UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

FOLDING PROTEINS IN-SILCO

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ALAN J. RAY

Norman, Oklahoma

2024

FOLDING PROTEINS IN-SILCO

A DISSERTATION APPROVED FOR DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY THE COMMITTEE CONSISTING OF

Dr. U.H.E. Hansmann, Chair

Dr. Kieran Mullen

Dr. Yitong Dong

Dr. Bayram Saparov

© Copyright by Alan J. Ray 2024

All Rights Reserved.

Contents

List of Figures

List of Tables

Abstract

Determination of how proteins transform into their biologically relevant structure is an extraordinarily complex research area. They fold into their native state as a population of conformations on time scales that experimental methods struggle to resolve. Molecular dynamic simulations can avoid some of these issues, providing theoretical answers to protein dynamics. To that effect, this work presents two enhanced sampling methods that aim to lower the computational cost of the Replica Exchange protocol with fewer replicas while avoiding the exchange bottleneck in an approach called Replica-Exchange-with-Tunneling. The method is used to simulate the folding switch of the metamorphic chemokine Lymphotactin. *Go*-model potentials bias replicas to fold as either the Ltn10 form or Ltn40 form. This study proposes that the conversion between the two forms is assisted by bifurcated hydrogen bonding. Resolution-Exchange-with-Tunneling (ResET) builds further on these advancements by reducing the number of replicas to the minimum of 2, while still avoiding exchange rejection and enhancing sampling. The folding performance of the method is shown to outperform other simulation approaches for folding the trpcage protein. It was also used to elicit the contribution that amino acid mutations have on the behavior of Alzheimer associated amyloid beta proteins.

Chapter 1

Concerning Proteins

1.1 Introduction

This dissertation is largely concerned with proteins, and from within its pages, the reader may discover how these molecular machines facilitate the biological processes necessary for life.

The instructions for creating a protein are encoded by Deoxyribonucleic acid (DNA). A DNA sequence is translated to produce an unique polypeptide chain, made via permutations of 20 amino acid compounds.[1] These unique chains may constitute what a protein is, but not what function the protein performs. Their biological role is a factor of their geometric shape, which can not be ascertained solely from the amino acid sequence. In a post-translation process known as protein folding the polypeptide chain undergoes conformation changes, transforming into its biologically significant structure, or native state.[2,3]

Though individual members of the amino acid chain, called residues, influence this structure, it is primarily the impact of the surrounding solvent environment that drives the folding process.[4] In fact, challenges to the Anfinsen's dogma of "one amino acid sequence equals a single folded protein structure" have been seen.[5–7] In Chapter 4, the metamorphic protein lymphotactin exhibits two separate biologically significant folds based on current cell solvent conditions. Chapter 5

studies the Alzheimer's disease related protein amyloid beta (Aβ). These proteins can be induced to adopt an alternative mis-folded structure by other mis-fold Aβ neighbors. The accumulation of these mis-folded proteins are the hallmark for Alzheimer's disease.[8–10] Understanding the molecular forces that drive these folding behaviors provide insights for therapeutic treatments or disease preventions, by developing methods that can target critical points in the folding of the polypeptide chain.[11]

Given the nano-scales in which a protein exist, with respect to both time and size, the study of protein folding is a non-trivial task using traditional experimental approaches. Depending on the method, the experimentally determined structures are only a snapshot of an equilibrium of native conformations, with limited explanation of the forces controlling a conformation's expression. To overcome these shortcomings the folding process of the experimentally resolved structure can be computationally simulated by using Molecular Dynamics.[12] By computing classical Hamiltonian equations of motion, Molecular Dynamic can explore an energy landscape to determine the probability of possible molecular configurations. This barrier can be alleviated by applying algorithmic improvements designed to enhance the sampling of the configuration space. Analyst of this statistical data provides insight in both the thermodynamics and kinetics of the folding process. Unfortunately the number of degrees of freedom that must be solved to properly sample a folding process are a barrier for even small proteins. This barrier can be alleviated by applying algorithmic improvements designed to enhance the sampling of the configuration space.

Numerous of such methods have been created over the years, each with their own approaches on how to efficiently and accurately simulate the folding phenomenon. The intention is to provide yet another method on folding proteins and how to find them.

1.2 Proteins

Proteins are the foundations on which cellular life is built. They are not only the structural scaffolding of the cell but also function in other roles such as transporting other molecules throughout the cell, participating in the defense of the cell against infection, either directly or signalling the need for larger immune response, and even acting as catalysis in reactions. These dynamic capabilities are determined by the three dimensional shape the protein expresses. This correlation between protein biological function and 3D structure is a factor of creating active interaction sites for recognition with other molecules. Given the impact of this relationship, eliciting the factors that control what structure a protein can exhibit is critical to understanding how cellular life operates.[13] Unfortunately this is not trivial.

The genesis of protein can be explained by Francis Crick's Central Dogma of Biology.[14] Though amendments have been made to the Dogma, the core tenets remain, which state that the genetic information encapsulated within the DNA code is transformed to create polypeptide polymers that are proteins.[15] The strands of DNA do not directly participate in the protein synthesis, rather a multi-step process occurs involving the transcription and translation of DNA by ribonucleic acids (RNA).

The DNA helix is first unwound, then replicated during the *transcription* step by messenger RNA (mRNA). By replicating, the genetic code can be used in the construction of larger biomolecules without consuming DNA. Transfer RNA (tRNA) then decodes the mRNA replicated DNA sequences during the *translation* stage to synthesis proteins by rendering an amino acid when their corresponding triplet of DNA codons are processed. A peptide bond (Figure 1.1) is formed between a carboxylic carbon and amide nitrogen linking individual amino acids into

the polypeptide chain. The sequence of the amino acids residues in a chain are unique to each protein, which acts as the first dimension of identity, known as the primary structure. These amino acids are differentiated by what side chain functional group is bound to the α-carbon. The composition of these side chains impact local protein structure through factors of steric hindrance and hydrophobicity.[16–18] In fact, as the translation stage proceeds, a freshly made segment will geometrically transform by folding into itself, forming a more compact state. This transition stage folding does not result in the final native protein structure, rather it is the initial transitions from a "random coil" of chain towards an intermediate molten globule state that lacks the complete organized structure found in a folded native state.[19] It is during this collapse, hydrogen bonding can occur between the peptide backbone carbonyl oxygen and the amide hydrogen, leading to the formation of local structural elements called secondary structure, the second dimension of identity. The shape of these secondary structure elements are a key factor in a protein's function as they form the major active sites. These structural motifs are shared across all protein families and are characterized by the hydrogen bonding patterns.[20–22] The most common and important of these motifs are known as α-

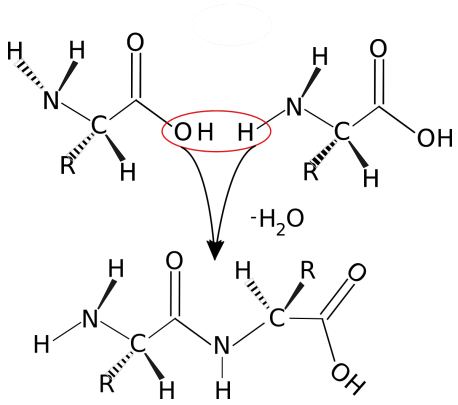

Figure 1.1: Amino acids are linked into a chain when a condensation reaction forms the covalent peptide bond

helix and β-sheets. The α-helix forms when a protein coils into a right turn helix, burying its hydrophobic residues into the core of the helix, creating a hydrophilic interface.[23,24] β-sheets form when strands of the protein are repeatedly pleated

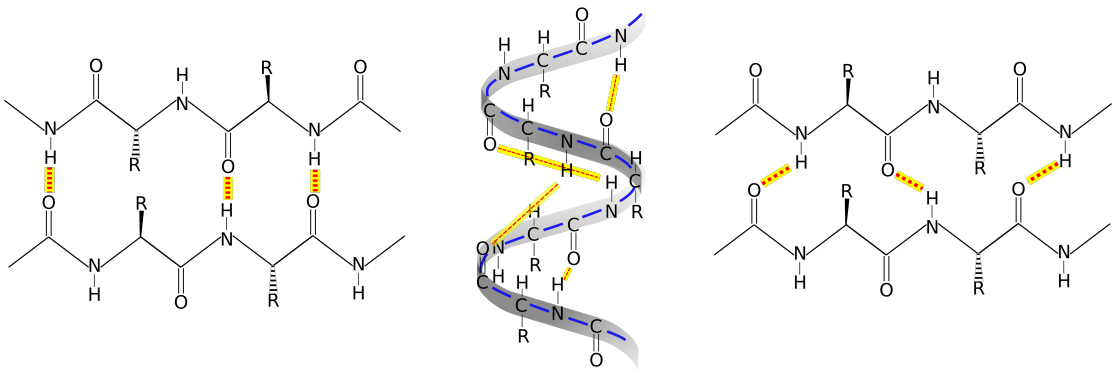

Figure 1.2: The classification of β-sheets is defined by their orientation with the another β strand, as anti-parallel (left) or parallel (right). The α helix between the two β-sheets show the hydrogen bonding pattern that occurs within the helical coil.

with itself. The hydrogen bonding occurs either in a parallel or anti-parallel fashion (Figure 1.2). The pleated nature of β-sheets results in a multitude of extended structures by deforming the strands via twisting and bending. This provides active sites for intra and inter interactions, either with the "face" of the β-sheet or at the edges of the strand.[25,26] These interaction options provide the β-sheet the ability to aggregate other molecular structures or form complexes.[27,28] Though they are fundamental structural elements, proteins are not built entirely of these secondary structures. Rather proteins are a series of different secondary structure motifs, connected by sections that lack the hydrogen bonding which permits for higher flexibility between the more organized secondary structure regions. Thus the complete protein is able to fold into a 3D shape that aligns the individual secondary structures to prime positions for interaction. This level of structural arrangement is the third dimension of protein identity, the tertiary structure. A protein must undergo the entire folding process into the proper tertiary structure to achieve its native fold state.

It should be noted that prior to these larger folding events, post-translation modification may occur, modifying protein's functionality with additional non-amino acid molecules such as sugars and lipids or adjustment of phosphate groups. After these modifications are complete, a protein will begin to exchange energy with its

surroundings as it traverses down a funnel shaped energy surface that represents all possible structures, minimizing its total free energy to enter a stable native state.

Proteins navigate this landscape at rates that deny the possibility that the folding behavior is the result of random motions.[29–32] This leads to what is known as the protein folding problem, which asks how the polypeptide polymer deduces the structure of its native state, the mechanisms that protein folding undergoes, and what can be predicted.[33] The answer to these questions reside within the folding funnel landscape.

1.3 Folding Landscape

The theory that protein folding occurs in a funnel like shape in a multi-dimensional position-energy space derives from early studies that sought to explain the driving forces of folding against the entropic penalties the process overcomes. As the protein forms into a compact state, the chain suffers a lost of entropy that is offset by increases in solvent entropy. A hydrophobic effect drives the non-polar residues to bury themselves within the chain that increases the solvent mobility, which stabilizes developing conformations. In fact a protein's native structure is most affected by mutating a residue with a non-polar amino acid, disrupting the hydrophobic effect.[34] Kauzman was able to demonstrate the hydrophobic effect contribution to the folding process in a study that showed a protein will enter an unfolded denatured state when removed from an aqueous cell like environment to a non-polar solvent, additionally that at temperatures greater than 55°C, where thermal fluctuations are too large for the folding process, and < 20°C when hydrophobic interactions are weak.[35,36] More importantly, this study showed that the denatured proteins would spontaneously refold when returned to proper sol-

vent conditions, and that protein self-assembly is encoded within the amino acid sequence but is influenced by the solvent environment. To model this behavior the forces and movement of the particle's are mapped unto a topological surface known as phase space.[37]

The space is constructed from the collection of all phase points defined by each particles position and momentum in the system. The construct provides an avenue to extract physical quantities by sampling from the ensemble of all possible configurations within the space. A trajectory path for folding events can be obtained by tracking the curves created by a series of phase points through this space. Any such path can not cross itself due the deterministic property of the classical equations of motions used to determine particle movement and the law of conservation of energy, thus each point has a unique "next" point that is determined completely by the previous point. Path crossing would create divergent curves at these points.[38] The dimensionality of the configuration space is commonly reduced to fewer degrees of freedom for projections into thermodynamic energy terms. The reduction simplifies the landscape, but it remains a wasteland of energies, with a surface that is neither perfectly smooth or always a singular funnel.[39]

Potential energy wells exist throughout the landscape that represent meta-stable structures along a folding pathway. Theses wells are defined by their barrier height facing the decent down the funnel, and are the free energy barriers of folding. As a result the protein structures in these will have a larger population, and therefore are targets of research interest.[40,41] They can provide insights into folding mechanisms or as therapeutic treatment targets when structural weak points are exposed.[42–44] The movement down the funnel which produce the protein folding event are dependent on overcoming these barriers that are entropic.

Folding occurs faster during the initial stages, due to greater conformational en-

tropy of an unfolded chain which can more easily overcome the barrier. The process slows as the formation of bonds that constitute the secondary structure motifs leads to a decrease in chain surface area, reducing the ability to expel the heats of formation associated with the new bonds. The thermal fluctuations between the solvent and protein trapped within a well eventually allowing it to overcome the barrier. Proteins atop these barriers proceed further down the funnel to potential wells that are more favorable with lower total energy.[19,45] This is repeated until the protein enters a global minimum, adopting its native conformation. To fold correctly, a pathway must be thermodynamically *stable* and *fast*.[46] Secondary structures are formed by the cooperatively of many weak bonding interactions, and stable intermediate states are necessary to prevent fluctuations from hindering their formation. On the other hand, longer folding times increases the probability for mis-folding during extended periods of exposed states. These two criteria for folding imply a proper pathway is also the path of minimum free energy due to low potential well barriers minimizing the global folding rate.

A mapping of the free energy landscape would reveal the potential wells as regions with the largest number of conformations, where a theoretical folding pathway could be proposed by tracing path between these regions to the global minimum.[47] That will not be a single structure, rather the representation for the ensemble of native conformations the protein can express (see Figure 1.3). Native structure determination from conformations can be complicated by proteins that exhibit multi-funnel landscapes by having more than one thermodynamically stable conformation. Two such classes of proteins are study here, mis-folded proteins and metamorphic.[7,39,48]

Mis-folded proteins can exhibit an alternative stable structure instead of the native conformation proscribed by the amino acid sequence. Not only will these proteins lack their biological function, but in cases of prion-like mis-folded proteins, they

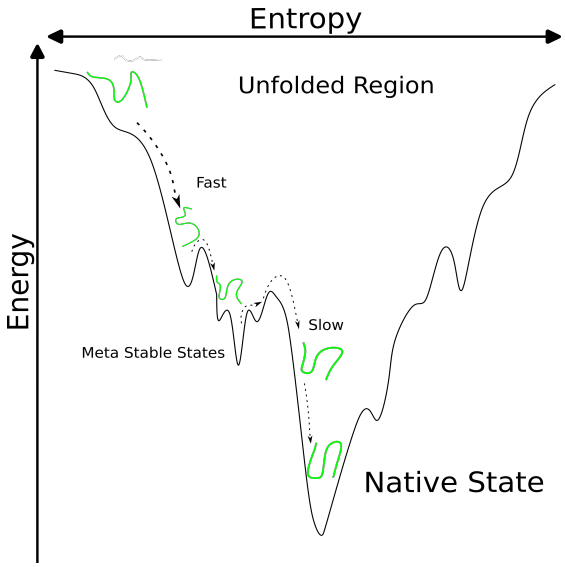

Figure 1.3: Folding begins at the top of this simple example funnel, rapidly condensing to encourage secondary structure formation. The process slows as the protein entering metastable intermediate states within a local energy minimum until overcoming the barriers to descend to the native conformation. Proteins with many conformations will contain a broad or second funnel basin.

will cause their correctly folded neighbors to adopt their mis-folded shape. This behavior has been associated in the pathology of several neurodegenerative disorders where the mis-folded proteins accumulated within the body, resisting attempts of disposal or propagate at rates that exceed removal.[49,50] The protein associated with the Alzheimer diseases amyloid-beta (Aβ) is one such example. Studies have shown that the accumulation of the mis-folded Aβ into insoluble plaque deposits inhibit neuroplasticity resulting in the onset of dementia.[51] In Chapter 5, we show that the mutation of a hydrophobic residue is a likely contributor for the protection against mis-folding of an Aβ mutant strongly correlated with Alzheimer disease.

The multi-funnel behavior can also be beneficial as in the case of metamorphic proteins. Metamorphic proteins have a single sequence with two unique folds, where separate biology functions are associated with each structure.[52] In the case of the chemokine protein lymphotactin, studied in Chapter 4, the two folds exist in near equilibrium and rapidly switch between folds as a function of current cell

needs, signalled via the environment.[53,54] This creates two near equal global minima separated by a free energy barrier that is not the same entropic barrier seen in the folding process, but an enthalpic one due to the thermodynamic work requirements of bond breaking to unfold, in order to reform new secondary structure bonds. Though these barriers are generally much smaller than most of the folding barriers, the changes in structure are pronounced. In lymphotactin, the conversion between folds inverts the protein core, inverting surface residues, and creates an entirely new secondary structure hydrogen bonding network. The method used in Chapter 4 is able to demonstrate that this process is assisted by bifurcated hydrogen bonds that initiate and stabilize the transformation.

To generate an accurate energy landscape that can explore how proteins fold or a system that can exist with a multi-structural state; the surface must be well-sampled. All configurations within energy minima and barrier peaks must be evaluated. This task is difficult to conduct experimentally, due to spectral method limitations and complexity. Examples of these limitations can be seen in X-ray crystallography and Nuclear Magnetic Resonance (NMR).

X-ray crystallography resolves well-defined structures, but its accuracy is diminished by the crystallization process that also reduces conformations to a single structure. Additionally, X-ray diffraction detects only the heavier atoms, and is unable to determine positions of the hydrogen that constitutes the secondary structure.[55] NMR spectra present an averaging of conformations that exist in solutions and cover a range of protein motions. Consequently such NMR data requires a deal of refinement to resolve structures, which includes statistical elimination. The method is limited by a variety of factors. Proteins must be less than 60 kDa. The interconversion rate between conformations states should not exceed spectral relaxation times. Lastly, the protein structure may be affected by the denatured solvent exchanging with the amide hydrogens.[56] Other spectral meth-

ods avoid some of these impediments, but they still share a common difficulty, that of independently producing a well sampled energy landscape for determining thermodynamic or kinetic properties of the folding process. But due to the protein folding protein being a question of thermodynamics, physics can be applied to model the molecular motion to determine behavior as a property of statistical mechanics

Chapter 2

Molecular Dynamics

2.1 Molecular Dynamics

The interactions and motions of all particles must be considered to properly model a protein. The atomistic behavior of these interactions imply that they should be treated quantum mechanically; fortunately this can be simplified into classical mechanics due to the consideration of systems with many identical particles on time and energy scales that exceed individual electronic contributions.[57] This consideration is the foundation of the method known as Molecular Dynamics.

The microscopic properties of many-body proteins are now defined by their positions and momenta using the classical Newtonian equation of motion, $\vec{F} = \frac{d}{dt}(m\vec{v})$. Any electronic contributions can then be simplified into averaged properties due to the low mass of electrons with the Born-Oppenheimer Approximation, that a model's energy can be expressed as a sum of the individual contributions.[58] This allows the total energy to be described by the Hamiltonian with phase space coordinates. States generated with this formulation follow the Boltzmann distribution, from which the probability density can connect the microscopic state properties to macroscopic thermodynamic ones. Exactness requires all points in a system's phase space to be included. Given the impossibility of such a task, a well sampled ensemble may be considered with a large population sampling throughout the space. This requires numerical optimization to compute the large number of

Hamiltonian terms.

Molecular Dynamic energies are calculated using a collection of functional forms with empirically derived constants known as "forcefields." These parameters vary between forcefields, but share the assumption that bonds do not break. Bond lengths are instead modeled by their atom-types equilibrium value, with specification of the atom-types characterizing the forcefield.[59–61] The general form for the sum of individual intramolecular and intermolecular contributions is:

$$E_{pot} = \sum_{Bonds} \frac{k_i}{2} \left(l_i - l_{i,0}\right)^2 + \sum_{Angles} \frac{k_i}{2} \left(\theta_i - \theta_{i,0}\right)^2 + \sum_{Torison} \frac{k_n}{2} \left(1 + \cos\left(n\varphi_i - \gamma\right)\right) + \sum_{non-Bonded} 4\epsilon_{ij} \left(\left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}}\right)^{6} \right) + \sum_{Charge} \frac{q_i q_j}{4\pi\epsilon_0 r_{i,j}} \tag{2.1}$$

Simplified models of motions can be used for the intramolecular forces of stretching, bending and rotation, due to the no-bond-breaking criteria. Bond lengths are assumed to have harmonic behavior (see Figure 2.1), that do not deviate beyond their equilibrium range. Values that deviate are penalized by comparing them with a reference term calculated for when all other forces are zero.

$$U(l_i) = \sum_i \frac{k_i}{2} \left(l_i - l_{i,0}\right)^2 \tag{2.2}$$

where: k = force constant, l_i = bond length, and $l_{i,0}$ = reference bond length.

The modeling of energy as the square of the difference from the reference $l_{i,0}$, maintains the bond close to its equilibrium length, due to high energies being required to significantly deviate.

The angles formed between 3 atoms parameterize the bending motions and can

similarly be modelled using harmonic potentials of the form:

$$U(\theta_i) = \sum_i \frac{k_i}{2} \left(\theta_i - \theta_{i,0}\right)^2 \tag{2.3}$$

where: k = force constant, θ_i = angle, $\theta_{i,0}$ = reference angle

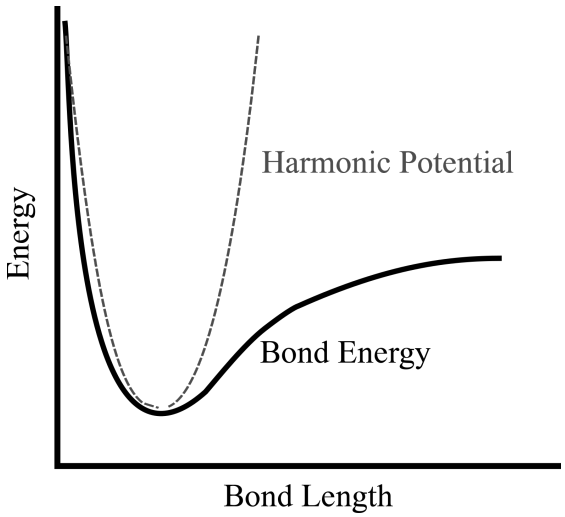

Figure 2.1: The solid black line shows bond energy as a function of length. The assumption of no-bond-breaking allows the tail of the potential bond energy curve to be removed and replaced with a parabola. The potential well at the base of both curves represent the ground-state of the bonds. The overlap between the models shows that a harmonic potential can reasonably model the behavior in this region, given a bond does not deviate far from the equilibrium.

The torsional term for bond rotations can not be consider with harmonics but with a periodic potential. The form varies between forcefields, with the example below shows a basic form that uses a single term for rotational periodicity:

$$U(\varphi_i) = \sum_{n=0}^{N} \frac{k_{n,i}}{2} \left(1 + \cos\left(n\varphi_i - \gamma\right)\right) \tag{2.4}$$

φ_i = torsion angle, γ = period minimum, n = multiplicity, $k_{i,n}$ = period height

Interactions between independent atoms that lack a physical bond connection are modelled as non-bonded energies in forcefields, which are classified as either long-range or short-range interactions. The non-bonded pair term uses them for the van der Waals forces of attraction and repulsion calculated as a Lennard-Jones

potential as function of distance in the form below:

$$U(r_{i,j}) = \sum_{i=1} \sum_{j=i+1} 4\epsilon_{i,j} \left[\left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{12} - \left(\frac{\sigma_{i,j}}{r_{i,j}} \right)^{6} \right] \tag{2.5}$$

where: ϵ = well depth, σ = separation distance at which energy is zero, the minimum occurring at $2^{\frac{1}{6}}\sigma$, $r_{i,j}$ = inter-atom distance

The potential well in Figure 2.2 provides a visual representation for these terms. Cut-off distances can also be applied to optimize computation by limiting the number of neighbors under consideration. Such methods are not applicable to long-range electrostatics. Ionic charges induce dipole moments that are not an additive term, or appropriate for cut-off distance due to an inverse distance relationship.

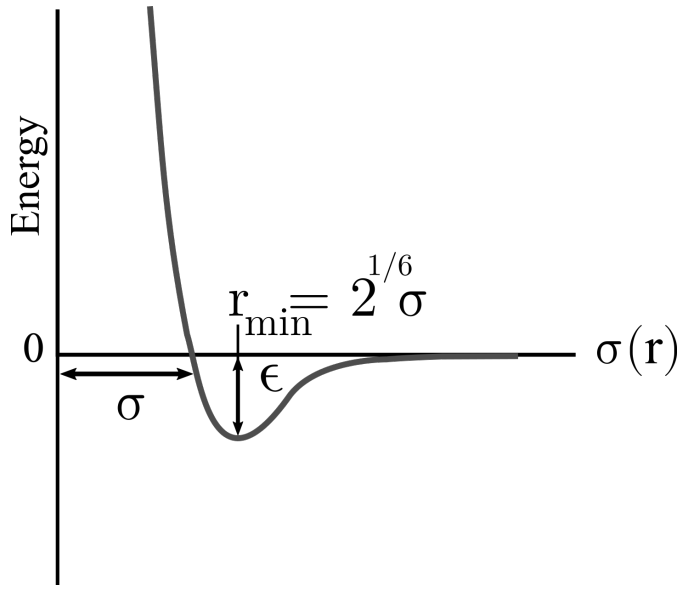

Figure 2.2: The Lennard-Jones potential for non-bonded energy as a function of interatomic separation distance.The terms ϵ is the potential well depth for the interaction, the distance from the well and the potential energy is 0 is measured by σ.

Coulomb pair-pair potentials are therefore applied to one of that neighbors such that:

$$U(r_{i,j}) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.6}$$

where: q_n = charge of n-neighbor, ϵ_0 = permittivity of free charge, and

r_{ij} = Distance between pairs i,j

These terms are evaluated numerically, often with an Ewald summation that stipulates for convergence, the overall system is neutral. The resulting forcefield energy term is used in the equations of motion to compute forces as the change in potential energy with respect to positions.

$$F = \frac{-\delta U}{\delta x_n} \tag{2.7}$$

New positions and velocities are solved with a numerical integrator, the Velocity-Verlet method is used extensively in this work.[62–64] The Velocity-Verlet algorithm contains all of the qualities required in a Molecular dynamic integrator: numerical stability, time reversibility and is also self-starting due to computing the new values on the same time points rather than "leap-frogging" over separate half time steps. This is performed using the following schema in determining the new positions and velocities:

$$Positions : \vec{x}(t+\delta t) = \vec{x}(t) + v(t)\delta t + \frac{1}{2}\, a(t)\delta t^2 \tag{2.8}$$

$$Velocites : \vec{v}(t+\delta t) = \vec{v}(t) + \frac{a(t) + a(t+\delta t)}{2}\delta t \tag{2.9}$$

Molecular Dynamics calculates properties across a constant energy ensemble, therefore the conservation of energy is critical in an integrator.[57] Computer rounding error causes the energy to drift with each trajectory update. Larger integrator time steps reduce this rate but are bounded by the highest frequency of motion, bond stretching. Constraining the bonds as rigid or semi-rigid can optimize the

step to the 2-6 femtosecond range, but are still ineffective at sampling for protein dynamics beyond millisecond ranges.[65–67] It then becomes beneficial to tweak the Molecular Dynamic method. Two general modifications are used in this work: reducing the degrees of freedom by coarse graining the models, and improved sampling methods that enhance configuration space explorations for well sampled ensembles. The following sections outline the theoretical basis for the precise methods implemented.

2.2 Model Granularity

Protein studies can require system sizes that exceed what computers can efficiently manipulate. Coarse graining the model can reduce the computation cost by combining degrees of freedom. The grouping can be represented by a bead of semi-empirical potential. The reduction produces smoother energy landscapes but the larger time steps results fewer computations to obtain a larger sample of states for calculating ensemble averages. Motions of high frequency, such as hydrogen bonds, are primary targets for grouping, but can also group under shared commonality. Recall from Chapter 1 that only backbone protein atoms participate in secondary structure bonding. Therefore coarse graining the entire side chain may still produce reasonable folding information.[68] The beading method is applied in Chapter 5 with the Martini-forcefield that combines four heavy atoms into a single bead, parametrizing each according to the intensity level of their interaction type: polar, non-polar, apolar, or charged.[69,70]

Chapter 4 uses a structure-base reduction called Go-models. This method retains much of the fine-grain quality, relying on an additional potential energy term to bias a model to *go* to its native state.[71] The *Go*-potential is parameterized from a folded structure's native contacts, which drive the model to reform them. Structure-based models therefore required resolved target structures with well-defined native contacts, which is strongly correlated with high secondary struc-

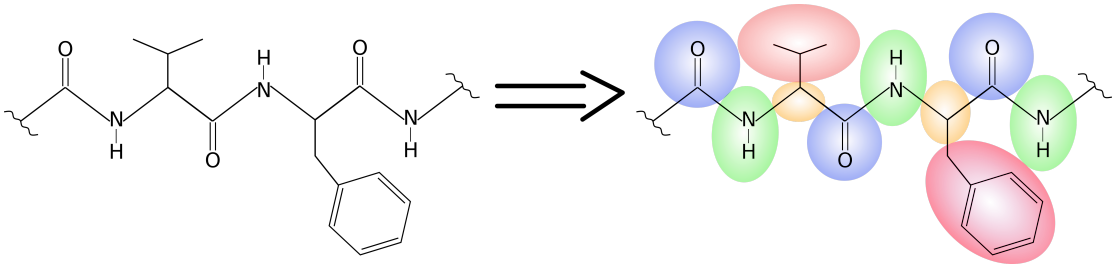

Figure 2.3: Coarse grain beading reduces the degrees of freedom with the color beads representing the all of the overlapped atoms into single parameterized potential.

ture content.[72,73] This limits structure methods to single target considerations, implicating a singular smooth folding funnel. Proteins with multiple native conformations or sizeable disordered regions can not be accurately modeled with this method. These drawbacks can be alleviated by combining with sampling algorithms that operate over multiple model-resolutions.[74,75]

2.3 Sampling Methods

Lack of efficiency in molecular dynamic sampling is often not an issue of physics and can be divided into two competing goals: system exploration (ergodicity) and global convergence (time). Many solutions have been proposed, and what follows is a brief introduction to fundamental method this work builds upon.

Local minimum trapping can make satisfying ergodicity difficult using deterministic equations of motions. In such cases, solutions may be found in stochastic methods such as Markov Chain Monte Carlo (MCMC).[76] The equations of motions are replaced by randomly choosing new positions to generate new configurations. Proposed moves are approved only if they are consistent with the Boltzmann distribution. The acceptance probability is determined in a criteria function comparing the difference the current and proposed state energy difference.

$$w = min(x, exp(-\beta \Delta E)) \tag{2.10}$$

where: x$\in$ random[0,1], $\beta = \frac{1}{k_B T}$

New states lower in energy are accepted with $w > 1$. Proposed states of higher energy are only stochastically accepted if the Boltzmann weight is higher than a random number between 0 and 1. The higher energy states represent uphill transitions that may be unfavorable but small energy difference can increase the probability of escaping the local minimum. Naturally, acceptance rates are then a factor of producing optimal ΔE terms, which hinder the method. Each additional degree of freedom increases the rejection rate exponentially, thus providing little benefit for proteins. To address this shortcoming, more advanced algorithms were developed that combine aspects of MCMC moves, with molecular dynamics motions,[77] the most relevant of these being replica exchange.[78,79]

The protocol simulates a number of identical replica systems using molecular dynamics. Each replica differs from other replicas by a manipulation of a system parameter or degree of freedom value, that can contribute to the folding of a protein when adjusted. The fundamental version of the method is known Replica Exchange Molecular Dynamics (REMD), and performs the method by serially increasing the system temperature over the replicas. The ordering creates a temperature ladder, with high temperature replicas escaping local minimum on one end, opposite of the low temperature replicas that find local minimum when trapped. To sample the N-independent local minimum, periodic exchanges of neighboring replicas are performed. Before any exchange move is performed, MD motions are conducted so the replicas can independently explore their own landscape using natural dynamics. This allows systems to evolve without the high rejection seen in MCMC, but return local minimum trapping issues. After a set length of an MD stage, an exchange move is proposed and accepted according to their Boltzmann weight:

$$w(i,j) = min(1, exp(-\Delta\beta_{i,j}\Delta E_{i,j})) \tag{2.11}$$

where: $\Delta\ \beta$ = Difference in inverse temperature, $\Delta E_{i,j}$ = Difference in potential energy, $i = i^{th}$ replica, and $j = j^{th}$ replica

Depending on the temperatures of the two replicas, the exchange rate can become vanishingly small. This can be alleviated with rescaling of the velocities with a scaling coefficient calculated from a ratio of the kinetic energy term from both replicas.[80]

$$(v_0, v_1) \rightarrow (r v_0, r v_1) \qquad r = \sqrt{\frac{T_0}{T_1}} = \sqrt{\frac{E_{kin}(v_0)}{E_{kin}(v_1)}} \tag{2.12}$$

These exchanges allow the simulation to walk randomly throughout the energy space, improving sampling at the cost of generating artificial trajectories. The performance of REMD is thus dependent on sufficient replica mixing, which is the method's bottleneck. Similar to the energy step size for a MCMC move, the temperature gap between replicas must be minimized to reduce rejection rates. The N-replicas required is unfortunately a factor of system size that scales poorly as a power series. Addressing this bottleneck is the fundamental aim in this work. Two replica exchange protocols are proposed that attempt to simultaneously improve the molecular dynamics stages and exchange while lowering the number of required replicas.

Neither methods use temperature to explore the replica space, instead specific degrees of freedom are targeted using a measure of structural differences between a fine-grain and coarse-grain model. The lack of cardinality between model resolutions force missing degrees of freedom to be reconstructed during communications. Resolution exchange methods lessen the difficulty of this task, reintroducing missing degrees with exchanges of increasing fine grain replicas, but shares all of the bottlenecks of REMD.[81] The pitfalls can be avoided with Multiscale-Essential-Sampling, which directly communicates the structural differences between granularity through the addition of restraining potential terms to the potential energy.

$$E_{pot} = E_{FG} + E_{CG} + \lambda E_{\lambda} \tag{2.13}$$

The λ potential can be scaled using a Hamiltonian Replica exchange variant, that uses the λ coefficient to exchange across a λ space of replicas.[82] The targeted

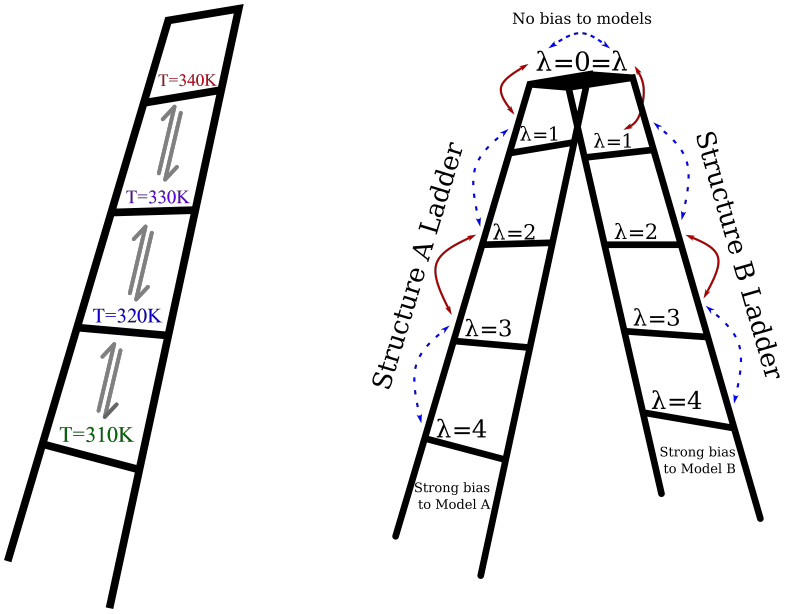

Figure 2.4: REMD has is a single replica ladder that only exchanges up and down with temperature. Replica-Exchange-with-Tunneling has a double ladder system using a Hamiltonian exchanges with λ. The sides do not interact, and are biased to their own model. The strength of the bias is scaled by λ and decreases until 0 at the top. Exchanges with neighbors occurs in pairs that shifts with attempts. First attempt shown as blue dashed arrows. Solid red for second.

behavior of this method requires a strong λ bias with a large replicas system to function. The Replica Exchange with Tunneling (RET) method discussed in Chapter 4 reduces these problems including a tunneling stage that improves exchange rates with fewer replicas and can be applied to two state systems like metamorphic proteins using a twin ladder replica system.[83] The Resolution-Exchange-with-Tunneling method in Chapter 5 uses even fewer replicas for a resolution exchange by communicating in both directions of model granularity The remaining chapters are the application and development of the two methods that utilize features discussed in this chapter to determine protein folding behavior outlined in Chapter 1. The goal of which is to produce a means to efficiently fold a protein and an effective way to find its many folds.

Chapter 3

Research Overview

3.1 Chapter summary

The first two chapters were dedicated to the introduction of the protein folding process and the computational methods to discern their behavior. The genesis of functional proteins described in Chapter 1 does not end with chain creation. The functionality of the molecule only emerges after the chain collapses into a proper structure. With an energy landscape that is populated with intermediate conformations critical to the success. And though this fold is entirely encoded within its amino acid sequence, it is susceptible to solvent effects that is difficult to determine by experiment or prediction.

Chapter 2 discussed how computational physics models can provide theoretical answers to these behaviors. The first half of the chapter established that Molecular Dynamics (MD) can simulate the motions of the proteins but that dynamics of protein folding occurs on time frames beyond the range of what modern computers can achieve. Solutions to this issues are presented in the second part in the forms of model simplifications and enhance sampling methods. Reducing model resolution improves simulation computing cost, therefore feasible trajectory length but at the cost of result details. Where as computational cost and energetic barriers are the constraints to the enhance sampling of energy landscape methods. These potential pitfalls can be alleviate by combining aspects of different methods for

efficient, faster, and accurate system modeling.

3.2 Project Summary

Transition rates times between proteins of two native conformation can exceed the capabilities of many methods. To this extent we applied the Replica-Exchange-with-Tunneling (RET) method (detail explanation in Chapter 4) to investigate the folding switch of the metamorphic chemokine lymphotactin. *Go*-model potentials tuned to each of lymphotactin's two native structures were broadcast to fine-grain system models from the *Go*-model systems. RET is able to perform the rapid conformation switch with exchanges over a replica ladder of scaling λ coefficients for the *Go* bias.

Advancements to the RET design are presented in Chapter 5 with Resolution-Exchange-with-Tunneling (ResET). The number of replicas required to avoid exchange bottlenecks is eliminated, requiring only two replicas of different resolutions. Design validation was performed using the folding benchmark protein trpcage, as well amyloid beta protein.

These chapters were taken from published works as following:

- Bifurcated Hydrogen Bonds and the Fold Switching of Lymphotactin, Journal of Physical Chemistry B 2020, 124(20), 6555- 6564 by Prabir Khauta, Alan J. Ray, Ulrich H. E. Hansmann
- Resolution Exchange with Tunneling for Enhanced Sampling of Protein Landscapes, Physical Review E 106, 015302 by Fatih Yasar, Alan J. Ray, and Ulrich H. E. Hansmann

Chapter 4

Bifurcated Hydrogen Bonds and the Fold Switching of Lymphotactin

The following chapter was published in the Journal of Physical Chemistry B with the dissertation author as the article; Bifurcated Hydrogen Bonds and the Fold Switching of Lymphotactin, Journal of Physical Chemistry B 2020, 124(20), 6555-6564 by Prabir Khauta, Alan J. Ray, Ulrich H. E. Hansmann. All text and figures are taken with the permission from the publisher.

4.1 Abstract

Lymphotactin (Ltn) exists under physiological conditions in an equilibrium between two interconverting structures with distinct biological functions. Using Replica-Exchange-with-Tunneling we study the conversion between the two folds. Unlike previously proposed, we find that the fold switching does not require unfolding of Lymphotactin, but proceeds through a series of intermediates that remain partially structured. This process relies on two bifurcated hydrogen bonds that connect the β_2 and β_3 strands and ease the transition between the hydrogen bond pattern by which the central three-stranded β-sheet in the two forms differ.

4.2 Introduction

Exhibiting a diverse spectrum of functions, ranging from transport of molecules to catalysis of biochemical reactions, proteins play a crucial role in the molecular machinery of cells. Protein function is determined by the three-dimensional structure. In the now classical model of protein folding the sequence of amino acids encodes an energy landscape that funnels folding pathways into a unique native state.[32,84] While this mechanism describes folding of many proteins, it needs to be modified for intrinsically disordered[85,86] or metamorphic proteins[7, 87] where the sequence encodes not only a single native fold but an ensemble of structures, allowing a single protein to take multiple functions. Hence, for understanding the role of intrinsically disordered and metamorphic proteins in cells, and the regulation of their function, it is necessary to establish the underlying multi-funnel energy landscape that leads to this ensemble of diverse structures, and to comprehend the mechanism by which these structures convert into each other.

This task can be most easily tackled for metamorphic proteins such as the 93-residue protein Lymphotactin (Ltn) which are observed in two well-defined structures. As a chemokine Ltn belongs to a family of signaling proteins whose primary function is to direct immune response leukocytes toward areas of inflammation.[53,88] Ltn has only one of the two disulfide bonds otherwise found in chemokines, allowing it to adopt and switch between two well-defined native folds with distinct functions. The first one (Ltn10) is a typical chemokine-like fold with a three-stranded β-sheet attached to a C-terminal α-helix (PDB ID: 2HDM;[89] see Figure 4.1(a)). When in its second form, Ltn40, the protein forms dimers and has a four-stranded β-sheet in a dimeric β-sandwich (PDB ID: 2JP1;[54] see Figure 4.1(b)). The two forms have distinct and complementary functions: Ltn10 activates the XCR1 receptor on the cell surface, while Ltn40 binds to heparin, a polysaccharide component of the extracellular matrix.[7,90] By being able to assume

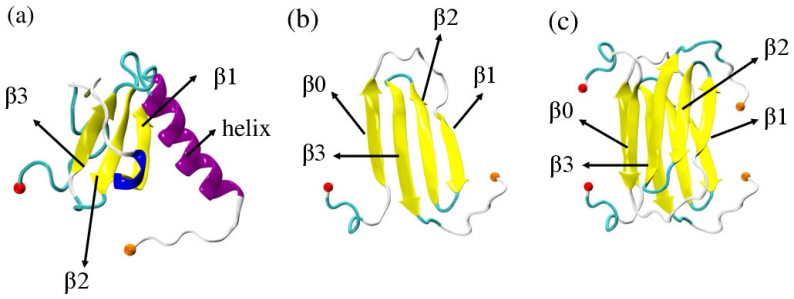

Figure 4.1: Lymphotactin chains can take two distinct structures, both deposited in the Protein Data Bank: (a) Ltn10 (PDB-ID: 2HDM) and Ltn40. The Ltn40 monomer is shown in (b) and derived from the experimentally observed dimer (PDB-ID: 2JP1) shown in (c). Labels identify the secondary structural elements, and the N-terminal and C-terminal C_α atoms are drawn as spheres in red and orange, respectively.

both motifs and perform disparate and complementing functions, lymphotactin bridges the two main functional aspects of chemokine physiology: activation of specific G-protein-coupled receptors (GPCRs) leading to chemotaxis,[91] and establishing a signaling gradient toward the target location by binding with extracellular matrix glycosaminoglycans (GAGs). Under physiological conditions, both forms rapidly interconvert and are equally populated. However, the presence of several basic amino acids (nine Arg and six Lys residues) makes Lymphotactin sensitive to solution conditions and temperature. For example, Ltn10 is the dominant conformation at lower temperature (10°C) and high salt concentration (200 mM NaCl), while the alternative fold Ltn40 is dominant at 40°C and no salt. This shift of the equilibrium with temperature and ionic strength was investigated in CHARMM simulations[92,93] in which an accumulation of sodium ions around the charged residues of the helical region increased with decreasing temperature. These computational results are in agreement with experimental observations that high salt concentration and low temperature stabilize the chemokine fold.[94]

While the structure and function of the two Lymphotactin motifs have been studied in detail,[54,89,90,94,95]the mechanism of interconversion is still unclear. Unlike

other metamorphic proteins such as mitotic arrest deficiency 2 protein (Mad2),[96–98] the fold switch requires a complete reorganization of core residues[7] making the conversion especially difficult to study in experiments. Volkman's group[99] examined the kinetic rates of the process by stopped-flow fluorescence. Their results suggest that the conversion process involves large-scale unfolding with a disruption of all stabilizing hydrogen bonds. However, such a mechanism is difficult to reconcile with the surprisingly low barrier separating Ltn10 and Ltn40 that rather suggests a conversion pathway going through intermediates with conserved local contacts (such as in three β-strands β_1, β_2 and β_3) found in both folds and only encountering minimal disruptions of bonds. Unfortunately, such crucial but transient intermediates are hard to resolve on the short-time scales by which Ltn10 and Ltn40 convert into each other, and may have been below the temporal resolution of the Volkman's experiments.[99] On the other hand, the interconversion time scales are still too long to be studied with sufficient statistics in constant temperature all-atom molecular dynamics simulations. Computational studies of fold switching have to rely instead on either enhanced sampling techniques[100,101] or structure-based models (also called Go-models) [102]. They were used, for instance, to study the fold switching in the transcription factor RfaH.[101,102] A two-funneled Go-model was also used in a recent computational study of Camilloni and Ludovico[103] to probe the fold-switching of Lymphotactin. While this study reported the presence of structured intermediates for the fold-switching, it is not clear how far the presence of the intermediates and the barrier heights reflect the details of the construction of this model rather than the physics of the system.

In this study, we aim to resolve the experimental discrepancy by studying the Lymphotactin conversion process in all-atom simulations relying on a physical force field. As it is difficult to obtain from regular molecular dynamics sufficient statistics, we utilize an enhanced sampling technique developed in our lab, that was designed specifically for the investigation of such switching processes. Our

technique, Replica-Exchange-with-Tunneling (RET),[83,100,101,104] allows us to observe the interconversion process with sufficient detail to characterize important intermediates. This in turn enables us to propose a conversion mechanism that is consistent with the experimentally observed low barrier separating Ltn10 and Ltn40. While the experimentally observed equilibrium is between Ltn10 monomers and Ltn40 homodimers, dimerization and conversion are separate processes, with the transition between Ltn10 monomers and Ltn40 monomers being the rate-limiting process.[99] For this reason, we consider only the conversion of monomers, but one should keep in mind that the Le Chatelier's principle would imply a shift of the equilibrium between Ltn10 and Ltn40 monomers toward the Ltn40 form if also the subsequent dimerization is considered. Our analysis indicates that the fold switch in Lymphotactin monomers occurs along a series of only partially unfolded intermediates, with the breaking and reformation of secondary structure relying on the presence of two bifurcated backbone hydrogen bonds that connect the β_2 and β_3 strands found in both motifs. We conjecture that these bifurcated hydrogen bonds are essential for fold switching, as they allow a repositioning of the β-strand forming residues without the need to cross high energy barriers.

4.3 Materials and Methods

4.3.1 Replica-Exchange-with-Tunneling

In order to understand the mechanism of fold switching in Lymphotactin by way of computer simulations, one has to sample the free energy landscape of the protein with high accuracy. However, the accessible time scales in all-atom molecular dynamics simulations in explicit solvent are even for small proteins with less than 100 residues only of order $\approx \mu s$, and therefore insufficient to obtain sufficient statistics. We have proposed in earlier work[83,100,101,104] a variant of the Hamilton Replica Exchange method[82, 105] as a way to overcome this sampling problem in studies of conformational transitions. Our approach relies on two ingredients.

First, a ladder of replicas is set up, where on each replica a "physical" model is coupled with a structure-based model. On one side of the ladder the structure-based model biases the physical system toward the Ltn10 state, on the other side toward Ltn40. The strength of the coupling (biasing) on each replica is controlled by a parameter λ which is maximal at the two ends, and zero for the central replica, where the physical model is therefore not biased by one of the structure-based models. Exchange moves between neighboring replicas induce a walk along the ladder by which the Lymphotactin configuration changes from one motif into the other. When these exchange moves are accepted or rejected with the criterium commonly used in Replica Exchange Sampling, the correct distribution according to the given λ value will be sampled on each replica. Hence, on the central replica, at $\lambda = 0$, the correct and unbiased distribution of the physical model of Lymphotactin will be sampled. While formally correct, this sometimes also called Multi Scale Essential Sampling (MSES)[74,106] is of limited use in studies of large systems, as the acceptance probability for exchange moves becomes vanishingly small. Hence, the second ingredient of our approach is to replace the canonical acceptance criterium with a new one that allows the system to "tunnel" through the unfavorable "transition state" generated by the exchange move. This tunneling is achieved by re-scaling the velocities of atoms in the two configurations in such a way that the total energy at a given λ value is the same before and after the exchange. The two replicas evolve then by microcanonical molecular dynamics, exchanging potential and kinetic energy, symbolized by the gradual color change in the schematic diagram of Figure 4.2. After a short time (a few picoseconds) the velocity distribution of each of the two replicas will approach the one that would be expected at the given temperature. At this point, the potential energies of the two configurations are compared with the corresponding energies before the exchange move, and either accepted or rejected. We have coined this approach as Replica-Exchange-with-Tunneling (RET), and described it and its limitations in detail in our earlier works.[83,100]

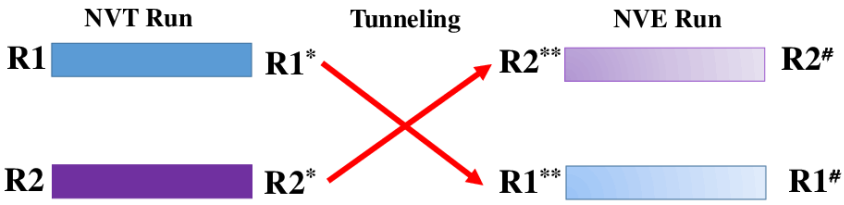

Figure 4.2: A schematic diagram of the RET approach. The new configuration R2** on replica 1 has after rescaling of velocities the same total energy as the old configuration R1*. While evolving to the state R2* potential and kinetic energies interconvert, symbolically noted by the shift in color. In a similar way have the configurations R1** and R2* on replica 2 the same energy, and does R1** evolve to R1.)

In previous work,[100,101,104] we could show that our above described approach leads indeed to an enhanced sampling of transition events and an improved statistics in the sampled energy landscape (which is generated from the unbiased central replica where $\lambda = 0$). Being specifically designed for simulating conversions between known structures, RET proved to be more efficient than other enhance sampling techniques in our previous studies;[100,101,104] however, the improvement is a quantitative one, not a qualitative one. As in generalized ensemble techniques, the system evolves in an RET simulation by an artifical kinetics, which does not neccessarily lead to physical trajectories; see also our discussion in Section 5.4 Results and Discussion. Hence, kinetic information has to be obtained in an indirect way by extracting it from the free-energy landscape by the transition path theory, Markov state model (MSM) analysis, or similar approaches. For instance, upper bounds on the transition rates can in principle be deduced from the height of the free-energy barriers using Kramers' theory, as proposed in ref 37.

4.3.2 Simulation Setup

In the present work we use our enhanced sampling method to study the conversion between Ltn10 and Ltn40 monomers. For this purpose, we simulate our system

with an energy function

$$E_{pot} = E_{phy}(q_{phy}) + E_{go}(q_{go}) + \lambda E_{\lambda}(q_{phy}, q_{go}) \tag{4.1}$$

where E_{pot} is the total potential energy of the system, $E_{phy}(q_{phy})$ and $E_{go}(q_{go})$ are the potential energies of the physical model and Go-model, respectively, and $E_{\lambda}(q_{phy}, q_{go})$ describes the coupling between the two models.

Interactions in the physical model are described by the CHARMM36m force-field[107] in combination with TIP3P explicit solvent[108], with an acetyl group cap on the N-terminus and a methylamine group cap for the C-terminus. The protein is then solvated with 9639 water molecules in a cubic box of length 67.5 Å. Each systems is neutralized by adding 8 chloride (Cl^-) ions. By choosing the box size comparable to the end-to-end distance of a polymer in a good solvent, we try to allow for the possibility that conversion between Ltn10 and Ltn40 requires unfolding of the protein (as was proposed in earlier work[99]), while at the same time minimizing computational costs. Re-scaling the masses of the all-atom physical models by 14.49 is required to match the temperature scales of the two models as the Go-models do not include hydrogen atoms, i.e., have a smaller number of degrees of freedom. The initial configurations (taken from the PDB) are randomized for 1 ns in high temperature molecular dynamics simulations at 1500 K. After visual inspection that the high temperature simulation did not lead to unphysical geometries, we cooled the system down to the target temperature of 310 K by performing an additional simulation of 1 ns duration. The resulting configuration was further minimized to generate the start configuration for the actual RET run.

The two structure-based (Go-)models (one biasing toward Ltn10, the other toward Ltn40) were generated using the SMOG-Server[109] at http://smog-server.org. While the wild-type form of Lymphotactin consists of 93 residues, most of the C-terminus tail is disordered in both conformations, 30% for Ltn10 and 42% for

Ltn40.[54] As the set-up of a structure-based energy function is meaningless for such unstructured regions, we have not considered this tail. Instead, we have restricted our simulations to a 75 residue fragment which describes the parts of Lymphotactin that are structured in at least one of the two folds. The length of this fragment matches the one of the NMR resolved Ltn10 forms (PDB ID: 2HDM[89]), but is larger than the Ltn40 form (PDB ID: 2JP1[54]) for which only 60 residues are resolved. For the generation of the SMOG parameter the remaining 15 residues were assumed to be in a random configuration and added using the PyMOL[110] mutagenesis tool.

The biasing energy is defined as[74,106]

$$E_\lambda = \begin{cases} \frac{1}{2}\left(\Delta^2(i,j)\right) & -ds < \Delta(i,j) < ds \\ A + \frac{B}{\Delta^S(i,j)} + f_{max}\Delta(i,j) & \Delta(i,j) > ds \\ A + \frac{B}{\Delta^S(i,j)}(-1)^S - f_{max}\Delta(i,j) & \Delta(i,j) < -ds \end{cases} \tag{4.2}$$

where ds marks the region in which Eλ is a quadratic function of $\Delta(ij) = \delta_{phy}(ij) - \delta_{go}(ij)$, i.e., of the difference of the distances between C_α-atoms i and j as measured in the two models. Guided by previous work we chose $ds = 0.3nm$. The control parameter f_{max} sets the maximum force when $\Delta(i,j) \to \infty$, and S determines how fast this value is realized. As discussed in earlier studies[74,106] it is convenient to choose $S = 1$ and $f_{max} = 0$. The parameter ds determines the region in which the E_λ is assumed to be a quadratic function of The parameters A and B ensure continuity of E_λ and its first derivative at $\Delta(i,j) = \pm ds$, and are given by

$$A = \left(\frac{1}{2} + \frac{1}{S}\right) ds^2 - \left(\frac{1}{S} + 1\right) f_{max} ds \quad and \quad B = \left(\frac{f_{max} - ds}{S}\right) ds^{S+1}. \tag{4.3}$$

The 24 replica systems were prepared with a λ distribution of λ = 0.1, 0.09, 0.08, 0.07, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01, 0.009, 0, 0, 0.009, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1. Here the E_λ term biases replica 0–10 toward the Ltn10 motif, and replica 13–23 toward Ltn40. In order to simplify our programming, we use two replicas (with indices 11 and 12) to represent the case where the physical model is not biased by any structure-based biasing-term, i.e., where $\lambda = 0$. One of the two replicas exchanges configurations with the neighboring replica in the Ltn10 branch, and the other with the neighbor replica of the Ltn40 branch. Since at $\lambda = 0$ physical and Go-model are independent can the configuration of the physical model be easily exchanged between the two replicas.

Our simulations rely on an in-house implementation of the above described approach into the Gromacs 4.6.5 package[111] that is available from the authors and an Github (*github.com/orgs/hansmann-lab*). The equations of motion are integrated with the Velocity Verlet algorithm,[64] with hydrogen bonds constrained by the LINCS algorithm[112], using a time step of 2 fs. The van der Waals and electrostatic cutoffs are set to 1.2 nm. Note that instead of simulating each of the replicas at a constant temperature, the temperature of the replicas is changed in steps of 0.01 K between 310 and 310.23 K due to the way our code has been implemented into Gromacs, where the temperature of the replicas is maintained by using the v-rescale thermostat.[113] Choosing the length of the microcanonical segment in the RET move as 1 ps, a 100 ns trajectory was generated. Thermodynamic quantities were calculated solely from replicas with $\lambda = 0$, i.e., without bias from the two structure-based models.

4.3.3 Observables

The free energy as a function of order parameter, λ, is defined as

$$\Delta G(\lambda) = -k_B T \left[\ln \rho(\lambda) - \ln \rho_{max}\right] , \tag{4.4}$$

where k_B and T denote the Boltzmann's constant and temperature, respectively. ρ is an estimate of the probability density function calculated from a histogram of the data, while ρ_{max} is the maximum of the density. The second term ensures that $\Delta G = 0$ for the lowest free energy minimum. Free energy values reported in this work are calculated at a temperature of 310 K.

The configurations as obtained from the equilibrated trajectories that correspond to unbiased trajectories ($\lambda = 0$) are characterized based on secondary structure pattern. The pattern are calculated using the STRIDE algorithm[114] as implemented in VMD software[115] and characteristic backbone hydrogen bonding pattern specific to each of the Ltn native forms (see Table 4.1). Ltn has four characteristic β-strands, β_0 (residue 10–15), β_1 (residue 25–31), β_2 (residue 34–41), β_3 (residue 44–51) and a C-terminal helix, H (residue 54–66). Based on the secondary structural pattern of these five regions and backbone hydrogen bonding pattern in $\beta_1 - \beta_2 - \beta_3$ region, we define three variables (B0, H, and B123) using following logical expression, which eventually helps us characterize the configurations and distinguish the pattern among them.

- **B**0: If there are at least two residues in β-strand in β_0 region, then the value of B0 is 1 i.e., β_0 exists; otherwise it is zero.
- **H**: If there are at least three residues in helix in C-terminal helix region, then the value of H is 1 i.e., C-terminal helix exists; otherwise it is zero.
- **B**123: If there is no characteristic hydrogen bonds (Ltn10-like or Ltn40-like) in $\beta_1 - \beta_2 - \beta_3$ region and there is less than two residues in β-strand in

$\beta_1-\beta_2$ or $\beta_2-\beta_3$ regions, B123 will be assigned to have a value of zero. If this condition is not satisfied, B123 will have hydrogen bonding pattern as that of Ltn10-like or Ltn40-like or Mixed (i.e., substantial existence of both type of hydrogen bonds) depending on the conditions: if the difference in number of characteristic Ltn10-like and Ltn40-like hydrogen bonds (see Table 4.1) is greater than or equal to two, B123 is Ltn10-like. Similarly, if the difference in number of characteristic Ltn40-like and Ltn10-like hydrogen bonds is greater than or equal to two, B123 is Ltn40-like. All other remaining configurations will be considered to have mixed hydrogen bonding pattern.

One quantity by which we have measured similarity of a given configuration to one of the two native Lymphotactin folds is the fraction of specific native contacts $Q_{spec,1}(X)$, defined as

$$Q_{spec,1}(X) = \frac{1}{N}\sum_{(i,j)} \frac{1}{1+\exp\left[\beta\left(r_{ij}(X) - \min(\lambda r_{ij}^1, r_{ij}^2)\right)\right]} \tag{4.5}$$

Here, 1 denotes one native fold of Lymphotactin (Ltn10 or Ltn40), while 2 represents the alternative one. Only contacts specific to either of the native folds are considered, i.e., contacts that are found in both native folds are excluded. Here, we define a contact in the native structure by the requirement that two backbone atoms on distinct residues are within 4.5 Å. Thus, N is the number of such contact pairs (specific to the one form of Lymphotactin native structure) of (i, j) backbone atoms i and j belonging to residues θ_i and θ_j. To avoid the contacts forming by adjacent residues, we have considered only the residues where $\mid \theta_i - \theta_j \mid > 3$. $r_{ij}(X)$ is the distance between the atoms i and j in conformation X, while r_{ij}^1, r_{ij}^2 are that distance in the native fold 1 and 2, respectively. β is used to smooth the distribution of the values and considered to be 5 Å^{-1}. The fluctuation of the contact formation is controlled by the minimum value of λ times r_{ij}^1 and r_{ij}^2, where λ is taken to be 1.8. By introducing such a minimum value between the above

mentioned two quantities, we set a range of fluctuation for each of the specific native contacts considered so that one could differentiate between the set of specific native contacts with respect to the two folds, and thus measure the similarity of a given configuration with respect to only one of the two native folds.

The transition pathway between Ltn10 and Ltn40 is derived from the free energy landscape projected on suitable coordinates by calculating the minimum energy pathway with the MEPSA software.[116] This user-friendly software utilized the graph-theory based Dijkstra's algorithm[117] to construct a minimum energy pathway between two given minima, identifying the barriers among different minima along such a pathway. Hence, the perspective here is one of transition state theory. Unlike competing approaches such as, transition path sampling, (autocite paper 50-53), string method, (cite paper 53,54) kinetic network model (cite paper 55), traveling-salesman-based automated path searching (TAPS), (cite paper 56) etc., it requires fewer resources but may not in all cases identify the optimal pathway.

4.4 Results and Discussion

Previous investigations into the conversion mechanism between the two Lymphotactin motifs were hampered by the computational difficulties of sampling the protein's free energy landscape with sufficient statistics. In regular molecular dynamics simulations, the protein will spend most of the time in one of the basins of attraction - either exploring configurations similar to Ltn10, or, in the other case, Ltn40-like configurations - with transitions between the two basins being rare events. We argue that these computational difficulties can be overcome with our variant of replica exchange sampling which was designed specifically for investigation of switching between two well-defined states. In order to demonstrate that our approach allows indeed a more accurate sampling of the free energy landscape by raising the rate of transitions between the two main basins, we show in Figure 4.3(a) the walk of a typical realization of our system along the ladder of

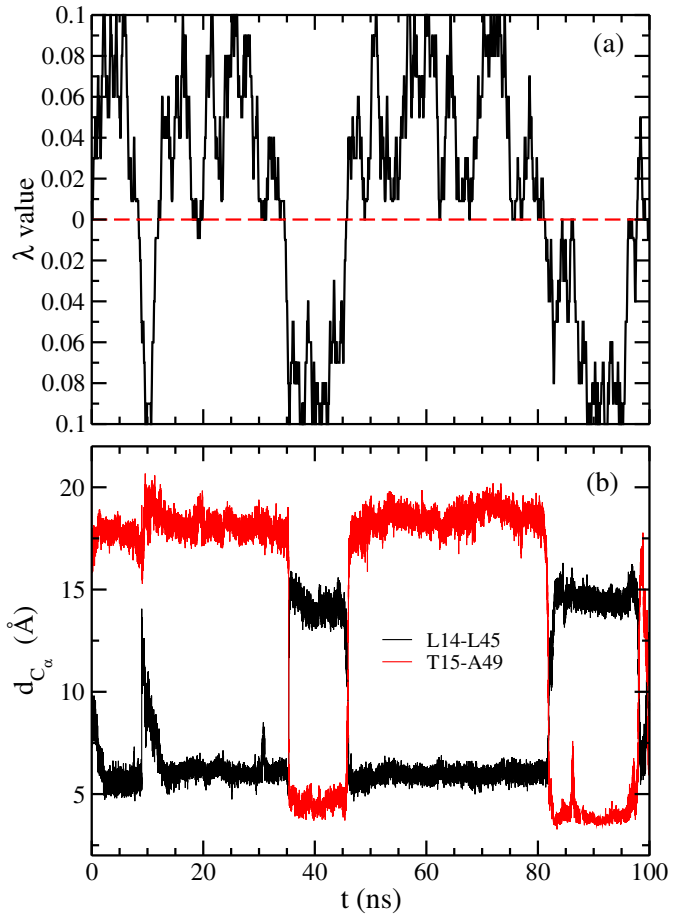

Figure 4.3: (a) A typical example of a replica walking through λ space starting from a replica, where the physical model is initially biased toward Ltn40 with $\lambda = 0.009$. While the system walks between replicas with bias toward Ltn10 and such with bias toward Ltn40, its configuration changes accordingly. This can be seen in (b) where we show the corresponding time evolution of C_α distances (d_{C_α}) between L14 and L45 (black) and that between T15 and A49 (red). The first distance is a measure for the similarity with Ltn40, and the second distance one for the similarity with Ltn10.

replicas. At start time ($t = 0$) the physical system sits on a replica where it is biased with $\lambda = 0.009$ toward the Ltn40 form. During the 100 ns of simulation this realization of Lymphotactin walks numerous times between the two end-points of the ladder. On the one end, the physical system will be biased maximally with coupling parameter $\lambda_{max} = 0.1$ toward the Ltn40 structure, and on the other end with a maximal $\lambda_{max} = 0.1$ toward Ltn10. The average exchange rate between neighboring replicas is $\sim 47\%$. In Figure 4.3(a) we show that this walk through λ-space induces indeed inter-conversion between the two forms. For this purpose, we characterize the state of a given configuration by the C_α distances between

two specific residue pairs. In the Ltn10 structure the two residues T15 and A49 form a hydrophobic contact, while they are separated by a large distance in Ltn40. The opposite relation is found for the pair L14 and L45, which form hydrophobic contacts in Ltn40 but are far away from each other in the Ltn10 structure. Both distances are shown in Figure 4.3(b) and evolve in anti-correlated fashion over the 100ns of simulation. If the system is on one side of the ladder and biased toward Ltn10, the distance between T15 and A49 (shown in red) will have small values and the distance between L14 and L45 (drawn in black) have large values, while the opposite is true once the system is on replica where the bias from the structure-based model is toward the Ltn40 structure.

A measure for the efficiency of our method, and a lower limit on the number of independent configurations sampled at the $\lambda = 0$ replicas, is the number of walks along the whole ladder, from the replica with maximal bias toward Ltn10 to the one with maximal bias toward Ltn40, and back. Such a walk along the whole ladder of replicas is called by us a tunneling event, and inversely related to the average time needed to cross the ladder (termed by us the tunneling time). The higher the number of tunneling events, and the shorter the tunneling time, the more efficient will be our approach. However, calculation of the number of tunneling events or the tunneling time gives only meaningful results after the system has approached equilibrium. This convergence of the simulation is checked by calculating the free energy as a function of the two distances introduced above, and comparing it for two non-overlapping different time intervals.

Resemblance of the data for these two intervals in Figure 4.4 suggests that the simulation has converged after 20 ns, and therefore, we use the last 80 ns of the simulated trajectories for our analysis. In this time span, we find a total of 31 tunneling events with an average tunneling time of $\sim$ 33 ns. We have demonstrated in earlier work[83,100,101,104] for smaller systems that the number of tunneling events

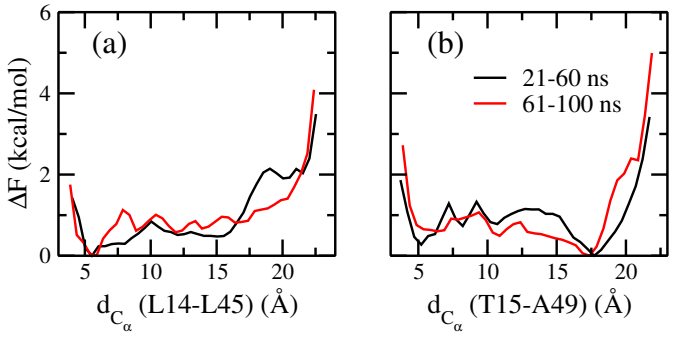

Figure 4.4: Free energy (ΔF) as a function of the two C_α distances (d_{C_α}) between (a) L14 and L45 (a measure for the similarity to Ltn40) and (b) between T15 and A49 (a measure for the similarity to Ltn10), as obtained from the unbiased replica where $\lambda = 0$. Shown are values as measured for different segments of the simulation.

is in our approach much higher than the ones found in regular Hamilton replica exchange simulations with comparable number of replicas and λ distribution. In the latter case, often not a single tunneling event could be detected. As Lymphotactin is larger than the previously studied systems, and the sampling difficulties increase exponentially with system size, we expect that for Lymphotactin the improvement over regular Hamilton replica exchange is even higher than seen in our previous work.

The increased efficiency of our approach, leading to 31 tunneling events, gives us confidence in the free energy landscape found at $\lambda = 0$, i.e., at a replica where the "physical" model of our system is not biased toward either Ltn10 or Ltn40. In order to measure the frequency of these two motifs for the unbiased replica, we define a configuration as Ltn10-like if the C_α distances between T15 and A49, d_{C_α} (T15-A49), is less than 8 Å and and that between L14 and L45, d_{C_α} (L14-L45), greater than 12 Å. This definition was derived from visual inspection of the landscape as obtained for the replica with maximal bias toward Ltn10. Guided in a similar way by a visual inspection of the landscape as obtained for the replica with maximal bias toward Ltn40, we define a configuration as Ltn40-like if d_{C_α} (L14-L45) is less than 8 Å and d_{C_α} (T15-A49) greater than 12 Å. Using the above definitions we find that on the unbiased replica about 17% of configurations are Ltn10-like, while

34% are Ltn40-like. This suggests that Ltn40 is the most stable form and about 50% of the configurations sampled in our simulation do not represent either fold. Visual inspection and secondary structure analysis indicate that most of those configurations are intermediates on the pathway between the two opposite folds. The slightly higher population of Ltn40 over Ltn10 is in accord with the experimental study,[94] where under physiological conditions 46% of the configurations are Ltn10-like, and 54% Ltn40-like. On the other hand, configurations that do not belong to either of the two motifs are not observed with the large frequency seen in our simulations. This difference can be explained by the low temporal resolution of the experiments which makes it difficult to characterize short-lived intermediates, i.e., the experimentally reported frequencies for Ltn10 and Ltn40-like configurations are relative frequencies resulting from the well-resolved signals corresponding to 12 residues collected from a two-dimensional ^{1}5N-^{1}H HSQC spectrum.

In order to explore the inter-conversion pathway, we show in Figure 4.5(a) the free energy landscape projected on the two characteristic distances introduced earlier: d_{C_α} (L14-L45) (measuring similarity with Ltn40) and d_{C_α} (T15-A49) (quantifying similarity to Ltn10). For calculating the landscape, we use only data sampled at the $\lambda = 0$ replicas, where the physical model is not biased by any Go-model term.

The so-drawn landscape is characterized by two prominent basins corresponding to either Ltn10-like or Ltn40-like configurations. Conversion events can be described by pathways connecting the two basins in the landscape. However, not all possible pathways are equally likely. Take as an example the pathway represented by a black line in Figure 4.5(a). This path is obtained by projecting onto the landscape the configurations sampled along the walk in λ space during a certain tunneling event.

This pathway describes a transition between the two Lymphotactin folds that requires crossing an energy barrier much higher than that reported in the experi-

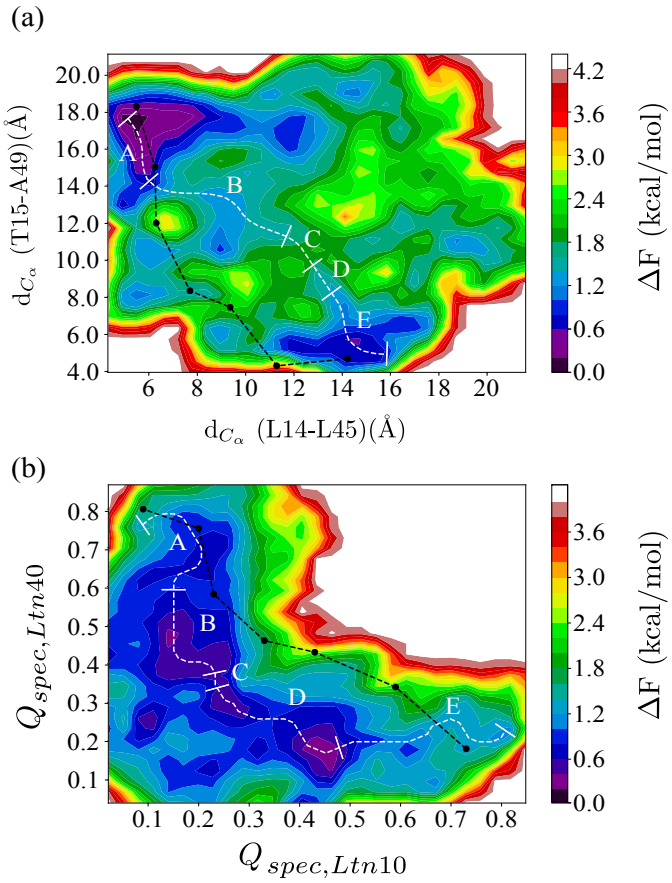

Figure 4.5: Free energy landscape as obtained from our RET simulation at replicas where the physical models are not biased by any Go-term. The landscape is in (a) projected on C_α distances between L14 and L45 ((d_{C_α} (L14-L45)), a measure for similarity to the Ltn40 structure) and between T15 and A49 ((d_{C_α} (T15-A49), a measure for the similarity to Ltn10). The black line shows a typical pathway as obtained during a tunneling event, i.e., from a walk in λ-space between the structure strongly biasing toward Ltn10 and that toward Ltn40. On the other side, the minimum energy pathway as obtained from MEPSA software[116] is drawn in white. The labels A-E mark five distinct regions that can be identified along this pathway (discussed in the text). For comparison, we show in (b) the same landscape, but projected on the fraction of specific native contacts (defined in Eqn. 4.5) with respect to Ltn10 ($Q_{spec,Ltn10}$) and Ltn40 ($Q_{spec,Ltn40}$) native structure. Both the minimum energy and the tunneling pathways are shown again, using the same color coding as in (a).

ments, indicating that this is not true pathway. On the other hand, we can obtain a thermodynamically reasonable pathway by using the MEPSA software[116] to construct the minimum energy pathway, which we have drawn as a white line in the landscape in Figure 4.5(a). The algorithm used to determine such a minimum energy pathway in the MEPSA software[116] and the advantages of this software have been discussed in Section 5.3.3. Unlike the tunneling pathway, this pathway

does not go through regions of the landscape characterized by unfolded configurations, but instead proceeds through a series of basins, with an effective energy barrier similar to that reported in experiments. This indicates that the interconversion process does not proceed by unfolding of the Ltn10 or Ltn40 structure but rather involves a series of intermediates or transition states. By construction the minimum energy pathway does not connect specific configurations but bins in the landscape each containing a certain number of configurations sampled throughout the simulations. These configurations can be characterized according to presence (or lack) of a C-terminal helix (found in the Ltn10 structure), the N-terminal β_0 strand (found in the Ltn40-structure), and the hydrogen bonding in the β_1 to β_3 region, which offers another way of distinguishing between the Ltn10 and Ltn40-like configurations. The procedure by which we attribute these three traits to a configuration is described in the method section. The frequency with which the three traits are observed allows one to identify five distinct regions along the pathway that correlate with the basins and barriers of the landscape. These segments are labeled as A to E in Figure4.5(a). A similar division is not possible for the pathway, derived from the tunneling event.

The difference between the two possible pathways does not depend on the specific coordinates on which the landscape is projected. This can be seen from Figure 4.5(b) where we project the free energy landscape on the fraction of specific native contacts with respect to each of the two folds, and overlay again both paths on the landscape. The pathway derived from the tunneling event is again not consistent with the landscape. Hence, the tunneling events in our RET approach cannot be used to derive a conversion mechanism as they rely on an artificial dynamics, designed to increase sampling efficiency. On the other hand, while the configurations in the minimum energy pathway calculated for the new landscape will differ from the one calculated for the other landscape, we can again identify the same five regions. We remark that the radius of gyration (a measure for the compactness

of configurations) of the central part, made up of $\beta_1 - \beta_2 - \beta_3$ in both Ltn10 and Ltn40, differs little along the pathway, which implies that the Lymphotactin configurations do not unfold and refold while assuming this pathway during the inter-conversion process.

The qualitative agreement between the minimum energy pathways found for the two landscapes suggests that these pathways describe indeed the conversion process. Hence, in order to determine the mechanism and to establish the separating free energy barrier, we have analyzed in more detail the pathway shown in Figure 4.5(a). When starting from the Ltn40 basin of region A, the β_0 strand gets dissolved when reaching basin B which has about 1.5 kcal/mol higher free energy than the Ltn40 fold of basin A. Upon crossing this barrier, the Lymphotactin configuration evolves further through a number of intermediates with little difference in free energy, forming the flat-floored valley of region C. Using visual inspection and the STRIDE algorithm[114] as implemented in VMD software[115] for secondary structure analysis, we observe a re-arrangement of backbone hydrogen bonds, until, when entering region D the hydrogen bond pattern of the remaining three β-strands (β_1, β_2, and β_3) becomes similar to that of the Ltn10 fold.

The Ltn10 basin (region E) has again about 1.5 kcal/mol lower free energy value than the transition state D and is reached when the helix slowly begins to form at the C-terminus (within the residues 54 to 66). A schematic diagram explaining the mechanism is shown in Figure 4.6.

From the above discussed conversion process, we have estimated a total energy barrier of $\sim$ 2 kcal/mol for Ltn40$\rightarrow$Ltn10 conversion, while slightly a lower value of $\sim$ 1.5 kcal/mol is the barrier for the reverse process.

While the higher energy barrier for the Ltn40$\rightarrow$Ltn10 conversion, implying a slower conversion rate than the reverse one, is in agreement with the experimen-

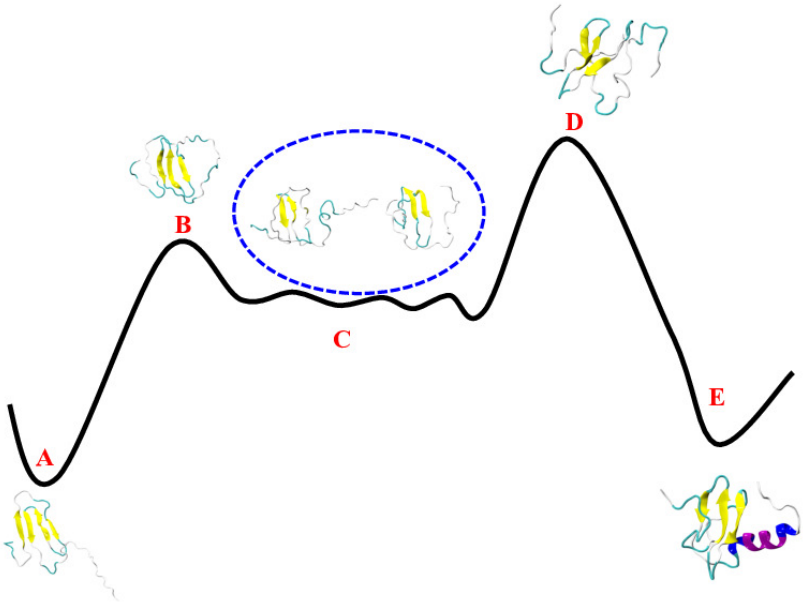

Figure 4.6: A schematic diagram of the interconversion mechanism between Ltn40 and Ltn10 with representative structures.

tally reported results,[90,99,118] the energies are higher than the ones measured in the experiments: 0.5 kcal/mol and 0.2 kcal/mol, respectively. This divergence may reflect again the difference between our computational setting (considering a transition between monomer structures) and the experimental setting that studies the transition from Ltn10 monomers to Ltn40 dimers. As the dimerization is the fast process, the Le Chatelier's principle will predict a shift of the equilibrium toward the Ltn40 motif, effectively lowering the barrier.

Unlike an earlier proposal,[99] our conversion mechanism does not require unfolding of the Ltn monomers. Instead it assumes partially conserved contacts as have also been reported in another previous computational study.[103] Our mechanism relies on a sequence of local changes with the main assumption that the intermediates keep part of the local ordering. Especially, we assume that the central three-stranded β-sheet (β_1, β_2, and β_3) is preserved. This sheet is found in both motifs, but with a shift in the backbone hydrogen pattern by one residue,[54] see Table 4.1. Not that residues on the central β_2 strand do not get shifted. Instead, they change their hydrogen bond forming partners on the adjacent strands (β_1 or β_3). Our pro-

Table 4.1: Characteristic Backbone Hydrogen Bonding Pattern among Different β-Strands for Ltn10 and Ltn40 Lymphotactin Native Forms as Obtained from Our Study and Reported in the NMR Structure.[54]

$\beta_1 \leftrightarrow \beta_2$		$\beta_2 \leftrightarrow \beta_3$	
Ltn10	Ltn40	Ltn10	Ltn40
T26 $\leftrightarrow$ I40	K25 $\leftrightarrow$ I40	–	R35 $\leftrightarrow$ D50
T28 $\leftrightarrow$ I38	Y27 $\leftrightarrow$ I38	V37 $\leftrightarrow$ A49	V37 $\leftrightarrow$ C48
T30 $\leftrightarrow$ A36	I29 $\leftrightarrow$ A36	F39 $\leftrightarrow$ V47	F39 $\leftrightarrow$ K46
–	E31 $\leftrightarrow$ L34	T41 $\leftrightarrow$ L45	T41 $\leftrightarrow$ G44

posed conversion mechanism therefore has to explain how this shift between these residues can proceed without the need of a complete unfolding. A more thorough analysis of the pathway shows that the unfolding of the β-sheet and the resulting high barriers are avoided by some residues forming bifurcated/bridged hydrogen bonds with two consecutive residues. For example, while in the Ltn40 configurations of basin A, residue F39 forms a backbone hydrogen bond with residue K46, and in the Ltn10 configurations of basin E backbone hydrogen bonds with residue V47. However, we observe that F39 (located on strand β_2) can also form simultaneously hydrogen bonds with residues K46 and V47 on strand β_3. In this case, the amide nitrogen of F39 only participates in forming hydrogen bonds and thus acts as a bifurcated donor, or both carbonyl oxygen and amide nitrogen of F39 participate in forming those hydrogen bonds, helping in forming bridged hydrogen bonds. Similarly, we find that in a similar manner residue T41 can also form hydrogen bonds with both residues G44 and L45. All six residues on the central β_2 strand (see Table 4.1) can form such bifurcated hydrogen bonds that bridge between Ltn10-like and Ltn40-like hydrogen bonding, but for most residues these bifurcated hydrogen bonds appear with a relative frequency of less than 5%, that is, less than 5% of all hydrogen bonds connecting a residue on strand β_2 with a partner residue on either β_1 or β_3 are bifurcated hydrogen bonds.

The exceptions are the already mentioned bifurcated hydrogen bonds F39-K46/V47,

which is observed with a relative frequency of about 24%, and T41-G44/L45, which appears with a relative frequency of about 11%. We remark that we also find similar frequencies for the minimum free energy path of the free energy landscape in Figure 4.5(b). Both bifurcated hydrogen bonds connect residues located at the start of the sheet formed by the β_2 and β_3 strand. In the Ltn10 motif, residue F39 forms a hydrogen bond with V47, and T41 one with L45. Transitioning to the Ltn40 hydrogen bonding of F39 with K46 and T41 with G44 will be eased by transient formation of the bifurcated hydrogen bonds at these locations as they avoid the energetic costs of dissolving and reforming hydrogen bonds. Correspondingly, the F39-K46/V47 is observed with highest relative frequency in the transition regions B (about 37%) and D (about 64%), and with about 16% relative frequency in the intermediate region C. The corresponding frequencies are lower for the T41-G44/L45 bifurcated hydrogen bond, which appears in 8% (11%) in the transition region B (D), and with a relative frequency of 14% in the intermediate region C.

We conjecture that formation of the bifurcated/bridged hydrogen bonds F39-K46/V47 and T41-G44/L45 at the turn region between the β_2 and β_3 strands is crucial for enabling the fold switch as it disturbs the geometry of the sheet and initiate a wave of successive re-arrangement hydrogen bonds in the three-stranded β-sheet which avoids large energy barrier that would otherwise arise from breaking and forming hydrogen bonds. Our conjecture could be tested in principle by mutation experiments where the mutated side chain would form contacts that lead to repositioning of the backbone atoms that would restrict formation of such bifurcated hydrogen bonds. Another possibility would be the use of deuterium which would also alter the relative frequency of bifurcated hydrogen bonds.

Interestingly, both bifurcated hydrogen bond pairs are also seen with substantial frequency in the region E (dominated by Ltn10-like configurations), where the F39-K46/V47 bond is observed in about 49% of the configurations, and the

T41-G44/L45 one in about 16% of configurations. Hence, in Ltn10-like configurations the hydrogen bond pairs F39-V47 and T41-L45 are easily replaced by the corresponding bifurcated hydrogen bonds. This is likely because the extension from the hydrogen bond F39-V47 to the bifurcated hydrogen bond F39-K46/V47 involves rotation of the side chains of the three residues that reduces the hydrophobic solvent accessible surface area (SASA) by about 26 Å^2, and increases the exposure of the charged K46 by about 7 Å^2. Both are energetically favorable changes. The effects are less pronounced for the bifurcated hydrogen bond T41-G44/L45 where there is only a small reduction in hydrophobic SASA $\approx$ 7 Å^2. On the other hand, in region A (where Ltn40-like configurations are found) only the T41-G44/L45 bifurcated hydrogen bond is observed with substantial, but much smaller, relative frequency (about 9%). Here, formation of a bifurcated hydrogen bond T41-G44/L45 would not change the solvent accessible surface area of the three involved residues, while the bifurcated hydrogen bond F39-K46/V47 would lead to a reduction of about 45 Å^2 of solvent exposure for the charged K46 that could not be compensated by the favorable loss of hydrophobic SASA of about 18 Å^2. However, we remark that unlike to the Ltn10 configuration, where the β_3 strand is stabilized by four contacts with the N-terminal helix, no such β_3-stabilizing contacts with the β_0 strand exist in the Ltn40 configuration. Hence, the barrier for unraveling the $\beta_2 - \beta_3$ hydrogen bonding is likely lower in the Ltn40 configurations than in the Ltn10 configurations where instead formation of the two bifurcated hydrogen bonds circumvents the otherwise higher barrier.

Differences in the number of stabilizing side chain contacts are also the reason why we observe bifurcated hydrogen bonds with appreciable frequency only between β_2 and β_3, but not between β_2 and β_1. In the Ltn10 configuration, four side chain contacts (A49-W55, A49-V56, D50-V56, P51-V56) with the helix stabilize the β_3 strand, but ten such side chain contacts (L24-W55, L24-C59, L24-M63, L24-K66, T26-M63, Y27-V60, Y27-R61, Y27-M63, Y27-D64, I29-V60) that stabilize

the β_1 strand. The difference in number of stabilizing side chain contacts is even larger for the Ltn40 configuration. In this motif are no side chain contacts with the β_0 strand that could stabilize the β_3 strand, but four such contacts stabilizing the β_1 strand. As formation of bifurcated hydrogen bonds requires repositioning of backbone atoms that in turn depends on suitable rotation of side chain atoms, such bifurcated hydrogen bonds are less likely between β_1 and β_2 than between β_2 and the more flexible β_3.

4.5 Conclusions

Using a variant of Replica-Exchange-with-Tunneling (RET), we have probed the interconversion of a metamorphic protein that switches biological function by altering its three dimensional structure between two well-defined native forms, Ltn10 and Ltn40. While Ltn10 is monomer with three-stranded β-sheet ending with a C-terminal helix, Ltn40 exists as a dimer with all β-sheet arrangements. In order to ease the numerical difficulties, we have only considered the conversion for a monomer. This is justified as conversion and dimerization are separate events, with the conversion the time-limiting process.[99] Our investigation relies on the use of RET, an enhanced sampling method developed in our group, that has enabled us to sample the free energy landscape of the protein with high precision. We find relative population frequencies that are consistent with experimental measurements, but our simulations predict a larger population of intermediate configurations than reported in the experiments. We reason that our method allows us to identify intermediates that due to their short-life time are difficult to observe in experiments. Analyzing the free energy landscape allows us to identify an conversion mechanism that relies on passage through a number of distinct structural intermediates, and involves breaking and reformation of the β_0-strand and the C-terminal helix and a re-arrangement of hydrogen bonds in the central three-stranded β-sheet made from β_1, β_2 and β_3. The associated high costs of breaking and forming hydro-

gen bonds are avoided by formation of bifurcated hydrogen bonds that naturally bridge between the characteristic hydrogen bond pattern in the three β-sheets common in both motifs. These pattern differs in both forms by being shifted by one residue. We surmise that formation of these bifurcated hydrogen bonds facilitates the switch between these two patterns, guiding in this way the conversion between the two motifs.

Chapter 5

Resolution Exchange with Tunneling for Enhance Sampling of Protein Landscape

The following chapter was published in Physical Review E with the dissertation author as the article; Resolution Exchange with Tunneling for Enhanced Sampling of Protein Landscapes, Physical Review E 106, 015302 by Fatih Yasar, Alan J. Ray, and Ulrich H. E. Hansmann. All text and figures are taken with permission from the publisher.

5.1 Abstract

Simulations of protein folding and protein association happen on timescales that are orders of magnitude larger than what can typically be covered in all-atom molecular dynamics simulations. Use of low-resolution models alleviates this problem but may reduce the accuracy of the simulations. We introduce a replica-exchange-based multiscale sampling technique that combines the faster sampling in coarse-grained simulations with the potentially higher accuracy of all-atom simulations. After testing the efficiency of our Resolution Exchange with Tunneling (ResET) in simulations of the Trp-cage protein, an often used model to evaluate sampling techniques in protein simulations, we use our approach to compare the landscape of wild type and A2T mutant Aβ_{1-42} peptides. Our results suggest a mechanism by that the mutation of a small hydrophobic Alanine (A) into a bulky

polar Threonine (T) may interfere with the self-assembly of Aβ-fibrils.

5.2 Introduction

While molecular dynamics is now commonly used to study folding, association and aggregation of proteins and other biological macromolecules,[119–127] biochemical processes such as the formation of amyloid fibers from monomers[123,127,128] often occur on timescales[128,129] that exceeds what can be covered in all-atom simulations. Coarse-graining, i.e., lowering the resolution of a system,[122,130–134] allows one to reduce the computational difficulties and to access timescales not obtainable to the fine-grained all-atom models,[122,130] but it often results in lower accuracy. This is because the smaller number of degrees of freedom lowers the entropy of the system, and it is difficult to compensate for this reduction by modifying the enthalpic contributions accordingly.[130] Multiscale techniques try to combine the advantages of fine-grained models (that are more accurate but costly to evaluate) with that of coarse-grained models (which are less detailed but enable larger time steps).

One example is Resolution Exchange[135] where the replica-exchange protocol [136] is used to induce a walk in resolution space. In the same way that for Replica-Exchange molecular dynamics (REMD)[136,137] the walk in temperature space leads to faster sampling at low temperatures, enables exploration of resolution space a faster convergence of simulations at an all-atom level.[135,138] However, the replica-exchange step requires reconstruction of the fine-grained degrees of freedom of a previously coarse-grained configuration, for instance, by adding side chains to a conformation that was described prior only by the backbone. Various approaches[75,135,138–140] have been developed to address this problem, but often they result in high energies of the proposal configuration (and therefore low acceptance rates)[75,138] or introduce biases [139, 140].

The dilemma can be alleviated by introducing a potential energy made of three terms:

$$E_{pot} = E_{FG} + E_{CG} + \lambda E_{\lambda} \ . \tag{5.1}$$

The first term is the energy E_{FG} of the protein system and the surrounding environment as described by an *all-atom* (*fine-grained*) model. The second term E_{CG} describes the same system by a suitable *coarse-grained* model. Both models are coupled by a system-specific penalty term E_{λ}[141,142] that measures the similarity between the configurations at both levels of resolution, with the strength of coupling controlled by a replica-specific parameter λ. Hence, Hamilton Replica Exchange[143,144] of the above defined multiscale system leads to an exchange of information between fine-grained and coarse-grained models, with measurements taken at the replica where $\lambda = 0$. However, while avoiding the problem of steric clashes in resolution exchange, the exchange probability is often still small,[145] and the resulting need to use multiple replica to bridge the two levels of resolution makes this approach not appealing.

As an alternative, we propose here a Resolution Exchange with Tunneling (ResET) approach that requires only two replicas. Working and efficiency of our approach is tested in simulations of the Trp-cage[146,147] miniprotein (Protein Data Bank (PDB) Identifier: 1L2Y), an often used model for testing new sampling techniques. As a first application we use in the second part ResET to compare the landscape of Aβ_{1-42} wild type peptides, implicated in Alzheimer's disease, with that of A2T mutants which seems to protect against Alzheimer's disease.[148–150] Our results suggest a mechanism by that the mutation of a small hydrophobic Alanine (A) into a bulky polar Threonine (T) may interfere with the self-assembly of Aβ-fibrils, decreasing the chance for formation of the Aβ-amyloids that are a hallmark of Alzheimer's disease.[151–153]

5.3 Resolution Exchange with Tunneling

Resolution Exchange with Tunneling (ResET) utilizes two replica, each containing both a coarse-grained and a fine-grained representation of the system. On each replica, both representations evolve separately by molecular dynamics. On the first replica, A, is the *coarse-grained model* in a configuration A_{CG} and has a potential energy $E^{pot}_{CG}(A_{CG})$ and a kinetic energy $E^{kin}_{CG}(A_{CG})$. On the other hand, the fine-grained model is in a configuration A_{FG} that has a kinetic energy $E^{kin}_{FG}(A_{FG})$ and a potential energy $E^{biased}_{FG}(A_{FG})$ which depends on the configuration A_{CG} of the coarse-grained model by $E^{biased}_{FG}(A_{FG}) = E^{pot}_{FG}(A_{FG})+\lambda_1 E_\lambda(A_{CG}, A_{FG})$. Hence, the two models on this replica interact only by the term $\lambda_1 E_\lambda(A_{CG}, A_{FG})$ that biases the fine-grained model, but are otherwise invisible to each other. The effect of this biasing term is that configurations of the fine-grained model are favored which resemble the coarse-grained model configuration, with the strength of the bias controlled by parameter λ_1. The opposite situation is found on the replica B. Here lives an independent fine-grained model with configuration B_{FG} that has a potential energy $E^{pot}_{FG}(B_{FG})$ and kinetic energy $E^{kin}_{FG}(B_{FG})$, while, on the other hand, the configuration B_{CG} of the coarse-grained model has a kinetic energy $E^{kin}_{CG}(B_{CG})$ and a potential energy $E^{biased}_{CG}(B_{FG}, B_{CG}) = E^{pot}_{CG}(B_{BG}) + \lambda_2 E_\lambda(B_{CG}, B_{FG})$ that depends on the fine-grained model by a term $\lambda_1 E_\lambda(A_{CG}, A_{FG})$. This biasing term now ensures that on replica B the coarse-grained configuration resembles the one of the fine-grained model.

While the time step for integrating fine-grained and coarse-grained models may differ, they have to be the same for the corresponding models on both replicas. This is because after a certain number of molecular dynamics steps a decision is made on whether to replace on the replica B the configuration B_{FG} in the unbiased fine-grained model by the configuration A_{FG} of the auxiliary (biased) fine-grained

model of the replica A. This replacement goes together with a re-weighting of the velocities $v_{FG}(A_{FG})$ such that $\hat{E}^{kin}_{FG}(A_{FG}) = E^{kin}_{FG}(B_{FG})$, and is accepted with probability:

$$w(B \to A) = \min\left(1, \exp(-\beta(E^{pot}_{FG}(A_{FG}) - E^{pot}_{FG}(B_{BG}) - \lambda_1 E_\lambda(A_{FG}, A_{CG}) - \Delta E^{kin}_{FG})\right) \tag{5.2}$$

with $\Delta E^{kin}_{FG} = E^{kin}_{FG}(A_{FG}) - E^{kin}_{FG}(B_{FG})$. The re-weighting of the velocities and the Metropolis-Hastings acceptance criterium accounts for the fact that the proposal configurations A_{FG} are generated on replica A by a biased process, i.e., it corrects for the resulting skewed probability with which the configuration A_{FG} is proposed as a replacement for B_{FG}.

At other times, the the coarse-grained configuration A_{CG} on replica A is replaced by the configuration B_{CG} of the biased coarse-grained model on replica B with probability:

$$w(A \to B) = \min\left(1, \exp(-\beta(E^{pot}_{CG}(B_{CG}) - E^{pot}_{CG}(A_{BG}) - \lambda_2 E_\lambda(B_{FG}, B_{CG}) - \Delta E^{kin}_{CG})\right) \tag{5.3}$$

with $\Delta E^{kin}_{CG} = E^{kin}_{CG}(B_{CG}) + E^{kin}_{CG}(A_{CG})$. Re-weighting the velocities of configuration B_{CG} such that $\hat{E}^{kin}_{CG}(B_{CG}) = E^{kin}_{CG}(A_{CG})$, and the Metropolis-Hastings acceptance criterium are again to correct for the skewed probability by which the configuration B_{CG} is proposed.

Note that the update of the unbiased coarse-grained configuration on replica A also changes the E_λ biasing term in the ancillary fine-grained configuration, as does the update of the unbiased fine-grained configuration on replica B changes the corresponding biasing term in the steered coarse-grained configuration. In order to minimize this disturbance, we also rescale the velocities in the biased models such that the change in kinetic energy compensates for the change in E_λ.

We remark that in software packages such as GROMACS[154] it is sometimes simpler to separate the biased and unbiased models onto different replicas. In this case one would have four replicas, with a possible distribution of the models sketched in the table below.

Table 5.1: ResET Model Distribution

Model	replica	Potential Energy	Kinetic Energy	Lambda	Lambda Energy
unbiased fine-grained model	0	P_0	K_0		
biased fine-grained model	1	P_1	K_1	λ_1	$E_\lambda(1,3)$
biased coarse-grained model	2	P_2	K_2	λ_2	$E_\lambda(0,2)$
unbiased coarse-grained model	3	P_3	K_3		

In this implementation, the replica 0 and 2, and replica 1 and 3, communicated during the molecular dynamics evolution of the configurations; and the ResET move replaces the configuration of replica 0 by that of replica 1, and/or the configuration on replica 3 by that of replica 2.

5.4 Material and Methods

5.4.1 Setup of the ResET Simulation

Our simulations utilize a modified version of the GROMACS[154] molecular package available from the authors. Initial tests of the working and efficiency are for the Trp-cage protein,[146,147] an often used system for evaluating new algorithms. In order to compare our simulations with previous studies, we follow closely the set-up of Han et al[155] for the coarse-grained model, and that of Kouza et al [156] for the fine-grained model. Hence, our coarse-grained Trp-cage protein model is described by PACE force-field,[122] with the uncapped protein solvated by 1118 MARTINI[157] coarse-grained water molecules, and buffered 0.15M Na^+ and Cl^- ions, in a cubic box of length 5.18 nm, leading to a total of 1313 coarse-grained particles. On the

other hand, in our fine grained model is the N-terminus capped by an acetyl group and at C-terminus by methylamine, leading to a total number of 313 atoms for the protein that are solvated with 2645 extended simple point charge (SPC/E[158]) water molecules in a cubic box with an edge length of 4.4 nm. One chlorine ion (Cl^-) is added to neutralize the system. Hence, the system contains 8249 fine-grained particles, with the interactions between them described by the AMBER94 force-field.[159]

As a first application we compare in the second part of this study the ensemble of configurations sampled by ResET simulation of Aβ_{1-42} wild type and A2T mutant peptides. While aggregates of the wild type Aβ-peptides are implicated in Alzheimer's disease, the A2T mutant appears to be protective, i.e, reducing the probability for acquiring the disease. We use in our simulations for both wild type and mutant as coarse-grained model the MARTINI force-field,[157] which is computationally efficient and has been already used earlier in Aβ simulations.[160,161] Here, the main chain of each amino acid is represented by one bead, and the side chains by up to four beads depending on the size of the amino acid. Our wild type protein thus contains 91 beads, and the mutant 92 beads. Each peptide is placed in a cubic box and solvated with the MARTINI-CG water molecules represented by single beads. Together with 3 Na^+ MARTINI-ion beads and a box size of 7.16 nm (wild type) and 7.24 nm A2T mutant) we arrive at 2925 and 3189 particles, respectively. On the other hand, the fine-grained representations of wild type and mutant peptides are modeled by the CHARMM36 force-field[162] which we found in previous work to be efficient for simulations of intrinsically disordered and amyloid-forming proteins. The N- and C-termini are capped with Methyl groups. The protein is placed in the center of a cubic box using a 1 nm distance between the atoms of the protein and box. The each system is solvated with TIP3 water molecules[163] and neutralized with 3 Na^+ ions. This leads to a box size of 7.5 nm and a total number of 41412 particles for the wild type. Correspondingly,

we get a box size of 7.6nm and a total number of 44509 particles for the mutant.

In simulations of both the Trp-cage protein and the Aβ-peptides we use for both fine-grained and coarse-grained models shift functions with a cut-off of 1.2 nm in the calculations of Coulomb and van der Waals interactions. Because of periodic boundary conditions we employe Particle mesh Ewald (PME)[164] summation to account for long-range electrostatic interactions. Hydrogen atoms and bond distances are constraint in the fine-grained model by the LINCS algorithm.[165] Equations of motion are integrated using a leap-frog algorithm, with a time step of 2 fs for both the fine-grained model and coarse-grained model. The v-rescale thermostat[166] with a coupling time of 0.01 ps is used to maintain the temperature in the coarse-grained models, while a Nose-Hoover[167,168] thermostat with the coupling time of 0.5 ps controls the temperature in the fine-grained models.

A key element of the ResET sampling technique is the restraining potential E_λ which quantifies the similarity between fine-grained and coarse-grained configurations. In our case, we choose a function of the form.[142]

$$E_\lambda(q_{FG}, q_{CG}) = \begin{cases} \frac{1}{2}\left(\Delta^2(i,j)\right) & -ds < \Delta(i,j) < ds \\ A + \frac{B}{\Delta^S(i,j)} + f_{max}\Delta(i,j) & \Delta(i,j) > ds \\ A + \frac{B}{\Delta^S(i,j)}(-1)^S - f_{max}\Delta(i,j) & \Delta(i,j) < -ds \end{cases} \tag{5.4}$$

where q_{FG} are the coordinates of atoms in the fine-grained model and q_{CG} the ones in the coarse-grained model. $\Delta(ij) = \delta_{FG}(ij) - \delta_{CG}(ij)$ is the difference between the distances ($\delta(ij)$) measured in either the fine-grained or the coarse-grained models between the C_α-atoms i and j. The control parameter f_{max} sets the maximum force as $\Delta(i,j) \to \infty$ and S determines how fast this value is realized. The parameters A and B are included to ensure continuity of $E_\alpha(q_{fg}, q_{cg})$ and it's first derivative at values where $\Delta(i,j) = \pm ds$, i.e., where the functional form of

Table 5.2: Simulation details

	Trp-cage			$A\beta_{1-42}$		
Method	Force-Field	Sampling No	Time (ns)	Force-Field	Sampling No	Time(ns)
Canonical FG	AMBER94	3	5000	CHARMM36	– – –	– – –
REMD FG	AMBER94	1	200	– – –	– – –	– – –
ResET FG+CG	AMBER94+PACE	6	200(1000)	CHARMM36+MARTINI v2.2	2	100(500)

5.4 changes. These parameters are thus computed by

$$A = \left(\frac{1}{2} + \frac{1}{S}\right) ds^2 - \left(\frac{1}{S} + 1\right) f_{max} ds \quad and \quad B = \left(\frac{f_{max} - ds}{S}\right) ds^{S+1}. \quad (5.5)$$

In the ResET simulations is the biased fine-grained model on replica A coupled to the unbiased coarse-grained model by a parameter $\lambda_1 = 0.5$, while on replica B the biased coarse-grained models is coupled to the free fine-grained models by a parameter $\lambda_2 = 2.5$. The ResET replacement move is tried every 250 ps, with the bias-correction factor $\lambda_1 E_\lambda(A_{FG}, A_{CG}) - \Delta E^{kin}_{FG}$ limited to the interval (0,100), and on replica B $\lambda_2 E_\lambda(B_{FG}, B_{CG}) - \Delta E^{kin}_{CG})$ to the interval (0,20), choices that we found in preliminary test runs leading to increased numerical stability.

Start structures for both fine-grained and coarse-grained models are generated by heating up the experimental structures of PDB-ID: 1L2Y (Trp-cage) and PDB-ID: 1Z0Q ($A\beta_{1-42}$)[151,169] to 500 or 1000 K in short molecular dynamics simulations under NVT conditions (0.5 ns and 1 ns), and cooling them down to the respective temperatures (with the exception of the REMD simulations is this 310 K). Simulations of the various systems start from the so-generated configurations and are performed in the NVT ensemble, with the simulation details listed in Table 5.2.

For most of our analysis we use GROMACS tools[154] such as gmx-rms which calculates the root-mean-square deviation (RMSD) and the root-mean-square fluctuations (RMSF) of residues with respect to an initial configuration. For visualization we use the VMD software,[170] which we also use to calculate the solvent accessible surface area (SASA) using a probe radius of 1.4 Å. Other quantities are calculated

with in-house programs and defined in the manuscript. An example are dynamic cross-correlation maps which are calculated using the definition of:[171,172]

$$C(i,j) = \frac{\langle \Delta \mathbf{r}_i . \Delta \mathbf{r}_j \rangle}{\langle \Delta \mathbf{r}_i^2 \rangle \langle \Delta \mathbf{r}_j^2 \rangle}. \tag{5.6}$$

where $\Delta \mathbf{r}_i$ and $\Delta \mathbf{r}_i$ are the displacement vectors of i-th and j-th residues of the system and angle brackets represent ensemble averages. Positive values mark correlated motions of the respective residues while negative values indicate anti-correlated motion.

5.5 Results and Discussion

5.5.1 Efficiency of ResET

In order to test the working and efficiency of our multiscale approach ResET, we perform first simulations of the Trp-cage[146,147] miniprotein, an often used model for testing sampling techniques. Choice of this system, with which we are familiar from previous work, therefore allows a direct comparison with past simulations. An example are the replica exchange molecular dynamics (REMD) simulations of Ref.,[156,173] where 40 replicas of equal volume are simulated at 40 temperatures spanning a range from T= 280 K to T=540 K. Configurations are exchanged between neighboring temperatures according to a generalized Metropolis criterium, leading to a random walk in temperature that allows replicas to find local minima (when at low temperatures) and escape out of them (when at high temperatures). The net-effect is an enhanced sampling at the target temperature. Defining a configuration as native-like if the root-mean-square deviation (RMSD) to the PDB-structure (PDB-ID: 1L2Y) of less than 2.5 Å, we find at T=310 K native-like configuration with a frequency of 87 %, using the more restrictive criterium of a RMSD smaller than 2.2 Å, the frequency reduces to 55%. Note, that these frequencies do not change beyond statistical fluctuations once the REMD simu-

lation has reached 50 ns, and we therefore neglect the first 50 ns of our 200 ns long trajectories when calculating the frequencies. While these frequencies are similar to the ones observed in earlier work,[156,173] we suspect that our values overestimate the frequency of folded configurations that reside at a certain time at T=310 K. This is because the systems are simulated at each temperature with the same volume. This volume, while sufficiently large at the target temperature may at the higher temperature suppress extended configurations, therefore artificially stabilizing folded configurations. For this reason, we prefer to compare our ResET simulations instead directly with regular constant temperature molecular dynamics, simulating the Trp-cage protein in three independent trajectories at T=310 K over 5000 ns, a value that is comparable to the experimental measured folding times of around 4μ.[174] The RMSD as function of time is shown for all three trajectories in Figure 5.1a.

Visual inspection of the three trajectories points to another problem. For a small protein such as Trp-cage is the RMSD not good measure for similarity as configurations that appear as similar by visual inspection may differ by relatively large RMSD values. This can be seen, for instance, in the second trajectory where at around 600 ns the RMSD increases from 2.0 Å to 3.6 Å, i.e., from native-like to configurations to one considered no longer native-like according to the above definition of a native configuration (i.e., having a RMSD of less than 2.5 Å). However, visual inspection shows that the molecule keeps its native-like fold, see the corresponding configurations also shown in the Figure. This contradiction between our RMSD-based definition and visual inspection made us configurations while the RMSD consider another quantity as measure for similarity. The two main characteristics of the Trp-cage native structure are its two helices (residues 2-9 and 11-14), and the contact between residues 6W (a Tryptophan) and residue 18P

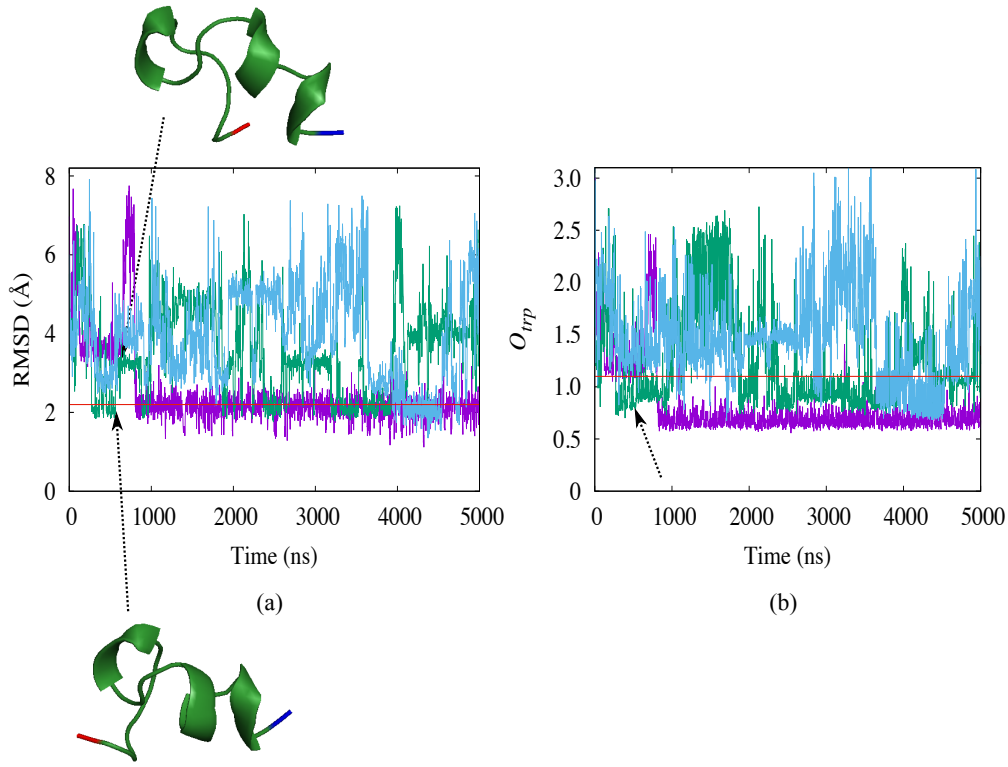

Figure 5.1: The time evolution of RMSD (a) and folding parameter O_{trp} (b) as measured in regular molecular dynamics simulations at T=310 K. Trajectory 1 is drawn in purple, trajectory 2 in green and trajectory 3 in blue. The two snapshots are taken from trajectory 2 at 601.0 ns (snapshot at the bottom) and 602.3 ns (snapshot on top). Both snapshots show similar configurations while the RMSD changes from 2.0 Å to 3.6 Å. N- and C-terminal residues in the snapshots are marked in blue and red color, respectively.

(a Proline). Hence we define as marker for Trp-cage folding a new quantity:

$$O_{trp} = d_{6-18} + 1/(n_H + 1) \tag{5.7}$$

Here, d_{6-18} is the difference between residues 6W and 18P, and n_H the number of residues that have dihedral angles as seen in a helix. The time evolution of this quantity in Figure 5.1b shows that the new coordinate allows indeed a better description of the folding transitions, as its behavior differs less from the visual inspection. Especially, we do not see for the second trajectory at 600 ns the false signal for non-native configurations that we see in the RMSD plot. Comparing O_{trp} as function of time with visual inspection of configurations along the trajec-

tories suggests that folded configurations are characterized by values of $O_{trp} < 1$, and we use in the following this definition to quantify frequencies of folded configurations.

With this definition, we observe the first folding event at t=11.6 ns (in trajectory 2), and the systems stays folded for about 600 ns before unfolding again. For trajectory 1 folding is observed at t=800 ns, and no folding is observed within 3500 ns in the third trajectory where the protein unfolds afterwards again at about 4500 ns. As a consequence, we find between 250 ns and 500 ns folded configurations with a frequency of about 26% and between 750 ns and 1000 ns, with about 49%. The frequencies increase only slowly as the simulations proceed, and between 3000 ns and 5000 ns we find native-like configurations with about 58%. The above numbers are consistent with the experimentally measured folding times of about 4μ.[174]

How does our new multiscale method fits in this discussion? The time evolution of our marker function $O_{trp}(t)$ is shown in Figure 5.2. Native-like configurations according to our criterium are observed after around 30 ns, and between 50 ns and 100 ns seen with a frequency of about 59%. The frequencies do not change much as the simulation progresses, and between 150 ns and 200 ns are native-like configurations observed with 65%. We remark that these frequencies do not depend on the choice of parameters with which we scale the λ energy contribution in the ResET update.

These frequencies for folded configurations are similar to what is seen in long-time canonical runs, but require shorter simulation times. Hence, our simulations of the Trp-cage protein indicate that our new multiscale simulation method leads indeed to an increase in sampling efficiency. If we take as criterium for the comparison the time it takes to have (on average) about 50% of configuration folded (about 800 ns for the canonical runs and 50 ns for the ResET run) we find that ResET is about 16 times faster than the canonical simulations. While the gain

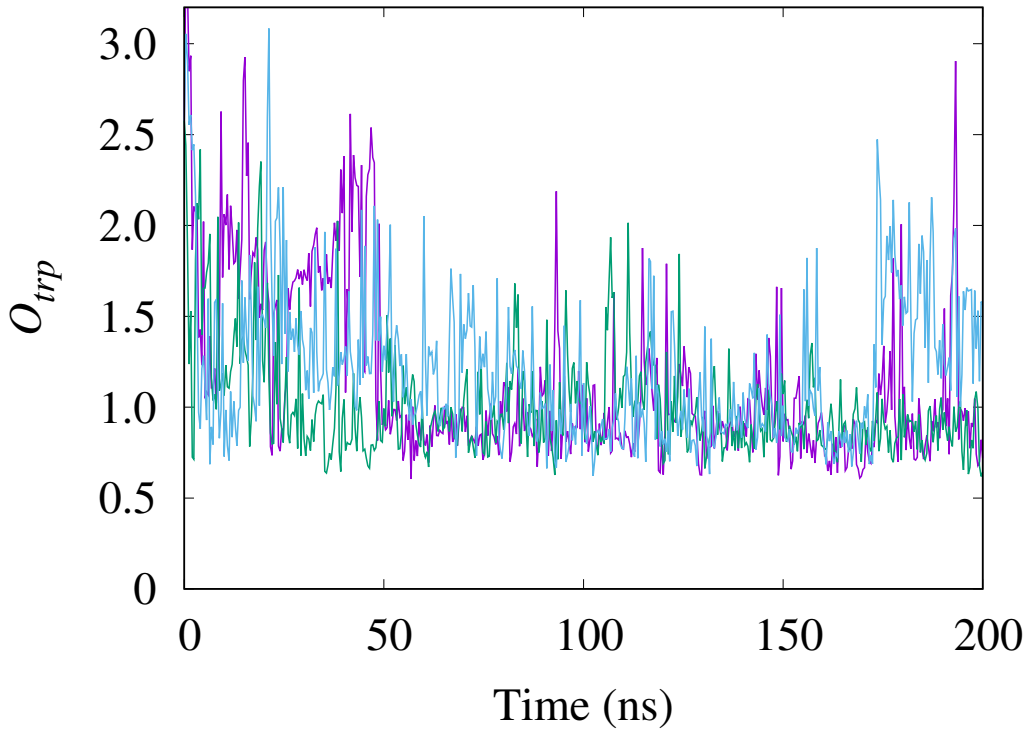

Figure 5.2: The time evolution of the order parameter for 100-20 kJ/mol. Trajectory 1 is drawn in purple, trajectory 2 in green and trajectory 3 in blue.

in efficiency will depend on the specifics of the coarse-grained model (i.e. how much faster it samples the configuration space) and its coupling to the physical force-field, our data demonstrate the faster sampling properties of our multiscale approach.

5.5.2 Comparing Aβ wildtype and A2T mutant

Our evaluation of the sampling efficiency of ResET relies on a rather simple test case. As a more interesting first application, we use in the second part our sampling technique to compare the ensembles of wild type and A2T mutant Aβ_{1-42} peptides. Fibrils containing Aβ_{1-40} or the more toxic Aβ_{1-42} are a hallmark of Alzheimer's disease and the focus of intense research.[148] A large number of familial mutations are known that worsens the symptoms of Alzheimer's disease or hasten its outbreak,[149,150] but there have been also mutations identified that are protec-

tive, i.e. lower the risk to fall ill with Alzheimer's disease. One example is the mutant A2T where the second residue (counted from the N-terminus) is changed from a small hydrophobic Alanine (A) into a bulky polar Threonine (T).[151] It has been not yet established why this mutation is protective [152, 153], but one possibility is that this mutation alters the pathway for amyloid formation, for instance, by making it more difficult to form aggregates. In order to test this hypothesis we simulate here Aβ_{1-42} wild type and A2T mutant monomers, and compare the ensembles of sampled configurations for their aggregation propensities.

Under physiological conditions are Aβ-peptides intrinsically disordered, and we do not expect the appearance of folded structures. Instead, we assume that the ensemble of configurations contains such with transiently formed β-strands that would encourage aggregation. We conjecture that such transient ordering appears more often for wild type Aβ_{1-42} than for the A2T mutant peptides. In order to identify these differences in local ordering, we have measured the root-mean-square-fluctuations (RMSF) of residues for both cases, taking as reference structure the corresponding start configuration, but discarding for the calculation of the RMSF the first 50 ns of the simulation. The RMSF is chosen because this quantity describes the flexibility of residues or segments of the protein, and the more flexible a segment is the less likely will it be involved in forming stable structures. Our data are shown in Figure 5.3, and while there are only small differences for the first 20 residues between wild type and mutant, the situation is different for the C-terminal half of the chain. For residues 21-37 is the RMSF considerably lower for the mutant than for the wild type. We remark that this picture does not change if we recalculate the RMSF, including now all heavy atoms (i.e, not only backbone but also side-chain atoms).

The lower flexibility of the segment 21-37 in the mutant is not correlated with increased secondary structure. Residues take dihedral angle values as in a helix

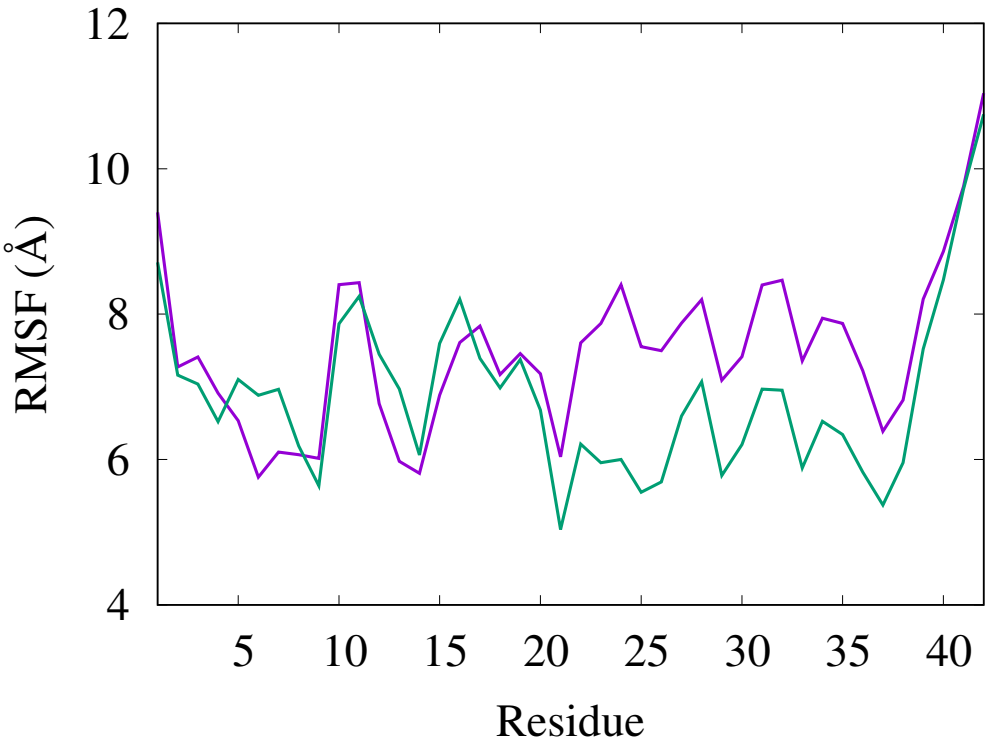

Figure 5.3: Root-mean-square-fluctuations of residues in either wild type (purple) or A2T mutant (green) Aβ_{1-42} peptides. Only heavy atoms are considered in the calculation, and the first 50 ns of the 100 ns trajectories discarded to allow for convergence of the simulations.

or a β strand with about 10% in both wild type and mutant. However, there is a change in the average radius of gyration (RGY, a measure for the volume), which with 10.6(1) Å is larger for the mutant than for the wild type where it is 10.5(1) Å. Similarly is the average solvent accessible surface area (SASA) of the peptide in the mutant with 38.0(1) nm^2 less fluctuating than in the wild type (38.0(3) nm^2), reflecting the gain in surface area resulting from the more bulky Threonine. However, the relation is different for the segment of residues 21-37, where the wild type has a SASA value of 18.2(3) nm^2 and the mutant a SASA of 18.1(2) nm^2 . The differences for the segment result from polar residues as the solvent accessible surface area of hydrophobic residues is with 4.1(1) nm^2 the same for both mutant and wild type. Hence, the differences in SASA values for this segment indicate that in the mutant polar residues, which are exposed to solvent in the wild type, form contacts with other residues. In order to understand the

differences between mutant and wild type in more detail, we have also analyzed the contacts and cross-correlations between residues, focusing again on the final 50 ns of the trajectories for both systems. The resulting maps for both systems are shown in Figure 5.4 a-b, with the coloring describing the degree of correlation between residues.

Unlike in the wild type are in the A2T mutant the disordered N-terminus (residues 1-9) and residues 27-33 correlated. This correlation results from electrostatic interactions, for instance between the NH3+ group of residue K28 (a Lysin) with negatively-charged COO- group of residue E7 (a Glutamic acid) seen in the snapshot shown in Figure 5.4 d. Hence, the replacement of the small hydrophobic Alanine by a bulky polar Threonine allows for the above electrostatic interactions in the mutant that do not exist in the wild type, and whose importance for inhibiting amyloid formation in the A2T mutant has been already noticed earlier in Ref..[175] These interactions likely stabilize not only the segment 27-33, but are responsible for the lower RMSF seen for residues 21-37. The interactions between N-terminus and residues 27-33 compete now in the A2T with hydrophobic interactions between the segment formed by residues 13-21, which include the central hydrophobic core ($L_{17}VFFA_{21}$), and the mostly hydrophobic C-terminus (residues 37-42), see the corresponding snapshot in Figure 5.4 c. As a result the two segments are correlated in the wild type but not in the mutants. These interactions between the peptide's two main hydrophobic domains are thought to be crucial for the self-assembly of Aβ-fibrils,[176,177] but are now missing in the A2T mutant, reducing the risk for aggregation.

5.6 Conclusion

We have described a replica-exchange-based multiscale simulation method, Resolution-Exchange with Tunneling (ResET), designed for simulation of protein-folding and

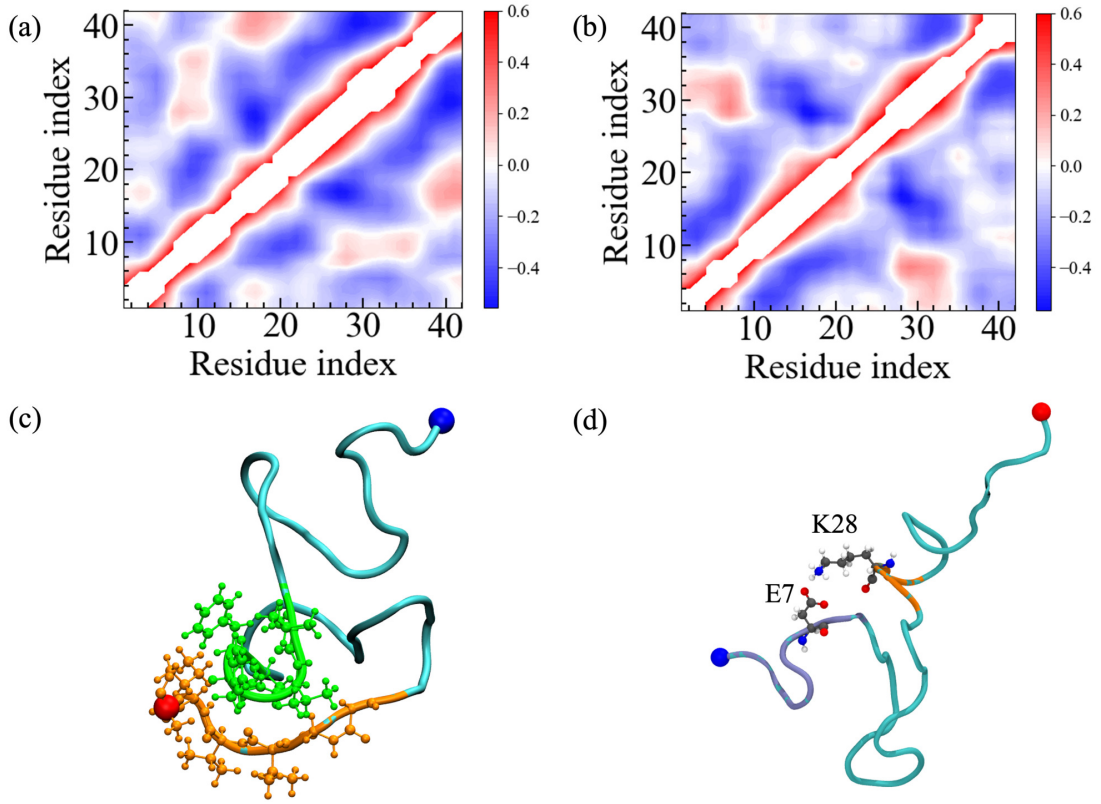

Figure 5.4: Two-dimensional dynamic cross-correlation map extracted from (a) wild type and (b) mutant Aβ_{1-42} ResET simulations. A representative snapshot obtained from the wild type simulations is shown in (c), where the central hydrophobic core $L_{17}VFFA_{21}$ and the C-terminal hydrophobic residues $G_{37}GVVIA_{42}$ are drawn in green and orange color, respectively. A corresponding snapshot from the mutant simulation is shown in (d), where the disordered N-terminus (residues 1-10) and residues 27-31 are colored in ice-blue and orange, respectively. N- and C-terminal residues are represented by blue and red spheres.

aggregation. Our approach combines the faster sampling in coarse-grained simulations with the potentially higher accuracy of all-atom simulations. It avoids the problem of low acceptance rates plaguing similar approaches and requires only few replica. After testing the accuracy and efficiency of our approach for the small Trp-cage protein by comparing our approach with long-scale (5 μs) regular molecular dynamic simulations, we use our new method to compare to compare the ensemble of Aβ_{1-42} wild type peptides, implicated in Alzheimer's disease, with that of A2T mutants which seems to protect against Alzheimer's disease. Our ResET simulations indicate that the replacement of a small Alanine (A) by a bulky Threonine (T) as residue 2 alters the pathway for amyloid formation by introducing steric

constraints on the mostly polar N-terminal residues that encourage electrostatic interactions with residues 27-33. These interactions reduce the flexibility of the extended segment 21-37, therefore contributing to the overall larger volume, more exposed surface and resulting higher solubility of the mutant. At the same time do this interactions also interfere with the hydrophobic interactions between the central hydrophobic core ($L_{17}VFFA_{21}$), and the mostly hydrophobic C-terminus (residues 37-42), known to be crucial for the self-assembly of Aβ-fibrils, decreasing therefore the chance of formation of Aβ-amyloids. Further contributing to this mechanism that may explain why the A2T mutant seems to protect the carrier against Alzheimer's disease, could be the larger exposed hydrophobic surface area that in connection with increase solubility may trigger faster degradation of the mutant. We plan to test this hypothesis by comparing the A2T mutant with suitable double mutants that interfere with this mechanism.

Chapter 6

ResET GPU

The following chapter is from unpublished work in preparation for a manuscript for publication.

6.1 Abstract

Molecular Dynamic simulations performed with Graphical Processing Units (GPU) have dramatically extend the time range for studying protein folding.[178] However, the sampling inefficacies inherent to Molecular Dynamics still remain. These barriers can be overcome with enhanced sampling methods. We present a new version of our Resolution-Exchange-with-Tunneling that has been combined with GPU computing power for improved simulation and sampling performance over traditional methods. The previously demonstrated ResET for GROMACS method has been completely redesigned as a python library for the OpenMM MD package.[60,179] The updated version is deployable on high performance computers (HPC) or personal workstations, using either CPU or GPU architectures. These improvements expand ResET applicability to larger systems and time scales, that were unfeasible using CPUs.

6.2 Reason for Revisions

The previous version of the ResET algorithm was built on the framework of the Replica-Exchange-with-Tunneling method (RET). The ordinal implementation of

RET was incorporated, at that time, the most current GROMACS version 4.6.5. The extension of the shard sub-routines from the previously proven RET algorithm simplified the initial presentation of ResET, the method was now limited by a dated GROMACS version. Rather than updating the program to a more recent GROMACS version, ResET has been rewritten as a python library for the molecular dynamic package OpenMM with the capacity to perform its replica exchange method on a single GPU. The change is a simpler implementation, that is faster and more widely applicable for folding studies.

6.3 Summary of Revisions

The redesign is now a Python library for the Molecular Dynamic package OpenMM. The implementation to a new software was done to take advantage of OpenMM's GPU capabilities and modular custom forces. These attributes have modernized ResET into a faster, more accessible and applicable method. GPU-ResET can now simulate at speeds up to 100 fold greater than the previous CPU version. Distributing the method as an add-on library removes the need for a modified installation that solely performs ResET. OpenMM custom forces tools have also expanded ResET's capabilities. Our new design permits choices of collective variable or coarse grain models for the multi-scale essential sampling part of the algorithm. We outline in this section the critical changes necessary for the improved of ResET.

While OpenMM is able to perform MD simulations with GPU's, their application with replica exchange protocols, present difficulties. Multiple GPUs can be used for a single MD simulation, but replica exchange methods are commonly assignment as 1 replicas to 1 GPU. This presents a resource barrier with n-number of identical GPUs required to operate n-number replicas in traditional parallel computing. This limitation and lock of replica exchange protocol in basic OpenMM, guided us to a serial computing method to overcome the barriers. Our approach lowers the resource requirements, as the method can be performed using a single

GPU. The design can still benefit from multiple GPUS, without overhead lost due to message interface passing (MPI).[180]

The serial method general work-follow operates by first setting an n-number of replica exchange moves to attempt. The number of attempts act as a master loop for the algorithm. During each iteration of the exchange loop, the replicas are simulated one at a time, creating exact checkpoints when completed, deleting the current replica system and loading the next replica. Once all replica systems are simulated, the current exchange loop ends, restarting the cycle. At no point during the cycle do replicas directly communicate, Plain text datafiles store all information. This includes files for storing all system parameters that enable for the repeated creation and deletion replicas, and all structural and energy information for the ResET algorithm.

The communication of the configuration bias term E_λ (discussed in Section 5) between fine-grain and coarse-grain models require the unbias models are performed first. Both unbias models types record their distance matrix into respective data files. The appropriate partner replica receive the data by loading the unbias distance map file into memory to calculate as a OpenMM Custom non-bonded force according to calculation show in Chapter 4. Thus removing any communication log or idling, while a slower replica system updates. Information to perform the exchange protocol outlined in Chapter 5 is stored in the same fashion. The algorithm evaluates the exchange with saved data files, to create a replica ID index for a who-has-what-now replica hash-table. In the case of a successful replacement, the exchange partner index is used to load accompanying checkpoint file, but setting the system parameters according to the original target replica, completing the exchange replacement. By interactions only with outputs or memory stored data, the simulations suffer less than 15% performance decrease versus MD simulation. The frequency of data output is the primary source for the lost, the exchange

procedure coast is nominal.

A Github repository is available for installation alongside an OpenMM python environments. This distribution contains all tools required to perform ResET simulations, with examples and analyst codes for processing ResET results as a user friendly platform.‘

6.4 Performance Evaluation

6.4.1 ResET Simulation Setup

In order to validate the new ResET program, we closely follow the amyloid beta (Aβ) simulations performed in our initial application with the GROMACS version 4.6.5. The same structural files for both fine-grain (FG) and coarse-grain (CG) models of Aβ_{42} model simulations were used for the validation test. The GROMACS MARTINI coarse grain model was converted into an OpenMM topology using the Martini-OpenMM library by MacCallum Lab.[181] The Charmm36 forcefield used for the GROMACS fine grain models was exported and explicitly referenced to OpenMM.[162] The Martini Aβ_{42} model was a 7.16 nm box with a 91 bead representation, containing 3 Na^+ Martini ion for a total particle number of 2925. The Aβ_{42} fine grain model N- and C- terminals were capped with acetyl and methyl groups in a 7.5 nm cubic box. The system was solvated with TIP3P water with 3 Na^+ ions for a total of 41412 particles.[182]

Due to difference in OpenMM and GROMACS exact operation and offerings, a few simulation parameters required adjusting. These variations should not cause significant deviation from the previous results but must be state with the goal to replicate previous results. The same shift function cut-offs of 1.2 nm for Coulomb and van der Waals interactions with a Particle Mesh Ewald were employed but hydrogen bond lengths were restricted using the OpenMM H-bond restraint method.

In lieu of the leap-frog integrator with the v-rescale thermostat for the CG-model and Nose-Hoover for the FG model in GROMACS, the OpenMM version used the Langevin Integrator with a collision frequency of 1 ps and Andersen Thermostat with frequency control of 0.5 ps at 310 K for both CG and FG simulations.[62,183–185] The CG model model used a timestep of 10 fs in contrast to the FG timestep of 2 fs. This allowed the coarse grain model to evolve further before broadcasting information to the FG partner. At the start of the simulation protocol, each model is prepared in separate warm-up simulations to bring the system to the correct temperature prior to production. The coupling bias is turned off during this stage but is enabled after warm-up stage using the previous factors of $\lambda_{CG} = 2.5$ and $\lambda_{FG} = 0.5$ with a bias structure correction frequency of 100 steps and 20 steps respectively. The ResET replacement attempts are performed every 250 ps. Simulations were performed using a single Intel Skylake 6130 CPU with a NVIDIA Quadro RTX6000 GPU.

6.4.2 Validation

The wild-type Aβ_{1-42} (PDBID:1Z0Q) was simulated for 100 ns using OpenMM ResET for comparison with like GROMACS ResET. In non-time series analyst, only data after 50 ns is take, this was done to only consider results after convergences, the same criteria used for the previously analyst. RMSD from the reference structures were measured for both simulations. The time series in both simulations show that the method maintains an elevated RMSD. A series average after the 50 ns convergence show the GROMACS version averaged $8\mathring{A}$, while the OpenMM GPU version averaged $10\mathring{A}$.

Though the OpenMM version averaged higher in RMSD, the difference in average RMSD falls within the deviation of the GROMACS results. Visual inspection of the 60-80 ns segment show two molecules show similar structures. The OpenMM simulation also produces a β-sheet formation between the strands, which did not occur in the GROMACS simulations.

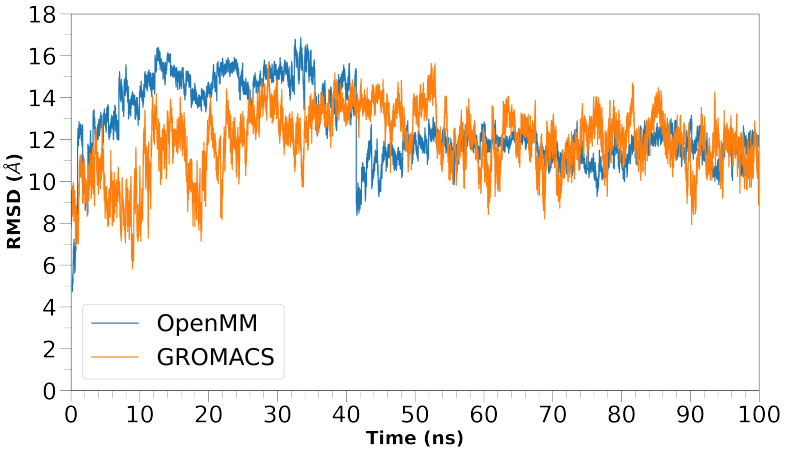

Figure 6.1: Initially higher, the OpenMM RMSD deviates less after 40 ns.

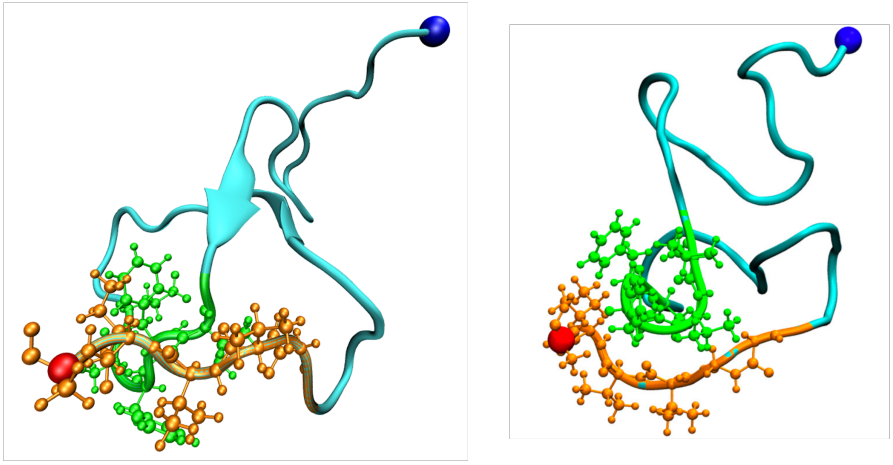

Figure 6.2: The contacts between residues $L_{17}VFFA_{21}$ and $G_{37}GVVIA_{42}$ that appear in Chapter 6 Figure 17(c),shown again on the right, occur again in the new simulations. Additional β-sheet formation also appears in the OpenMM simulation, shown on the left.

Though these simulations are not mirror results they follow the same trends as were expected in our previous study. Simulating this model in physiological conditions folded structures were not expected but rather configuration ensembles for beta strand aggregation, so the additional appearance is a positive result. The new simulations appear produce this behavior and at higher rates. We can therefore conclude our ReSET method has been successfully recreated in OpenMM and can be simulated with GPU hardware. The performance increases allows the method to be applied to complex larger systems with ease, that can additionally be fine

tuned to investigate specific questions with the updated modular design.

Chapter 7

Conclusion

7.1 Conlcusion and Outlook

The enhanced sampling methods presented in this dissertation demonstrated notable advancements in reducing algorithm difficulties that replica exchange methods face. The RET method was able to explore a multi-funnel landscape basin in the lymphotactin folding switch that simpler method could not explore. The double ladder biasing scheme showed that the switch between Ltn10 and Ltn40 relies on bifurcated hydrogen bonds to avoid the energy cost from reconstructing the hydrogen bonding network between the fold's secondary structure. The high frequencies of these interactions and rapid conversion between states would impede other methods due to the short-life spans of any intermediate states. The exchange-with-tunneling method was able to capture these moments, allowing us to propose a transition pathway.

The performance of the Resolution exchange with Tunneling was benchmarked using the mini-protein trpcage. By coupling behavior in both directions of model resolution, ResET was not able to only fold the difficult protein but also match and outperformed the accompanying cases with a fraction of simulation time. Applicability of ResET to study larger systems of research interest was also presented with Aβ proteins. The results implicated the mutation of the hydrophobic Alanine residue to the bulkier polar Threonine as a possible factor to the protective nature of the A2T Aβ mutant. This was determined by construct an energy landscape for

both variants that show divergent bonding behaviors around the mutated regions. Again the method was able to produce these results with fewer resources with shorter trajectories than canonical simulations would require.

Both methods are large contributions forward in solving protein folding problems but further advancements must be met before solved. The systems studied here, were of moderate size with only monomeric representations of the proteins. Additionally, they were performed using unoptimized computer hardware, which impeded feasible total simulation time and system sizes. The future of protein folding studies will need to consider larger systems, containing multiple protein to properly model phenomenon such as aggregation or dimerization, events that occur with the proteins studied here. The outlook for this is positive with the upcoming publication of the OpenMM GPU ResET discussed in Chapter 6, which further extents the capabilities of the sampling method to increasing complex systems. Future work will seek to develop a algorithmic method for parameter determination.

Bibliography

[1] Harvey F Lodish. *Molecular cell biology*. Macmillan, 2008.

[2] Scott Freeman. *Biological science*. Pearson education, Inc., 2008.

[3] Massimo Stefani and Christopher M Dobson. "Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution". In: *Journal of molecular medicine* 81 (2003), pp. 678–699.

[4] Philip Charles Nelson and Philip Nelson. *Biological physics*. WH Freeman New York, 2004.

[5] Christian B Anfinsen. "Principles that govern the folding of protein chains". In: *Science* 181.4096 (1973), pp. 223–230.

[6] Lauren L Porter, Irina Artsimovitch, and César A Ramırez-Sarmiento. "Metamorphic proteins and how to find them". In: *Current Opinion in Structural Biology* 86 (2024), p. 102807.

[7] Alexey G Murzin. "Metamorphic proteins". In: *Science* 320.5884 (2008), pp. 1725–1726.

[8] Prabir Khatua, Alan J Ray, and Ulrich HE Hansmann. "Bifurcated hydrogen bonds and the fold switching of lymphotactin". In: *The Journal of Physical Chemistry B* 124.30 (2020), pp. 6555–6564.

[9] Fatih Yasar, Alan J Ray, and Ulrich HE Hansmann. "Resolution exchange with tunneling for enhanced sampling of protein landscapes". In: *Physical Review E* 106.1 (2022), p. 015302.

[10] David Eisenberg and Mathias Jucker. "The amyloid state of proteins in human diseases". In: *Cell* 148.6 (2012), pp. 1188–1203.

[11] Julie S Valastyan and Susan Lindquist. "Mechanisms of protein-folding diseases at a glance". In: *Disease models & mechanisms* 7.1 (2014), pp. 9–14.

[12] Andrew R Leach. *Molecular modelling: principles and applications*. Pearson education, 2001.

[13] Brian Kuhlman and Philip Bradley. "Advances in protein structure prediction and design". In: *Nature reviews molecular cell biology* 20.11 (2019), pp. 681–697.

[14] Francis Crick. "Central dogma of molecular biology". In: *Nature* 227.5258 (1970), pp. 561–563.

[15] Matthew Cobb. "60 years ago, Francis Crick changed the logic of biology". In: *PLoS biology* 15.9 (2017), e2003243.

[16] Joel Janin et al. "Conformation of amino acid side-chains in proteins". In: *Journal of molecular biology* 125.3 (1978), pp. 357–386.

[17] Donard S Dwyer. "Electronic properties of the amino acid side chains contribute to the structural preferences in protein folding". In: *Journal of Biomolecular Structure and Dynamics* 18.6 (2001), pp. 881–892.

[18] Ken Dill and Sarina Bromberg. *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience*. Garland Science, 2010.

[19] Kerson Huang. *Lectures on statistical physics and protein folding*. World Scientific, 2005.

[20] Patrick A Alexander et al. "The design and characterization of two proteins with 88% sequence identity but different structure and function". In: *Proceedings of the National Academy of Sciences* 104.29 (2007), pp. 11963–11968.

[21] Mallika Iyer et al. “What the protein data bank tells us about the evolutionary conservation of protein conformational diversity”. In: *Protein Science* 31.7 (2022), e4325.

[22] Andrew CR Martin et al. “Protein folds and functions”. In: *Structure* 6.7 (1998), pp. 875–884.

[23] Linus Pauling. “The discovery of the alpha helix”. In: *Culture of Chemistry: The Best Articles on the Human Side of 20th-Century Chemistry from the Archives of the Chemical Intelligencer* (2015), pp. 161–167.

[24] Rajeev Aurora and George D Rosee. “Helix capping”. In: *Protein Science* 7.1 (1998), pp. 21–38.

[25] James S Nowick. “Exploring β-sheet structure and interactions with chemical model systems”. In: *Accounts of chemical research* 41.10 (2008), pp. 1319–1330.

[26] Ning Zhang et al. “New insights regarding protein folding as learned from beta-sheets”. In: *EXCLI journal* 11 (2012), p. 543.

[27] Takami Tomiyama et al. “A mouse model of amyloid β oligomers: their contribution to synaptic alteration, abnormal tau phosphorylation, glial activation, and neuronal loss in vivo”. In: *Journal of Neuroscience* 30.14 (2010), pp. 4845–4856.

[28] Yifat Miller, Buyong Ma, and Ruth Nussinov. “Polymorphism of Alzheimer’s Aβ17-42 (p3) oligomers: the importance of the turn location and its conformation”. In: *Biophysical journal* 97.4 (2009), pp. 1168–1177.

[29] Martin Karplus. “The Levinthal paradox: yesterday and today”. In: *Folding and design* 2 (1997), S69–S75.

[30] Robert J Good. “Surface free energy of solids and liquids: Thermodynamics, molecular forces, and structure”. In: *Journal of colloid and interface science* 59.3 (1977), pp. 398–419.

[31] Joseph D Bryngelson et al. “Funnels, pathways, and the energy landscape of protein folding: a synthesis”. In: *Proteins: Structure, Function, and Bioinformatics* 21.3 (1995), pp. 167–195.

[32] Peter E Leopold, Mauricio Montal, and José N Onuchic. “Protein folding funnels: a kinetic approach to the sequence-structure relationship.” In: *Proceedings of the National Academy of Sciences* 89.18 (1992), pp. 8721–8725.

[33] Ken A Dill et al. “The protein folding problem”. In: *Annu. Rev. Biophys.* 37 (2008), pp. 289–316.

[34] Christopher J Cramer. *Essentials of computational chemistry: theories and models*. John Wiley & Sons, 2013.

[35] Walter Kauzmann. “Some factors in the interpretation of protein denaturation”. In: *Advances in protein chemistry*. Vol. 14. Elsevier, 1959, pp. 1–63.

[36] Robert L Baldwin. “Dynamic hydration shell restores Kauzmann’s 1959 explanation of how the hydrophobic factor drives protein folding”. In: *Proceedings of the National Academy of Sciences* 111.36 (2014), pp. 13052–13056.

[37] Kerson Huang. *Statistical mechanics*. John Wiley & Sons, 2008.

[38] Vladimir Igorevich Arnol’d. *Mathematical methods of classical mechanics*. Vol. 60. Springer Science & Business Media, 2013.

[39] Leonor Cruzeiro. “Proteins multi-funnel energy landscape and misfolding diseases”. In: *Journal of Physical Organic Chemistry* 21.7-8 (2008), pp. 549–554.

[40] Ruth Nussinov. *Introduction to protein ensembles and allostery*. 2016.

[41] Ken A Dill. “Dominant forces in protein folding”. In: *Biochemistry* 29.31 (1990), pp. 7133–7155.

[42] Maddalena D Caiati et al. “PrPC controls via protein kinase A the direction of synaptic plasticity in the immature hippocampus”. In: *Journal of Neuroscience* 33.7 (2013), pp. 2973–2983.

[43] Sasha B Ebrahimi and Devleena Samanta. “Engineering protein-based therapeutics through structural and chemical design”. In: *Nature Communications* 14.1 (2023), p. 2411.

[44] Suzanne Hermeling et al. “Structure-immunogenicity relationships of therapeutic proteins”. In: *Pharmaceutical research* 21 (2004), pp. 897–903.

[45] Jean-Marc Dubois, Gilles Ouanounou, and Béatrice Rouzaire-Dubois. “The Boltzmann equation in molecular biology”. In: *Progress in biophysics and molecular biology* 99.2-3 (2009), pp. 87–93.

[46] Kresten Lindorff-Larsen et al. “How fast-folding proteins fold”. In: *Science* 334.6055 (2011), pp. 517–520.

[47] Eric Alm and David Baker. “Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures”. In: *Proceedings of the National Academy of Sciences* 96.20 (1999), pp. 11305–11310.

[48] Acacia F Dishman et al. “Evolution of fold switching in a metamorphic protein”. In: *Science* 371.6524 (2021), pp. 86–90.

[49] David R Riddell et al. “Impact of apolipoprotein E (ApoE) polymorphism on brain ApoE levels”. In: *Journal of Neuroscience* 28.45 (2008), pp. 11445–11453.

[50] Saeed Sadigh-Eteghad et al. “Amyloid-beta: a crucial factor in Alzheimer’s disease”. In: *Medical principles and practice* 24.1 (2015), pp. 1–10.

[51] Guo-fang Chen et al. “Amyloid beta: structure, biology and structure-based therapeutic development”. In: *Acta Pharmacologica Sinica* 38.9 (2017), pp. 1205–1235.

[52] Muralikrishna Lella and Radhakrishnan Mahalakshmi. "Metamorphic proteins: emergence of dual protein folds from one primary sequence". In: *Biochemistry* 56.24 (2017), pp. 2971–2984.

[53] Gregory S Kelner et al. "Lymphotactin: a cytokine that represents a new class of chemokine". In: *Science* 266.5189 (1994), pp. 1395–1399.

[54] Robbyn L Tuinstra et al. "Interconversion between two unrelated protein folds in the lymphotactin native state". In: *Proceedings of the National Academy of Sciences* 105.13 (2008), pp. 5057–5062.

[55] Jan Drenth. *Principles of protein X-ray crystallography*. Springer Science & Business Media, 2007.

[56] Gordon S Rule and T Kevin Hitchens. *NMR spectroscopy*. Springer, 2006.

[57] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[58] Max Born, Klaus Fuchs, and Edmund Taylor Whittaker. "The statistical mechanics of condensing systems". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 166.926 (1938), pp. 391–414.

[59] Viktor Hornak et al. "Comparison of multiple Amber force fields and development of improved protein backbone parameters". In: *Proteins: Structure, Function, and Bioinformatics* 65.3 (2006), pp. 712–725.

[60] David Van Der Spoel et al. "GROMACS: fast, flexible, and free". In: *Journal of computational chemistry* 26.16 (2005), pp. 1701–1718.

[61] Jay W Ponder and David A Case. "Force fields for protein simulations". In: *Advances in protein chemistry* 66 (2003), pp. 27–85.

[62] Michel A Cuendet and Wilfred F van Gunsteren. "On the calculation of velocity-dependent properties in molecular dynamics simulations using the leapfrog integration algorithm". In: *The Journal of chemical physics* 127.18 (2007).

[63] Ernst Hairer, Christian Lubich, and Gerhard Wanner. "Geometric numerical integration illustrated by the Störmer–Verlet method". In: *Acta numerica* 12 (2003), pp. 399–450.

[64] William C Swope et al. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". In: *The Journal of chemical physics* 76.1 (1982), pp. 637–649.

[65] Hans C Andersen. "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations". In: *Journal of computational Physics* 52.1 (1983), pp. 24–34.

[66] Berk Hess et al. "LINCS: A linear constraint solver for molecular simulations". In: *Journal of computational chemistry* 18.12 (1997), pp. 1463–1472.

[67] Vincent Kräutler, Wilfred F Van Gunsteren, and Philippe H Hünenberger. "A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations". In: *Journal of computational chemistry* 22.5 (2001), pp. 501–508.

[68] Sebastian Kmiecik et al. "Coarse-grained protein models and their applications". In: *Chemical reviews* 116.14 (2016), pp. 7898–7936.

[69] Siewert J Marrink et al. "The MARTINI force field: coarse grained model for biomolecular simulations". In: *The journal of physical chemistry B* 111.27 (2007), pp. 7812–7824.

[70] Luca Monticelli et al. "The MARTINI coarse-grained force field: extension to proteins". In: *Journal of chemical theory and computation* 4.5 (2008), pp. 819–834.

[71] Trinh Xuan Hoang and Marek Cieplak. "Molecular dynamics of folding of secondary structures in Go-type models of proteins". In: *The Journal of Chemical Physics* 112.15 (2000), pp. 6851–6862.

[72] Robert S DeWitte, Alexey V Ishchenko, and Eugene I Shakhnovich. "SMoG: de novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies in molecular design". In: *Journal of the American Chemical Society* 119.20 (1997), pp. 4608–4617.

[73] Antonio B de Oliveira Jr et al. "SMOG 2 and OpenSMOG: Extending the limits of structure-based models". In: *Protein Science* 31.1 (2022), pp. 158–172.

[74] Kei Moritsugu, Tohru Terada, and Akinori Kidera. "Scalable free energy calculation of proteins via multiscale essential sampling". In: *The Journal of chemical physics* 133.22 (2010).

[75] P. Liu et al. "Reconstructing atomistic detail for coarse-grained models with resolution exchange". In: *J. Chem. Phys.* 129 (2008), p. 114103.

[76] David Landau and Kurt Binder. *A guide to Monte Carlo simulations in statistical physics*. Cambridge university press, 2021.

[77] Charles J Geyer. "Markov chain Monte Carlo maximum likelihood". In: (1991).

[78] Ulrich HE Hansmann. "Parallel tempering algorithm for conformational studies of biological molecules". In: *Chemical Physics Letters* 281.1-3 (1997), pp. 140–150.

[79] Yuji Sugita and Yuko Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical physics letters* 314.1-2 (1999), pp. 141–151.

[80] Jinan Wang et al. "Velocity-scaling optimized replica exchange molecular dynamics of proteins in a hybrid explicit/implicit solvent". In: *The Journal of chemical physics* 135.8 (2011).

[81] Edward Lyman, F Marty Ytreberg, and Daniel M Zuckerman. "Resolution exchange simulation". In: *Physical review letters* 96.2 (2006), p. 028105.

[82] Hiroaki Fukunishi, Osamu Watanabe, and Shoji Takada. "On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction". In: *The Journal of chemical physics* 116.20 (2002), pp. 9058–9067.

[83] Fatih Yaşar, Nathan A Bernhardt, and Ulrich HE Hansmann. "Replica-exchange-with-tunneling for fast exploration of protein landscapes". In: *The Journal of Chemical Physics* 143.22 (2015).

[84] José Nelson Onuchic and Peter G Wolynes. "Theory of protein folding". In: *Current opinion in structural biology* 14.1 (2004), pp. 70–75.

[85] A Keith Dunker et al. "Intrinsically disordered protein". In: *Journal of molecular graphics and modelling* 19.1 (2001), pp. 26–59.

[86] H Jane Dyson and Peter E Wright. "Intrinsically unstructured proteins and their functions". In: *Nature reviews Molecular cell biology* 6.3 (2005), pp. 197–208.

[87] Philip N Bryan and John Orban. "Proteins that switch folds". In: *Current opinion in structural biology* 20.4 (2010), pp. 482–488.

[88] Devora Rossi and Albert Zlotnik. "The biology of chemokines and their receptors". In: *Annual review of immunology* 18.1 (2000), pp. 217–242.

[89] Robbyn L Tuinstra et al. "An engineered second disulfide bond restricts lymphotactin/XCL1 to a chemokine-like conformation with XCR1 agonist activity". In: *Biochemistry* 46.10 (2007), pp. 2564–2573.

[90] Brian F Volkman, Tina Y Liu, and Francis C Peterson. "Lymphotactin structural dynamics". In: *Methods in enzymology* 461 (2009), pp. 51–70.

[91] Marcus Thelen and Jens V Stein. "How chemokines invite leukocytes to dance". In: *Nature immunology* 9.9 (2008), pp. 953–959.

[92] Mark S Formaneck, Liang Ma, and Qiang Cui. "Effects of temperature and salt concentration on the structural stability of human lymphotactin: Insights from molecular simulations". In: *Journal of the American Chemical Society* 128.29 (2006), pp. 9506–9517.

[93] Liang Ma and Qiang Cui. "The Temperature Dependence of Salt- Protein Association Is Sequence Specific". In: *Biochemistry* 45.48 (2006), pp. 14466–14472.

[94] E Sonay Kuloğlu et al. "Structural rearrangement of human lymphotactin, a C chemokine, under physiological solution conditions". In: *Journal of Biological Chemistry* 277.20 (2002), pp. 17863–17870.

[95] Francis C Peterson et al. "Identification and characterization of a glycosaminoglycan recognition element of the C chemokine lymphotactin". In: *Journal of Biological Chemistry* 279.13 (2004), pp. 12598–12604.

[96] Xuelian Luo et al. "The Mad2 spindle checkpoint protein has two distinct natively folded states". In: *Nature structural & molecular biology* 11.4 (2004), pp. 338–345.

[97] Marina Mapelli et al. "The Mad2 conformational dimer: structure and implications for the spindle assembly checkpoint". In: *Cell* 131.4 (2007), pp. 730–743.

[98] Xuelian Luo and Hongtao Yu. "Protein metamorphosis: the two-state behavior of Mad2". In: *Structure* 16.11 (2008), pp. 1616–1625.

[99] Robert C Tyler et al. "Native-state interconversion of a metamorphic protein requires global unfolding". In: *Biochemistry* 50.33 (2011), pp. 7077–7079.

[100] Nathan A Bernhardt et al. "Simulating protein fold switching by replica exchange with tunneling". In: *Journal of chemical theory and computation* 12.11 (2016), pp. 5656–5666.

[101] Nathan A Bernhardt and Ulrich HE Hansmann. "Multifunnel landscape of the fold-switching protein RfaH-CTD". In: *The Journal of Physical Chemistry B* 122.5 (2018), pp. 1600–1607.

[102] César A Ramırez-Sarmiento et al. "Interdomain contacts control native state switching of RfaH on a dual-funneled landscape". In: *PLOS Computational Biology* 11.7 (2015), e1004379.

[103] Carlo Camilloni and Ludovico Sutto. "Lymphotactin: how a protein can adopt two folds". In: *The Journal of chemical physics* 131.24 (2009).

[104] Huiling Zhang et al. "Fibril–barrel transitions in cylindrin amyloids". In: *Journal of Chemical Theory and Computation* 13.8 (2017), pp. 3936–3944.

[105] Wooseop Kwak and Ulrich HE Hansmann. "Efficient sampling of protein structures by model hopping". In: *Physical review letters* 95.13 (2005), p. 138102.

[106] Weihng Zhang and Jianhan Chen. "Accelerate sampling in atomistic energy landscapes using topology-based coarse-grained models". In: *Biophysical Journal* 106.2 (2014), 250a.

[107] Robert B Best et al. "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles". In: *Journal of chemical theory and computation* 8.9 (2012), pp. 3257–3273.

[108] William L Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935.

[109] Jeffrey K Noel et al. "SMOG@ ctbp: simplified deployment of structure-based models in GROMACS". In: *Nucleic acids research* 38.suppl_2 (2010), W657–W661.

[110] Warren L DeLano et al. "Pymol: An open-source molecular graphics tool". In: *CCP4 Newsl. Protein Crystallogr* 40.1 (2002), pp. 82–92.

[111] Berk Hess et al. “GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation”. In: *Journal of chemical theory and computation* 4.3 (2008), pp. 435–447.

[112] Berk Hess. “P-LINCS: A parallel linear constraint solver for molecular simulation”. In: *Journal of chemical theory and computation* 4.1 (2008), pp. 116–122.

[113] Giovanni Bussi, Davide Donadio, and Michele Parrinello. “Canonical sampling through velocity rescaling”. In: *The Journal of chemical physics* 126.1 (2007).

[114] Dmitrij Frishman and Patrick Argos. “Knowledge-based protein secondary structure assignment”. In: *Proteins: Structure, Function, and Bioinformatics* 23.4 (1995), pp. 566–579.

[115] William Humphrey, Andrew Dalke, and Klaus Schulten. “VMD: visual molecular dynamics”. In: *Journal of molecular graphics* 14.1 (1996), pp. 33–38.

[116] Inigo Marcos-Alcalde et al. “MEPSA: minimum energy pathway analysis for energy landscapes”. In: *Bioinformatics* 31.23 (2015), pp. 3853–3855.

[117] EW DIJKSTRA. “A Note on Two Problems in Connexion with Graphs.” In: *Numerische Mathematik* 1 (1959), pp. 269–271.

[118] Robert C Tyler et al. “Electrostatic optimization of the conformational energy landscape in a metamorphic protein”. In: *Biochemistry* 51.45 (2012), pp. 9067–9075.

[119] Ken A. Dill and Justin L. MacCallum. “The protein-folding problem, 50 years on”. In: *Science* 338 (2012), pp. 1042–6.

[120] Ken A. Dill et al. “The Protein Folding Problem”. In: *Annu. Rev. Biophys.* 37.1 (2008), pp. 289–316. DOI: 10.1146/annurev.biophys.37.092707.153558.

[121] C. M. Dobson. "Protein folding and misfolding". In: *Nature* 426 (2003), pp. 884–90.

[122] W. Han, C. K. Wan, and Y. D. Wu. "PACE Force Field for Protein Simulations. 2. Folding Simulations of Peptides". In: *J. Chem. Theory Comput.* 6 (2010), pp. 3390–402.

[123] C. M. Chiti F. Dobson. "Protein misfolding, functional amyloid, and human disease". In: *Annu Rev Biochem* 75 (2006), pp. 333–66.

[124] James C. Stroud et al. "Toxic fibrillar oligomers of amyloid-β have cross-β structure". In: *Proc. Natl. Acad. Sci. U.S.A.* 109.20 (2012), pp. 7717–7722. ISSN: 0027-8424. DOI: 10.1073/pnas.1203193109.

[125] W. M. Berhanu, F. Yasar, and U. H. Hansmann. "In silico cross seeding of Abeta and amylin fibril-like oligomers". In: *ACS Chem. Neurosci.* 4 (2013), pp. 1488–500.

[126] Todd M. Doran et al. "Role of amino acid hydrophobicity, aromaticity, and molecular volume on IAPP(20–29) amyloid self-assembly". In: *Proteins* 80.4 (2012), pp. 1053–1065. DOI: https://doi.org/10.1002/prot.24007.

[127] M. Eisenberg D. Jucker. "The amyloid state of proteins in human diseases". In: *Cell* 148 (2012), pp. 1188–203.

[128] Y. S. Eisele. "From soluble abeta to progressive abeta aggregation: could prion-like templated misfolding play a role?" In: *Brain Pathol* 23 (2013), pp. 333–41.

[129] K. Weise et al. "Interaction of hIAPP with model raft membranes and pancreatic beta-cells: cytotoxicity of hIAPP oligomers". In: *Chembiochem* 11 (2010), pp. 1280–90.

[130] Sebastian Kmiecik et al. "Coarse-Grained Protein Models and Their Applications". In: *Chem. Rev.* 116.14 (2016), pp. 7898–7936. DOI: 10.1021/acs.chemrev.6b00163.

[131] Wei Han and Yun-Dong Wu. "Coarse-Grained Protein Model Coupled with a Coarse-Grained Water Model: Molecular Dynamics Study of Polyalanine-Based Peptides". In: *J. Chem. Theory Comput.* 3.6 (2007), pp. 2146–2161. DOI: 10.1021/ct700151x.

[132] Carol A. Rohl et al. "Protein Structure Prediction Using Rosetta". In: *Numerical Computer Methods, Part D.* Vol. 383. Methods in Enzymology. Academic Press, 2004, pp. 66–93. DOI: https://doi.org/10.1016/S0076-6879(04)83004-0. URL: http://www.sciencedirect.com/science/article/pii/S0076687904830040.

[133] A. Kolinski. "Protein modeling and structure prediction with a reduced representation". In: *Acta. Biochim. Pol.* 51 (2004), pp. 349–71.

[134] A. Liwo et al. "A unified coarse-grained model of biological macromolecules based on mean-field multipole-multipole interactions". In: *J. Mol. Model.* 20 (2014), p. 2306.

[135] E. Lyman, F. M. Ytreberg, and D. M. Zuckerman. "Resolution exchange simulation". In: *Phys. Rev. Lett.* 96 (2006), p. 028105.

[136] U. H. E. Hansmann. "Parallel Tempering Algorithm for Conformational Studies of Biological Molecules". In: *Chem. Phys. Lett.* 281.1-3 (Dec. 1997), pp. 140–50. ISSN: 0009-2614. DOI: 10.1016/S0009-2614(97)01198-6.

[137] Y. Sugita and Y. Okamoto. "Replica-exchange Molecular Dynamics Method for Protein Folding". In: *Chem. Phys. Lett.* 314.1-2 (Nov. 1999), pp. 141–51. ISSN: 0009-2614. DOI: 10.1016/S0009-2614(99)01123-9.

[138] E. Lyman and D. M. Zuckerman. "Resolution Exchange Simulation with Incremental Coarsening". In: *J. Chem. Theory Comput.* 2 (2006), pp. 656–66.

[139] P. Liu and G. A. Voth. "Smart resolution replica exchange: an efficient algorithm for exploring complex energy landscapes". In: *J. Chem. Phys.* 126 (2007), p. 045106.

[140] R. Lwin T. Z. Luo. "Overcoming entropic barrier with coupled sampling at dual resolutions". In: *J. Chem. Phys.* 123 (2005), p. 194904.

[141] Kei Moritsugu, Tohru Terada, and Akinori Kidera. "Scalable Free Energy Calculation of Proteins Via Multiscale Essential Sampling". In: *J. Chem. Phys.* 133.22 (Dec. 2010), p. 224105. ISSN: 0021-9606. DOI: 10.1063/1.3510519.

[142] Weihong Zhang and Jianhan Chen. "Accelerate Sampling in Atomistic Energy Landscapes Using Topology-Based Coarse-Grained Models". In: *J. Chem. Theor. Comp.* 10.3 (Mar. 2014), pp. 918–23. ISSN: 1549-9618. DOI: 10.1021/ct500031v.

[143] H Fukunishi, O Watanabe, and S Takada. "On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction." In: *J. Chem. Phys.* 116.20 (2002), pp. 9058–67.

[144] W.; Kwak and Ulrich H. E. Hansmann. "Efficient Sampling of Protein Structures by Model Hopping". In: *Phys. Rev. Lett.* 95.13 (2005), p. 138102.

[145] W. Nadler and U. H. Hansmann. "Dynamics and optimal number of replicas in parallel tempering simulations". In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 76 (2007), p. 065701.

[146] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. "Designing a 20-residue protein". In: *Nat. Struct. Biol.* 9.6 (June 2002), pp. 425–430. ISSN: 1072-8368. DOI: 10.1038/nsb798.

[147] C. Simmerling, B. Strockbine, and A. E. Roitberg. "All-atom structure prediction and folding simulations of a stable protein". In: *J. Am. Chem. Soc.* 124 (2002), pp. 11258–9.

[148] J. Hardy and DJ Selkoe. "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics." In: *Science* 297 (2002), pp. 353–356. DOI: 10.1126/science.1072994.

[149] Maho Morishima-Kawashima and Yasuo Ihara. “Alzheimer’s disease: -Amyloid protein and tau”. In: *J. Neurosci. Res.* 70.3 (2002), pp. 392–401. DOI: https://doi.org/10.1002/jnr.10355.

[150] Dennis J. Selkoe and Marcia B. Podlisny. “Deciphering the Genetic Basis of Alzheimer’s Disease”. In: *Annu. Rev. Genomics Hum. Genet.* 3.1 (2002), pp. 67–99. DOI: 10.1146/annurev.genom.3.022502.103022.

[151] T. Jonsson et al. “A mutation in APP protects against Alzheimer’s disease and age-related cognitive decline”. In: *Nature* 488 (2012), p. 96.

[152] J. A. Maloney et al. “Molecular mechanisms of Alzheimer disease protection by the A673T allele of amyloid precursor protein.” In: *J. Biol. Chem.* 289.45 (2014), p. 30990. DOI: doi.org/10.1074/jbc.M114.589069.

[153] I. Benilova et al. “The Alzheimer disease protective mutation A2T modulates kinetic and thermodynamic properties of amyloid- (A) aggregation.” In: *J. Biol. Chem.* 289 (2014), p. 30977.

[154] Berk Hess et al. “GROMACS 4: Algorithms for Highly Efficient, Load-balanced, and Scalable Molecular Simulation”. In: *J. Chem. Theory Comput.* 4.3 (Mar. 2008), pp. 435–47. ISSN: 1549-9618. DOI: 10.1021/ct700301q.

[155] Wei Han and Klaus Schulten. “Characterization of Folding Mechanisms of Trp-Cage and WW-Domain by Network Analysis of Simulations with a Hybrid-Resolution Model”. In: *J. Phys. Chem. B* 117.42 (2013), pp. 13367–13377. DOI: 10.1021/jp404331d.

[156] M. Kouza and U. H. Hansmann. “Velocity scaling for optimizing replica exchange molecular dynamics”. In: *J. Chem. Phys.* 134 (2011), p. 044124.

[157] Siewert J. Marrink et al. “The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations”. In: *J. Phys. Chem. B* 111.27 (2007), pp. 7812–7824. DOI: 10.1021/jp071097f.

[158] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. "The missing term in effective pair potentials". In: *J. Phys. Chem.* 91.24 (1987), pp. 6269–6271. DOI: 10.1021/j100308a038.

[159] Wendy D. Cornell et al. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules". In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197. DOI: 10.1021/ja00124a002.

[160] Mingzhen Zhang et al. "Polymorphic Associations and Structures of the Cross-Seeding of A1–42 and hIAPP1–37 Polypeptides". In: *J. Chem. Inf. Model.* 55.8 (2015), pp. 1628–1639.

[161] Adolfo B. Poma et al. "Mechanical and thermodynamic properties of A42, A40, and -synuclein fibrils: a coarse-grained method to complement experimental studies". In: *Beilstein J. Nanotechnol.* 10 (2019), pp. 500–513. ISSN: 2190-4286. DOI: 10.3762/bjnano.10.51.

[162] R. B. Best et al. "Optimization of the Additive CHARMM All-atom Protein Force Field Targeting Improved Sampling of the Backbone Phi, Psi and Side-chain Chi(1) and Chi(2) Dihedral Angles". In: *J. Chem. Theory Comput.* 8 (9) (2012), 3257–73.

[163] William L. Jorgensen et al. "Comparison of simple potential functions for simulating liquid water". In: *J. Chem. Phys.* 79 (1983), p. 926.

[164] U. Essmann et al. "A Smooth Particle Mesh Ewald Method." In: *J. Chem. Phys.* 103 (1995), pp. 8577–8593.

[165] Berk Hess. "P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation". In: *J. Chem. Theory Comput.* 4.1 (2008), pp. 116–122.

[166] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical Sampling Through Velocity Rescaling". In: *J. Chem. Phys.* 126.1 (Jan. 2007), p. 014101. ISSN: 0021-9606. DOI: 10.1063/1.2408420.

[167] Shūichi Nosé. "A molecular dynamics method for simulations in the canonical ensemble". In: *Mol. Phys.* 52.2 (1984), pp. 255–268. DOI: 10.1080/00268978400101201.

[168] William G. Hoover. "Canonical dynamics: Equilibrium phase-space distributions". In: *Phys. Rev. A* 31 (3 Mar. 1985), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695. URL: https://link.aps.org/doi/10.1103/PhysRevA.31.1695.

[169] Simona Tomaselli et al. "The -to- Conformational Transition of Alzheimer's A-(1–42) Peptide in Aqueous Media is Reversible: A Step by Step Conformational Analysis Suggests the Location of Conformation Seeding". In: *ChemBioChem* 7.2 (2006), pp. 257–267. DOI: https://doi.org/10.1002/cbic.200500223.

[170] W Humphrey, A Dalke, and K. Schulten. "VMD: Visual Molecular Dynamics". In: *J. Mol. Graph.* 14.1 (1996), pp. 33–8, 27–8.

[171] S. Swaminathan, W. E. Harte, and David L. Beveridge. "Investigation of domain structure in proteins via molecular dynamics simulation: application to HIV-1 protease dimer". In: *J. Am. Chem. Soc.* 113.7 (1991), pp. 2717–2721. DOI: 10.1021/ja00007a054.

[172] Wenhua Wang, Prabir Khatua, and Ulrich H. E. Hansmann. "Cleavage, Downregulation, and Aggregation of Serum Amyloid A". In: *J. Phys. Chem. B* 124.6 (2020), pp. 1009–1019. DOI: 10.1021/acs.jpcb.9b10843.

[173] D. Paschek, H. Nymeyer, and A. E. Garcia. "Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water". In: *J. Struct. Biol.* 157 (2007), pp. 524–33.

[174] Linlin Qiu et al. "Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4 s". In: *J. Am. Chem. Soc.* 124.44 (2002), pp. 12952–12953.

[175] P. Das, B. Murray, and G. Belfort. "Alzheimer's protective A2T mutation changes the conformational landscape of the A monomer differently than does the A2V mutation." In: *Biophys. J.* 108.3 (2015), p. 738. DOI: 10.1016/j.bpj.2014.12.013.

[176] S. Zhang et al. "The Alzheimer's Peptide A Adopts a Collapsed Coil Structure in Water". In: *J. Struct. Biol.* 130.2 (2000), pp. 130–141. ISSN: 1047-8477. DOI: https://doi.org/10.1006/jsbi.2000.4288. URL: https://www.sciencedirect.com/science/article/pii/S1047847700942886.

[177] Summer L. Bernstein et al. "Amyloid -Protein: Monomer Structure and Early Aggregation States of A42 and Its Pro19 Alloform". In: *J. Am. Chem. Soc.* 127.7 (2005), pp. 2075–2084. DOI: 10.1021/ja044531p.

[178] James C Phillips et al. "Scalable molecular dynamics on CPU and GPU architectures with NAMD". In: *The Journal of chemical physics* 153.4 (2020).

[179] Peter Eastman et al. "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics". In: *PLoS computational biology* 13.7 (2017), e1005659.

[180] Edgar Gabriel et al. "Open MPI: Goals, concept, and design of a next generation MPI implementation". In: *Recent Advances in Parallel Virtual Machine and Message Passing Interface: 11th European PVM/MPI Users' Group Meeting Budapest, Hungary, September 19-22, 2004. Proceedings 11.* Springer. 2004, pp. 97–104.

[181] Justin L MacCallum et al. "An implementation of the Martini coarse-grained force field in OpenMM". In: *Biophysical Journal* 122.14 (2023), pp. 2864–2870.

[182] Pekka Mark and Lennart Nilsson. "Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K". In: *The Journal of Physical Chemistry A* 105.43 (2001), pp. 9954–9960.

[183] Denis J Evans and Brad Lee Holian. “The nose–hoover thermostat”. In: *The Journal of chemical physics* 83.8 (1985), pp. 4069–4074.

[184] Jesús A Izaguirre et al. “Langevin stabilization of molecular dynamics”. In: *The Journal of chemical physics* 114.5 (2001), pp. 2090–2098.

[185] EA Koopman and CP Lowe. “Advantages of a Lowe-Andersen thermostat in molecular dynamics simulations”. In: *The Journal of chemical physics* 124.20 (2006).