Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Chemistry

Katrina Selavko

Protein Folding of Chignolin

Master's thesis in Chemistry
Supervisor: Titus van Erp
Co-supervisor: Anders Lervik
June 2022

NTNU
Norwegian University of
Science and Technology

Katrina Selavko

Protein Folding of Chignolin

Master's thesis in Chemistry
Supervisor: Titus van Erp
Co-supervisor: Anders Lervik
June 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Chemistry

NTNU
Kunnskap for en bedre verden

Abstract

Chignolin, a 10 residue mini-protein, is a model system for β-hairpin turns. To determine the rate and folding method of chignolin, chignolin was simulated using RETIS, a method of simulation for rare events, that tracks a reaction over an order parameter. Several order parameters are used in RETIS simulations including single distances between residues in the protein, combinations of distances between residues in the protein and the RMSD value. The order parameters were judged on whether or not chignolin was properly folded by the last interface. The rate constant was then also calculated using the PyRETIS python library used for the simulations. These were then adjusted for the non-folded trajectories in the order parameter.

The trajectories were analyzed to determine the folding mechanism. For this, plots comparing how the distances between different residues change as the transition progressed were analyzed to determine hydrogen bond formation order. In addition, principal component analysis, decision tree classification and path density plotting were used to locate any common misfolded configurations.

None of the order parameters presented in this work solely result in trajectories that properly fold. However, insights into potential order parameters are also made. The rate constant was calculated to be between $2.776 \times 10^{-4} ps^{-1}$ and $1.021 \times 10^{-5} ps^{-1}$, which is in reasonable agreement with previous experiments. The ASP3O-GLY7N order parameter resulted in 5% folded chignolin, and the ASP3O-THR8N order parameter resulted in 11% folded chignolin. The other trajectories were misfolded. Features of chignolin other than the hydrogen bonds play a large role in whether chignolin is folded or misfolded. The hydrogen bonds present in chignolin usually form around the same time, and this usually occurs at the same time or soon after the hydrophobic core is formed.

Sammendrag

Chignolin, et miniprotein med 10 aminosyrer, er et modellsystem for β-tråder. For å bestemme hyppigheten og foldemetoden til chignolin, ble chignolin simulert ved hjelp av RETIS, en simuleringsmetode for sjeldne hendelser, som sporer en reaksjon over en ordensparameter. Flere ordensparametere brukes i RETIS-simuleringer inkludert enkeltavstander mellom aminosyrer i proteinet, kombinasjoner av avstander mellom aminosyrer i proteinet og RMSD-verdien. Ordensparametrene ble bedømt på om chignolin var korrekt foldet ved slutten av overgangen i RETIS simuleringene. Reaksjonshastighetskonstant ble da også beregnet ved hjelp av PyRETIS-programmet som ble brukt til simuleringene. Disse ble deretter justert for de ikke-foldede banene i ordensparameteren.

Banene ble analysert for å bestemme foldemekanismen. For dette ble plotter, som sammenlikner hvordan avstandene mellom forskjellige aminosyrer endres etter hvert som overgangen gikk, analysert for å bestemme dannelsesrekkefølgen for hydrogenbindinger. I tillegg ble hovedkomponentanalyse, beslutningstreklassifisering og banetetthetsplotting brukt for å lokalisere vanlige feilfoldede konfigurasjoner.

Ingen av ordensparametrene som presenteres i dette arbeidet resulterer utelukkende i baner som foldes korrekt. Arbeidet inneholder imidlertid også innsikt i potensielle ordreparametere. Reaksjonshastighetskonstant ble beregnet til å være mellom $2.776 \times 10^{-4} ps^{-1}$ og $1.021 \times 10^{-5} ps^{-1}$, noe som er i rimelig overensstemmelse med tidligere eksperimenter. ASP3O-GLY7N-ordensparameteren resulterte i 5 % foldet chignolin, og ASP3O-THR8N-ordensparameteren resulterte i 11 % foldet chignolin. De andre banene ble feilfoldet. Andre egenskaper ved chignolin enn hydrogenbindingene spiller en stor rolle i om chignolin er foldet eller feilfoldet. Hydrogenbindingene som er tilstede i chignolin dannes vanligvis rundt samme tid, og dette skjer vanligvis på samme tid eller kort tid etter at den hydrofobisk kjernen er dannet.

Acknowledgements

I would like to thank my supervisors, Titus van Erp and Anders Lervik, for their support and insights throughout this project. It would not have been possible without them. Titus provided great help in understanding the theory behind simulating rare events and simulations. Anders provided great support with using PyRETIS and analyzing the data. They were both incredibly insightful in helping me understand the results obtained throughout this project and guiding me in the whole process.

Contents

Figures

Tables

Acronyms and Symbols

[0-], [0+], [1+], ... Ensemble notation for RETIS simulation.

ASP, D Aspartic acid.

GLU, E Glutamic Acid.

GLY, G Glycine.

MC Monte Carlo.

MD Molecular Dynamics.

OP Order Parameter.

PCA Principal Component Analysis.

PRO, P Proline.

RETIS Replica Exchange Transition Interface Sampling.

RMSD Root mean square deviation.

THRE ,T Threonine.

TIS Transition Interface Sampling.

TPS Transition Path Sampling.

TRP, W Tryptophan.

TST Transition State Theory.

TYR, Y Tyrosine.

Chapter 1

Introduction

Chignolin is a man-designed miniprotein. It has been the center of many studies[1–22], both experimental and simulations, due to its unique characteristics. Despite its small size, only 10 residues, chignolin can be considered a protein. It is stable in water at room temperature and exhibits a two-state transition. Chignolin was designed based on the β-hairpin turn of the popularly studied G-peptide and statistics from 100 other proteins with a similar hairpin shape [1]. This allows chignolin to act as a model of β-hairpin turns.

The folding of chignolin, like all protein folding, is a rare event. Being a rare event means that it occurs infrequently because of a high free energy barrier. When simulated with MD, the system stays unfolded for a long time, the transition happens quickly and then the system remains in the folded state for a long time. This short time of the transition and the long waiting times for transitions to occur make it computationally expensive to use techniques like MD to model the system. To see one or more transitions of chignolin folding or unfolding, the length of a regular MD simulation would need to be on the microsecond scale [19].

Because of the long simulation times need to acquire one transition in regular MD, special techniques need to be implemented to focus on simulating the transition. One of these techniques is RETIS [23], which is built upon the ideas of TIS [24] and TPS [25]. When using RETIS, an order parameter must be determined. An order parameter needs to describe state A and B in a way that there is no overlap between the two states. As the system gets more complicated, determining the order parameter becomes more complicated as well. In the case of chignolin, another research group used a combination of two hydrogen bond distances, ASP3O-GLY7N and ASP3N-THR8O, as the order parameter in a study using TPS [13].

Once the order parameter is determined, interfaces are set along the order parameter to track the progression of the reaction. RETIS uses the probabilities of crossing different interfaces set across a transition to calculate the total combined probability of the transition. RETIS, and more specifically PyRETIS [26], a python library designed for easy implementation of RETIS, has been used to study many different systems biological and otherwise. This includes proton transfer, water

autoionization, proteins binding to DNA and conformational studies of other proteins.

1.1 Chignolin

Chignolin is a small man-designed protein of 10 residues that can act as a model system for β-hairpin turns. Chignolin's residue sequence is originally based on G-peptide, the β-hairpin segment of protein GB1 domain. This was chosen because G-peptide is a well studied peptide with a two-state transition. To design chignolin, the center 8 residues in G-peptide were used as the basis of statistical analysis. A database of 8 residue sequences was collected from 100 proteins in a collection in PDB and then used to identify sequences with similar backbone geometry to G-peptide. From this, statistical analysis was performed to find the most frequent residues in each position of the sequence. Adding a terminal glycine to each side to remove any terminal charge effects gives the residue sequence GYDPETGTWG now known as chignolin [1].

Despite its small size of only 10 residues, chignolin displays many of the characteristics ascribed to proteins, rather than just a string of amino acids. Chignolin's ability to keep its structure and that it displays thermal transition indicate that chignolin is a protein. Chignolin displays a two-state transition between the folded and unfolded states, though there is a misfolded state as well [1]. It is unusual for proteins of a similar size to fold into a stable structure. This stability may indicate that β-hairpin folds could be a nucleus for folding in larger proteins β-hairpins are a part of. Some refer to chignolin as a miniprotein rather than a protein because of the small size.

Chignolin can be seen as a model system as it is similar to a large number of other β-hairpin turns. This similarity in structure to other β-hairpins opens the possibility of chignolin, or a similar protein, being an ancient protein that has evolved into many present day proteins. Another possibility is that the similar structure is a result of convergent evolution, as the sequence and similar sequences are stable [1]. Chignolin could be used as a basis to design other proteins containing β-hairpins because it is stable and folds into a defined structure. Since its creation, chignolin has been the subject of many studies. The small size of chignolin makes it easier to use in computational simulations, and its similarity to other β-hairpin turns also makes it a model system which can be used to learn about β-hairpin turns and protein folding more generally.

Chignolin is shown to exhibit a native folded state as well as a misfolded state in simulation. There is some evidence to support that this misfolded state exists in experiments, but that it can be more prevalent in simulation. This could be a result of force field properties. The native state of chignolin can be defined using several parameters, which can be seen in Table 1.1. In the native state, there are specific hydrogen bonds present. Chignolin usually exhibits hydrogen bonds between ASP3O-GLU5N, ASP3O-GLY7N, ASP3N-THR8O and ASP3O-THR8N. Hydrogen bonds between the residues ASP3-THR8 are characteristic of the native

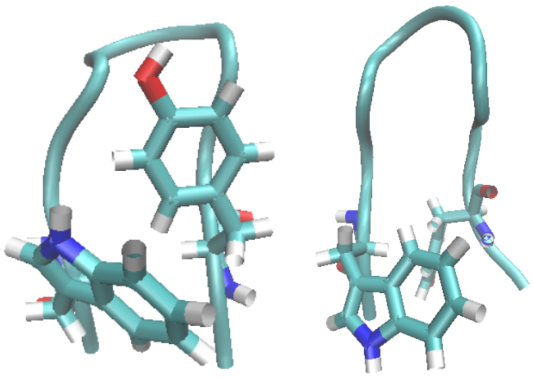


Figure 1.1: The left shows properly folded chignolin with the correct pi-turn, hydrogen bonds and hydrophobic interactions. The right shows misfolded chignolin where the hydrophobic core is not formed and instead the residues are on opposite sides of the protein. There is also a change in the shape of the backbone.

Table 1.1: Potential Reaction Coordinates from Literature

Literature Reaction Coordinates	**Folded or Misfolded**
Hydrophobic Interactions	
TYR2-PRO4	Folded [1]
TYR2-TRP9	Folded [1, 5, 7, 10, 17]
Hydrogen bonds	
ASP3O-GLY7N	Folded [1, 5, 10, 17]
ASP3N-GLY7O	Misfolded [6, 10]
ASP3N-THR8O	Folded [1, 5, 10]
THR6-THR8	Folded [6, 7]
ASP3O-THR8N	Folded [1, 5, 17]
GLU5N-ASP3O	Folded [1, 5]
GLY1O-THR9N	Misfolded [6]
GLY1O-GLY10N	Folded [6]
Turn/Angles	
π-turn ASP3 to THR8	Folded [5, 7]
α-turn ASP3 to GLY7	Misfolded [7]
Psi of GLY7	Folded [9]
RMSD	
Alpha C RMSD <0.18nm	Folded [5, 18]
Alpha C RMSD 0.18nm<x<0.32nm	Misfolded [18]

state while a bond between ASP3N-GLY7O is more likely to indicate the misfolded state. Folded chignolin exhibits a π-turn from residues PRO4-GLY7, while the misfolded state has an α-turn between residues ASP3-GLY7. Chignolin has a hydrophobic core made with the TYR2 and TRP9 residues. In the native folded state, the rings of these residues are perpendicular to each other. Folded and misfolded chignolin can be seen in Figure 1.1.

Other methods of defining the native state are the RMSD value and the fraction of native contacts. Both of these methods compare the conformation of chignolin in simulation to the conformation determined by NMR methods. A RMSD value of 18Å is used to specify the protein being in the native folded state, while a value between 18Å-22Å can indicate the misfolded state. Fraction of native contacts is another measure of similarity in structure that takes into account all atoms in residues more than 3 residues away from each other; it calculates how close atoms are to the positions expected in the reference structure. Fraction of native contacts is often a good folding coordinate for small proteins. However the fraction of native contacts may not be as reliable for chignolin due to its small size as the fraction of native contact calculation is recommended for use on proteins with at least 20 residues [27].

Chignolin has been studied by experiments [1] and simulations , including a variety of MD methods [1–11] and TIS [12, 13]. These studies give insight to how β-hairpin turns fold. There are two main theories for the method by which chignolin and β-hairpin turns fold. The zipper method and the hydrophobic collapse method. Some studies have pointed to a mixture of the two theories, or another method entirely. There is also a discussion over what part of folding is the rate determining step, the formation of the turn or the hydrogen bonds.

The zipper method would be characterized by hydrogen bonds along the length of the β-hairpin turn forming in succession. Some studies indicate that the formation of the turn in β-hairpin turns is an important step and may initiate folding [14, 15]. The formation of the turn significantly increases the chances of the protein folding. There is also evidence that suggests that interaction near the turn may effect the rate constant more than interactions at the ends of the strand, which may play more of a role in stabilization [21]. Some research also points to the hydrophobic core providing stability and acting as the zipper [22].

The hydrophobic collapse method would be characterized by a hydrophobic collapse of certain residues and then the forming of hydrogen bonds. While some studies indicate that the formation of the turn occurs first, there are also some that indicate that hydrophobic collapse occurs first [3, 14]. One study indicates that the hydrophobic core formation is more important than the hydrogen bond formation but occurs at the same time in β-hairpin turns [4].

1.2 Techniques for Modelling Chignolin

As chignolin is a model system for β-hairpin turns and given its small size, chignolin has been the subject of many studies using simulation techniques. Here a few are

quickly described. There is a short discussion of why each is or is not ideal to model chignolin.

Molecular Dynamics

Molecular Dynamics is a popular way to model systems; however, most interesting systems are out of reach for regular Molecular Dynamics simulations as a lot of computational time is used in the folded and the unfolded states waiting for a transition to occur. Especially when modeling a rare event, the computational cost adds up quickly when creating enough transitions to analyze. A way past this is to use specialized supercomputers, such as Anton which was designed and built specifically for proteins and biomolecules [28]. Chignolin has been simulated on Anton before with the data being used in conjunction with other proteins to determine the general order of protein folding steps [2]. Chignolin had to be excluded from several of the analyses, for example the native contacts analysis, due to its small size. For this project, the necessary supercomputing power was not available to be able to use MD for most of the simualtions.

Replica Exchange MD

A modified version of MD, Replica Exchange MD (REMD) has also been used to study proteins including chignolin [8, 9, 29]. REMD combines both MD simulations with Monte Carlo methods. In this method, several parallel simulations are run of the same system using either different temperatures or different Hamiltonians. Each replica is then swapped with a probability based on Metropolis rules to create a general path ensemble for the system. The swapping between different temperature/Hamiltonian replicas allows the system to avoid getting stuck on one side of high energy barriers and allows a more complete look at path space [30]. REMD does not require previous knowledge of the system, but it can be an expensive method if the system is not suited for it and can lose information that is temperature dependant.

Coarse-Grained Modeling

Coarse-grained modeling follows the idea that not every atom is important in determining the collective variables from biological systems. The number of degrees of freedom in a model are reduced, and as a result, the MD simulations are then sped up. However there is not one set way in which to create a coarse-grained model. Researchers using this method must be sure to preserve the system's important features and those they wish to study [31, 32]. Coarse-grain modelling also effects time in the simulation, which can alter the kinetics of the system; however, this can be fixed with a scaling factor [33]. Chignolin has been studied using coarse-grained modeling [10]. For chignolin, there are several factors that complicate using coarse-grained modeling, including the prior knowledge needed

for this simulation method to give good results. Additionally, chignolin is already a quite small molecule and a lot of detail could be lost.

Metadynamics

Metadynamics can be used with a system with a free energy surface (FES) that is characterized by several local minima separated by high energy barriers, as is often the case with rare events. In this method, the minima are 'filled', a process which allows the FES to be explored [34]. It can be difficult to choose a order parameter for Metadynamics without prior knowledge of the system and how it folds [29], which is not available in this case.

Markov State Model

In Markov State Modeling many comparatively short MD trajectories are produced and then joined together. These small trajectories can be produced by for example a system like Folding@home [35]. This method using Folding@home has been used to study chignolin specifically before [11]. Each simulation is begun at a different configuration. Each frame is taken as a discrete state and since there is overlap between the trajectories, the pathway between two states can be determined as well as the transition probability. Thermodynamic and kinetic information about the system are also available [36].

Transition Path Sampling

Transition path sampling is the precursor to RETIS, which is used in this study. TPS methods have been used to study β-hairpins before, with one-way shooting [12] and the weighted ensemble method [13]. In one of the TPS experiments of chignolin, two hydrogen bond distances are used as the order parameter [13]. TPS and RETIS do not necessarily require prior knowledge of the system to begin. RETIS also allows for more efficient calculation of the rate constant and the use of swapping moves to increase the efficiency of the sampling.

1.3 Research Question

The aim of this study is threefold. First is to determine to determine a good order parameter to use for RETIS simulations of chignolin. This is complicated by the factors mentioned above including that state A and B must be well defined, which is difficult for biological systems like proteins. At the moment there is not a general order parameter available for chignolin or other small β-hairpin turns.

A second goal is to determine the rate constant of chignolin folding with RETIS. RETIS allows for the calculating of the rate constant to be more efficient. This is due to the replica exchange move and the flux calculation from the [0-] and

[0+] ensembles. With RETIS, the rate constant should be able to be calculated in fewer simulation cycles than when using TPS or TIS more generally.

The final goal is to describe the folding mechanism of chignolin and give insights into the differences between folded and unfolded chignolin. The order in which the hydrogen bonds form will for example be looked at to see if either the zipper or hydrophobic collapse theories can be supported. Additionally analyzing the differences more generally can help inform determining a good order parameter as well.

To accomplish these goals in this paper, the structure is set as follows. First, a background on molecular modeling, including RETIS and related methods will be given. Then a short description of the method used in this study is followed by the results. A discussion comes with the results before the thesis is concluded with a summary, including suggestions for future direction.

Chapter 2

Theory

2.1 Molecular Dynamics

Molecular Dynamics (MD) simulations is one common way to simulate molecular systems. In MD, atomistic information of dynamical processes is determined for the system. MD can be used to determine equilibrium and transport properties. To complete a MD simulation, several steps are used. A system is set up and then Newton's equations of motion are solved. When the system is equilibrated, then the desired quantities can be measured through calculations from the positions and momenta of the particles in the system [37]. For example, the temperature of the system can be calculated as follows:

$$T(t) = \sum_{i=1}^{N} \frac{m_i * v_i^2(t)}{k_B * N_f}, \tag{2.1}$$

where m_i is mass, v_i is velocity, k_B is the Boltzmann constant and N_f is the number of degrees of freedom.

To begin a MD simulation parameters must be set and the system should be initialized [37]. These parameters include the initial temperature and density of the system, the number of particles and the time step to be used. To initialize the system, the initial positions and velocities of each particle is chosen. The initial positions used depend on the system to be simulated, but should be possible and not include any overlap between atoms. The velocities are chosen from a uniform distribution and then are shifted and scaled so that the kinetic energy is as expected for the system. The forces on each atom are then computed which is the most computationally heavy part of a MD simulation. For this the interactions between every pair of atoms within a certain cut-off range is calculated. Here periodic boundary conditions are used.

Once all of the forces are calculated, the equations of motion must be integrated. There are several algorithms that can be used to do this; a popular choice is the velocity-Verlet algorithm. The integrator used must follow several specifications, it must be area-preserving and time reversible as well as it should be accur-

ate for long time steps and use little memory. Time reversible means that future and past phase space coordinates play an equal role in the algorithm. In the case of an algorithm that is not time reversible, reversing the momenta would not trace back the original trajectory. Area preserving indicates that an algorithm preserves the volume of a system. A non-area preserving algorithm would over time expand the volume of the system, which would affect the energy conservation of the system. An algorithm that can only use short time steps requires the forces to be calculated more often, which increases the computational cost. The averages of the desired quantities can then be calculated by averaging over time [37]. This averaging over time is important because of Lyapunov instability. Lyapunov instability results in two systems that start very similarly to be vastly different after a short simulation time, or approximately 1000 MD steps.

2.2 Monte Carlo Importance Sampling

Monte Carlo (MC) is another method used to simulate molecular systems. In Monte Carlo simulations, states are generated based on Boltzmann distribution rather than reproducing the dynamics of a system like in MD. A series of steps are repeated to complete a Monte Carlo simulation.

The first step in MC is to create a configuration and calculate the energy. Next a random displacement is made to create a trial configuration of the system. The energy for this new configuration is then also calculated. The trial move is accepted or rejected, using for example the Metropolis scheme [37].

Importance sampling is the idea of sampling mostly where the Boltzmann factor is large and not where it is negligible; this creates a sampling according to statistical weight. There are also a few requirements for MC, including that it must fulfill detailed balance, must be ergodic and must be markovian. Detailed balance means that the probability of making the new path form the old path is the same as the probability of making the old path from the new path. Ergodicity is when all points in space are able to be reached in a finite number of steps. Markovian means that there is no memory of previous moves.

2.3 Modelling Rare Events

2.3.1 Transition State Theory

Rare events are processes that occur infrequently, usually due to a high free energy barrier or entropic bottleneck. In a rare event, the time spent in transition is much shorter than the time spent before or after the transition. This can cause issues when simulating the system to study the transition. Using regular molecular dynamics, which follows a path the molecules could expect to move in, the transition would be sampled rarely with most of the time spent sampling the states present before and after the transition occurs. To be able to gather enough data

about the transition itself, the simulation would have to be very long. This results in most interesting systems being out of reach for MD simulations [38].

Transition State Theory (TST) is the basis on which many methods for modeling rare events are built upon. Introduced in the 1930s by Wigner and Eyring, the main idea of TST is that the transition from reactant to product always follows a path on the free energy surface that goes through the transition state. The transition state is a multidimensional surface in the free energy profile at the local maximum. The transition state is equally likely to go to the reactant state as the product state. This method can be used to calculate the rate constant, or transition probability, which can then be compared with experimental findings. The rate constant thus allows a comparison between simulations and experiments.

An assumption in TST is that once the transition state has been crossed, the transition state will not be recrossed. However, this is often not the case. Recrossings can occur for a variety of reasons such as the chosen reaction coordinate not describing the reaction well. The reaction coordinate is chosen to differentiate between the reactants and products. A correction, or the transmission coefficient, can be applied to TST to make its calculations more accurate. This transition coefficient was first developed by Keck. Bennet, Chandler and several others also developed methods of calculating this transmission coefficient [37].

A two step method is now used to calculate the rate constant using TST. The first step is to calculate the free energy as a function of the chosen reaction coordinate and then the second step is when the transmission coefficient is calculated. The transition state is where the free energy maximum is. Starting trajectories from the transition state and following them to see whether they end in the reactant or product state results in the transmission coefficient. While the transmission coefficient allows the rate constant to be corrected, this method's efficiency is highly dependent on a good choice for reaction coordinate which can be difficult to identify especially for complex systems like proteins [39].

2.3.2 Transition Path Sampling

Transition Path Sampling (TPS) gathers the potential transition pathways from a reactant, state A, to a product, or state B. To do this Monte Carlo importance sampling is used in trajectory space rather than configuration space [40]. Once the collection of paths, or path ensemble, is created, information about the transition, like rate and mechanism can be analyzed.

In TPS, instead of using the reaction coordinate that describes the dynamics of the transition, an order parameter is used. The order parameter must describe state A and state B. Additionally, the order parameter must be descriptive of the entire states A and B, and there must be no overlap between them.

The path ensemble is defined as: $X = [x_0, x_1, x_2, ..., x_L]$, where the continuous path has been discretized into L time slices. The probability of finding a given path can be given as the sum of the probability of initial conditions and the probability of reaching each of the next time slices.

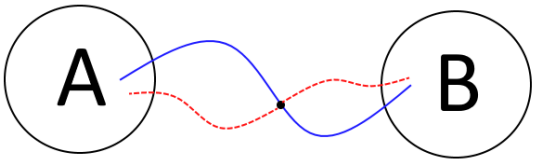


Figure 2.1: The initial path is shown in blue. A time slice, represented with the black dot, is chosen along the path. A change in position or momentum is given. The new path (red dashed line) is produced from the changed time slice by integrating the equations of motion forwards and backwards in time. The new path does not necessarily connect states A and B as shown here.

$$P[X] = \rho(x_0)\prod_{i=0}^{L-1} p(x_i \to x_{i+1}) \tag{2.2}$$

This probability should then be further constrained to only the paths that connect state A and state B by using characteristic functions $h_A(x)$ and $h_B(x)$. The characteristic functions are unity if x is in state A or state B, respectively, and are otherwise 0. This results in a path having no weight if it does not start or end in state A or B. This results in the probability of:

$$P_{AB}[X] = h_A(x_0)P[X]h_B(x_L)/Z_{AB}, \tag{2.3}$$

where Z_{AB} is a normalization factor [40].

Using this probability, a path ensemble can be created using Monte Carlo importance sampling. Often a technique called shooting is employed, but new moves are regularly developed to improve efficiency. A visual description of the shooting move can be seen in Figure 2.1. To begin, a path that connects states A and B is used. This path can be created using a long MD simulation, a high temperature MD simulation or by applying a series of constraints onto paths leaving state A. The random walk obeys detailed balance, which means that the probability to generate and accept the new path from the old path is equal to that of generating and accepting the old path from the new path.

A time slice from the old path is chosen and then modified randomly in its position and velocity. This step is either accepted or rejected based on the energy difference between the shooting point before and after modification. Rejecting a move here saves computation time as the trajectory generation is the expensive part.

If the energy difference is accepted, then the equations of motion are integrated forwards and backwards in time until the path reaches a certain length. This can be a fixed or a variable length, n. Variable length paths are more efficient as the path generation can end when the path reaches state A or B rather than continuing for a set amount of steps regardless. The move as a whole is accepted or rejected. The probability of the whole move can be calculated with the following:

$$P_{acc}[x^o \rightarrow x^n] = h(x^n) \times min[1, \frac{\exp(-\beta E(x^n_{shoot}))}{\exp(-\beta E(x^o_{shoot}))}] \times min[1, \frac{n^o}{n^n}], \tag{2.4}$$

where $\beta = \frac{1}{k_B T}$.

To improve efficiency even further, the last part of the equation, $min[1, \frac{n^o}{n^n}]$, can be used to calculate the maximum acceptable length before the integration, and then this can be used as a cutoff. If at any point in the process the move is rejected, the old path is recounted.

To calculate the rate using TPS, the correlation function must be calculated. The correlation function oscillates before plateauing. The plateau value is used to calculate the rate constant.

$$k_{AB} = \frac{d}{dt} C(t) = \frac{< h_A(x_0) h_B(x_t) >}{< h_A(x_0) >}, \tag{2.5}$$

where $< h_A(x_0) h_B(x_t) >$ is the conditional probability to be in B at time t given that you were in A at t=0. This is a very small number that is calculated by combining path sampling and umbrella sampling.

2.3.3 Transition Interface Sampling

Transition Interface Sampling (TIS) is a more efficient way to calculate the rate constant. TIS is more efficient than TPS due to Effective Positive Flux. In TPS, the rate constant oscillates and is a combination of positive and negative fluxes. By introducing overall state *A* and overall state *B*, this is avoided in TIS.

Overall state ***A*** is everything inside stable state A as well as all the phase points that are more recently in state A when the equation of motions are integrated backward. Overall state ***B*** is therefore everything in stable state B and everywhere else in phase space that was more recently state B. The characteristic functions and correlation function are also updated to include the overall states.

$$C(t) = \frac{< h_{\boldsymbol{A}}(x_0) h_{\boldsymbol{B}}(x_t) >}{< h_{\boldsymbol{A}}(x_0) >} \tag{2.6}$$

As a result of the use of the overall states and, therefore, the lack of fluctuations in the equation, the rate constant can be derived from the correlation function at t=0.

Similarly to TPS, there is an order parameter that separates state A and state B. With TIS, more so than in TPS, the order parameter must describe the stable states well as there is no cancellation of terms in the rate constant equation [24]. Additionally, here intermediate interfaces are introduced. Phase space is broken up by a set of interfaces. Using these interfaces, the crossing probability is calculated as a conditional probability.

$$k_{AB} = f_A P(\lambda_B|\lambda_A) = f_A \prod_{i=0}^{n-1} P_A(\lambda_{i+1}|\lambda_i) \tag{2.7}$$

Here f_A is the flux of trajectories through the first interface. This is usually calculated with a long MD run. $P_A(\lambda_{i+1}|\lambda_i)$ is calculated by generating trajectories that start in state A and end in state A or B with a crossing with the λ_i interface. $P_A(\lambda_{i+1}|\lambda_i)$ is the percentage of those paths that also crosses the λ_{i+1} interface. It is the conditional probability that a trajectory will cross a given interface given that it has crossed the first interface. Each individual probability of crossing an interface is higher than the overall crossing probability and using the factorization lowers the computational cost [38].

The efficiency of this method can be improved by placing the interfaces so that approximately 20% of paths in the [i+] ensemble cross λ_{i+1}. Still, TIS has evolved into several new methods, including Partial Path TIS (PPTIS) and Replica Exchange TIS. PPTIS is designed for diffusive processes with flat, wide free energy barriers. As a result the trajectories produced are much shorter and assume that the memory of the trajectory is lost at each interface. This makes PPTIS approximate [41]. For these reasons, RETIS is a better choice for studying chignolin than PPTIS.

2.3.4 Replica Exchange Transition Interface Sampling

Replica Exchange Transition Interface Sampling (RETIS) is one of many methods that build upon TIS. The two major improvements from TIS to RETIS is using a [0-] ensemble to calculate the flux and the introduction of the swapping move where trajectories are swapped between different path ensembles.

In RETIS, instead of using a long MD simulation to calculate the flux, a [0-] ensemble is created instead. The [0-] ensemble is made up of paths that start on the left-most interface and then proceed in the direction opposite of the reaction before returning to the left interface. the [0+] ensemble contains all of the paths that cross from the left to he right interface and those that cross the second interface before returning to the first (left) interface. The flux is calculated from the average path lengths of the [0-] and [0+] ensemble with the equation shown below [41].

$$f_A = \frac{1}{< t^{[0^+]} > + < t^{[0^-]} >} \tag{2.8}$$

RETIS can find multiple reaction channels because of the swapping move. RETIS does not use different temperature simulations to swap between, but rather exchanges trajectories from different path ensembles. For example, a trajectory that crosses the [2+] ensemble interface may also cross the [3+] ensemble interface and vice versa. This swapping move increases efficiency and allows several reaction channels to be found. Additionally, when the order parameter is non-

ideal, correlation between the paths can be less of a problem than with other simulation methods [41].

RETIS was chosen as the simulation method for this study for several reasons. First, RETIS is more efficient and faster than TPS and TIS [41]. RETIS is also less sensitive to the order parameter chosen. A poorly chosen order parameter will take more cycles to accurately determine the rate constant, but has less effect overall on RETIS than other simulation methods. The flux calculation accounts for some mistakes in the order parameter as well. Lastly, RETIS does not change the dynamics of the system, like some of the other methods previously discussed. This gives a clear picture of the transition of unfolded to folded chignolin.

Chapter 3

Method

3.1 Initial Molecular Dynamics Simulations

The structure of chignolin was downloaded from the Protein Data Bank (PDB ID 1UAO). The PDB structure was acquired using NMR and is in the folded configuration. The system is at a neutral pH, with the ASP and GLU residues negatively charged and two sodium atoms added to the water to balance out this charge [5–7, 42]. The C and N-termini were left unprotected [1, 5]. The force fields chosen for the experiment were the OPLS-AA/M force field [43] and the CHARMM27 force field [44] as they are common force fields for protein simulations. There is also evidence that the CHARMM27 force field decreases the frequency at which the misfolded configuration of chignolin is formed due to the glycine properties of this force field [9].

As transition path sampling approaches require an initial trajectory, regular molecular dynamics simulations were performed using GROMACS [45]. To prepare the system for each force field, the NMR structures were solvated in water with two sodium ions added to balance out the negatively charged residues in chignolin. They were then energetically minimized and then equilibrated under a NVT ensemble and then a NPT ensemble. Periodic boundary conditions were used. The coulombic interactions interactions were obtained using particle mesh ewald method. The cutoff for short-range coulombic interactions was 1.0nm; the short-range van der Waals cutoff was also set to 1.0nm. The temperature was coupled to 300K using the velocity rescaling themostat, a modified Berendsen thermostat, with a coupling parameter of 0.1ps, and pressure was coupled using the isotropic Parrinello-Rahman barostat with a coupling parameter of 2.0ps. Dispersion corrections were applied to the pressure and the energy terms to account for the cutoffs. The time step was set to 2fs. After the system was prepared, production MD was run for the two force fields. MD simulations were run for 5ns at 300K, 400K, 500K and 600K. For each of these, the temperature and pressure was adjusted accordingly while the other parameters remained the same.

3.2 Determining Potential Order Parameters

Using the information from the initial MD simulations and information collected from literature, suitable potential order parameters were determined. Plots of the potential order parameter vs time step were made for the MD simulations. The 300K and 500K MD simulations were used for the analysis to compare how each changed over time. Principal component analysis (PCA) of all of the distances between residues was also done for the 500K MD simulation, as well as simulations at 300K started from the folded and unfolded configuration, to determine which distances were the most important for distinguishing between folded and unfolded chignolin. Decision tree classification was also used to suggest combinations of factors that could be used as an order parameter.

3.3 RETIS Simulations

All simulations were run on a computer cluster (Idun at the NTNU [46]) using the PyRETIS python library [26]. The initial 500K MD simulation was used as the initial trajectory in the PyRETIS simulations. PyRETIS uses GROMACS to run the molecular simulations. The Gromacs engine available in PyRETIS was originally used before switching to the Gromacs2 engine in PyRETIS to increase simulation speed. The differences between these engines the way GROMACS and PyRETIS interface.

1000 cycles were run using only shooting moves to remove some of the effects of the initial trajectory being produced at 500K. This could have been increased to completely remove memory of the initial trajectory. Then the simulations were begun again from the last accepted trajectory, this time introducing swapping and time-reversal moves. The frequency of these was set to 0.5. Example RETIS input files can be seen in Appendix A.

The PyRETIS Analysis tool was used to monitor the simulations and calculate the rate constant. The order parameter interfaces were adjusted separately for each force field as necessary to improve efficiency. The aim was to have the average [0-] ensemble path length be approximately 4 times longer than the average [0+] ensemble path length and approximately 20% of paths in the [i+] ensemble cross λ_{i+1} (between 10% and 50%). An overview of the order parameters and interfaces discussed in this thesis can be found in Table 3.1. The force field is indicated by C for CHARMM27 and O for OPLS-AA/M. Additional interfaces were used throughout this project to arrive at the ones mentioned and discussed here.

3.4 Analysis of Data

3.4.1 Analysis of Order Parameters

Analysis on the trajectories from the RETIS simulations included checking if the trajectories that crossed from the left to the right interface where in fact correctly

Table 3.1: Overview of Order Parameters and Interfaces

Order Parameter	Force Field	Interfaces
ASP3O-GLY7N	C/O	[-0.5, -0.47, -0.35]
	C/O	[-0.5, -0.47, -0.38, -0.3]
ASP3O-THR8N	O	[-0.57, -0.5, -0.4, -0.3]
	C	[-0.55, -0.4, -0.3]
	C/O	[-0.5, -0.4, -0.3]
ASP3N-THR8O	C	[-0.43, -0.34, -0.28]
ASP3O-GLY7N/TYR2-TRP Add	C/O	[-0.85, -0.72, -0.58]
	C/O	[-1, -0.72, -0.58]
ASP3O-GLY7N/TYR2-TRP If	C/O	[D-G(-0.5, -0.4, -0.3), Y-W(-0.4)]
ASP3O-THR8N/ASP3N-THR8O	C/O	[-0.5, -0.3, -0.1]
	C	[-0.5, -0.45, -0.3, -0.1]
ASP3O-GLY7N/ASP3N-THR8O Mult	C/O	[-0.5, -0.3, -0.09]
ASP3O-GLY7N/ASP3N-THR8O If	C	[D-G(-0.6, -0.47, -0.38, -0.3), D-T(-0.29)]
RMSD	C	[-0.4, -0.35, -0.29, -0.2, -0.14, -0.1]
	O	[-0.4, -0.35, -0.28, -0.2, -0.15, -0.1]

folded by the end of the trajectory. This was done by visual inspection with VMD [47] as well as graphing the RMSD values and different distances in the molecule to see how the trajectories produced by PyRETIS with the order parameter compared to what is expected from folded chignolin, which was obtained from a 300K MD GROMACS simulation. This served as a method of evaluating the quality of the order parameter to see whether or not the order parameter separated state A and state B completely.

Additionally, graphs were created to look at the [0-] ensemble paths, or the path that are started from the leftmost interface and are supposed to explore the unfolded area before returning to the interface. This is to determine if the sampling was from state A to state B (unfolded to folded) or if it was entirely in state B, or always folded. The interfaces for an order parameter must allow exploration of both state A and state B for the best sampling. This information was used to adjust the interfaces and the order parameters.

3.4.2 Analysis of Folding Pathways

The trajectories that folded were collected and used to analyze the bond formation order. The distance between the atoms involved in the hydrogen bonds in the backbone of chignolin and the RMSD values were plotted against simulation time to see the order the bonds formed and if this affected the RMSD value. This gives insight into the folding mechanism and if the zipper or hydrophobic collapse folding method is more likely to occur. Path density plots of different variables were also made for several different order parameters. This was used to find patterns in how chignolin folds and to see if there were any conformations that occur regularly while chignolin is folding. The folded chignolin and unfolded chignolin conformations were then compared using PCA. Decision tree classification was also used to separate properly and not properly folded chignolin. These analyses allow for insights into the structure of chignolin.

Chapter 4

Results and Discussion

4.1 Order Parameter

Through a literature search and analysis of MD simulations, including principal component analysis (PCA) and decision tree analysis, potential order parameters were determined. In the Table 1.1, many of the potential reaction coordinates mentioned in literature are listed.

Plots of the potential order parameters vs time step were made from the initial MD simulations. Two 5ns long MD simulations at 300K, starting in a folded and a unfolded configuration as well as a 5ns long 500K MD simulation that exhibits a transition were used for this. A sample of these plots for the hydrogen bonds and other specific distances mentioned in literature are found in Figure 4.1. As can be seen in the graphs for the ASP3O-GLU5N and THR6-THR8 distances, not every bond in the molecule would be useful as an order parameter. An order parameter must separate the folded and unfolded states completely. Overlap in the distances between the atoms, as seen when the green and blue lines cross in the figure, means there is not separation between the folded and unfolded state on this axis, and PyRETIS would not be able to use this to tell the folded and unfolded configurations apart.

PCA can show which variables are positively and negatively correlated to the classification between categories, here folded and unfolded chignolin. The distances between the alpha carbons in chignolin and the RMSD values were used for this analysis. The scores and loadings can be seen in Figure 4.2. The distance between residues ASP3-GLY7 and TYR2-GLY7 as well as the distances between residues TYR2-THR8 and ASP3-THR8 are highly correlated. Of the highly correlated distances between residues, ASP3-GLY7 and ASP3-THR8 were chosen to be used as order parameters in this study due to their mention as important hydrogen bonds in literature. These distances also explain some of the variance in folded and unfolded chignolin along with the distance between residues ASP3-THR6 and PRO4-GLY7, the distance between the residues in the hydrophobic core, TYR2-TRP9, and the RMSD value. Further classification was performed by creating decision trees with the python library sci-kit learn. As can be seen in the decision

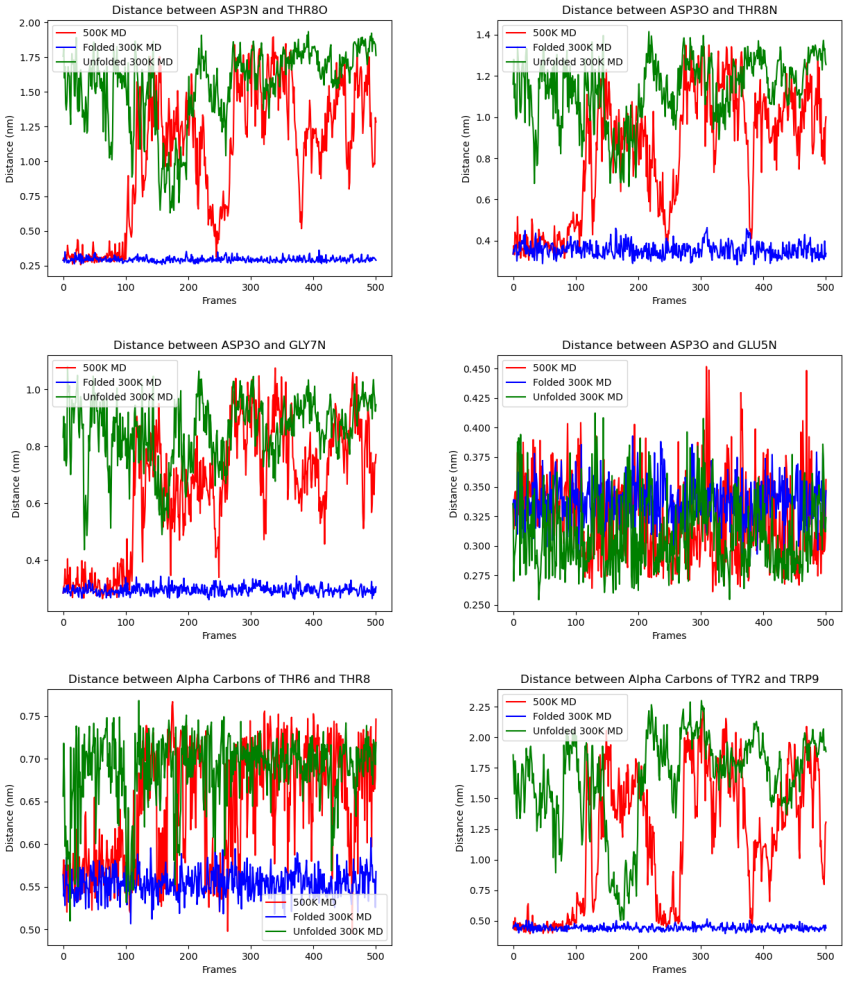


Figure 4.1: Here potential order parameters are plotted against simulation time. A 300K MD simulation from a folded configuration (blue) shows the expected value for folded chignolin; a 300K MD simulation starting from an unfolded configuration (green) shows the range of distances that can occur while chignolin is unfolded. A 500K MD simulation (red) shows the transition between these two. Ideally there would be no overlap between the folded and unfolded values.

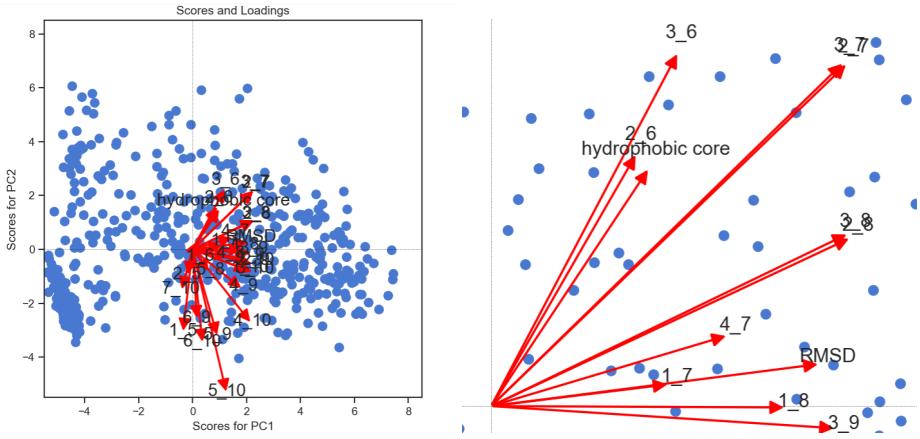


Figure 4.2: PCA was performed for the distances between the alpha carbons of each residue. An inset of the quadrant of the graph opposite to the folded points of folded chignolin can be seen on the right. Several of the distances are highly correlated, for example the TYR2-GLY7 distance and the ASP3-GLY7 distance, as they are very close together and overlapping. These distances as well as the ASP3-TRP6 distance, the TYR2-THR8 distance, the ASP3-THR8 distance and the RMSD value all explain variance between folded and unfolded chignolin.

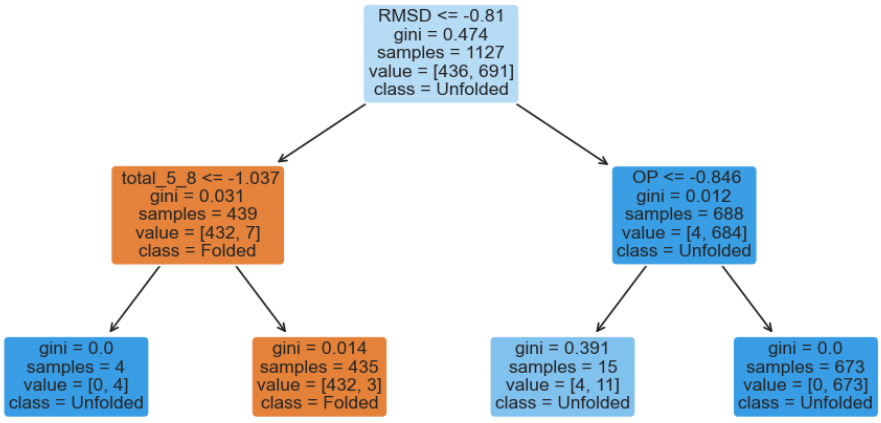


Figure 4.3: This shows the decision tree created from the distances between the alpha carbons of each residue and the RMSD values. The values were scaled before using scikit-learn decision tree classification. The RMSD value separates most of the folded and unfolded configurations, while an additional distance (GLU5-THR8) is used to further classify the folded configuration. Here the blue color represents that it is likely to be unfolded and orange represents folded. The deeper the color; the more likely it is be folded/unfolded. The values, which shows how many in each group are folded or unfolded, and the gini value show how well the decision tree is split.

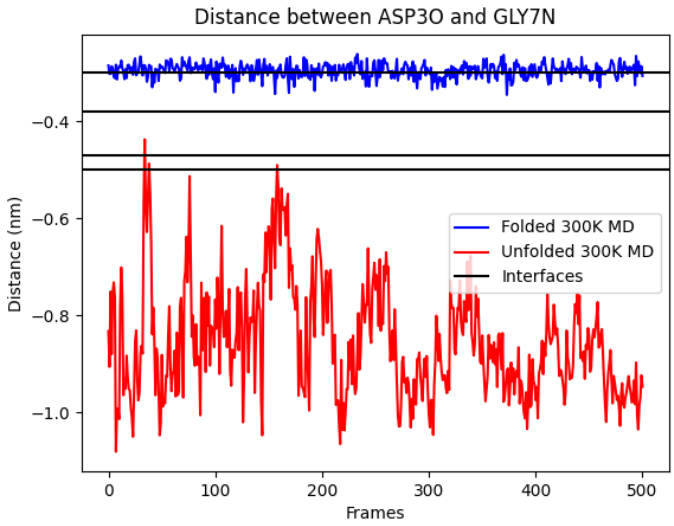


Figure 4.4: The interfaces for the ASP3O-GLY7N order parameter are displayed over the distances expected while folded (blue) and unfolded(red). These expected distances are separated, and the interfaces span across both states.

tree in Figure 4.3, the RMSD value can be used to classify most configurations as folded or misfolded.

Several order parameters were attempted throughout the project including: distances between ASP3O and GLY7N, ASP3O and THR8N, and ASP3N and THR8O, combinations of these distances, the fraction of native contacts and RMSD.

4.1.1 Distance Between Atoms as Order Parameters

The basic program for how to calculate the distance between two particles is a modified version of the Distance Order Parameter available in PyRETIS. This can be found in Appendix B. The major change is that here the negative distance is taken as the atoms need to get closer while RETIS requires the order parameter to increase rather than decrease.

ASP3O-GLY7N Negative Distance

The distance between ASP3O and GLY7N was chosen as one order parameter to test with the RETIS simulations. For the CHARMM27 force field, the interfaces used were [-0.5, -.47, -.38, -.30]. As shown in Figure 4.4, the -0.3 interface is inside the expected folded region. Figure 4.5 shows the path density of the trajectories in the [0-] ensemble. The distance between ASP3O and GLY7N extends to -0.9nm in some of the trajectories. This shows that while the [0-] ensemble does explore part of the unfolded region, the whole region is not successfully explored when the

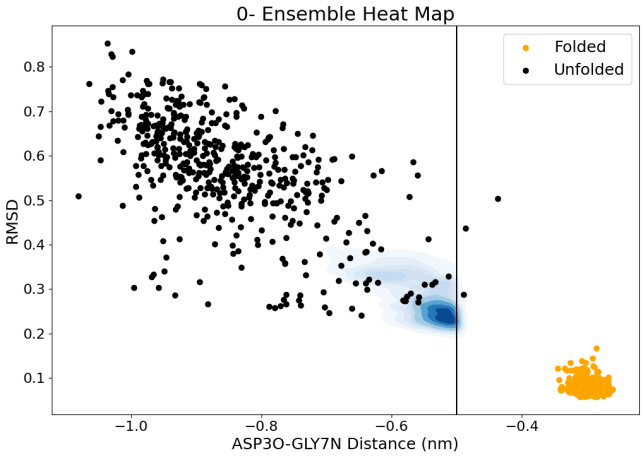


Figure 4.5: The heat map of the [0-] ensemble for the ASP3O-GLY7N order parameter shows that the trajectories explore beyond the interface. However, the entire folded region is not explored while the interface is -0.5, shown as the vertical line.

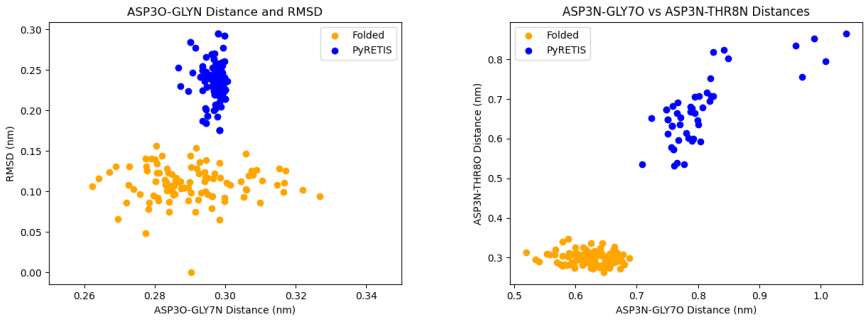


Figure 4.6: On the left, it can be seen that the RMSD values are higher than expected for folded chignolin in the trajectories with the ASP3O-GLY7N order parameter. On the right in the turn plot, the orange represent folded chignolin while the blue represents the last frames of the PyRETIS simulations. This shows the turn region in chignolin is not formed properly and indicates that chignolin in these trajectories become misfolded.

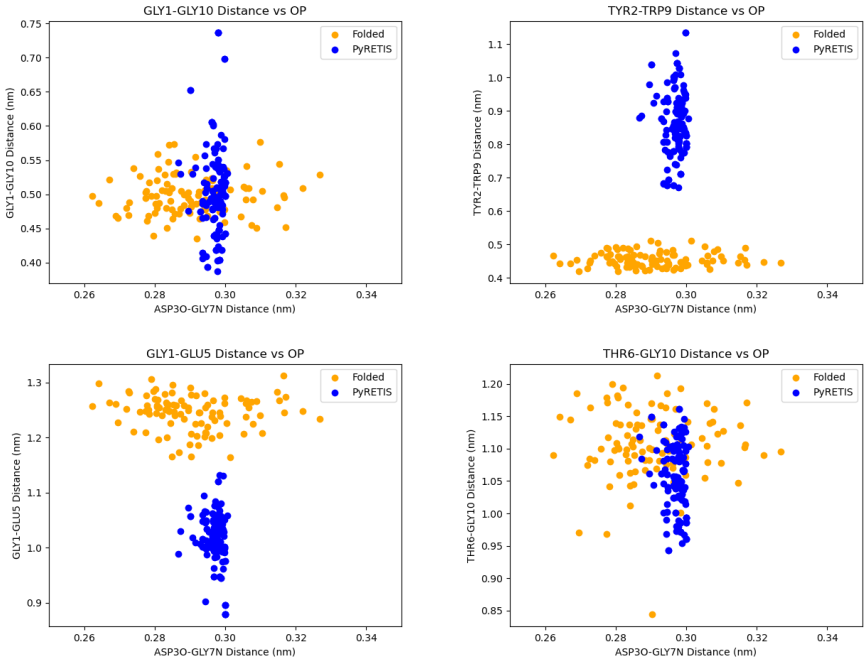


Figure 4.7: Here the distances between select residues in chignolin are shown for the ASP3O-GLY7N order parameter. The orange shows what is expected for folded chignolin (from a 300K MD run), and the blue shows the distances of the last frame for each LMR trajectory created from PyRETIS. It can be seen that several of the distances are different from the expected, while some are as expected.

first interface is placed at -0.5. This interface can be moved closer to the attractive basin of the unfolded state to facilitate complete exploration of the unfolded state.

PyRETIS produces trajectories by shooting off of the initial trajectory. Some of the trajectories are accepted, while others are not depending on if they cross specific interfaces. The accepted trajectories are either LML trajectories or LMR trajectories. The LML trajectories cross the left interface, a given middle interface and then return to the left interface. These trajectories do not fold. The LMR trajectories cross from the left interface to the right interface where chignolin is folded properly, according to the order parameter. For the analysis of if the trajectories are in fact properly folded, the LMR trajectories are used.

Analysis of the LMR trajectories showed that chignolin was not folded correctly at the end of the trajectories. Generally, a RMSD value of less than 0.18nm with respect to the crystal NMR structure of chignolin is considered folded. Additionally, the distances between ASP3-GLY7 and ASP3-THR8 are an indication of if the turn in chignolin is formed properly. This can help distinguish between the folded configuration, which has a π-turn, and the misfolded configuration, which has an α-turn. The plot of these distances is referred to as the turn plot in this thesis. In these trajectories, the ASP3-GLY7 and ASP3-THR8 distances are not in the folded region and the RMSD values were too high as seen in Figure 4.6. Additionally, the graph of the turn indicates that these trajectories have become misfolded rather than properly folded. More differences in the structure of chignolin from this simulation can be seen in Figure 4.7, which shows the distances between different atoms in the backbone of chignolin. The distance between the end residues in chignolin (GLY1-GLY10) match what is expected from folded chignolin, but the distances from the turn to one of the ends(GLY1-GLU5) and the distance between the residues in the hydrophobic core (TYR2-TRP9) are different than would be expected.

The LMR trajectories produced in PyRETIS were extended for 1 ns to see if chignolin would finish folding or unfold given more simulation time. Approximately 5% of the trajectories fold during this extension while the rest unfold.

For the OPLS-AA/M force field, the interfaces were placed at [-0.5, -0.47, -0.35] and later moved to [-0.5, -0.47, -0.38, -0.3]. The [0-] ensemble trajectories were able to explore as far as -0.85nm, which is approximately as far into the unfolded region as the CHARMM27 trajectories with the same interface. The simulations using the OPLS-AA/M force field is slower in performing cycles, likely due to the different water models recommended to be used with the CHARMM27 and OPLS-AA/M force fields. Only a few trajectories were therefore collected, with only one of these potentially being from unfolded to folded. The LMR trajectory that was saved does seem to end in the folded configuration from RMSD analysis and visual inspection.

Given that very few trajectories fold with this order parameter, the ASP3O-GLY7N hydrogen bond may not be the most important feature for the formation of chignolin. The other hydrogen bonds, or other interactions in chignolin, may play a more important role. However this bond could still be important for stabilizing

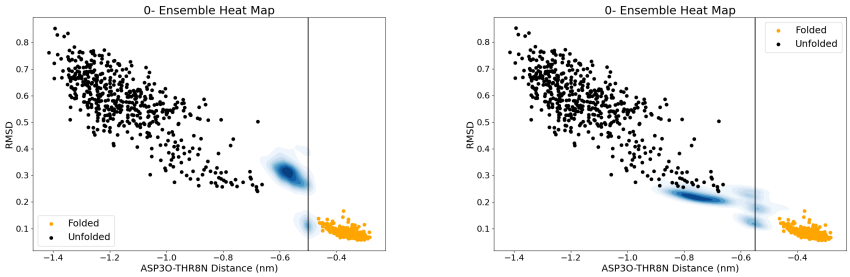


Figure 4.8: This is the [0-] ensemble heatmap for the ASP3O-THR8N order parameter and CHARMM27 force field. The interfaces -0.5 (left) and -0.55 (right) are represented by the vertical lines. The CHARMM27 -0.55 interface shows that state A, unfolded chignolin is reached and partially explored. The -0.5 interface shows that the trajectories do not successfully explore state A and remain close to state B or folded chignolin.

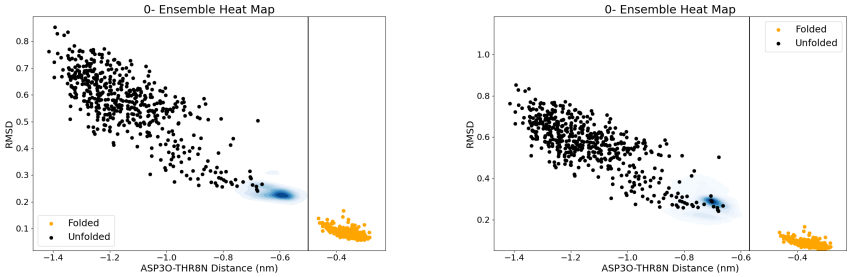


Figure 4.9: The first interfaces -0.5 (left) and -0.57 (right) are represented by the vertical lines for the ASP3O-THR8N order parameter. The OPLS-AA/M -0.57 interfaces show that state A, unfolded chignolin is reached and partially, but not completely, explored. The -0.5 interface does not allow almost any exploration of the unfolded region at all.

once the folded conformation of chignolin is formed. Driving the reaction on this distance may not be a good order parameter due to its similarity to the ASP3N-GLY7O hydrogen bond that indicates misfolded chignolin has formed.

ASP3O-THR8N Negative Distance

The negative distance of ASP3O-THR8N was also attempted as an order parameter. The last interface was placed at -0.3 to be in the folded region, and many different first interfaces were tried to increase the efficiency of sampling.The first interface was moved to increase the length of the [0-] ensemble trajectories compared to the [0+] ensemble trajectories. Additionally, the first interface must be close enough to state A (unfolded) so that PyRETIS could sample trajectories from state A (unfolded) to state B (folded) rather than sampling from state B to state B, or remaining entirely in the folded region.

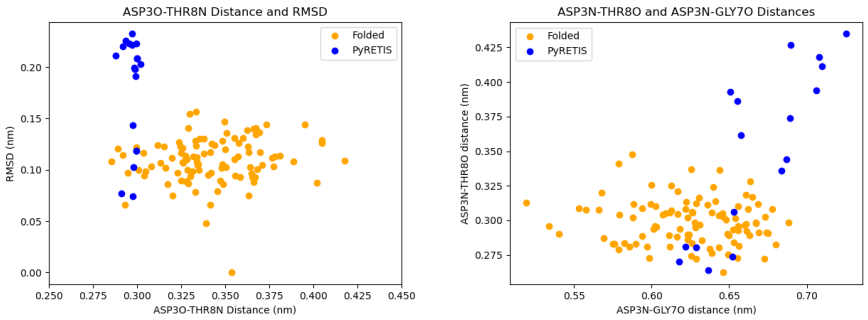


Figure 4.10: The left shows the RMSD plot for the ASP3O-THR8N order parameter; the right shows the turn plot. The last frame of the LMR trajectories (blue) are compared to folded chignolin (orange). Some of the trajectories end with a RMSD value in the expected range for folded chignolin, while still some trajectories do not. Additionally some trajectories have the correct configuration in the turn region for folded and not misfolded chignolin.

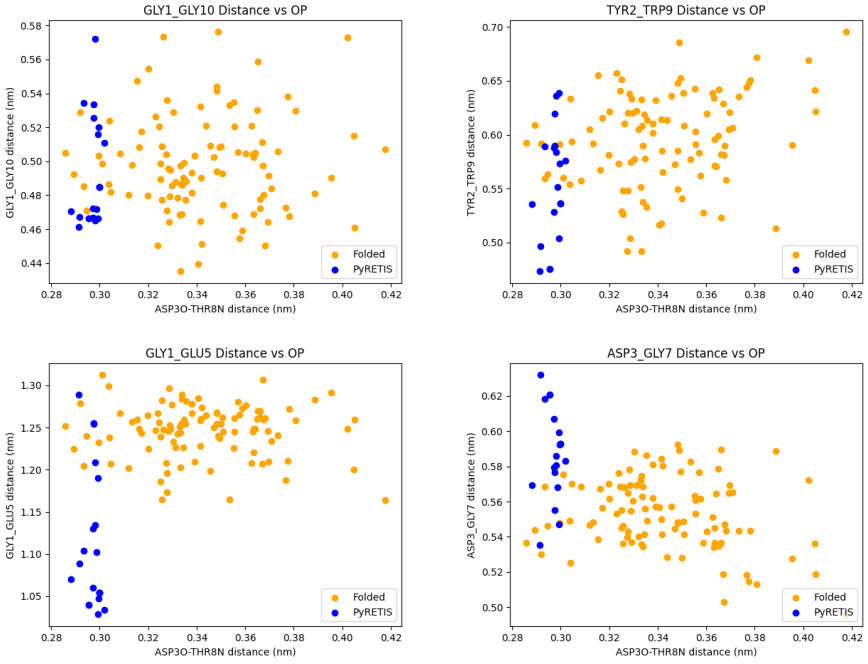


Figure 4.11: These plots shows the distances between specific residues for the ASP3O-THR8N order parameter. The orange shows what is expected for folded chignolin, and the blue shows the distances of the last frame for each trajectory created from PyRETIS simulations. Here it can be seen that there is a difference in the distances for the GLY1-GLU5 distance as well as differences in the distances for ASP3-GLY7.

Figure 4.8 shows two of the first interfaces used for the CHARMM27 force field, -0.5 and -0.55. When the interface is -0.5, the trajectories in the [0-] ensemble are not stuck directly on the interface. They can explore towards the unfolded region; however, the trajectories do not explore much inside the unfolded region. When the first interface is -0.55. The trajectories do explore farther away from the interface towards the unfolded region. This gives a more accurate sampling of trajectories that go from unfolded to folded chignolin. Moving the first interface farther towards the unfolded region would result in even better sampling of the entire transition.

For the OPLS-AA/M force field a variety of first interfaces was also tested. Figure 4.9 shows interfaces -0.5 and -0.57. With the OPLS-AA/M force field, the regions explored in the [0-] ensemble are further away from the interface than with the CHARMM27 force field, where there is a large overlap with the interface. This indicates that the trajectories in this force field spend a lot of the time near the interface as well as exploring farther away. In OPLS-AA/M, the -0.5 interface trajectories explore a region closer to the unfolded region, and the -0.57 interface explores part of the unfolded region. This shows these trajectories spend less time around the interface and more time exploring. These different behaviors show the difference between the two force fields. The properties of the force field can affect sampling. In both cases, the first interface could still be moved farther over to increase the area in the unfolded region that is explored.

This order parameter was an improvement on the ASP3O-GLY7N order parameter as there were trajectories that did fold properly by the end of the simulation. This was determined by visual inspection, RMSD values, which are shown in Figure 4.10 and distances between specific residues as shown in Figure 4.11. Some of the RMSD values are lower than 0.18nm indicating that the trajectory ends with properly folded chignolin; however, not all of the trajectories end properly folded. The second graph in Figure 4.10 shows that the turn is formed correctly in some trajectories. Additionally, the distances between, for example ASP3-GLY7 and GLY1-GLU5, are not within the range expected. This shows there is room for improvement in the order parameter used to distinguish the folded and unfolded state and potentially a second parameter should be added to improve upon it. The success of this order parameter in producing trajectories that properly fold may indicate that the ASP3O-THR8N hydrogen bond may have a important role in the folding mechanism.

ASP3N-THR8O Negative Distance

The final single distance order parameter attempted was the distance of ASP3N to THR8O with the interfaces [-0.43, -0.34, -0.28]. This order parameter also resulted in folded chignolin. The RMSD values shown in Figure 4.12 and the notable distances between residues in Figure 4.13 indicate that there is folded chignolin. However not all of the trajectories end properly folded as can be seen by the deviation from expected values in the distances between residues and a few tra-

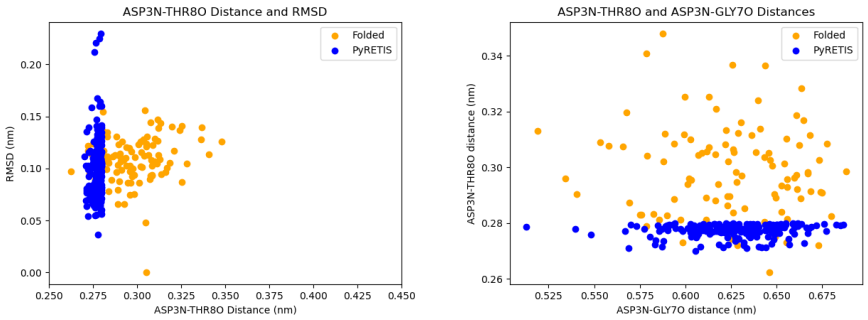


Figure 4.12: The graph on the left shows the RMSD plot for the last frames of the trajectories produced by PyRETIS with the ASP3N-THR8O order parameter. Many, but not all, of the trajectories end folded. A few trajectories have a RMSD value over 0.2nm. The turn plot on the right is able to differentiate between folded and misfolded chignolin. The orange represent folded chignolin while the blue represents the last frames of the PyRETIS simulations. The turn seems to be properly folded in all trajectories.

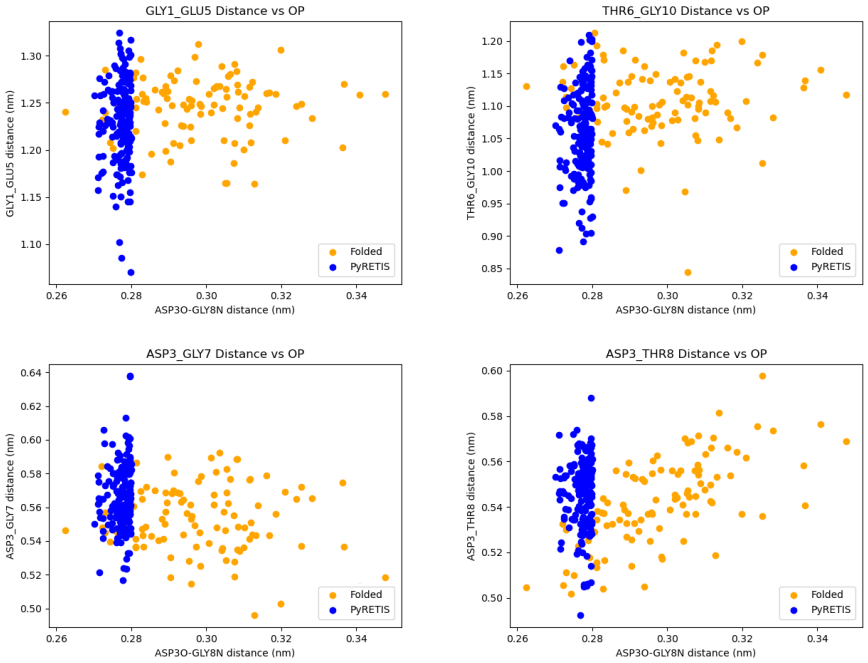


Figure 4.13: Here the distances between some of the residues in chignolin are shown for the ASP3N-THR8O order parameter trajectories. The orange shows what is expected for folded chignolin (from a 300K MD run), and the blue shows the distances of the last frame for each trajectory created from PyRETIS simulations with ASP3N-THR8O as the order parameter. It can be seen that while the PyRETIS simulations mostly match what is expected, there is some deviations where the trajectories are incorrectly folded.

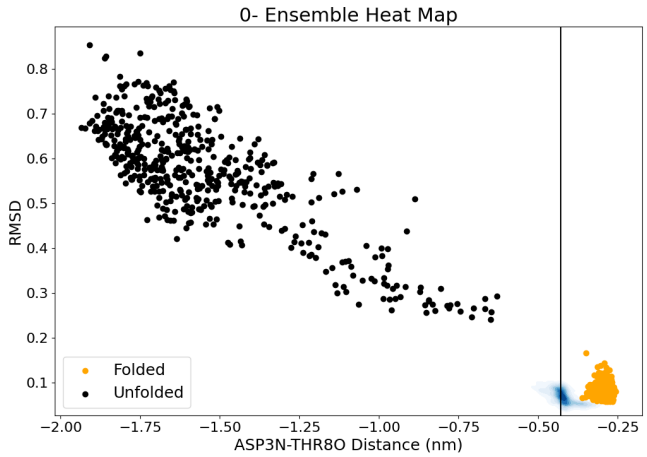


Figure 4.14: Heat maps of the [0-] ensemble trajectory for the ASP3N-THR8O order parameter with the CHARMM27 force field. The folded (orange) and unfolded (black) regions are shown with the 0.42 interface. The [0-] ensemble is not within, but also does not explore far beyond, the folded region. This interface placement does not allow proper sampling from unfolded to folded.

jectories that end with RMSD values of above 0.18nm. The ASP3N-THR8O and ASP3N-GLY7O graph, which as previously mentioned can help indicate if the turn in chignolin is formed properly, shows that the turn is formed properly with this order parameter. The difference in RMSD in the trajectories that did not fold therefore likely comes from another section of the structure of chignolin. The few trajectories that did not fold during the simulation were then extended to determine if they would fold quickly after; the extended trajectories instead unfolded completely.

The [0-] ensemble, a heatmap of which can be seen in Figure 4.14, shows that the unfolded region was not explored. The trajectories were not able to extend far beyond the interface. Given this, that some of the trajectories ended with improperly folded chignolin shows that this is not a good parameter to be using as the order parameter and as a result, the interfaces were not further adjusted. In this case, the first interface was placed too close to the folded state and not in the unfolded state. The interface not being in the unfolded region does not allow proper sampling of the entire transition or calculation of the rate constant. The trajectories produced here consist almost entirely of folded chignolin and not the transition.

Overall, no single distance between the residues is a good order parameter for chignolin. Despite its small size, it is too complicated for a simple order parameter like a single distance. Additionally, there were many problems determining where the interfaces should be placed. In an attempt to increase the length of the average [0-] ensemble path length,the interfaces were moved closer to the folded region.

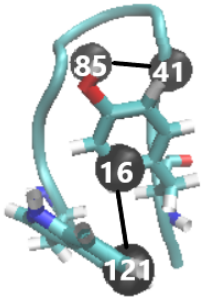


Figure 4.15: The two distances used as the ASP3O-GLY7N/TYR2-TRP9 order parameter are shown here with the indexes of the atoms. The distance between ASP3O-GLY7N (indexes 41 and 85) was chosen to help determine if the turn is formed, while the distance between TYR2-TRP9 (indexes 16 and 121) show if the hydrophobic core is formed successfully.

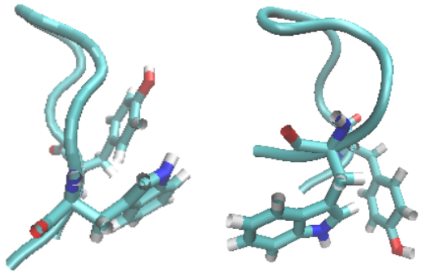


Figure 4.16: On the left is properly folded chignolin; on the right chignolin is not folded properly and is not stable. There is a bend in the protein and the residues in the hydrophobic core have the wrong orientation.

The interfaces should have instead been moved farther into the unfolded region. The attraction to the folded conformation was stronger than that to the unfolded conformation where the interfaces were placed here. The first interface should be placed in the basin of attraction for the unfolded region (state A). This interface placement issue affects all of the order parameters used in this study. RETIS tries to correct for this by using the flux, which considers the average path lengths of the [0-] and [0+] ensembles, in the rate constant calculation. However, the trajectories produced do not fully represent the transition from completely unfolded to folded.

4.1.2 Combinations of Distances as Order Parameter

No single distance in chignolin is sufficient as a order parameter. Therefore, combinations of several distances was also used. The negatives of the values are used so that the order parameter increases as the protein folds.

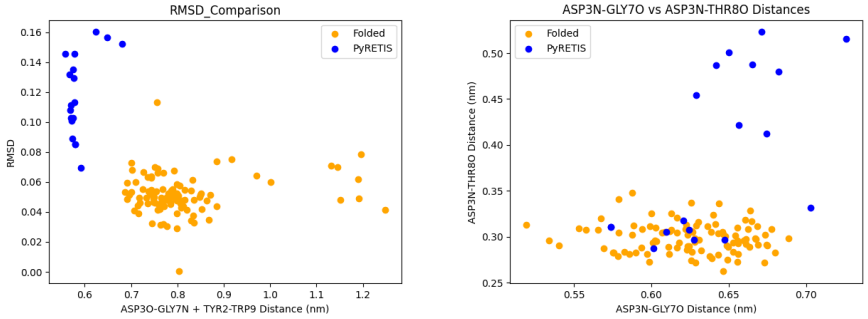


Figure 4.17: The graph on the left shows the RMSD values for the additive ASP-GLY/TYR-TRP order parameter with the CHARMM27 force field. The RMSD values are higher than expected with the CHARMM27 force field. It can also be seen that the last interface was placed farther than would normally be expected with this order parameter. The graph to the right shows two distances that can indicate if the turn region is formed properly. Approximately half seem to be properly formed while the rest are not.

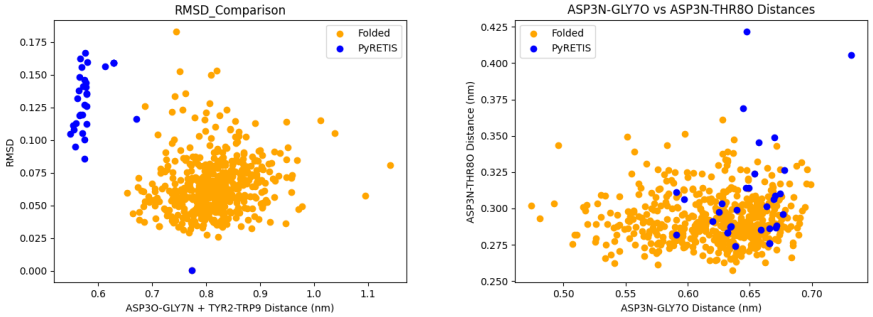


Figure 4.18: The graph on the left shows the RMSD values for additive distance between ASP-GLY/TYR-TRP order parameter with the OPLS-AA/M force field. While the RMSD values are as expected with the OPLS-AA/M force field, the distances do not match what is expected in chignolin. The interface was incorrectly placed. The graph to the right shows two distances that can indicate if the turn region is formed properly. Most seem to be properly formed while a few are not.

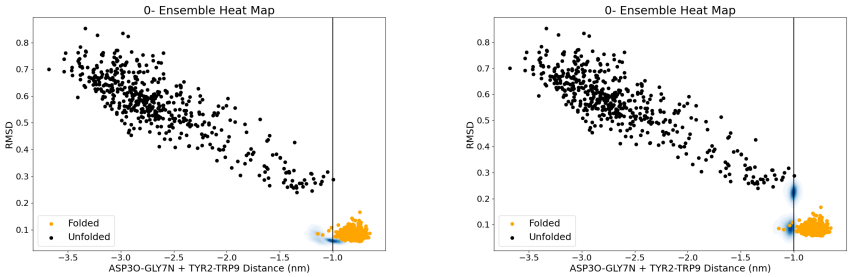


Figure 4.19: Heat map of the trajectories of the [0-] ensemble for the ASP3O-GLY7N/TYR2-TRP9 (Addition) order parameter with CHARMM27 (left) and the OPLS-AA/M (right) force fields with a first interface of -1. The trajectories are almost entirely in the folded region and do not explore.

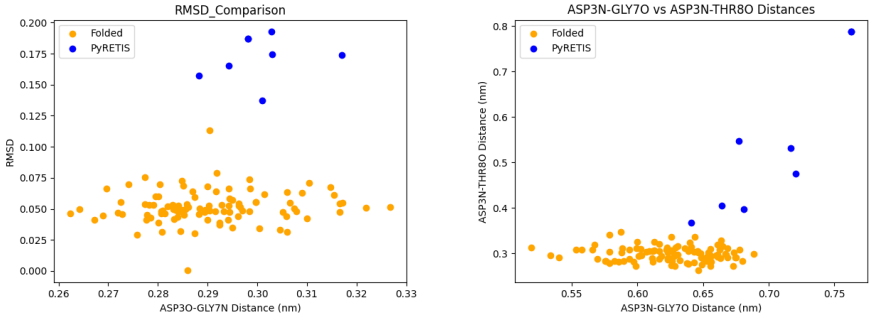


Figure 4.20: RMSD values for the ASP3O-GLY7N/TYR2-TRP9 (If/then) order parameter are to the left and the GLY-THR graph to the right. The turn plot shows that the turn is likely not formed properly in these trajectories, but the ASP-GLY and TYR-TRP distances appear to be more in line with what is expected for folded chignolin even though the RMSD values are still too high.

ASP3O-GLY7N and TYR2-TRP9

ASP3O-GLY7N alone was not able to move the trajectory from unfolded to completely folded chignolin so an additional distance was found to add to this order parameter. The distance between TYR2-TRP9 was chosen due to the importance of the hydrophobic core. These two distances, which can be seen in Figure 4.16, were added together to create a combination to be a new order parameter with the interfaces [-1, -0.72, -0.58] for both the CHARMM27 and the OPLS-AA/M force field. The trajectories that move from the left interface to the right interface are not folded in the end as determined by visual inspection. Many of the trajectories ended with chignolin with an uncharacteristically strong bend in the backbone, which can be seen in Figure 4.16.

This bend found in misfolded chignolin also highlights the difference between the two force fields here. The RMSD value can be calculated from comparing the trajectory frame directly to the NMR crystal structure acquired from PDB. It can also be compared to folded chignolin form simulations with either force field. When this bent conformation of chignolin is compared to folded chignolin with the OPLS-AA/M force field, it is different from the RMSD calculated with CHARMM27 or the crystal structure. This suggests differences in the force fields abilities to replicate the system and highlights that the force field is an important choice and about how its effects should be considered.

Figures 4.17 and 4.18 also show that chignolin is unfolded at the end of these trajectories. While the RMSD values are low enough to be considered folded, the distances are not correct for folded chignolin. The interface here was improperly placed to far to the right. This value was gotten by adding the average expected values for each distance together and did not take into consideration any variation. Since the interface is not in the basin of attraction for the folded region none of the trajectories properly folded. If the interface was moved towards the center of the expected region, it is possible that sampling would be better and properly folded trajectories would be produced. Still, in over half of the trajectories, the turn in chignolin seems to have formed properly. Given that it is not stable even with the turn formed, there are other factors that need to be considered in an order parameter for chignolin.

Figure 4.19 shows the path density of the [0-] ensemble for both the CHARMM27 and the OPLS-AA/M force fields with this order parameter. The trajectories for the CHARMM27 trajectories with the interface -1 do not explore outside the folded region as the interface is already within the folded region and is not able to easily leave. However, the trajectories did not end with folded chignolin which showed that this order parameter does not successfully separate folded and unfolded chignolin. The OPLS-AA/M force field with the interface of -1 is not contained within the folded region as the RMSD value becomes that of unfolded, or misfolded, chignolin. However, the trajectories still does not explore very much of the unfolded region. This interface should have also been moved. It would need to be further into the unfolded region to improve sampling.

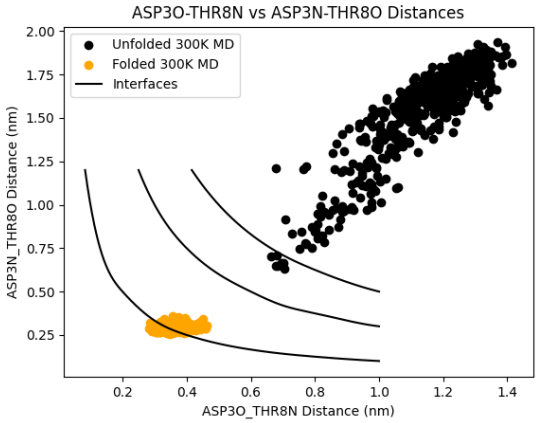


Figure 4.21: The interfaces for the multiplicative ASP3O-THR8N/ASP3N-THR8O order parameter are displayed over a graph of the distances in folded chignolin as well as in unfolded chignolin.

As the linear combination used for these two distances did not result in folded chignolin, another formulation of these two distances using an if statement in the program to check the TYR2-TRP9 distance after the ASP3O-GLY7N distance reaches the expected value (over -0.3nm). The interfaces are first [-0.5, -0.4, -0.3] for the ASP3O-GLY7N distance, then if that condition is met, the TYR2-TRP9 distance is considered to see if it crosses the interface [-0.4]. This would be more strict with each of the distances individually. The [0-] ensemble here would be expected to be similar to that of the ASP3O-GLY7N order parameter and sufficiently explore the unfolded region.

The trajectories produced with this order parameter also do not fold correctly. This was determined by visual inspection using VMD and from the RMSD value which are higher than for folded chignolin. Most of the trajectories ended incorrectly folded in similar ways to when using the addition version of this order parameter. The RMSD values, shown in Figure 4.20, also indicate the trajectories misfold. The turn in chignolin is also not correctly formed. This suggests that the bend found in the trajectories with the linear combination is not caused by too much freedom in the order parameter, but in that the distances here can be satisfied without chignolin being in the correct configuration.

ASP3O-THR8N and ASP3N-THR8O

Another order parameter attempted was the distance between ASP3O-THR8N and ASP3N-THR8O. First a multiplicative order parameter was attempted. The positive version of the interfaces for the OPLS-AA/M force field, [-0.5, -0.3, -0.1], can be seen in Figure 4.21. Due to the small number of trajectories, the CHARMM27 and OPLS-AA/M trajectories were combined for this analysis. The [0-] ensemble trajectories, the heat map of which is visible in Figure 4.22 do not explore the

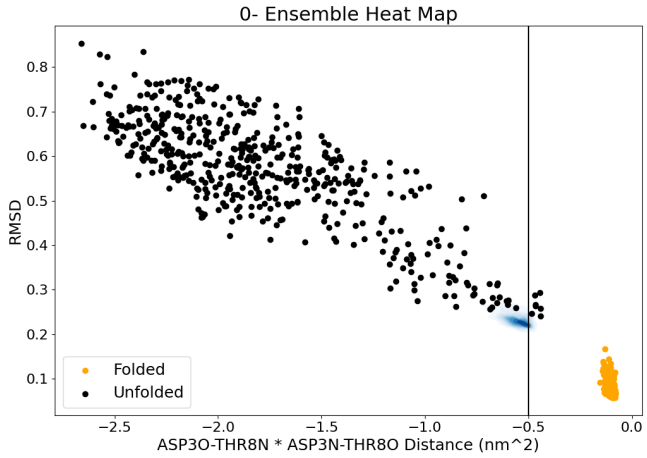


Figure 4.22: The [0-] ensemble trajectories for the ASP3O-THR8N/ASP3N-THR8O order parameter are shown stuck near the interface and not exploring the unfolded region. For better sampling, the interface should be moved to facilitate exploration of the unfolded region.

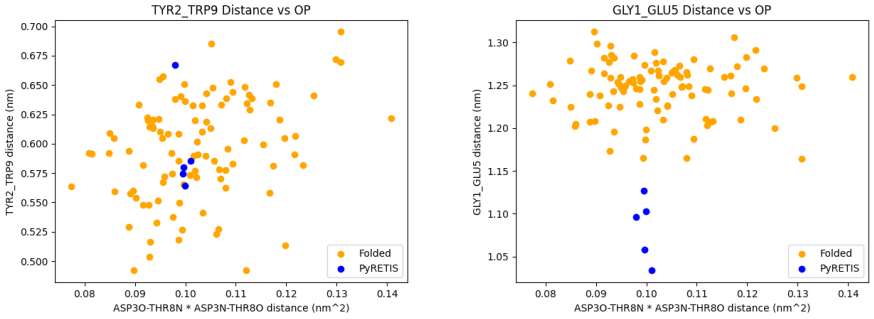


Figure 4.23: Here the distances between specific residues in chignolin are shown for the ASP3O-THR8N/ASP3N-THR8O order parameter. The orange shows what is expected for folded chignolin (from a 300K MD run), and the blue shows the distances of the last frame for each trajectory created from PyRETIS simulations with ASP3N-THR8OxASP3O-THR8N as the order parameter. It can be seen that several of the distances match what is expected. However GLY1-GLU5 distance is different from the expected in different trajectories produced.

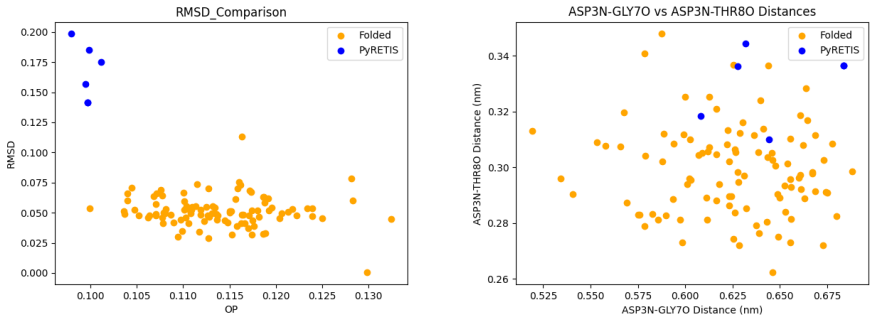


Figure 4.24: The left graph shows the RMSD values for the ASP3O-THR8N/ASP3N-THR8O order parameter. The right graph shows the distances between two residues that show if the turn is formed properly for folded chignolin. The RMSD values are too high, but the turn is formed properly.

unfolded region; they instead stay close to the interface. To increase the quality of the sampling and rate constant calculation, the first interface should be moved farther into the unfolded region to where the majority of the points are on the plot. Additionally, when using this order parameter, the LMR trajectories do not necessarily correctly fold by the end of the trajectory.

Figure 4.23 shows that while many of the distances in the backbone of chignolin are as expected, some of them, including the distance GLY1-GLU5 differ. This changes the shape of chignolin and affects the RMSD values as well. The RMSD values are too high for folded chignolin as can be seen in Figure 4.24. This difference in RMSD is found in parts of chignolin outside the turn as the other graph in Figure 4.24 shows that the turn is formed properly in these trajectories. Also of note, the distance between TYR2-TRP9, shown in Figure 4.23, is as expected for folded chignolin. This distance would indicate that the hydrophobic core is in the correct distance to be have the appropriate hydrophobic interactions. The difference between this misfolded chignolin and folded chignolin must be somewhere other than the hydrophobic core.

To determine if this misfolded conformation that chignolin finds itself in in these trajectories is an intermediate between unfolded and folded chignolin, the LMR trajectories were extended for 1ns using GROMACS. Upon extension, these trajectories do not fold, but completely unfold quickly. Therefore, it is not an intermediate state. This supports that something other than the turn and hydrophobic core is also important for the stabilization of chignolin. Looking at the other distances between the residues and path density plots, as is done later in Section 4.3.1, could give some insight into what else besides the turn is necessary for folded chignolin to form and be stable.

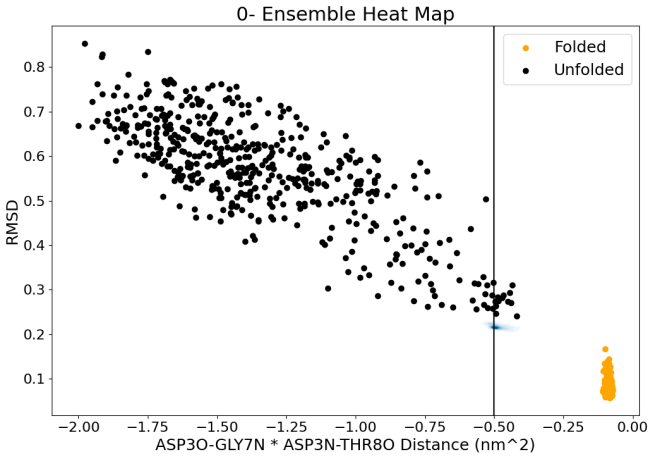


Figure 4.25: The heatmap of the [0-] ensemble for the ASP3O-GLY7N/ASP3N-THR8O order parameter with the CHARMM27 force field shows that the trajectories cannot explore far beyond the interface. The unfolded region is not explored and the trajectories produced with these interfaces does not represent a full transition from unfolded to folded. The OPLS-AA/M heatmap (not pictured) shows a similar lack of exploration.

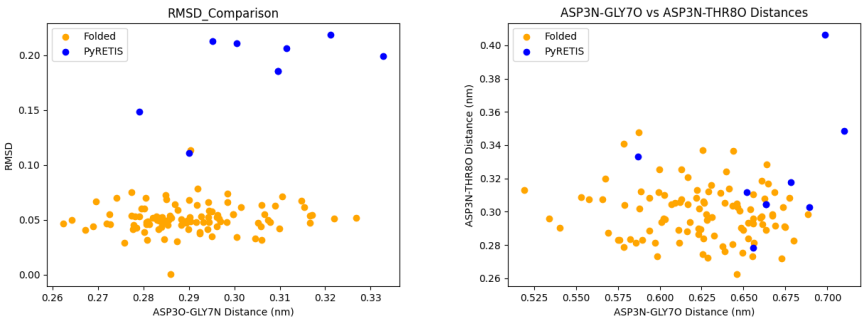


Figure 4.26: The left shows the RMSD values from the PyRTEIS simulations with the ASP3O-GLY7N and ASP3N-THR8O order parameters and the right graph shows a plot which helps show if the turn is formed properly in chignolin. The RMSD values are too high for folded chignolin, but some of the turns are formed properly.

ASP3O-GLY7N and ASP3N-THR8O

The final combination of distances order parameter is the combination of the distances between ASP3O-GLY7N and ASP3N-THR8O. These two distances were used as the order parameter in a different TPS study of chignolin and is similar to the plot mentioned before that can show if the turn id formed correctly or not. First a multiplicative combination was attempted with a the interfaces [-.5, -.3, -.09]. This is similar to the ASP3O-THR8N/ASP3N-THR8O order parameter above. The [0-] ensemble, shown in Figure 4.25, shows that the unfolded region is not explored. If this order parameter was to be used further, the first interface would have to be moved.

After the multiplicative order parameter, another order parameter, the method of checking the first distance until it reaches the desired distance before checking if the 2nd distance is also within a specific distance, was employed. The ASP3O-GLY7N distance had the interfaces [-0.6, -0.45, -.35] after the ASP3N-THR8O distance was checked if it was less than 0.29nm. The [0-] path ensemble trajectories here were able to explore a little more of the unfolded region than the ASP3O-GLY7N single distance order parameter because the interface was moved farther into the unfolded area.

Both of these formulations of these distances as order parameter do not produce trajectories that fold into properly folded chignolin. This is evident by the high RMSD values despite the similarity in the formation of the turn seen in Figure 4.26. These trajectories were also extended to see if they were a intermediate state between folded and unfolded chignolin. Upon extension using GROMACS, some, but not all, of the trajectories fold correctly within 1 ns. This again points to other factors other than the turn and hydrogen bonds also being important in the stabilization of chignolin.

After looking at several single and combine distance order parameters, it is apparent that the order parameter must be much more descriptive of the protein to be able to ensure that chignolin is properly folded by the end of the trajectory. Therefore, methods of looking at the general shape of the molecule were considered for the order parameter.

4.1.3 RMSD as Order Parameter

The Kabsch algorithm [48, 49] was used to create a RMSD order parameter script that could work with the Gromacs2 engine. This can be found in Appendix B. Chignolin is folded when the RMSD value is less than 0.18nm and unfolded when the RMSD value is higher. The RMSD values of folded chignolin and a trajectory with transitions between folded and unfolded chignolin can be seen in Figure 4.27 with the interfaces [-0.4 , -0.35, -0.29, -0.2, -0.14, -0.1] chosen for the CHARMM27 force field and the interfaces [-0.4, -0.35, -0.28, -0.2, -0.15, -0.1] for the OPLS-AA/M force field. The [0-] ensemble explores from the interface of -0.4nm to -0.63nm. Completely unfolded chignolin would have an RMSD value of around -0.8, here negative to match the order parameter. This interface for the

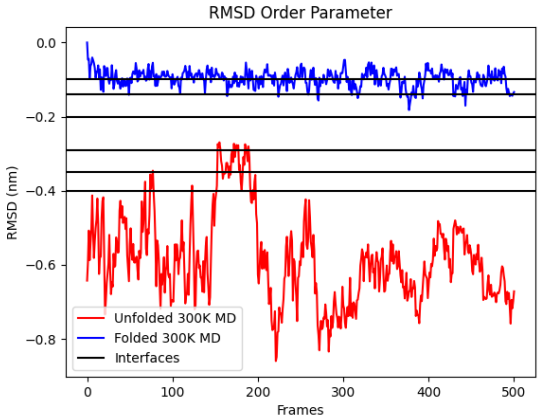


Figure 4.27: The interfaces for the RMSD order parameter displayed over the distances expected while folded (blue) and unfolded (red). The interfaces cross from the unfolded region to the folded region.

RMSD order parameter does not explore the entire unfolded region, but it does explore a larger area of it than the other order parameters discussed earlier. The interface can still be moved to increase the area of the unfolded region explored.

While RMSD is a good indicator of whether or not chignolin is folded, it is not an absolute. Visual inspection was used to determine if the structure is folded. Another method to determine if chignolin is actually folded, and would therefore likely be stable, the trajectories were extended using GROMACS. Chignolin, when properly folded, is stable at this temperature; therefore if the the trajectory is extended and does not unfold, then it can be considered stable and folded. The trajectories produced with this order parameter were not all folded. Due to time constraints very few trajectories were created in the simulation, resulting in few to analyze to determine if this is a good order parameter and how to potentially improve it.

From the RMSD order parameter trajectories, it was found that other key factors play a part in stabilizing chignolin that the RMSD alone cannot necessarily replicate. A combination of these factors and RMSD could be designed to take the benefits of both potential order parameters. This would let chignolin get the general shape required for chignolin (which is needed for a good RMSD value) and also specific stabilizing interactions.

Table 4.1: Order Parameter, interfaces and rate constants

Order Parameter	**Interfaces**	**Rate Constant (ps^{-1})**	**Adjusted RC (ps^{-1})**
ASP3O-GLY7N (C)	[-0.5, -0.47, -0.38, -0.3]	$1.7661 \times 10^{-2} \pm 51\%$	9.35×10^{-4}
ASP3O-GLY7N (O)	[-0.5, -0.47, -0.38, -0.3]	$2.9696 \times 10^{-3} \pm 145\%$	
ASP3O-THR8N (C)	[-0.5, -0.4, -0.3]	$2.3603 \times 10^{-3} \pm 53\%$	2.776×10^{-4}
ASP3O-THR8N (O)	[-0.5, -0.4, -0.3]	$1.285 \times 10^{-2} \pm 70\%$	
ASP3N-THR8O (C)	[-0.43, -0.34, -0.28]	$2.0476 \times 10^{-2} \pm 33\%$	1.99×10^{-2}
ASP3O-GLY7N/TYR2-TRP9 (C)	[D-G(-0.5, -0.4, -0.30), Y-W(-0.4)]	$6.935 \times 10^{-5} \pm 117\%$	
ASP3O-THR8N/ASP3N-THR8N (C)	[-0.5, -0.45, -0.3, -0.1]	$3.076 \times 10^{-4} \pm 155\%$	
ASP3O-THR8N/ASP3N-THR8N (O)	[-0.45, -0.3, -0.275, -0.2, -0.1]	$8.526 \times 10^{-6} \pm 154\%$	
ASP3O-THR8N/ASP3O-GLY7N Mult (C)	[-0.5, -0.3, -0.09]	$6.504 \times 10^{-4} \pm 176\%$	
ASP3O-GLY7N/ASP3O-THR8N If (C)	[D-G(-0.6, -0.47, -0.38, -0.3) D-T(-0.29)]	$3.063 \times 10^{-5} \pm 229\%$	1.021×10^{-5}
ASP3O-THR8N/RMSD (C)	[D-T(-0.7, -0.6, -0.5, -0.4) RMSD(-0.2, -0.1)]	$1.860 \times 10^{-4} \pm 177\%$	

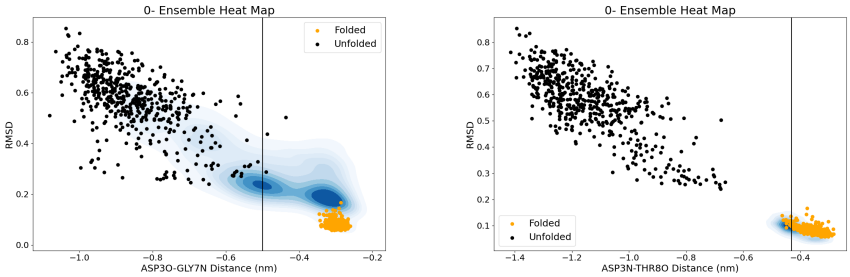


Figure 4.28: Trajectories were started in GROMACS from the first interfaces of two order parameters. The interface is -0.5 for the ASP3O-GLY7N order parameter (left). Some of the trajectories explore the unfolded region and some of them go towards the misfolded region. A few of the trajectories end properly folded. The interface for the ASP3N-THR8O order parameter is -0.43 (right), and the trajectories go towards the folded region.

4.2 Rate Constant

For each of the PyRETIS simulations performed the rate constant was calculated as described in the theory section using the pyretisanalyse function in PyRETIS. Table 4.1 shows the order parameters and the rate constants calculated from those simulations.

The rate constant for chignolin is previously estimated to be of the order 10^{-5}-$10^{-6} ps^{-1}$ [16, 50]. Comparing the results from these order parameters to that, it can be seen that not all of the order parameters and interfaces result in a rate constant that agrees with these estimations. There are several reasons proposed for why this is the case.

First, several of the PyRETIS runs, for example those with the ASP3O-GLY7N order parameter, did not end with properly folded chignolin. This results in an order parameter that is higher than expected as it could be more likely to reach a not folded state compared to the folded conformation. The originally calculated rate constant for the ASP3O-GLY7N order parameter is around $2.9696 \times 10^{-3} ps^{-1}$ for the OPLS-AA/M force field and around $1.7661 \times 10^{-2} ps^{-1}$ for the CHARMM27 force field. This order parameter with the CHARMM27 force field was adjusted by extending all of the accepted LMR trajectories using GROMACS to determine if they would later fold properly or not. Approximately 5% did fold with in 1ns, and this value was used to adjust the rate constant calculated in PyRETIS to be around $9.35 \times 10^{-4} ps^{-1}$. This is closer, but not exactly as expected. The OPLS-AA/M force field rate constant was not adjusted due to a lack of trajectories to extend and analyze.

The difference between the rate constant calculated for the OPLS-AA/M force field and the CHARMM27 force field could be caused by differences in force field behavior or simply by the number of trajectories used to calculate each rate constant. The OPLS-AA/M rate constant here only had a little over 200 trajectories to

calculate from, while the CHARMM27 rate constant had over 2,500 trajectories. The large difference in cycles comes from that using the OPLS-AA/M force field resulted in slower calculations. This is likely due to the water model used with each force field. The more cycles completed, the more accurate the rate constant calculation is. The PyRETIS analysis tool plots the rate constant calculated versus the cycle number. Neither force field rate constant had completely plateaued in these plots, which means that given more time and cycles, the rate constant likely would have changed.

A similar approach to adjust the rate constant for the number of non-folded trajectories counted as folded in the original rate constant calculation was also done for the ASP3O-THR8N and the ASP3N-THR8O order parameters. These trajectories did produce folded chignolin, as well as misfolded chignolin. The percentage of these trajectories that actually ended folded was used to adjust the rate constant in these two instances. The change is more drastic with the ASP3O-THR8N rate constant which with 11% of the trajectories actually being folded resulted with a rate constant of around $2.776 \times 10^{-4} ps^{-1}$. Most of the ASP3N-THR8O order parameter trajectories were folded so this adjustment does not affect the rate constant too much; however this order parameters rate constant is too high by several orders. This is likely due to another effect on the rate constants.

The rate constants, especially, for example, the ASP3N-THR8O order parameter, may be too high because they are not properly sampling from State A (unfolded) to State B (folded). This is evident with the ASP3N-THR8O order parameter, because as seen previously in Figure 4.14, the [0-] ensemble is stuck in the folded state. While not as extreme as in this case, most of the first interfaces used in this study should have been moved further towards the unfolded configuration of chignolin to improve sampling and calculation of the flux. In Figure 4.28, it can be seen that the ASP3N-THR8O order parameter with these first interface of -0.43 has virtually no trajectories that leave from this interface and travel towards the unfolded state rather than to the folded state. The ASP3O-GLY7N order parameter with the first interface of -0.5 interface has only 4 of 83 trajectories go to the folded state. Many trajectories do visit the the misfolded region, where the RMSD value is higher but the distance measurement is similar, and the unfolded region as well.

Another source of discrepancy could be temperature. There may be temperature effects from the initial trajectory used in the simulations which was created by a 500K MD simulation. This effect should decrease with the number of cycles performed. Perhaps more than 1000 cycles of shooting moves should have been used to eliminate correlations with the initial trajectory; 1000 cycles may not be long enough for the system to completely forget the initial trajectory. Additionally, the previously estimated rate constants were not all conducted at 300K, the temperature at which these trajectories were created. This affects the rate constant as well. At higher temperatures, it is more likely for the transition to take place.

While initially the rate constants calculated for the two variable order parameters seem to be more correct based on similarity to the literature value, these

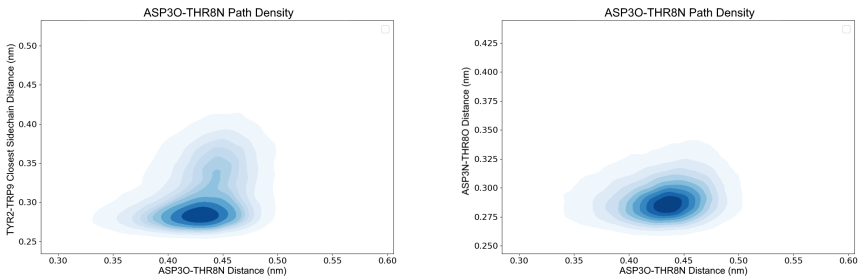


Figure 4.29: Path density plots created from all ensembles of the ASP3O-THR8N order parameter. The left plot shows the residues involved in the hydrophobic core. The distance between these residues decreases as the the ASP3O-THR8N distance decreases. On the right is an example of the rest of the path density plots, The trajectories do not necessarily follow any path.

rate constants are not accurate to chignolin. The ASP3O-GLY7N/TYR2-TRP9, ASP3O-THR8N/ASP3N-THR8N and ASP3O-THR8N/ASP3O-GLY7N order parameters do not properly fold by the end of the trajectories as shown in the sections above discussing them. The rate constants calculated here are therefore not accurate for chignolin.

Approximately 1/3 of the trajectories extended from the supposedly folded trajectories with the ASP3O-GLY7N/ASP3O-THR8N order parameter fold properly, which gives an adjusted rate constant of $1.021 \times 10^{-5} ps^{-1}$. This order parameter is similar to the ASP3O-GLY7N order parameter, except the first interface is placed at -0.6 instead of -0.5 for the ASP3O-GLY7N distance. In both cases, some of the extended trajectories fold, and it is possible that moving this interface improves the sampling and therefore also the calculation of the rate constant in a smaller number of cycles. Given all of these factors, the rate constant can be estimated to be of the order of 10^{-4}-$10^{-5} ps^{-1}$.

4.3 Analysis of Trajectories

4.3.1 Path Density

Path density plots were made using all of the ensembles for specific order parameters. Different variables are plotted against the order parameter to see how it develops over time. Generally, in a path density plot it would be expected that there is at least two dense areas, one in state A, or unfolded, and one in state B, or folded. There would then be the paths between these two areas. Here, due to the interface placement, the unfolded state is not represented on the plots. In most cases, these plots show misfolded conformations. Still, information about chignolin, folded and misfolded can be learned from these plots.

The ASP3O-THR8N order parameter path density plots, shown in Figure 4.29, do not show that there is any specific pattern in most of the residue distances ex-

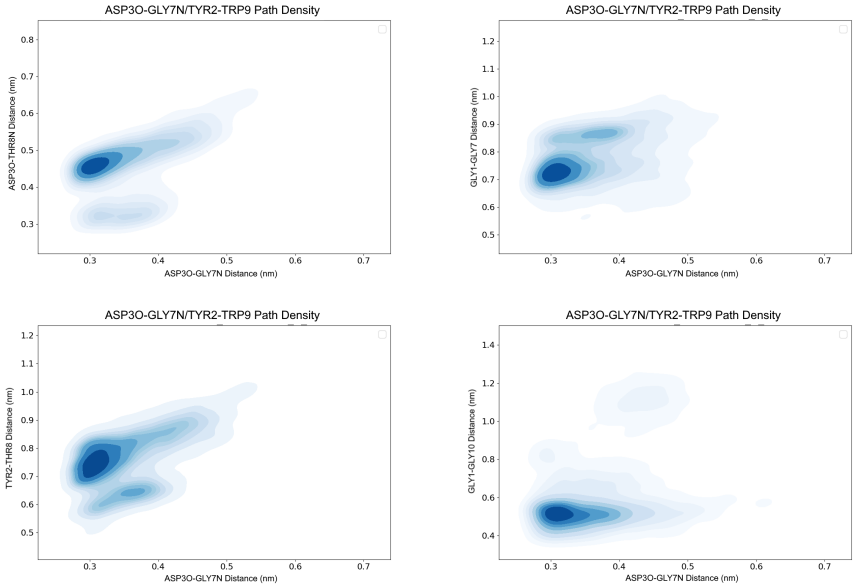


Figure 4.30: Path density plots created from all ensembles of the ASP3O-GLY7N/TYR-TRP If statement order parameter. The darker areas on the plot represent where a higher concentration of frames in the trajectories are spent. In these plots, there are two areas of higher density.

cept for the distance between the closest heavy atoms in the sidechains of TYR2 and TRP9. This is used to represent the hydrophobic core in chignolin. The distance between these two residues decreases as the trajectories progress and as the ASP3O-THR8N distance decreases as well. The other plots, where not much can be determined is likely due to the placement of the interfaces. The trajectories for this order parameter have similar configurations.

The ASP3O-GLY7N/TYR-TRP If statement order parameter path density plots shown in Figure 4.30 do show some patterns. Different distances between residues are plotted against the ASP3O-GLY7N distance which is a major part of this order parameter. Due to the interface placement, these do not show the transition from folded to unfolded, but instead give a look at different misfolded configurations.

Two major potential conformations can be seen in these plots as there is two areas of high path density in several of them. The plots of ASP3O-THR8N, GLY1-GLY7, TYR2-THR8, TYR2-TRP9, TYR2-GLY10 all have these two areas of high density. When looking at individual trajectories, there is a pattern that for each of these distances, there are two separate states. The two potential distances are not randomly in each trajectory. For example, if the ASP3O-THR8N distance is higher, the GLY1-GLY7 distance is lower and the TYR2-TRP8 distance is also higher. This pattern, and the vice versa, tracks for these trajectories. None of the other plots, except for the GLY1-GLY10 plot, show any discernible pattern. The GLY1-GLY10 plot shows a distinct decrease in this distance as the transition progresses. This in-

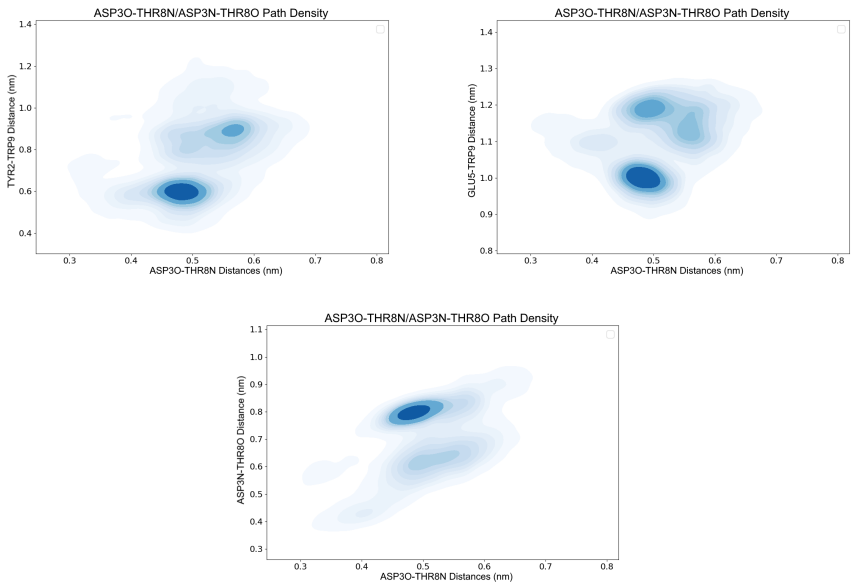


Figure 4.31: Path density plots created from all ensembles of the ASP3O-THR8N/ASP3N-THR8O order parameter. The darker areas on the plot represent where a higher concentration of frames in the trajectories are spent. There are several areas with higher density.

dicates that there could be several misfolded conformations that are more likely to form. Looking at the energy of the conformations would be interesting to compare to the folded conformation.

The ASP3O-THR8N/ASP3N-THR8O order parameter also shows some interesting patterns in the path density plots, shown in Figure 4.31. It does not represent the whole transition from unfolded to folded, but there are again some areas that show two or more areas of higher density. These could also be conformations that are frequently visited in the misfolded state. The path density plots show that while the trajectories are changing in the order parameter they are also changing in the distances between other residues, for example TYR2-TRP9 and GLU5-TRP9. Most of the residue pairs showed a change as the order parameter gets closer to the last interface. If the interfaces had been placed to allow a more thorough look at the entire transition more patterns in this changing could be found and used to help determine the folding mechanism.

4.3.2 Bond Formation Order

When looking at the LMR trajectories produced in this study, it can be seen that many of them, as previously discussed, are not properly folded. Of the ones that are folded at the end of the trajectory, some of them began folded due to poor interface placement. There are still several trajectories that do exhibit a transition to properly folded chignolin. These trajectories that have a transition in them are

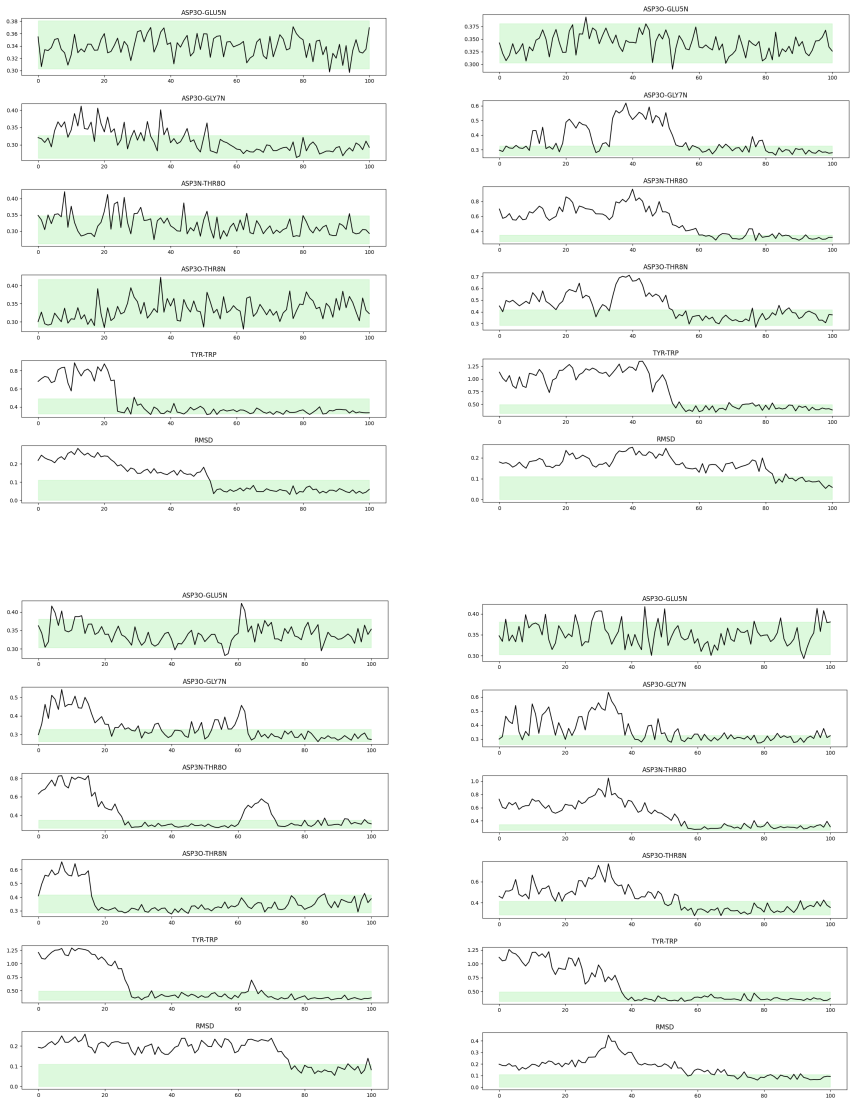


Figure 4.32: These plots are able to show the order that different hydrogen bonds and hydrophobic interactions occur. The RMSD values and the distances between the hydrogen bonds found in chignolin are plotted against the frames in each unfolded to folded region. The range of values expected when chignolin is folded, collected from a 300K MD simulation, is highlighted in green. This shows the order in which bonds are formed. The trajectories that change from a higher to a lower RMSD value are chosen to show the transition from unfolded to folded. These plots are continued in the next figure.

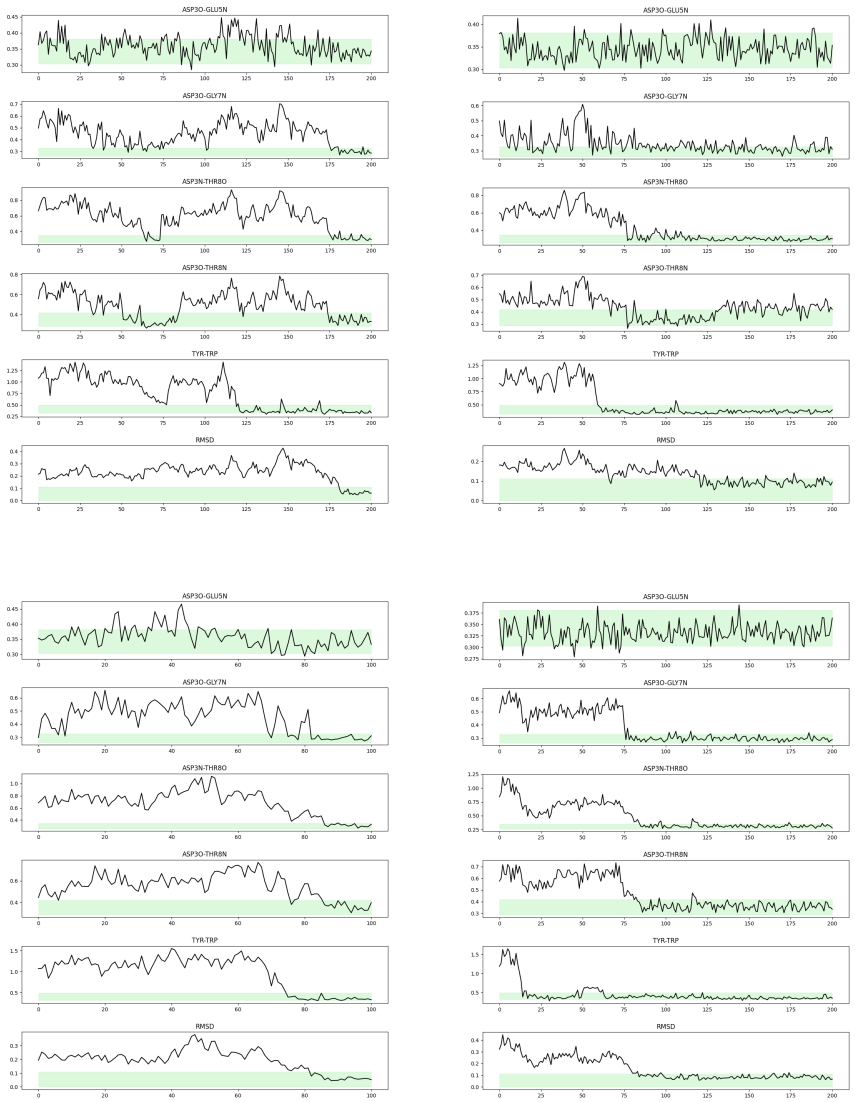


Figure 4.33: These plots are able to show the order that different hydrogen bonds and hydrophobic interactions occur. The RMSD values and the distances between the hydrogen bonds found in chignolin are plotted against the frames in each unfolded to folded region. The range of values expected when chignolin is folded, collected from a 300K MD simulation, is highlighted in green. This shows the order in which bonds are formed. The trajectories that change from a higher to a lower RMSD value are chosen to show the transition from unfolded to folded. This is a continuation from the previous figure.

shown in Figure 4.33. This allows the order of the formation of the hydrogen bonds to be compared to other factors, like the RMSD value and the hydrophobic core.

Looking at these plots, the ASP3-GLU5 bond is at the correct distance the whole time. This is likely due to the placement of the interfaces which does not require chignolin to become completely unfolded where chignolin would more resemble a straight line than a pin. Additionally, due to the placement on the molecule there is not as much range that this distance can be.

Most of the trajectories here have the 3 hydrogen bonds ASP3O-GLY7N, ASP3O-THR8N and ASP3N-THR8O reach the correct distances at approximately the same time. However, this does not always occur at the same time as when the RMSD value becomes lower than 0.18nm. This can happen before or after these bonds form. This indicates that it is not only these bonds that are important to the shape and stability of chignolin. The TYR-TRP distance also seems to usually form around the same time as the backbone bonds or a little before. Though in at least one instance it seems to have reached the appropriate distance after the backbone bonds.

For the zipper method, the expected result would have been that each hydrogen bond is formed after the other, like a zipper being pulled up. This description is not consistent with the results presented here. One thing to consider here is the resolution of the data of the transition. If the resolution is not good enough, the minute difference in the time it takes for each of the bonds to form would not be apparent. However, since the hydrogen bonds seem to form almost simultaneously, usually after the hydrophobic core is already in place, this could indicate support of the hydrophobic collapse method for the folding mechanism of chignolin. In this method the molecule collapses together and then forms the necessary bonds. An addition caveat to these results is that the whole transition is not observed from unfolded to folded. Therefore, there is a lot of changes in the molecule that happen during the transition that is not modelled here, so it is impossible to use this data to definitively conclude what model is supported in this data. It is unknown if the turn formation, hydrophobic collapse or something else entirely occurs first.

4.3.3 Properties of Folded Chignolin

In addition to looking at the transition itself, analyzing folded and unfolded chignolin can also give insights into how chignolin folds and what order parameters could potentially be used in the future. For this a tool called Protein-Ligand Interaction Profiler (PLIP) [51], PCA and decision tree classification were used.

Protein Interactions

PLIP is a tool that allows users to upload .pdb files of biological molecules and get information about the molecule and its interactions. As chignolin is one chain, the intra-chain interactions like hydrogen bonding, hydrophobic interactions and

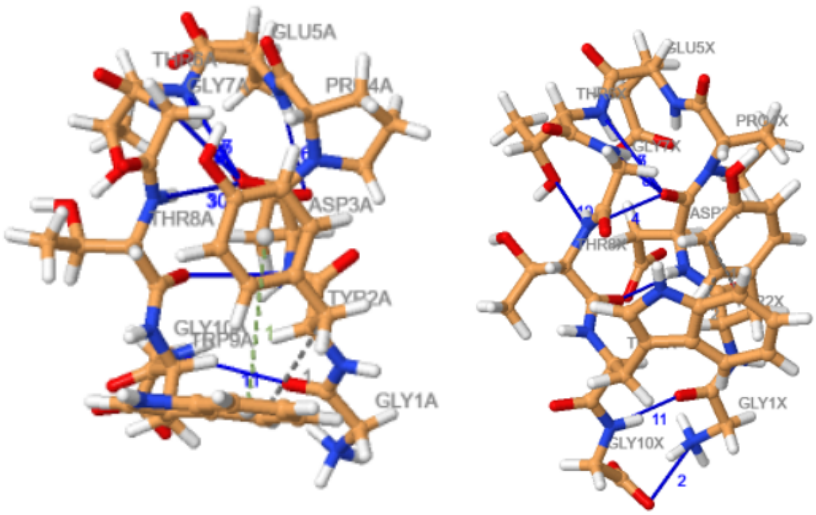


Figure 4.34: This figure shows the output of the PLIP analysis on folded chignolin. The left shows the crystal structure of chignolin acquired from PDB. Here there is π-stacking (green dotted line), hydrogen bonds (blue line) and hydrophobic interactions (grey dotted line). The right shows an example of folded chignolin from a simulation with the CHARMM27 force field. Here hydrogen bond and hydrophobic interactions are seen, but not the π-stacking. Differences between these two may come from force field properties.

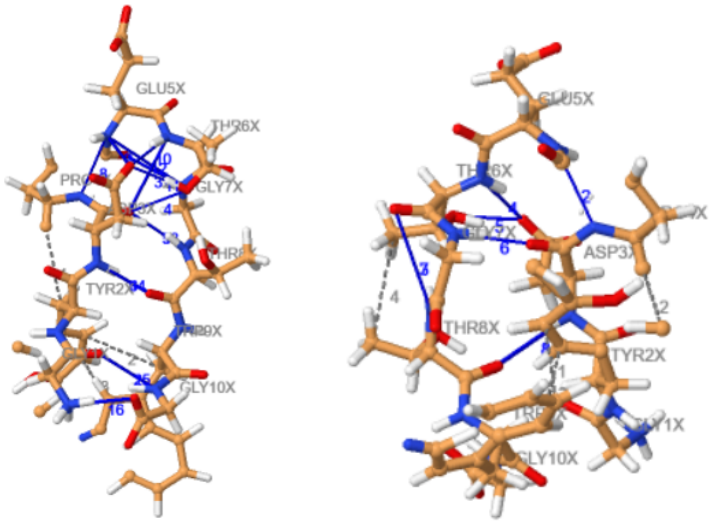


Figure 4.35: This figure shows the output of the PLIP analysis on misfolded chignolin. Here the hydrogen bonds (blue line) and hydrophobic interactions (grey dotted line) present in chignolin that is not properly folded are shown. Chignolin that is misfolded can have more or less hydrogen bonds and interactions than folded chignolin.

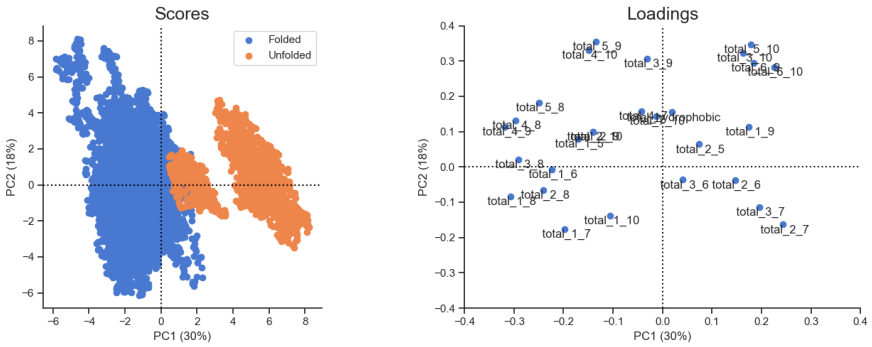


Figure 4.36: PCA on the folded and unfolded trajectories from the ASP3N-THR8O order parameter results in these scores and loadings. The distance between residues ASP3-GLY10, GLU5-GLY10, THR6-TRP9 and THR6-GLY10 are correlated with each other. Misfolded chignolin has a higher value for these distances than folded chignolin. Misfolded chignolin also has a smaller distance for the distances for residues GLY1-THR8, TYR2-THR8, GLY1-THR6, GLY1-GLY7, GLY1-GLY10 and ASP3-THR8.

π-stacking were determined. In Figure 4.34, the crystal structure of chignolin exhibits hydrogen bonds, hydrophobic interactions and π-stacking that would be suggested to be present in literature. The folded CHARMM27 chignolin does not have π-stacking. This is likely due to the properties of the CHARMM27 force field.

Figure 4.35 shows a few examples of incorrectly folded chignolin. In these examples we can see that misfolded chignolin can differ from folded chignolin in a variety of ways. In folded chignolin, PLIP usually finds 2-3 hydrophobic interactions and 10-11 hydrogen bonds. Some of these bonds were mentioned in the literature search. The results form PLIP can be found in Appendix C. In misfolded chignolin, there is much more variation in the interactions and bonds found. Some conformations have too few hydrogen bonds while others have too many. The number of hydrophobic interactions between TYR2 and TRP9 are different as well. There is no single way that the improperly folded chignolin existed.

Classification of Folded and Misfolded Chignolin

While in the simulations performed in this study, few proper transition between unfolded and folded chignolin were collected, a large amount of data of each folded and misfolded chignolin was collected. Analyzing this and and the differences between the two can lead to insights that can be used in the future to determine order parameter or even shows methods of stabilization in chignolin.

PCA was used to analyze the trajectories from the ASP3N-THR8O order parameter that fold compared to those that did not fold. This order parameter resulted a a fair amount of both folded and misfolded chignolin. The scores and loadings are seen in Figure 4.36. The RMSD values for each of these groups is different; the folded trajectories have a lower RMSD value. The general shape of the backbone of chignolin is different between folded and misfolded chignolin. The space

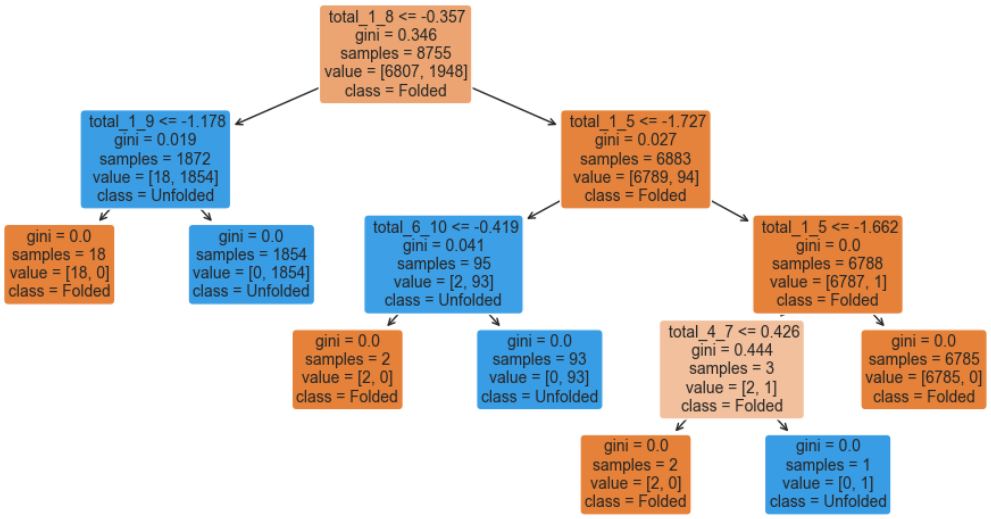


Figure 4.37: The python library sci-kit learn was used to create a decision tree to classify folded and misfolded chignolin. This decision tree uses the distances GLY1-THR8, GLY1-TRP9, GLY1-GLU5, THR6-GLY10 and PRO4-GLY7 to separate the states of chignolin. The orange color represents folded chignolin and the blue represents chignolin that is not folded. The darker the color, the more likely it is that the classification is correct.

between the two sides of chignolin is larger when chignolin is properly folded. This could indicate that the distances used for the interfaces for the order parameter were too small for chignolin to necessarily form correctly. There was also a difference in the length of one of the sides of chignolin. The distance from the 6th to the 10th residue is longer in misfolded chignolin. This could indicate a bend or angling in the chain between these residues is present in folded chignolin.

Using this data to create a decision tree results in Figure 4.37. This decision tree is more complicated than the decision tree made from the MD simulations. This decision tree is the difference between folded chignolin and chignolin that is misfolded instead of folded vs completely unfolded chignolin. This shows the parameters that ensure chignolin is properly folded and stable rather than just folded into any random configuration. The distances between residues GLY1-THR8, GLY1-TRP9, GLY1-GLU5, THR6-GLY10 and PRO4-GLY7 were used to separate the two states of chignolin. None of these distances are the hydrogen bonds that are supposed to stabilize chignolin. A few of the distances are around the ends of the molecule like GLY1-THR8 and GLY1-TRP9. These could correlate to the hydrophobic core between the residues TYR2-TRP9. Additionally, the PRO4-GLY7 distance is related to the turn in chignolin. These residues are involved in the turn that if it is a π-turn it is supposed to be folded while if it is an α-turn it is misfolded. the other two distances are the sides of chignolin GLY1-GLU5 and THR6-GLY10. The sides of chignolin when properly folded do display specific bends that allow

the proper bonds and interactions to take place. This was also noted from the PCA analysis as being an important parameter in determining folded chignolin from misfolded chignolin.

These results from PCA and decision tree analysis show that the general shape of chignolin is important like the hydrogen bonds and hydrophobic interactions are. It is not a given that if the hydrogen bonds and interactions are present that chignolin will be properly folded. Moving forward combining these non-hydrogen bond or hydrophobic interactions with these bonds and interactions would likely be able to create a better order parameter that can successfully separate folded and misfolded chignolin and calculate an accurate rate constant.

Chapter 5

Conclusion

5.1 Order Parameter

None of the order parameters attempted in this study are perfect order parameters. A single distance, or even two distances are not enough to successfully distinguish between folded and unfolded chignolin. The ASP3O-THR8N order parameter resulted in 11% folded chignolin, without the extension. The ASP3O-GLY7N order parameter extension simulations resulted in 5% folded chignolin, and the ASP3O-GLY7N/ASP3O-THR8N order parameter extension simulations resulted in about 33% of the trajectories ending with properly folded chignolin. Constraining the ASP3O-THR8N part of this order parameter increases the percent of correctly folded. The other trajectories were misfolded.

An order parameter would likely need to contain parameters for the shape of the molecule, perhaps using the angles between residues, as well the hydrogen bonds. The RMSD value could also be used for this, but it alone is not specific enough to guarantee folded chignolin. The RMSD value could perhaps be combined with other import factors to make an order parameter.

The difficulty with finding an order parameter for this relatively small and uncomplicated biological system shows the importance of finding general order parameters for biological systems like the fraction of native contacts which unfortunately cannot be used in a system this small. A general order parameter would improve the ease at which RETIS could be used to simulate proteins.

The interfaces used for the order parameters in this study were not ideally placed for proper sampling of the entire transition from folded to unfolded. It is important to ensure that the first interface is in state A, where it can fully explore the unfolded region. The last interface also needs to be placed ideally in the basin of attraction of state B, or the folded state. This was not achieved in this study. In future experiments, it would be wise to ensure this early in the process of test in order parameters. This incorrect interface placement also limits what can be determined about the folding mechanism, given that chignolin does not start completely unfolded and instead is already in an intermediate state. However, the rate constant can still be calculated given RETIS uses the flux calculated from the

[0-] and [0+] ensembles to correct for this kind of issue.

5.2 Rate Constant

Due to the issues with the order parameter a specific rate constant could not be accurately calculated in PyRETIS, but an estimation could be made. With the adjusted rate constants calculated with the percentage of trajectories that fold or that fold upon extension, an estimation of between $2.776 \times 10^{-4} ps^{-1}$ and $1.021 \times 10^{-5} ps^{-1}$ can be made from different order parameters. This value is still affected by effects like potential temperature effects and effects from the lack of exploration of the unfolded region. However, this result is in reasonable agreement with the value of previous experiments which is in the order of $1 \times 10^{-5} ps^{-1}$.

Ideally, the rate constant would be calculated from longer simulations with more cycles. Due to time and computing constraints, not all of the order parameters used were able to have a rate constant calculated from enough cycles that the rate constant had stopped changing after each new cycle and plateaued. This can be improved with more time and simulation cycles.

5.3 Folding Mechanism

The LMR trajectories from all order parameters were collected and used for an analysis to determine information about the folding mechanism of chignolin. The path density plots did show some areas of higher density, unfortunately these are not folded chignolin due to the small number of trajectories that properly fold chignolin. These could be instead common misfolded configurations. Information about these conformations can be used to determine an order parameter that successfully separates folded and misfolded chignolin.

This study was not able to definitively determine a hydrogen bond formation order in part due to the small number of properly folding trajectories. The majority of the trajectories that fold seem to have the ASP3O-GLY7N, ASP3O-THR8N and ASP3N-THR8O hydrogen bonds form simultaneously at this resolution. There were several instances where the TYR2-TRP9 hydrophobic core may have formed first or at the same time as the hydrogen bonds. The RMSD value does not seem to depend on the hydrogen bond formation and more generally depends on the shape of chignolin. All of these factors are required for chignolin to properly fold.

PCA and decision tree analysis of the folded vs misfolded configurations allowed a closer look at the shape of the residues that is important in properly folded chignolin. The majority of the trajectories ended in a misfolded state that was similar in shape to the folded state, but was not stable. This analysis showed that in folded chignolin there is a bend between the residues ASP6-GLY10 that in not present while chignolin is misfolded. Also, chignolin, while still shaped like a pin, has the sides further apart when properly folded rather than misfolded. There

are several other distance requirements in chignolin that do not directly relate to the hydrogen bonds or the hydrophobic interactions.

5.4 Further Work

More research should still be done to determine a good order parameter for chignolin, and potentially other small β-hairpin turns or proteins more generally. The work here eliminates some possibilities and further sheds light on potential future order parameters to be used in RETIS simulations of chignolin. Single distances are not specific enough to successfully distinguish folded and unfolded chignolin. Using two distances was mildly more successful and so was the using RMSD in combination with other factors. Combining something like the RMSD value with more strict parameters regarding certain distances or angles may be a good place to start developing new order parameters.

Finding a universal order parameter, like fraction of native contacts, that can also be applied to small proteins, such as chignolin, would be a benefit. Determining order parameters for every system, small protein or otherwise, that is simulated is time consuming and potentially obfuscates any trends in protein folding in general. This difficulty limits the application of techniques like RETIS which could further research on protein folding in general.

Further analysis of this data could include looking further into trajectories with a low RMSD value can provide further insight into the important factors that go into distinguishing folded and unfolded chignolin. Seeing what bonds are and are not present in the majority of folded and unfolded configurations could help deduce which bonds are most influential in stabilizing chignolin. In addition to bonds, the angles between the residues that give chignolin its shape should also be explored.

Additionally more research could be done into how the force field chosen affects the simulation of chignolin. The properties of a given force field can drastically affect behavior as could be seen in some of the results presented here, notably the lack of π-stacking. Comparing these results to the force fields could help better understand how these results apply to chignolin in bench experiments.

Once a good order parameter is determine, a rate constant calculation can also be more accurately performed. This should then be compared to literature values for simulations as well as to experimental results. More trajectories that exhibit a transition can be collected and analyzed for a more complete look at the entire transition from unfolded to folded chignolin. After that, what is learned from this model system can be applied to other β-hairpin turns and proteins.

Bibliography

[1] S. Honda, K. Yamasaki, Y. Sawada and H. Morii, '10 Residue Folded Peptide Designed by Segment Statistics,' en, *Structure*, vol. 12, no. 8, pp. 1507–1518, Aug. 2004, ISSN: 09692126. DOI: 10.1016/j.str.2004.05.022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0969212604002424 (visited on 05/04/2022).

[2] K. Lindorff-Larsen, S. Piana, R. O. Dror and D. E. Shaw, 'How Fast-Folding Proteins Fold,' en, *Science*, vol. 334, no. 6055, pp. 517–520, Oct. 2011, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1208351. [Online]. Available: https://www.sciencemag.org/lookup/doi/10.1126/science.1208351 (visited on 05/04/2022).

[3] J. Lee and S. Shin, 'Understanding β-Hairpin Formation by Molecular Dynamics Simulations of Unfolding,' en, *Biophysical Journal*, vol. 81, no. 5, pp. 2507–2516, Nov. 2001, ISSN: 00063495. DOI: 10.1016/S0006-3495(01)75896-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006349501758961 (visited on 05/04/2022).

[4] B. Zagrovic, E. J. Sorin and V. Pande, 'β-hairpin folding simulations in atomistic detail using an implicit solvent model 1 1Edited by F. Cohen,' en, *Journal of Molecular Biology*, vol. 313, no. 1, pp. 151–169, Oct. 2001, ISSN: 00222836. DOI: 10.1006/jmbi.2001.5033. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022283601950338 (visited on 05/04/2022).

[5] S. Enemark and R. Rajagopalan, 'Turn-directed folding dynamics of -hairpin-forming de novo decapeptide chignolin,' *Phys. Chem. Chem. Phys.*, vol. 14, pp. 12 442–12 450, 36 2012. DOI: 10.1039/C2CP40285H. [Online]. Available: http://dx.doi.org/10.1039/C2CP40285H.

[6] Y. Maruyama and A. Mitsutake, 'Analysis of Structural Stability of Chignolin,' en, *The Journal of Physical Chemistry B*, vol. 122, no. 14, pp. 3801–3814, Apr. 2018, ISSN: 1520-6106, 1520-5207. DOI: 10.1021/acs.jpcb.8b00288. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.jpcb.8b00288 (visited on 10/04/2022).

[7] Y. Maruyama, S. Koroku, M. Imai, K. Takeuchi and A. Mitsutake, 'Mutation-induced change in chignolin stability from -turn to -turn,' en, *RSC Advances*, vol. 10, no. 38, pp. 22 797–22 808, 2020, ISSN: 2046-2069. DOI: 10.1039/D0RA01148G. [Online]. Available: http://xlink.rsc.org/?DOI=D0RA01148G (visited on 10/04/2022).

[8] P. Shaffer, O. Valsson and M. Parrinello, 'Enhanced, targeted sampling of high-dimensional free-energy landscapes using variationally enhanced sampling, with an application to chignolin,' en, *Proceedings of the National Academy of Sciences*, vol. 113, no. 5, pp. 1150–1155, Feb. 2016, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1519712113. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1519712113 (visited on 05/04/2022).

[9] P. Kührová, A. De Simone, M. Otyepka and R. B. Best, 'Force-Field Dependence of Chignolin Folding and Misfolding: Comparison with Experiment and Redesign,' en, *Biophysical Journal*, vol. 102, no. 8, pp. 1897–1906, Apr. 2012, ISSN: 00063495. DOI: 10.1016/j.bpj.2012.03.024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006349512003359 (visited on 05/04/2022).

[10] R. Harada and A. Kitao, 'Exploring the folding free energy landscape of a -hairpin miniprotein, chignolin, using multiscale free energy landscape calculation method,' *The Journal of Physical Chemistry B*, vol. 115, no. 27, pp. 8806–8812, 2011. DOI: 10.1021/jp2008623.

[11] Y. Miao, F. Feixas, C. Eun and J. A. McCammon, 'Accelerated molecular dynamics simulations of protein folding,' *Journal of Computational Chemistry*, vol. 36, no. 20, pp. 1536–1549, 2015. DOI: 10.1002/jcc.23964.

[12] Z. F. Brotzakis and P. G. Bolhuis, 'A one-way shooting algorithm for transition path sampling of asymmetric barriers,' en, *The Journal of Chemical Physics*, vol. 145, no. 16, p. 164 112, Oct. 2016, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4965882. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.4965882 (visited on 05/04/2022).

[13] H. Fujisaki, K. Moritsugu and Y. Matsunaga, 'Exploring Configuration Space and Path Space of Biomolecules Using Enhanced Sampling Techniques—Searching for Mechanism and Kinetics of Biomolecular Functions,' en, *International Journal of Molecular Sciences*, vol. 19, no. 10, p. 3177, Oct. 2018, ISSN: 1422-0067. DOI: 10.3390/ijms19103177. [Online]. Available: http://www.mdpi.com/1422-0067/19/10/3177 (visited on 05/04/2022).

[14] A. Kolinski, B. Ilkowski and J. Skolnick, 'Dynamics and Thermodynamics of β-Hairpin Assembly: Insights from Various Simulation Techniques,' en, *Biophysical Journal*, vol. 77, no. 6, pp. 2942–2952, Dec. 1999, ISSN: 00063495. DOI: 10.1016/S0006-3495(99)77127-4. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0006349599771274 (visited on 05/04/2022).

[15] V. Muñoz, P. A. Thompson, J. Hofrichter and W. A. Eaton, 'Folding dynamics and mechanism of β-hairpin formation,' en, *Nature*, vol. 390, no. 6656, pp. 196–199, Nov. 1997, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/36626. [Online]. Available: http://www.nature.com/articles/36626 (visited on 05/04/2022).

[16] S.-H. Ahn, A. Ojha, R. Amaro and J. McCammon, 'Gaussian accelerated molecular dynamics with the weighted ensemble method: A hybrid method improves thermodynamics and kinetics sampling,' 2021. DOI: 10.33774/chemrxiv-2021-2j2zr.

[17] A. Suenaga, T. Narumi, N. Futatsugi, R. Yanai, Y. Ohno, N. Okimoto and M. Taiji, 'Folding dynamics of 10-residue -hairpin peptide chignolin,' *Chemistry – An Asian Journal*, vol. 2, no. 5, pp. 591–598, 2007. DOI: 10.1002/asia.200600385.

[18] D. van der Spoel and M. M. Seibert, 'Protein folding kinetics and thermodynamics from atomistic simulations,' *Physical Review Letters*, vol. 96, no. 23, 2006. DOI: 10.1103/physrevlett.96.238102.

[19] A. Mitsutake and H. Takano, 'Relaxation mode analysis and markov state relaxation mode analysis for chignolin in aqueous solution near a transition temperature,' *The Journal of Chemical Physics*, vol. 143, no. 12, p. 124 111, 2015. DOI: 10.1063/1.4931813.

[20] H. Wu, M. R. Ghaani, Z. Futera and N. J. English, 'Effects of externally applied electric fields on the manipulation of solvated-chignolin folding: Static- versus alternating-field dichotomy at play,' *The Journal of Physical Chemistry B*, vol. 126, no. 2, pp. 376–386, 2022. DOI: 10.1021/acs.jpcb.1c06857.

[21] V. Muñoz, E. R. Henry, J. Hofrichter and W. A. Eaton, 'A statistical mechanical model for β-hairpin kinetics,' en, *Proceedings of the National Academy of Sciences*, vol. 95, no. 11, pp. 5872–5879, May 1998, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.95.11.5872. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.95.11.5872 (visited on 05/04/2022).

[22] J. Tsai and M. Levitt, 'Evidence of turn and salt bridge contributions to β-hairpin stability: MD simulations of C-terminal fragment from the B1 domain of protein G,' en, *Biophysical Chemistry*, vol. 101-102, pp. 187–201, Dec. 2002, ISSN: 03014622. DOI: 10.1016/S0301-4622(02)00198-9. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0301462202001989 (visited on 05/04/2022).

[23] T. S. van Erp and P. G. Bolhuis, 'Elaborating transition interface sampling methods,' *ChemInform*, vol. 35, no. 47, 2004. DOI: 10.1002/chin.200447267.

[24] T. S. van Erp, D. Moroni and P. G. Bolhuis, 'A novel path sampling method for the calculation of rate constants,' *The Journal of Chemical Physics*, vol. 118, no. 17, pp. 7762–7774, 2003. DOI: 10.1063/1.1562614. eprint: https://doi.org/10.1063/1.1562614. [Online]. Available: https://doi.org/10.1063/1.1562614.

[25] C. Dellago, P. G. Bolhuis, F. S. Csajka and D. Chandler, 'Transition path sampling and the calculation of rate constants,' *The Journal of Chemical Physics*, vol. 108, no. 5, pp. 1964–1977, 1998. DOI: 10.1063/1.475562.

[26] E. Riccardi, A. Lervik, S. Roet, O. Aarøen and T. S. Erp, 'Pyretis 2: An improbability drive for rare events,' *Journal of Computational Chemistry*, vol. 41, no. 4, pp. 370–377, 2019. DOI: 10.1002/jcc.26112.

[27] R. B. Best, G. Hummer and W. A. Eaton, 'Native contacts determine protein folding mechanisms in atomistic simulations,' en, *Proceedings of the National Academy of Sciences*, vol. 110, no. 44, pp. 17 874–17 879, Oct. 2013, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1311599110. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1311599110 (visited on 05/04/2022).

[28] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles and S. C. Wang, 'Anton, a special-purpose machine for molecular dynamics simulation,' en, *Communications of the ACM*, vol. 51, no. 7, pp. 91–97, Jul. 2008, ISSN: 0001-0782, 1557-7317. DOI: 10.1145/1364782.1364802. [Online]. Available: https://dl.acm.org/doi/10.1145/1364782.1364802 (visited on 05/04/2022).

[29] D. Wang, Y. Wang, J. Chang, L. Zhang, H. Wang and W. E., 'Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics,' en, *Nature Computational Science*, vol. 2, no. 1, pp. 20–29, Jan. 2022, ISSN: 2662-8457. DOI: 10.1038/s43588-021-00173-1. [Online]. Available: https://www.nature.com/articles/s43588-021-00173-1 (visited on 05/04/2022).

[30] R. Qi, G. Wei, B. Ma and R. Nussinov, 'Replica Exchange Molecular Dynamics: A Practical Application Protocol with Solutions to Common Problems and a Peptide Aggregation and Self-Assembly Example,' in *Peptide Self-Assembly*, B. L. Nilsson and T. M. Doran, Eds., vol. 1777, New York, NY: Springer New York, 2018, pp. 101–119, ISBN: 9781493978090. DOI: 10.1007/978-1-4939-7811-3_5. [Online]. Available: http://link.springer.com/10.1007/978-1-4939-7811-3_5 (visited on 05/04/2022).

[31] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé and C. Clementi, 'Machine Learning of Coarse-Grained Molecular Dynamics Force Fields,' en, *ACS Central Science*, vol. 5, no. 5, pp. 755–

767, May 2019, ISSN: 2374-7943, 2374-7951. DOI: 10.1021/acscentsci.8b00913. [Online]. Available: https://pubs.acs.org/doi/10.1021/acscentsci.8b00913 (visited on 05/04/2022).

[32] J. Wang, N. Charron, B. Husic, S. Olsson, F. Noé and C. Clementi, 'Multi-body effects in a coarse-grained protein force field,' en, *The Journal of Chemical Physics*, vol. 154, no. 16, p. 164 113, Apr. 2021, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/5.0041022. [Online]. Available: https://aip.scitation.org/doi/10.1063/5.0041022 (visited on 05/04/2022).

[33] P. K. Depa and J. K. Maranas, 'Speed up of dynamic observables in coarse-grained molecular-dynamics simulations of unentangled polymers,' *The Journal of Chemical Physics*, vol. 123, no. 9, p. 094 901, 2005. DOI: 10.1063/1.1997150.

[34] A. Laio and M. Parrinello, 'Escaping free-energy minima,' en, *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 562–12 566, Oct. 2002, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.202427399. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.202427399 (visited on 05/04/2022).

[35] Mar. 2022. [Online]. Available: https://foldingathome.org/.

[36] K. A. McKiernan, B. E. Husic and V. S. Pande, 'Modeling the mechanism of CLN025 beta-hairpin formation,' en, *The Journal of Chemical Physics*, vol. 147, no. 10, p. 104 107, Sep. 2017, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4993207. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.4993207 (visited on 05/04/2022).

[37] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, 2nd ed, ser. Computational science series 1. San Diego: Academic Press, 2002, ISBN: 9780122673511.

[38] T. S. van Erp, *Dynamical rare event simulation techniques for equilibrium and non-equilibrium systems*, 2011. DOI: 10.48550/ARXIV.1101.0927. [Online]. Available: https://arxiv.org/abs/1101.0927.

[39] C. Dellago and P. G. Bolhuis, 'Transition path sampling simulations of biological systems,' in *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*, M. Reiher, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 291–317, ISBN: 978-3-540-38085-6. DOI: 10.1007/128_085. [Online]. Available: https://doi.org/10.1007/128_085.

[40] C. Dellago and P. G. Bolhuis, 'Activation Energies from Transition Path Sampling Simulations,' en, *Molecular Simulation*, vol. 30, no. 11-12, pp. 795–799, Sep. 2004, ISSN: 0892-7022, 1029-0435. DOI: 10.1080/08927020412331294869. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08927020412331294869 (visited on 05/04/2022).

[41] R. Cabriolu, K. M. Skjelbred Refsnes, P. G. Bolhuis and T. S. van Erp, 'Foundations and latest advances in replica exchange transition interface sampling,' en, *The Journal of Chemical Physics*, vol. 147, no. 15, p. 152 722, Oct. 2017, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.4989844. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.4989844 (visited on 09/04/2022).

[42] A. S. Kamenik, P. H. Handle, F. Hofer, U. Kahler, J. Kraml and K. R. Liedl, 'Polarizable and non-polarizable force fields: Protein folding, unfolding, and misfolding,' en, *The Journal of Chemical Physics*, vol. 153, no. 18, p. 185 102, Nov. 2020, ISSN: 0021-9606, 1089-7690. DOI: 10.1063/5.0022135. [Online]. Available: http://aip.scitation.org/doi/10.1063/5.0022135 (visited on 10/04/2022).

[43] M. J. Robertson, J. Tirado-Rives and W. L. Jorgensen, 'Improved peptide and protein torsional energetics with the opls-aa force field,' *Journal of Chemical Theory and Computation*, vol. 11, no. 7, pp. 3499–3509, 2015. DOI: 10.1021/acs.jctc.5b00356.

[44] N. Foloppe and A. D. MacKerell Jr., 'All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data,' *Journal of Computational Chemistry*, vol. 21, no. 2, pp. 86–104, 2000. DOI: 10.1002/(sici)1096-987x(20000130)21:2<86::aid-jcc2>3.0.co;2-g.

[45] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, 'Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,' *SoftwareX*, vol. 1-2, pp. 19–25, 2015. DOI: 10.1016/j.softx.2015.06.001.

[46] M. Själander, M. Jahre, G. Tufte and N. Reissmann, *EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure*, 2019. arXiv: 1912.05848 [cs.DC].

[47] W. Humphrey, A. Dalke and K. Schulten, 'VMD – Visual Molecular Dynamics,' *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[48] W. Kabsch, 'A solution for the best rotation to relate two sets of vectors,' *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, 1976. DOI: 10.1107/s0567739476001873.

[49] W. Kabsch, 'A discussion of the solution for the best rotation to relate two sets of vectors,' *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, 1978. DOI: 10.1107/s0567739478001680.

[50] Z. F. Brotzakis, M. Vendruscolo and P. G. Bolhuis, 'A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations,' *Proceedings of the National Academy of Sciences*, vol. 118, no. 2, 2020. DOI: 10.1073/pnas.2012423118.

[51] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt and M. Schroeder, 'PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA,' *Nucleic Acids Research*, vol. 49, no. W1, W530–W534, May 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkab294. eprint: https://academic.oup.com/nar/article-pdf/49/W1/W530/38841758/gkab294.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkab294.

Appendix A

RETIS Input Files

Code listing A.1 shows an input file for the ASP3O-GLY7N order parameter. The interfaces are given under "Simulation". Under "Engine settings", the timestep and the subcycles are given. Also whether gromacs or gromacs2 is used is selected here; this input file specifies gromacs, the original engine. In the "TIS settings" and "RETIS settings" the frequency of different moves are specified. The initial path used is retrieved from a .trr file and how to calculate the order parameter is retrieved from a file called "op.py" specified under the "Orderparameter" section. The path to where the order parameter program is stored and the name of the function to calculate the order parameter are specified there. "Output" determines how often the trajectory files are saved and how often data about the order parameter and energy are written to a file.

Code listing A.1: RETIS Input File

```
Chignolin
================

Simulation
----------
task = retis
steps = 10000
interfaces = [-0.5, -0.47, -0.38, -0.30]


System
------
units = gromacs

Particles
---------
position = {'file': 'extended.gro'}
velocity = {'generate': 'maxwell',
            'temperature': 2.0,
            'momentum': True,
            'seed': 0}
name = ['Chignolin', 'Water']
type = [0]

Forcefield
```

```
----------
description = CHARMM

Engine settings
---------------
class = gromacs
gmx = gmx_mpi
mdrun = gmx_mpi mdrun
input_path = gromacs_input
timestep = 0.002
subcycles = 5
maxwarn = 1
write_vel = True
write_force = False
gmx_format = gro

TIS settings
------------
freq =  0.5
maxlength = 200000
aimless = True
allowmaxlength = False
zero_momentum = False
rescale_energy = False
sigma_v =  -1
seed = 0


RETIS settings
--------------
swapfreq = 0.5
relative_shoots = None
nullmoves = True
swapsimul = True


Initial-path settings
---------------------
method = load
load_folder = load


Orderparameter
--------------
class = ASP_GLY
module = 'op.py'


Output
------
trajectory-file = 1
order-file = 1
energy-file = 10
```

Code listing A.2 shows the input file for the RMSD order parameter. This input file shows the same parameters as above in the "Simulation" section. The engine here is gromacs2 rather than gromacs; how to change this is under the "Engine settings" section. In this simulation there was only shooting moves, so the frequencies of swapping and other moves are 0. Additional information required for the order parameter, like the reference structure for RMSD, can also be specified in the "Orderparameter" section.

Code listing A.2: RETIS Input File with RMSD as Order Parameter

```
Chignolin
=================

Simulation
----------
task = retis
steps = 100
interfaces = [-0.25, -0.15, -0.1]


System
------
units = gromacs

Particles
---------
position = {'file': 'extended.gro'}
velocity = {'generate': 'maxwell',
            'temperature': 2.0,
            'momentum': True,
            'seed': 0}
name = ['Chignolin', 'Water']
type = [0]

Forcefield
----------
description = CHARMM

Engine settings
---------------
class = gromacs2
gmx = gmx_mpi
mdrun = gmx_mpi mdrun
input_path = gromacs_input
timestep = 0.002
subcycles = 5
maxwarn = 1
write_vel = True
write_force = False
gmx_format = gro

TIS settings
------------
freq =  0.0
maxlength = 20000
aimless = True
allowmaxlength = False
zero_momentum = False
```

```
rescale_energy = False
sigma_v =  -1
seed = 0


RETIS settings
--------------
swapfreq = 0.0
relative_shoots = None
nullmoves = True
swapsimul = True


Initial-path settings
---------------------
method = load
load_folder = load
load_and_kick = True

Orderparameter
--------------
class = RmsdOrderParameter
module = 'rmsd_op.py'
reference_system_path = 'ref.gro'
periodic = True


Output
------
trajectory-file = 1
order-file = 1
energy-file = 100
```

Appendix B

Order Parameter Programs

The original code for the order parameters was taken from PyRETIS and adapted to suit the needs of this experiment. Code listing B.1 shows the Negative distance order parameter. The indexes of the atoms to calculate the distance between is retrieved from the RETIS input files seen above in Appendix A.

The order parameter in Code listing B.2 is an example of the order parameter using an if statement to check two distances separately. The indexes of the atoms involved in both distances are given in the RETIS input file. The distances for the first distance are set in the input file as well. The distance requirement for the the second distance is set in the order parameter here. Once that is met, the order parameter is set to 0, so that it crosses the final interface.

The order parameter in Code listing B.3 is based on the Kabsch algorithm for calculating RMSD. The path to the reference structure is taken from the RETIS input file. The other parameters needed are retrieved from the PyRETIS system. This program fixes molecules broken over the periodic boundary. This code applies to systems with a cubic box.

Code listing B.1: Negative Distance Order Parameter

```
import sys
import numpy as np
from pyretis.orderparameter import OrderParameter

class Negative_Distance(OrderParameter):
    """A distance order parameter.

    This class defines a very simple order parameter which is just
    the negative of the scalar distance between two particles.

    Attributes
    ----------
    index : tuple of integers
        These are the indices used for the two particles.
        `system.particles.pos[index[0]]` and
        `system.particles.pos[index[1]]` will be used.
    periodic : boolean
        This determines if periodic boundaries should be applied to
        the distance or not.
```

```
    """

    def __init__(self, index, periodic=True):
        """Initialise order parameter.

        Parameters
        ----------
        index : tuple of ints
            This is the indices of the atom we will use the position of.
        periodic : boolean, optional
            This determines if periodic boundary conditions should be
            applied to the position.

        """

        pbc = 'Periodic' if periodic else 'Non-periodic'
        txt = f'{pbc} distance, particles {index[0]} and {index[1]}'
        super().__init__(description=txt, velocity=False)
        self.periodic = periodic
        self.index = index

    def calculate(self, system):
        """Calculate the order parameter.

        Here, the order parameter is just the negative of the distance between two
        particles.

        Parameters
        ----------
        system : object like :py:class:`.System`
            The object containing the positions and box used for the
            calculation.

        Returns
        -------
        out : list of floats
            The negative distance order parameter.

        """
        particles = system.particles
        delta = particles.pos[self.index[1]] - particles.pos[self.index[0]]
        if self.periodic:
            delta = system.box.pbc_dist_coordinate(delta)
        lamb = np.sqrt(np.dot(delta, delta))
        lamb2 = lamb*-1
        return [lamb2]
```

Code listing B.2: If Statement Order Parameter

```
import sys
import numpy as np
from pyretis.orderparameter import OrderParameter

class If_Statement(OrderParameter):
    def __init__(self, index, index2, periodic=True):
        pbc = 'Periodic' if periodic else 'Non-periodic'
        txt = f'{pbc} distance, particles {index[0]} and {index[1]}'
        super().__init__(description=txt, velocity=False)
        self.periodic = periodic
        self.index = index
        self.index2 = index2

    def calculate(self, system):
        particles = system.particles
        delta = particles.pos[self.index[1]] - particles.pos[self.index[0]]
        delta2 = particles.pos[self.index2[1]] - particles.pos[self.index2[0]]
        if self.periodic:
            delta = system.box.pbc_dist_coordinate(delta)
            delta2 = system.box.pbc_dist_coordinate(delta2)
        lamb = np.sqrt(np.dot(delta, delta))
        lamb2 = lamb*-1
        lamb3 = np.sqrt(np.dot(delta2, delta2))
        lamb4 = lamb3 *-1
        lamb5 = lamb2
        if lamb2 >= -0.301:
            if lamb4 >= -0.5:
                lamb5 = 0
        return [lamb5]
```

Code listing B.3: RMSD Order Parameter

```
import os
import sys
import numpy as np
from numpy.linalg import svd, det
from pyretis.orderparameter import OrderParameter
from pyretis.inout.formats.gromacs import (
    read_gromacs_gro_file,
    write_gromacs_gro_file,
)

class RmsdOrderParameter(OrderParameter):
    def __init__(self, reference_system_path, periodic=True):
        """
        Initialise order parameter.

        Parameters
        ----------
        reference_system_path: string
            Absolute path to the reference system .gro file.
            This will determine the distances.

        periodic : boolean, optional
            This determines if periodic boundary conditions should be
            applied to the position.
        """

        pbc = 'Periodic' if periodic else 'Non-periodic'
        txt = f'{pbc} distance'
        super().__init__(description=txt, velocity=False)
        self.periodic = periodic

        self.reference_system = read_gromacs_gro_file(reference_system_path)

    def calculate(self, system):
        # box1 has to be format [size_x, size_y, size_z]
        xyz1, box1 = self.get_values_from_system(system)
        xyz1 = np.array([xyz1[i] for i in range(138)])
        xyz1 = self.fix_broken_brute(xyz1, box1)
        frame2, xyz2, vel2, box2 = self.reference_system  # reference

        idx = [4,11,32,46,58,73,87,94,108,132]
        xyz1 = np.array([xyz1[id] for id in idx])
        xyz2 = np.array([xyz2[id] for id in idx])

        xyz1 = xyz1 - self.get_com(xyz1)
        xyz2 = xyz2 - self.get_com(xyz2)
        rotation = self.kabsch(xyz1, xyz2)
        new_xyz = np.dot(xyz1, rotation)

        diff = new_xyz - xyz2
        rmsd2 = np.sqrt((diff * diff).sum() / len(idx))
        return [rmsd2 * -1]

    def get_values_from_system(self, system):
        """
        Read position and box from system.

        Parameters
```

```
        ----------
        system: pyretis.core.System
            The current system received from Pyretis.

        Returns
        ----------
        box: numpy.array
            Box dimensions extracted from Pyretis box cell.
        """

        if not system.particles.pos.any() or len(system.particles.pos[0]) != 3:
            raise Exception("Particle␣positions␣were␣absent␣or␣invalid.")

        if not system.box.cell.any() or len(system.box.cell) != 3:
            raise Exception("System␣box␣was␣absent␣or␣invalid.")

        return system.particles.pos, system.box.cell #np.transpose(particles?)

    def fix_broken_brute(self, xyz, box):
        """Fix broken coordinates by trying all possibilities - so not too smart..."""
        box_length = np.array([box[0], box[1], box[2]])  # box lengths for x, y, and z.
        inv_box_length = 1.0 / box_length  # inverse box length
        half_length = 0.5 * box_length  # half box length
        # assume atom no 0 is correct
        new_xyz = []
        for idx, xyzi in enumerate(xyz):
            if idx == 0:  # Skip first atom
                new_xyz.append(xyzi)
                continue
            distance = xyzi - xyz[0]  # Distance vector
            k = np.where(np.abs(distance) > half_length)[0]
            distance[k] -= np.rint(distance[k] * inv_box_length[k]) * box_length[k]
# PBC distance vector
            new_xyz.append(xyz[0] + distance)
        return np.array(new_xyz)

    def kabsch(self, mobile, target):
        """
        Find optimal rotation matrix of pmat unto qmat.

        See: https://en.wikipedia.org/wiki/Kabsch_algorithm
        """
        # 2) Covariance matrix:
        cov = np.dot(np.transpose(mobile), target)
        # 3) Find rotation:
        V, _, Wt = svd(cov)
        if det(V) * det(Wt) < 0.0:
            V[:, -1] *= -1
        rotation = np.dot(V, Wt)
        return rotation

    def get_com(self, xyz):
        """Return geometric center."""
        return np.mean(xyz, axis=0)
```

Appendix C

PLIP Folded Chignolin Results

Prediction of noncovalent interactions for PDB structure PROTEIN

Created on 2022/05/28 using PLIP v2.2.2

If you are using PLIP in your work, please cite: Adasme,M. et al. PLIP 2021: expanding the scope of the protein-ligand interaction profiler to DNA and RNA. Nucl. Acids Res. (05 May 2021), gkab294. doi: 10.1093/nar/gkab294 Analysis was done on model 1.

GLY:X:1 (GLY-TYR-ASP-PRO-GLU-THR-GLY-THR-TRP-GLY) - INTRA

+ TYR:X:2
+ ASP:X:3
+ PRO:X:4
+ GLU:X:5
+ THR:X:6
+ GLY:X:7
+ THR:X:8
+ TRP:X:9
+ GLY:X:10

Interacting chain(s): X

(See interactions on the next pages. The first is folded chignolin, and the second is an example of misfolded chignolin.)

Hydrophobic Interactions

RESNR	RESTYPE	RESCHAIN	RESNR_LIG	RESTYPE_LIG	RESCHAIN_LIG	DIST	LIGCARBONIDX	PROTCARBONIDX	LIGCOO	PROTCOO
2	TYR	X	4	PRO	X	3.67	49	20	22.640, 20.520, 28.060	25.830, 22.110, 27.180
2	TYR	X	9	TRP	X	3.52	121	14	28.580, 26.950, 24.330	25.610, 25.760, 25.790
2	TYR	X	9	TRP	X	3.47	120	25	29.000, 25.740, 23.780	26.570, 23.530, 24.910

Hydrogen Bonds

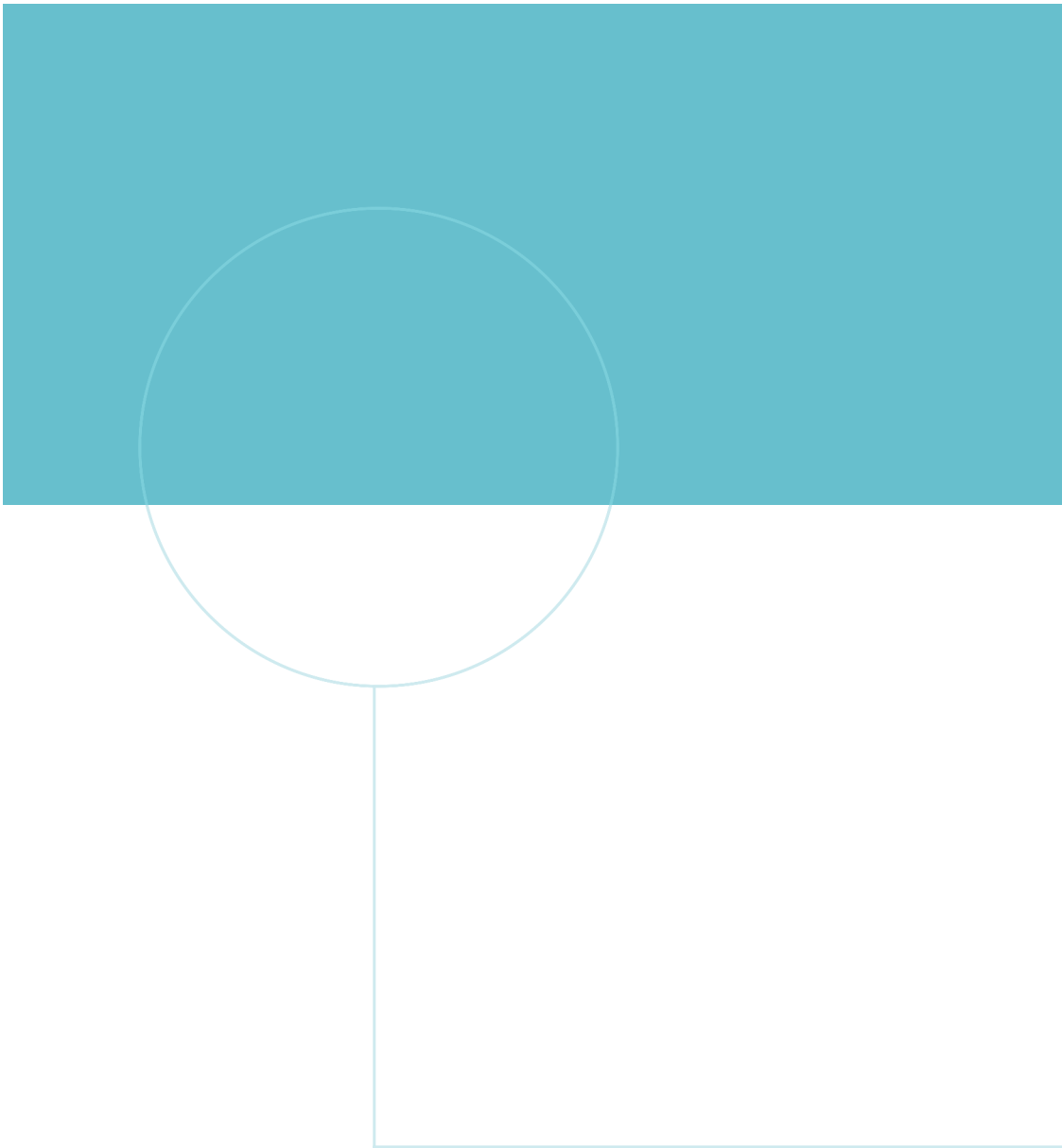
RESNR	RESTYPE	RESCHAIN	RESNR_LIG	RESTYPE_LIG	RESCHAIN_LIG	SIDECHAIN	DIST_H-A	DIST_D-A	DON_ANGLE	PROTISDON	DONORIDX	DONORTYPE	ACCEPTORIDX	ACCEPTORTYPE	LIGCOO	PROTCOO
1	GLY	X	10	GLY	X	False	1.88	2.89	161.84	True	1	N3+	137	O.co2	24.630, 30.190, 20.340	22.650, 30.040, 22.440
1	GLY	X	10	GLY	X	False	1.99	2.97	165.15	False	131	Nam	9	O2	25.730, 27.660, 19.920	23.990, 27.640, 22.330
3	ASP	X	6	THR	X	False	3.15	3.76	120.62	False	72	Nam	42	O2	20.840, 18.680, 23.050	23.210, 21.380, 24.170
3	ASP	X	7	GLY	X	False	2.04	2.93	147.17	False	86	Nam	42	O2	23.650, 18.930, 22.630	23.210, 21.380, 24.170
3	ASP	X	8	THR	X	False	2.04	2.93	147.55	True	31	Nam	106	O2	25.030, 24.010, 22.040	23.040, 24.230, 24.180
3	ASP	X	8	THR	X	False	2.28	3.23	158.10	False	93	Nam	42	O2	25.420, 21.170, 21.830	23.210, 21.380, 24.170
6	THR	X	3	ASP	X	False	3.15	3.76	120.62	True	72	Nam	42	O2	23.210, 21.380, 24.170	20.840, 18.680, 23.050
7	GLY	X	3	ASP	X	False	2.04	2.93	147.17	True	86	Nam	42	O2	23.210, 21.380, 24.170	23.650, 18.930, 22.630
8	THR	X	3	ASP	X	False	2.04	2.93	147.55	False	31	Nam	106	O2	23.040, 24.230, 24.180	25.030, 24.010, 22.040
8	THR	X	3	ASP	X	False	2.28	3.23	158.10	True	93	Nam	42	O2	23.210, 21.380, 24.170	25.420, 21.170, 21.830
10	GLY	X	1	GLY	X	False	1.99	2.97	165.15	True	131	Nam	9	O2	23.990, 27.640, 22.330	25.730, 27.660, 19.920

Hydrophobic Interactions

RESNR	RESTYPE	RESCHAIN	RESNR_LIG	RESTYPE_LIG	RESCHAIN_LIG	DIST	LIGCARBONIDX	PROTCARBONIDX	LIGCOO	PROTCOO
2	TYR	X	4	PRO	X	3.66	46	14	30.910, 16.260, 15.500	32.840, 18.500, 17.650
2	TYR	X	9	TRP	X	4.00	111	18	31.520, 22.680, 20.680	33.560, 20.900, 17.740
2	TYR	X	9	TRP	X	3.27	115	20	34.000, 21.630, 21.070	34.630, 21.840, 17.870

Hydrogen Bonds

RESNR	RESTYPE	RESCHAIN	RESNR_LIG	RESTYPE_LIG	RESCHAIN_LIG	SIDECHAIN	DIST_H-A	DIST_D-A	DON_ANGLE	PROTISDON	DONORIDX	DONORTYPE	ACCEPTORIDX	ACCEPTORTYPE	LIGCOO	PROTCOO
1	GLY	X	10	GLY	X	False	1.97	3.00	174.92	True	1	N3+	138	O.co2	31.090, 17.980, 24.840	33.650, 17.130, 23.530
1	GLY	X	10	GLY	X	False	2.20	2.97	132.62	False	131	Nam	9	O2	31.220, 20.580, 23.080	31.830, 18.060, 21.640
3	ASP	X	6	THR	X	False	3.44	3.78	102.92	False	72	Nam	42	O2	24.480, 17.800, 14.710	27.970, 18.520, 15.980
3	ASP	X	6	THR	X	False	3.16	3.46	100.11	False	78	O3	42	O2	24.900, 18.160, 17.540	27.970, 18.520, 15.980
3	ASP	X	8	THR	X	False	2.04	2.97	156.30	False	93	Nam	42	O2	27.330, 20.930, 17.600	27.970, 18.520, 15.980
3	ASP	X	8	THR	X	False	1.81	2.73	154.22	True	31	Nam	106	O2	29.200, 19.700, 19.410	29.880, 17.400, 18.100
4	PRO	X	7	GLY	X	False	2.82	3.19	102.62	False	86	Nam	56	O2	25.760, 20.110, 15.370	27.140, 18.940, 12.740
4	PRO	X	5	GLU	X	False	2.30	2.90	118.32	True	43	Nam	57	Nam	26.660, 16.750, 13.250	29.130, 17.030, 14.740
5	GLU	X	3	ASP	X	False	2.89	3.38	111.32	True	57	Nam	39	O.co2	25.920, 15.870, 16.430	26.660, 16.750, 13.250
6	THR	X	3	ASP	X	False	1.98	2.96	168.29	True	72	Nam	39	O.co2	25.920, 15.870, 16.430	24.480, 17.800, 14.710
6	THR	X	3	ASP	X	True	1.79	2.74	174.65	True	78	O3	39	O.co2	25.920, 15.870, 16.430	24.900, 18.160, 17.540
7	GLY	X	5	GLU	X	False	3.27	4.07	138.36	True	86	Nam	57	Nam	26.660, 16.750, 13.250	25.760, 20.110, 15.370
8	THR	X	3	ASP	X	False	2.04	2.97	156.30	True	93	Nam	42	O2	27.970, 18.520, 15.980	27.330, 20.930, 17.600
8	THR	X	3	ASP	X	False	1.81	2.73	154.22	False	31	Nam	106	O2	29.880, 17.400, 18.100	29.200, 19.700, 19.410
10	GLY	X	1	GLY	X	False	2.20	2.97	132.62	True	131	Nam	9	O2	31.830, 18.060, 21.640	31.220, 20.580, 23.080
10	GLY	X	1	GLY	X	True	1.97	3.00	174.92	False	1	N3	138	O2	33.650, 17.130, 23.530	31.090, 17.980, 24.840



NTNU

Norwegian University of
Science and Technology