C) folding propagation with a near native arrangement of all three chains formed up to the middle of the complex. (D) Completely folded collagen triple helix (the 3 residues indicated in grey have the same residue number in each strand, are spatially close in the native structure and represent a residue triplet).

from a simulation started from the folded conformation (figures A.1 and A.3). Once folded the triple helix is stable for the rest of the simulation. In contrast to a mostly two-state folding of globular proteins with little accumulation of stable intermediates, the folding propagation process of the collagen triple helix follows a stepwise process (figures 3.2, 3.3, A.4 and A.5). Starting from the C-terminus the RMSD with respect to the folded state of the following segments decreases progressively along the strands with intermediates that differ in length by $\sim$3 residues until the complete triple helix has formed.
Similarly, the near-native contacts increase progressively over time with occasional steep rise during folding once a propagation step is completed (figure 3.3). The time between each successive folding progression of $\sim$3 residues varies between a few ns and a few hundred ns (figures 3.3, A.4 and A.5).

If one looks separately at the backbone dihedral transitions of the individual chains, another repeating pattern becomes visible: In a folding propagation step a segment of three residues in two strands typically forms first a transient associated native-like structure with the residues from the third chain still unfolded.

simulation N[a]	wild-type[b] Å (ns)	weak restraints Å (ns)	G7aA Å (ns)	G7aT Å (ns)	G7abcA Å (ns)	G7abcT Å (ns)
1	1.4 (44)	14.7 (-)	1.6 (879)	5.9 (-)	3.9 (398)	4.8 (677)
2	2.5 (960)	10.1 (-)	6.8 (-)	3.0 (475)	10.8 (-)	15.4 (-)
3	1.6 (319)	1.6 (706)	7.6 (-)	5.0 (-)	10.4 (-)	9.2 (-)
4	1.4 (138)	2.5 (139)	9.4 (-)	2.9 (280)	3.9 (907)	4.8 (-)
5	1.5 (96)	11.2 (-)	2.4 (769)	2.9 (494)	8.4 (-)	18.7 (-)
6	1.4 (367)	9.0 (-)	13.3 (-)	1.9 (992)	4.2 (83)	6.2 (-)
7	1.7 (331)	7.9 (-)	4.6 (-)	9.8 (-)	4.4 (555)	13.9 (-)
8	7.8 (-)	13.3 (-)	11.8 (-)	1.7 (494)	8.6 (-)	5.1 (-)
9	1.4 (95)	1.9 (89)	2.0 (905)	3.1 (428)	3.7 (837)	14.1 (-)
10	1.6 (45)	1.7 (390)	2.7 (635)	6.2 (-)	4.0 (540)	9.1 (-)

Table 3.1: **RMSD and folding time of triple helices starting from 10 different start structures**
[a] N indicates the starting structure or simulation number.
[b] Each column corresponds to the wild-type (WT) or mutated collagen sequence; G7aA means mutation only in chain A, G7abcA: mutation in every chain; RMSD is given in Å vs. native triple helix structure for the finally sampled frame (at 1 μs). In brackets the time (in ns) is given, after which the peptide was (first) completely folded.

The short two-strand segment forms a template or binding site for the third chain to fold along the template to complete the folding of a 3-residue segment for all three chains.
This mechanism is illustrated in figure 3.4 as a sequence of successive backbone dihedral transitions. In the unfolded state a broad range of Ψ/Φ dihedral angle combinations (including near-native states) are sampled but for a folded segment only a narrow range of Ψ_{Pro}/Φ_{Gly} combinations characteristic for a triple helix are sampled (figure 3.4).
Furthermore, on another example the process is illustrated by recording the pairwise RMSD of the 7^{th} residue of two selected chains (figure A.5). Here, first chain B and C align to each other (resulting in low and stable RMSD of the 7^{th} residue relative to the native triple helix) but they unfold again (chain A is not folding). However, at a later time point chain A and C align to each other followed shortly later by folding of chain B on this template to the stable triplet position including complete formation of all native contacts between the chains (figure A.5). This example demonstrates that each chain alone or parts of it can adopt the correctly folded state several times before it eventually reaches the stable state next to the two other chains. Thus the population of dihedral angles in the "folded" state area in figure 3.4 occurring in the unfolded segments (red background) can be explained by the transient folding of individual chains. The transient formation of near native structures for individual chains is also illustrated in figure A.6. In short time intervals of few ns individual chains can

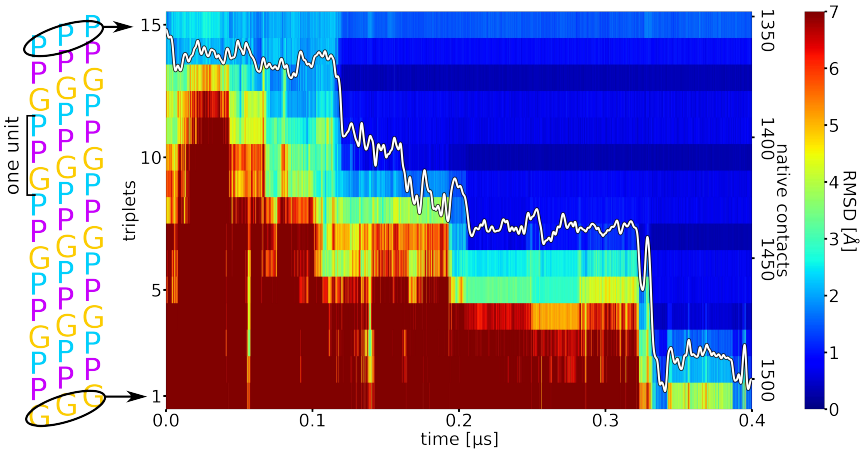


Figure 3.3: **Time course of triple helix formation during an MD simulation.** The simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (see figure 3.2 A). Each labelled stripe (1–15) represents a residue triplet (residues with same number in each chain, indicated as black ellipse in the left panel) along the three strands. The root-mean-square deviation (RMSD) of each residue triplet relative to the native folded structure is indicated by a color-code (color bar on the right side). A blue color represents sampled states close to native, whereas red color corresponds to an unfolded triplet structure. The y-axis on the right side of the plot gives the number of formed native contacts (white line in the plot).

indeed reach an RMSD (non-hydrogen atoms) <1.5 Å relative to the native structure.

The grouped propagation process arises from the repeating sequence of each chain (schematically illustrated in figure 3.5). Since every third residue is a Gly, these are the most flexible points of the sequence. The backbone dihedral angles of the Pro residues are nearly unaltered comparing folded and unfolded peptides except for the Ψ angle at the connection to a Gly. For Gly the Φ angle differs most significantly between unfolded and folded state (figure A.7). Taking all successful folding simulations into account the folding propagation was on average completed after ∼290 ns. This translates to an average formation time of one repeating 3-residue unit (forming basically the elementary propagation step) of ∼75 ns. It also justifies our maximum simulation time of 1 μs that is ∼15 times longer than the elementary folding step. Reducing the restraints on the backbone C_α atoms at the C-terminus (to 0.05 kcal·mol^{-1}·Å^{-2}, which allows in principle fluctuations of the restraint atoms by up to 4 Å within a mean energy change of RT = 0.6 kcal·mol^{-1}, R: gas constant, T: temperature of 310 K) led to highly mobile C-terminal start arrangements and to a significantly smaller fraction of successful folding propagation processes (Table 3.1 and figures 3.6 and A.8).

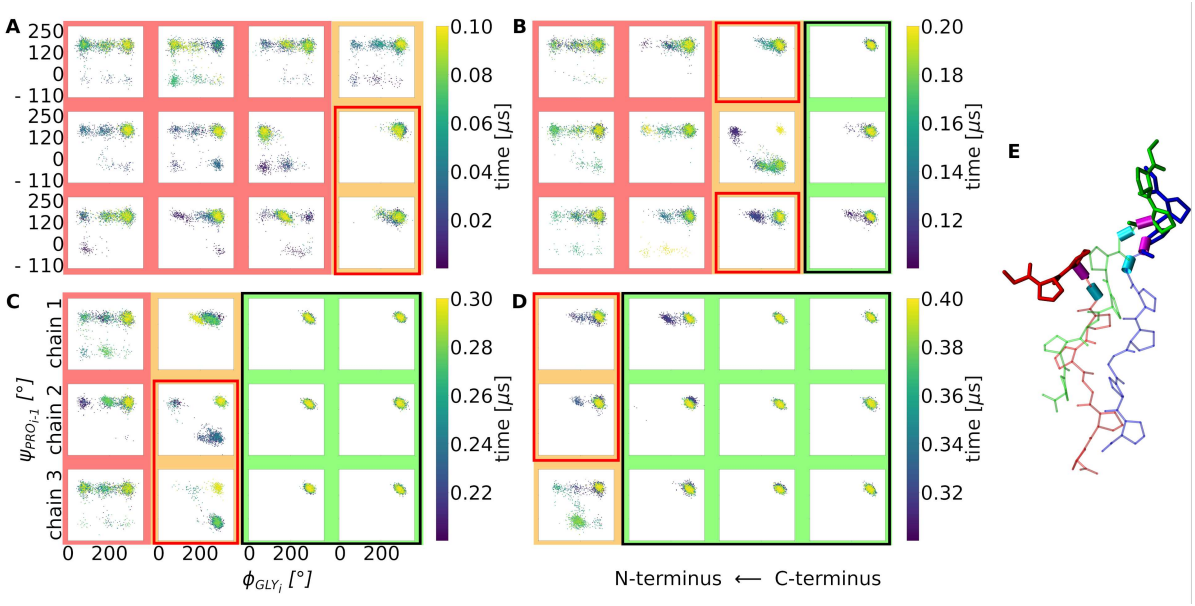


Figure 3.4: **Sequential collagen folding process in terms of successive dihedral angle transitions.** (A-D) 2D-Ramachandran type plot of the most flexible backbone dihedral angles (Φ of Gly_i and Ψ of neighbouring Pro_{i-1}) in each chain of the triple helix during subsequent time periods of a folding simulation (simulation time increases from A to D). The already folded part of the triple helix, starting from the right side (C-terminus) is framed in black and highlighted by a green background (in A-D) with both dihedral angles sample the native conformation in the upper right corner of the plots. In contrast, dihedral angles cover a broader distribution in the still unfolded, disordered segments of each chain (red background). The part in between (orange) indicates an intermediate state, where two chains adopt already a native dihedral angle configuration while one is still in a non-native configuration. The process is repeated in a step-wise manner in A-D in the direction of the N-terminus of the chains. (E) Simulation snapshot indicating a partially folded triple helix with two chains overhanging by three residues in near-native geometry (blue and green) and one strand with the corresponding three residues still unfolded (red sticks). The rotatable bonds that define the Φ of Gly_i and Ψ of Pro_{i-1} are shown as cyan and magenta cylinders, respectively.

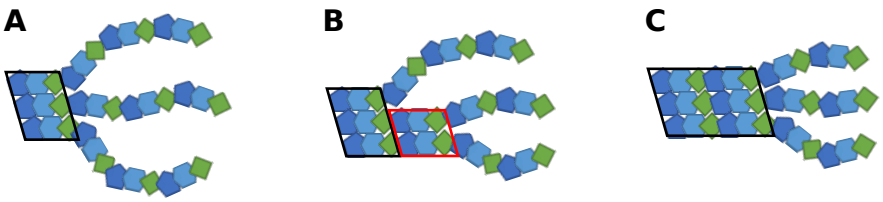


Figure 3.5: **Schematic illustrations of the sequential folding propagation process.** (A) Starting with an already folded part of the triple helix (framed by a black box in A, each residue is represented by a coloured bead). (B) Followed by the formation of a near-native structure of two chains (red box in B) and subsequent folding of the third chain (framed box in C) to form an extension of the triple helix by three residues of each chain.

This observation indicates the importance of a stable folded nucleus close to the consensus triple helix near the C-terminus for further propagation. In this case only 3 of 10 starting arrangements folded correctly to a structure in close agreement with the experimental reference triple helical structure (figure 3.6). However, for the successful trajectories that formed a stable nucleus with a growing tip the same stepwise mechanism and similar folding propagation times were observed (figures 3.6 and A.8). It emphasizes the importance of a correctly folded and stable nucleus close to the consensus triple helix near the C-terminus. Once a stable nucleus has formed stepwise propagation can proceed rapidly as has been indicated also in previous experimental studies [44]. Nevertheless, especially in the simulations with a highly mobile C-terminus several transiently stable misfolded and intermediate conformations were observed. It includes the formation of bulged loops formed by one strand during propagation steps. Interestingly, the loop regions always consist of multiples of three residue segments. Hence, three or a multiple of three residues (ending with Gly) can loop out and the triple helix eventually continues following the consensus structure after shifting the alignment with respect to the looped out strand (figure 3.7 A-C). Accidental formation of such a bulge loop is critical because if propagation eventually proceeds there is no way to resolve this misfolded loop.

The fibril is on average kinked at the bulge (figure 3.8). This can of course strongly affect its ability to associate with other fibrils to form a stable bundle. In other cases, an unstable nucleus resulted in a non-native association of the chains during propagation that was stable for the rest of the simulation (figure 3.7 D and E). An unstable C-terminal nucleus also results frequently in another type of misfolded triple helices that are locally arranged in a native-like triple helical structure but the chains are shifted relative to each other.

These shifts occur generally in units of the sequence repeat (example cases observed in simulations with an unstable nucleus are illustrated in figure 3.9). A misfolded stable structure with a loop was only observed once in the case of reasonable stabilization of the initial folding nucleus (simulation 8, see Table 3.1),

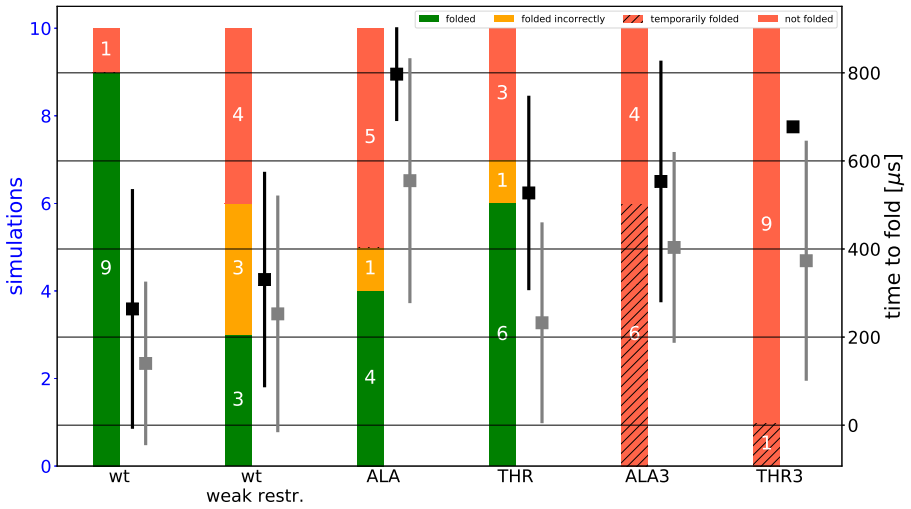


Figure 3.6: **Efficiency and folding times of all simulations.** Each column includes 10 simulations. Successful simulations are coloured green. Simulations reaching only partly helical and stable folded peptides are coloured orange and all simulations which resulted in an unfolded peptide are coloured red. Red hatched bars indicate simulations which temporarily folded but subsequent unfolding. Black squares represent the average time to observe a completely folded triple helical structure. Grey squares represent the same time to result in folding up to half of the fully folded structure. Weaker restraints (positional restraints with respect to the native structure on the C_α atoms of the last 2 residues of 0.05 kcal·mol^{-1}·Å^{-2} vs. 0.5 kcal·mol^{-1}·Å^{-2} for the standard set of simulations) do not modulate the folding propagation time but the efficiency and the appearance of misfolded structures. A single point mutation of G→A/T at the center of one chain already increases the folding time and lowers the efficiency. Three point mutations of G→A/T (one at the center of each chain) do not result in a single successfully and completely folded triple helix.

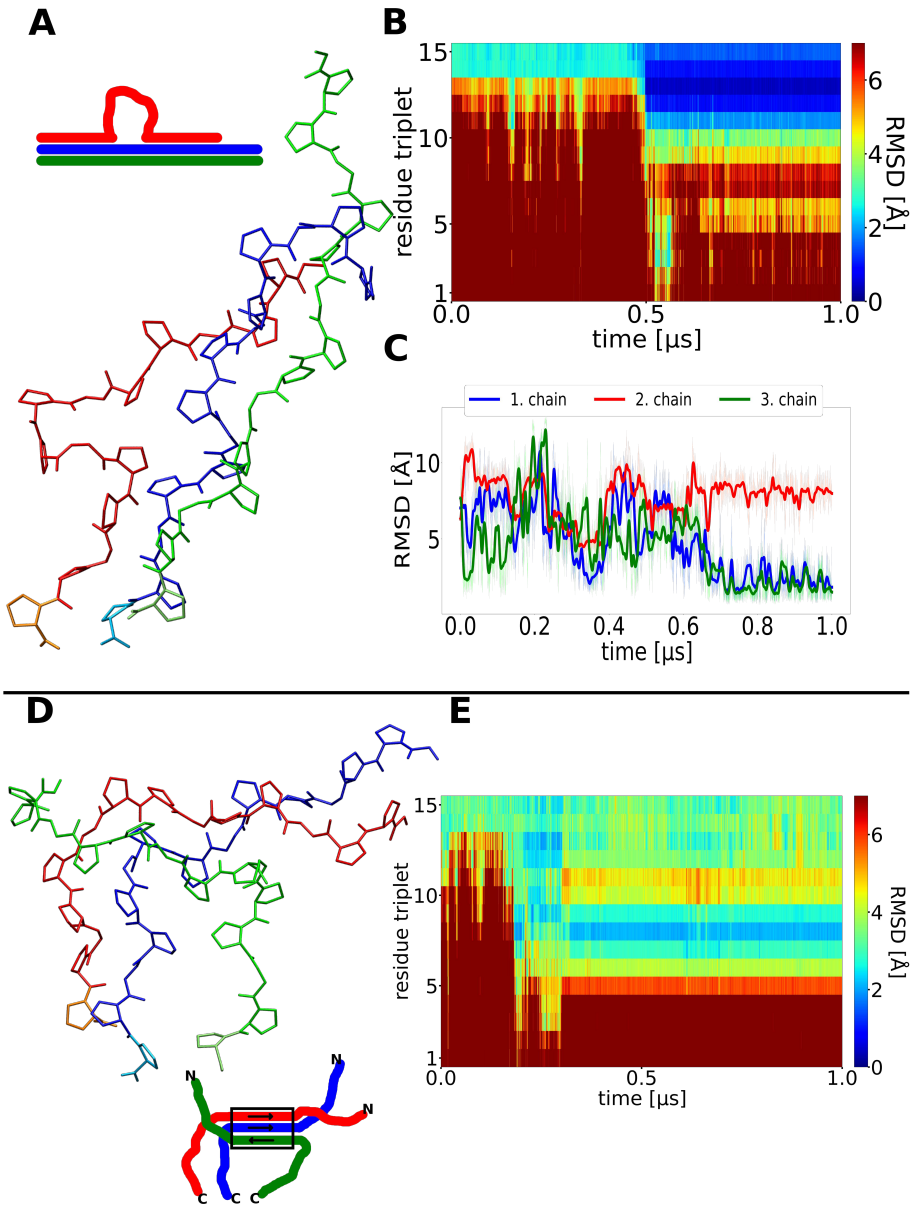


Figure 3.7: **Misfolding of collagen triple helices during simulations.** (A) Snapshot of a finally sampled conformation with one chain (red) forming a loop but near-native conformation of the neighbouring segments observed in a simulation that did not reach the native fold. (B) Time evolution of the RMSD of triplets along the simulation, after 0.5 μs a correct folding of a first segment up to the loop is observed (blue regime in the plot). (C) RMSD with respect to the native folded triple helix of each chain vs. simulation time. (D) Snapshot at the final stage of a simulation with very weak restraints to stabilize the initial folding nucleus at the C-terminus of the triple helix. The green chain has detached from the other chains at the C-terminus and binds to the other two chains in a non-native arrangement. (E) The non-native arrangement remains stable for the rest of the simulation (after ∼0.4 μs).

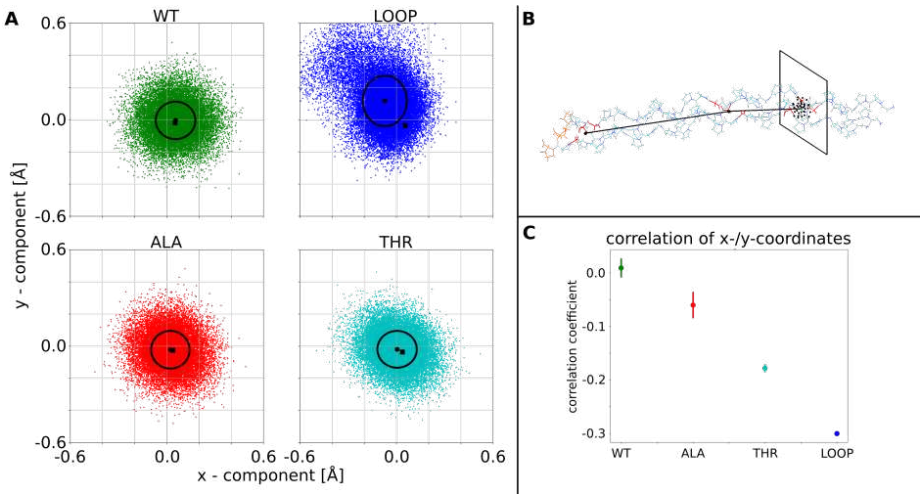


Figure 3.8: **Influence of central G→A/T substitution or loop in one chain on the flexibility of the triple helix.** (A) Each point in the 4 plots is the projection of a vector from the center of the folded triple helix to the N-Terminus on the plane perpendicular to the vector pointing from the C-Terminus to the center of the triple helix. WT indicates the natively folded triple helix with an isotropic distribution indicating a fluctuation around the straight triple helix (standard deviation of the distribution indicated as black circle). The LOOP, ALA, THR label indicate simulations with a central loop segment of one chain (see also figure 3.7 A), a central G→A or G→T mutation in one strand, respectively. The LOOP case indicates a significant directional bending fluctuation. (B) illustration of the vectors and projection. (C) correlation of x/y-displacement for the WT, the single G7aA and G7aT mutations and the triple helix with the looped out segment in one chain. Increased correlation indicates anisotropic bending fluctuations of one end of the triple helix relative to the other end.

but several times for the case of a weakly stabilized starting segment indicating again the critical importance of a stable folding start point to proceed to the correct folding propagation.
Since collagens are the product of very long genes it is likely that mutations occur at some positions and may accumulate during evolution. Mutations containing amino acids with larger side chains can lead to more significant alterations in the structure [37], [42]. Since a number of mutations have been linked to important connective tissue diseases, a thorough understanding of factors that affect the folding of collagen is of particular interest [47], [60], [61]. Most critical is the substitution of the flexible Gly (G) by other residues and efforts were focused on investigating the effect of G→A and G→T substitutions in either one chain or all three collagen triple helix forming chains. For instance, peptides containing a G→A mutation can form triple-helical structures that contain a local distortion at the site of the mutation and a disruption of the normal collagen hydrogen bonding pattern [62], [63].
We performed 10 folding propagation simulations with one chain containing a central Gly7Ala (G7aA) or a Gly7aThr (G7aT) mutation (figure A.8). In both cases only about half of the simulations resulted in successful propagation to the fully folded triple helix (and typically with a higher final RMSD relative to the experimental structure, table 3.1). Some of the cases that did not propagate to the full triple helix still reached folded conformations up to the position of the substitution but the rest of the chains remained in an unfolded state (figures A.9 and A.10).

Interestingly, the average folding time in the successful cases rose to 900 ns in the G7aA case and 590 ns in the G7aT case qualitatively consistent with experimental result on decreased folding rates of G→A replacements in collagen model peptides [64]. Since the side chains of the mutated residues in G7aA and G7aT point inwards towards the center of the triple helix the larger side chains create a sterical barrier that hinders and delays the propagation process. Such an arrangement is also seen in the X-ray structure of a collagen peptide with Gly→Ala substitutions [63].
Substitution of the central Gly in all chains at the same position finally made it impossible for the three peptide chains to fold to a stable triple helical structure (figures 3.10, A.12 and A.13). However, in 6 out of 10 simulations of the all G7abcA case at least transiently a structure close to a full triple helix was observed that unfolded again after a short period (∼80 ns for the case illustrated in figure 3.10). The result is consistent with the observed strongly reduced melting temperature of collagen-modelling triple helices due to Gly→Ala substitutions [63], [64]. For the Gly→Thr substitution in all chains only one simulation temporarily reached a folded state, none of the 10 simulated peptides stayed in this state until the end of the simulations. In this case the measured "folding times" represent the moments when the RMSD of the N-terminus was the first time at the level of the folded state. For both cases the average was in the range observed for the mutations in a single strand.
Based on an energetic analysis of the folding process using a continuum solvent model we also estimated the mean energy change during the propagation process.

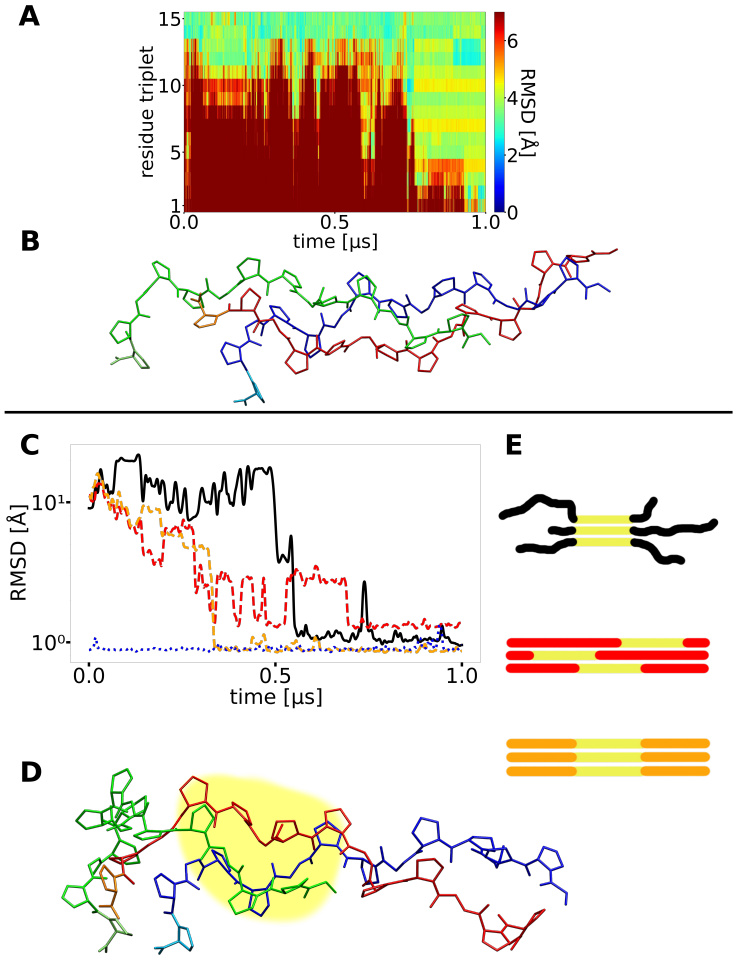


Figure 3.9: **Misfolding due to chain shifts.** (A) A second type of misfolding observed with very weak restraints on the C-terminal folding nucleus with the RMSD of all triplets reaching uniformly ~ 4 Å with respect to the native structure at ~ 1 µs simulation time. (B) A snapshot of the final conformation reveals, that large parts of the triple helix are formed, but the strands are partially shifted relative to each other causing a significant deviation of each triplet from the placement in the native structure. (C) The RMSD of segments of 3x6 residues (yellow part in I and J) of different simulations are plotted and illustrated. The black curve shows a simulation which formed a partially helix-like structure that prevented the rest of the helix from proper folding (snapshot in D and schematic illustration in E (top)). The red and orange curves show different segments of a successfully folded simulation (illustrated in E). The blue dotted example represents a helix folded from the beginning. The black curve gets close to the blue and orange one, indicating the helical structure of the yellow part. At the same time the RMSD of the same segment/residues in a folded helix is higher (red curve), the here formed structure (D) is different from the native helix.

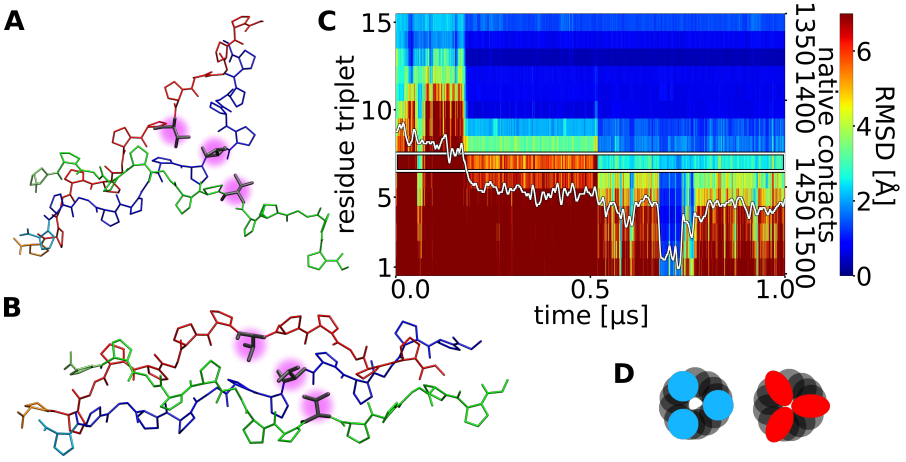


Figure 3.10: **Residue substitutions can strongly affect triple helix folding.** (A) Snapshot of an intermediate state with the triple helix formed up to the center of the chains that each contain a central G→T substitution (magenta sphere). (B) Short-lived sampled snapshot of the same chains, temporarily folded to a triple helical conformation, but the middle part around the mutations adopting an expanded structure. (C) RMSD plot of residue triplets along the triple helix (one residue of each chain per triplet) and native contacts. The blue part of the plot indicates correctly folded triple helix, at ∼0.7 µs the complete near-native folded triple helix is observed but unfolds again at ∼0.78 µs. (D) schematic cross section of helices at the position of the mutations to illustrate how the mutated residues (red) sterically collide with the other chains, compared to the WT helix (blue).

For the WT case a drop of the mean energy by 80 kcal·mol^{-1} comparing the fully folded ensemble vs. the unfolded ensemble was obtained (Table 3.2 and A.1 and figure A.13). It translates to ∼16 kcal·mol^{-1} mean energy change upon adding one folding propagation unit to the triple helix. Note, that conformational entropy effects (that favour the unfolded state) are not considered. The substitution Gly→Ala and Gly→Thr resulted in a significant drop of the folding energy and a further strong reduction was observed for the case substitutions in all three strands (Table 3.2).

Simulation Set	**Wild-Type**	**Weak restraints**	**G7aA**	**G7aT**	**G7abcA**	**G7abcT**
Mean energy change (kcal·mol^{-1})	-83 ± 11	-61 ± 24	-64 ± 19	-72 ± 18	-50 ± 14	-30 ± 16

Table 3.2: **Mean energy difference between unfolded and folded ensembles of collagen triple helices**
Mean energies were obtained using the Molecular Mechanics Generalized Born Surface Area (MMGBSA) method.

3.4 Conclusion and Perspective

We successfully simulated the molecular folding propagation process of collagen-like peptides during multiple simulations. Although the initiating nucleus formation of the three chains of the helix was enforced by artificial restraints, our simulations can depict the process starting from a partially folded triple helix. The mechanism can be described as a zipper-like stepwise assembly of groups of multiples of three amino acids, caused by the repeating occurrence of the small and flexible Gly residue. Each repeating step-wise assembly is initiated by an approximate alignment of two chains in a conformation with the main chain dihedral angles already in the near-native regimes. This arrangement forms the template for the third chain to bind to the template for completing the propagation step of three residues in each chain. While the completed extension of all three chains combined to the triple helix forms a stable structure (a fraying of a properly folded structure was not observed), the template arrangement is unstable and can also unfold before the third chain attaches to complete the propagation step. Hence, a possible mechanism with two chains forming initially and transiently a dimeric template much longer than one repeating unit followed by folding of the third strand [45] is not supported by the simulations.
Simulations with very weak restraints on the C-terminal triple helix nucleus indicate that such an unstable nucleus can result in the accumulation of misfolded structures. A looping out of segments that are multiples of the repeating unit was observed. Further propagation beyond such structures cannot be resolved and could cause formation of deformed and less stable collagen fibrils. However,

even in case of weak restraints to form the nucleus for the successful folding cases a similar propagation kinetic was observed as for cases with a stabilized nucleus.
In future studies it might also be possible to follow and identify folding nuclei directly from unrestrained simulations [65]. If propagation has started, the conditions on how the nucleus has been formed are not relevant. It emphasizes also the role of chaperone proteins such as HSP47 that bind at regular intervals of an already formed collagen helix and may further stabilize the structure to allow for correct propagation.
In contrast, the simulations on chains that contain substitutions of the central Gly residue (but starting with a stable nucleus) clearly longer folding times but no increase in the type and number off misfolded conformations was found. Also in this case stabilization of already formed triple helix segments seems to be important to rapidly propagate the process. The observed longer folding times agree qualitatively with experimental results on folding of collagens with Gly→Ala substitutions [64].

Chapter 4

Energetic Analysis of Transthyretin Mutations

In the previous chapter mainly the folding process of a protein (namely collagen) was examined using Molecular Dynamics (MD) simulations. Furthermore possible effects of mutations within the protein's sequence on the folding process were demonstrated clearly. Due to proteins' abundance and complexity it is inevitable that mutations occur, but not all mutations prohibit the formation of a protein. Some just slightly alter their structure and thus function. Some even improve their stability or efficiency, which basically makes evolution possible. On the other hand proteins are complex machines, so a mutation which does not prevent its formation probably has a negative effect on its function. This ranges from slightly reduced efficiency to complete dysfunctionality or to even harmful behaviour.
The content of this chapter has been previously published in a similar form [66].[1]

4.1 Role of Transthyretin

Transthyretin (TTR) is a transport protein for thyroid hormone thyroxine (T4) and retinol binding protein found [67]. It is produced mainly in the liver, but also ($< 5\%$) in the choroid plexus of the brain and the retinal pigment epithelium [68].
TTR forms a symmetric tetramer, consisting of four identical monomers. Each monomer is characterized by a large predominance of β-strands. The tetramer is formed by the association of two dimers that are in equilibrium with the monomeric proteins. [69]–[72] Andrade [73] and Falls et al. [74] first described the occurrence of amyloidosis of TTR in 1952 and 1955 respectively, which means the pathological formation of long and stable fibrils consisting of misfolded parallel aligned TTR molecules[75]. Amyloidosis occurs also among other proteins causing severe diseases like Alzheimer's disease or Parkinson's disease.[76] Today

[1]

three different types of TTR amyloidosis (ATTR) are known, which all result in severe diseases [68], [77]–[79]. The wild-type (WT) form of ATTR is also called Senile Systemic Amyloidosis (SSA) since it normally affects elderly people. The fibrils deposit mainly in the heart causing stiffness and thickening of the muscle. Autopsies of supercentenarians exposed SSA being responsible for 70% of their deaths [74]. The other two disease forms, Familial Amyloid Cardiomyopathy (FAC) and Familial Amyloid Polyneuropathy (FAP), are hereditary since they are caused by gene mutations of the TTR gene resulting in disease causing point mutations of the protein [80]–[82]. Similar to SSA the heart and kidneys are affected, but in contrast to SSA these diseases manifest much earlier within the affected person's span of life.
In order to form amyloid aggregates, the TTR tetramer has to dissociate to the monomeric form that unfolds and can form aggregates (see figure 4.1). Hence, both mutations that increase the intrinsic propensity for β-aggregation or stabilize the amyloid plaque structure but also mutations that destabilize the TTR tetramer or folded monomer may cause increased TTR amyloidosis. Indeed, more than 100 mutations are known that affect TTR amyloidosis [81], [83]. However, there are also known mutations which prevent amyloid production by enhancing the stability of the tetramer, monomer or reduce the β-aggregation tendency. An improved understanding of the mechanism how mutations affect TTR amyloidosis could be helpful for further development of therapeutic drug treatments.[84] Since recently the only therapy was the transplantation of the heart and / or the liver of affected patients. Today there are several promising approaches of medical treatments. Some aim on stopping protein production of harmful TTR-variants by blocking Ribonucleic Acid (RNA)-translation with short interfering RNA sequences binding specifically to corresponding Messenger Ribonucleic Acid (mRNA) in TTR-producing cells. [85], [86] Other drug molecules bind to TTR molecule itself and stabilize its natural tetrameric structure. A promising binding site for these drugs seems to be the binding pocket for thyroxine (e.g. Tafamidis [87], [88] or Diflunisal [89], [90]), which reduces the ability to transport thyroxine, one of TTR's purposes. A third way is to remove the harmful fibrils, which normally are not degraded by the body / cell itself. To enable this procedure, monoclonal antibodies are designed, which bind to misfolded monomers, fibrils or pre-fibrillar TTR and induce phagocytic clearance by macrophages. [91]–[94]

4.2 Analysis of TTR Mutations

Experimental crystal structures of a number of TTR variants in the folded tetrameric form have been determined [70]–[72], [95]–[97]. These structural studies indicate small variations in loop regions or local segments around the mutation site but no major conformational changes that may directly explain a reduced tendency of unfolding or tetramer formation. However, for some TTR mutations a reduced tetramer association compared to WT has been found using

biophysical techniques [98]. Combined experimental and molecular simulation approaches on the globular TTR indicate that indeed the tetramer stability and the unfolding tendency of the TTR monomer are of key importance for the tendency of a TTR mutant to form amyloids [71], [83], [99]. However, these studies so far did not include the TTR amyloid structure. Recently, a cryogenic electron microscopy (cryo-EM) structure of the TTR amyloid has been determined [100]. Together with the crystal structures of the globular tetrameric form this allows one to study the effect of mutations on the stability of the globular form, the unfolded form and the amyloid form using free energy simulations. In principle, among the most accurate methods are alchemical free energy simulations to transform amino acid side chains in the different TTR forms and to record associated free energy changes. However, such techniques are computationally quite demanding and often allow to study only few mutations [83]. The aim of this study is to employ a less demanding end-point free energy method to investigate systematically a large set of TTR mutations. The method is based on running short MD simulations of the wild-type (WT) and the mutated proteins in the different conformational states and evaluate the generated trajectories using a continuum solvent model (Molecular Mechanics Generalized Born Surface Area (MMGBSA) method) [22], [59], [101]. To avoid large conformational changes during the simulations, weak positional restraints on the structures were included, assuming that the mutations do not alter the structure significantly compared to WT. Indeed, crystal structures of TTR mutations indicate only small conformational changes in the globular form compared to WT [71], [95], [96]. The approach was applied to 36 TTR mutations for which either an increased or decreased tendency for amyloidosis has been reported experimentally. Since experimentally only a tendency for amyloidosis is known it allowed us only a qualitative comparison with calculated stability changes. For 30 out of 36 mutations the performed calculations agreed qualitatively with the experimental tendency and could be used to identify the origins of this tendency. The simulations indicate that mutations can both stabilize or destabilize the globular tetrameric form but also the amyloid structure to various degrees. Overall, the tendency of mutations to promote increased or decreased amyloidosis correlates strongly with the destabilization of the globular or dimeric/tetrameric forms but much less or not with the calculated destabilization/stabilization of the amyloid structure. Similar results were also obtained using alternative methods to evaluate protein stability such as FoldX [102]. The influence of neighbouring residues and structural and energetic origins of the tendencies are discussed. The rapid methodology can also be used to systematically analyse mutation effects in other amyloid forming systems assuming that the structural changes are small.

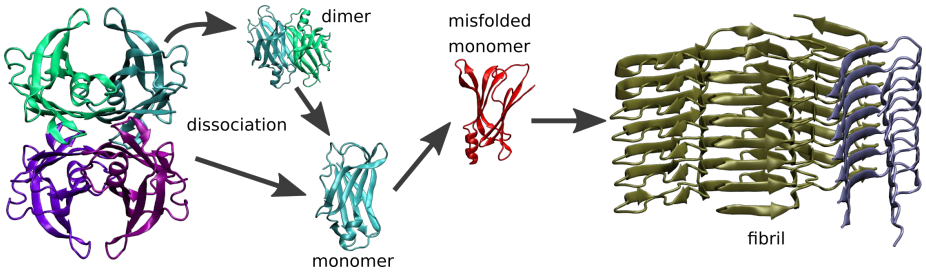


Figure 4.1: **Schematic pathway to TTR amyloid fibril formation.** For amyloidosis the native TTR tetramer structure (cartoon representation with different colors for each monomer) dissociates into dimers and finally monomers which then unfold (or at least partially unfold). The unfolded monomers align to each other forming a fibril structure where each layer arises from one monomer. Note, the structures of some residues are not resolved in the cryo-EM fibril structure, leading to two fragments per layer (visualized by different colors).

4.3 Setup and Parameters

As starting structure for the globular TTR protein the protein data base entry Protein Data Bank (also file format .pdb) (PDB) 6e6z[69] was used. It represents the tetrameric structure. The complexed Tafamidis molecules were deleted from the structure file. The specific structure was selected by best possible overlap in sequence with the given fibril structure. A comparison with other PDB-structures of the tetramer revealed a conformity about 0.7 Å - 1.2 Å of backbone root-mean-square deviation (RMSD). The mutations were created by replacing the corresponding side chains and adjustment of side chain structures by selecting the sterically best fitting rotamer and energy minimization. The structure of the amyloid fibril form corresponds to the entry PDB 6sdz. [100] The mutations of the fibril were created in the same way as for the globular protein. A fibril was represented by 7 protein chains to minimize the effect of the protein water boundary at both ends of the oligomeric fibril.
Employing the Amber18 software [103] package solvated starting structures were generated, using the ff14SB force field [56] for proteins and the Optimal Point Charge (OPC) water model [14]. (see 2.4.1) The globular protein structures were embedded in octahedral water boxes with a minimum distance of 10.0 Å between the protein and the box boundary. For the fibrils cubic boxes were found to be optimal with respect to overall system size. Sodium and chloride ions were added to neutralize the charge of the systems and to obtain a ∼100mM salt concentration. After a minimization run of 500 steps (250 steepest descent algorithm and 250 conjugate gradient algorithm) the systems were heated up linearly to 300 K within 200 ps followed by 50 ps equilibration with constant volume periodic boundary conditions and restraints on all heavy atoms (2.0 $kcal \cdot mol^{-1} \cdot Å^{-2}$). Subsequently the systems were relaxed for 100 ps at constant pressure (1.0 bar) and the positional restraints were reduced in four steps of

0.5 kcal·mol^{-1}·Å^{-2}. Each time the system was simulated for 50 ps with the final state of the previous simulation taken as reference for the positional restraints, allowing the mutated structures to adjust. This leads to conformationally relaxed backbone structures to accommodate the mutated side chain. At a final restraint level of 0.1 kcal·mol^{-1}·Å^{-2} the system relaxed for 3 ns before the data gathering run of 2 ns was performed. For all simulations a time step of 2 fs and a 10.0 Å real space cut-off of was chosen (the particle mesh Ewald (PME) method was used to account for long range electrostatic interactions). All bonds involving hydrogens were constrained by SHAKE [58]. The SETTLE algorithm was used to constraint bond length in water molecules [104]. Data gathering simulations were performed at constant pressure of 1 bar and temperature of 300 K using a Langevin thermostat.

The end-point free energy calculations were performed using the MMGBSA tools of the Amber18 package. [59] The explicit water molecules and ions were removed from the trajectories and the sampled conformations were re-evaluated using a Generalized Born (GB) continuum model using the mbondi3 radii and parameters from Nguyen et al. [105] (igb=8 in Amber) in combination with a surface area dependent tension model to account for non-polar solvation (surface tension coefficient γ = 0.005 kcal·mol^{-1}·Å^{-2}) was used. Between 100-10000 trajectory frames were used to obtain mean energy contributions. No conformational entropy changes were considered, hence, it was assumed that the change in mobility due to mutation is similar in the globular vs fibril form. For representing the energy change associated with a mutation in an unfolded protein the central residue of a tripeptide in an extended conformation was considered (includes at most nearest neighbour effects in the unfolded chain upon residue mutations).

For comparison all mutations were also analysed using the FoldX modelling program suite [102], which uses an empirical force field. Within the BuildModel tool FoldX employs an internal structural optimization upon mutagenesis and evaluates single conformations of structures using a knowledge-based combination of energy terms. The effect of a mutation is obtained as score relative to the WT sequence. It was applied to all mutations in globular and fibril form.

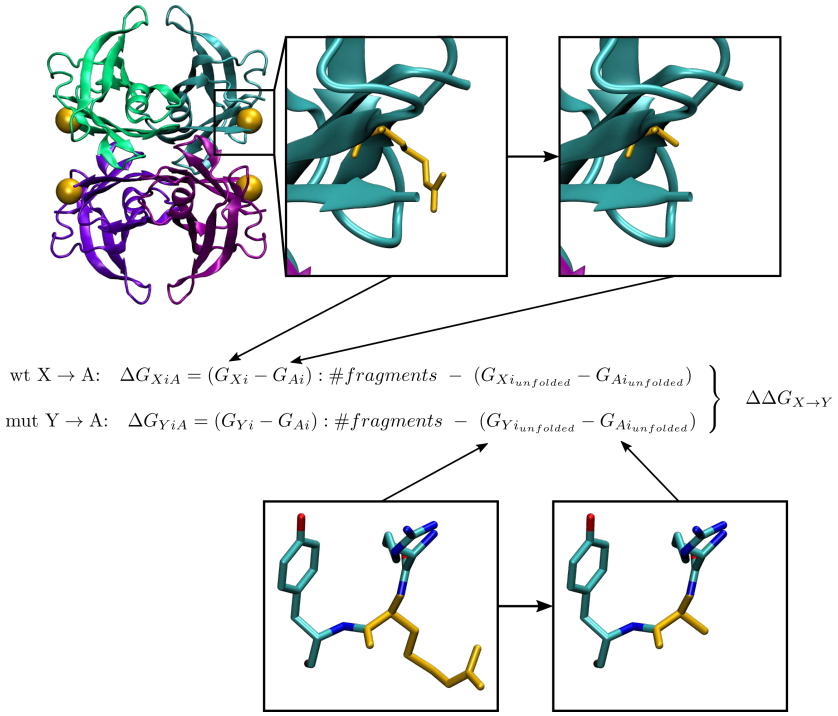


Figure 4.2: **Schematic illustration of the MMGBSA calculations of single residue substitutions** (marked yellow, in the tetramer the position is indicated as yellow sphere). To obtain mean energy differences of single point mutations, first the energy contribution of a specific residue is calculated by cutting all atoms after the C_β-atom of this residue (Ala-scan) and subtracting the mean energy of the unfolded sequence by calculating the energy of the specific residue and its next neighbours. This is done for the WT and the mutated protein resulting in the free energy contribution difference of the single residue / side chain.

4.4 Results of the New Method

The method presented here was applied to 36 mutations of TTR and examined in several aspects. The first question was how the stability of the whole molecule is affected by the mutations. This was done for both the globular and fibrillar form and then both were compared. Furthermore the stability within the globular tetramer was analysed. Additionally the method was compared with another computational method (FoldX).

4.4.1 Influence of Mutations on Fibril and Tetramer Stability

The formation of TTR amyloid structures requires the dissociation of the TTR tetramer, unfolding of the monomer and formation of the amyloid arrangement (figure 4.1). Hence, a mutation can influence the free energy change associated with each step. The application of the MMGBSA method [22] is a relatively fast computational approach that allows one to investigate the contribution of each conformational transition on a large set of TTR mutations (for comparison the even faster FoldX method [102] was also used). Calculations were performed for the mutations in the fibril structure and the monomer, dimer and tetramer structures (illustrated in figure 4.2). The computer time (on a single core) for running the simulations of the globular and amyloid forms of each TTR mutation and MMGBSA evaluation takes $\sim$ 2-3h computer time (using 500 trajectory frames for MMGBSA analysis, however, for the cases investigated here we analysed in each case 10000 frames). For the present study 36 TTR variants were investigated for which experimental data on the amyloid forming tendency is available (location of the mutations is presented in figure 4.3).

In order to directly compare the calculations to the amyloid forming tendency we consider first the difference of the mean energy contribution of the selected side chain mutation for forming the tetramer structure vs. forming the fibril structure (4.4). The energies are calculated with respect to the unfolded solvated side chain (represented as central residue in a solvated tripeptide).

Each mutant was initially created based on the same tetrameric template structure (PDB 6e6z[69], see Methods section) or the same fibril structure (PDB 6sdz). Hence, we assume that both the globular as well as the fibril structure are similar for all mutations and do not cause major conformational changes or formation of a new fibril structure. Also, experimental (quantitative) data on how a mutation changes the amyloid fibril stability is not available, only an amyloidosis tendency for forming fibrils of each TTR variant can be obtained experimentally. Hence, one can distinguish mutations that increase (red background in figure 4.4) or decrease (yellow background in figure 4.4) the tendency for amyloid fibril formation. For most, that is 28 of the 36 cases, our calculations correctly reproduce the experimentally obtained fibril formation tendency (figure 4.4).

Interestingly, for most known amyloidogenic mutations V30G[106], D38A[78], E54G[81], [107], E54K[82], [107], L55P[71], [108]–[111], L58H[71], [112], T60A[70], [71], [82], [113], E61K[114], S77Y[70], [82], [115], Y78F[116], I84A[117], I84S[82], [97], [117], [118], H88S[119], E89K[82], [120], Y114C[121]–[125] and Y114H[71], [96], [99]) our results indicate a significant overall destabilizing effect on the tetramer. Note, that at this analysis stage influences on monomer stability, dimer formation and energetics of dimer association to tetramers are all included (decomposition see next paragraph). Interestingly, the calculations predict that the fibril structure is sometimes even slightly or considerably destabilized or less stabilized for most of these variants compared to the WT.

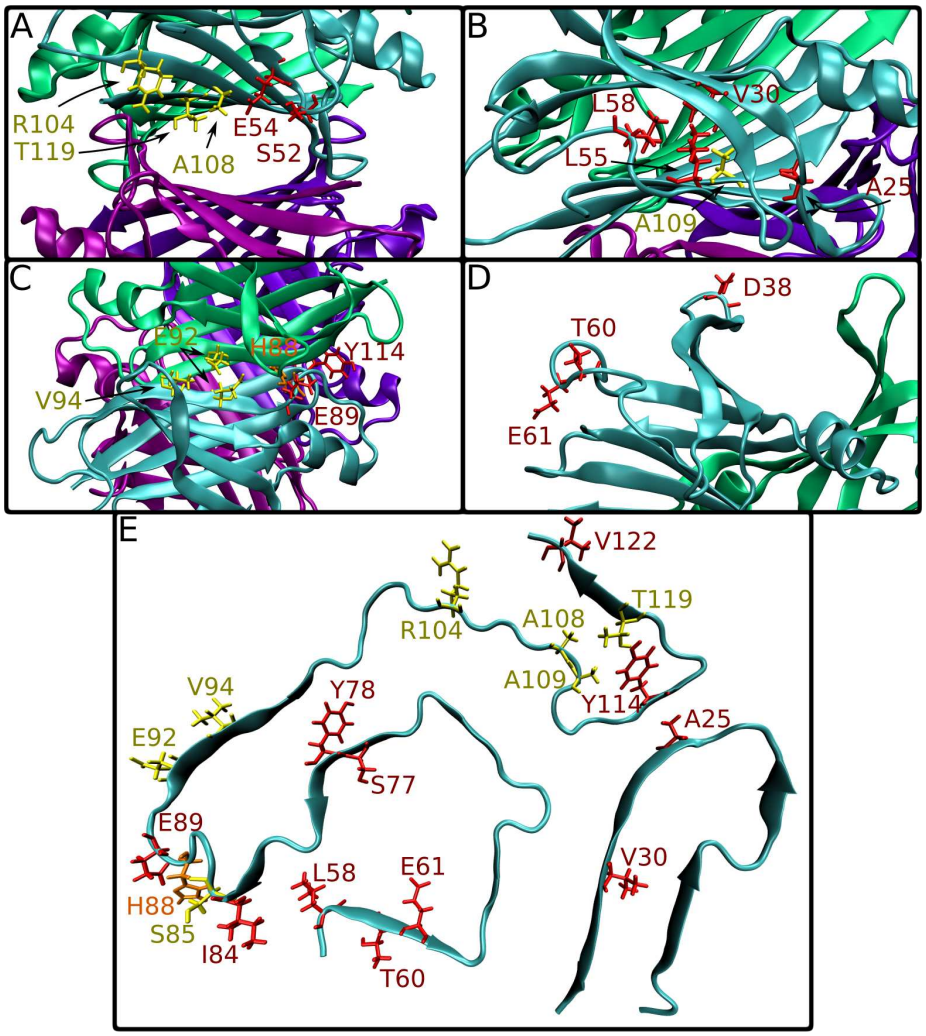


Figure 4.3: **Location of mutations within the globular tetramer** (A-D, PDB 6e6z) **and fibril** (E, PDB 6sdz). Relevant side chains (of the WT) are shown as sticks with red coloured residues represent mutation positions that were found to increase amyloidosis tendency (yellow: opposite behaviour). Within the tetramer mutation positions can be oriented towards the cavity between the two dimers (shown in A), the buried region between the monomer's β-sheets (B), the interface between the monomers within each dimer (C, both at the interface and in the rim region around the interface) or are solvent exposed in the globular tetramer (D, also exposed in the dimer and monomer). Mutation positions are also mapped on the known fibril structure (E, PDB 6sdz, only one layer is shown for clarity).

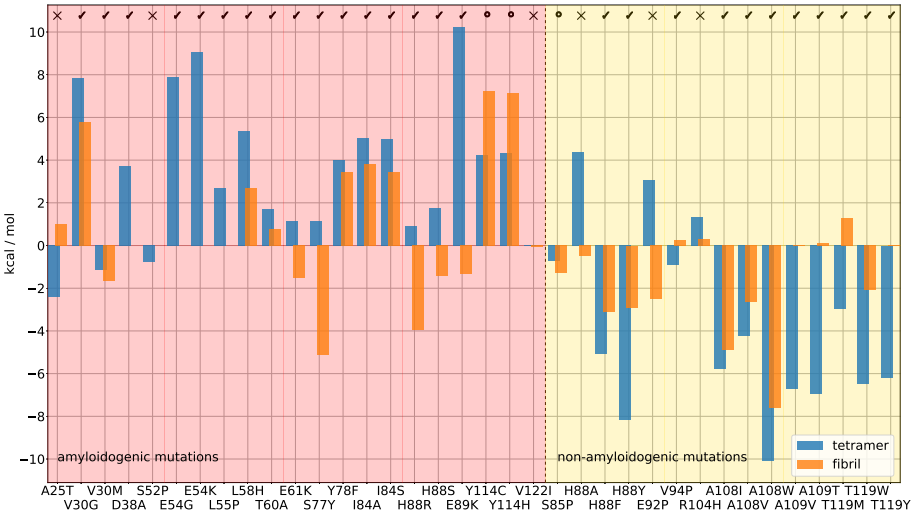


Figure 4.4: **Energy contribution (boxes) of single point mutations of TTR in tetrameric (blue) and fibril (orange) form.** The mutations are indicated on the x-axis. A negative/positive contribution implies a stabilization/destabilization of the structure. All energies are per monomer/layer, however all fragments / layers were included in the calculations and the result was divided by theirs number (4 or 7 for the tetramer or the fibril respectively). Since the sequence of the fibril includes a gap compared to the tetramer, there are no fibril energy values for some variants. A red background indicates that this mutation is known for its increased amyloidogenicity whereas a yellow background means that the mutant inhibits its amyloidosis or stabilizes the tetrameric form. The symbols at the top depict if the result matches to experimental data (check mark) or partially matches (circle) or does not match (cross). Note the variants D38A, S52P, E54G, E54K and L55P show no energy value for the fibril since the site of the mutation was not part of the fibril's sequence.

Among the known non-amyloidogenic mutations the calculations yield for most cases including S85P[69], H88F[119], H88Y[119], V94P[69], A108I[126], A108V[126], A108W[69], A109V[126], A109T[127], [128], T119M[70], [71], [109], T119W[69] and T119Y[69] an energetically stabilizing effect on the tetramer compared to WT. Surprisingly, several of the non-amyloidogenic mutations are predicted to stabilize also the amyloid form albeit to a lesser degree than the stabilization of the tetramer (figure 4.4, yellow part). In several of the non-amyloidogenic cases, however, no or only very small effects on the fibril stability relative to WT were found. Overall, the results indicate a rather strong correlation between calculated stabilizing/destabilizing effect of a mutation on tetramer formation and experimentally observed tendency for fibril formation. In contrast, no good correlation was found between known amyloid forming tendency of the mutations and relative energetic stabilization of the fibril structure (only for the V30M, the E61K and the S77Y a significant stabilization of the fibril form is predicted that outscores the effect on the globular form).
In addition to MMGBSA calculations we also employed the FoldX approach [102] to predict the effect of mutations on the stability of globular vs. fibril conformation of TTR. The BuildModel option in FoldX allows conformational optimization upon mutation and energetic evaluation based on a knowledge-based optimal combination of energy terms (on single conformations). Interestingly, very similar to the MMGBSA analysis it also predicts for almost all amyloidogenic mutation cases a destabilization of the globular form and for most non-amyloidogenic cases a stabilization of the globular tetrameric form (figure B.1). FoldX predicts for most mutations a small destabilizing effect on the TTR fibril structure and for several of the non-amyloidogenic mutations an extremely strong fibril destabilization (10-20 kcal·mol^{-1} per mutated residue). These predictions possibly overestimate the destabilization since FoldX has been designed and trained for the application on globular protein structures. Nevertheless, the predictions by FoldX show qualitatively the same trend and confirm the results obtained with the MMGBSA approach.
Since the calculated stability changes of mutations in the fibril form showed only little or no correlation with the experimentally observed TTR amyloidosis tendency we calculated the mean (absolute) MMGBSA energy per residue for the globular (tetramer) form and for the fibril form. It turned out that the MMGBSA energy is $\sim$2 kcal·mol^{-1} more negative per residue than the globular form. This result varies for the mutations but is overall similar to WT (figure B.2). It indicates that according to the MMGBSA calculations the fibril structure is energetically much more stable than the globular form. This in turn implies that the small change due to mutation of a single residue may not change the large favouritism of the fibril form. Hence, once the globular form is unfolded at sufficiently high concentration the fibril form is in all cases energetically strongly favoured. However, the energetic favouritism of the fibril form depends on the number of peptide layers included in the MMGBSA evaluation (figure 4.5). Considering just one layer in the fibril conformation results in a mean energy per residue in favour of the globular form. In this case the fibril backbone of the protein is fully solvent exposed without forming hydrogen bonds

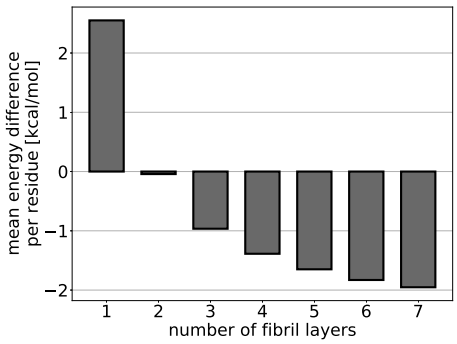


Figure 4.5: **Calculated mean MMGBSA energy difference between tetramer and fibril.** The mean energy per residue of fibrils of different lengths (1-7 layers) was compared to the globular tetramer TTR. While just one layer is energetically unfavoured, two layers are already energetically equally favourable and three or more layers are more stable than the tetramer.

to a neighbouring layer. However, already 3 layers forming a fibril results in a calculated mean energy per residue in favour of the fibril eventually reaching a level of $\sim$ -2 kcal·mol^{-1} per residue (4.5 and B.6).

For some mutations the MMGBSA calculations do not agree with the experimental trend of TTR amyloidosis. For example, for the variant A25T [109] the calculations predict a decreased stability of the fibril and increased stability of the tetramer although experimentally this mutation causes increased TTR amyloidosis. Visual inspection shows that the A25 is located at a narrow buried interface between two β-strand segments in the fibril structure (figure 4.3). Indeed, replacement by a larger polar threonine can perturb this interface. It is possible that such variant forms an altered fibril topology with an increased space in the fibril not considered in the present study. It is also possible that some mutations promote fibril formation by stabilizing intermediates during amyloid formation. Such mutation can change the kinetics of fibril formation without stabilizing the final fibril structure. The possibility of an altered fibril structure could also explain the results for mutations Y114C and Y114H. The mutations to smaller residues destabilize the globular form but more strongly also the fibril structure. An altered fibril topology or adjustments of the fibril structure not considered in the present study may reduce the predicted destabilization effect on the fibril. Also the H88R variant [119], known to promote amyloidosis, is predicted to strongly stabilize the globular tetramer and also but to a lesser degree the fibril structure. In this case the protonation states of the WT histidine (we assume the standard neutral protonation state) may influence the stability. Experimental studies have shown, that at neutral pH His88 is neutrally protonated, but at lower pH values changes to double protonation state, which destabilizes the WT structure. [129] The results of our calculations revealed a slight destabilization of the tetramer, which was sensitive to the exact

positioning of the histidine side chain.
Indeed, the H88A [119], E92P [118] (artificial) and R104H [71] variants are known to reduce amyloidosis but the calculations predict a destabilization of the tetramer possibly due to changes in protonation states not considered in the calculations. For example, for the V30M variant [71], [118], [130] the calculations indicate a slight stabilization of the tetramer and a stronger stabilization of the fibril. In this case a rather small calculated effect is expected because both methionine and valine are hydrophobic side chains of not very different size. A similar explanation may hold for mutation V122I [100], [108] for which the calculations predicted no change in amyloidogenic tendency. Indeed, the variant V30G corresponding to the loss of a whole non-polar side chain shows a strong effect that agreed qualitatively with experiment. The result of the amyloidogenic variant S52P [131], [132] also does not fit to experimental data. This could be due to changes in the backbone structure that may affect the fibril topology due to replacement by proline. However, in this case other known in vivo influences (not considered by our calculations) like an enhanced proteolytic cleavage between Lys48-Thr49, favoured by this mutation, may drive the formation of fibrils. [133]
The MMGBSA analysis allows us also to separate the effect of a mutation into energetic contributions to both for the globular tetramer and the fibril structure (figures B.3, B.4, B.5 and B.6). For the bonded energy contributions destabilizing and stabilizing effects are observed with no clear correlation to the amyloidogenic effect of the mutations (figure B.3). Also, no clear distinction between amyloidogenic and non-amyloidogenic variants is observed for electrostatic and van der Waals contributions except that both contributions show a significant anti-correlation (figure B.4, B.5). The surface area dependent non-polar solvation term (figure B.6) correlates strongly with the van der Waals contribution. This is expected since reduction of surface area often also leads to stronger packing energies. Interestingly, if a mutation leads to decreased or increased van der Waals interaction this is typically seen then for both the globular and the fibril TTR form.

4.4.2 Influence of Mutations on tetra-, di- and monomer formation

Depending on the position of a mutation, its impact may affect the folding stability of a single monomer, formation of a dimer or formation of the tetrameric complex formed by two dimers (figure 4.1 and 4.6). As reference state for the calculations the transition in the unfolded structure (represented as central residue in a tripeptide) was used. The calculated relative stability changes for the tetramer are the same as given in figure 4.4 and the stability change is given per monomer unit. Thus, in the plots of figure 4.6, identical values for monomer, dimer and tetramer indicate that the mutation changes only the folding of the monomer but has no further influence on the dimer and tetramer association. This concerns most of the investigated mutations (see figure 4.4). However, some mutations, especially of residue 88 and 89 and 114 show a modest

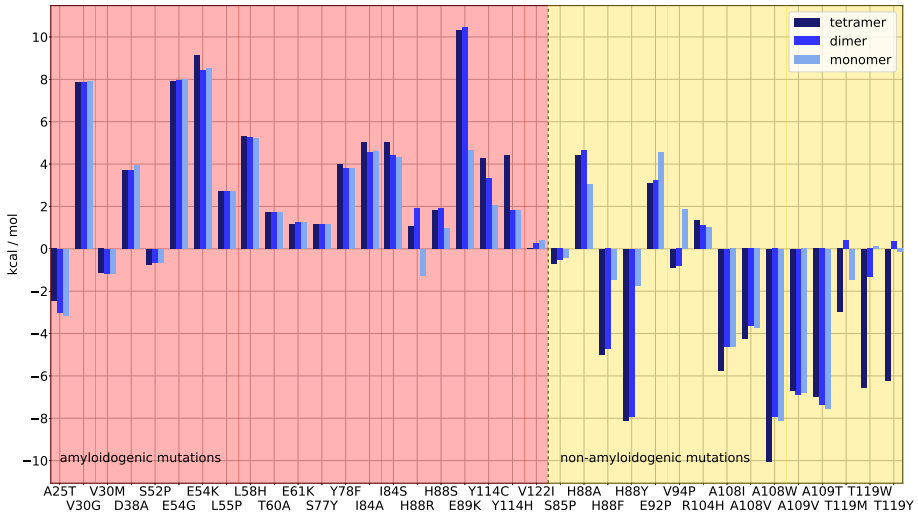


Figure 4.6: **Comparison of calculated energy changes for tetramers, dimers and monomers upon mutation** (relative to WT). The mutations are indicated on the x-axis. A negative/positive contribution implies a stabilization/destabilization of the structure. All energies are per monomer/layer.

change in monomer stability but a significant change in relative free energy of dimer formation (no or only slight further stability change upon tetramer formation). Among the mutations that promote amyloid formation there are five cases (E54K, I84A, I84S, Y114C, Y114H) for which a destabilization of the tetramer is observed and vice versa there are also six cases among the mutations that reduce the TTR amyloidosis amyloidosis (A108I, A108V, A108W, T119M, T119W, T119Y) with a predicted increase in tetramer stability. Indeed, residues 108, 119 are located at the dimer-dimer interface (figure 4.3) and large residues can fill empty space at the interface. In some of the latter cases the stabilization of the globular forms is also due to increases folding stability. Overall, the calculations indicate that stabilization or destabilization of each step up to the tetrameric globular form can influence the TTR amyloidosis tendency. The effects on dimer and tetramer formation can be correlated to the location of the residues at or close to an interface between the monomers in the dimer H88, Y114, E92, V94) or the interface between dimers in the tetramer (A108, T119, see figure 4.3).

4.4.3 Optimizing the efficiency of the calculations

In the above MMGBSA calculations we analysed for each mutation case 10000 frames of each simulation. This resulted in calculated errors of the mean well below 1.0 kcal·mol^{-1} for the calculations of the whole tetramer as well as the fibril. However, the MMGBSA calculations on these many frames exceeds (con-

siderably) the simulation time for generating the trajectory frames. Predictions based on FoldX that took ∼24 h for all 36 mutations are faster mainly because no MD simulation to generate an ensemble of conformations is required and gave qualitatively similar results (except for some mutations in the fibril structure that resulted in very high penalties).
An ensemble of structures allows one to estimate an error of the calculated energies. In order to speed up the overall calculations and balance each part we systematically reduced the number of frames used for the MMGBSA approach down to only 10 (distributed equally over the whole simulation). Interestingly, despite an increase in calculated error of the mean (>10 kcal·mol^{-1} in some cases for the evaluation of only 10 frames) the mean calculated energy values changed only very little (figure B.7). The error per selected residue (7 copies in case of 7 fibril layers) is also lower than for the whole molecule and allows one to speed up the MMGBSA calculation. Hence, it indicates that for a rapid estimation of mutation effects the evaluation of just a few hundred frames might be sufficient and the generation and evaluation (including simulation and MMGBSA calculation) is then a matter of minutes for each mutation. Note also that the evaluation of each mutation can be performed independently in parallel which further speeds up the evaluation of mutation effects.

4.5 Evaluation of the Results

A significant number of human proteins can undergo amyloidosis resulting in unfolding and formation of amyloidogenic fibrils that cause various degenerative diseases [68], [81], [127]. Recently, the structures of both the globular forms and the amyloid fibril forms of several of these proteins have been determined. Especially, the rapidly growing number of fibril structures also formed under different in vivo or ex vivo conditions becomes available due to progress in the structure determination by cryo-EM techniques. However, there is still very little understanding why certain protein sequences possess a strong tendency for fibril formation, why certain mutations increase or decrease the tendency of fibril formation and which structural form (folded, unfolded or fibril) has the largest influence[134]. Simulation studies can be helpful to obtain insight into the molecular details and also energetics of fibril formation and the influence of mutations. However, available methods to quantify relative stability changes are often time consuming especially in case of systematic applications. A second goal of the present study was to evaluate the possibility of using a Molecular Mechanics Generalized Born Surface Area (MMGBSA) method – or even simpler FoldX – approach to rapidly estimate mean energy changes due to mutations applied to the TTR system for which experimental data on many mutations is available. For almost 80% of the 36 tested mutations the calculations predicted a tendency in correct agreement with experiment. Only a qualitative comparison is possible because experimentally only the qualitative increase or decrease of TTR amyloidosis tendency is available.

The calculations allowed us also to extract some important conclusions concerning the influence of the mutations on the TTR amyloidosis. The systematic application to 36 TTR mutations indicate that it is indeed the destabilization of the globular forms that strongly correlates with TTR amyloidosis tendency in agreement with previous studies based on studying limited sets of mutations [71], [83]. Similar results were obtained for application of FoldX. The effect, however, cannot be attributed to one of the possible sub-equilibria such as monomer folding, dimer formation or tetramer formation but can be caused by influencing either one or several of these steps. The analysis of energetic contributions of each mutation also did not identify a single energy term responsible for modulating the tendency of fibril formation of a given mutation.
Interestingly, mutations that are predicted to increase or decrease the fibril stability relative to WT are approximately equally distributed among those that either show enhanced or reduced amyloidogenic tendencies. It indicates that according to the calculations once the TTR is monomeric and (partially) unfolds a reduced stability of the fibril (relative to WT) has only little influence of TTR amyloidosis (the residual stability of the fibril is still sufficient to drive fibril formation). Calculations on the mean energy per residue in the globular vs. fibril form indicate indeed a significant energetic stabilization of the amyloid fibril form compared to the globular structure offering a direct explanation for the above conclusion. Interestingly, the energetic favouritism of the fibril is predicted to depend significantly on the number of layers included in the calculations. Formation of an initial single layer is energetically strongly unfavoured compared to the globular form but already a fibril composed of just 3 layers has a mean free energy (MMGBSA) per residue that favours the fibril structure.
Finally, we demonstrated that the approach is rapid enough for the systematic application on large numbers of mutations for a given globular and fibril protein structure. It was found that approaches such as FoldX are also useful to evaluate the tendency but since the method is based on an empirically optimized weighting of different energy terms (for globular proteins) it might be useful to extend this also to studies on amyloid fibril structures. It has been shown recently that some peptide or protein sequences can adopt several different fibril topologies depending on sequence and experimental condition for amyloid fibril formation[135]. In cases where such alternative fibril topologies are available, it is also possible to evaluate the preference of a mutation to promote formation of different fibrils.

Chapter 5

Dynamics of RNA Bulge Loops

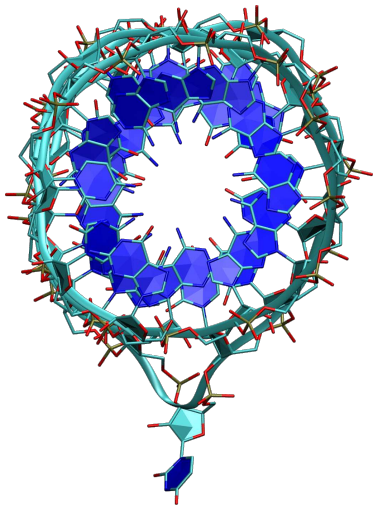


Figure 5.1: **Top view of double-stranded Ribonucleic Acid (dsRNA) with bulge base looped out.** A bulge is the easiest form of a loop in one strand of a dsRNA. Basically it is a single additional base in one strand. (For a schematic visualisation see fig 5.6.)

While the previous chapters examined mutations in proteins, this chapter is neither about proteins nor about mutations in terms of replacing something. Instead we will analyse a certain structural aspect of a type of molecule which is actually essential for proteins to be built in our cells, the Ribonucleic Acid (RNA). Beside proteins, lipids and carbohydrates, nucleotides form the fourth group of most important biomolecules. The group of nucleotides consists of Deoxyribonucleic

Acid (DNA) and RNA, which are relatively similar in structure, but have different tasks. While DNA's purpose is mainly to conserve genetic information, RNA's features are more versatile. As DNA it can conserve and transport genetic information but it also catalyses specific chemical reactions.[136]–[138] Something mainly proteins are known for.
Therefore different types of RNA exist. Messenger Ribonucleic Acid (mRNA) serves as a transporter of genetic information from one place (e.g. nucleus or RNA virus) to the ribosome, a cellular complex where proteins are built. The information of a protein's sequence is stored sequentially within mRNA, making them long thread-like molecules. Transfer Ribonucleic Acid (tRNA) molecules translate the mRNA sequence to amino acids and fuse them to proteins within the ribosomes.[139] The ribosome itself is a ribozyme (**ribo**nucleic acid en**zyme**) and consists of proteins, which provide structure and stability and Ribosomal Ribonucleic Acid (rRNA) which catalyses the peptide bond formation within the ribosome.[140]–[143]
The overlapping characteristics of RNA with DNA on the one side and proteins on the other side, motivated the theory of the RNA-world. It is assumed, that the first forms of life on earth were not protein-based but self-reproducing RNA structures.[144]–[147]

5.1 Structure of RNA

Like DNA, RNA consists of a linear sequence of nucleotides. (see figure 5.2) Each nucleotide has three subunits, ribose, phosphate and a nucleobase. The ribose, a cyclic monosaccharid with five carbon atoms, is connected by phosphodiester linkages at its third and fifth C-atom to phosphate groups, which again bind to ribose rings at the corresponding positions. This alternating chain forms the backbone of the RNA sequence. Since the C-atoms of ribose are named by numbers with a prime (1' to 5'), RNA sequences have a 3'-end and a 5'-end, according to the connecting C-atoms C3' and C5'.
At the C1' atom one out of four different nucleobases is bound to the ribose ring, generating the code of the sequence. (see figure 5.2) The four nucleobases are adenine, cytosine, guanine and uracil. The latter replaces thymine among the otherwise identical nucleobases of DNA.
Another difference to DNA is the hydroxy group at the ribose's C2' position, which makes RNA less stable than DNA due to easier self-cleavage processes. Therefore DNA is better capable to store long genetic codes. On the other hand the additional hydroxy group acts as further partner for hydrogen bonds, which enables much more versatile RNA structures due to non-canonical base pairs [148]–[150] and compact helix packing.[138]

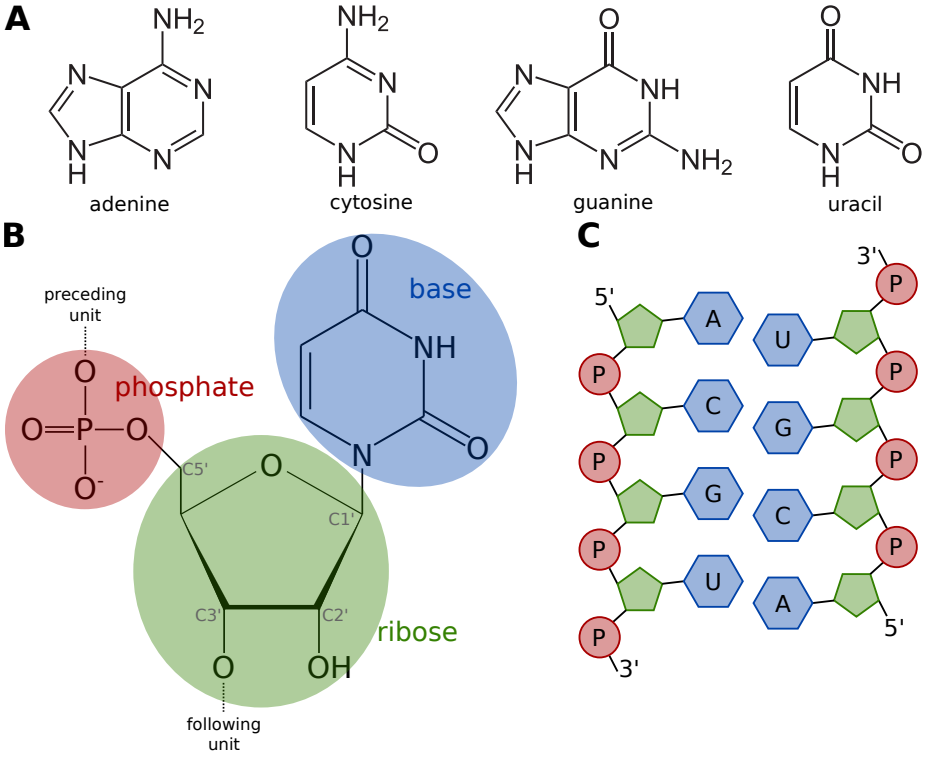


Figure 5.2: **Composition of RNA** (A) The four aromatic bases occurring in RNA sequences. (B) One repetitive unit of a sequence. The phosphate (red) and ribose sugar (green) form the backbone. The base (blue) can be one of the four listed in A. (C) Multiple units of B fuse to a single RNA strand, with 3' and 5' end, which can form a double strand with a complementary strand. Standard base pairs are adenine-uracil and cytosine-guanine.

Similar to proteins, RNA's structure is described by different levels of "scale", ranging from primary to tertiary structure. The primary structure only includes the sequence of the nucleotides lined up. This sequence is in general generated by transcription of DNA code, but also undergoes post-transcriptional modifications.[151]
The secondary structure describes the interaction of the nucleotides' bases (A, C, G and U). Here, a distinction is essentially made between bases that interact canonically via Watson-Crick-Franklin (WCF) base pairing [152] (see figure 5.3) and those that interact non-canonically. Watson-Crick-Franklin pairing in RNA only occurs between adenine and uracil (A–U) or cytosine and guanine (C–G) bases. These pairings of so called complementary base pairs, formed by hydrogen bonds, are particularly stable. [138] Each WCF pair stabilizes an RNA helix by 1-3 kcal/mol, driving single-stranded RNA to fold back and form double-stranded sections containing WCF base pairs. [138], [153]

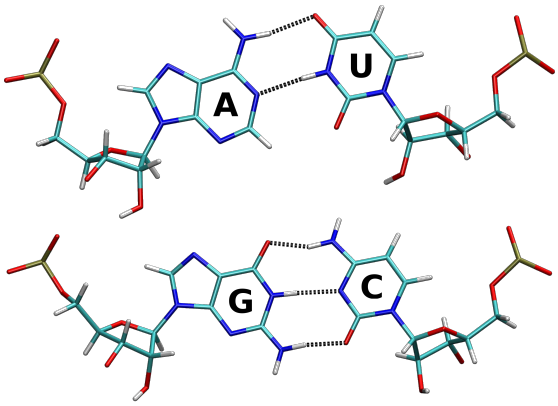


Figure 5.3: **Canonical base pairing** also called WCF pairing with Hydrogen bonds (black) between canonical base pairs adenine (A) – uracil (U) and guanine (G) – cytosine (C).

Schematically the secondary structure is shown in 2D drawings (see figure 5.4). Generally, RNA appears as single or double-stranded, where double-stranded means that only WCF paired nucleotides occur. Normally there is a mixture of both, leading to typical forms like hairpin loops, bulge loops, internal loops or junctions.
While canonically paired helical sequences contribute energetically stronger to RNA's stability, the unpaired bases (e.g. in loops) are not necessarily unbound. Rather there are several non-canonical pairings, not represented in this visualization.
However, the stabilizing effect of canonically bound base pairs is significantly stronger than the stabilizing interactions in protein secondary structure. [154], [155]

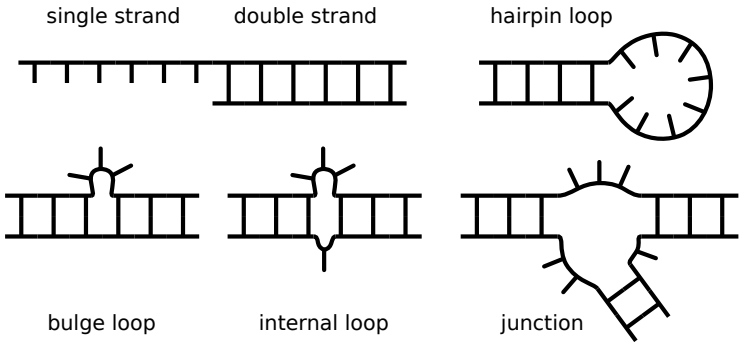


Figure 5.4: **Typical secondary 2D RNA structures**

While the secondary structure basically just determines which nucleotides are canonically (WCF) paired, the tertiary structure describes the non-canonical interactions of unbound bases, sugar rings and phosphate groups, determining the spherical (3D) structure of an RNA molecule. [138], [156], [157] These parts can make up to 35 - 50 % of the RNA structure. [157] As for WCF parings, hydrogen bonds play an important role within these interactions, often including the 2'-OH group of the sugar ring. [138]

In contrast to DNA, RNA usually exists in shorter sections of double-stranded helices connected by other conformational structures like loop regions. For instance the longest uninterrupted continuous helical sequence within *E. coli* 16S RNA is just 12 base pairs long. [157] The typical appearance of dsRNA is called A-form, with a deep major groove and a shallow minor groove. (see figure 5.5) Compared to the more abundant B-form of DNA, the A-form is more compact and the bases are tilted towards the helical axis. Also the major groove is narrower [158], which makes it less accessible for binding ligand proteins. [159]

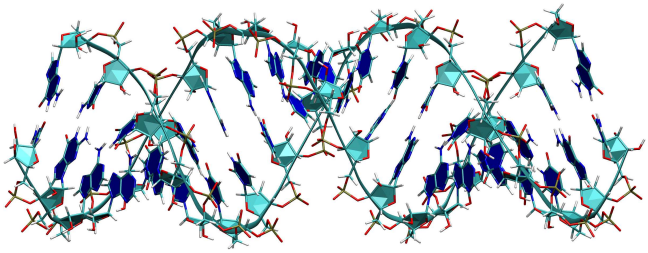


Figure 5.5: **Double-stranded RNA helix.** A-form RNA helix of 19 base pairs. In the central part of the helix the major groove is visible on the bottom whereas the minor groove is located at the top side of the image.

5.2 Simulating a Bulge

From the many occurring secondary structures of RNA we will now select a very simple one, the bulge loop (5.4). To make it even simpler, we chose a bulge loop containing only one base. So what we are actually looking at is a double-stranded helix with an extra base in one of the two strands. From the point of view of the double helix, the bulge is a disturbance in the helical structure. One strand is a little bit longer than its complementary one, which leads to slight bending of the whole structure or strong bending of the bulge strand's backbone. Regarding the title of the thesis, we interpret this bulge as a disturbance of an RNA helix. Our goal was to observe the dynamics of this bulge. Does it move along the helix, or does it stay at one position? Does it linger inside the helix or loop out to make room for the other paired bases?
If we choose the same type of nucleotide within each strand (only A - U pairs

or only C - G pairs), there seems to be no predefined position for the surplus nucleotide, since all possible WCF pairs are the same. So a single-base bulge loop within such a helix can move through the helix, if the single unpaired bulge base pairs to the opposing base of one of its neighbouring WCF pairs, replacing its direct neighbour of the same kind, which then becomes the new, now moved, bulge base. Only at the rim parts, where the helix is open, the bulge could "escape" the helical structure so to speak and become a single-base single-stranded Ribonucleic Acid (ssRNA) attached to one of the helix's strands. As simulating a very long uniform dsRNA is too computationally expensive regarding simulation time, we try to counter this possible "escape" by putting two C-G pairs at each end of an A-form dsRNA framing a series of 15 A-U pairs. Cytosine - guanine (C-G) pairs are more stable than adenine - uracil (A-U) pairs, because they include an additional hydrogen bond. (see figure 5.3).
Since this is a quite artificial structure, we built it from scratch using the AMBER [103] *NAB* tool (*fd_helix()* routine) [160], based on the structural parameters from Arnott et al [161]. This generates a Protein Data Bank (also file format .pdb) (PDB) file. Within this file one adenine (or uracil) was deleted in the middle of the helix, creating a single base U-bulge (or A-bulge) framed by 7 A-U pairs and 2 C-G pairs on each side. (see figure 5.6)

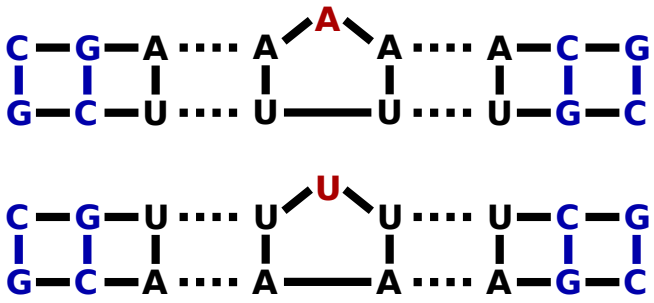


Figure 5.6: **2D structures of single A/U - base bulge.** Double stranded RNA helices with a single base bulge. Each end is stabilized by two C-G base pairs framing 14 A-U base pairs and an additional bulge base in the middle.

Now AMBER's *leap* module was used to create AMBER topology and coordinate files. Here the AMBER *OL3* force field was used to parametrize nucleotides' interactions [162]–[164]. Around the RNA helix a cubic box was built with a minimum distance of 8 Å between the box walls and the helix. The empty space inside the box was filled with water molecules, described by the *OPC* force field [14] and salt ions (sodium and chloride) to neutralize the system and to create a physiological concentration of ∼0.1 M.
Furthermore the masses repartitioning option for hydrogens and heavy atoms was used to allow larger time steps of 4 fs during simulations [57]. After energy minimization (1000 steps steepest gradient method followed by 1500 steps conjugate gradient method) the system was evenly heated to 300 K in

200 picoseconds simulation time, followed by 100 ps equilibration. During this phase all non-hydrogen atoms of the RNA helix were stabilized by positional restraints with a force constant of 4.0 kcal·mol^{-1}·Å^{-2}. The box condition was set to constant volume mode while the system heated up to adjust the pressure. Long-range electrostatic interactions were calculated using the particle mesh Ewald method [20], [21] in combination with a 10Å real space cut-off. During simulations all bonds involving hydrogen atoms were constrained by SHAKE [58]. The temperature of the system was controlled by a Langevin thermostat. (see 2.3.1)
After equilibration at 300 K the pressure was kept constant and the restraints of the RNA helix were reduced stepwise to zero in four simulations, each lasting 80 ps. (4.0, 2.0, 0.5, 0.5 kcal·mol^{-1}·Å^{-2}) During the last step 0.5 kcal·mol^{-1}·Å^{-2} were applied only on the bulge region, to allow the rest of the helix to equilibrate while keeping the bulge fixed in the middle of the helix. (see figure 5.7)
Finally a production run was performed over 1 µs. Every 16 ps a frame was saved. The whole protocol was done for 10 simulations of an dsRNA containing an A-bulge loop and 10 simulations of an dsRNA containing a U-bulge loop.

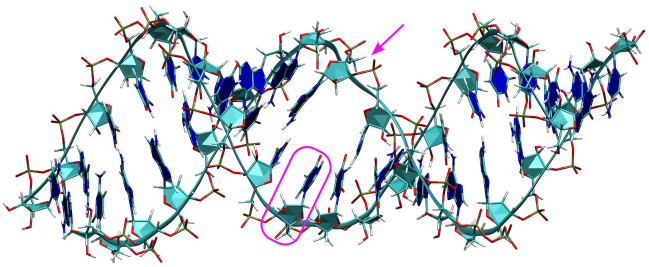


Figure 5.7: **Double-stranded RNA helix with bulge** In the middle of the RNA helix a single unpaired nucleotide (marked with magenta frame) causes a slight deformation of the helix's backbone (magenta arrow).

5.3 Detecting the Bulge

While screening the simulations visually, it is normally always possible to recognize the bulge location within the helix, like in figure 5.7. Sometimes, when the bulge moves, two bases of the bulge strand seem to "share" one opposite base of the non-bulge strand at the same time, but it is still clear, that the bulge is at this position. To automatically scan thousands of frames and detect the position of the bulge, we need measurable variables, which allow to identify the bulge reliably.

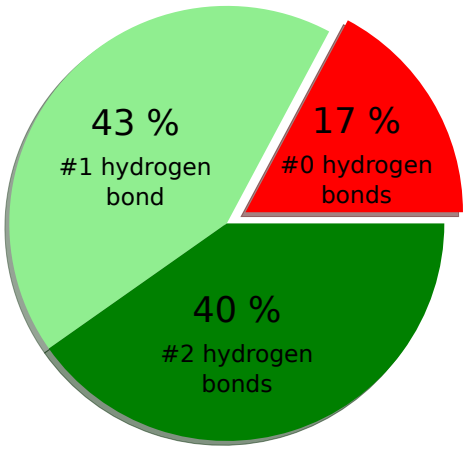


Figure 5.8: **Abundance of hydrogen bonds in dsRNA**. Statistic of the amount of hydrogen bonds being present between A-U base pairs of dsRNA during 2 μs of simulation. 17 % of the "WCF pairs" showed no bond at all, leaving them unpaired. (parameters for hydrogen bonds: distance = 3.0Å, angle = 135°)

Since the difference between dsRNA and ssRNA is by definition the presence of WCF base pairs, which in turn are formed by hydrogen bonds of opposing bases, it is obvious to look for these bonds first. The identification of the bulge should work by finding the base, which forms none of these specific hydrogen bonds to another base. A scan of simulations of A - U dsRNA without any bulge quickly reveals the problem of this approach. Opposing bases in a double-stranded helix don't bind to each other with all "required" WCF hydrogen bonds or sometimes don't bind at all. Only 43 % of all A - U base pairs bear two expected hydrogen bonds. 17 % even showed no hydrogen bond at all, which actually classified them as unbound. (see also figure C.1) This is not too surprising, because a free RNA helix is not an inflexible structure. It wobbles around and can even open up randomly, especially at the ends. Two opposing bases still stay vis-à-vis most of the time, because the whole helix is primarily stabilized by base-stacking between the aromatic rings of the nucleotides. [165], [166] So even in an unperturbed double-stranded helix (without bulge loop) opposing nucleobases are not continuously WCF paired, but break up occasionally.

To identify the bulge base anyway we try to assign each base of the non-bulge strand a corresponding base of the bulge strand. The base that is left at the end without an assigned partner base is the bulge base. An intuitive way to assign opposing bases to each other, is the minimum distance between them. Since every base pair is slightly shifted due to the helix's twist, this measurement can also lead to wrong allocations if one base moves up or down (along the helix's axis) and thus gets closer to the opposing base of its neighbouring base. (see figure 5.9)

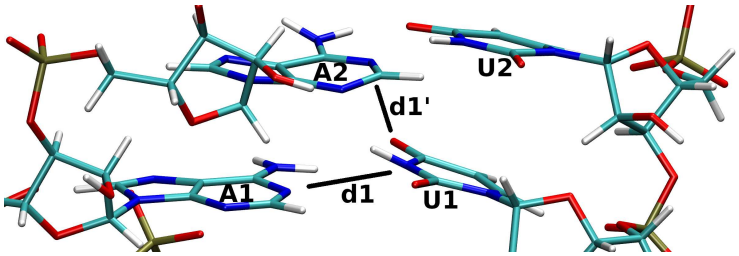


Figure 5.9: **Distances between dsRNA base pairs.** Base U1, which is actually paired with base A1, temporarily moves closer to base A2 of base pair A2 - U2. In this situation the minimum distance d1' between U1 and A2 can be shorter than the minimum distance d1 between U1 and A1.

To avoid such wrong allocations, we introduce another measurement. A base which approaches the "wrong" opposing base from below / above can only bind to this base (and replace its neighbour) if it also tilts itself towards this base to form hydrogen bonds. If they stay parallel, the angle between acceptor and donor of the possible hydrogen bond is not small enough. We use the orientation of the bases' planes, defined by the ring atoms, to get another variable to assign bases to each other. (see figure 5.10) The intersection line of these planes should be located close to the middle of both bases. So we measure the mean distance of both bases to this intersection line as an additional measure. A base which is orientated like an opposite one, but displaced by a small offset, produces an intersection line, which is far away.

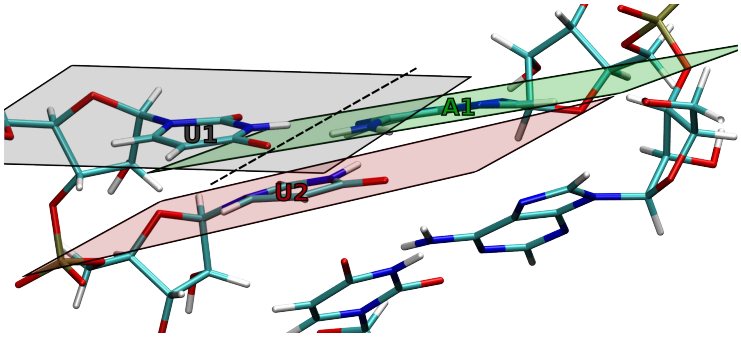


Figure 5.10: **Intersection of fitted planes.** Bulge base U2 (red) is closer to base A1 (green) than the opposing base U1 (grey) of A1. Instead of direct distances, the distance to the intersection line of base planes can be measured. The base planes are fitted to bases' rings. Since A1 and U2 are nearly parallel, the intersection line of their planes is far away from both bases.

The two described measures were used to find partner bases for those cases, which were not paired by WCF hydrogen bonds and finally identify the bulge base within each frame of the simulation automatically.

Dihedral Angles of Nucleobases

Nucleobases, which are not paired to an opposing base, are more likely to leave the usual double helix conformation. The only way for them to dodge is within the approximate plane that describes the ring system of the base, circling around the ribose sugar ring, which is built-in the RNA strand's backbone. If the bulge base loops out of the helix, it does not clash with the other WCF base pairs any more and is better accessible to potential binding molecules. On the other hand this twist leads to a deformation of the helix's backbone and the base has to leave the stabilizing stacking conformation with its neighbouring bases. Furthermore the bulge can't move along the helix axis no longer, when looped out, since it cannot pair to an opposing base. So we are also interested in the state of the current bulge base.

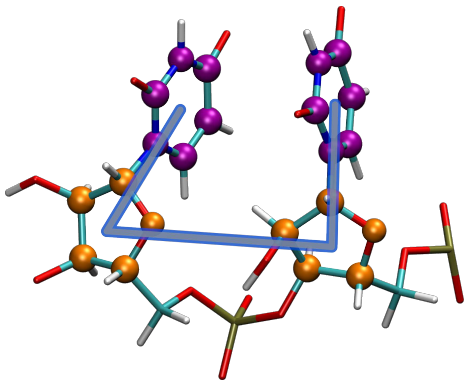


Figure 5.11: **Twist between nucleotides.** The centres of the nucleobases and the riboses are used to calculate the dihedral angle between two nucleotides.

To measure this state, a dihedral angle between neighboured bases is useful. Therefore we define two points for each nucleobase. One is the geometric center of the sugar ring and the other one is the center of a base ring. With these two points of each nucleotide we can calculate dihedral angles between two neighbouring nucleotides. (see figure 5.11)

Due to the twist of the double helix, every nucleobase is shifted slightly relative to its neighbours by $14.7^\circ \pm 4.5^\circ$. This corresponds to the mean reference value of two 1 µs simulations of dsRNA. (see figure C.2) The reference values were subtracted from all dihedrals of the bulge RNA helix. Since there are two nucleotides needed for one dihedral angle, each nucleotide has two dedicated dihedrals, one to each neighbour. To get one value for each residue the average of both dedicated dihedrals was calculated.

A problem of this method is, that we don't know which one of the two nucleobases, that are included in one dihedral value, is looped out. Or in other words, a looped out nucleobase also falsifies its neighbours' dihedral values.In order to prevent this error, we also look at the dihedrals of the neighbours' neighbours of one nucleobase. (see figure 5.12) If the dihedral angle of a base's neighbour is outside a reference interval, while the dihedral angle of the other neighbour is within the reference interval, we chose only the dihedral angle of the unaltered side instead of averaging both sides. As the reference interval we chose 90% of the dihedral angles of the dsRNA simulations mentioned above. (figure C.2)

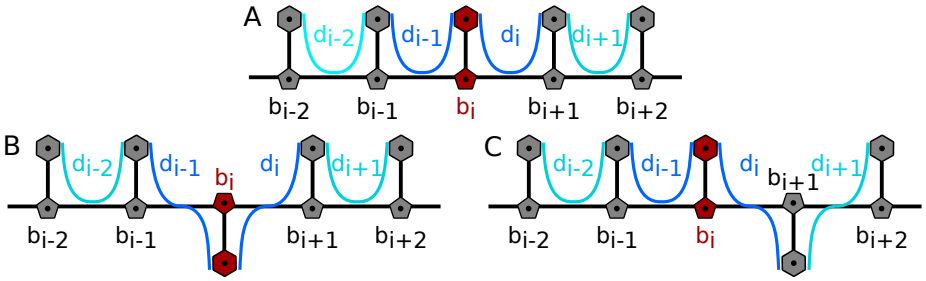


Figure 5.12: **Averaging of neighbouring dihedral angles.** A) To calculate the "dihedral angle" of one nucleobase b_i, both neighbouring dihedral angles, of which b_i is part of, d_{i-1} and d_i are averaged, if the dihedral angles next to these ones are within the reference interval. B) If b_i loops out, it alters $di-1$ and d_i by the same absolute value. C) If for example b_{i+1} loops out, it alters only one neighbouring dihedral angle of b_i, but also d_{i+1}. If d_{i-2} is within the reference interval, but d_{i+1} not, just the left dihedral angle of b_i, d_{i-1} is chosen as "dihedral angle" of b_i instead of the average.

If a bulge loops out of the helix, it can deform the backbone of the bulge strand. (figure 5.13) This deformation can cause problems in the method we used to calculate the dihedral angles. If the four points that define a dihedral angle slip too much on one line, slight shifts in one point can cause large fluctuations in the dihedral. To prevent this, we used an imaginary point between the two ribose rings next to the looped out bulge base. Expecting that these two nucleotides stay within the helix's pattern, this corresponds to an interpolated point within the backbone at the bulge's position.

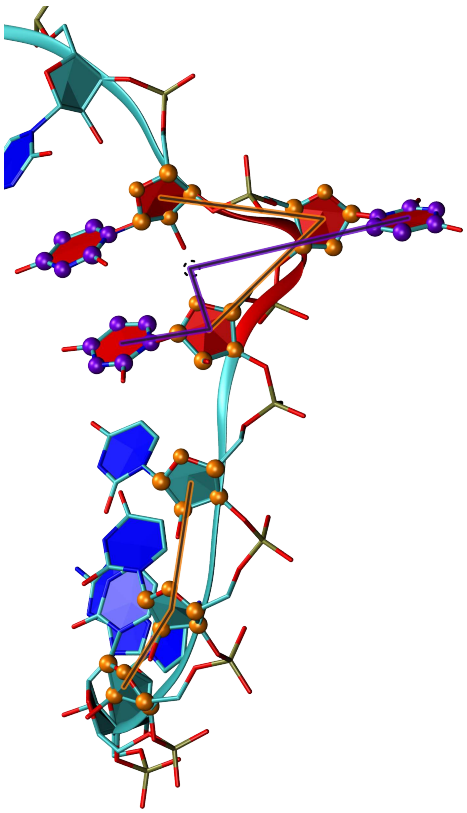


Figure 5.13: **Backbone deformation.** A looped out bulge deforms the RNA backbone (red part). To improve values of dihedral d (purple), an imaginary point (dashed black circle) in the middle of the two ribose rings next to the bulge residue was used if backbone angle (a2) deviated from the reference value (a1).

The deformation was detected by measuring the angles between the ribose rings along the backbone. If this angle was outside of a reference interval the correction was applied. The interval included 90% of all angles of 2 µs dsRNA simulations.

5.4 Dynamics of the Bulge

The first goal was to locate the bulge along the helix. Since the bulge strand of the helix is 19 residues long and we always assign the bulge base to one position, we received distinct bulge positions ranging from 1 to 19. Assuming that the outer C-G pairs prevent the escape of the bulge, since they are bound stronger, we expect bulge positions between 3 and 17.

A distribution of the bulge position along the RNA helix for all 10 simulations (10 µs total) of an A-bulge and a U-bulge respectively is shown in figure 5.14. Each bulge started at position 10, which explains the local maximum in the middle of the distribution for both types of bulges. What strikes us first is, that the U-bulge tends clearly to the helix's 3' end, where it lingers half of the simulation time. It is rarely located at other positions, except the starting position and position 13.
The A-bulge also has higher residence time at the borders, with a bias to the 5' end. However it lingers clearly more time on most positions in-between the helix's ends than the U-bulge, without any other preferred positions. In general the A-bulge distribution is more spread than the U-bulge ones.

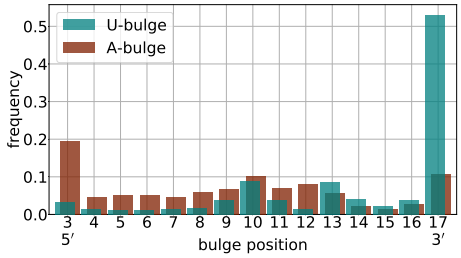
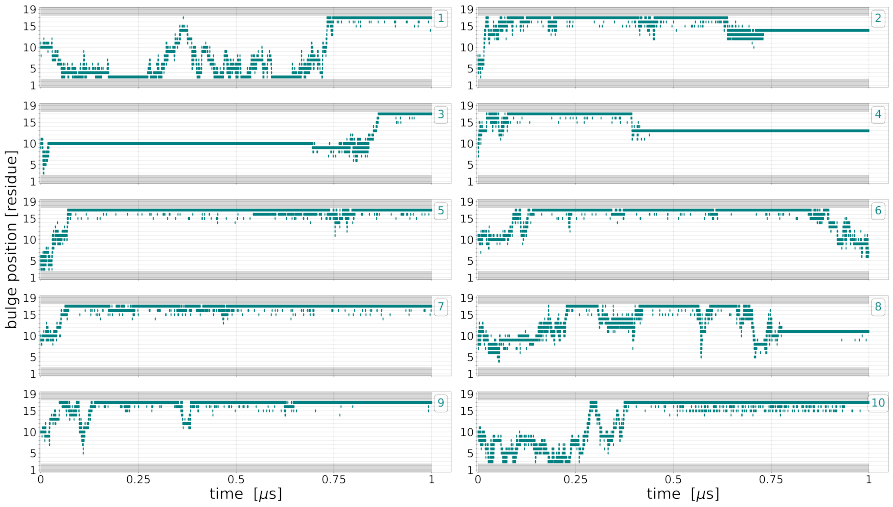


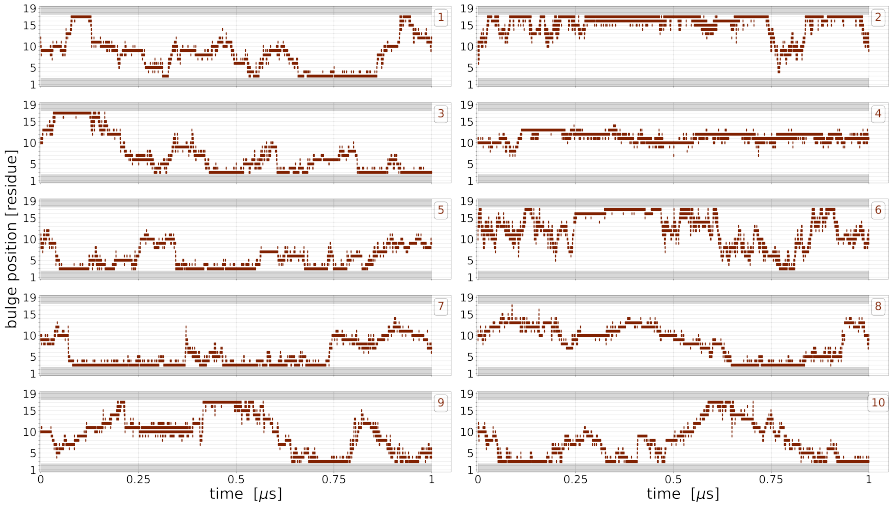
Figure 5.14: **Distribution of bulge positions.**

Looking at the time-resolved series of bulge positions for each simulation, reveals another difference in behaviour between both bulge types. The U-bulge (see figure 5.15a) stays at a single position for long periods, whereas the A-bulge (see figure 5.15b) seems to be much more mobile, changing direction several times.
It has to be mentioned, that bulges of both setups only move between few time steps of the simulation due to the high time resolution. U-bulges change their position only in 3.5% of all frames, A-bulges in 3.8%. A detailed example of bulge movement during a time window of 1.6 ns (100 frames) (figure C.3) illustrates this behaviour. The time window was chosen around the highest mobility of each bulge (most jumps between time steps). In case of the U-bulge its position moves in 50% of all time steps, the A-bulge moves in 64%, but most of the

movement is flipping forward and backward between two positions.
Since the U-bulge stays at certain positions for long periods of time, it is actually more mobile during the rest of the time, than the A-bulge (in contrary to the impression of figure 5.15).



(a) **10 simulations of a U-bulge.**



(b) **10 simulations of an A-bulge.**

Figure 5.15: The location of 10 U-bulges (teal) and 10 A-bulges (cedar) within the helix is shown on the y-axis during the simulation time of 1 µs. The two C-G pairs at each end of the helix are marked by grey areas, which form stronger WCF bonds and thus confine the bulge within the helix.

To better understand the different behaviour of the two bulges, let's look at the dihedral angles of each type of bulge. As an example one simulation of each type of bulge is shown in figure 5.16. An overview of all simulation is given in figure C.4.

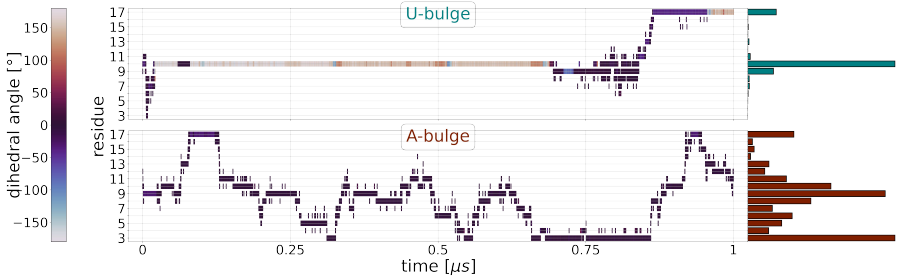


Figure 5.16: **Position and dihedral angle.** The position (residue) of one U-bulges (teal) and one A-bulges (cedar) is shown. The color of the curves represents the dihedral angle of the bulge. A distribution of the frequency, each position is taken, is shown on the right side.

The example U-bulge shows some quick movement at the beginning where it travels to the 5' end but quickly returns to the middle of the helix. There it stays nearly half of the simulation time at residue 10. Finally it starts moving again around this position for ∼150ns, before it moves to the 3' end, where it stays with very short interruptions until the end of the simulation. The distribution of the positions shows, that the bulge takes mainly three positions during the whole simulation. The mobility (jumps between residues compared to all time steps) of the U-bulge is 2.0%.
The A-bulge example seems to be much more agile, taking most of the positions for some time. However its mobility is also 2.0%
Additionally to the position of the bulge, the dihedral angle of the current bulge residue is displayed by the color of the plot markers. What we can see, is that the U-bulge's dihedral angle stays low around 0°, while it is moving, but during the inflexible periods the dihedral angle reaches high values according to absolute degrees. So during these periods of motionlessness the bulge loops out most of the time. An example structure of a bulge with low dihedral angle and one of a looped out bulge is shown in figure 5.18 (A,B).
If we neglect the states of the U-bulge, when it is looped out, its mobility raises, as expected. Choosing only states of the U-bulge which dihedral angles are within an interval containing 99% (95%, 90%) of the dihedral angles of the A-bulge, results in a U-bulge mobility of 9.4% (11.2%, 12.1%), while the A-bulge's mobility stays at 2.0% if we apply the same interval of dihedral angles. So the U-bulge is actually much more volatile than the A-bulge, if it is not looped out. Otherwise it stays at one position.

Analysing the distribution of the bulges' dihedral angles (see figure 5.17) gives a better imagination of the different behaviour of U- and A-bulge. The A-bulge's dihedral stays most of the time around 0° in an interval of $\sim \pm 25°$. This state corresponds to a WCF paired base within a dsRNA helix, which is not looped out. An example is presented in figure 5.18 A).
A second state of the A-bulge can be recognized around 35°. This one occurs much more infrequently than the main state around 0°. It corresponds to a slight loop out of the base towards the minor groove. As shown by the example in figure 5.18 C), it nearly does not deform the helix's backbone. Since the base gets closer to the opposing U-base of its neighbouring base, it forms additional hydrogen bonds with this base, stabilizing this position. From this dihedral state the A-bulge cannot move neither towards the 3' end nor the 5' end. (see histograms a) - c) of figure 5.17)
The U-bulge shows three distinct dihedral states. One is nearly identical with the main state of the A-bulge around 0° and corresponds to a not looped out state. Compared to the A-bulge the U-bulge takes this state much rarer. Most of the time it is in a state around -40° towards the minor groove. By contrast with the corresponding A-bulge the U-bulge loops out a little bit further and the base is tilted from its normal orientation. (see figure 5.18 D) It also forms a hydrogen bond with the opposing RNA strand.
The third state of the U-bulge is a broad one close to ± 180°. It corresponds to a completely looped out base (see figure 5.18 B). The RNA backbone is heavily deformed in this state, which only occurred for the U-bulge.

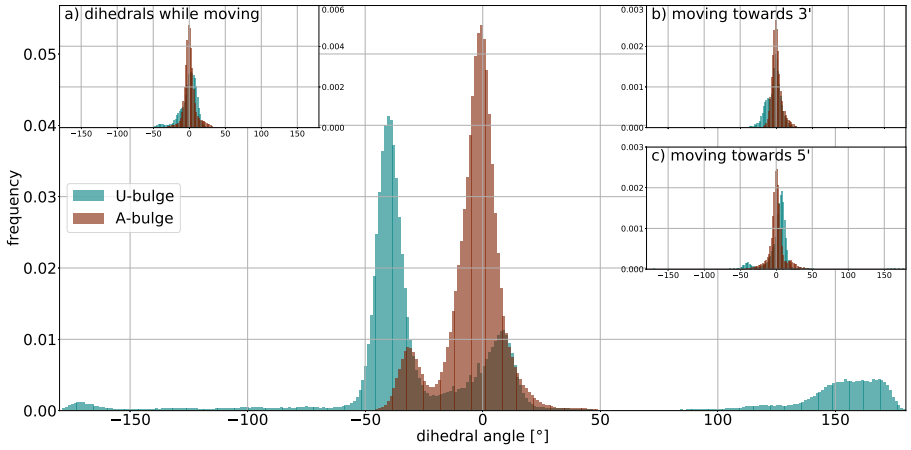


Figure 5.17: **Distribution of dihedral angles.** The main histogram illustrates the distribution of all dihedral angles of the current bulge residue (U-bulges (teal) / A-bulges (cedar)). Histogram a) shows only dihedrals of bulges, which are changing their position within the next simulation step. Histograms b) and c) show dihedrals of bulges which are moving towards the 3' or 5' end respectively.

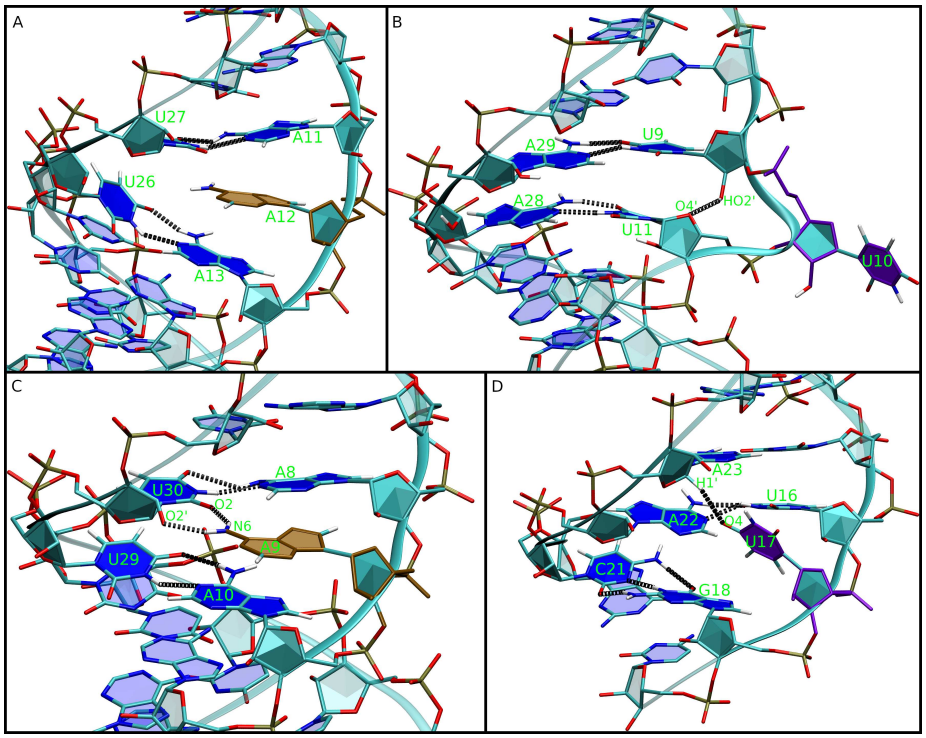


Figure 5.18: **Snapshots of different states of a bulge.** A) A-bulge (brown) of dihedral angle around 0° and no hydrogen bonds (black dotted lines) to opposing or neighbouring residues. B) U-bulge (magenta) in completely looped out state and distinctly deformed RNA backbone. C) A-bulge slightly looped out towards the minor groove and hardly deformed RNA backbone. Bulge base A9 forms additional hydrogen bonds with base U30, which is WCF paired to base A8. D) U-bulge slightly looped out towards the minor groove and hardly deformed RNA backbone. Bulge base U17 is tilted compared to its neighbours and forms an additional hydrogen bond with opposing base A23.

In general both bulges only move if their dihedral angle is close to 0°. The U-bulge shows a slight tendency of negative values (loop out into minor groove) when moving towards the 3' end. Looping out to this side gets the base closer to the opposing base which is shifted to the 3' end. On the other hand when the U-bulge moves towards the 5' end a tendency of positive dihedral angles (loop out into major groove) can be recognized. Consistently this direction gets the base closer towards the opposing base, which is shifted to the 5' end.

5.5 Discussion

As shown by the presented results, we are able to identify and track a U-bulge and an A-bulge within an 19 residues long dsRNA helix during 10 simulations of 1µs for each type of bulge. The bulge is prevented to escape the helix by two C-G pairs at each end of the helix, quite effectively. Only few times these C-G pairs opened for short time, but never when the bulge was close to them.
Although conditions for U-bulges and A-bulges are the same, they exhibit quite different behaviour. While the A-bulge stays most of the time in place as if paired with a counterpart, the U-bulge loops out much more often and further. The reason is probably the higher stacking energy of adenine (A) bases between each other due to their larger ring system. So an adenine nucleobase which is not WCF paired to an opponent uracil nucleobase is still held within the regular dsRNA helix structure by its direct adenine neighbours, more than an uracil nucleobase by its uracil neighbours.
Since a looped out base cannot switch to a neighbour base for obvious reasons, U-bulges present a less volatile behaviour than A-bulges, because these looped out states are quite persistent (see figure 5.15). During the not looped out intervals, the U-bulges clearly behave more mobile than A-bulges. Again the lower stacking energy of the uracil (U) bases may be the cause. It allows U-bases to move around more than A-bases. (see figure C.5) To pair to a neighbour's opposing residue and replace the neighbour, a nucleobase has to shift sideways to get in front of the new WCF partner base. Therefore a more mobile nucleobase can lead to a more mobile bulge.
The here presented RNA setup was build from scratch and is completely artificial without a known appearance in nature. However the nucleobases and helical structure occurs in nature plenty of times, just the long uniform sequence of A/U-pairs is not common. So the characteristics of an adenine or uracil bulge in nature may be similar, but they are affected by many more parameters, like different neighbouring nucleobases or a more complex tertiary structure. For instance a uracil bulge may loop out more often and further than an adenine bulge under similar circumstances. Whereas the movement of a bulge along the RNA helix we observed is probably not very relevant for natural RNA structures, since the bulge would meet a C/G-pair quite fast.
A field of interest could be the engineering of DNA / RNA origamis. These are also artificial structures, so a homogeneous sequence of A/U-pairs could be created. The exploit of a bulge in such a sequence could be the transportation of a signal along the helix, like an uracil bulge starting at the 3' end and provoking a reaction when reaching the 5' end. For this purpose it would be very useful to be able to manipulate the bulge's movement. Further research in this direction would be interesting and could be done perfectly with our bulge setup. Probably it would be reasonable to first analyse additional measures of the bulge and the helix to better understand what influences affect the bulge's behaviour in which way. Thinkable interventions to manipulate the bulge would be for instance to stretch / press or tilt the helix. Other possibilities would be to act on the helix by global measures like temperature or salt concentration.

Of course further effort could also be put into improving the applied detection algorithms of the bulge. A 'temporary' disorder of the helix, meaning several residues unpairing and arranging in a different way, can disturb the detection, especially if it appears close to the 'real' bulge residue.

Chapter 6

Summary and Outlook

As we have seen particularly clearly in the past few years of the Covid pandemic, understanding biological processes at the molecular level can be very helpful in improving our lives. We have a large number of modern experimental instruments at our disposal for this purpose. However even if these are continuously improved and advance into smaller dimensions and time scales, they still reach limitations. We can further shift these limits with the help of Molecular Dynamics (MD) simulations, which were presented in the beginning of this work.
An important area of research is the folding of proteins, which is constantly taking place in our bodies. The most common protein in our body is collagen, which forms elongated triple helices. The process of folding into this helical structure could be simulated in detail using multiple simulations of short collagen peptides. A zipper-like mechanism was observed, which is caused by segmentation of the sequence into groups of three amino acids by the periodic appearance of glycine. Furthermore, the simulations showed that only the union of all three collagen strands leads to stable structures and that a stable initial nucleus is important for the subsequent folding. Rare eversion of a strand was also observed. The resulting loop – a perturbation during folding — caused the helix to bend. As further perturbations, mutations of individual glycine residues were simulated, which clearly impeded the folding process.
Each examined aspect offers many approaches for further investigation. More simulation and longer peptides could reveal other rare wild type misfoldings, such as loops of different sizes for instance. The positional restraints, which mimic the stabilising HSP47 proteins could be omitted by simulating these proteins in complex with collagen to stay closer to reality. Finally, further mutations or post-translational modifications of the wild type could be tested.
As the first example showed, it happens that proteins fold more or less well into their natural structure despite a local mismatch during the folding process. If it is not an isolated random but a systematic mismatch, caused in turn by a mutation in the protein sequence, the effects can be serious. Sometimes the stability of the protein can be so severely impaired that it can no longer fulfil its function. In special cases, this leads to the protein unfolding again and forming

new undesired structures (e.g. amyloids). Consequently, it is of interest to know and understand the consequences of a mutation. A method for this was presented using the transport protein Transthyretin (TTR). The combination of MD simulations and free energy calculations made it possible to correctly assess the tendency towards or against amyloid formation in almost 80% of 36 selected mutations of this protein. The results could confirm the assumption of former studies that the destabilization of the wild type correlates with the amyloid formation. Furthermore, it became clearly visible how much the amyloid fibrils are energetically more favourable compared to the globular protein, which explains their high stability.
The presented method is strongly designed for speed to quickly analyse many cases in order to gain an advantage over experimental measurements. Presumably it can be further accelerated by finer adjustment of its parameters. In addition, there are of course many more mutations that can be examined and countless other proteins with other mutations.
A mutation in a protein or a Deoxyribonucleic Acid (DNA) sequence does not always have to have negative consequences. As we know, evolution is based on the fact that every now and then a mutation leads to an improvement. In the third system examined within this thesis, a disturbance in the form of a bulge loop was inserted into an artificial double-stranded Ribonucleic Acid (dsRNA). The intention was to study the behaviour of this perturbation and to give ideas for possible applications such as signalling along the helix or a molecular switch. Clearly different behaviour between adenine and uracil bulge loops was observed. While the former usually remains within the helix structure and moves around there rather non-directionally, the latter preferably moves to the 3' end of the helix or often loops out of the helix, which severely deforms its backbone and immobilizes the bulge.
In order to make the implementation of such a system – e.g. in a DNA / Ribonucleic Acid (RNA) origami – more attractive, it would be useful if one could actively influence the bulge, for instance to migrate in a certain direction, or to loop in or out of the helix. Future studies will have to find out whether, for example twisting or stretching the helix has such an effect.

The application of MD simulations to describe processes in biological systems at the molecular level is nowadays a frequently used method. Even if some might doubt the resilience of such theoretical simulations, this thesis was able to show how MD can serve to gain new insights that would either not be possible today or would require much more effort. Of course, simulations are better the more experimentally confirmed data are included. The entire underlying knowledge for their implementation is based on experimentally gained knowledge. In the best case, experimental measurements and computer simulations support each other. The computer hardware and algorithms available today are sufficient enough, so that even on a standard PC – preferably with a modern GPU – interesting simulations can be run. Powerful but less available computing clusters and supercomputers naturally expand the possibilities enormously and are constantly being improved. The simultaneous development of powerful artificial intelligence

systems will probably further speed up development, so that at some point it will be difficult to recognise what is real.

Appendix A

Mechanism of collagen folding propagation

Structure of triple helical peptide

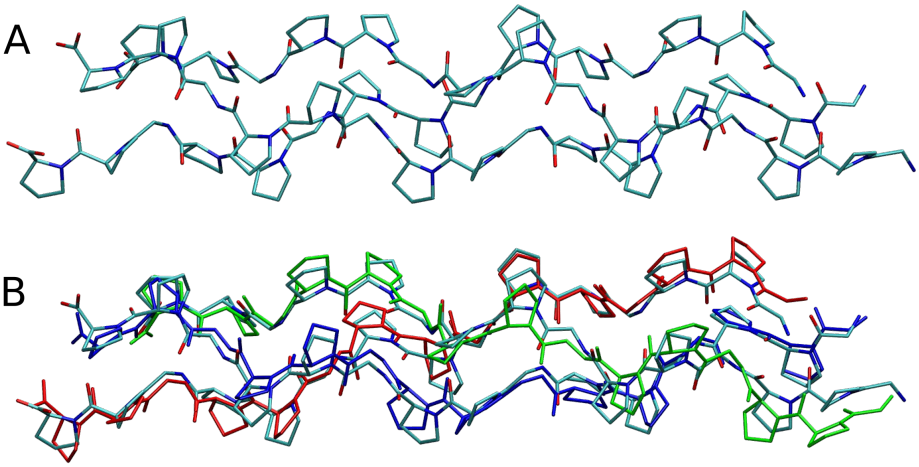


Figure A.1: **Structure of triple helical peptide** (A) Structure of a folded triple helical peptide of 3 x (Gly-Pro-Pro)$_5$ sequence (B) Superposition of crystal structure (light blue) and final peptide (each chain in different color) after folding simulation.

Comparison of fluctuations of free and restrained peptide

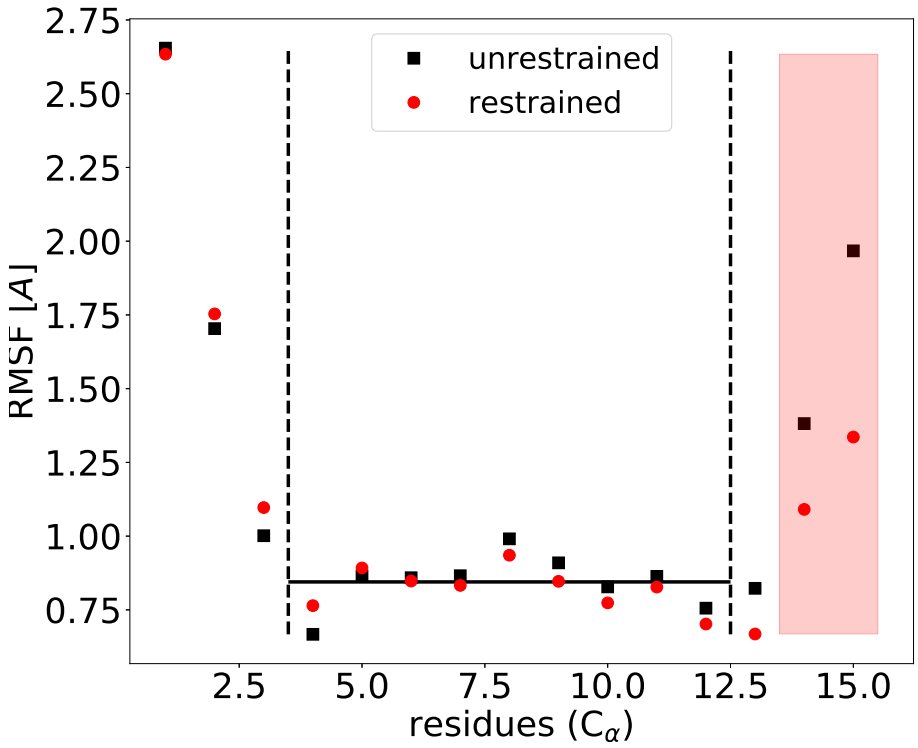


Figure A.2: **Comparison of fluctuations of free and restrained peptide.** root-mean-square fluctuation (RMSF) of an unrestrained and a restrained triple helix: The black squares represent the RMSF of the C_α atoms of the residues along the helix (mean of three residues, one of each strand). The N-terminus (left) and the C-terminus (right) show increased flexibility. The red circles represent the RMSF of a triple helix which was restrained on the two terminal residues at the C-terminus (red area). Clearly visible is that the restrained C-terminus is less flexible than the unrestrained terminus, but still more flexible than the middle part of the helix (between the dashed lines).

RMSD of folding / misfolding / mutated / folded example

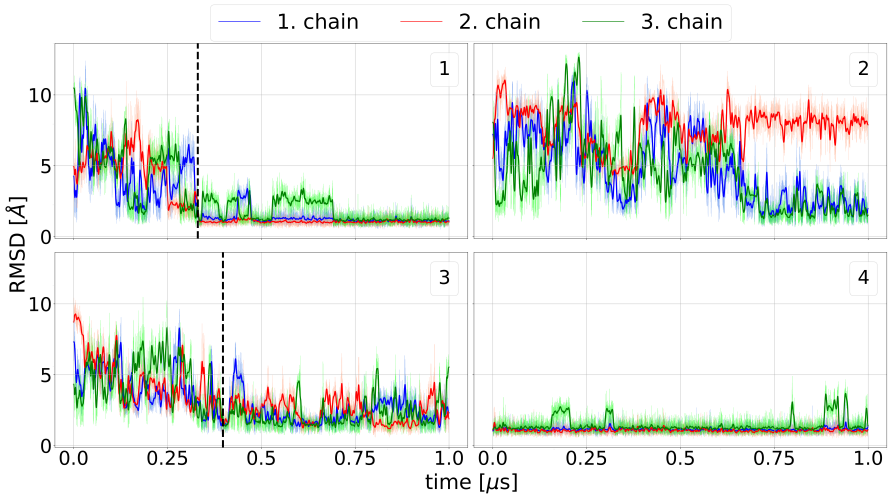


Figure A.3: **RMSD of folding / misfolding / mutated / folded example.** root-mean-square deviation (RMSD) of individual chains (with respect to the native collagen structure) vs. simulation time for 3 folding simulations (1-3) and one simulation starting from folded triple helix (4). The RMSD of each strand is visualized separately (blue, red, green). The thicker line shows the smoothed data (Gaussian filter, σ=25). The vertical dashed line marks the time point of (first) folding (RMSD of all chains below 2 Å). Example 1 indicates a successful folding of the wild type sequence. One chain in example 2 formed a loop in the second (red) strand in final part of the simulation and the RMSD of the chain remains at a high level. Example 3 shows a mutant with the 3 central Gly residues replaced by Ala. Although the RMSD temporarily reaches low values, no stable folding is observed. Finally, the 4th example shows the simulation starting from the native folded triple helix.

RMSD-triplets of all 10 WT simulations

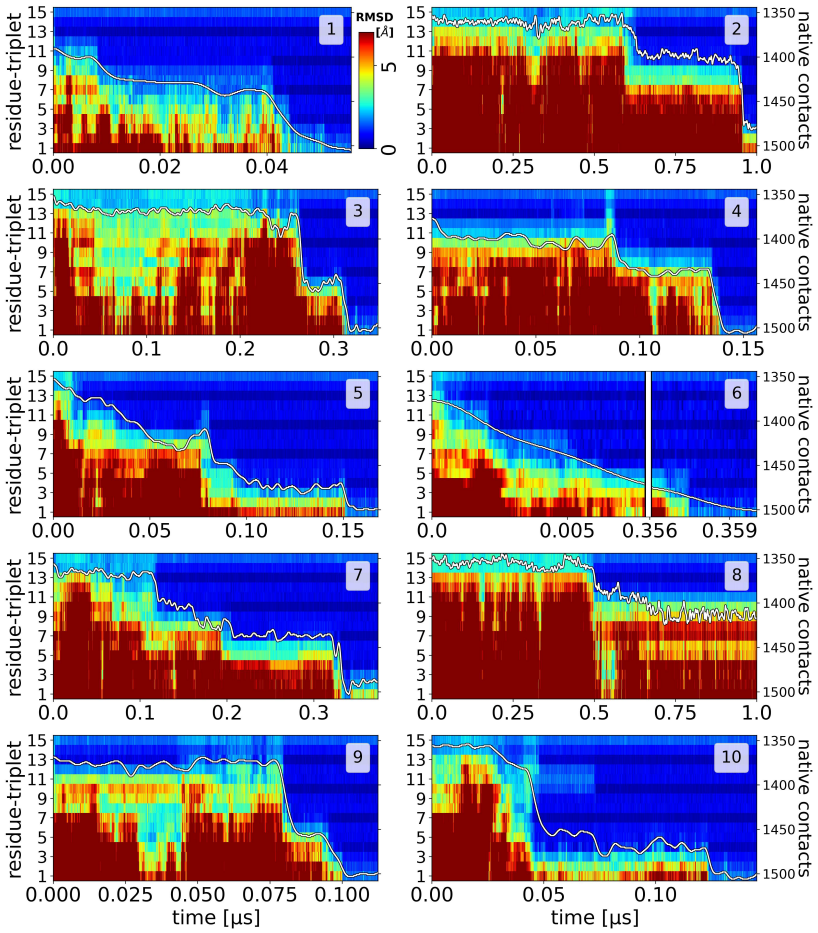


Figure A.4: RMSD of residue triplets of all WT simulations (see Fig. 3.2). Each simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (top). Each labelled stripe (1-15) in the central plot represents a residue triplet (one residue with the same number of each strand, e.g. 3 Pro with the same number that are spatially close in the native state) along the three strands. The RMSD of each residue triplet relative to the native folded structure is indicated by a color-code (color bar in the first plot). A blue color represents sampled states close to native, whereas red color corresponds to an unfolded triplet structure. The y-axis on the right side of the plot represents the number of native contacts between the three strands (white line in the plot). All simulation ran for 1 μs. To present the folding process in detail the time range of each graph was adjusted.

Detailed exemplary folding process

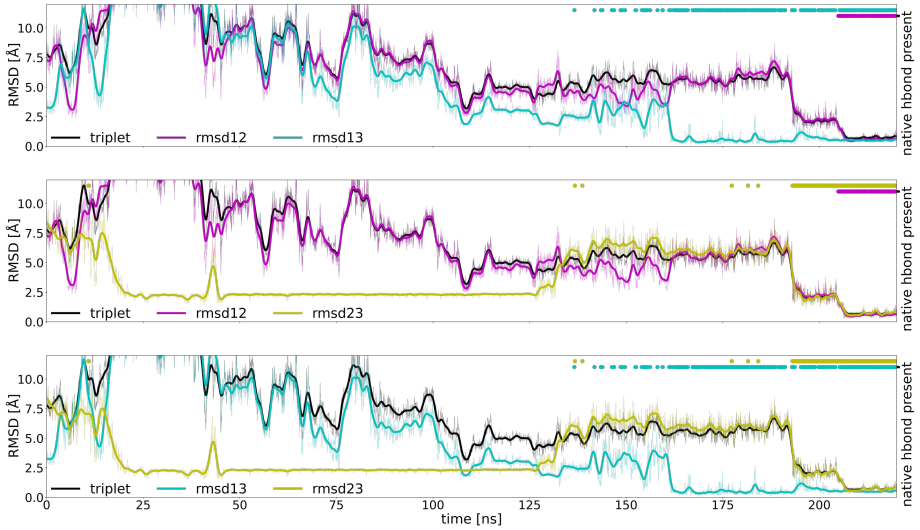


Figure A.5: **Detailed example of a folding process** at one point along the triple helix (residue 7 of chain A, B, C). Each plot visualizes the situation of one strand of the helix. The black curve shows the RMSD of all three strands. The coloured curves indicate the pairwise RMSD of 2 selected chains. (magenta: 1& 2, cyan: 1& 3, yellow: 2& 3). On the top of each plot the presence of the corresponding hydrogen bond is indicated by dots in the same colors. In the 1^{st} half the 2^{nd} and 3^{rd} strand align to a metastable conformation, but the first strand does not join them. Also the corresponding hydrogen bond is not formed. After 125 ns both strands separate again. In the last third strand 1 and 3 align to each other and the hydrogen bond is formed. Around 200 ns the 3^{rd} strand follows and the final and stable conformation is formed.

Distribution of dihedral angles folded vs unfolded

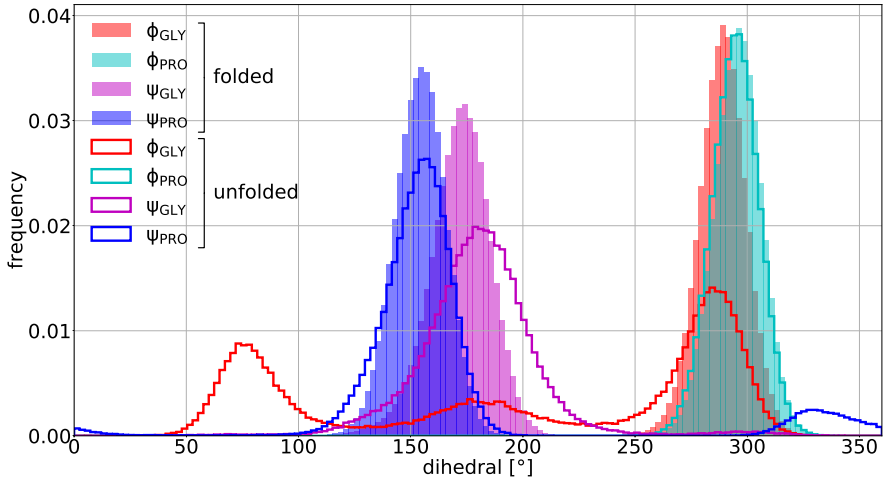


Figure A.6: **Distribution of dihedral angles for two cases.** In the folded case (filled areas) all angles populate a single state. In an unfolded case (framed areas) especially the Φ-angle of GLY shows a broad variation. The Ψ-angle of GLY is slightly shifted. The Φ-angle of PRO does not change due to its ring structure, whereas the Ψ-angle shows a second state around 330˚.

WT simulations with reduced restraints

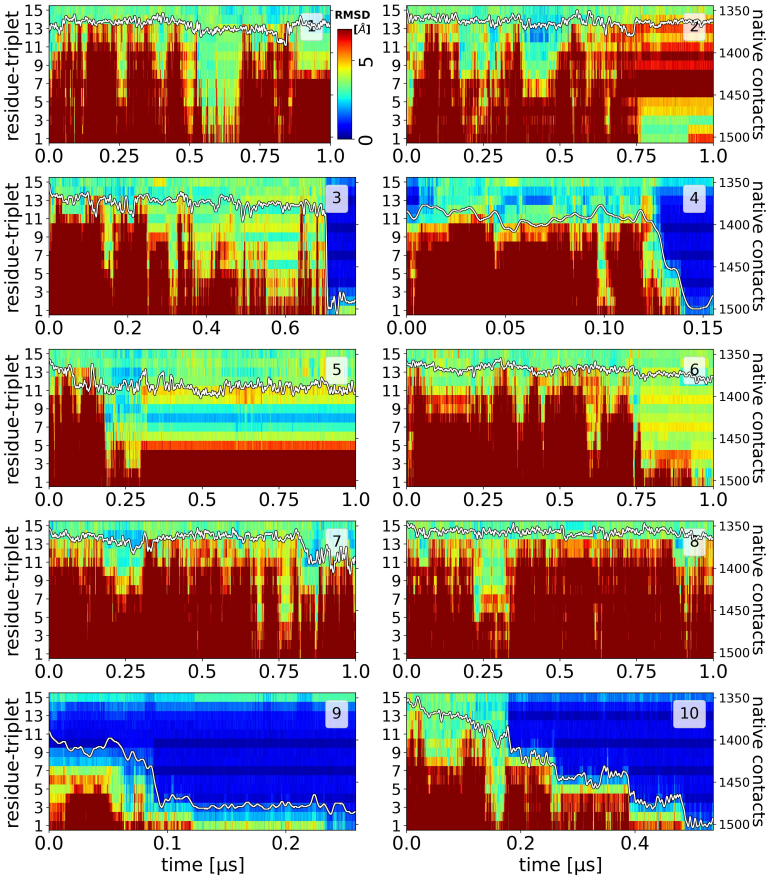


Figure A.7: **RMSD of residue triplets of all WT simulations** (like Figure 3.2) but with **reduced restraints at the C-Terminus.** Each simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (top). Each labelled stripe (1-15) in the central plot represents a residue triplet (one residue with the same number of each strand, e.g. 3 Pro with the same number that are spatially close in the native state) along the three strands. The RMSD of each residue triplet relative to the native folded structure is indicated by a color-code (color bar in the first plot). A blue color represents sampled states close to native, whereas red color corresponds to an unfolded triplet structure. The y-axis on the right side of the plot represents the number of native contacts between the three strands (white line in the plot). All simulation ran for 1 μs. To present the folding process in detail the time range of each graph was adjusted.

Snapshots of mutated area

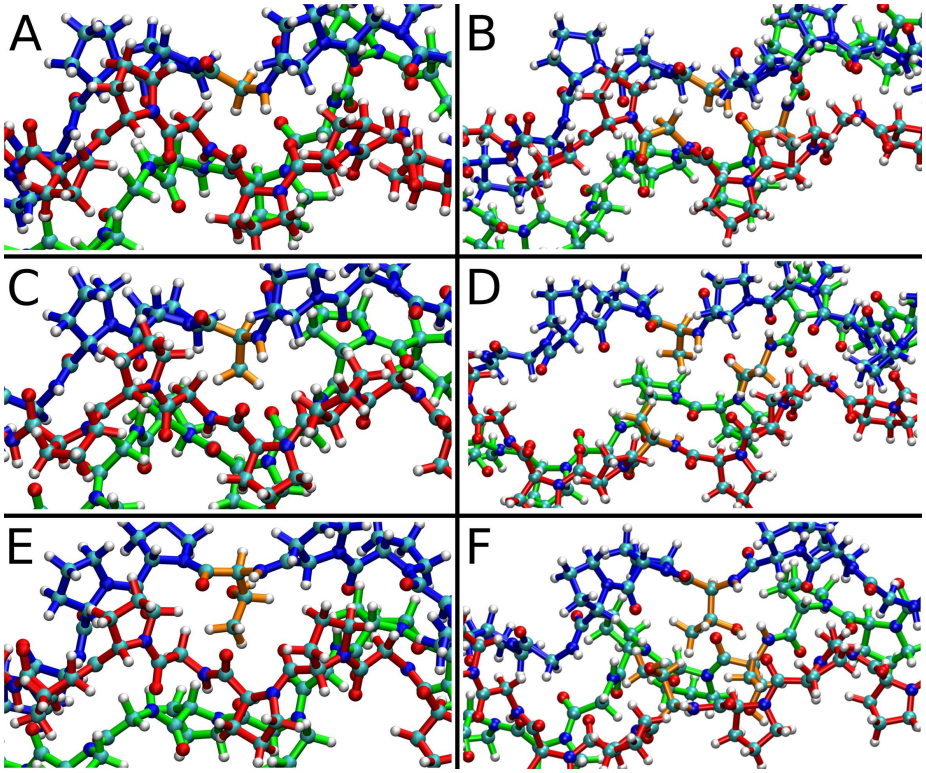


Figure A.8: **Snapshots of the mutated area of the helix.** Mutated residues are coloured orange. A: one GLY of WT, B: three GLY of WT forming a triplet, C: one mutation G7aA, D: three mutations G7abcA, E: one mutation G7aT, F: three mutations G7abcT.

RMSD-triplets of G7A mutants

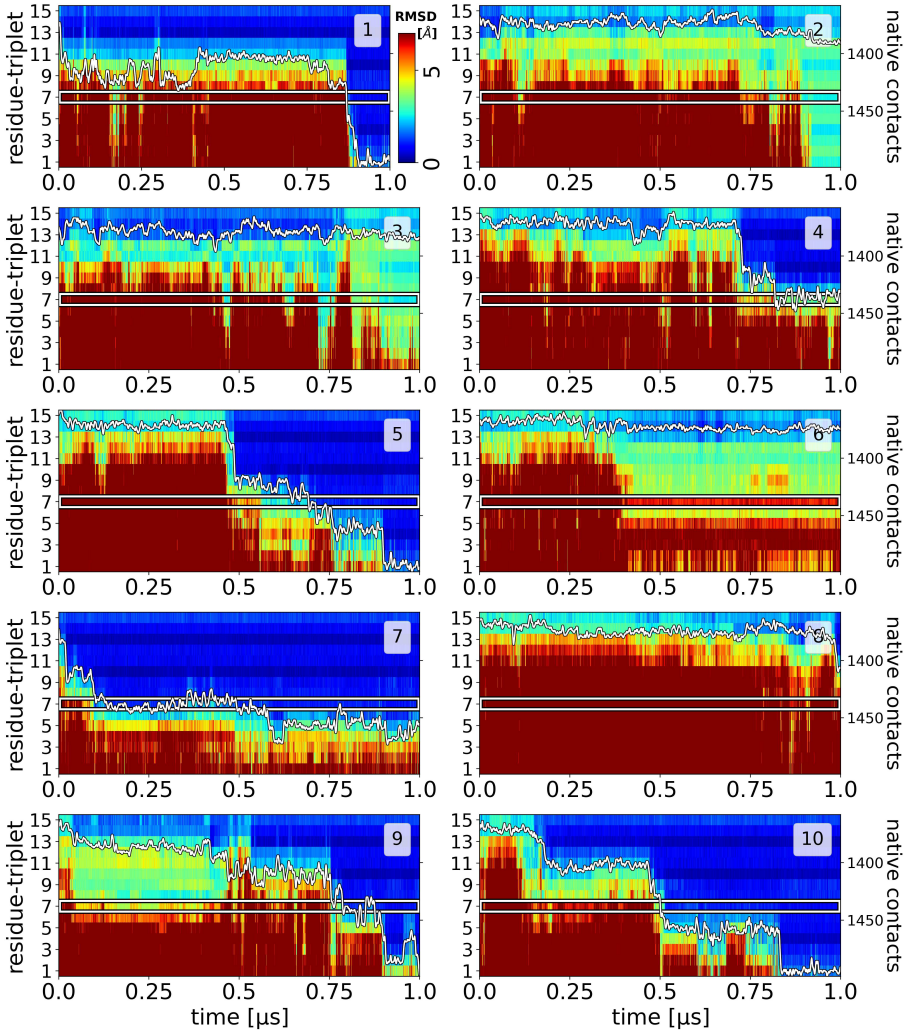


Figure A.9: **RMSD of residue triplets** of all simulations with a **single G7aA mutation** (marked by white frame). Each simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (top). Each labelled stripe (1-15) in the central plot represents a residue triplet along the three strands. The RMSD of each residue triplet relative to the native folded structure is indicated by a color-code (color bar in the first plot). A blue color represents sampled states close to native, whereas red color corresponds to an unfolded triplet structure. The y-axis on the right side of the plot represents the number of native contacts between the three strands (white line in the plot).

RMSD-triplets of G7T mutants

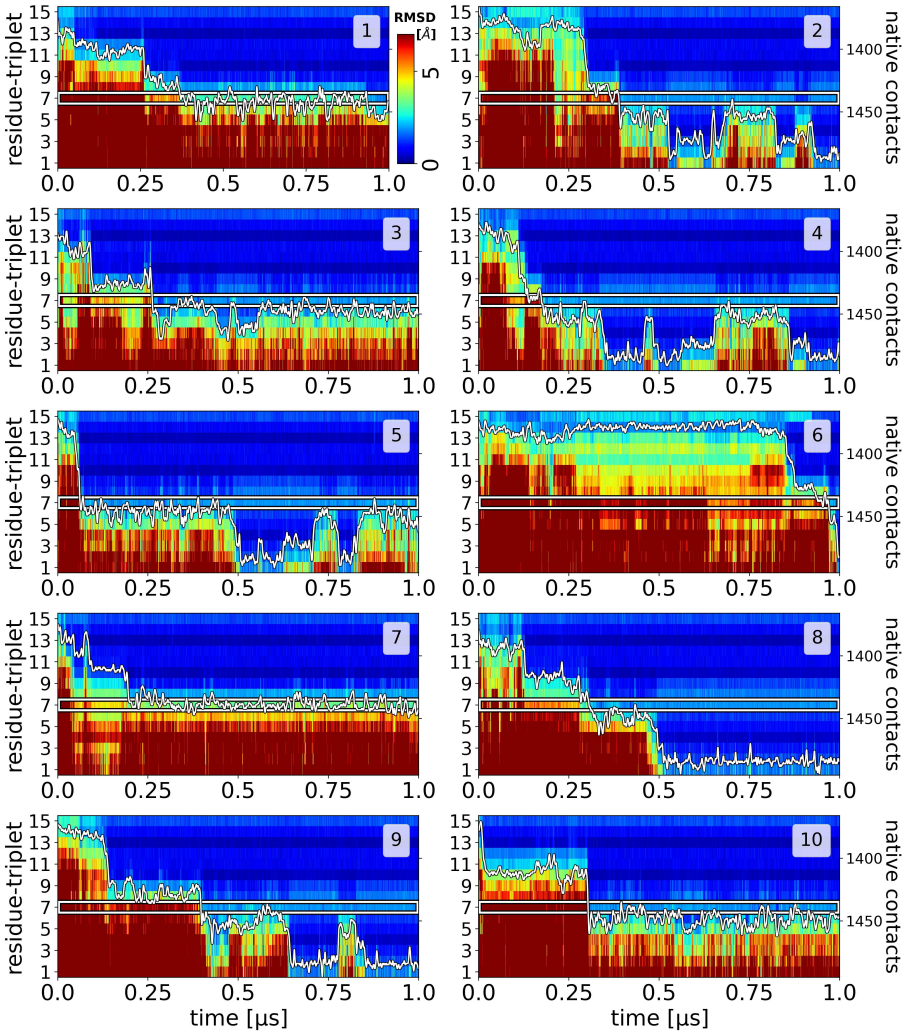


Figure A.10: **RMSD of residue triplets** of all simulations with a **single G7aT mutation** (marked by white frame). The simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (top). See legend of Figure A.9.

RMSD-triplets of G7/22/37A mutants

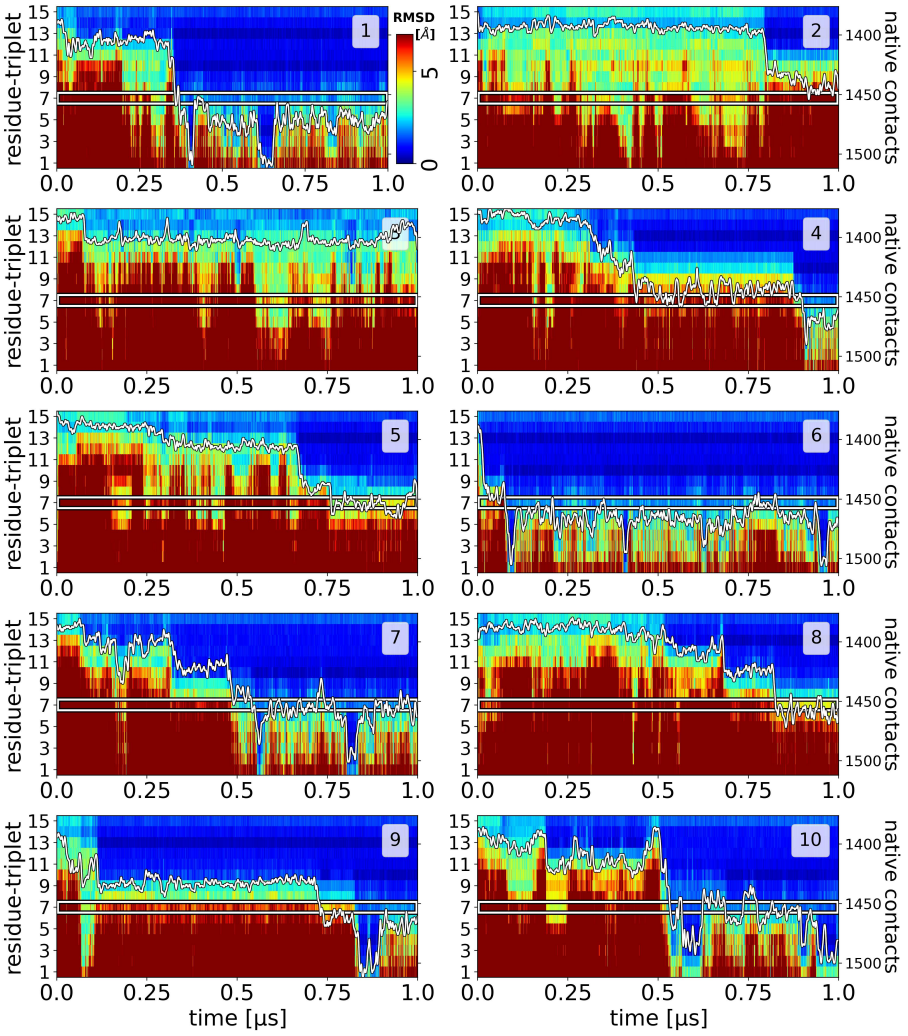


Figure A.11: **RMSD of residue triplets** of all simulations with **three mutations G7abcA** (indicating position 7 in each chain A,B,C; marked by white frame). The simulation starts from an unfolded collagen peptide with an already formed nucleus at the C-terminus (top). Same as legend of Figure A.9.

RMSD-triplets of G7/22/37T mutants

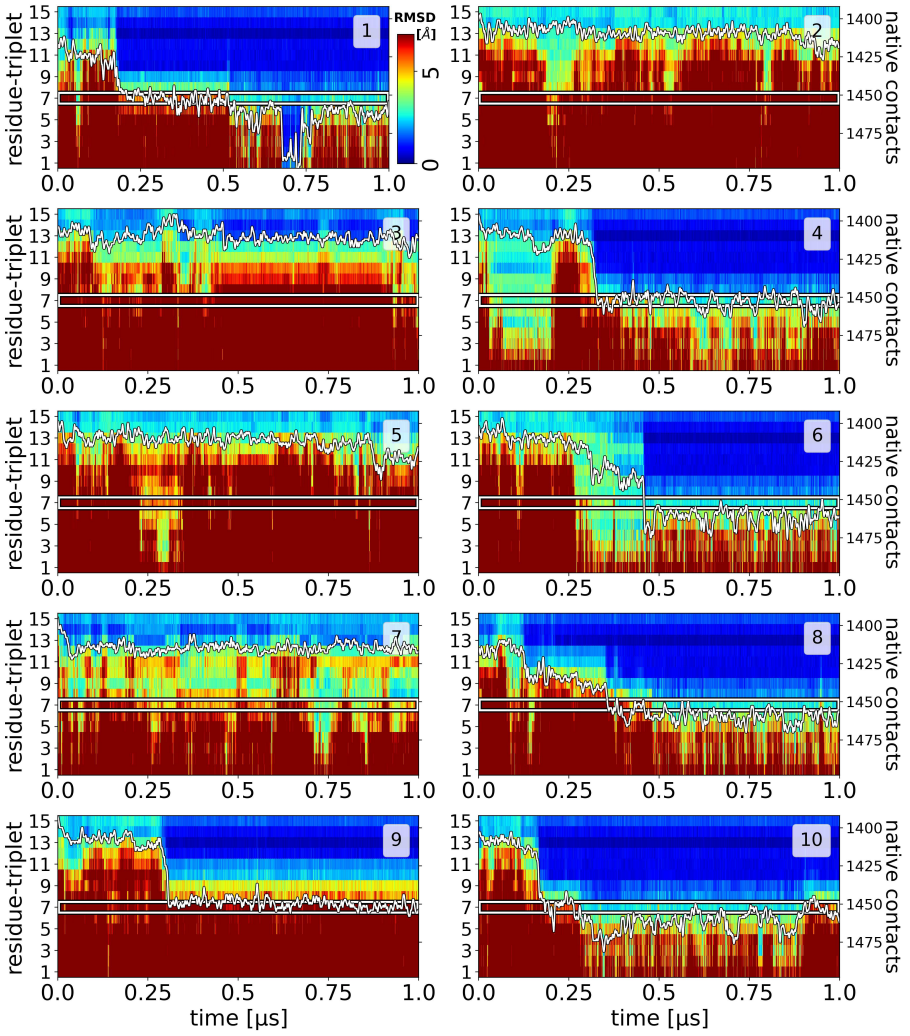


Figure A.12: Same as Figure A.11 but for simulations with **three mutations G7abcT** (marked by white frame).

Free energy variation in time

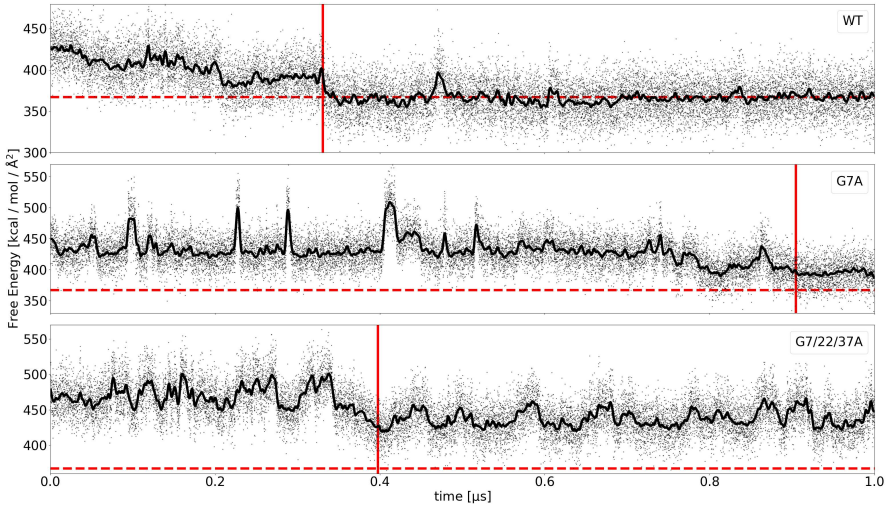


Figure A.13: **Time course of the MMGBSA energy during three exemplary simulations.** Red vertical lines indicate the time point when the RMSD of the N-Terminus reached values of a folded helix for the first time. The red dashed line displays the mean energy of a folded helix. In the first example a successful folding process of a wild-type peptide is shown (Figure 3.2 and Figure A.4 (example 7)). The second example presents a peptide with one G7aA mutation in one strand (the letter a indicates chain A of the triple helix), which folded around 900 ns, but the final energy was increased compared to a folded wild-type. The third example shows a peptide with three mutations (at position 7 of each chain) which temporarily got close to the folded state, but unfolded subsequently.

Temporary alignment of one strand

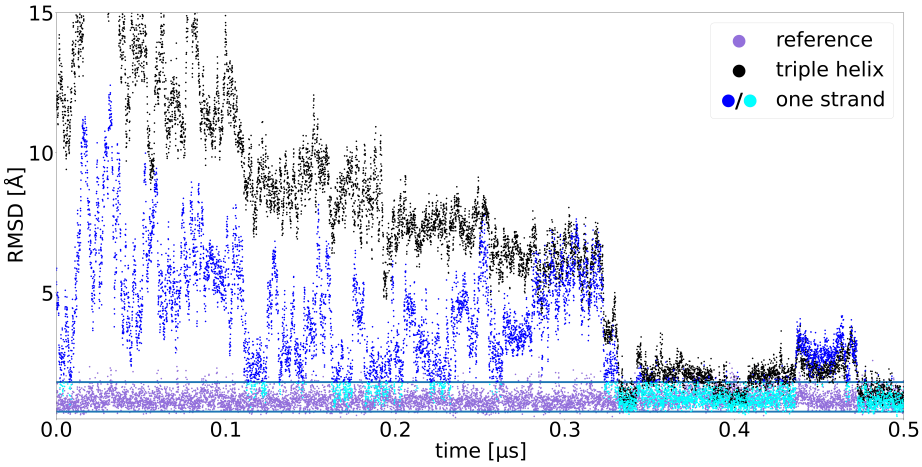


Figure A.14: **Temporary alignment of one strand.** The RMSD of one strand of a folded reference triple helix is shown in purple. The two horizontal lines mark the range wherein 95% of the reference data points are. The black series shows the RMSD of a folding triple helix, which reaches the level of the reference the first time around 330 ns, when the whole helix is folded. The RMSD of a single strand (blue) reaches this level several times (cyan) before the whole helix is folded for short periods indicating a temporarily correct but unstable alignment of this strand.

Free energy differences between beginning and end of simulations (kcal/mol)

simulation	wild-type	weak restraints	G7aA	G7aT	G7abcA	G7abcT
1	-70.7	-36.1	-80.2	-58.6	-57.8	-21.7
2	-82.3	-57.9	-76.5	-72.9	-57.4	-6.6
3	-94.0	-89.9	-36.2	-49.9	-16.6	-40.9
4	-68.9	-70.8	-56.8	-90.9	-58.6	-41.4
5	-104.8	-64.3	-73.9	-64.3	-44.4	-78.0
6	-87.4	-77.7	-28.4	-92.0	-51.9	-62.3
7	-77.9	-25.9	-66.3	-50.7	-44.5	-25.6
8	-68.3	-19.0	-58.7	-86.3	-68.6	-36.2
9	-81.3	-74.5	-95.8	-101.7	-38.6	-28.0
10	-92.0	-91.6	-71.1	-54.7	-62.6	-32.8
average	-83±11	-61±24	-64±19	-72±18	-50±14	-30±16

Table A.1: Free energy differences between beginning and end of simulations ($kcal \cdot mol^{-1}$)

Appendix B

Energetic Analysis of Transthyretin mutations

Results of FoldX Suite

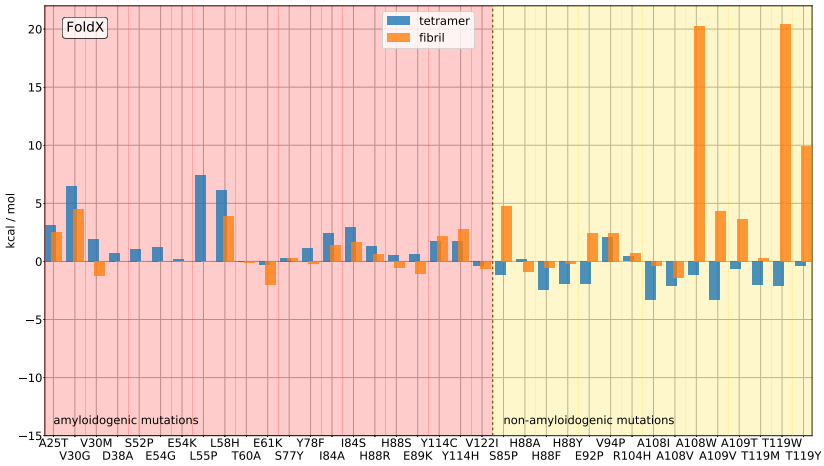


Figure B.1: **Results of FoldX Suite.** Energy contribution (boxes) of single point mutations of transthyretin in tetrameric (blue) and fibril (orange) form using the FoldX program(1). The mutations are indicated on the x-axis. A negative/positive contribution implies a stabilization/destabilization of the structure. All energies are per monomer/fibril layer. Since the sequence of the fibril includes a gap compared to the tetramer, there are no fibril energy values for some variants. A red background indicates that this mutation is known for its increased amyloidogenicity whereas a yellow background means that the mutant inhibits its amyloidosis or stabilizes the tetrameric form.

Total energy per residue of full molecule

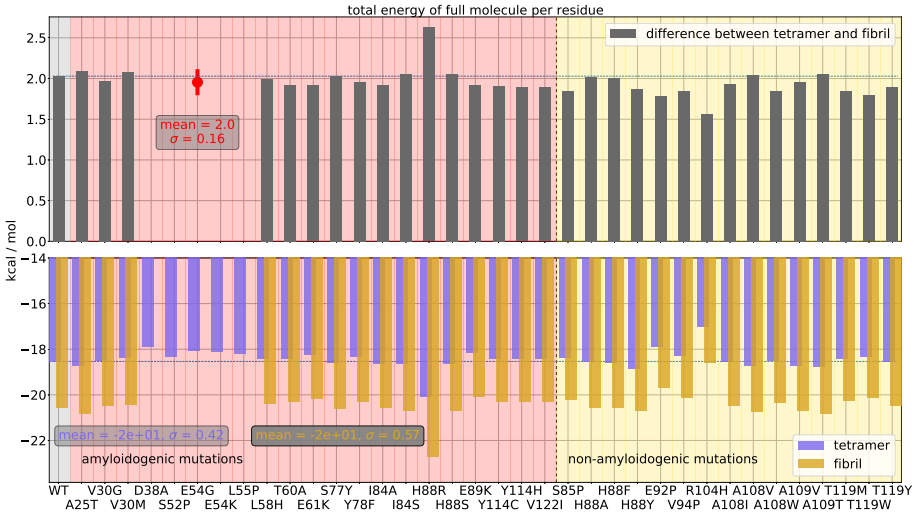


Figure B.2: **Total energy per residue of full molecule.** Calculated mean energy difference per residue between globular and fibril TTR structures. Mean energies were obtained as averages over 10000 trajectory frames. In case of the globular structure the mean energy differences per residue were obtained by dividing the calculated mean total energies by 4·115 (4: number of globular proteins, 115: number of residues). In case of the fibril the total energy was divided by 7·93 (7: number of layers, 93: number of residues resolved in the fibril structure).

Energy contribution of differnt types of interactions

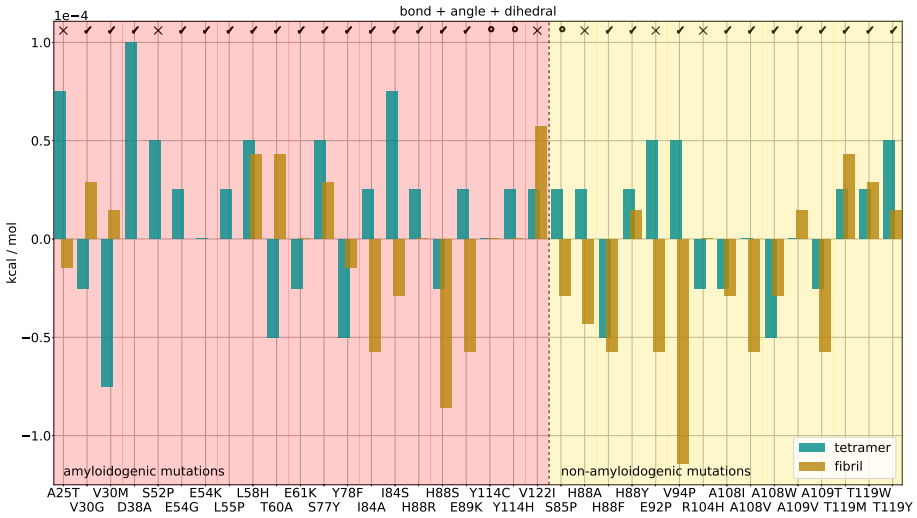


Figure B.3: **Bonded energy contributions** (sum of bond length, bond angle and dihedral angle contributions, boxes) of single point mutations of Transthyretin in tetrameric (blue) and fibril (orange) (MMGBSA). The mutations are indicated on the x-axis. A negative/positive contribution implies a stabilization/destabilization of the structure. All energies are per monomer/layer. Since the sequence of the fibril includes a gap compared to the tetramer, there are no fibril energy values for some variants. A red background indicates that this mutation is known for its increased amyloidogenicity whereas a yellow background means that the mutant inhibits its amyloidosis or stabilizes the tetrameric form.

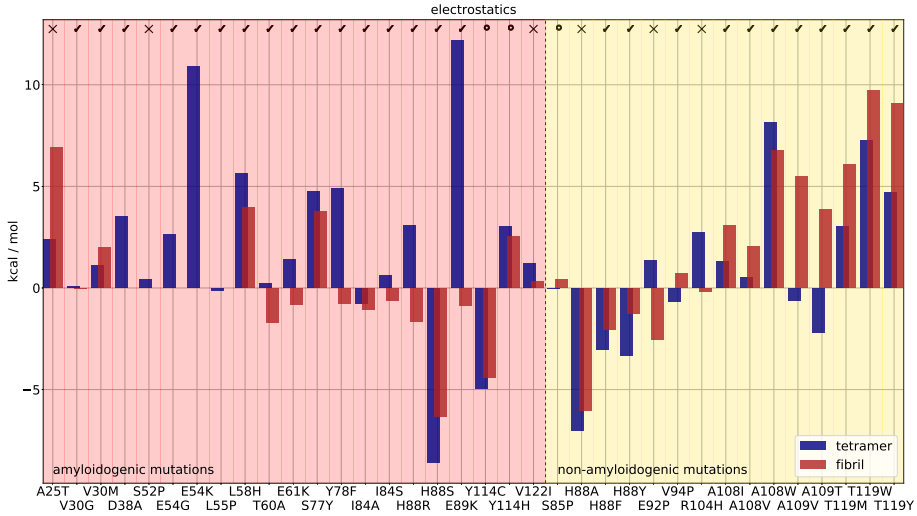


Figure B.4: Same as B.3 but for electrostatic contribution (sum of Coulomb and Generalized Born terms) due to each mutation.

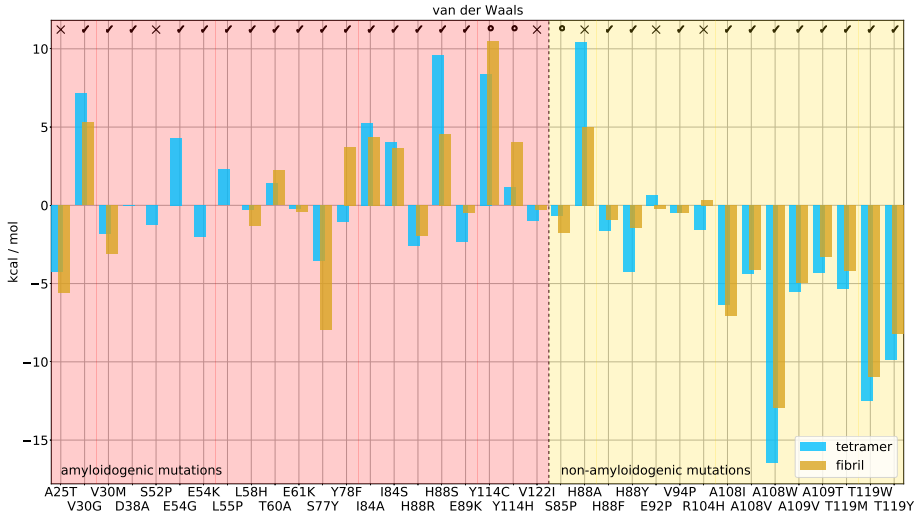


Figure B.5: Same as B.3 but for the van der Waals contribution (Lennard-Jones term) due to each mutation.

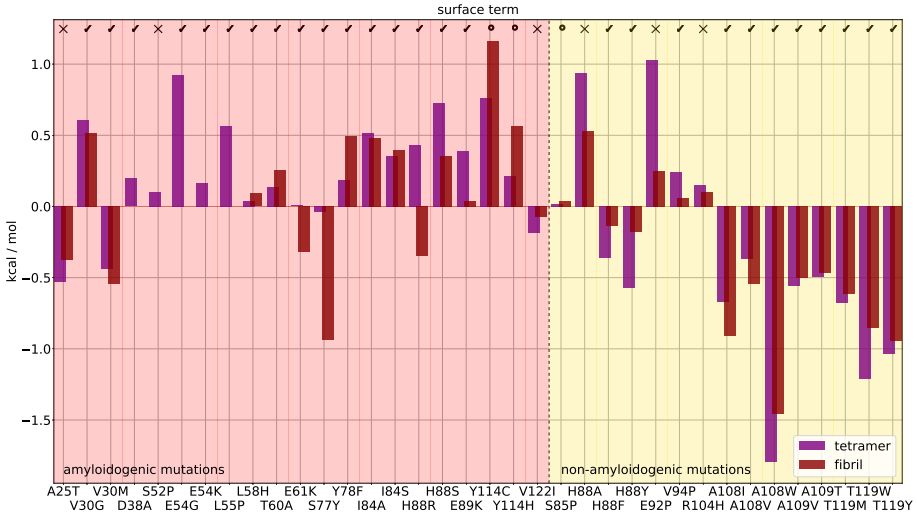


Figure B.6: Same as B.3 but for non-polar surface area dependent solvation contribution due to each mutation.

Comparison of different amounts of frames

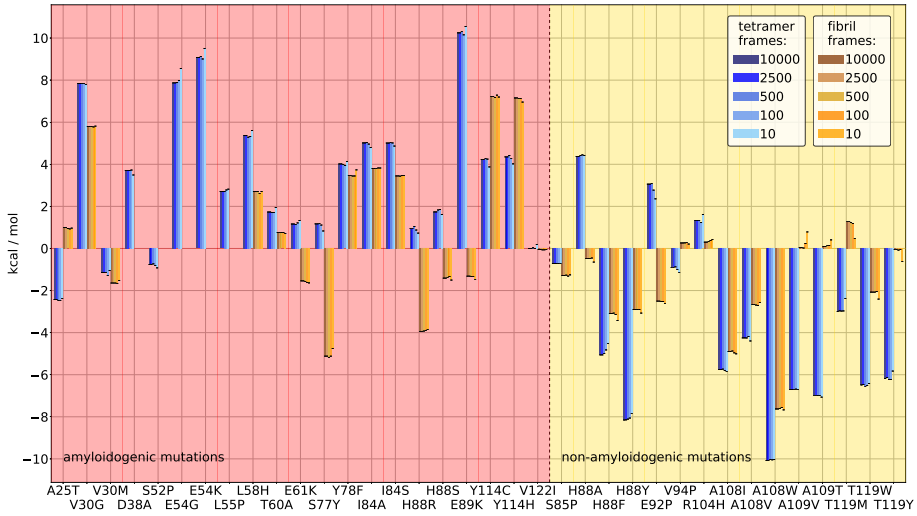


Figure B.7: **Comparison of different numbers of frames used for the MMGBSA calculations.** The number of trajectory frames used for the MMGBSA post-processing calculations was varied between 10 frames (evenly and randomly distributed over the whole data gathering phase) and 10000 frames. The changes in MMGBSA energies are indicates as boxes in different colors (see panel).

Energies for differnt amount of layers

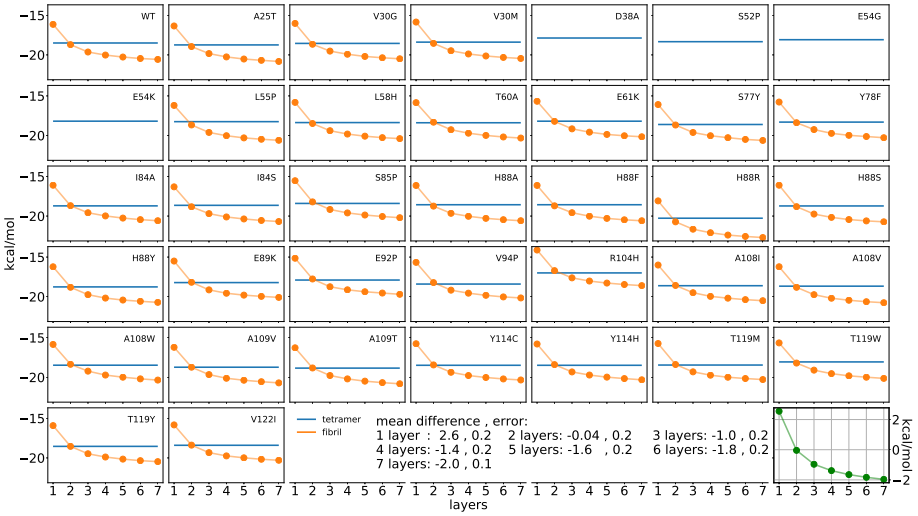


Figure B.8: **Energy per residue for all variants.** The blue line represents the values for the globular protein. The orange graph visualizes how the values develop for a rising number of fibril layers. The green graph in the lower right shows the mean energy difference between globular and fibril protein.

Appendix C

Dynamics of RNA Bulge Loops

Reference WC Hydrogen Bonds

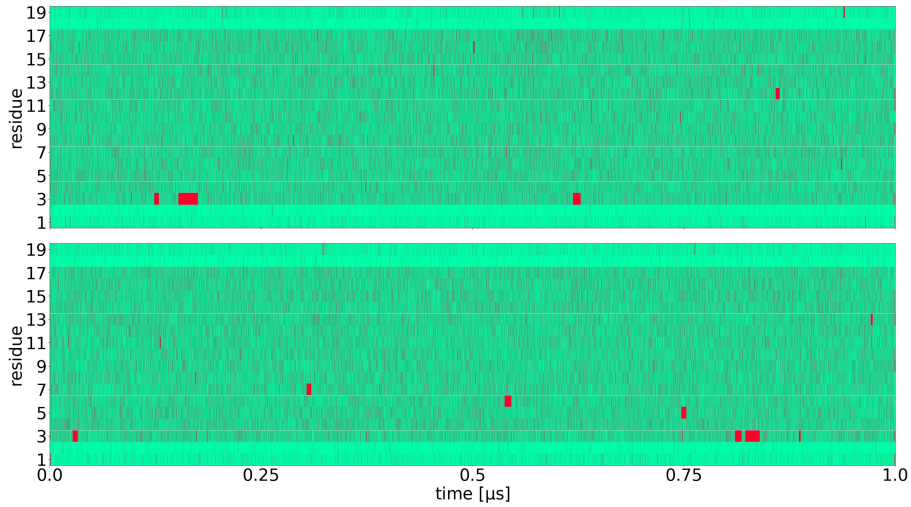


Figure C.1: **Appearance of Watson-Crick-Franklin (WCF) hydrogen bonds in two reference simulations.** The course of every base pair (y-axis) is shown over time (x-axis). Frames with at least one WCF bonds are coloured cyan, frames without any WCF bond are coloured red. During the whole simulation time unpaired base "pairs" occur quite frequently. The two stronger bound C-G pairs at each end of the helix (bottom and top) show less frames without any hydrogen bond. (parameters for hydrogen bonds: distance = 3.0Å, angle = 135°)

Reference Dihedral Angles

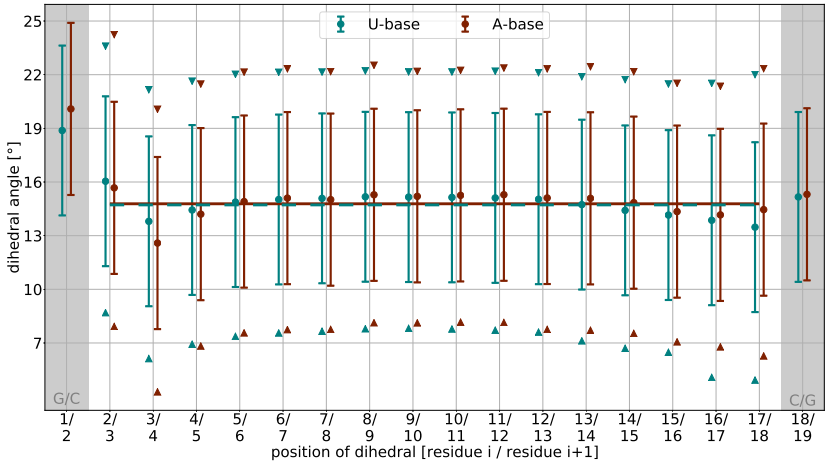


Figure C.2: **Reference dsRNA helices of uracil (U) and adenine (A) bases without bulge.** The mean values (circles) and standard deviation of the dihedral angles between residues are shown. The triangles indicate the interval which contains 90% of the values, which was used as reference. The horizontal line represents the mean value of all U- or A-bases respectively.

Examples of High Mobility

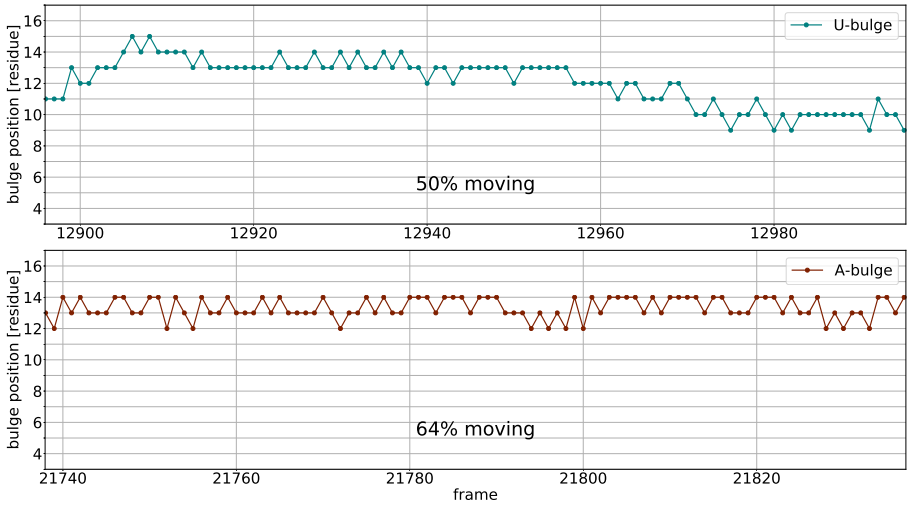
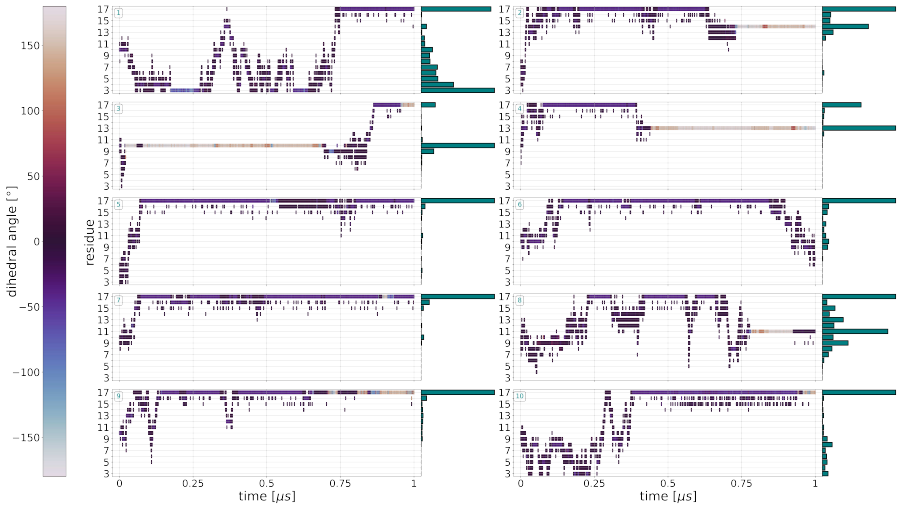
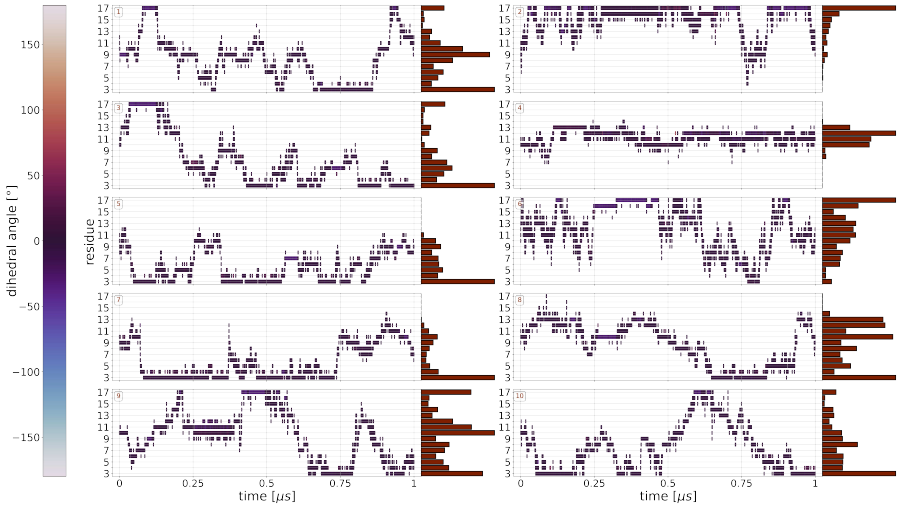


Figure C.3: **Two examples of high mobility of a U-bulge (top) and an A-bulge (bottom).** The position of the bulge within the RNA helix (y-axis) is drawn against the frame (x-axis) of the simulation. The window contains 100 frames. The mobility, meaning time steps while the bulge changes position relative to all 100 steps, is 50% for the U-bulge and 64% for the A-bulge.

Overview U/A-bulge Dihedral Angles



(a) **10 simulations of a U-bulge.**



(b) **10 simulations of an A-bulge.**

Figure C.4: **Bulge location and dihedral angle.** The location of 10 U-bulges (teal) and 10 A-bulges (cedar) within the helix is shown on the y-axis during the simulation time of 1 μs. The two C-G pairs at each end of the helix are marked by grey areas, which the bulge is not expected to pass.

Fluctuations of A/U-bases

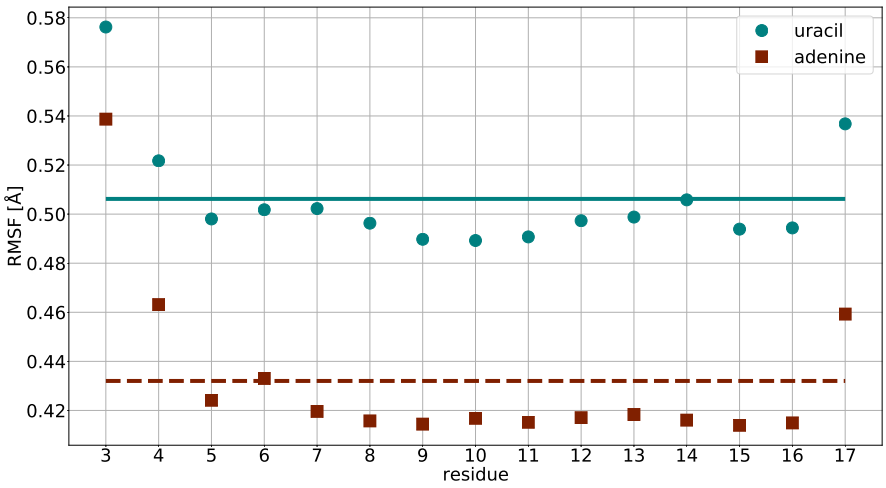


Figure C.5: **Fluctuations within reference dsRNA.** Fluctuations (RMSF) of uracil (U) bases and adenine (A) bases along the reference dsRNA helix (2x 1μs simulation). The markers show the fluctuation of each residue's nucleobase (U: circles, A: squares). Both ends of the helix appear more volatile than the middle part of the helix. The lines illustrate the mean of all residues (U: solid, A: dashed).

Acknowledgements

Ich danke Martin Zacharias für die Einführung in die Welt der Molekulardynamik und die Betreuung während meiner Zeit am Lehrstuhl T38. Seine freundliche und optimistische Einstellung war stets motivierend. Wenn ich sein immer offenes Büro mit einem Problem aufsuchte, verließ ich es meist wieder mit einem vielversprechendem Lösungsansatz oder zumindest schlauer als vorher. Dabei hatte ich immer das Gefühl, dass es ihm Spaß macht, Wissen zu vermitteln.
Mein Dank gilt auch Sonja für ihre gute Organisation rauschender Weihnachtsfeiern, schöner Winterschulen und aller bürokratischen Angelegenheiten.
Für seine Unterstützung gerade während meiner Anfangszeit danke ich Florian. Ich hab mich nicht auf seinen Platz getraut, als er das Büro verließ. Dafür kam Richard, der mir mindestens genauso geholfen hat und Shu-Yu, der sogar mit auf verschneite Berggipfel gekraxelt ist. Allen dreien danke ich für eine immer angenehme Büroatmosphäre.
Ich danke Paul für viele anregende Diskussionen, gute Stimmung und Spaß. Ich wünsche ihm, dass es ihm bald wieder besser geht und er wieder zu alter Stärke findet (die Hüttenwanderung steht noch aus ;).
Meinem SFB-Kollegen Danial danke ich für nette Gesellschaft auf mehreren Konferenzen und Workshops.
Besonderer Dank gilt auch unserem Rechencluster "Cow", ohne den ich immer noch simulieren würde... und seinem kompetenten Administrator Jonathan, der bei Problemen stets zur Stelle war.
Ebenso danke ich Cows großem Bruder SuperMuc vom LRZ und allen Mitarbeitern, die es einem so leicht machen, mit einem Supercomputer zu arbeiten (wenn man vorher schon mit seinem kleinen Bruder und dessen Betreuer üben durfte).
Des Weiteren danke ich allen aktuellen und ehemaligen Mitarbeitern des Lehrstuhls T38 für eine schöne Zeit an der Uni, in den Bergen oder an der Isar.
Ich danke auch der TUM und all ihren Mitarbeitern im Hintergrund, die mir meine Arbeit ermöglicht haben. Der DFG und dem SFB 1035 danke ich für die Finanzierung meiner Arbeit und spannende Retreats.
Zu guter Letzt bedanke ich mich bei meiner Familie, meinem Bruder Laurin, meiner Mutter Eva, meiner Frau Kerrin und meiner Tochter Fenja für ihre immerwährende Unterstützung und dafür, dass sie da sind. Der Mensch, ohne den ich diese Arbeit vermutlich nie begonnen hätte, ist leider nicht mehr dabei, wenn ich sie jetzt beende. Ich danke dir Wolfgang für alles! Du warst für mich nicht nur Vater sondern auch der Physiker, der mein Interesse an den Naturwissenschaften immer befeuert hat.

List of Abbreviations

Ala alanine
ATTR TTR amyloidosis
CPU Central processing unit
cryo-EM cryogenic electron microscopy
DNA Deoxyribonucleic Acid
dsRNA double-stranded Ribonucleic Acid
FAC Familial Amyloid Cardiomyopathy
FAP Familial Amyloid Polyneuropathy
GB Generalized Born
Gly glycine
GPU Graphics processing unit
HSP47 heat shock protein 47
Hyp hydroxyproline
MD Molecular Dynamics
MMGBSA Molecular Mechanics Generalized Born Surface Area
MMPBSA Molecular Mechanics Poisson-Boltzmann Surface Area
mRNA Messenger Ribonucleic Acid
OPC Optimal Point Charge
PBC periodic boundary conditions
PC personal computer
PDB Protein Data Bank (also file format .pdb)
PB Poisson-Boltzmann
pH potential of hydrogen
PME particle mesh Ewald
Pro proline
RMSD root-mean-square deviation
RMSF root-mean-square fluctuation

RNA Ribonucleic Acid

rRNA Ribosomal Ribonucleic Acid

SASA solvent-accessible surface area

SSA Senile Systemic Amyloidosis

ssRNA single-stranded Ribonucleic Acid

Thr threonine

TIP3P Transferable Intermolecular Potential 3-Point

tRNA Transfer Ribonucleic Acid

TTR Transthyretin

vdW van der Waals

WCF Watson-Crick-Franklin

WT wild-type

List of Figures

List of Tables

Bibliography

[1] (). National institute of general medical sciences, National Institute of General Medical Sciences (NIGMS), [Online]. Available: `https://nigms.nih.gov/` (visited on 04/16/2023).

[2] J.-I. Choe and B. Kim, "Determination of proper time step for molecular dynamics simulation," *Bulletin of the Korean Chemical Society*, vol. 21, Apr. 20, 2000.

[3] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters," *Journal of Chemical Physics*, vol. 76, pp. 637–649, Jan. 1, 1982, ADS Bibcode: 1982JChPh..76..637S.

[4] D. Beeman, "Some multistep methods for use in molecular dynamics calculations," *Journal of Computational Physics*, vol. 20, no. 2, pp. 130–139, Feb. 1, 1976.

[5] D. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A.Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R.Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M.Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S.Zhao, and P.A. Kollman, Amber 2021, 2021.

[6] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," *Chemical Reviews*, vol. 106, no. 5, pp. 1589–1615, May 1, 2006, Publisher: American Chemical Society.

[7] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *Journal of Chemical Physics*, vol. 81, pp. 3684–3690, Oct. 1, 1984, ADS Bibcode: 1984JChPh..81.3684B.

[8] D. J. Sindhikara, S. Kim, A. F. Voter, and A. E. Roitberg, "Bad seeds sprout perilous dynamics: Stochastic thermostat induced trajectory synchronization in biomolecules,"
Journal of Chemical Theory and Computation, vol. 5, no. 6,
pp. 1624–1631, Jun. 9, 2009, Publisher: American Chemical Society.

[9] B. P. Uberuaga, M. Anghel, and A. F. Voter, "Synchronization of trajectories in canonical molecular-dynamics simulations: Observation, explanation, and exploitation," *The Journal of Chemical Physics*, vol. 120, no. 14, pp. 6363–6374, Apr. 8, 2004.

[10] H. C. Andersen, "Molecular dynamics simulations at constant pressure and/or temperature," *The Journal of Chemical Physics*, vol. 72, no. 4, pp. 2384–2393, Feb. 15, 1980, Publisher: American Institute of Physics.

[11] n. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Physical Review. A, General Physics*, vol. 31, no. 3, pp. 1695–1697, Mar. 1985.

[12] W. L. Jorgensen, "Quantum and statistical mechanical studies of liquids. 10. transferable intermolecular potential functions for water, alcohols, and ethers. application to liquid water,"
Journal of the American Chemical Society, vol. 103, no. 2, pp. 335–340, Jan. 1, 1981, Publisher: American Chemical Society.

[13] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, p. 926, Aug. 31, 1998, Publisher: American Institute of PhysicsAIP.

[14] S. Izadi, R. Anandakrishnan, and A. V. Onufriev, "Building water models: A different approach,"
The Journal of Physical Chemistry Letters, vol. 5, no. 21, pp. 3863–3871, Nov. 6, 2014, Publisher: American Chemical Society.

[15] D. Frenkel and B. Smit,
"Understanding molecular simulation: From algorithms to applications,"
in, vol. 1, Jan. 1, 2001, p. 664.

[16] M. Tuckerman,
Statistical Mechanics: Theory and Molecular Simulation.
OUP Oxford, Feb. 11, 2010, 719 pp.

[17] J. Kolafa and J. W. Perram, "Cutoff errors in the ewald summation formulae for point charge systems,"
Molecular Simulation, vol. 9, no. 5, pp. 351–368, Jan. 1, 1992, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08927029208049126.

[18] J. J. Benedetto and G. Zimmermann, "Sampling multipliers and the poisson summation formula," *Journal of Fourier Analysis and Applications*, vol. 3, no. 5, pp. 505–523, Sep. 1, 1997.

[19] E. M. Stein and G. Weiss, Introduction to Fourier Analysis on Euclidean Spaces (PMS-32), Volume 32. Princeton University Press, Jun. 2, 2016, 310 pp., Google-Books-ID: xnIwDAAAQBAJ.

[20] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. explicit solvent particle mesh ewald," *Journal of Chemical Theory and Computation*, vol. 9, no. 9, pp. 3878–3888, Sep. 10, 2013, Publisher: American Chemical Society.

[21] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker, "Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. generalized born," *Journal of Chemical Theory and Computation*, vol. 8, no. 5, pp. 1542–1555, May 8, 2012.

[22] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo, "Recent developments and applications of the MMPBSA method," *Frontiers in Molecular Biosciences*, vol. 4, 2018.

[23] A. Onufriev, "Chapter 7 - implicit solvent models in molecular dynamics simulations: A brief overview," in Annual Reports in Computational Chemistry, R. A. Wheeler and D. C. Spellmeyer, Eds., vol. 4, Elsevier, Jan. 1, 2008, pp. 125–137.

[24] T. Schlick, Molecular Modeling and Simulation. Springer, 2002.

[25] A. Shrake and J. A. Rupley, "Environment and exposure to solvent of protein atoms. lysozyme and insulin," *Journal of Molecular Biology*, vol. 79, no. 2, pp. 351–371, Sep. 15, 1973.

[26] N. A. Baker, "Improving implicit solvent simulations: A poisson-centric view," *Current Opinion in Structural Biology*, vol. 15, no. 2, pp. 137–143, Apr. 2005.

[27] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, no. 16, pp. 6127–6129, Aug. 1, 1990, Publisher: American Chemical Society.

[28] H. Kamberaj, Molecular Dynamics Simulations in Statistical Physics: Theory and Applications. Springer, 2020.

[29] J. Srinivasan, M. W. Trevathan, P. Beroza, and D. A. Case, "Application of a pairwise generalized born model to proteins and nucleic acids: Inclusion of salt effects," *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, vol. 101, no. 6, pp. 426–434, May 25, 1999.

[30] C. Chothia, "Principles that determine the structure of proteins," *Annual Review of Biochemistry*, vol. 53, no. 1, pp. 537–572, 1984, _eprint: https://doi.org/10.1146/annurev.bi.53.070184.002541.

[31] A. Kessel and N. Ben-Tal, Introduction to Proteins: Structure, Function, and Motion. CRC Press, Dec. 17, 2010, 623 pp., Google-Books-ID: cMMgypIrxocC.

[32] J. Hartmann and M. Zacharias, "Mechanism of collagen folding propagation studied by molecular dynamics simulations," *PLOS Computational Biology*, vol. 17, no. 6, e1009079, Jun. 8, 2021, Publisher: Public Library of Science.

[33] M. D. Shoulders and R. T. Raines, "Collagen structure and stability," *Annual Review of Biochemistry*, vol. 78, pp. 929–958, 2009.

[34] G. N. Ramachandran and G. Kartha, "Structure of collagen," *Nature*, vol. 174, no. 4423, pp. 269–270, Aug. 7, 1954.

[35] A. Rich and F. H. Crick, "The structure of collagen," *Nature*, vol. 176, no. 4489, pp. 915–916, Nov. 12, 1955.

[36] P. M. Cowan and S. McGAVIN, "Structure of poly-l-proline," *Nature*, vol. 176, no. 4480, pp. 501–503, Sep. 1955, Number: 4480 Publisher: Nature Publishing Group.

[37] K. Beck, V. C. Chan, N. Shenoy, A. Kirkpatrick, J. A. M. Ramshaw, and B. Brodsky, "Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine," *Proceedings of the National Academy of Sciences*, vol. 97, no. 8, pp. 4273–4278, Apr. 11, 2000, Publisher: Proceedings of the National Academy of Sciences.

[38] C. J. Jones, C. Cummings, J. Ball, and P. Beighton, "Collagen defect of bone in osteogenesis imperfecta (type i). an electron microscopic study," *Clinical Orthopaedics and Related Research*, no. 183, pp. 208–214, Mar. 1984.

[39] P. Beighton, A. De Paepe, B. Steinmann, P. Tsipouras, and R. J. Wenstrup, "Ehlers-danlos syndromes: Revised nosology, villefranche, 1997. ehlers-danlos national foundation (USA) and ehlers-danlos support group (UK)," *American Journal of Medical Genetics*, vol. 77, no. 1, pp. 31–37, Apr. 28, 1998.

[40] K. E. Morrison, M. Mariyama, T. L. Yang-Feng, and S. T. Reeders, "Sequence and localization of a partial cDNA encoding the human alpha 3 chain of type IV collagen.," *American Journal of Human Genetics*, vol. 49, no. 3, pp. 545–554, Sep. 1991.

[41] B. G. Hudson, R. Kalluri, S. Gunwar, M. Weber, F. Ballester, J. K. Hudson, M. E. Noelken, M. Sarras, W. R. Richardson, and J. Saus, "The pathogenesis of alport syndrome involves type IV collagen molecules containing the alpha 3(IV) chain: Evidence from anti-GBM nephritis after renal transplantation," *Kidney International*, vol. 42, no. 1, pp. 179–187, Jul. 1992.

[42] A. Buevich and J. Baum, "Nuclear magnetic resonance characterization of peptide models of collagen-folding diseases," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 356, no. 1406, pp. 159–168, Feb. 28, 2001.

[43] A. V. Buevich, T. Silva, B. Brodsky, and J. Baum, "Transformation of the mechanism of triple-helix peptide folding in the absence of a c-terminal nucleation domain and its implications for mutations in collagen disorders *," *Journal of Biological Chemistry*, vol. 279, no. 45, pp. 46 890–46 895, Nov. 5, 2004, Publisher: Elsevier.

[44] A. Bachmann, T. Kiefhaber, S. Boudko, J. Engel, and H. P. Bächinger, "Collagen triple-helix formation in all-trans chains proceeds by a nucleation/growth mechanism with a purely entropic barrier," *Proceedings of the National Academy of Sciences*, vol. 102, no. 39, pp. 13 897–13 902, Sep. 27, 2005, Publisher: Proceedings of the National Academy of Sciences.

[45] S. Boudko, S. Frank, R. A. Kammerer, J. Stetefeld, T. Schulthess, R. Landwehr, A. Lustig, H. P. Bächinger, and J. Engel, "Nucleation and propagation of the collagen triple helix in single-chain and trimerized peptides: Transition from third to first order kinetics," *Journal of Molecular Biology*, vol. 317, no. 3, pp. 459–470, Mar. 29, 2002.

[46] P. H. Byers, E. M. Click, E. Harper, and P. Bornstein, "Interchain disulfide bonds in procollagen are located in a large nontriple-helical COOH-terminal domain.," *Proceedings of the National Academy of Sciences*, vol. 72, no. 8, pp. 3009–3013, Aug. 1975, Publisher: Proceedings of the National Academy of Sciences.

[47] C. Widmer, J. M. Gebauer, E. Brunstein, S. Rosenbaum, F. Zaucke, C. Drögemüller, T. Leeb, and U. Baumann, "Molecular basis for the action of the collagen-specific chaperone hsp47/SERPINH1 and its structure-specific client recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 33, pp. 13 243–13 247, Aug. 14, 2012.

[48] D. L. Bodian, R. J. Radmer, S. Holbert, and T. E. Klein, "Molecular dynamics simulations of the full triple helical region of collagen type i provide an atomic scale view of the protein's regional heterogeneity," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 193–204, 2011.

[49] J. W. Bourne and P. A. Torzilli, "Molecular simulations predict novel collagen conformations during cross-link loading," *Matrix Biology*, vol. 30, no. 5, pp. 356–360, Jun. 1, 2011.

[50] A. Gautieri, S. Vesentini, A. Redaelli, and M. J. Buehler, "Osteogenesis imperfecta mutations lead to local tropocollagen unfolding and disruption of h-bond network," *RSC Advances*, vol. 2, no. 9, pp. 3890–3896, Apr. 10, 2012, Publisher: The Royal Society of Chemistry.

[51] S. Park, T. E. Klein, and V. S. Pande, "Folding and misfolding of the collagen triple helix: Markov analysis of molecular dynamics simulations," *Biophysical Journal*, vol. 93, no. 12, pp. 4108–4115, Dec. 15, 2007.

[52] C. M. Stultz, "The folding mechanism of collagen-like model peptides explored through detailed molecular simulations," *Protein Science : A Publication of the Protein Society*, vol. 15, no. 9, pp. 2166–2177, Sep. 2006.

[53] M. Tang, T. Li, E. Pickering, N. S. Gandhi, K. Burrage, and Y. Gu, "Steered molecular dynamics characterization of the elastic modulus and deformation mechanisms of single natural tropocollagen molecules," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 86, pp. 359–367, Oct. 1, 2018.

[54] K. Okuyama, K. Miyama, K. Mizuno, and H. P. Bächinger, "Crystal structure of (gly-pro-hyp)(9) : Implications for the collagen molecular model," *Biopolymers*, vol. 97, no. 8, pp. 607–616, Aug. 2012.

[55] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1668–1688, Dec. 2005.

[56] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "Ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb," *Journal of Chemical Theory and Computation*, vol. 11, no. 8, pp. 3696–3713, Aug. 11, 2015, Publisher: American Chemical Society.

[57] C. W. Hopkins, S. Le Grand, R. C. Walker, and A. E. Roitberg, "Long-time-step molecular dynamics through hydrogen mass repartitioning," *Journal of Chemical Theory and Computation*, vol. 11, no. 4, pp. 1864–1874, Apr. 14, 2015.

[58] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327–341, Mar. 1, 1977.

[59] B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg, "MMPBSA.py: An efficient program for end-state free energy calculations," *Journal of Chemical Theory and Computation*, vol. 8, no. 9, pp. 3314–3321, Sep. 11, 2012.

[60] S. Ito and K. Nagata, "Biology of hsp47 (serpin h1), a collagen-specific molecular chaperone," *Seminars in Cell & Developmental Biology*, vol. 62, pp. 142–151, Feb. 2017.

[61] ——, "Roles of the endoplasmic reticulum-resident, collagen-specific molecular chaperone hsp47 in vertebrate cells and human disease," *The Journal of Biological Chemistry*, vol. 294, no. 6, pp. 2133–2141, Feb. 8, 2019.

[62] J. Bella, J. Liu, R. Kramer, B. Brodsky, and H. M. Berman, "Conformational effects of gly-x-gly interruptions in the collagen triple helix," *Journal of Molecular Biology*, vol. 362, no. 2, pp. 298–311, Sep. 15, 2006.

[63] J. Bella, M. Eaton, B. Brodsky, and H. M. Berman, "Crystal and molecular structure of a collagen-like peptide at 1.9 a resolution," *Science (New York, N.Y.)*, vol. 266, no. 5182, pp. 75–81, Oct. 7, 1994.

[64] M. Bhate, X. Wang, J. Baum, and B. Brodsky, "Folding and conformational consequences of glycine to alanine replacements at different positions in a collagen model peptide," *Biochemistry*, vol. 41, no. 20, pp. 6539–6547, May 21, 2002.

[65] M. Meli, G. Morra, and G. Colombo, "Simple model of protein energetics to identify ab initio folding transitions from all-atom MD simulations of proteins," *Journal of Chemical Theory and Computation*, vol. 16, no. 9, pp. 5960–5971, Sep. 8, 2020, Publisher: American Chemical Society.

[66] J. Hartmann and M. Zacharias, "Analysis of amyloidogenic transthyretin mutations using continuum solvent free energy calculations," *Proteins: Structure, Function, and Bioinformatics*, vol. n/a, n/a Jul. 16, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.26399.

[67] M. D. Benson and T. Uemichi, "Transthyretin amyloidosis," *Amyloid*, vol. 3, no. 1, pp. 44–56, Jan. 1, 1996, Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/13506129609014354.

[68] M. A. Gertz, M. D. Benson, P. J. Dyck, M. Grogan, T. Coelho, M. Cruz, J. L. Berk, V. Plante-Bordeneuve, H. H. J. Schmidt, and G. Merlini, "Diagnosis, prognosis, and therapy of transthyretin amyloidosis," *Journal of the American College of Cardiology*, vol. 66, no. 21, pp. 2451–2466, Dec. 1, 2015.

[69] L. Saelices, L. M. Johnson, W. Y. Liang, M. R. Sawaya, D. Cascio, P. Ruchala, J. Whitelegge, L. Jiang, R. Riek, and D. S. Eisenberg, "Uncovering the mechanism of aggregation of human transthyretin*," *Journal of Biological Chemistry*, vol. 290, no. 48, pp. 28 932–28 943, Nov. 27, 2015.

[70] N. Schormann, J. R. Murrell, and M. D. Benson, "Tertiary structures of amyloidogenic and non-amyloidogenic transthyretin variants: New model for amyloid fibril formation,"
Amyloid, vol. 5, no. 3, pp. 175–187, Jan. 1, 1998, Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/13506129809003843.

[71] L. Cendron, A. Trovato, F. Seno, C. Folli, B. Alfieri, G. Zanotti, and R. Berni, "Amyloidogenic potential of transthyretin variants: Insights from structural and computational analyses," *The Journal of Biological Chemistry*, vol. 284, no. 38, pp. 25 832–25 841, Sep. 18, 2009.

[72] J. A. Hamilton, L. K. Steinrauf, B. C. Braden, J. Liepnieks, M. D. Benson, G. Holmgren, O. Sandgren, and L. Steen, "The x-ray crystal structure refinements of normal human transthyretin and the amyloidogenic val-30–¿met variant to 1.7-a resolution.," *Journal of Biological Chemistry*, vol. 268, no. 4, pp. 2416–2424, Feb. 5, 1993.

[73] C. ANDRADE, "A PECULIAR FORM OF PERIPHERAL NEUROPATHY: FAMILIAR ATYPICAL GENERALIZED AMYLOIDOSIS WITH SPECIAL INVOLVEMENT OF THE PERIPHERAL NERVES,"
Brain, vol. 75, no. 3, pp. 408–427, Sep. 1, 1952.

[74] H. F. FALLS, J. JACKSON, J. H. CAREY, J. G. RUKAVINA, and W. D. BLOCK, "Ocular manifestations of hereditary primary systemic amyloidosis," *A.M.A. Archives of Ophthalmology*, vol. 54, no. 5, pp. 660–664, Nov. 1, 1955.

[75] D. R. Jacobson, R. D. Pastore, R. Yaghoubian, I. Kane, G. Gallo, F. S. Buck, and J. N. Buxbaum, "Variant-sequence transthyretin (isoleucine 122) in late-onset cardiac amyloidosis in black americans," *The New England Journal of Medicine*, vol. 336, no. 7, pp. 466–473, Feb. 13, 1997.

[76] P. H. Nguyen, A. Ramamoorthy, B. R. Sahoo, J. Zheng, P. Faller, J. E. Straub, L. Dominguez, J.-E. Shea, N. V. Dokholyan, A. De Simone, B. Ma, R. Nussinov, S. Najafi, S. T. Ngo, A. Loquet, M. Chiricotto, P. Ganguly, J. McCarty, M. S. Li, C. Hall, Y. Wang, Y. Miller, S. Melchionna, B. Habenstein, S. Timr, J. Chen, B. Hnath, B. Strodel, R. Kayed, S. Lesné, G. Wei, F. Sterpone, A. J. Doig, and P. Derreumaux, "Amyloid oligomers: A joint experimental/computational perspective on alzheimer's disease, parkinson's disease, type II diabetes, and amyotrophic lateral sclerosis,"
Chemical Reviews, vol. 121, no. 4, pp. 2545–2647, Feb. 24, 2021.

[77] P. Westermark, K. Sletten, B. Johansson, and G. G. Cornwell, "Fibril in senile systemic amyloidosis is derived from normal transthyretin.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 7, pp. 2843–2845, Apr. 1990.

[78] H. J. Cho, J. Y. Yoon, M. H. Bae, J. H. Lee, D. H. Yang, H. S. Park, Y. Cho, S. C. Chae, and J. E. Jun, "Familial transthyretin amyloidosis with variant asp38ala presenting with orthostatic hypotension and chronic diarrhea," *Journal of Cardiovascular Ultrasound*, vol. 20, no. 4, pp. 209–212, Dec. 2012.

[79] V. Planté-Bordeneuve and G. Said, "Familial amyloid polyneuropathy," *The Lancet. Neurology*, vol. 10, no. 12, pp. 1086–1097, Dec. 2011.

[80] L. S. Coles and R. D. Young, "Supercentenarians and transthyretin amyloidosis: The next frontier of human life extension," *Preventive Medicine*, Dietary Nutraceuticals and Age Management Medicine, vol. 54, S9–S11, May 1, 2012.

[81] M. M. Reilly, D. Adams, D. R. Booth, M. B. Davis, G. Said, M. Laubriat-Bianchin, M. B. Pepys, P. K. Thomas, and A. E. Harding, "Transthyretin gene analysis in european patients with suspected familial amyloid polyneuropathy," *Brain*, vol. 118, no. 4, pp. 849–856, Aug. 1, 1995, Publisher: Oxford Academic.

[82] J. D. Gillmore, T. Damy, M. Fontana, M. Hutchinson, H. J. Lachmann, A. Martinez-Naharro, C. C. Quarta, T. Rezk, C. J. Whelan, E. Gonzalez-Lopez, T. Lane, J. A. Gilbertson, D. Rowczenio, A. Petrie, and P. N. Hawkins, "A new staging system for cardiac transthyretin amyloidosis," *European Heart Journal*, vol. 39, no. 30, pp. 2799–2806, Aug. 7, 2018.

[83] A. W. Yee, M. Aldeghi, M. P. Blakeley, A. Ostermann, P. J. Mas, M. Moulin, D. de Sanctis, M. W. Bowler, C. Mueller-Dieckmann, E. P. Mitchell, M. Haertlein, B. L. de Groot, E. Boeri Erba, and V. T. Forsyth, "A molecular mechanism for transthyretin amyloidogenesis," *Nature Communications*, vol. 10, no. 1, p. 925, Dec. 2019.

[84] M. L. Müller, J. Butler, and B. Heidecker, "Emerging therapies in transthyretin amyloidosis – a new wave of hope after years of stagnancy?" *European Journal of Heart Failure*, vol. 22, no. 1, pp. 39–53, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejhf.1695.

[85] C. F. Bennett, "Therapeutic antisense oligonucleotides are coming of age," *Annual Review of Medicine*, vol. 70, pp. 307–321, Jan. 27, 2019.

[86] C. J. Niemietz, V. Sauer, J. Stella, L. Fleischhauer, G. Chandhok, S. Guttmann, Y. Avsar, S. Guo, E. J. Ackermann, J. Gollob, B. P. Monia, A. Zibert, and H. H.-J. Schmidt, "Evaluation of therapeutic oligonucleotides for familial amyloid polyneuropathy in patient-derived hepatocyte-like cells," *PloS One*, vol. 11, no. 9, e0161455, 2016.

[87] T. Coelho, L. F. Maia, A. Martins da Silva, M. Waddington Cruz, V. Planté-Bordeneuve, P. Lozeron, O. B. Suhr, J. M. Campistol, I. M. Conceição, H. H.-J. Schmidt, P. Trigo, J. W. Kelly, R. Labaudinière, J. Chan, J. Packman, A. Wilson, and D. R. Grogan, "Tafamidis for transthyretin familial amyloid polyneuropathy: A randomized, controlled trial," *Neurology*, vol. 79, no. 8, pp. 785–792, Aug. 21, 2012.

[88] G. Merlini, V. Planté-Bordeneuve, D. P. Judge, H. Schmidt, L. Obici, S. Perlini, J. Packman, T. Tripp, and D. R. Grogan, "Effects of tafamidis on transthyretin stabilization and clinical outcomes in patients with non-val30met transthyretin amyloidosis," *Journal of Cardiovascular Translational Research*, vol. 6, no. 6, pp. 1011–1020, Dec. 2013.

[89] J. L. Berk, O. B. Suhr, L. Obici, Y. Sekijima, S. R. Zeldenrust, T. Yamashita, M. A. Heneghan, P. D. Gorevic, W. J. Litchy, J. F. Wiesman, E. Nordh, M. Corato, A. Lozza, A. Cortese, J. Robinson-Papp, T. Colton, D. V. Rybin, A. B. Bisbee, Y. Ando, S.-i. Ikeda, D. C. Seldin, G. Merlini, M. Skinner, J. W. Kelly, P. J. Dyck, and Diflunisal Trial Consortium, "Repurposing diflunisal for familial amyloid polyneuropathy: A randomized clinical trial," *JAMA*, vol. 310, no. 24, pp. 2658–2667, Dec. 25, 2013.

[90] Y. Sekijima, M. A. Dendle, and J. W. Kelly, "Orally administered diflunisal stabilizes transthyretin against dissociation required for amyloidogenesis," *Amyloid: The International Journal of Experimental and Clinical Investigation: The Official Journal of the International Society of Amyloidosis*, vol. 13, no. 4, pp. 236–249, Dec. 2006.

[91] K. Bodin, S. Ellmerich, M. C. Kahan, G. A. Tennent, A. Loesch, J. A. Gilbertson, W. L. Hutchinson, P. P. Mangione, J. R. Gallimore, D. J. Millar, S. Minogue, A. P. Dhillon, G. W. Taylor, A. R. Bradwell, A. Petrie, J. D. Gillmore, V. Bellotti, M. Botto, P. N. Hawkins, and M. B. Pepys, "Antibodies to human serum amyloid p component eliminate visceral amyloid deposits," *Nature*, vol. 468, no. 7320, pp. 93–97, Nov. 4, 2010.

[92] Y. Su, H. Jono, M. Torikai, A. Hosoi, K. Soejima, J. Guo, M. Tasaki, Y. Misumi, M. Ueda, S. Shinriki, M. Shono, K. Obayashi, T. Nakashima, K. Sugawara, and Y. Ando, "Antibody therapy for familial amyloidotic polyneuropathy," *Amyloid: The International Journal of Experimental and Clinical Investigation: The Official Journal of the International Society of Amyloidosis*, vol. 19 Suppl 1, pp. 45–46, Jun. 2012.

[93] M. Phay, V. Blinder, S. Macy, M. J. Greene, D. C. Wooliver, W. Liu, A. Planas, D. M. Walsh, L. H. Connors, S. R. Primmer, S. A. Planque, S. Paul, and B. O'Nuallain, "Transthyretin aggregate-specific antibodies recognize cryptic epitopes on patient-derived amyloid fibrils," *Rejuvenation Research*, vol. 17, no. 2, pp. 97–104, Apr. 2014.

[94] S. A. Planque, Y. Nishiyama, M. Hara, S. Sonoda, S. K. Murphy, K. Watanabe, Y. Mitsuda, E. L. Brown, R. J. Massey, S. R. Primmer, B. O'Nuallain, and S. Paul, "Physiological IgM class catalytic antibodies selective for transthyretin amyloid," *The Journal of Biological Chemistry*, vol. 289, no. 19, pp. 13 243–13 258, May 9, 2014.

[95] A. M. Damas, S. Ribeiro, V. S. Lamzin, J. A. Palha, and M. J. Saraiva, "Structure of the val122ile variant transthyretin - a cardiomyopathic mutant," *Acta Crystallographica. Section D, Biological Crystallography*, vol. 52, pp. 966–972, Pt 5 Sep. 1, 1996.

[96] G. Zanotti, C. Folli, L. Cendron, B. Alfieri, S. K. Nishida, F. Gliubich, N. Pasquato, A. Negro, and R. Berni, "Structural and mutational analyses of protein-protein interactions between transthyretin and retinol-binding protein,"
The FEBS journal, vol. 275, no. 23, pp. 5841–5854, Dec. 2008.

[97] J. A. Hamilton, L. K. Steinrauf, B. C. Braden, J. R. Murrell, and M. D. Benson, "Structural changes in transthyretin produced by the ile 84 ser mutation which result in decreased affinity for retinol-binding protein," *Amyloid*, vol. 3, no. 1, pp. 1–12, Jan. 1, 1996, Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/13506129609014349.

[98] D. E. Jenne, K. Denzel, P. Blätzinger, P. Winter, B. Obermaier, R. P. Linke, and K. Altland, "A new isoleucine substitution of val-20 in transthyretin tetramers selectively impairs dimer-dimer contacts and causes systemic amyloidosis," *Proceedings of the National Academy of Sciences*, vol. 93, no. 13, pp. 6302–6307, Jun. 25, 1996, Publisher: National Academy of Sciences Section: Research Article.

[99] Y. Sekijima, R. I. Campos, P. Hammarström, K. P. R. Nilsson, T. Yoshinaga, K. Nagamatsu, M. Yazaki, F. Kametani, and S.-i. Ikeda, "Pathological, biochemical, and biophysical characteristics of the transthyretin variant y114h (p.y134h) explain its very mild clinical phenotype," *Journal of the peripheral nervous system: JPNS*, vol. 20, no. 4, pp. 372–379, Dec. 2015.

[100] M. Schmidt, S. Wiese, V. Adak, J. Engler, S. Agarwal, G. Fritz, P. Westermark, M. Zacharias, and M. Fändrich, "Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis,"
Nature Communications, vol. 10, no. 1, p. 5008, Nov. 1, 2019.

[101] T. Siebenmorgen and M. Zacharias, "Computational prediction of protein–protein binding affinities,"
WIREs Computational Molecular Science, vol. 10, no. 3, e1448, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1448.

[102] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The FoldX web server: An online force field.,"
Nucleic Acids Res, vol. 33, Web Server issue 2005.

[103] D. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A.Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R.Harris, S. Izadi, S.A. Izmailov, C. Jin, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M.Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xue, D.M. York, S.Zhao, and P.A. Kollman, Amber 2021, 2018.

[104] S. Miyamoto and P. A. Kollman, "Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models," *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 952–962, 1992, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.540130805.

[105] H. Nguyen, D. R. Roe, and C. Simmerling, "Improved generalized born solvent model parameters for protein simulations," *Journal of Chemical Theory and Computation*, vol. 9, no. 4, pp. 2020–2034, Apr. 9, 2013, Publisher: American Chemical Society.

[106] R. B. Petersen, H. Goren, M. Cohen, S. L. Richardson, N. Tresser, A. Lynn, M. Gali, M. Estes, and P. Gambetti, "Transthyretin amyloidosis: A new mutation associated with dementia," *Annals of Neurology*, vol. 41, no. 3, pp. 307–313, 1997, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.410410305.

[107] M. Miyata, T. Sato, M. Mizuguchi, T. Nakamura, S. Ikemizu, Y. Nabeshima, S. Susuki, Y. Suwa, H. Morioka, Y. Ando, M. A. Suico, T. Shuto, T. Koga, Y. Yamagata, and H. Kai, "Role of the glutamic acid 54 residue in transthyretin stability and thyroxine binding," *Biochemistry*, vol. 49, no. 1, pp. 114–123, Jan. 12, 2010, Publisher: American Chemical Society.

[108] D. R. Jacobson, D. E. McFarlin, I. Kane, and J. N. Buxbaum, "Transthyretin pro55, a variant associated with early-onset, aggressive, diffuse amyloidosis with cardiac and neurologic involvement," *Human Genetics*, vol. 89, no. 3, pp. 353–356, May 1992.

[109] M. J. Saraiva, M. R. Almeida, I. L. Alves, M. J. Bonifácio, A. M. Damas, J. A. Palha, G. Goldsteins, and E. Lundgren, "Modulating conformational factors in transthyretin amyloid," *Ciba Foundation Symposium*, vol. 199, 47–52, discussion 52–57, 1996.

[110] E. P. C. Azevedo, H. M. Pereira, R. C. Garratt, J. W. Kelly, D. Foguel, and F. L. Palhano, "Dissecting the structure, thermodynamic stability, and aggregation properties of the a25t transthyretin (a25t-TTR) variant involved in leptomeningeal amyloidosis: Identifying protein partners that co-aggregate during a25t-TTR fibrillogenesis in cerebrospinal fluid,"

Biochemistry, vol. 50, no. 51, pp. 11 070–11 083, Dec. 27, 2011, Publisher: American Chemical Society.

[111] A. F. Castro-Rodrigues, L. Gales, M. J. Saraiva, and A. M. Damas, "Structural insights into a zinc-dependent pathway leading to leu55pro transthyretin amyloid fibrils," *Acta Crystallographica. Section D, Biological Crystallography*, vol. 67, pp. 1035–1044, Pt 12 Dec. 2011.

[112] W. C. Nichols, J. J. Liepnieks, V. A. McKusick, and M. D. Benson, "Direct sequencing of the gene for maryland/german familial amyloidotic polyneuropathy type II and genotyping by allele-specific enzymatic amplification," *Genomics*, vol. 5, no. 3, pp. 535–540, Oct. 1989.

[113] M. R. Wallace, F. E. Dwulet, P. M. Conneally, and M. D. Benson, "Biochemical and molecular genetic characterization of a new variant prealbumin associated with hereditary amyloidosis," *The Journal of Clinical Investigation*, vol. 78, no. 1, pp. 6–12, Jul. 1986.

[114] T. Murakami, T. Yokoyama, M. Mizuguchi, S. Toné, S. Takaku, K. Sango, H. Nishimura, K. Watabe, and Y. Sunada, "A low amyloidogenic e61k transthyretin mutation may cause familial amyloid polyneuropathy," *Journal of Neurochemistry*, vol. 156, no. 6, pp. 957–966, 2021, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.15162.

[115] M. R. Wallace, F. E. Dwulet, E. C. Williams, P. M. Conneally, and M. D. Benson, "Identification of a new hereditary amyloidosis prealbumin variant, tyr-77, and detection of the gene by DNA analysis," *The Journal of Clinical Investigation*, vol. 81, no. 1, pp. 189–193, Jan. 1988.

[116] H. Terazaki, Y. Ando, R. Fernandes, K.-i. Yamamura, S. Maeda, and M. J. Saraiva, "Immunization in familial amyloidotic polyneuropathy: Counteracting deposition by immunization with a y78f TTR mutant," *Laboratory Investigation; a Journal of Technical Methods and Pathology*, vol. 86, no. 1, pp. 23–31, Jan. 2006.

[117] N. Pasquato, R. Berni, C. Folli, B. Alfieri, L. Cendron, and G. Zanotti, "Acidic pH-induced conformational changes in amyloidogenic mutant transthyretin," *Journal of Molecular Biology*, vol. 366, no. 3, pp. 711–719, Feb. 23, 2007.

[118] G. Zanotti, L. Cendron, C. Folli, P. Florio, B. P. Imbimbo, and R. Berni, "Structural evidence for native state stabilization of a conformationally labile amyloidogenic transthyretin variant by fibrillogenesis inhibitors," *FEBS letters*, vol. 587, no. 15, pp. 2325–2331, Aug. 2, 2013.

[119] T. Yokoyama, Y. Hanawa, T. Obita, and M. Mizuguchi, "Stability and crystal structures of his88 mutant human transthyretins," *FEBS letters*, vol. 591, no. 13, pp. 1862–1871, Jul. 2017.

[120] M. Nakamura, K. H. Asl, and M. D. Benson, "A novel variant of transthyretin (glu89lys) associated with familial amyloidotic polyneuropathy,"
Amyloid, vol. 7, no. 1, pp. 46–50, Jan. 1, 2000, Publisher: Taylor & Francis _eprint: https://doi.org/10.3109/13506120009146824.

[121] S. Ueno, T. Uemichi, S. Yorifuji, and S. Tarui, "A novel variant of transthyretin (tyr114 to cys) deduced from the nucleotide sequences of gene fragments from familial amyloidotic polyneuropathy in japanese sibling cases," *Biochemical and Biophysical Research Communications*, vol. 169, no. 1, pp. 143–147, May 31, 1990.

[122] S. Ueno, H. Fujimura, S. Yorifuji, Y. Nakamura, M. Takahashi, S. Tarui, and T. Yanagihara, "Familial amyloid polyneuropathy associated with the transthyretin cys114 gene in a japanese kindred," *Brain: A Journal of Neurology*, vol. 115 (Pt 5), pp. 1275–1289, Oct. 1992.

[123] E. Haagsma, J. Post, A. DeJager, P. Nikkels, B. Hamel, and B. Hazenberg, "A dutch kindred with familial amyloidotic polyneuropathy associated with the transthyretin cys 114 mutant," *Amyloid*, vol. 4, no. 2, pp. 112–117, Jun. 1997.

[124] T. Eneqvist, A. Olofsson, Y. Ando, T. Miyakawa, S. Katsuragi, J. Jass, E. Lundgren, and A. E. Sauer-Eriksson, "Disulfide-bond formation in the transthyretin mutant y114c prevents amyloid fibril formation in vivo and in vitro," *Biochemistry*, vol. 41, no. 44, pp. 13 143–13 151, Nov. 5, 2002.

[125] A. Karlsson, A. Olofsson, T. Eneqvist, and A. E. Sauer-Eriksson, "Cys114-linked dimers of transthyretin are compatible with amyloid formation," *Biochemistry*, vol. 44, no. 39, pp. 13 063–13 070, Oct. 4, 2005.

[126] R. Sant'Anna, M. R. Almeida, N. Varejão, P. Gallego, S. Esperante, P. Ferreira, A. Pereira-Henriques, F. L. Palhano, M. de Carvalho, D. Foguel, D. Reverter, M. J. Saraiva, and S. Ventura, "Cavity filling mutations at the thyroxine-binding site dramatically increase transthyretin stability and prevent its aggregation,"
Scientific Reports, vol. 7, p. 44 709, Mar. 24, 2017.

[127] G. Y. Park, A. Jamerlan, K. H. Shim, and S. S. A. An, "Diagnostic and treatment approaches involving transthyretin in amyloidogenic diseases," *International Journal of Molecular Sciences*, vol. 20, no. 12, p. 2982, Jan. 2019, Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[128] L. K. Steinrauf, J. A. Hamilton, B. C. Braden, J. R. Murrell, and M. D. Benson, "X-ray crystal structure of the ala-109–¿thr variant of human transthyretin which produces euthyroid hyperthyroxinemia," *The Journal of Biological Chemistry*, vol. 268, no. 4, pp. 2425–2430, Feb. 5, 1993.

[129] T. Yokoyama, M. Mizuguchi, Y. Nabeshima, K. Kusaka, T. Yamada, T. Hosoya, T. Ohhara, K. Kurihara, K. Tomoyori, I. Tanaka, and N. Niimura, "Hydrogen-bond network and pH sensitivity in transthyretin: Neutron crystal structure of human transthyretin," *Journal of Structural Biology*, vol. 177, no. 2, pp. 283–290, Feb. 1, 2012.

[130] G. Holmgren, U. Hellman, H.-E. Lundgren, O. Sandgren, and O. B. Suhr, "Impact of homozygosity for an amyloidogenic transthyretin mutation on phenotype and long term outcome," *Journal of Medical Genetics*, vol. 42, no. 12, pp. 953–956, Dec. 2005.

[131] S. C. Penchala, S. Connelly, Y. Wang, M. S. Park, L. Zhao, A. Baranczak, I. Rappley, H. Vogel, M. Liedtke, R. M. Witteles, E. T. Powers, N. Reixach, W. K. Chan, I. A. Wilson, J. W. Kelly, I. A. Graef, and M. M. Alhamadsheh, "AG10 inhibits amyloidogenesis and cellular toxicity of the familial amyloid cardiomyopathy-associated v122i transthyretin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 24, pp. 9992–9997, Jun. 11, 2013.

[132] A. J. Stangou, P. N. Hawkins, N. D. Heaton, M. Rela, M. Monaghan, P. Nihoyannopoulos, J. O'Grady, M. B. Pepys, and R. Williams, "PROGRESSIVE CARDIAC AMYLOIDOSIS FOLLOWING LIVER TRANSPLANTATION FOR FAMILIAL AMYLOID POLYNEUROPATHY: Implications for amyloid fibrillogenesis," *Transplantation*, vol. 66, no. 2, pp. 229–233, Jul. 27, 1998.

[133] P. P. Mangione, R. Porcari, J. D. Gillmore, P. Pucci, M. Monti, M. Porcari, S. Giorgetti, L. Marchese, S. Raimondi, L. C. Serpell, W. Chen, A. Relini, J. Marcoux, I. R. Clatworthy, G. W. Taylor, G. A. Tennent, C. V. Robinson, P. N. Hawkins, M. Stoppini, S. P. Wood, M. B. Pepys, and V. Bellotti, "Proteolytic cleavage of ser52pro variant transthyretin triggers its amyloid fibrillogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 4, pp. 1539–1544, Jan. 28, 2014.

[134] F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," *Annual Review of Biochemistry*, vol. 75, pp. 333–366, 2006.

[135] M. Kollmer, W. Close, L. Funk, J. Rasmussen, A. Bsoul, A. Schierhorn, M. Schmidt, C. J. Sigurdson, M. Jucker, and M. Fändrich, "Cryo-EM structure and polymorphism of aβ amyloid fibrils purified from alzheimer's brain tissue," *Nature Communications*, vol. 10, no. 1, p. 4760, Oct. 29, 2019.

[136] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 8, 1970.

[137] K. Kruger, P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech, "Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena,"
Cell, vol. 31, no. 1, pp. 147–157, Nov. 1982.

[138] J. Šponer, G. Bussi, M. Krepl, P. Banáš, S. Bottaro, R. A. Cunha, A. Gil-Ley, G. Pinamonti, S. Poblete, P. Jurečka, N. G. Walter, and M. Otyepka, "RNA structural dynamics as captured by molecular simulations: A comprehensive overview,"
Chemical Reviews, vol. 118, no. 8, pp. 4177–4338, Apr. 25, 2018.

[139] J. M. Ogle, A. P. Carter, and V. Ramakrishnan, "Insights into the decoding mechanism from recent ribosome structures,"
Trends in Biochemical Sciences, vol. 28, no. 5, pp. 259–266, May 2003.

[140] T. R. Cech, "Structural biology. the ribosome is a ribozyme,"
Science (New York, N.Y.), vol. 289, no. 5481, pp. 878–879, Aug. 11, 2000.

[141] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 a resolution,"
Science (New York, N.Y.), vol. 289, no. 5481, pp. 905–920, Aug. 11, 2000.

[142] P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz, "The structural basis of ribosome activity in peptide bond synthesis,"
Science (New York, N.Y.), vol. 289, no. 5481, pp. 920–930, Aug. 11, 2000.

[143] G. W. Muth, L. Ortoleva-Donnelly, and S. A. Strobel, "A single adenosine with a neutral pKa in the ribosomal peptidyl transferase center,"
Science (New York, N.Y.), vol. 289, no. 5481, pp. 947–950, Aug. 11, 2000.

[144] M. P. Robertson and G. F. Joyce, "The origins of the RNA world,"
Cold Spring Harbor Perspectives in Biology, vol. 4, no. 5, a003608, May 1, 2012, Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[145] L. E. Orgel, "Evolution of the genetic apparatus,"
Journal of Molecular Biology, vol. 38, no. 3, pp. 381–393, Dec. 1968.

[146] F. H. Crick, "The origin of the genetic code,"
Journal of Molecular Biology, vol. 38, no. 3, pp. 367–379, Dec. 1968.

[147] C. R. Woese, D. H. Dugre, W. C. Saxinger, and S. A. Dugre, "The molecular basis for the genetic code.,"
Proceedings of the National Academy of Sciences of the United States of America, vol. 55, no. 4, pp. 966–974, Apr. 1966.

[148] N. B. Leontis and E. Westhof, "Geometric nomenclature and classification of RNA base pairs,"
RNA (New York, N.Y.), vol. 7, no. 4, pp. 499–512, Apr. 2001.

[149] N. B. Leontis, J. Stombaugh, and E. Westhof, "The non-watson-crick base pairs and their associated isostericity matrices," *Nucleic Acids Research*, vol. 30, no. 16, pp. 3497–3531, Aug. 15, 2002.

[150] J. Stombaugh, C. L. Zirbel, E. Westhof, and N. B. Leontis, "Frequency and isostericity of RNA base pairs," *Nucleic Acids Research*, vol. 37, no. 7, pp. 2294–2312, Apr. 2009.

[151] R. T. Batey, R. P. Rambo, and J. A. Doudna, "Tertiary motifs in RNA structure and folding," *Angewandte Chemie International Edition*, vol. 38, no. 16, pp. 2326–2343, 1999, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-3773%2819990816%2938%3A16%3C2326%3A%3AAID-ANIE2326%3E3.0.CO%3B2-3.

[152] J. D. Watson and F. H. C. Crick, "Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid," *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953, Number: 4356 Publisher: Nature Publishing Group.

[153] D. H. Turner, N. Sugimoto, and S. M. Freier, "RNA structure prediction," *Annual Review of Biophysics and Biophysical Chemistry*, vol. 17, pp. 167–192, 1988.

[154] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner, "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure," *Journal of Molecular Biology*, vol. 288, no. 5, pp. 911–940, May 21, 1999.

[155] D. H. Mathews and D. H. Turner, "Prediction of RNA secondary structure by free energy minimization," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 270–278, Jun. 2006.

[156] C. L. Zirbel, J. E. Šponer, J. Šponer, J. Stombaugh, and N. B. Leontis, "Classification and energetics of the base-phosphate interactions in RNA," *Nucleic Acids Research*, vol. 37, no. 15, pp. 4898–4918, Aug. 2009.

[157] B. A. Sweeney, P. Roy, and N. B. Leontis, "An introduction to recurrent nucleotide interactions in RNA," *Wiley interdisciplinary reviews. RNA*, vol. 6, no. 1, pp. 17–45, Feb. 2015.

[158] K. M. Weeks and D. M. Crothers, "Major groove accessibility of RNA," *Science (New York, N.Y.)*, vol. 261, no. 5128, pp. 1574–1577, Sep. 17, 1993.

[159] M. Zacharias, "Simulation of the structure and dynamics of nonhelical RNA motifs," *Current Opinion in Structural Biology*, vol. 10, no. 3, pp. 311–317, Jun. 2000.

[160] T. J. Macke and D. A. Case, "Modeling unusual nucleic acid structures," American Chemical Society, 1998.

[161] S. Arnott, D. W. Hukins, S. D. Dover, W. Fuller, and A. R. Hodgson, "Structures of synthetic polynucleotides in the a-RNA and a'-RNA conformations: X-ray diffraction analyses of the molecular conformations of polyadenylic acid–polyuridylic acid and polyinosinic acid–polycytidylic acid,"
Journal of Molecular Biology, vol. 81, no. 2, pp. 107–122, Dec. 5, 1973.

[162] A. Pérez, I. Marchán, D. Svozil, J. Sponer, T. E. Cheatham, C. A. Laughton, and M. Orozco, "Refinement of the AMBER force field for nucleic acids: Improving the description of alpha/gamma conformers," *Biophysical Journal*, vol. 92, no. 11, pp. 3817–3829, Jun. 1, 2007.

[163] L. Wickstrom, A. Okur, and C. Simmerling, "Evaluating the performance of the ff99sb force field based on NMR scalar coupling data," *Biophysical Journal*, vol. 97, no. 3, pp. 853–856, Aug. 5, 2009.

[164] M. Zgarbová, M. Otyepka, J. Šponer, A. Mládek, P. Banáš, T. E. I. Cheatham, and P. Jurečka, "Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles,"
Journal of Chemical Theory and Computation, vol. 7, no. 9, pp. 2886–2902, Sep. 13, 2011, Publisher: American Chemical Society.

[165] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix,"
Nucleic Acids Research, vol. 34, no. 2, pp. 564–574, Jan. 1, 2006.

[166] R. F. Brown, C. T. Andrews, and A. H. Elcock, "Stacking free energies of all DNA and RNA nucleoside pairs and dinucleoside-monophosphates computed using recently revised AMBER parameters and compared with experiment," *Journal of Chemical Theory and Computation*, vol. 11, no. 5, pp. 2315–2328, May 12, 2015.