MAX-PLANCK-GESELLSCHAFT

JOHANNES GUTENBERG UNIVERSITÄT MAINZ JG|U

Johannes-Gutenberg Universität Mainz

Master thesis

Interactions and phase behaviour of Ubiquilin-2

Hanne Zillmer
2712948

supervised by
Prof. Dr. Andreas Hildebrandt
Prof. Dr. Kurt Kremer

June 16, 2021

Abstract

Ubiquitin and Ubiquilin-2 are both part of the protein quality control mechanism in the cell. Mutations, especially in the PXX domain of Ubiquilin-2, have been found as one cause of amyotrophical lateral sclerosis (ALS). Experimental results have shown that Ubiquilin-2 undergoes liquid-liquid phase separation and that this process is disrupted by Ubiquitin. This Master thesis investigates the interactions of Ubiquilin-2 and Ubiquitin using molecular dynamics simulations. As all-atom simulations are too computationally expensive to model the time scales necessary to characterize multi-chain interactions of these proteins, a coarse-grained model that represents side chains with a single site and treats solvent interactions implicitly, PLUM, is employed. This work extends the PLUM model to more accurately describe multi-domain proteins with ordered, partially-disordered and disordered regions, using ideas from traditional elastic network models. In particular, additional harmonic interactions are implemented, and the corresponding spring constants are iteratively optimized to reproduce fluctuations of all-atom reference simulations of the relevant folded domains. Thereby, the extended model facilitates the stabilization of Ubiquitin and the UBA domain of Ubiquilin-2 and ensures their stability in the condensed phase, as determined experimentally. As Ubiquilin-2 is partially disordered, this extension is applied locally only to the folded UBA domain. The simultaneous modeling of disordered and partially ordered domains is enabled via the PLUM model. The developed model is then validated by determining the binding affinity using umbrella sampling. The results for the umbrella sampling fall within the order of magnitude of experimental measurements. Preliminary results for simulations of aggregation suggest that the PLUM is limited in its application to larger systems. Thus, it is prohibitively computational expensive to derive full phase diagrams of the interaction of Ubiquilin-2 and Ubiquitin using the current implementation of the PLUM model. Taken together, the results of this Master thesis represent a significant step towards investigations of the LLPS of Ubiquilin-2 using coarse-grained molecular dynamics simulations, and open several directions for future research.

Acknowledgement

I would like to thank Prof. Dr. Kurt Kremer for giving me the possibility to write my Master thesis in the theory group of the MPIP Mainz. I felt well supported throughout my whole time at the institute.

I would like to thank Prof. Dr. Andreas Hildebrandt for his interest in my Master thesis and his willingness to act as my supervisor from the chair of Bionformatics of the University Mainz.

Special thanks to Dr. Joseph Rudzinski from the Biosimulation group of the MPIP Mainz who always had an open door—open Slack channel—for questions and problems. Thank you for helping me during my first steps into a new, exciting field of research, your comprehensive explanations and for creating a constructive, error-friendly atmosphere where I could pose any question—and especially for your patience with my bad internet connection.

Thanks to the whole Biosimulation group who made me feel welcome from the first day in this difficult Corona times!

Contents

Abbreviations **4**

1 Introduction **5**

1.1 Biological background 5

1.2 MD simulations of intrinsically disordered proteins 8

2 Theory and methods **10**

2.1 Basic physical concepts of MD simulations 10

2.2 Coarse-graining and the PLUM model 13

2.3 Calculation of observables from MD simulations 17

2.4 Elastic network models and stabilization of protein structure 19

2.5 Determination of binding affinity using umbrella sampling 20

2.5.1 Estimation of the potential of mean force 21

2.5.2 Facilitated convergence through additional restraints 22

2.6 Implementation details 23

2.6.1 Preparation of structure files 23

2.6.2 All-atom simulations 24

2.7 Coarse-grained simulations 24

2.7.1 Stabilization of protein structure 25

2.7.2 Umbrella sampling 25

2.7.3 Simulation of aggregation 26

3 Results **27**

3.1 Stabilization of protein structure 27

3.1.1 Stabilization of an Ubiquitin monomer 27

3.1.2 Stabilization of the UBA domain of Ubiquilin-2 30

3.2 Umbrella sampling 36

3.3 Simulation of aggregation 42

4 Discussion **43**

4.1 Limitations of the PLUM model that have to be further investigated 43

4.2 Stabilization of protein structure 44

4.3 Umbrella sampling 47

4.4 Intermolecular interactions between multiple chains 50

5 Conclusions and outlook **51**

References **53**

Supplementary information **62**

Declaration of independence **70**

Abbreviations

AA	All-atom
ABF	Adaptive biasing force
ALS	Amyotrophic lateral sclerosis
CG	Coarse-grained
COM	Center of mass
ENM	Elastic network model
IDP	Intrinsically disordered protein
LLPS	Liquid-liquid phase separation
LCST	Lower critical solution temperature
MD	Molecular dynamics
NMR	Nuclear magnetic resonance
PMF	Potential mean force
$\mathbf{R_g}$	Radius of gyration
RMSD	Root-mean-square deviations
RMSF	Root-mean-square fluctuations
SI	Supporting information
mUb	Ubiquitin monomer
UBA	Ubiquitin-associating
UBL	Ubiquitin-like
UBQLN2	Ubiquilin-2
US	Umbrella sampling
UCST	Upper critical solution temperature

1 Introduction

1.1 Biological background

The structure-function paradigm asserts that proteins need a stable three-dimensional structure to be able to carry out a biological function. However, continually growing evidence has shown that there is a large fraction of proteins, known as intrinsically disordered proteins (IDPs), that lack a stable three-dimensional structure but are still involved in biological processes. These proteins are not necessarily fully disordered, but may contain disordered regions along with structured domains [1]. Although IDPs do not have a single, well-defined equilibrium structure, some IDPs obtain a stable structure through binding to other proteins. However, this disorder-to-order transition is not a prerequisite for an IDP to be biologically active. Indeed, IDPs are characterized not by one stable structure but by a spectrum of conformations which range from extended coils to collapsed globules [2]. Furthermore, these disordered regions are connected to different characteristics of the amino acid sequence. In contrast to structured proteins, IDPs tend to have an amino acid sequence with a low complexity — some IDPs just contain a few amino acid types. While IDPs contain lower percentage of hydrophobic amino acids (Ile, Trp, Val), they are rich in polar and charged amino acids (Arg, Lys, Ser, Glu, Pro) [3]. Therefore, it may be possible to predict the disordered propensity of proteins based on different features of their amino acid sequence.

IDPs take part in many different biological processes such as signaling pathways, regulation of transcription and translation and the cell cycle. Due to their high multivalency, IDPs can form many contacts simultaneously. These contacts are important for cellular organization through the formation of membraneless organelles, also called bodies or granules, via liquid-liquid phase separation (LLPS). These phase-separated condensates have liquid-like properties including internal diffusion and inter-condensate fusion upon contact [4]. Through this compartmentalization, certain proteins are concentrated in order to take part in biological functions. Proteins phase separate when macromolecule-macromolecule and water-water interactions are energetically more favorable than macromolecule-water interactions. This depends not only on the concentration of phase-separating proteins, but also on environmental factors including temperature, pH, salt concentration and others [5]. These factors may change in response to heat shock and other stress factors, modulating LLPS.

Since biomolecular condensates formed by LLPS are important for many biological processes, aberrant phase separation may contribute to various complex human diseases including cancer, infectious diseases, and neurodegeneration. Neurodegeneration is thought to be driven mainly by regional aggregation of cytosolic or nuclear proteins [6]. Different studies showed that many of the aggregates in neurodegenerative diseases contain proteins that are part of the protein quality control mechanism [7], [8]. This control mechanism detects misfolded proteins in the endoplasmic reticulum and ensures their degradation in the ubiquitin-proteasome system. Misfolded proteins are marked by Ubiquitin, which is highly conserved throughout different species. Therefore, a lysine residue of the misfolded proteins is bound to the C-terminus of Ubiquitin monomers

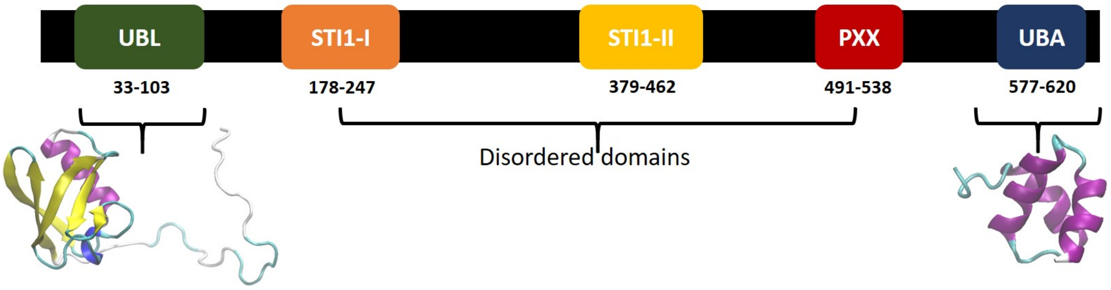

Figure 1: Domain architecture of UBQLN2. Proline-rich domain (PXX) contains most of the ALS-linked mutations. The figure was adapted from [10].

(mUb) or polyubiquitin chains via a reaction chain that includes three different ligases — E1, E2 and E3. The ubiquitinated substrate is then recognized by the proteasome which is responsible for the proteolysis. The guidance to the proteasome is facilitated by so-called shuttle proteins. While the Ubiquitin-like domains (UBL) are recognized by and subsequently bound to subunits of the proteasome, the Ubiquitin-associated (UBA) domain of these proteins binds to the ubiquitin of the target proteins. The binding between the UBA domain and mUb is induced by a hydrophobic patch on mUb. This hydrophobic patch consists of three different residues, namely residue 8 (Leu), residue 44 (Ile) and residue 70 (Val) and is critical for the ubiquitin-proteasome system [9]. The UBA and UBL domains of different shuttling proteins are, like mUb, conserved, and have been well characterized. Ubiquilin-2 (UBQLN2), a member of the protein family of Ubiquitin-like proteins, is one example of a shuttle protein for the Ubiquitin-proteasome system.

There are four Ubiquitin-like proteins in human cells (Ubiquilin 1 - 4). All of them are part of the protein quality control system and play a role in biological processes such as protein degradation, autophagy and stress response. Regarding their domain architecture 1, Ubiquilins contain multiple ST1-like domains that take part in dimerization and lead to homodimers and heterodimers in addition to the UBA and UBL domain. Furthermore, UBQLN2 contains a unique, proline rich PXX domain. While the C-terminal and the N-terminal region of UBQLN2 are predominantly ordered, the central region is predicted to be mainly disordered [10].

Dao et al. showed that UBQLN2 not only co-localizes with stress granules in response to stress factors such as oxidative stress, translational inhibition and osmotic stress, but also undergoes LLPS under physiological conditions [10]. Using different deletion constructs, they were able to identify the STI1-II domain as a crucial domain for LLPS of UBQLN2. According to their results, this domain drives the LLPS through multivalent interactions with the PXX domain and the UBA domain resulting in the oligomerization of UBQLN2 monomers. This oligomerization can be interrupted by Ub. The results of Dao et al. suggest that mUb and polyubiqitin chains bind specifically to the UBA domain. The authors determined a binding affinity of ≈ 30 kJ/mol for the interaction between UBQLN2:UBA and mUb. This result is supported by experimental results of Zhang et al [11]. They characterized the interaction interface of Ub and the UBA domain of Ubiquilin-1, and also showed the high affinity binding between this UBA domain and Ub. Due to their high sequence identity, the experimental findings for the UBA domain of Ubiquilin-1 might be transferable to the UBA domain of UBQLN2 [11]. The binding of Ubiquitin to UBQLN2:UBA appears to disrupt its interaction with the STI1-II domain and, as a result, alters the LLPS of UBQLN2 [10].

Gaining further insight into the LLPS of UBQLN2 and its interactions with Ubiquitin is important because it has been linked to the neurodegenerative disease amyotrophic lateral sclerosis (ALS). ALS is incurable, affects the upper and lower motor neurons and causes among other symptoms an increased muscle tone, muscle fasciculations, and muscle cramps [12]. Most of the mutations relevant for ALS are found in the PXX domain of UQBLN2 (Figure 1). These mutations, first identified in 2011 [8], likely affect the interactions between the molecular components and perturb the protein degradation. As a consequence, the propensity for aggregation might change.

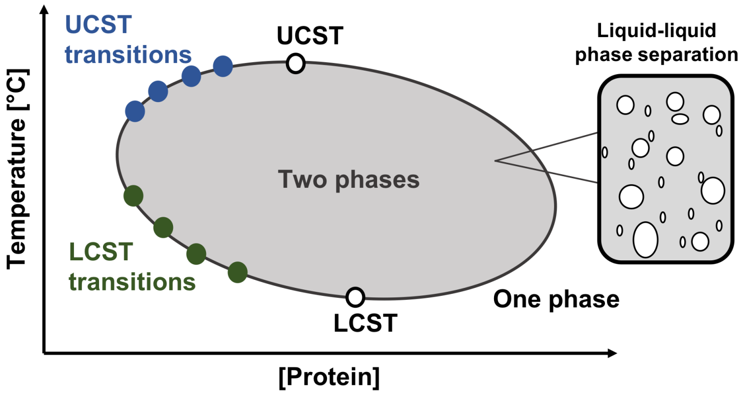

Figure 2: Yang et al. propose a closed loop phase diagram. The figure was adapted from [13].

Yang et al. further analyzed the interactions that drive LLPS of UBQLN2, employing the sticker-spacer perspective [13]. Sticker residues are associative motifs that drive LLPS. They are mostly hydrophobic and polar amino acids, but the chemical basis of sticker interactions varies across phase separating systems. On the other hand, spacer residues connect the sticker motifs along the backbone without significant attractions between them. Yang et al. examined the phase transition of the wild type and the effect of single point mutations on different sticker and spacer residues. Regarding the wildtype, the authors observed two phase transitions as a function of temperature [13]. Performing mutation assays for all 20 proteinogenic amino acids, their results show that hydrophobic mutations lower the temperature threshold in sticker but not in spacer residues. Furthermore, aromatic mutations decrease the phase separation temperature.

Overall, this data suggests a closed loop phase diagram for the phase separation of UBQLN2. This means that there is a lower critical solution temperature (LCST) below which the protein is always in solution and there is an upper critical solution temperature (UCST) above which the protein is always in solution. LCST transitions are driven by weaker hydrogen bonds at higher temperatures and the resulting dehydration of polymer chains [14] while UCST transitions are governed by changes in the strength of molecular interactions. [15]. Depending on the protein concentration, there are other LCST and UCST transitions possible at different temperatures. (Figure 2).

The study of Yang et al. gives some insight into the effect of mutations on LLPS of UQBLN2. However, the mechanisms of ALS progression remain still unknown. Therefore, understanding the molecular details of the relationship between mutations and aberrant phase separation is one route for active therapies. However, comprehending LLPS is not a trivial task since there are many different factors including post-translational modification and protein modifiers that affect this process. Because of the lack of a stable tertiary structure, amino acids are less packed in IDPs. This results in a higher accessibility of amino acids for post-translational modification which is important for the regulation of LLPS [16]. There is a large number of possible post-translational modifications, including phosphorylation and hydroxylation among others. In addition to post-translational modification and environmental factors, LLPS can be also influenced by interactions with other biomolecules such as Ubiquitin.

The different modifications result in altered chemical properties of the amino acids, such as charge state or excluded volume, and either enhanced or weakened interactions between the molecules [17]. Understanding different interactions during LLPS is crucial in order to learn more about pathological protein aggregates in neurodegenerative diseases, such as ALS. Charge-charge and π-interactions as well as hydrophobic contacts and hydrogen bonds contribute to the emergent properties, e.g. aggregation, of IDPs. However, it is difficult to determine the relative importance of distinct interactions to LLPS because the formed droplets are highly dynamic and, like UBQLN2, many components are at least partially disordered [16]. Hence, it is hard to to obtain structural properties of this IDP experimentally and understand the forces that drive its LLPS using solely experimental methods [18].

1.2 MD simulations of intrinsically disordered proteins

Experiments provide limited insights into LLPS of macromolecules due to the spatiotemporal resolution of individual techniques. This often makes it difficult to simultaneously probe the relevant length and time scales while capturing microscopic details. Although some ensemble techniques, such as NMR, have a finer spatiotemporal resolution, they still lose details about individual molecules which complicates the investigation of IDPs with broad conformational ensembles [19]. Therefore, it is helpful to complement experimental measurements with computational methods, e.g., molecular dynamics (MD) simulations.

MD is a simulation method that is used to analyze the behavior of molecular systems. By numerically solving Newton's equations of motion in a step-wise manner, trajectories are generated that display the time evolution of the system. Using these trajectories, different observables can be calculated that characterize the molecules, e.g., their compactness and flexibility among others. In addition to analyzing properties of single molecules, it is also possible to investigate the interactions of several molecules, e.g., to calculate their contact probability or binding affinity. By providing microscopic details, this type of study can lead to insights about the relevant driving forces of LLPS. Additionally, results of MD simulations can be used to design and specify further experiments to reinforce the simulation data.

Depending on the resolution of the studied system, MD simulations are categorized into three classes: *ab-initio* (quantum), all-atom (AA) and coarse-grained (CG) MD simulations [20]. While quantum MD simulations model the electrons of the system explicitly, the latter two employ so-called force fields to model the forces acting on the particles in the system. As the name already implies, AA MD simulations include all particles of a system without the electrons, which results in a reduction of the computational resources needed. They are a powerful tool for investigating specific chemical driving forces of biomolecular processes. Although some studies have investigated the correlation between single-chain properties and phase behavior using AA simulations, these models are these models are prohibitively expensive for the large scales necessary to investigate LLPS for more than one chain. [21].

Therefore, CG approaches that represent groups of atoms with a single CG bead are needed. CG models decrease the number of degrees of freedom and the required computational resources. Due to this reduction in degrees of freedom, there are less force calculations in each MD step and the free energy landscape is smoothed. Hence, the motion through phase space is accelerated and conformational changes occur faster. This possibility of simulating larger systems comes with a loss of accuracy. However, CG models focus on essential interactions and driving forces while getting rid of less important details of the examined system.

Regarding MD simulations of IDPs, they are not trivial to conduct for AA as well as CG models and there are different approaches. While AA simulations are often used to model IDP structures and their binding, CG simulations are typically applied to model assemblies of IDPs. One of the simplest CG model for IDPs are Gō-models, which represent each amino acid with one bead. Gō-models are only relevant for coupled folding and binding processes. The energy functions of these models are constructed based on contacts found in experimentally determined structures such that the native structure is the global minimum of the model. These models are only applicable to systems where the structure of the proteins in the bound complex is known. Further developed Gō-models—hybrid Gō-models—integrate amino-acid specific interactions and non-native interactions. Therefore, these hybrid Gō-models allow for the investigation of the impact of specific interactions to the folding process [22]. Other models have been developed to investigate the specific binding of two IDPs whose functions do not require folding. Theses models need to reproduce the conformational heterogeneity of proteins without a stable three-dimensional structure within fuzzy IDP complexes [19]. One example is a CG model which has a simple contact potential to represent cation-π interactions. Via these interactions aromatic amino acids, like phenylalanine, can bind to charged amino acids, such as lysine. They have been proposed as being important for binding partner recognition [23].

Other approaches allow a more general view at IDP interactions without requiring any specific binding partners and also allow potentially the simulation of many molecules. The formation of granules has been primarily investigated using CG models that do not have a native-structure bias by employing simple physics-based models [16]. In these models the solvent is modeled implicitly. For example, one study investigated the temperature-dependent LLPS of various proteins [18] and examined sequence determinants for LLPS based on a model sequence with various charge distributions. Furthermore, there have been recent efforts to improve these models through a more accurate reproduction of experimental observables, e.g. radius of gyration [24]. However, some interactions, such as cation-π interactions, which are regarded as important for LLPS, are not included explicitly in these models. In addition, most of the studies focus on completely disordered proteins [24].

In this Master thesis, the interactions between mUb and UBQLN2 and their relevance for LLPS is investigated using CG MD simulations. In order to represent the disordered and partially ordered domains of UBQLN2, the PLUM model, parameterized to balance α-helix and β-sheet propensity, was employed. While the PLUM model can stabilize some tertiary structures, e.g., three-helix bundles, it does not perfectly model mUb, which has a more complex structure. Therefore, the PLUM model was extended with additional, harmonic interactions that facilitated the simulation of mUb and ensured a stable structure of UBQLN2:UBA upon interaction with other proteins. This extension was based on the idea of an elastic network model. The harmonic interactions were established based on a C_α contact map and their strength was iteratively optimized in a set of simulations. To investigate the interactions of mUb and UBQLN2 in aggregation simulations, it was necessary to guarantee that the two proteins interact reasonably in the PLUM-ELM model. This was verified using umbrella sampling. Aggregation simulations were then performed to investigate the interactions between UBQLN2 and mUB and preliminary analyzed using different observables. The Master thesis is organized as follows. First, an introduction to the basic physical concepts of MD simulations, the PLUM model and the idea behind elastic network models and umbrella sampling is given. This section is concluded with the relevant implementation details. Next, the results of the different MD simulations are summarized. These results are then discussed and set in context to other studies before providing the conclusions from the study and an outlook for future work.

2 Theory and methods

2.1 Basic physical concepts of MD simulations

MD simulations are computational calculations based on physical concepts. Because *ab initio* modeling through quantum mechanics would require too many computational resources for more complex systems, such as proteins, classical MD simulations are needed. During these simulations, particles are evolved by using Newton's equations of motion. These movements and interactions can be described by the forces applied to them and the derived energies. Overall, classical MD simulations result in trajectories that provide information about the most favorable energetic states and pathways to them. A MD cycle typically includes the following steps:

1. Coordinates and velocities are initialized.

2. Forces are calculated based on the coordinates and the provided interactions.
3. Coordinates and velocities are updated based on the forces.
4. Observables, such as positions, energies, temperature, are computed and written as output at predefined time steps.
5. Is the termination condition fulfilled? If not, the simulation is continued with step 2.

First an input configuration needs to be generated including all atoms that should be simulated. This file can be obtained using experimental results, like structures from NMR experiments, or the initial coordinates can be created manually or based on the sequence with the help of prediction tools, e.g. the I-TASSER server [25]. In order to precisely describe all forces acting within and between molecules in the system, the Schrödinger equation has to be solved which is almost impossible for system sizes exceeding a few atoms. Thus, empirical force fields are employed to approximate these forces and energies and model the interactions in molecules . These modeled interactions include bonded and non-bonded interactions. An example of a force field, the PLUM model, and its implementation is given in section 2.2.

The bonded interactions typically consist of bonds, angles and dihedral interactions. Chemical bonds form a covalent link between two atoms that vibrate around a equilibrium distance. The simplest way to model these bonds is with a harmonic potential which depends on the distance between the bonded particles. Three bonded atoms form an angle that, like a covalent bond, pulsates around an equilibrium position. These angles are also mostly modeled using harmonic potentials. Finally, dihedral angle interactions control the relative orientation of the planes defined by adjacent bond angles. This dihedral term is generally modeled using some form of a cosine series [20].

The non-bonded interactions consist of electrostatic and van der Waals interactions and act between atoms that are not linked by a covalent bond. Electrostatics describe the interaction between atoms due to their partial charges. These charges interact through the Coulomb potential and their interactions are considered long range because they decay slowly with distance. In contrast, the van der Waals interactions are short ranged. They include an attractive and a repulsive part. If the distance between atoms is short, the van der Waals interactions will be strongly repulsive due to the Pauli principle. On the other hand, the electron clouds of two atoms will polarize each other when they are close. This results in an induced-dipole and the attractive part of the van der Waals interactions. These interactions are often modeled using the Lennard-Jones potential which is commonly expressed by:

$$V(r) = 4\epsilon[\frac{\sigma}{r_{ij}}^{12} - \frac{\sigma}{r_{ij}}^{6}], \tag{1}$$

where ϵ is the depth of the energy minimum, r_{ij} is the distance between two particles i and j and σ is the distance between atom i and j at which the attractive and the repulsive contributions of the Lennard-Jones potential balance out.

The non-bonded and bonded interactions add up to the total energy of the system. As all of the above described interactions are energies, the corresponding force on each atom is obtained by taking the negative gradient of the energy. The calculated forces are then used in the next step to update the coordinates and velocities of the atoms. For this update, the classical equations of motion are solved in a step-by-step manner. Newton's second law implies that the acceleration of an object is proportional to the force acting on it:

$$\vec{F}(\vec{r},t) = m\vec{a}(t) = m\frac{d}{dt}\vec{v}(t) = m\frac{d^2}{dt^2}\vec{r}(t) \tag{2}$$

Since it is a daunting task to solve this second-order differential equation analytically, integrators are employed that solve Newton's equations of motion numerically, e.g. the leap-frog algorithm [26]:

$$\vec{v}(t+\frac{\Delta t}{2}) = \vec{v}(t-\frac{\Delta t}{2}) + \vec{a}(t)\Delta t \tag{3}$$

$$\vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t+\frac{\Delta t}{2})\Delta t, \tag{4}$$

where $\vec{a}$ is the acceleration, $\vec{v}$ is the velocity and $\vec{r}$ is the position at a certain time point t. It is called the leap-frog algorithm because the position and the velocities are updated at shifted time points which causes them to "leapfrog" over each other. Another integrator is the leap-frog stochastic dynamics integrator which is based on Langevin dynamics. Considering a number of solute particles and a much greater number of solvent particles, it would take a lot of computational resources to model the motion of both. As the time scale for solvent motion is much faster than for dispersion particles and one is not interested in the motion of the solvent itself, the impact of the solvent on the dispersion particles can be treated collectively [27]. Therefore, a friction constant and a noise term are added to Newton's equations of motion in Langevin dynamics. The noise term represents a stochastic process and aims to resemble the effects on a system that are caused by high velocity collisions with the solvent. Hence, it partially incorporates solvent effects into systems that are modeled in vacuum.

Depending on the properties of interest, there are different statistical ensembles that can be sampled from. An ensemble is defined as a set of microstates that are compatible with a given macrostate. The microcanonical ensemble (NVE) describes an isolated system in which the number of particles (N), the volume (V) and the energy (E) do not change over time and no energy is exchanged with the environment. In a canonical ensemble (NVT) the number of particles (N), the volume (V) and the temperature (T) are kept constant. In contrast to the microcanonical ensemble, the energy will change over time because the system can exchange energy with a heat bath. Another important ensemble is the isobaric-isothermal ensemble (NPT). This ensemble implies a constant number of particles (N), pressure (P) and temperature (T). All simulations in this Master thesis were conducted either in the NVT or NPT ensemble.

In order to sample from the desired ensemble, thermostats and barostats are needed to maintain the correct average temperature or pressure of the system, respectively. In MD simulations the temperature is computed based on the kinetic energy of the system.

Hence, it is possible to modify a system's temperature by velocity rescaling. Since simple velocity rescaling leads to instabilities, more sophisticated algorithms were developed to control the temperature of a system. For example, the above mentioned stochastic dynamics integrator can be used as thermostat. Another algorithm is the Berendsen thermostat [28] which rescales the velocity as followed:

$$\vec{v}_i(t) \rightarrow \vec{v}_i^{\prime}(t) = \sqrt{1 + \gamma(\frac{T_f}{T(t)} - 1)}\vec{v}_i(t), \tag{5}$$

where $\vec{v}_i$ and $\vec{v}_i^{\prime}$ correspond to the velocity and the rescaled velocity, T_f is the target temperature and the time constant γ controls the decay time. The Berendsen thermostat is able to relax systems to a desired temperature efficiently. On the other hand, it is not correctly sampling the NVT ensemble because it prevents the kinetic energy from fluctuating. As the Berendsen thermostat approximates the canonical ensemble, it is often used for equilibration of larger systems. After that, the thermostat is switched to a more accurate one.

There is also the Berendsen barostat [28] which works in a similar way as the thermostat. Like the temperature of a system, its pressure can be also derived from the kinetic energy. To keep the average pressure constant, barostats rescale the distance between particles instead of the velocity. The Berendsen barostat is often used for equilibration for the same reasons as the the Berendsen thermostat. To ensure correct sampling after that, other barostats are used, such as the Parrinello-Rahman barostat [29]. This barostat not only allows changes to the volume of the simulation box but also to the box shape.

Before starting the aforementioned MD cycle, the initial structure needs to be locally energy-minimized. If there was no energy-minimization and the starting configuration was energetically unfavorable, e.g. due to steric clashes, the system would have a high potential energy. This energy would then be mostly translated into kinetic energy. The resulting high velocities might destabilize the simulation or even lead to a crash. One approach to energy-minimize systems is the steepest descent algorithm. This algorithm minimizes the energy of a structure by calculating the gradient of the energy in each step. Stepping in the opposite direction of this gradient, is a step in the direction of the locally, steepest descent. This is repeated until the algorithm converges to a local minimum.

The energy-minimization is followed by an equilibration phase which ensures to relax the system to the right temperate and pressure. At the end of the equilibration phase the mean of potential energy of the system should not change significantly. After that, the production run is started to sample the properties of interest.

2.2 Coarse-graining and the PLUM model

Simulating biological systems with more than just a few amino acids requires a lot of computing power. In addition, the relevant time scales of biological processes differ from ps to s or even longer. This makes AA MD simulations often impracticable for processes on longer time scales, such as LLPS. Hence, methods, such as coarse-graining, are needed to reduce the computational resources and enable simulations at time scales that are inaccessible for AA models.

CG approaches reduce the resolution by representing groups of atoms with a single bead. There are different levels of coarse-graining ranging from grouping together the atoms of single amino acids to combining several amino acids to one bead. This simplification aims to focus on the essential interactions and driving forces while getting rid of less important details. In order to ensure that a CG model is physically realistic and useful, the removed degrees of freedom must be chosen carefully [30]:

1. They must not be indispensable for the investigated process.
2. Removing these degrees of freedom should result in a sufficient reduction of computational resources to compensate the loss in accuracy.
3. The interactions that are important for the removed degrees of freedom should mostly not be coupled with the remaining degrees of freedom.
4. Removing these degrees of freedom, should enable a simple and efficient representation of the remaining degrees of freedom.

The reduction in degrees of freedom leads to less force calculations in each MD step. Furthermore, the free energy landscape is typically smoothed which accelerates the motion through phase space. Thus, conformational changes occur faster and the time measured for conformational changes in the CG model does not correspond to the time that it would take the real protein to undergo the same structural changes.

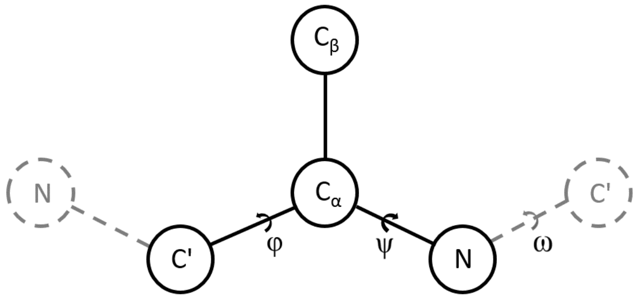

Figure 3: Mapping scheme of the PLUM model. Figure was adapted from [31]

There are different ways for developing a CG model. Two common approaches are top-down and bottom-up coarse-graining. Top-down models rely on experimental data while bottom-up models use an underlying model with higher resolution of the same system. One commonly used CG model is the MARTINI model which employs a four-to-one mapping [32]. This means that approximately four heavy atoms are mapped to one bead. The MARTINI model is a top-down model, parameterized using experimental partitioning coefficients. In the MARTINI force field water molecules are CG as well. Other force fields do not model the solvent explicitly, but account for its effects on the studied system implicitly in the implemented interactions.

The CG model used in this Master thesis, the PLUM model, is hybrid. This means that the some of the interactions are parameterized bottom-up, while others are derived from experiments. However, the interaction forms are all physics-based. In this implicit solvent CG approach, the backbone is represented with near-atomistic resolution [31]. Each amino acid is modeled by 3 (glycine) to 4 (all other amino acids) beads in the PLUM model. There is one bead for the amide group (N), one for the central carbon (C_α), one for the carbonyl (C′) group and one (for non-glycine residues) for the side chain (C_β). The side chain bead is responsible for the amino acid specificity and is placed at the location of the first carbon of the side chain. It is directly connected to the backbone (Figure 3). Since secondary structures, such as α-helices and β-sheets, are stabilized by backbone-backbone hydrogen bonds between the amide and the carboxyl group, the high resolution of the backbone allows the PLUM model to form well-defined secondary structures. All amino acids, except glycine, are given the same van der Waals radii for the C_β bead. Because glycine residues do not have a C_β bead in the PLUM model, they are highly flexible and occupy different regions in the Ramachandran plot [33]. Using the Ramachandran plot, the backbone dihedral angles ϕ and ψ are plotted against each other to visualize their energetically favorable positions. In addition, identical van der Waals radii might lead to problems if dense packing of secondary structures is important, e.g., in globular proteins [31]. Regarding the force field parameterization, the bond and angle potentials are chosen to be harmonic:

$$V_{bond}(r) = \frac{1}{2}k_{bond}(r - r_0)^2 \tag{6a}$$

$$V_{angle}(\theta) = \frac{1}{2}k_{angle}(\theta - \theta_0)^2, \tag{6b}$$

where k_{bond} is the spring constants for the bond and r and r_0 are the current distance and the equilibrium distance for the bonded beads, respectively. The variable k_{angle} refers to the spring constant of the angle and θ and θ_0 are the current angle and the equilibrium angle, respectively.

The structural flexibility of the backbone beads enters through the dihedrals—ϕ ($C'NC_\alpha C'$), ψ ($NC_\alpha C'N$) and ω ($C_\alpha C'NC_\alpha$). While ϕ and ψ are very flexible and are important for local structure elements, ω is located at the peptide bond. The rotation around this bond is resembled by a symmetric potential with two minima, *cis* and *trans* conformation. As *cis* conformations are sterically unfavored for all amino acids except prolines, the authors employed a potential with one minimum around the *trans* confirmation for ω:

$$V_{dih}(\omega) = k_n[1 - cos(n\omega - \omega_{n,0})], \tag{7}$$

with coefficient k_n and phase $\omega_{n,0}$. A potential with two minima is employed for a peptide bond right before a proline residue.

Regarding ϕ and ψ, they added a potential that takes into consideration the interacting dipoles of carbonyl and amide groups in a peptide bond:

$$V_{dip}(\Phi, \Psi) = k_{dip}[(1 - cos(\Phi)) + (1 - cos(\Psi))], \tag{8}$$

where k_{dip} is the force constant for this potential which accounts for a non-bonded dipole interaction.

The remaining non-bonded interactions in the PLUM model include backbone and side chain interactions and implicitly modeled hydrogen bonds. To model the steric interactions of the backbone, Bereau et al. used the purely repulsive Weeks-Chandler-Anderson potential which represents a local excluded volume:

$$V_{bb}(r) = \begin{cases} 4\epsilon_{bb}[(\frac{\sigma_{ij}}{r})^{12} - (\frac{\sigma_{ij}}{r})^{6} + \frac{1}{4}] & , r \leq r_c \\ 0 & , r > r_c, \end{cases} \tag{9}$$

where $r_c = 2^{1/6}\sigma_{ij}$ and σ_{ij} is the arithmetic mean between the two bead sizes. The same ϵ_{bb} is used for backbone-backbone and backbone-side chain interactions. The excluded volume is only calculated between beads that are further than two bonds apart.

The interactions between C_β beads are modeled based on the interaction strength for every pair of amino acids which were determined by Miyazawa and Jernigan via analysis of the PDB [34]. These are rescaled to values between 0, for the most hydrophilic residue, and 1, for the most hydrophobic residue and multiplied by the free parameter ϵ_{hp}. For the overall side chain interaction, Bereau et al. used a Lennard-Jones potential for the attractive part which was combined with a purely repulsive Weeks-Chandler-Andersen potential at smaller distances.

$$V_{hp} = \begin{cases} 4\epsilon_{hp}[(\frac{\sigma_{C_\beta}}{r})^{12} - (\frac{\sigma_{C_\beta}}{r})^{6}] + \epsilon_{hp} - \epsilon'_{ij} & , r \leq r_c, \\ 4\epsilon_{hp}\epsilon'_{ij}(\frac{\sigma_{C_\beta}}{r})^{12} - (\frac{\sigma_{C_\beta}}{r})^{6}] & , r_c \leq r \leq r_{hp,cut}, \\ 0 & , r > r_{hp,cut} \end{cases} \tag{10}$$

In order to implicitly model hydrogen bonds, the authors decided to use a 12-10 Lennard-Jones potential in combination with an angular term:

$$V_{hb}(r, \Theta_N, \Theta_C) = \epsilon_{hb}[5(\frac{\sigma_{hb}}{r}^{12}) - 6(\frac{\sigma_{hb}}{r})^{10}] \times \begin{cases} cos^2(\Theta_N)cos^2(\Theta_C) & , |\Theta_N|, |\Theta_C| < 90^\circ, \\ 0 & , \text{otherwise,} \end{cases} \tag{11}$$

where r is the distance between the N and C′ beads, σ_{hb} is the bead radius and the angles Θ_N and Θ_C are formed by the atoms HNC′ and NC′O, respectively. According to the authors, the implemented hydrogen bonds were only sufficient to stabilize β-sheets in combination with the aforementioned dipole interaction [31].

The implemented interactions aim to reproduce the balance of α/β structural propensity without being biased towards a native structure. This allows simulation of proteins with a changing structure or no dominant structure at all, e.g. IDPs. So far, the model has been shown to stabilize β-sheet structures [31], [35] and to fold several helical peptides [31], [36], [37]. Testing the transferability of the PLUM model to IDPs, Rutter et al. found that it over-stabilizes α-helices for a short disordered peptide [38]. In order to represent destabilization of helices in IDPs, Rutter et al. reduced the strength of the backbone-backbone hydrogen bond interaction strength parameter ϵ_{hb} to 94.5% of its original value [38]. Bereau et al performed a manual parametrization of the free energetic parameters (ϵ_{bb} , ϵ_{hp}) and the bead diameters (σ_{C_α} , σ_{C_β} , $\sigma_{C'_C}$, σ_{C_N}) while keeping the cutoff values for the different potentials and the equilibrium distance of a hydrogen bond fixed, in an attempt to reproduce the probability distribution functions of dihedral angles and large-scale properties, such as folding [31].

The PLUM model has its natural units [31]. Since coarse-graining a molecule speeds up its motion through phase space, the time that the CG molecule needs to undergo conformational changes is not equal to the time that the molecule would need to undergo these changes in reality. $\epsilon = 1$ Å is the length. The thermal energy at room temperature is given by $\epsilon = k_B T_r = 1.38 * 10^{-23}$ J K^{-1}300 K ≈ 0.6 kcal $*$ mol^{-1} and $M \approx 4.6 * 10^{-26}$ kg is the mass. $\tau = L\sqrt{\frac{M}{\epsilon}} \sim 0.1$ ps is the natural time unit of the PLUM model. The temperature is reported in Kelvin in this Master thesis. However, a temperature of 300 K corresponds to T* = 1 in PLUM. In order to match the real time and τ, an associated speedup factor needs to be determined. This speedup factor has not been determined for the PLUM model yet. Therefore, the natural units of the PLUM model will be used in this Master thesis.

2.3 Calculation of observables from MD simulations

MD simulations produce high-dimensional trajectories. To understand the conformations sampled by the system, a range of low-order, usually 1-dimensional, observables are employed. In this Master thesis the root-mean-square deviation (RMSD), radius of gyration (R$_\text{g}$), root-mean-square fluctuations (RMSF), the helical fraction and the β-sheet fraction were used as observables for each single chain simulation to measure convergence and to compare to other simulations and experimental results. It is important to mention that all of these observables have their advantages and disadvantages. Hence, regarding one of these observables alone results in an incomplete picture. All observables have to be taken into consideration in combination instead.

The RMSD is a measurement of the similarity between two structures. More specifically, for a configuration of the system at time t, the RMSD is determined by calculating the mean squared distance between the Cartesian coordinates at time t of atom i and the Cartesian coordinates of the same atom in a reference structure:

$$\rho^{\text{RMSD}}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i(t) - r_i^{\text{ref}})^2}, \tag{12}$$

where N is the number of atoms, t is the current time point and $r_i(t)$ and r_i^{ref} are the coordinates at time t and reference coordinates of atom i, respectively. Before calculating the RMSD, the current structure is superimposed to the reference structure by least-square fitting. An experimental structure is often used as reference for proteins.

The compactness of a protein can be measured through the R$_\text{g}$. It can indicate structural changes and is defined as the root-mean-square deviation from the center of mass (COM):

$$R_g(t) = \sqrt{\frac{1}{M}\sum_{i=1}^{N} m_i(r_i - R_{COM})^2}, \tag{13}$$

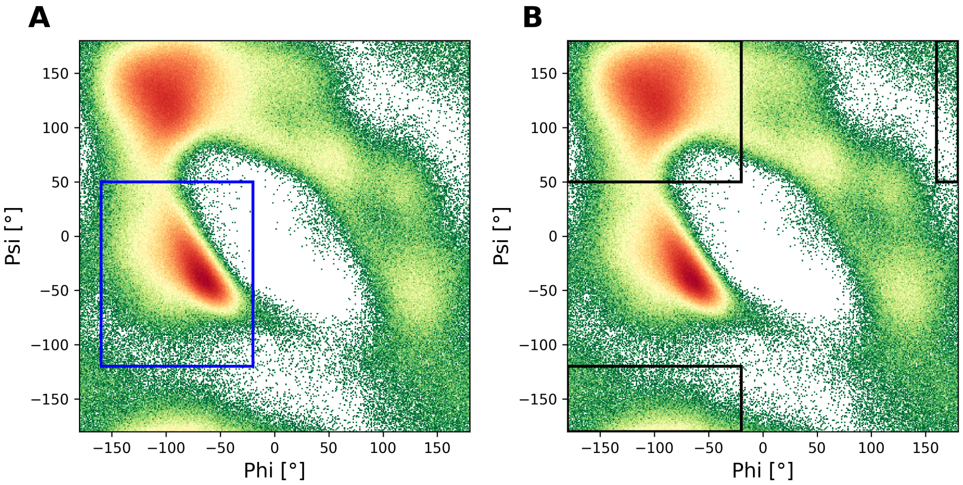

Figure 4: Ramachandran plot of a simulation of mUb with the PLUM-ELM model. (**A**) The blue box indicates the helical region. (**B**) The black boxes indicate the β-sheet regions

where M is the total mass, N is the number of particles, m_i is the mass of particle i, r_i are the Cartesian coordinates of particle i and $R = \frac{1}{M} \sum_{i=1}^{N} m_i \, r_i$ is the COM of the protein.

The RMSF measures the flexibility of each residue within a protein. For proteins it is common to use C_α atoms as a probe of the residue fluctuations. The RMSF is determined by the square root of the variance around the average position:

$$\rho_i^{RMSF}(t) = \sqrt{\langle (r_i(t) - \langle r_i \rangle)^2 \rangle}, \tag{14}$$

where $\langle ... \rangle$ indicates the average and r_i and $\langle r_i \rangle^2$ are the Cartesian coordinates at time point t and the average Cartesian coordinates of atom i, respectively. It can be compared to the experimentally determined Debye-Waller factor, also called B factor.

To examine the secondary structure of the proteins during the simulations, the helical and β-sheet fractions were calculated. In α-helices and β-sheets the dihedral angles of the backbone atoms (ϕ, ψ) fall within specific regions of the Ramachandran plot (Fig. 4), since the backbone flexibility largely depends on these two observables. Hence, it is possible to assign a secondary structure element to each residue based on the values of these dihedral angles. In this Master thesis the helical region is defined as $\phi \in [-160, -20]$ and $\psi \in [-120, 50]$ (Blue box in fig. 4A) [39]. For each residue in each time frame it was determined whether a residue is helical — within the helical (h) region — or non-helical (n) — outside the helical region. If (ϕ, ψ) of one residue fall within the specified region, this does not necessarily imply that this residue is part of a α-helix. It is the correlated positioning of multiple, consecutive dihedrals that indicates a secondary structure element. A helical segment is a segment where at least three consecutive residues lie within the helical region in a time frame, e.g. **...nhhhn...** is the shortest possible helical segment. The helical fraction of a given residue is then given by the fraction of time that this residue spends within a helical segment.

The β-sheet fraction was calculated in the same way as the helical fraction. There are three different regions in a Ramachandran plot which are occupied by β-sheets. They are shown in figure 4B (Black boxes): 1. $\phi_1 \in [-180, -20]$, $\psi_1 \in [50, 180]$, 2. $\phi_2 \in [-180, -20]$, $\psi_2 \in [-180, -120]$ and 3. $\phi_3 \in [160, 180]$, $\psi_3 \in [50, 180]$ [40]. For each residue in each time frame it was checked whether the dihedral angles are found in a β-sheet region (b) or non-beta (n) — outside the β-sheet regions. A β-sheet segment is a segment where at least five consecutive residues lie within the β-sheet regions in a time frame, e.g. **...nbbbbbn...** is the shortest possible β-sheet segment. The β-sheet fraction of a given residue is then given by the fraction of time that this residue spends within a β-sheet segment.

2.4 Elastic network models and stabilization of protein structure

To ensure the stabilization of tertiary structure in a CG model, some studies have used elastic network models (ENM). ENMs are simple CG models in which all atom pairs within a cutoff distance in the native structure are linked with a harmonic spring of rest length equal to the atoms' distance [41]. It has been demonstrated that this model can reproduce the frequency spectrum of a folded protein. The additional springs of ENMs can be also used on top of another, more complex CG model to stabilize a given reference structure. This approach does not consider the fact that the atoms might be positioned within the cut-off radius but are the still not able to interact with particular atom. Consider the case that atoms 2 and 3 are within the cutoff radius of atom 1 (Figure 5). A more sophisticated approach to determine the contact list that factors in these shadow contacts has been previously developed [42]. It excludes bonds that would establish unphysical contacts.

However, solely avoiding these contacts does still not take into consideration the different fluctuations between the pairs of atoms and the need for a unique spring constant for each pair, which is typically chosen uniformly for all springs. To solve this problem, Globisch et al. established an iterative scheme which updates each spring constant separately in a series of simulations [43]. Starting from an initial simulation with uniform spring constants, the spring constants are updated after each simulation according to:

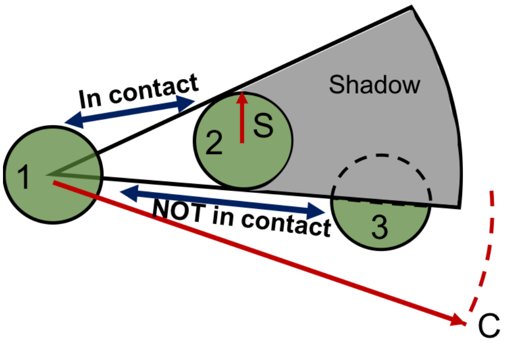

Figure 5: Scheme of how contacts are determined using the shadow contact map tool.The figure was adapted from [42]

$$K_{ij,n+1} = K_{ij,n} - \alpha \frac{k_B T}{R_C^4} [\sigma^2_{d,AA}(i,j) - \sigma^2_{d,CG}(i,j)], \quad (15)$$

where $K_{ij,n}$ and $K_{ij,n+1}$ are the current and the new spring constant, respectively, between particles i and j, α is the scaling factor, R_C is the cut-off distance, and $\sigma^2_{d,AA}(i,j)$ and $\sigma^2_{d,CG}(i,j)$ are the variances in distances of particles i and j in AA and CG simulation, respectively. The scheme is said to be converged, if the difference in variances of AA and CG simulation is sufficiently small [43]. To emphasize in which simulation additional springs were used, the resulting combination of the PLUM model and the additional spring network will be subsequently referred to as PLUM-ELM.

2.5 Determination of binding affinity using umbrella sampling

In addition to determining the properties of single chains, it is also possible to characterize the conformational behavior of multiple molecules using MD simulations. The overall interaction strength between molecules is determined through the binding affinity. Deriving the binding affinity of two proteins from unbiased MD simulations requires the occurrence of multiple binding and unbinding events. While the association rate of protein complexes is mostly limited by diffusion and is on the order of the magnitude of 10 M $*$ s^{-1} [44], [45], the dissociation rate is mainly governed by the strength of short-range non-covalent interactions, such as hydrogen bonds, van der Waals forces, hydrophobic interactions, and ionic bonds [46]. This leads to a wide range of dissociation constants for different protein complexes which makes it difficult to ensure several binding and unbinding events within the time scales accessible to MD simulations.

The binding affinity is a function of the free energy difference between the bound and unbound states. This free energy difference is linked to the equilibrium constant of the reversible binding process $L + P \rightleftharpoons LP$ which is defined as:

$$K_{eq} = \frac{[L] + [P]}{[LP]}, \tag{16}$$

where [...] is the concentration of each species. For an easy nomenclature one protein is named P while the other is referred to as L for ligand. The smaller the dissociation constant, the higher the free energy difference between the bound and the unbound state and the more tight the binding between two proteins or a protein and a ligand.

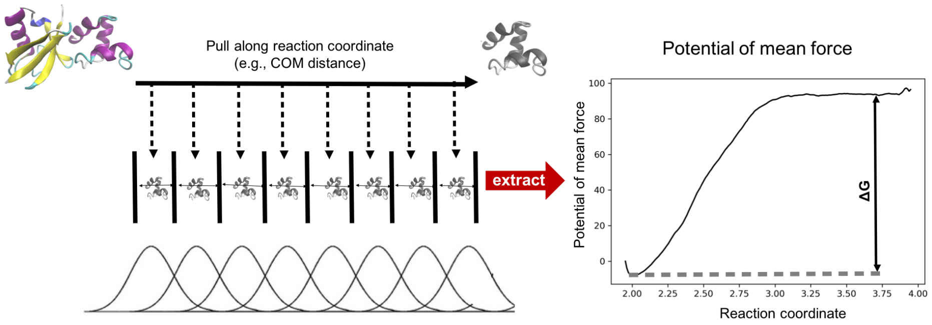

Figure 6: Scheme that describes the process of umbrella sampling.

To enable the crossing of free energy barriers, enhanced sampling techniques are needed, e.g. umbrella sampling (US) [47]. US aims to enhance sampling of energetically unfavorable configurations by adding a biasing potential $w(\xi)$ along a one- or more dimensional reaction coordinate ξ, e.g. the COM distance. Different choices are possible for the ξ. However, the bound and unbound states must be clearly distinguishable along the chosen ξ. $w(\xi)$ could be, for example, a harmonic function of the form:

$$w_i(\xi) = \frac{1}{2}k(\xi - \xi_i)^2 \tag{17}$$

In order to calculate the binding affinity, the unbiased potential of mean force (PMF)—the free energy surface along ξ — needs to be extracted. As $w(\xi)$ ensures that only a defined region along ξ is sampled, a single simulation allows an estimate for a small part of the desired PMF. Hence, several, successive simulations have to be run that sample different regions of ξ. The input files for these simulations are often created via a pulling simulation. During the pulling simulation, the ligand is pulled out of the binding site along ξ. Using the pulling trajectory, structure files are saved in the desired regions of ξ. The US is then conducted with these input structures and a biased histogram is produced for each umbrella window. The unbiased probability distribution of each simulation, and subsequently an estimate of the unbiased PMF, is then obtained using a reweighting procedure. A scheme of how an US works is illustrated in fig. 6.

2.5.1 Estimation of the potential of mean force

One method to unbias US simulations is the weighted histogram analysis method (WHAM) [48]. Using WHAM, the unbiased probability distribution $P(\xi)$ is calculated as followed:

$$P(\xi) = \frac{\sum_{i=1}^{N_w} g_i^{-1} h_i(\xi)}{\sum_{(j=1)}^{N_w} n_j g_j^{-1} \mathrm{e}^{-\beta(w_j(\xi) - f_j)}} \tag{18a}$$

$$\mathrm{e}^{-\beta f_j} = \int d\xi \mathrm{e}^{-\beta w_j(\xi)} P(\xi), \tag{18b}$$

where $\beta = 1/k_B T$ is the inverse temperature, $h_i(\xi)$ is an umbrella histogram representing the biased probability distribution $P_i^b(\xi)$, n_j is the number of data points in histogram h_j and f_i is the total free energy including the biasing potential. The statistical inefficiency g_i is given by $g_i = 1 + 2\tau_i$ with τ_i being the autocorrelation time. It only cancels out of the equation if all umbrella windows have the same autocorrelation time. Otherwise, windows with longer τ will be assigned with lower weights [49]. As there are two unknown parts in equations 18a and 18b, $P(\xi)$ and f_i, they have to have to be solved self-consistently starting from a initial guess for the free energy constants f_j. The PMF along ξ can then be calculated using $P(\xi)$:

$$\mathcal{W}(\xi) = \frac{1}{-\beta} \ln(P(\xi)/P(\xi_0)), \tag{19}$$

where ξ_0 is arbitrary reference point at which the PMF, $\mathcal{W}(\xi_0)$, is set to 0. The WHAM equations do not allow direct inference of an error estimates, but different bootstrapping methods can be used to compute them [50].

Another example of an estimator for calculating free energy differences between multiple states is the multistate Bennett acceptance ratio (MBAR) [51]. If the bin size in the WHAM equation is set to zero, the MBAR equation is obtained [52].

$$A_i = -\beta^{-1} ln \sum_{k=1}^{K} \sum_{n=1}^{N_k} \frac{\mathrm{e}^{-\beta U_i(\vec{r}_{kn})}}{\sum_{k'=1} N_k N_{k'} \mathrm{e}^{\beta A_{k'} - \beta U_{k'}(\vec{r}_{kn})}}, \tag{20}$$

where i takes values from 1 to K (the number of intermediate states), A_i is the free energy of state i, $\vec{r}_{kn}$ is the n^{th} sample from the k^{th} state and U_i is the potential of state i. In contrast to WHAM, MBAR allows for direct error estimation. In addition, MBAR does not suffer from a histogram bias because no histograms are used. However, the differences in free energy for MBAR and WHAM should be insignificant if the number of bins used for WHAM calculation is sufficiently large [53]. Hence, WHAM can be used to obtain reasonable error estimates of the free energy calculations if long autocorrelation times do prevent direct error estimation with MBAR.

2.5.2 Facilitated convergence through additional restraints

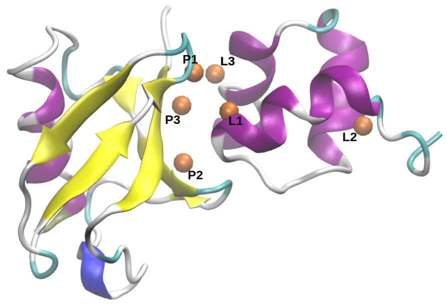

Figure 7: Definition of the local reference frame. Two triplets of beads P1, P2, P3 and L1, L2, L3, one for each molecule, are selected as reference beads. In order to restrict the relative orientation, the Euler angles $\{\theta\ (P3-L1-L2), \phi\ (P2-P1-L1-L2), \psi\ (P1-L1-L2-L3)\}$ are restrained.

Sampling the full phase-space of the molecules with all possible orientations is daunting. Hence, additional restraints can help to better sample and to reduce the simulation time needed to reach convergence of the US [54]. Roux et al. suggest different restraints that might reduce the simulation time needed, such as restraining the RMSD. They also establish a local frame of reference using three sites from each binding partner which restricts the relative orientation u_o of the two proteins. In addition, the relative position can be restrained not only based on distance but also on spherical coordinates using the local frame of reference. This restricts the movements of the ligand on the sphere that is determined by the protein-ligand distance. The relative orientation is specified by the Euler angles—a set of three angles (Θ, Φ, Ψ)—and can be used to preserve the orientation with respect to the bound state. The beads building the local frame of reference for the system studied in this Master thesis are shown in fig. 7. Restraining the relative orientation of the binding partners, K_{eq} is defined as:

$$K_{eq} = \frac{\int_{site} dR_L \int dR_P \mathrm{e}^{-\beta U}}{\int_{site} dR_L \int dR_P \mathrm{e}^{-\beta[U+u_o]}} \tag{21a}$$

$$\times \frac{\int_{site} dR_L \int dR_P \mathrm{e}^{-\beta[U+u_o]}}{\int_{bulk} dR_L \delta(r_{COM} - r^*_{COM}) \int dR_P \mathrm{e}^{-\beta[U+u_o]}} \tag{21b}$$

$$\times \frac{\int_{bulk} dR_L \delta(r_{COM} - r^*_{COM}) \int dR_P \mathrm{e}^{-\beta[U+u_o]}}{\int_{bulk} dR_L \delta(r_{COM} - r^*_{COM}) \int dR_P \mathrm{e}^{-\beta U}} \tag{21c}$$

There are different contributions to K_{eq} that influence the final binding affinity. The largest contribution comes from the differences in the PMF along the reaction coordinate:

$$I^* = \int_{site} dr\ \mathrm{e}^{-\beta\ [\mathcal{W}(r) - \mathcal{W}(r^*)]}, \tag{22}$$

where $\mathcal{W}(r)$ is the PMF along the reaction coordinate.

The term S^* describes the surface area on the sphere with radius r^* that is accessible to the ligand. This sphere is centered on the binding site:

$$S^* = (r^*)^2 \int_0^{2\pi} sin(\theta)\, d\theta \int d\phi\ \mathrm{e}^{-\beta\ u_a(\theta,\phi)}, \tag{23}$$

where r^* is a reference point in the bulk which is arbitrary chosen and θ and ϕ are the angles used to restrain the relative position of the binding partners.

At last, the contribution to the PMF of restraining the relative orientation has to be taken into account. Thus, K_{eq} can be as well written as [54]:

$$K_{eq}^{PMF} = S^* I^*\ \mathrm{e}^{-\beta(G_o^{bulk} - G_o^{site})}, \tag{24}$$

where G_o^{bulk} and G_o^{site} are the contribution of the orientational restraints in the bulk and the site, respectively.

A variety of additional restraints can be applied to a system as long as their contribution to the PMF is accounted for in the determination of $\mathrm{K_{eq}}$ via subsequent free energy calculations at the end states — bound and unbound state. The resulting binding free energy is then given by:

$$\Delta G^{\circ}_{bind} = -kTlog(K_{eq}C^{\circ}), \tag{25}$$

where C° is the standard state concentration. It is common to assume a C° of 1 mol/L (≡ 1/1661 Å^3) [54].

2.6 Implementation details

2.6.1 Preparation of structure files

To our knowledge, there is no structure file of UBQLN2's N-terminus published in the PDB database. Therefore, we used the published structure of Ubiquilin-1's UBA domain—residue 536 to 587— (PDB: 2JY5) which shares a sequence identity of 98 % with UBQLN2-UBA according the a BLASTp alignment [55]. In order to match the sequence of UBQLN1:UBA to UBQLN2:UBA, residue 555 was mutated from Serine to Asparagine using PyMol [56] and the first six residues (536 to 541) and the last residue (587) were deleted. For the simulations of the C-terminus (residues 450 to 624), the input structure was generated using the I-Tasser server [25].

For the simulations of Ub, the last two residues had to be removed using PyMol because they are glycine residues which are not compatible with the CAP group — first and last residue of each molecule — of the PLUM model.

2.6.2 All-atom simulations

The AA MD simulations were performed using the GROMACS simulation package version 5.1.2 [57] with periodic boundary conditions applied in all three dimensions. For each simulation the AMBER99SB-ILDN force field [58] and the TIP3P water model [59] were employed. The proteins were placed in a octahedron box shape with at least 1.0 nm distance to the box edges to avoid interaction with periodic images. UBQLN2:UBA and mUb were solvated using 3040 and 6525 water molecules, respectively.

If necessary, the system was neutralized by adding sodium or chloride ions. In all simulations, the leap-frog algorithm [26] was used as an integrator with a time step of 2 fs and hydrogen bonds were constrained using the LINCS algorithm [60]. The system was initially energy-minimized with the steepest descent algorithm. Subsequently, the system was equilibrated for 5 ns under constant volume and temperature (300 K). The system was first heated up from 0 K to 300 K over 0.5 ns using GROMACS implementation for simulated annealing. The temperature was maintained using the Berendsen thermostat with a friction coefficient of 0.5 ps^{-1}. The NVT equilibration was followed an NPT equilibration of 10 ns where the weakly-coupling Berendsen thermostat and barostat was used to maintain temperature at 300 K and pressure at 1 bar, respectively. In both cases, a coupling constant of 0.1 ps^{-1} was used. For the production phase, the simulations were performed under constant temperature (300 K) using velocity rescaling with a stochastic term and a friction constant of 0.1 ps^{-1}. In addition, the pressure (1 bar) was kept constant using the isotropic Parrinello-Rahman barostat with a friction constant of 2 ps^{-1}. For a more accurate thermostat, the temperature coupling for protein and non-protein atoms was separated. During the simulation time of 100 ns, the short range van-der-Waals and electrostatic interactions within a cut off of 1.2 nm were updated every time step. The long-range electrostatics were calculated with the PME method with a grid spacing of 0.08 nm and interpolation order 6. The neighborlist was updated using a grid and a cut off distance of 1.2 nm for short-range neighbor list.

From these single chain simulations the RMSD, RMSF, R_g, the helical fraction and, for mUb, the β-sheet fraction were calculated. For RMSD and R_g calculations only the backbone atoms were used to ensure comparability with the CG simulations. In the case of mUb, the last two residues were excluded from these calculations for the same reason.

2.7 Coarse-grained simulations

All simulations of single chains using the PLUM model, the US and the simulations of aggregation, were performed using the GROMACS simulation package version 4.5.4

[61] extended with the additional patches for the PLUM model [31]. All simulations employed the leap-frog stochastic dynamics integrator with a time step of 0.01 τ.

2.7.1 Stabilization of protein structure

Short single chain simulations were run for a total simulation time of 20 000 τ with periodic boundary conditions applied in all three dimensions. The systems were simulated at 300 K with a coupling constant of 1 ps. Semi-isotropic pressure coupling was used with a fixed box size in the x- and y-direction and the Berendsen barostat in the z-direction, as these simulations were initial simulations for the planned slab simulations described in section 4.4. A coupling constant of 1 ps and a reference pressure of 0 bar in the z-direction were used. The reference pressure in the x- and y-directions was 1 bar. The neighborlist was updated every 10$^{\text{th}}$ step using a grid search and a cut off of 1.5 nm. There is no explicit treatment of electrostatics in the PLUM implementation and the cut off for van der Waals forces was set to 1.5 nm. The first 500 τ of the trajectories were omitted as equilibration and were not included in further analysis of the observables.

Initially, different contact cut off values with a uniform spring constants were tested to determine how many springs are needed to stabilize the protein's structure. The SMOG server was used to the create relevant contact lists [62]. While the cut off distance differed for each of these lists the other parameters were kept the same. The default atom radius of 1 Å was used and only residues that were more than least three residues apart on the backbone sequence were considered. For using the SMOG server all hydrogen atoms had to be removed from the original PDB file. Since the SMOG server establishes a contact between two residues as soon as any two atoms of these residues are found within the contact cut off distance. The contact lists were modified and only Cα-Cα contacts within the radius were kept. Then, short simulations were run using the updated contact lists as additional bonds. The equilibrium distance for these bonds was chosen to be the distance in the original PDB file and the spring constant was varied. One of the contact cut off trials was chosen for each molecule and used for the initial model for the iterative scheme (see section 2.4. From this uniform starting point for each spring constant the constants were updated after every 20 000 τ long simulation according to the iterative scheme described in equation 15 with $\alpha = 1050$ and R_C corresponding to the contact cut off of the initial simulation. In the case of a negative spring constant, the value was set to zero for the current iteration.

2.7.2 Umbrella sampling

In order to generate the input structures for the US, the CG protein complex was placed in a rectangular box (12 nm x 5.94 nm x 4.84 nm) which made sure that the molecules could not interact with their periodic images—even at the maximal separation distance. After an energy minimization using the steepest descent algorithm, a pulling simulation was performed where UBQLN2:UBA was pulled over the course of 20 000 τ using a time step of 0.01 τ and a pulling rate of 0.001 $\frac{\text{nm}}{\text{ps}}$ along the reaction coordinate — here the x-component of the distance of center of masses (COM) — while Ubiquitin was position restrained at its initial position. Periodic boundary conditions were applied in all three dimensions and the neighbor list was updated every 10^{th} step using a grid search and a cut off of 1.5 nm The temperature was kept constant at 300 K with a coupling constant of 0.1 ps, and initial velocities were generated at this temperature as well. A cut off of 1.5 nm was used for the van der Waals forces. From the pulling simulation, configurations were saved approximately every 0.05 nm resulting 34 input configurations for the umbrella simulations. The initial distance between the two molecules was $\approx$ 2.05 nm and the configuration with greatest distance was saved at $\approx$ 3.7 nm.

Each input configuration was energy-minimized using the steepest descent algorithm and then a 400 000 τ long simulation was performed in the NVT ensemble with a harmonic biasing potential of 2500 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ while the pulling rate was set to 0. All other run parameters were equal to the above mentioned parameters for the pulling simulation except for the time step which was increased to 0.02 τ. The biasing potential was applied to all three dimensions of the COM of the protein. To help convergence, a local frame of reference was established consisting of three beads for each molecule. For mUb, the three reference beads were chosen to be the C_β beads of L8, I44 and V70—P1, P2 and P3, respectively. These residues are known to form a hydrophobic patch which is important for the binding of UBA and other Ubiquitin binding domains. [63]. UBQLN2:UBA contains two binding sites which play a role in binding to mUb [11]. Therefore, the C_β beads of residues 16 (L1) and 40 (L3) belonging to the two binding sites and residue 25 (L2) which is part of the second α-helix were chosen as reference beads in UBQLN2:UBA. The PMF was then be extracted using the GROMACS built-in function *g_wham* for WHAM calculations [49] and the python library *pymbar* for MBAR calculations [51]. While *pymbar* allows for direct error estimations, the errors of *g_wham* calculations were computed by bootstrapping trajectories [49]. To check the convergence of the PMF, the PMF was calculated for the first quarter, the first half, first three-quarters of the simulation as well as for the full-length simulation separately and the results were compared. The contribution to the PMF from restraining the Euler angles of the reference frame in the bound state was calculated using *pymbar* and the influence in the bulk was calculated directly as an angular integral. The restraining Euler angles over time were calculated using the MDtraj tools for angle computation [64]. In addition, the running average of the COM, RMSD, RMSF and R_g was calculated for each frame and compared to single simulations.

2.7.3 Simulation of aggregation

For the simulations of aggregation, 20 configurations were randomly chosen and arbitrarily placed in a 23 x 23 x 23 nm simulation box from each single chain simulation — C-terminus of UBQLN2 and mUb. The configurations were placed subsequently while making sure that the newly placed molecule was not overlapping with any other molecule in the box. For these starting configurations, short benchmark simulations were run in order to determine the ideal number of cores for each setup.

Then, the input structure was energy-minimized using the steepest-descent algorithm.

The energy-minimized structure was then heated from 0 to 500 K over 20 000 τ using simulated annealing and equilibrated at 500 K in the NVT ensemble for 300 000 τ. Before the 1 000 000 τ long production simulation was started, the system's temperature was reduced to 300 K over 10 000 τ. All other run parameters were kept the same as for US simulations 2.7.2 — except for the time constant for temperature coupling, which was reduced to 0.05 ps.

3 Results

3.1 Stabilization of protein structure

3.1.1 Stabilization of an Ubiquitin monomer

For each simulation of mUb the backbone RMSD, R_g, helical fraction, β-sheet fraction and the RMSF of the C_α beads were calculated. It is important to keep in mind that the AA model is still a model that has its inaccuracies. Therefore, the aim may not be to force the CG model to exactly reproduce the AA model. Rather, AA simulations are used as a reference in combination with experimental results.

To determine the right contact cut off value for the shadow map contacts, simulations with different contact cut off values and different, but uniform, spring constants, K, were performed. Figure 8A shows the RMSD distribution for an AA simulation, a simulation with the unrestrained PLUM model and simulations with a variety of PLUM-ELM models. The RMSD distribution of the AA simulation contains values between 0.05 and 0.2 nm which are lower than the CG simulations. The AA model gives rise to two distinct, stable conformations, as can be seen by the two modes of the RMSD distribution, centered at 0.09 nm and 0.12 nm. Regarding the RMSD over time, the molecule has a RMSD of $\approx$ 0.1 nm for the first 30 ns. Subsequently, the RMSD increases to $\approx$ 0.15 nm for the last 70 ns.

The unrestrained PLUM model also shows two peaks at $\approx$ 0.34 nm and at $\approx$ 0.4 nm in its RMSD distribution which have approximately the same height. However, the two RMSD distributions do not overlap. The PLUM-ELM model seems to stabilize only one state for mUb for all contact cut off values. The mean of the RMSD distribution decreases upon adding the springs with contact cutoff 0.7nm, and further decreases as the the contact cutoff is increased to 0.9nm. Finally, this distribution is slightly shifted to lower RMSD values as the spring constants are increased from 50 to 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$. The distributions from PLUM-ELM models with contact cutoff = 0.9 nm overlap with the more prominent mode of the AA distribution.

Regarding the R_g distributions for the different models, all distributions are found within a range of 0.1 nm (Fig. 8B). The R_g of the AA simulation is on average higher than for all CG simulations. The R_g distribution of the simulation with the unrestrained PLUM model is broader and the protein less compact in this simulation as compared to the AA simulation. This simulation shares a small overlap with the AA simulation. Applying a contact cut off of 0.7 nm results in even lower values for the R_g and subsequently a smaller overlap with the AA simulation. The R_g distributions from PLUM-ELMs with a

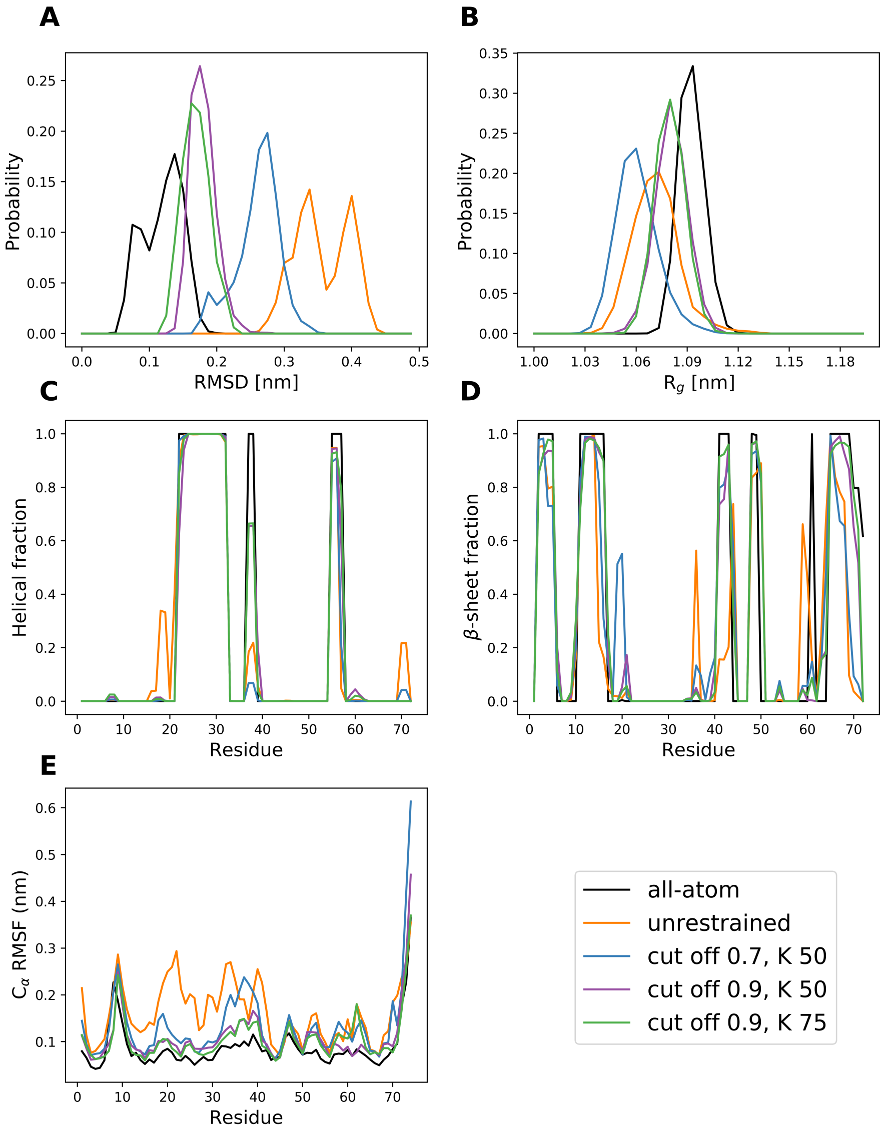

Figure 8: Results for the chosen observables for an AA simulation (black), a simulation with the unrestrained PLUM (orange) and simulations with the PLUM-ELM model with different contact cut off values and different, but uniform spring constants—contact cut off 0.7 nm and K = 50 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (blue), contact cut off 0.9 nm and K = 50 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (purple), contact cut off 0.9 nm and K = 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (green). (**A**) For the RMSD distributions only those simulations with a contact cut off of 0.9 nm do overlap with the distribution of the AA simulations. (**B**) The R_g distributions of all simulations do at least partially overlap with the R_g distribution of the AA simulation. The greatest overlap have the simulation with a contact cut off value of 0.9 nm. (**C**) The simulations with contact cut off 0.9 nm do correctly stabilize the helical regions. (**D**) In contrast, to the simulations with contact cut off 0.9 nm, the unrestrained PLUM does not stabilize all β-sheet regions correctly. (**E**) The simulations with contact cut off 0.9 nm resemble the AA RMSF the best.

contact cut off of 0.9 nm are closest to the AA simulation. The peak of both distributions is only ≈ 0.1 nm lower than the peak of the AA distribution.

Given the structure in the unrestrained PLUM model implied by the differences in the RMSD and R_g, the results suggest that the unrestrained PLUM is not able to stabilize the tertiary structure of the mUb on its own. The same applies to the simulation using a contact cut off of 0.7 nm for the PLUM-ELM model. On the other hand, enough springs are established with a contact cut off of 0.9 nm to have a good overlap with the distributions of the AA simulation. From the RMSD and R_g distributions, which give information about the tertiary structure of a protein, it is unclear to what extend the secondary structure is stabilized. Therefore, the α-helical and β-sheet fractions were calculated.

The calculated fractions for the AA simulation are in good accordance with the experimentally assigned secondary structure in the PDB file (see fig. 8C and D). The unrestrained PLUM model stabilizes well the largest two α-helical segments (residues 22 to 32 and 55 to 57). It also forms additional, short helical segments around residues 17 and 70 which are not stable over the full simulation length. Regarding the and β-sheet segments, the unrestrained PLUM model misses parts of the longer β-sheet segments (residues 11 to 16 and 65 to 72). The PLUM-ELM model with a contact cut off of 0.7 nm resembles the secondary structure of mUb well except for a short, additional β-sheet segment (residues 19 to 21) and a short, missing helical segment around residue 39. Using a contact cut off of 0.9 nm, the ELM-PLUM model stabilizes all secondary elements regardless of whether a force constant of 50 or 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ was used.

Finally, the RMSF of each simulation was analyzed in order to identify regions of higher flexibility in the proteins. The RMSF for the AA simulation of mUb shows a somewhat flexible N-terminus and a highly flexible C-terminus (Fig. 8E). The second highest peak is found around residue 8 at 0.22 nm. Residue 8 is a Leucine residue and is part of a hydrophobic patch which has been shown to be important for the interaction of mUb with UBA domains [11] and other binding partners. According to the structure, L8 is part of a turn between two β strands. The hydrophobic patch also includes two other residues, an isoleucine (residue 44) and a valine (residue 70). While V70 is part of the highly flexible tail at the C-terminus, I44 is part of the third highest peak. Our AA simulation is in good accordance with previously reported RMSF values [65], [66]. They all share the flexible C-terminus, the two next highest peaks and the plateau between residues 30 and 40. Looking at the experimental results for mUb, the B-factors also suggest a highly flexible C-terminus and that residue 8 is the most flexible internal residue. The rest of the B-factor profile differs somewhat from the shape of the RMSF presented here [67]. This might be due to the fact that Vijay-Kumar et al. calculated the B-factors from the whole amino acids while, here, only the C_α atoms were used.

In comparison to the AA simulation, the unrestrained PLUM model captures the fluctuations of the N-terminus including the peak around residue 8. It also qualitatively reproduces the peaks around residue 44 and the highly flexible tail. However, this model does not accurately represent the fluctuations along the middle section of the chain. In the unrestrained CG model, this part includes several peaks as high as the peak of residue 8 and it does not include the plateau between residue 30 and 40. The fluctuation towards the C-terminus lowers in the unrestrained CG model as well, but is not as low as in the AA simulation. The largest difference in fluctuations between the AA and the unrestrained PLUM model is to be found at residue 22 where the two simulations differ by ≈ 0.23 nm.

A contact cut off of 0.7 nm for the shadow map contacts seems to not include enough contacts to resemble the RMSF of the AA simulation. Although the fluctuations in the middle part are lowered as compared to the unrestrained PLUM model, these residues still fluctuate too much and the shape of the AA RMSF profile is not resembled in this region. Using a contact cut off of 0.9 nm and a spring constant of 50 or 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ results in an RMSF profile with all major maxima and minima of the AA RMSF, although the fluctuations of some residues remain a little too high.

Taken together, these results suggest that a contact cut off of 0.7 nm is not sufficient to stabilize mUb's protein structure. Therefore, a contact cut off of 0.9 nm and a spring constant of 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ was used as an initial model for the iterative scheme. Regarding the error of the RMSF (see SI 5), there is a large decrease in the error upon applying the uniform spring constants. During the iterative scheme the error of the RMSF fluctuates between 0.075 and 0.035 while it tends to get a bit smaller overall. Therefore, it was decided to not further optimize the spring constants after iteration 7. In that iteration four spring constant were equal to zero, the highest spring constant was 252.0 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ and the lowest non-zero spring constant was 17.8 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$. Figure 9 shows the calculated observables for selected iterations. In each plot the observable of the AA and the initial simulation are plotted as reference. The results for the other iterations can be found in the supporting information (SI) 5.

Regarding the results for the iterative scheme for mUb, there are no major changes in the observables for all iterations, except iteration 6. Adjusting each spring constant individually only marginally reduces the differences between the AA and the PLUM-ELM models. This suggests that a uniform spring constant is already sufficient to resemble AA observables in the PLUM-ELM model. Nevertheless, the mean R_g is ≈ 0.2 nm lower in iteration 6 as compared to all other iterations. This numerical difficulty might arise due to the independent treatment of the springs. This does not take into consideration that different springs established at the same C_α bead might influence each other. However, the properties of the observables are recovered in the PLUM-ELM model of iteration 7. Taken together, these results suggest that probably any of these PLUM-ELM models would sufficiently stabilize mUb in further simulations. Thus, iteration 7 was chosen to be used for the subsequent simulations.

3.1.2 Stabilization of the UBA domain of Ubiquilin-2

In addition to the above discussed simulations of mUb, single chain simulations of UBQLN2:UBA were run. Since the RMSF in fig. 10D suggest that the N- and C-terminus of UBQLN2:UBA are highly flexible, the first three residues and the last residue were excluded from RMSD calculations because they might have a significant impact on the

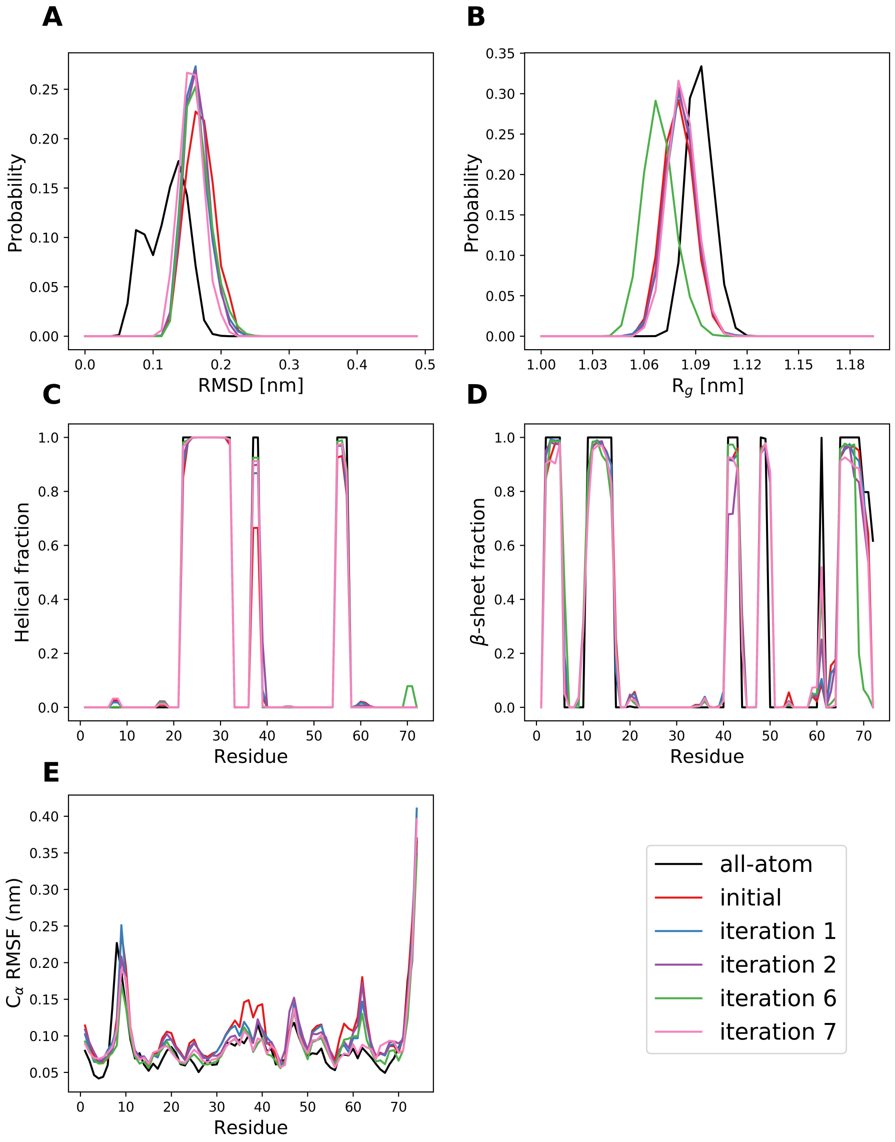

Figure 9: Results for chosen observables for an AA simulation of mUb (black), the chosen initial simulation (red) and iteration 1, 2, 6 and 7 (blue, purple, green, pink, respectively).(**A**) The RMSD distribution stays approximately the same for all iterations. (**B**) Except for iteration 6, the R_g distributions of all iterations are only slightly shifted from the AA simulation. (**C**) All iterations stabilize the helical regions. (**D**) The five β-sheets are sustained in each iteration. (**E**) While the initial simulation does not resemble all peaks of the RMSF, iteration 7 resembles most major peaks and the general shape of the RMSF.

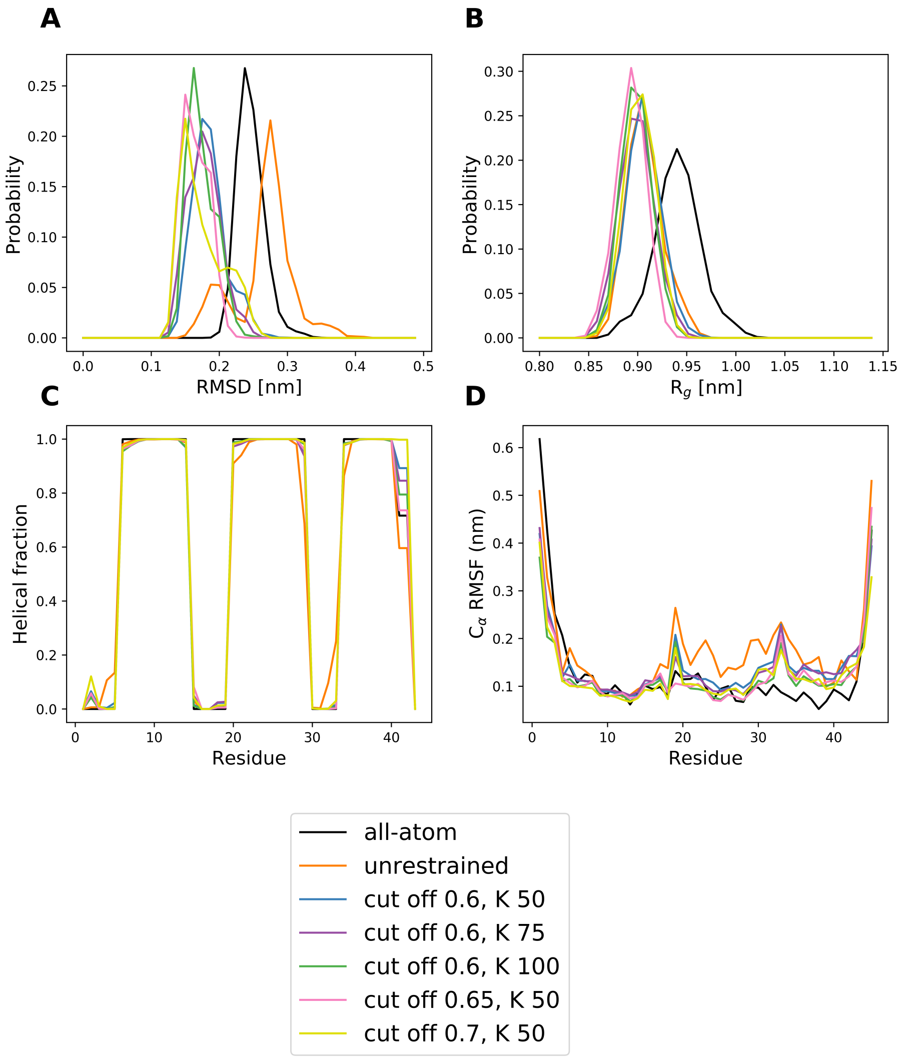

Figure 10: Results for the chosen observables for an AA simulation of UBQLN2:UBA (black), a simulation with the unrestrained PLUM (orange) and simulations with the PLUM-ELM model with different contact cut off values and different, but uniform spring constants—contact cut off 0.6 nm and K = 50 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (blue), contact cut off 0.6 nm and K = 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (purple), contact cut off 0.6 nm and K = 100 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (green), contact cut off 0.65 nm and K = 50 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (pink) and contact cut off 0.65 nm and K = 50 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ (light green)

overall RMSD distribution. This might result in high RMSD although the rest of the structure might be very similar. In comparison to mUb, the RMSD of the AA simulation of UBQLN2:UBA is higher, which might indicate more structural flexibility, e.g. in the arrangement of the three-helix bundle (Fig. 10A). The distribution of the AA simulation contains values from 0.2 to 0.3 nm and only has a single peak. The distribution of the unrestrained PLUM model is slightly shifted to higher RMSD values and includes a second, lower peak which might indicate a structural rearrangement throughout the simulation. The RMSD distribution of all simulations with the PLUM-ELM model is shifted to lower RMSD values as compared to the AA distribution. The single peaks of the PLUM-ELMs model with a contact cut off of 0.6 nm and a spring constant of 50 and 75 $\frac{\mathrm{kJ}}{\mathrm{mol} * \mathrm{nm}^2}$ are closest to the AA distribution. The higher the contact cut off or the spring constant of the PLUM-ELM models, the more the RMSD distributions are shifted to lower values. The PLUM-ELM model with a contact cut off of 0.7 nm and spring constant of 50 $\frac{\mathrm{kJ}}{\mathrm{mol} * \mathrm{nm}^2}$ has its peak for the RMSD distribution at the lowest value. In contrast to the other PLUM-ELM models, this setup has small plateau region in its RMSD distribution around 0.22 nm.

Fig. 10B shows the $\mathrm{R_g}$ distribution for the different simulations of UBQLN2:UBA. The $\mathrm{R_g}$ distribution of the AA simulation contains values from 0.85 to 1.03 nm and is broader than the $\mathrm{R_g}$ distribution of mUb which underlines the possible higher structural flexibility of the UBA domain. Overall, the values for the three-helix bundle are lower than for mUb which might be simply due to the different sequence length of the two molecules. Comparing the AA $\mathrm{R_g}$ to the CG simulations, the protein seems to be more compact in the PLUM model. For all CG simulations, the single peak of the $\mathrm{R_g}$ distribution lies 0.05 nm lower, at ≈ 0.9 nm, compared with the distribution of the AA simulation. The overall $\mathrm{R_g}$ distribution of the CG simulation is more narrow which might indicate less structural rearrangement as compared to the AA simulation.

Regarding the secondary structure of the three-helix bundle, the three α-helices are stabilized in all three classes of simulations. While the unrestrained PLUM model elongates the first and the third α-helix in some parts of the simulation, a contact cut off of 0.7 nm for the PLUM-ELM model over-stabilizes the C-terminus of the UBA domain. But both of these offsets are quite small as compared to the experimental results of Dao et al. [10].

The RMSF of the AA simulation suggests a highly flexible N- (residue 1 to 4) and C-terminus (residues 43 and 44). In comparison with mUb, there are no such clear peaks around the binding sites — residue 16 to 18 and residue 37 to 43 — as for mUb. The two highest peaks — except the termini — are found at residues 19 and 22, which is the region directly before and after the first binding site and at the beginning of the second α-helix. Before and after that regions the overall fluctuations are lower and range around 0.1 nm. The unrestrained PLUM model resembles the fluctuations of the C_α beads belonging to the first α-helix quite well, but allows too large fluctuations for the rest of the residues. All simulations using the PLUM-ELM lower the RMSF values as compared to the unrestrained PLUM model but a contact cut off value greater than 0.6 seems to establish too many springs as the RMSF values become lower than the RMSF of the AA simulation for residue 15 to 25. On the other hand, all simulations with the PLUM-ELM model have a peak at residue 33 where the results differ by around 0.14 nm as compared to the AA simulation.

Because, to our knowledge, no MD simulations have been previously conducted for UBQLN2:UBA, simulations of other UBA domains are used to compare the RMSF of UBQLN2:UBA. Although the UBA domains of UBQLN2 and the autophagosome cargo protein p62 hardly share any sequence identity, the RMSF profiles of these two UBA domains do have a similar shape [68], [69]. Evans et al. performed NMR spectroscopy as well as MD simulations in their investigations. Both methods suggest a highly flexible N- and C-terminus as in our results for UBQLN2:UBA. While they find a residue of higher flexibility between the first and the second α-helix in their simulation, there are no major peaks throughout the RMSF profile for their NMR results with values between 0.05 and 0.1 nm. Similar results were published by Teixeira et al. [69]. Overall, this suggests a low flexibility of the Cα beads which implies a high stability of the secondary structure underlined by the results for the helical fraction.

As all tried combinations of contact cut off values and uniform spring constants yield similar results, and there are no major outliers, the setup with contact cut off of 0.6 nm and a spring constant of 75 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ was chosen as initial model for the iterative scheme because it best resembled the RMSF and had one of the greatest overlaps with RMSD distribution of the AA simulation. Regarding the error of the RMSF (see SI 5), the initial decrease of the error upon adding springs is smaller than for mUb. The error of the RMSF fluctuates between 0.17 and 0.05 throughout the iterative scheme. However, as the RMSD distribution tended to be narrowed with each iteration, the iterative scheme for UBQLN2:UBA was stopped after 10 iterations. In this iteration the highest spring constant was 554.2 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ and the lowest, non-zero spring constant was 13.6 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$. Three spring constants were set to 0 in this iteration. The calculated observables for some selected iterations are shown in fig. 11. The plots for all other iterations can be found in the SI 5.

The RMSD distributions for iterations 1, 2, 5, 9 and 10 can be seen in fig. 11A. As compared to the initial and the AA simulation, the RMSD distribution for all iterations—except iteration 1—are more narrow and slightly shifted to the left with a single peak around 0.15 nm. This indicates that the additional springs whose stiffness tend to get higher throughout the iterative scheme reduce the structural flexibility of the UBA domain. In constrast, iteration 1 has two peaks in its RMSD distribution. Looking at the RMSD over time, the molecule seems to visit two states. First it is found in the same state as the molecule is in the other iterations. Then it switches to the second state for about 5000 τ before going back to the initial state. This indicate either that there might be another stable state which is not visited in the AA simulation or that the springs might cause some unphysical rearrangement which is fixed in the following iterations.

Comparing the R_g distributions, the distributions of the simulations of the iterative scheme are a little broader than the one of the initial simulation and the peaks for iteration

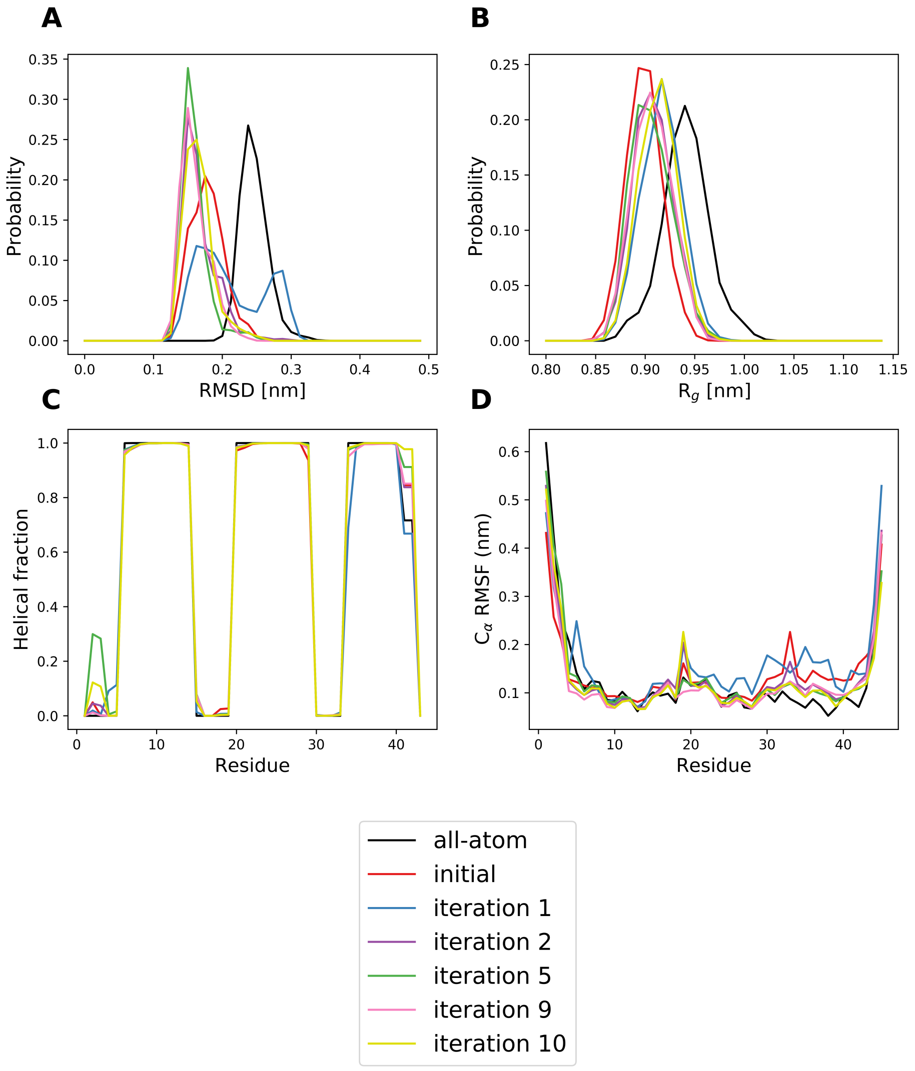

Figure 11: Results for chosen observables for an AA simulation (black), the chosen initial simulation (red) and iteration 1, 2, 5, 9 and 10 (blue, purple, green, pink, yellow respectively). While iteration 1 does not properly resemble the observables of the AA simulations, escpecially the RMSD, the difference of the other iterations to the AA simulations is smaller. Iteration 9 seems to be the best fit.

1 and 10 are slightly shifted to the right, closer to the AA distribution (Fig. 11B). Similar to mUb, the secondary structure was already well stabilized in the unrestrained PLUM and the PLUM-ELM model with uniform spring constant. Therefore, it was not expected to see much change among the different simulations of the iterative scheme and the initial simulation. Small outliers are seen in iteration 5 and 10 where a small helical segment is formed before the first α-helix during 30 % and 12 % of the simulation time. Furthermore, the end of the third α-helix is over stabilized in all iterations except iteration 1.

Regarding the RMSF, the C_α beads fluctuate less as compared to the initial simulation in all iterations except iteration 1. While the shape of AA RMSF profile is well resembled for the first 12 residues and residues 22 to 30, residue 19 fluctuates more than in the AA simulation in iterations 1, 2 and 10. The flexibility of all C_α beads following residue 30 is higher than in the AA simulation for all iterations with the biggest difference in iterations 1 and 2. In general, all observables for the PLUM-ELM model of iteration 1 suggest a similar numerical difficulty as it was observed in the iterative scheme of mUb in iteration 6.

Taken together, the results for the single chain simulations of UBQLN2:UBA suggest that the unrestrained PLUM model can reasonably stabilize the structure of the three-helix bundle. While the secondary structure is nicely stabilized there is a small offset for the RMSD and R_g. The biggest difference can be seen in the RMSF which is reduced by the iterative scheme. However, it was unclear whether these results would be transferable into the condensed phase. Hence, the additional springs were still applied for UBQLN2:UBA to ensure the stability of this domain during the simulations of aggregation, which has been shown experimentally. As all simulations of the iterative scheme showed similar results and all simulations, except iteration 1, reasonably stabilized the structure of UBQLN2:UBA, the following simulations, US and simulations of aggregation, were all run with the set of spring constants from iteration 9 to stabilize UBQLN2:UBA's structure.

3.2 Umbrella sampling

After stabilizing the structure of the proteins, the binding affinity of mUb and UBQLN2:UBA was calculated using US. Several attempts were needed to identify a biasing force that was high a enough to sufficiently restrain the COM distance between the two proteins. A histogram plot for a biasing force that was too low can be found in fig. 12 A. In this plot, the histograms are not evenly distributed along the reaction coordinate. The COM distance, especially of the middle windows, does not fluctuate around the set equilibrium position, but UBQLN2:UBA moves towards mUb throughout the simulation. Although not all regions along the reaction coordinate seem to be sampled sufficiently, the resulting PMF is continuous. However, the larger error bars around a COM distance of 3.25 nm are likely to indicate that more sampling is needed in that region. (Fig. 12B).

Fig. 13 shows the histogram and the running average for the full simulation length of 400 000 τ and the PMF. For all US simulations a biasing force of 2500 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ was used. To extract the PMF from the biased simulations, WHAM and MBAR were employed. Normally, MBAR calculations would allow for direct error estimation. In

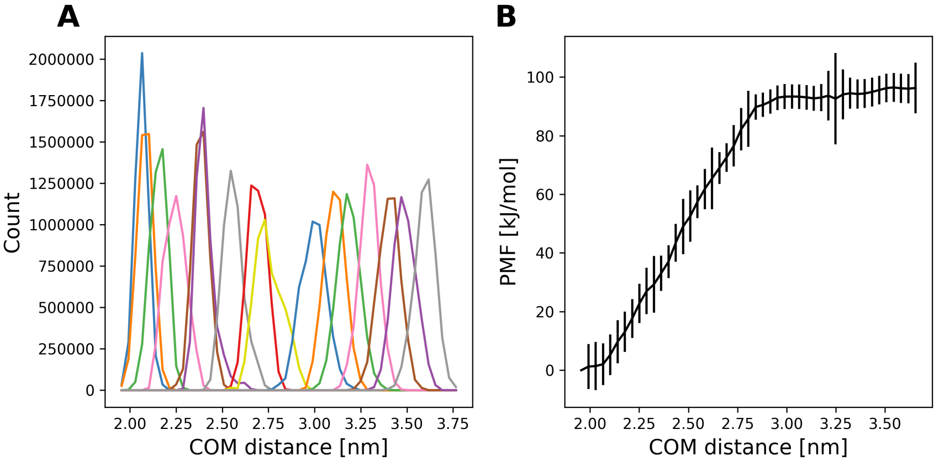

Figure 12: Histograms (**A**) und PMF profile (**B**) for an umbrella sampling with a biasing force of 750 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$ and no additional restraints.

this study, it was not possible to obtain an uncorrelated subsample from the trajectories due to long autocorrelation times. This is a prerequisite for reasonable error estimates for MBAR. Nonetheless, the error bars for the approximate PMF could be calculated using bootstrapping [50]. As this method is already implemented in *g_wham*, WHAM calculations were used to obtain the PMF and its error estimates. To ensure that a sufficiently small bin size was chosen for WHAM [53], the results were compared to the MBAR calculations with the correlated sample. They were both the same within error. Therefore, the results for the WHAM calculations are presented here and the results for the MBAR calculations can be found in the SI 5.

The histograms for all windows can be seen in figure 13A. All histograms do overlap with their neighboring windows which is important for a sufficient sampling along the reaction coordinate. The peaks of the histograms are evenly distributed over the considered region of the reaction coordinate. This indicates that the chosen spring constant is high enough although some distribution of COM distances are broader than others.

Regarding the running averages, they seem to stabilize for all windows after ≈ 20 000 τ (Fig. 13B). For most windows the running average is found around the initial COM distance. But in some simulations a shift of the running average towards a smaller COM distance is observed during the first 20 000 τ of the simulation. This is especially the case for the windows between 2.25 nm and 2.6 nm. After that drop in the running average, it stabilizes between the initial COM distance and the shortest distance between mUb and UBQLN2:UBA. Complementary, the average COM distance and its standard deviation of each window can be found in the SI 5. The average COM distance confirms the initial shift for some frame, especially for the first three windows. The distance between these windows shrank from 0.5 nm to ≈ 0.2 nm.

The resulting PMF profiles for 100 000, 200 000, 300 000 and 400 000 τ differ slightly in their shape while the difference between the lowest and the highest PMF value stays the

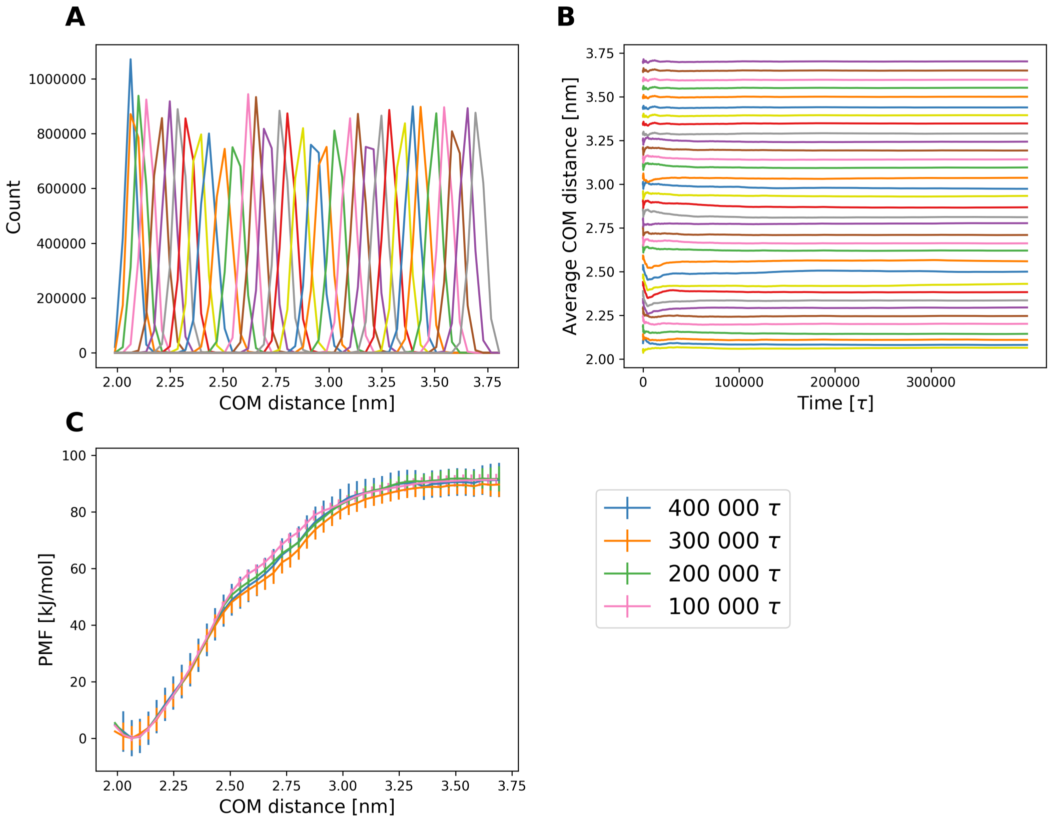

Figure 13: Histograms (**A**), running averages (**B**) over 400 000 τ and PMF calculations (**C**) based on WHAM using a biasing force of 2500 $\frac{\text{kJ}}{\text{mol} * \text{nm}^2}$

same within error (Fig. 13C). The minimum of the PMF for all parts of the simulations lies at 2.06 nm and the plateau is reached at ≈ 3.2 nm. The first three-quarters of the umbrella simulation have the lowest values in their PMF profile. The difference between the minimum and the plateau region of this PMF profile is ≈ 89 kJ/mol Calculating the PMF profile for the other parts of the simulations, the average PMF value in the plateau region of the PMF profile increases by ≈ 2 kJ/mol. Although the shape of the PMF profile of these parts still differs a little, the values at the end points of the US — the bound and unbound state — are the same. The errors for each part of the simulation were calculated using bootstrapping and range between 4.5 to 7.2 kJ/mol.

Fig. 14 shows the comparison of the RMSD and RMSF of the bound state, the unbound state and the single chain simulations. One would expect that the results of the single chain simulations are recovered as soon as the two proteins are not interacting any more — in the unbound state. While this expectation is met for UBQLN2:UBA, the results for mUb differ more from the single chain simulation in the unbound state. The peak for the RMSD distribution of mUb slightly shifts to the right. But the two distributions still share an overlap. In the bound state, the RMSD distribution is a bit narrower with the a single

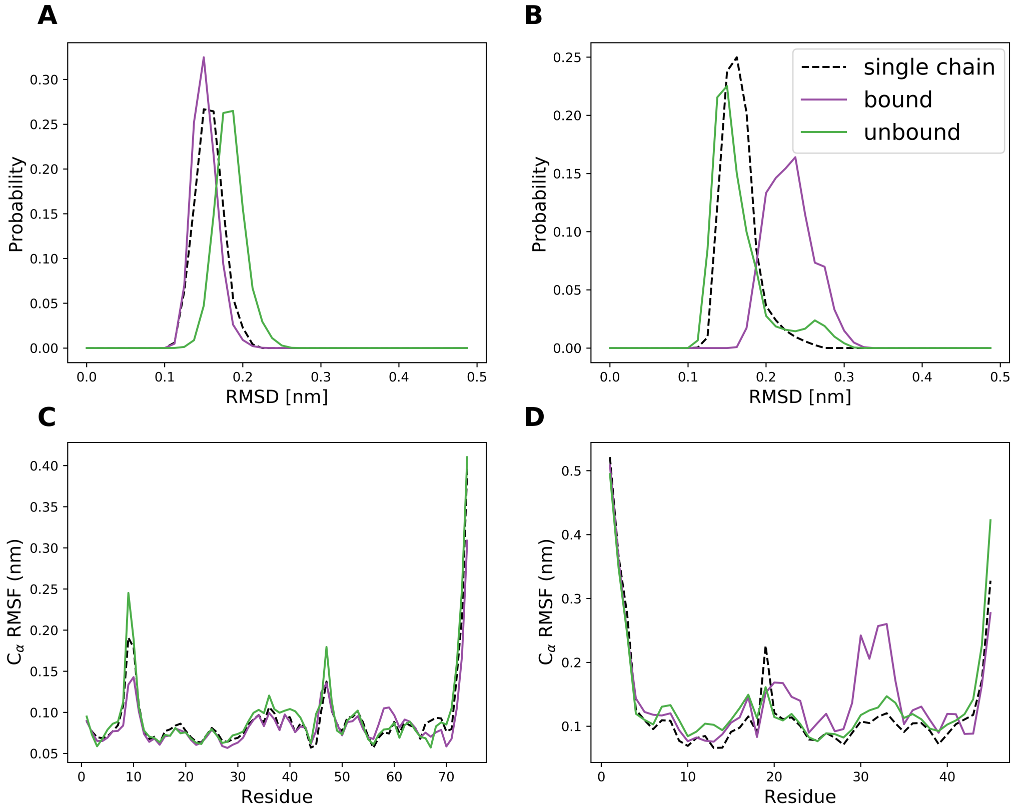

Figure 14: RMSD and RMSF for the bound and unbound state compared to the results for single chain simulations. (**A**) The RMSD for mUb slightly increases in the unbound state as compared to the single chain simulation and the bound state. (**B**) The RMSD for UBQLN2:UBA is higher in the bound state than in the unbound state and for the single chain. (**C**) The difference between the RMSF for the bound and unbound state and the single chain simulation is small. (**D**) The RMSF of the bound state differs more from the single chain simulation than the RMSF of the unbound state.

peak that is almost not shifted as compared to the single chain simulation (Fig. 14A). Regarding the RMSF, the most flexible residue and the region towards the C-terminus fluctuate a bit less in the bound state, while C_α beads around residue 55 to 60 fluctuate a bit more. In the unbound state, the two highest peaks of the RMSF profile are a bit higher than in the single chain simulation. The RMSF profile of the bound and the unbound state is still very similar to the RMSF profile of the single chain simulation (Fig. 14C).

The differences between the bound and the unbound state are bigger for UBQLN2:UBA. While the RMSD distribution for the unbound state is in good agreement with the distribution of the single chain simulation, the bound state only shares little overlap with it. The RMSD of the bound state is higher than the RMSD of the single chain (Fig.14B). Comparing the RMSF for the different windows, the C_α beads fluctuate more for the bound as well as for the unbound state. In the unbound state, the shape of the RMSF profile of the single chain simulation is mostly resembled. The biggest differences can be seen around

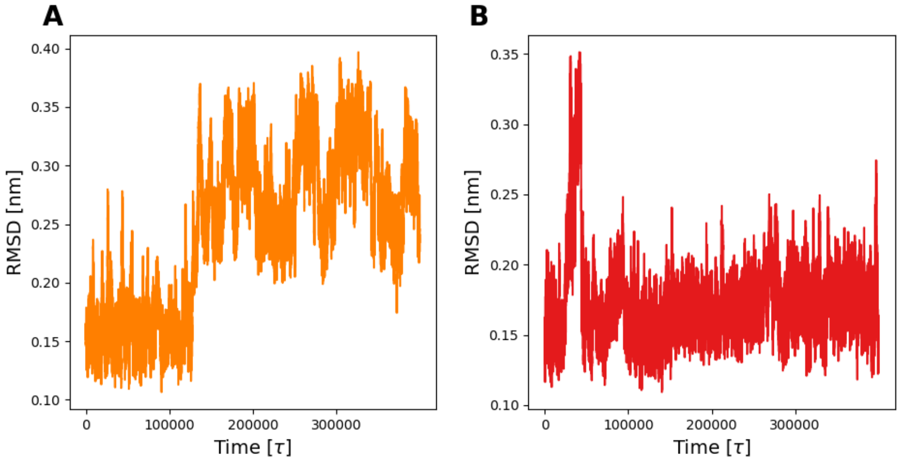

Figure 15: RMSD over time for the first and second run of the window with average COM distance 3.095 nm. (**A**) Simulation was run with the default random seed (1993) for the integrator. (**B**) RMSD of the single chain is mostly recovered using 619230 as random seed.

residue 10 to 20 and he region towards the C-terminus. In contrast, the RMSF differs a lot for the bound state from the single chain simulation. There are two higher peaks around residue 30 to 35 that are not found the profile of the single chain simulation.

The fact that the differences of the observables between the bound and the unbound state are smaller for mUb, is in accordance with the experimental results of Zhang et al. [11] for the UBA domain of Ubiquilin-1. Their results suggest that the structure of mUb is not changing upon binding to the UBA domain while the α-helices are slightly rearranging. Hence, the changes in the RMSD might indicate this structural rearrangement. Regarding these observables, one always has to keep in mind that they are obtained from biased simulations. These simulations are designed to calculate the binding affinity. Therefore, one cannot trust the apparent pathways to the bound state because restraints are added to the simulations which affect the resulting available structures.

In addition, the RMSD of the single chain simulation was recovered for most, but not for all of the windows in the unbound state. For UBQLN2:UBA it seems like there are three different states of the RMSD: the one of the single chain simulation, one around 0.3 nm and one around 0.22 nm that matches with the distribution for the bound the state. In some windows, the UBQLN2:UBA visits all three states throughout the simulation. In other simulations UBQLN2:UBA is first found in the state around 0.15 nm and then switches to a state with a higher RMSD. While UBQLN2:UBA only spends a short part of the simulation in this state for some windows, it stays in this state with a higher RMSD for other windows. The RMSD distributions for mUb are more stable between the different windows. The RMSD distribution for most windows is either very close to the distribution of the single chain simulation or slightly shifted to the right.

For the windows that did not recover the RMSD of the single chain simulations, the simulations were re-run using different random seeds. Fig. 15 shows the RMSD of UBQLN2:UBA over time for the window with the average COM distance of 3.095 nm.

The plots for the other simulations that were re-run can be found in the SI. Using the default random seed for the integrator (1993), the RMSD of UBQLN2:UBA fluctuates around 0.15 nm for the first 100 000 τ. After that, the RMSD ranges between 0.25 and 0.35 nm (Fig. 15A). For the second run of this window 619230 was used as random seed for the integrator. In contrast to the first simulation, the RMSD of this simulation mostly recovers the RMSD of the single chain simulation. UBQLN2:UBA only switches states for a short part of the simulation around 50 000 τ (Fig. 15B). In order to check the effect of this difference in the RMSD on the PMF, the PMF was re-calculated including both simulations for the windows that were re-run. The resulting PMFs only differ within error (see SI 5).

Euler angle	unrestrained [°]	restrained [°]
Θ	93.55 ± 12.67	95.01 ± 2.92
Φ	91.89 ± 18.24	92.58 ± 4.41
Ψ	4.50 ± 29.63	1.15 ± 3.93

Table 1: Averages and standard deviations of the Euler angles in the restrained and unrestrained bound state.

The long autocorrelation times suggested that an additional equilibration time might be needed. This need was underlined by the initial shifts of some frame seen in the plot of the running average 13B. For this purpose, the average COM distance was calculated over the a equilibration time of 50 000 τ and the last frame at this average was saved as input conformation of the production phase. While this procedure worked for most of the umbrella windows, another configuration had to be saved for some of the windows because the initially saved configuration led to an unstable simulation. It is expected that is problem has to do with the additional restraints of the Euler angles. It might be avoided more robustly by taking into consideration the averages of these angles as well for choosing an input structure for the production phase of the US. Preliminary results for the US simulations with a preceding phase of equilibration show that autocorrelation times are shortened for some umbrella windows while they are elongated for others. However, they are still too long to allow reasonable error estimates from the MBAR calculations. The results for the PMF are the same within error as for the US simulations without the equilibration. As the additional equilibration time seems to make no difference in the results for the PMF, the results without the additional equilibration were used for the final calculation of the binding affinity.

Table 1 shows the averages and the standard deviation of the Euler angles in the restrained and the restrained bound state over the full simulation length. These were calculated to check whether the Euler restraints are working correctly. The standard deviations are significantly smaller for all Euler angles in the restrained bound state. This suggests that the restraints of the Euler angles were correctly applied.

The contribution of the orientational restraints in the site and the bulk are written in table 2. The contribution of these restraints to the PMF in the unbound state was calculated by numerical integration as the bulk is isotropic. Their contribution to the PMF in the bound state was calculated via MBAR. In addition to the long autocorrelation times, the PLUM-ELM model had problems to stabilize the Euler angles in the bound state without additionally restraining them (see SI 5). Hence, the calculated contributions in the bound state are an approximation to the correct values. As S^* and I^* are both dependent on the choice of the reference point r^*, S^*I^* was calculated using three different reference points (see SI 5). They were found to differ ≈ 0.49 kJ/mol at most. Hence, r^* was chosen to be 34.71 Å based on these calculations. In addition, I^* depends on the region of integration, which should include the binding site only. This binding site is not well-defined. However, integrating over the full range of $\mathcal{W}(r)$ or + 4 Å from the minimum, results in a difference of $\approx 2.5 * 10^{-7}$ kJ/mol in the final binding affinity because the ascent of the PMF is steep.

Restraint	Contribution to PMF [kJ/mol]
G_o^{site}	-0.152 ± 0.155
G_o^{bulk}	12.2

Table 2: Contribution to the PMF of the orientational restraints in the bound and unbound state.

Taking together all contributions to the PMF, the final binding affinity is -77.31 kJ/mol. This is about 46 kJ/mol higher than the experimentally determined binding affinity of Dao et. al [10]. Although the binding affinity of the PLUM-ELM model does not perfectly align with the experimental results, they are still in the right order of magnitude.

3.3 Simulation of aggregation

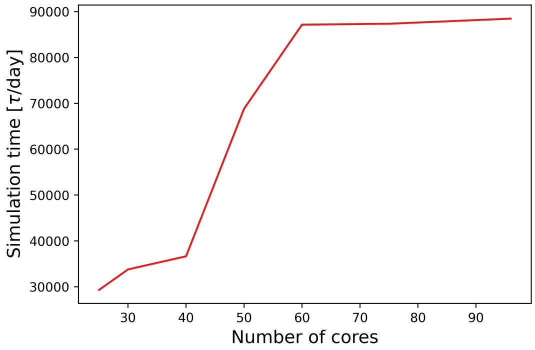

Figure 16: Average simulation time in ns/day in benchmark simulations using different numbers of cores

.

Unfortunately, a full analysis of the simulations of aggregation was out of scope for this Master thesis. Some of the preliminary results are presented in this section. Fig. 16 shows the plot for the short benchmarking simulations for the largest studied system—20 molecules of UBQLN2:C-terminus (residues 450 to 624) and 20 molecules of mUb. This system has 28 260 beads. First, the performance is slowly rising from 25 to 40 cores. The biggest rise can be seen between 40 and 60 cores. After that, the performance only increases little with a higher number of cores. In combination with the rising load imbalance for more cores, this suggests that using 60 to 65 cores is a reasonable choice for this system.

Examining the simulations visually in VMD, it was observed that all molecules diffuse slowly even at a temperature of 500 K during equilibration. There are some changes and new interactions throughout the production simulation which was performed at 300 K over 1 000 000 τ. But the formation of these contacts takes a long time. Fig. 17 shows the contact map for a simulation of UBQLN2:UBA domains. The highest contact probability is observed between residues 4 to 6 and 16 to 18. Residue 4 to 6 have a higher probability to interact with residues towards the C-terminus as well. Both of these regions have been identified as potentially important for the interactions between UBQLN2:UBA and mUb [11]. This motivates future studies on these interactions using the PLUM-ELM model.

4 Discussion

4.1 Limitations of the PLUM model that have to be further investigated

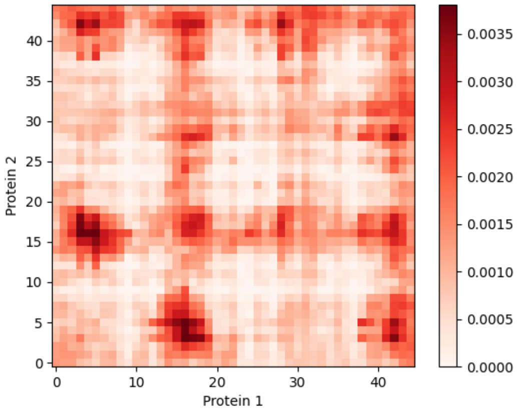

Figure 17: The residue-residue contact map of a simulation with UBQLN2:UBA domains shows an overall low contact probability for the first part of the simulation.

.

Like all other AA and CG models, PLUM is a model that has its advantages and potential pitfalls. Because the PLUM model is an implicit solvent model, it is not able to properly model the hydrophobic effect. The hydrophobic effect describes the apparent attraction between hydrophobic solute molecules due to the entropically-preferred rearrangement of (polar) solvent molecules. In an implicit solvent model these interactions have to be represented based on the solute's coordinates only which limits the ability to describe changes in this effective interaction, e.g., as a function of thermodynamic state [30]. For example, the current implementation of the PLUM model cannot capture the temperature dependence of this hydrophobic effect. With rising temperature, the free energy of solvation of residues changes which leads to the collapse of some IDPs upon increasing temperature. However, this change in the effective interactions cannot be modeled by a simple, fixed interaction potential. The representation of LCST transitions with implicit sovent models can be achieved through either temperature dependent interactions between solute molecules [70] or more sophisticated functional forms for the interactions that depend on properties of the environment, such as the solvent accessible surface area [71].

Moreover, environmental factors, e.g. the ionic concentration, are important in modulating LLPS of proteins. Modeling the influence of these factors in MD simulations requires the explicit treatment of electrostatics. Since there are no explicit electrostatics included in the PLUM model [31], the current version of the PLUM model is not able to capture the influence of the ionic concentration and other environmental factors. Future work should incorporate explicit electrostatics into the PLUM model, e.g., using the

Debye-Hueckel formalism.

Bereau et al. hypothesized that their implementation of the PLUM model might have packing problems for stabilizing the tertiary in globular proteins [31] which is underlined by the results for mUb. All side chain beads have the same van der Waals radius. This does not take into account the size differences of different amino acids. While alanine only has a methyl group as side chain, the side chain of arginine consists of a 3-carbon aliphatic straight chain ending in a guanidino group. Different bead sizes could be incorporated in the PLUM model via parameter refinement based on experimental results [72] or atomistic simulations [73], [30].

In addition, Rutter et al. identified problems in sampling the (ϕ, ψ) coordinates correctly for prolines [38]. The side chain of proline is cyclic and is bonded to the C_α and N atom of the backbone. The results of Rutter et al. show that the PLUM model is not able to stabilize a left-handed polyproline helix. This secondary structure element is not stabilized through hydrogen bonds but rather through the interaction of neighbored amide groups [74]. As the percentage of proline residues in mUb and UBQLN2:UBA are low, 4 % and 2 %, respectively, the two proteins do not form polyproline helix and this inaccuracy of the PLUM model should not influence the results of the folded domains presented here. However, this issue should be kept in mind when studying the PXX domain of UBQLN2 because 31 % of the residues in this domain are prolines. At the same, the structural propensity of this domain is low [10].

Another downside of the force field that has arisen during this Master thesis are the slow dynamics of larger systems. This slow dynamics are at least partially caused by the high resolution of hydrogen bonds in the PLUM model and limit the application of it to much larger systems. Regarding the setup that includes 20 molecules of the C-terminus of UBQLN2 and mUb (28 620 beads), $\approx$ 67 000 τ/day can be simulated with the current implementation of the PLUM model. Aiming at a simulation length of 1 000 000 τ for the simulations of aggregation, this results in a computing time of about 2 weeks for one simulation. This simulation would have to be conducted at different temperatures to allow deriving a phase diagram. In this Master thesis GROMACS version 4.5 was employed. Implementing the PLUM model for a newer GROMACS version, would require changes in the source code as the hydrogen bonding interactions are specialized. Therefore, the original implementation of the PLUM model was used in this study. However, the large amount of computational resources needed for the planned study (see section 4.4) suggest to implement PLUM on a faster platform in future studies, such as LAMMPS [75].

4.2 Stabilization of protein structure

The results of this thesis demonstrate that the PLUM-ELM model can effectively stabilize the structure of mUb and UBQLN2. The additional springs can be applied only locally which enables the simulation of ordered and disordered domains at the same time.

While proteins with a simpler structure are already reasonably modeled by the unrestrained PLUM model, the PLUM-ELM model facilitates the modeling of mUb, a more complex, globular protein. One reason for the lower accuracy of PLUM for mUb might be that the PLUM model was mainly parameterized using three-helix bundles. On the other hand, mUb contains several short helical segments and five β-strands that have to be correctly arranged in space. Both of these secondary structure elements are stabilized by backbone hydrogen bonds. Despite the relatively sophisticated hydrogen bonding interactions in the PLUM model, the limitations of the PLUM model discussed above prevent an accurate modeling of β-sheets in mUb. The results for the unrestrained PLUM model show that there might be some additional parameter tuning needed in order to fully stabilize this secondary structure element.

According to Bereau et al., another explanation for the difficulties in the arrangement of tertiary structures could be the problem of realistic packing in globular proteins [31]. Adding enough additional springs to model, fixes this problem to some extent. These springs harmonically restrain the distance of specified C_α beads around a chosen equilibrium. This procedure allows structural rearrangements to some extent while keeping the structure as close to the initial structure as needed. The balance between those two can be managed by the number and the strength of the additional springs. A higher number of springs with higher spring constant result in a more restricted structure.

Hence, the aim should be to use enough restraints to resemble the RMSF while, at the same time, not restraining the structure too much, since this would influence the RMSD distribution. This seems to be especially important for more flexible proteins such as UBQLN2:UBA. While the RMSD distribution of the unrestrained PLUM model is close to the AA simulation, it is significantly lowered when applying the PLUM-ELM model. At the same time, the fluctuations of the C_α beads are better resembled when the additional springs are applied. This suggests that the impact of the number of springs is higher than the particular spring constants used. Indeed, a set of uniform spring constants is already sufficient to stabilize the structure of the two proteins studied in the PLUM-ELM model.

Here, the shadow map approach [42] was used to determine the number of contacts for the addition of springs. Evaluating the PLUM-ELM model, one possibility would be to try different contact maps. Globisch et al., for example, establish a contact between two C_α beads when their average distance is smaller than a cut off and they are at least two residues apart. In addition, they checked whether the correlation coefficient of the covariance of an atom pair is higher than a predefined minimum or the variance of their is smaller than a maximum [43]. The advantage of this procedure is that it is not solely based on the initial structure and avoids erroneous bonds. Using the overlap criterion, a native contact is formed if the van der Waals spheres of any heavy atoms of different residues overlap [76]. Another approach that takes into account structural and chemical properties is the repulsive contact of structure unit. It differentiates between proper and destabilizing contacts and a contact is only established if the number of proper contacts is higher than the number of destabilizing contacts [77]. In the present study, all of these options would result in slightly different contact maps, as they focus on different features. This might lead to different results and a better or worse stabilization of protein structure as the number of stabilizing springs would differ for each approach.

Poma et al. used the latter two contact maps for their extension of the MARTINI model [78]. Instead of connecting beads in contact with harmonic bonds, they used

Lennard Jones potential interactions. This idea was already implemented for Gō-models and allows more structural flexibility in the MARTINI model. Earlier versions of the MARTINI model did not allow for this flexibility because the structure was kept close to the native structure using a stiff elastic network [79], [43]. Poma and co-workers state that their results for proteins with a highly conserved structure, like mUb, are still in good agreement with AA simulations. Hence, this might be one possibility to avoid restricting the RMSD too much. On the other hand, the results of this Master thesis show that the PLUM-ELM allows for some structural flexibility.

Investigating the phase-separating behavior of UBQLN2 in combination with mUb using the PLUM-ELM, is one aim of future studies. As compared to other CG models for studying LLPS, the structural flexibility in the PLUM-ELM might be an advantage. Most of these models treat the ordered regions as rigid bodies that do not have any structural flexibility. For example, Dignon et al. developed a sequence-specific bead-spring type model where one amino acid is mapped to one bead [18]. There are two different versions of the model. One is based on a hydrophobicity scale and the other one uses the paramaterized functional form of the Miyazawa Jerningan potential [34], [80]. There are three reasons why both version of this model are computational more efficient than the PLUM(-ELM) model. First, less interactions need to be calculated due to the lower resolution of the model of Dignon et al. Second, their model allows for greater time steps than the PLUM model. They were using a time step of 10 fs, while the greatest time step that was employed for the PLUM model so far is 3 fs [38]. At last, the model of Dignon et al. has a smoother free energy landscape than the PLUM-ELM model. The high resolution of hydrogen bonds results in a more rugged free energy landscapes, e.g. slower transitions between structures.

The higher computational efficiency of the model of Dignon et al. is traded against a lower structural accuracy. As the model does not distinguish between the backbone and the side chain of amino acids, it is limited in the distinction of different types of interactions [72]. This sensitivity might be important because further analysis with a modified hydrophobicity scale suggested that backbone interactions are likely to play a role in driving LLPS [72]. On the other hand, the PLUM-ELM model is capable of distinguishing these interactions and investigating the role of sidechain-sidechain and sidechain-backbone interactions in LLPS.

Furthermore, the more generic model of Dignon et al. does not form any secondary structure [18]. As this further reduces the computational resources needed, it denies that disordered domains can still have a structural propensity and might temporarily form secondary structure. Whereas, the high resolution of the backbone in the PLUM model enables secondary structure formation for ordered as well as disordered domains. This results in a higher structural accuracy of the PLUM model.

Comparing PLUM-ELM with the model of the Dignon et al, electrostatics are treated explicitly in the latter. This enables the investigation of environmental factors, such as salt concentration, and the simulation of LCST in LLPS. The current implementation of the PLUM model is lacking these interactions. However, they could be added to the PLUM model using a Debeye-Hückel potential for future studies.

The PLUM-ELM extension is based on a paper of Globisch et al [43]. They applied the iterative scheme to a protein employing the MARTINI model. Lower uniform spring constants are needed for the PLUM-ELM model than for the MARTINI model to initially stabilize the structure. The reason for that might be that MARTINI model, in contrast to the PLUM model, is not able stabilize secondary structure on its own in its current implementation. Hence, stabilizing this structure via an elastic network model is a prerequisite in order to model stable proteins with the MARTINI model. In the PLUM-ELM model, on the other hand, the high resolution of the backbone allows secondary structure formation and the additional springs are a supportive factor that ensure stabilization, even in untested environments.

To interpret the results presented in this thesis, it is important to consider the limitations of the particular observables considered. Regarding the RMSD, a flexible tail or structural differences in one loop already cause a high RMSD although the two structures might be almost identical otherwise [81]. Another problem of the RMSD is that many configurations can be found at the same distance of the given reference structure, although they are not the same. To further investigate what structural changes are indicated by particular RMSD values, one could use a weighted RMSD or calculate the all-to-all RMSD—this means that the RMSD of each time frame is taken with respect to all others in the trajectory [82]. Another possibility to compare the structural similarity would be calculating the RMSD of the dihedral angles [81].

4.3 Umbrella sampling

The results presented here for US underline that one needs to carefully assess multiple observables of each simulation to ensure sufficient sampling in each window. To ensure that the binding affinity of mUb and UBQLN2:UBA was calculated accurately for the PLUM-ELM model, different observables, such as the RMSD and the running average of the COM, were calculated. Using WHAM as an estimator, even poor sampling might result in a continuous PMF profile which is not converged and where the molecules did not sample the desired distances along the reaction coordinates. There are different potential sources of error that have to be taken into consideration to determine the accuracy of the calculated binding affinity. While some of these errors, like an insufficient sampling time, can be avoided easily, others are harder to detect.

First, an unfortunate choice of reference points for umbrella distance restraints might be made, e.g. an input structure where the two molecules are not interacting as they would in reality. This would lead to configurations used for the US that do not represent the path of unbinding. It has been shown that this offset is avoided if the orientation of the pulled protein is restrained closely to the bound state [83]. This was accomplished in this Master thesis using the local frame of reference. Another source of error is the used estimator, such as WHAM, MBAR or umbrella integration. In order to quantify the offset caused by the used estimator, the results presented here could be compared to results for umbrella integration. Here, WHAM and MBAR were used as estimators. Normally, one would prefer to use MBAR over WHAM because a binning error is introduced in WHAM calculations [51]. However, due to long autocorrelation times of some windows it was not possible to obtain an uncorrelated subsample for MBAR calculations and hence, to benefit from the direct error estimation of MBAR. Therefore, the PMF and its error estimates were calculated via WHAM and the already implemented bootstrapping. A sufficiently small bin size for these WHAM calculations was ensured by comparing the results to the results of the MBAR calculations with the correlated sample.

Although the PMF presented here seems to be converged, changing only within the statistical error over 400 000 τ, the long autocorrelation times suggest that more sampling is needed to get the correct free energy differences. These long autocorrelation times might arise due to insufficient sampling of coordinates that are orthogonal to the reaction coordinate [84]. In the literature, it is suggested to sample approximately 50 times the integrated autocorrelation time [53]. Regarding the longest autocorrelation time 160 000 τ of the investigated, the sampling in this study corresponds to only ≈ 2.5 times the autocorrelation times. This makes it difficult to achieve the suggested simulation length within reasonable usage of computational resources. Hence, the presented results for the binding affinity have to be interpreted with caution as they are only an approximation for the real free energies of the model.

If the histograms were not converged, the integrated autocorrelation time might be still underestimated because transitions might have not have occurred during the simulations [49], [84]. Taken together with the RMSD results for re-running different frames, this suggests that it might be more efficient and accurate to run multiple, shorter simulations for each umbrella window instead of running one long simulation for each window. These simulations should be performed starting from different input structure for each distance with different random seeds. Bootstrapping complete histograms from these simulations has been shown to be give more accurate estimates of the errors than bootstrapping complete trajectories if the phase space is not completely sampled [49].

Regarding the structural observables of the umbrella windows, the RMSD of the single chains is not recovered. This might be due to the additional restraints of the Euler angles. Here, single beads were used as a local reference frame. Hence, the orientational restraint is only applied to these beads. One way to ensure an even distribution of the additional forces among several beads is to use virtual sites [54]. This method allows to group together several particles to one virtual bead and avoid an impact of the restraints on the overall structure of the proteins.

In principle, it is possible to add a variety of restraints to help convergence and reduce the simulation time needed for US. As these restraints are, like coarse-graining, removing certain degrees of freedoms of the system, they have to be chosen carefully and their contribution has to be accounted for in the final calculations of the binding affinity. However, it is not always trivial to identify observables that might be hindering convergence and restrain them. This is especially the case if convergence impeded by specific interactions between pairs of residues [85]. To solve this problem, enhanced sampling techniques, such as replica-exchange US, can be employed [86], [54], [87]. In this approach the biasing potential of adjacent windows is exchanged. This exchange is accepted or not based on a Metropolis criterion. The aim of replica-exchange US is to make free energy calculations independent from the starting conformation of each window.

In addition to general sources of error, there are inaccuracies of the PLUM model in representing the interactions between UBQLN2 and mUb. The binding affinity of two molecules and the stability of a bound state are influenced by non-covalent interactions, such as van der Waals forces, hydrogen bonds and electrostatic interactions. One potential issue is again the lack of explicit electrostatics in the PLUM model, although these interactions are partially accounted for within the Miyazawa-Jernigan matrix [34]. Another source of inaccuracy might be the implicit representation of solvent in the model. Water-mediated hydrogen bonds can influence the binding of proteins [88]. As an implicit solvent model, the PLUM model treats water-mediated hydrogen bonds only implicitly through the backbone hydrogen bonds and other effective interactions in the model.

US is not the only possibility to calculate the binding affinity of two proteins. Another PMF-based method is adaptive biasing force (ABF) [89]. The ABF algorithm incorporates the average force acting along the reaction coordinate in the equations of motion. Therefore, the free energy barrier between the bound and unbound state is flattened adaptively [90]. As compared to US, ABF would allow insights into the dissociation pathway in future studies.

In addition, there are also methods for free energy calculations that do not rely on the extraction of the PMF, but on alchemical transformation. In these approaches the interactions of the ligand with the binding partner and/or solvent are switched off progessively using unphysical intermediate states. One of these methods is free energy perturbation (FEP) [91]. The greatest drawback of alchemical transformation methods is that their convergence is currently limited to small ligands [85]. As extensive sampling is needed, there are only few studies that tried to calculate the binding affinity of protein-protein systems [92], [93]. On the other hand, alchemical perturbation methods, such as FEP, are favored over PMF-based methods if the binding site is not easily accessible and there is no simple association path [94]. However, the results of Zhang et al. suggest a specific binding of UBQLN2:UBA to mUb through hydrophobic patches on the surface of both proteins [11]. Taken together the specific binding of mUb and UBQLN2:UBA and the size of the studied proteins, using US to determine the binding affinity between mUb and UBQLN2:UBA, was a valid choice.

Despite these limitations of the PLUM-ELM model, the results for the binding affinity of UBQLN2:UBA and mUb presented in this Master thesis are still within the order of magnitude experimental results [10], [11]. The experimental results suggest a binding affinity of 31.5 ± 0.1 kJ/mol. This is about 46 kJ/mol lower than the binding affinity that was determined in this Master thesis. Potential reasons for this difference were discussed above. The binding affinity calculated here and the contribution of the additional restraints to it were determined from not converged simulations and hence, might differ if the simulations are elongated until convergence. However, so far only small changes in the PMF were observed upon incorporating additional simulations for certain windows. Compared to the accuracy of another implicit solvent CG model that was employed for free energy calculations, the difference of the PLUM-ELM model to the experimental results is smaller [54]. The estimated binding affinity of the CG model used by Woo et al. was ≈ 300 kJ/mol larger than the experimentally determined binding affinity for their system. In contrast, other studies suggested a binding affinity can be calculated in good accordance with experimental results for other studied systems when AA models for US are used [54], [95]. The error of this calculated binding affinity is on the order of ≈ 4 kJ/mol. With this in mind, the results for the umbrella sampling indicate that the PLUM-ELM model provides a reasonable representation of the interactions between mUb and UBQLN2:UBA, and can be further employed to investigate the aggregation of UBQLN2.

4.4 Intermolecular interactions between multiple chains

Regarding the intermolecular interactions between multiple chains, simulations over a simulation time of 1 000 000 τ were run. For these simulations different domains of UBQLN2 with and without mUb were placed randomly in a simulation and their interactions were observed over time. A full analysis of these simulations was unfortunately out of scope of this Master thesis. This analysis would have included inter- and intra-molecular observables. As intra-molecular observables the change in RMSD, RMSF and R_g as compared to single chain simulations would have been analyzed. The solvent accessible surface area would have been calculated as an additional intra-molecular observable. Regarding the interactions among different molecules, the radial distribution function of COM of single molecules would have been computed and the formation of clusters over time would have been observed. Here, only the results for the contact map were presented. This contact map gives an initial insight of how different UBQLN2:UBA domains interact with each other in the PLUM-ELM. The influence of mUb on the contact map of UBQLN2, would have been of special interest because mUb interrupts the LLPS of UBQLN2 experimentally. For the same reason it would have been interesting to obtain further information on the interaction of the UBA and the STI1-II domain of UQBLN2.

Given the slow dynamics of larger systems, it was not intended to reach a fully equilibrated system in the simulations of aggregation. The simulations should be rather seen as a starting point for further investigation. The slow dynamics make it difficult to apply this model to simulations of aggregation without using enhanced sampling techniques. One of these techniques that might help sampling is Monte Carlo simulations. A Monte Carlo simulation consists of a series of random steps in the phase space modifying degrees of freedoms of the system. These steps are accepted or rejected with a probability that is based on the change in energy. In contrast to MD simulations, Monte Carlo simulations do not investigate the time evolution of a system, they produce a set of possible configurations in the equilibrium ensemble. This procedure facilitates crossing free energy barriers because energetically unfavorable conformations might be still accepted. These regions might not be visited in MD simulations, as the system might get trapped in local energy minima. Furthermore, other studies suggest that simulations should be elongated to larger time scales than 1 000 000 τ to see proper droplet formation [18]. This is not possible with the current implementation of the PLUM model and a reasonable usage of computational resources. Hence, using Monte Carlo simulations might be one way to better sampling of the phase space and observe LLPS in addition to improvements of the PLUM implementation as they already have been employed to study ensembles of IDPs before [96].

The overall aim of the simulations of aggregation would have been to derive a full phase diagram, as Dignon et al. did in their study employing the slab method [18]. In this approach, a rectangular simulation box that is equal in length for the x- and y-direction, but elongated in the z-direction is used. The high-density (concentrated) phase with surfaces normal to the z-axis is modeled in equilibrium with the low-density (dilute) phase. This allows the determination of the equilibrium concentration of proteins in each phase and consequently, a phase-diagram if several simulations at different temperatures are conducted.

The idea was to perform this type of simulations for six different setups. First, there would have been one simulation of the UBA domain solely. In two subsequently simulations, the PXX and STI1-II domain would have been added, respectively. Then, all of these three setups would have been repeated with additional mUb molecules. Deriving a phase diagram from each of the simulations, would have allowed to further investigate the impact of mUb and the disordered domains of UBQLN2 on the LLPS of UBQLN2 and the interactions between the two molecules. Experimental results suggest that only the full C-terminus of UBQLN2 is phase-separating. This LLPS is disrupted by mUb via binding to the UBA domain [10]. Unfortunately, conducting all of the needed simulations was out of scope of this Master thesis.

5 Conclusions and outlook

This Master thesis employed a multiscale approach to investigate the role of mUb in the aggregation of UBQLN2. Both of these proteins have been found to play a role in the pathological mechanism of ALS [10], [8]. In order to enable accurate modeling of ordered and disordered domains of UBQLN2, a physics-based CG model, the PLUM model, was extended with harmonic interactions to stabilize structured domains that are known to remain folded under significant environmental perturbations. This extension, the PLUM-ELM model, was tested on two different proteins, mUb and UBQLN2. The results of this Master thesis underline the capability of the PLUM-ELM model to model highly conserved proteins as well as molecules with a higher structural flexibility. As compared to the PLUM model, the PLUM-ELM model improves the stabilization of globular domains. Additionally, the PLUM-ELM model can be easily transferred to other molecules of interest with short, additional simulations. These simulations do not require a lot of computational resources. The most computing power is needed for the AA simulation of the system of interest. These systems might include other proteins associated with the ubiquitin-proteasome system that are also found in inclusions of ALS patients, e.g. p62.

Employing the PLUM-ELM model, the binding affinity between mUb and UBQLN2:UBA was determined. To obtain the difference in the PMF between the bound and the unbound state, umbrella sampling was used and additional restraints were employed to facilitate convergence. It was found that the PLUM-ELM model is able to give an estimate within the order of magnitude of experimental results. This implies that mUb and UBQLN2:UBA are interacting reasonably within the PLUM-ELM model despite the limitations of the PLUM model including the implicit representation of solvation, coarse description of amino acid side chains, and lack of explicit electrostatic interactions. Regarding the long autocorrelation times of the US simulations, the results presented here are an initial estimate of the binding affinity. Using replica-exchange US, is one approach for facilitating convergence with reasonable computational requirements in future studies [85]. In addition, using virtual sites instead of single beads as local frame of reference might further decrease the autocorrelation time of simulations. Moreover, it would be possible to add other additional restraints, e.g. restraining the RMSD or the relative position using spherical coordinates. Roux et al. could show that each of their additional restraints significantly helps convergence [85], [94], [54].

In addition, preliminary simulations of the aggregation of UBQLN2 were performed. These simulations identified limitations of the PLUM-ELM model and its implementation for larger systems and longer time scales. This suggests the implementation of PLUM on a faster platform, such as LAMMPS. In addition, using Monte Carlo instead of MD simulation, might enable the sampling needed to see the formation of droplets. By performing such simulations at different temperatures, a full phase diagram for the aggregation of UBQLN2 could be characterized [18]. Testing the impact of different domains of UBQLN2 on the phase diagram, could complement the experimental results of Dao et al. [10]. Additionally, the PLUM-ELM modle could applied to study mutations, which have been demonstrated experimentally to impact the aggregation of UBQLN2 [13].

Other improvements that can be made include the incorporation of electrostatics in the PLUM-ELM mode. This would allow the investigation of the impact of environmental factors, such as ionic concentration, on the LLPS of UBQLN2 and other proteins. In the current implementation of the PLUM model side chain interactions implicitly incorporate average electrostatic effects, but they have no explicit dependence on the solution conditions. Using an explicit Debye-Hueckel iteraction, would allow to model changes in the concentration.

In conclusion, the results presented in this Master thesis are a good starting point for further research. They open prospects in several directions: (i) further investigation of the capabilities and limits of the PLUM-(ELM) model, (ii) analysis of the impact of partially disordered domains, mUb interactions and mutations on the LLPS of UBQLN2, (iii) transfer of the PLUM-ELM model to other proteins. This might lead to further insights into the driving forces of LLPS and the mechanisms that lead to degenerative diseases, e.g. ALS.

References

[1] V. N. Uversky, Introduction to intrinsically disordered proteins (idps), Chemical Reviews 114 (2014) 6557–6560. doi:10.1021/cr500288y.

[2] H. J. Dyson, P. E. Wright, Intrinsically unstructured proteins and their functions, Nature Reviews Molecular Cell Biology 6 (2005) 197–208. doi:10.1038/nrm1589.

[3] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, A. K. Dunker, Sequence complexity of disordered protein, Proteins: Structure, Function and Bioinformatics 42(2000) . doi:https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3.

[4] A. A. Hyman, C. A. Weber, F. Jülicher, Liquid-liquid phase separation in biology, Annual review of cell and developmental biology 30 (2014) 39–58. doi:10.1146/annurev-cellbio-100913-013325.

[5] S. Alberti, G. Amy, M. Tanja, Considerations and challenges in studying liquid-liquid phase separationand biomolecular condensates, CellPress 176 (2019) 419–434. doi:10.1016/j.cell.2018.12.035.

[6] S. Alberti, D. Dormann, Liquid–liquid phase separation in disease, Annual Review of Genetics 53 (2019) 171–194. doi:10.1146/annurev-genet-112618-043527.

[7] N. B. Nedelsky, J. P. Taylor, Bridging biophysics and neurology: aberrant phase transitions in neurodegenerative disease, Nature Reviews Neurology 15(2019) . doi:10.1038/s41582-019-0157-5.

[8] H.-X. Deng, W. Chen, S.-T. Hong, K. M. Boycott, G. H. Gorrie, N. Siddique, Y. Yang, F. Fecto, Y. Shi, H. Zhai, H. Jiang, M. Hirano, E. Rampersaud, G. H. Jansen, S. Donkervoort, E. H. Bigio, B. R. Brooks, K. Ajroud, R. L. Sufit, J. L. Haines, E. Mugnaini, M. A. Pericak-Vance, T. Siddique, Mutations in ubqln2 cause dominant x-linked juvenile and adult-onset als and als/dementia, Nature 477(2011) . doi:10.1038/nature10353.

[9] R. Beal, Q. Deveraux, G. Xia, M. Rechsteiner, C. Pickart, Surface hydrophobic residues of multiubiquitin chains essential for proteolytic targeting., Proceedings of the National Academy of Sciences 93(1996) . doi:10.1073/pnas.93.2.861.

[10] T. P. Dao, R. M. Kolaitis, H. J. Kim, K. O'Donovan, B. Martyniak, E. Colicino, H. Hehnly, J. P. Taylor, C. A. Castañeda, Ubiquitin modulates liquid-liquid phase separation of ubqln2 via disruption of multivalent interactions, Molecular Cell 69 (2018) 965–978.e6. doi:10.1016/j.molcel.2018.02.004.

[11] D. Zhang, S. Raasi, D. Fushman, Affinity makes the difference: non-selective interaction of the ubadomain of ubiquilin-1 with monomeric ubiquitin and polyubiquitinchains, Journal of Molecular Biology 377 (2008) 162–180. doi:10.1038/jid.2014.371.

[12] L. Dellefave, Amyotrophic lateral sclerosis overview clinical manifestations of als, Amyotrophic Lateral Sclerosis (2007) 1–26.

[13] Y. Yang, H. B. Jones, T. P. Dao, C. A. Castañeda, Single amino acid substitutions in stickers, but not spacers, substantially alter ubqln2 phase transitions and dense phase material properties, Journal of Physical Chemistry B 123 (2019) 3618–3629. doi:10.1021/acs.jpcb.9b01024.

[14] Q. Zhang, C. Weber, U. S. Schubert, R. Hoogenboom, Thermoresponsive polymers with lower critical solution temperature: from fundamental aspects and measuring techniques to recommended turbidimetry conditions, Materials Horizons 4(2017) . doi:10.1039/C7MH00016B.

[15] E. W. Martin, T. Mittag, Relationship of sequence and phase separation in protein low-complexity regions, Biochemistry 57(2018) . doi:10.1021/acs.biochem.8b00008.

[16] G. L. Dignon, R. B. Best, J. Mittal, Biomolecular phase separation: From molecular driving forces to macroscopic properties, Annual Review of Physical Chemistry 71 (2020) 53–75. doi:10.1146/annurev-physchem-071819-113553.

[17] M. Hofweber, D. Dormann, Friend or foe—post-translational modifications as regulatorsof phase separation and rnp granule dynamics, Journal of Biological Chemistry 294 (2017) 7137–7150. doi:10.1074/jbc.TM118.001189.

[18] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, J. Mittal, Sequence determinants of protein phase behavior from a coarse-grained model, PLoS Computational Biology 14(2018) . doi:10.1371/journal.pcbi.1005941.

[19] R. B. Best, Computational and theoretical advances in studies of intrinsically disordered proteins, Current Opinion in Structural Biology 42 (2017) 147–154. doi:10.1016/j.sbi.2017.01.006.

[20] C. Lorenz, N. L. Doltsinis, Molecular dynamics simulation: From “ab initio” to “coarse grained” (2012). doi:10.1007/978-94-007-0711-5_7.

[21] G. L. Dignon, W. Zheng, J. Mittal, Simulation methods for liquid–liquid phase separation of disordered proteins, Current Opinion in Chemical Engineering 23 (2019) 92–98. doi:10.1016/j.coche.2019.03.004.

[22] Z. A. Levine, J. E. Shea, Simulations of disordered proteins and systems with conformational heterogeneity, Current Opinion in Structural Biology 43 (2017) 95–103. doi:10.1016/j.sbi.2016.11.006.

[23] J. Song, S. C. Ng, P. Tompa, K. A. W. Lee, H. S. Chan, Polycation-π interactions are a driving force for molecular recognition by an intrinsically disordered oncoprotein family, PLoS Computational Biology 9(2013) . doi:10.1371/journal.pcbi.1003239.

[24] K. Kasahara, H. Terazawa, T. Takahashi, J. Higo, Studies on molecular dynamics of intrinsically disordered proteins and their fuzzy complexes: A mini-review, Computational and Structural Biotechnology Journal 17 (2019) 712–720. doi: 10.1016/j.csbj.2019.06.009.

[25] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The i-tasser suite: Protein structure and function prediction, Nature Methods 12 (2014) 7–8. doi:10.1038/nmeth.3213.

[26] R. Hockney, S. Goel, J. Eastwood, Quiet high-resolution computer models of a plasma, Journal of Computational Physics 14(1974) . doi:10.1016/0021-9991(74) 90010-2.

[27] A. Friedman, Brownian dynamics simulations of colloidal dispersion (1992). doi: 10.1007/978-1-4615-7405-7_15.

[28] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, Molecular dynamics with coupling to an external bath, The Journal of Chemical Physics 81(1984) . doi:10.1063/1.448118.

[29] M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, Journal of Applied Physics 52(1981) . doi:10.1063/1.328693.

[30] S. Riniker, J. R. Allison, W. F. van Gunsteren, On developing coarse-grained models for biomolecular simulation: a review, Physical Chemistry Chemical Physics 14(2012) . doi:10.1039/c2cp40934h.

[31] T. Bereau, M. Deserno, Generic coarse-grained model for protein folding and aggregation, Journal of Chemical Physics 130(2009) . doi:10.1063/1.3152842.

[32] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The martini force field: coarse grained model for biomolecular simulations, The Journal of Physical Chemistry B 111(2007) . doi:10.1021/jp071097f.

[33] G. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, Journal of Molecular Biology 7(1963) . doi:10.1016/S0022-2836(63)80023-6.

[34] S. Miyazawa, R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, Macromolecules 18(1985) . doi:10.1021/ma00145a039.

[35] K. L. Osborne, M. Bachmann, B. Strodel, Thermodynamic analysis of structural transitions during gnnqqny aggregation, Proteins: Structure, Function and Bioinformatics 81 (2013) 1141–1155. doi:10.1002/prot.24263.

[36] T. Bereau, M. Bachmann, M. Deserno, Interplay between secondary and tertiary structure formation in protein folding cooperativity, Journal of the American Chemical Society 132 (2010) 13129–13131. doi:10.1021/ja105206w.

[37] T. Bereau, M. Deserno, M. Bachmann, Structural basis of folding cooperativity in model proteins: Insights from a microcanonical perspective, Biophysical Journal 100 (2011) 2764–2772. doi:10.1016/j.bpj.2011.03.056.

[38] G. O. Rutter, A. H. Brown, D. Quigley, T. R. Walsh, M. P. Allen, Testing the transferability of a coarse-grained model to intrinsically disordered proteins, Physical Chemistry Chemical Physics 17 (2015) 31741–31749. doi:10.1039/c5cp05652g.

[39] R. B. Best, G. Hummer, Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides, Journal of Physical Chemistry 113 (2009) 9004–9015. doi:10.1021/jp901540t.

[40] D. Caballero, J. Määttä, A. Q. Zhou, M. Sammalkorpi, C. S. O'Hern, L. Regan, Intrinsic α-helical and β-sheet conformational preferences: A computational case study of alanine, Protein Science 23(2014) . doi:10.1002/pro.2481.

[41] M. M. Tirion, Large amplitude elastic motions in proteins from a single-parameter, atomic analysis, Physical Review Letters(1996) doi:https://doi.org/10.1103/PhysRevLett.77.1905.

[42] J. K. Noel, P. C. Whitford, J. N. Onuchic, The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function, Journal of Physical Chemistry B 116 (2012) 8692–8702. doi:10.1021/jp300852d.

[43] C. Globisch, V. Krishnamani, M. Deserno, C. Peter, Optimization of an elastic network augmented coarse grained model to study ccmv capsid deformation, PLoS ONE 8(2013) . doi:10.1371/journal.pone.0060582.

[44] D. Shoup, A. Szabo, Role of diffusion in ligand binding to macromolecules and cell-bound receptors, Biophysical Journal 40(1982) . doi:10.1016/S0006-3495(82)84455-X.

[45] O. G. Berg, P. H. von Hippel, Diffusion-controlled macromolecular interactions, Annual Review of Biophysics and Biophysical Chemistry 14(1985) . doi:10.1146/annurev.bb.14.060185.001023.

[46] G. Schreiber, Kinetic studies of protein–protein interactions, Current Opinion in Structural Biology 12(2002) . doi:10.1016/S0959-440X(02)00287-7.

[47] G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling, Journal of Computational Physics 23 (1977) 187–199. doi:https://doi.org/10.1016/0021-9991(77)90121-8.

[48] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, P. A. Kollman, The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method, Journal of Computational Chemistry(1992) doi:https://doi.org/10.1002/jcc.540130812.

[49] J. S. Hub, B. L. D. Groot, D. V. D. Spoel, G-whams-a free weighted histogram analysis implementation including robust error and autocorrelation estimates, Journal of Chemical Theory and Computation 6 (2010) 3713–3720. doi:10.1021/ct100494z.

[50] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Statistical Science 1 (1986) 54–75.

[51] M. R. Shirts, J. D. Chodera, Statistically optimal analysis of samples from multiple equilibrium states, The Journal of Chemical Physics 129(2008) . doi:10.1063/1.2978177.

[52] Z. Tan, E. Gallicchio, M. Lapelosa, R. M. Levy, Theory of binless multi-state free energy estimation with applications to protein-ligand binding, The Journal of Chemical Physics 136(2012) . doi:10.1063/1.3701175.

[53] M. R. Shirts, D. L. Mobley, An introduction to best practices in free energy calculations (2013). doi:10.1007/978-1-62703-017-5_11.

[54] H. Woo, B. Roux, Calculation of absolute protein – ligand binding free, Proceedings of the National Academy of Sciences 102 (2005) 6825–6830. doi:https://doi.org/10.1073/pnas.0409005102.

[55] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, Journal of Molecular Biology 215(1990) . doi:10.1016/S0022-2836(05)80360-2.

[56] S. LLC, The pymol molecular graphics system, version 1.8, pyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC. (11 2015).

[57] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindah, Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX 1-2 (2015) 19–25. doi:10.1016/j.softx.2015.06.001.

[58] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, Improved side-chain torsion potentials for the amber ff99sb protein force field, Proteins: Structure, Function, and Bioinformatics 78(2010) . doi:10.1002/prot.22711.

[59] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water, The Journal of Chemical Physics 79(1983) . doi:10.1063/1.445869.

[60] B. Hess, H. Bekker, H. J. C. Berendsen, J. G. E. M. Fraaije, Lincs: A linear constraint solver for molecular simulations, Journal of Computational Chemistry 18(1997) . doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[61] B. Hess, C. Kutzner, D. V. D. Spoel, E. Lindahl, Grgmacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation, Journal of Chemical Theory and Computation 4 (2008) 435–447. doi:10.1021/ct700301q.

[62] J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, P. C. Whitford, Smog 2: A versatile software package for generating structure-based models, PLoS Computational Biology 12(2016) . doi:10.1371/journal.pcbi.1004794.

[63] R. E. Beal, D. Toscano-Cantaffa, P. Young, M. Rechsteiner, C. M. Pickart, The hydrophobic effect contributes to polyubiquitin chain recognition, Biochemistry 37(1998) . doi:10.1021/bi972514p.

[64] R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L. P. Wang, T. J. Lane, V. S. Pande, Mdtraj: A modern open library for the analysis of molecular dynamics trajectories, Biophysical Journal 109 (2015) 1528–1532. doi:10.1016/j.bpj.2015.08.015.

[65] A. C. Fogarty, R. Potestio, K. Kremer, Adaptive resolution simulation of a biomolecule and its hydration shell: Structural and dynamical properties, Journal of Chemical Physics 142(2015) . doi:10.1063/1.4921347.

[66] N. J. Marianayagam, S. E. Jackson, Native-state dynamics of the ubiquitin family: Implications for function and evolution, Journal of the Royal Society Interface 2 (2005) 47–54. doi:10.1098/rsif.2004.0025.

[67] S. Vijay-Kumar, C. E. Bugg, W. J. Cook, Structure of ubiquitin refined at 1.8 Å resolution, Journal of Molecular Biology 194(1987) . doi:10.1016/0022-2836(87)90679-6.

[68] C. L. Evans, J. E. Long, T. R. Gallagher, J. D. Hirst, M. S. Searle, Conformation and dynamics of the three-helix bundle uba domain of p62 from experiment and simulation, Proteins: Structure, Function and Genetics 71 (2008) 227–240. doi:10.1002/prot.21692.

[69] A. L. Teixeira, N. A. Alves, The high stability of the three-helix bundle uba domain of p62 protein as revealed by molecular dynamics simulations, Journal of Molecular Modeling 27(2021) . doi:10.1007/s00894-021-04698-0.

[70] G. L. Dignon, W. Zheng, Y. C. Kim, J. Mittal, Temperature-controlled liquid–liquid phase separation of disordered proteins, ACS Central Science(2019) doi:10.1021/acscentsci.9b00102.

[71] A. Vitalis, R. V. Pappu, Absinth: A new continuum solvation model for simulations of polypeptides in aqueous solutions, Journal of Computational Chemistry 30(2009) . doi:10.1002/jcc.21005.

[72] T. Dannenhoffer-Lafage, R. B. Best, A data-driven hydrophobicity scale for predicting liquid–liquid phase separation of proteins, The Journal of Physical Chemistry B 125(2021) . doi:10.1021/acs.jpcb.0c11479.

[73] W. G. Noid, Perspective: Coarse-grained models for biomolecular systems, The Journal of Chemical Physics 139(2013) . doi:10.1063/1.4818908.

[74] P. Wilhelm, B. Lewandowski, N. Trapp, H. Wennemers, A crystal structure of an oligoproline ppii-helix, at last, Journal of the American Chemical Society 136(2014) . doi:10.1021/ja507405j.

[75] S. Plimpton, Fast parallel algorithms for short-range molecular dynamics, Journal of Computational Physics 117 (1995) 1–19. doi:https://doi.org/10.1006/jcph.1995.1039.

[76] M. Cieplak, T. X. Hoang, Universality classes in folding times of proteins, Biophysical Journal 84(2003) . doi:10.1016/S0006-3495(03)74867-X.

[77] K. Wołek, Àngel Gómez-Sicilia, M. Cieplak, Determination of contact maps in proteins: A combination of structural and chemical approaches, The Journal of Chemical Physics 143(2015) . doi:10.1063/1.4929599.

[78] A. B. Poma, M. Cieplak, P. E. Theodorakis, Combining the martini and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins, Journal of Chemical Theory and Computation 13(2017) . doi:10.1021/acs.jctc.6b00986.

[79] X. Periole, M. Cavalli, S.-J. Marrink, M. A. Ceruso, Combining an elastic network with a coarse-grained molecular force field: Structure, dynamics, and intermolecular recognition, Journal of Chemical Theory and Computation 5(2009) . doi:10.1021/ct9002114.

[80] Y. C. Kim, G. Hummer, Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding, Journal of Molecular Biology 375(2008) . doi:10.1016/j.jmb.2007.11.063.

[81] I. Kufareva, R. Abagyan, Methods of protein structure comparison (2011). doi:10.1007/978-1-61779-588-6_10.

[82] A. Grossfield, P. N. Patrone, D. R. Roe, A. J. Schultz, D. Siderius, D. M. Zuckerman, Best practices for quantification of uncertainty and sampling quality in molecular simulations [article v1.0], Living Journal of Computational Molecular Science 1(2019) . doi:10.33011/livecoms.1.1.5067.

[83] D. Markthaler, S. Jakobtorweihen, N. Hansen, Lessons learned from the calculation of one-dimensional potentials of mean force [article v1.0], Living Journal of Computational Molecular Science 1(2019) . doi:10.33011/livecoms.1.2.11073.

[84] F. Zhu, G. Hummer, Convergence and error estimation in free energy calculations using the weighted histogram analysis method, Journal of Computational Chemistry 33 (2012) 453–465. doi:10.1002/jcc.21989.

[85] J. C. Gumbart, B. Roux, C. Chipot, Standard binding free energies from computer simulations: What is the best strategy?, Journal of Chemical Theory and Computation 9 (2013) 794–802. doi:10.1021/ct3008099.

[86] Y. Sugita, A. Kitao, Y. Okamoto, Multidimensional replica-exchange method for free-energy calculations, The Journal of Chemical Physics 113(2000) . doi:10.1063/1.1308516.

[87] H. Oshima, S. Re, Y. Sugita, Replica-exchange umbrella sampling combined with gaussian accelerated molecular dynamics for free-energy calculation of biomolecules, Journal of Chemical Theory and Computation 15(2019) . doi:10.1021/acs.jctc.9b00761.

[88] P. L. Kastritis, A. M. J. J. Bonvin, On the binding affinity of macromolecular interactions: daring to ask why proteins interact, Journal of The Royal Society Interface 10(2013) . doi:10.1098/rsif.2012.0835.

[89] E. Darve, A. Pohorille, Calculating free energies using average force, The Journal of Chemical Physics 115(2001) . doi:10.1063/1.1410978.

[90] J. Comer, J. C. Gumbart, J. Hénin, T. Lelièvre, A. Pohorille, C. Chipot, The adaptive biasing force method: Everything you always wanted to know but were afraid to ask, The Journal of Physical Chemistry B 119(2015) . doi:10.1021/jp506633n.

[91] R. W. Zwanzig, High-temperature equation of state by a perturbation method. i. nonpolar gases, The Journal of Chemical Physics 22(1954) . doi:10.1063/1.1740409.

[92] A. J. Clark, T. Gindin, B. Zhang, L. Wang, R. Abel, C. S. Murret, F. Xu, A. Bao, N. J. Lu, T. Zhou, P. D. Kwong, L. Shapiro, B. Honig, R. A. Friesner, Free energy perturbation calculation of relative binding free energy between broadly neutralizing antibodies and the gp120 glycoprotein of hiv-1, Journal of Molecular Biology 429(2017) . doi:10.1016/j.jmb.2016.11.021.

[93] H. Park, Y. H. Jeon, Free energy perturbation approach for the rational engineering of the antibody for human hepatitis b virus, Journal of Molecular Graphics and Modelling 29(2011) . doi:10.1016/j.jmgm.2010.11.010.

[94] Y. Deng, B. Roux, Computations of standard binding free energies with molecular dynamics simulations, The Journal of Physical Chemistry B 113(2009) . doi:10.1021/jp807701h.

[95] M. S. Lee, M. A. Olson, Calculation of absolute protein-ligand binding affinity using path and endpoint approaches, Biophysical Journal 90(2006) . doi:10.1529/biophysj.105.071589.

[96] L. M. Pietrek, L. S. Stelzl, G. Hummer, Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations, Journal of Chemical Theory and Computation 16 (2020) 725–737. doi:10.1021/acs.jctc.9b00809.

Supplementary information

Stabilization of protein structure

The error in the RMSF was chosen as criterion to stop the RMSF. Table 3 shows the error for the RMSF for the iterative scheme of mUb and UBQLN2:UBA.

Iteration	mUb	UBQLN2:UBA
0	0.06998	0.13859
1	0.04838	0.16513
2	0.02967	0.05037
3	0.04160	0.13942
4	0.06144	0.04507
5	0.07526	0.02973
6	0.03999	0.08109
7	0.03462	0.06838
8	-	0.06485
9	-	0.05611
10	-	0.05280

Table 3: The error of the RMSF for the iterative scheme of mUb and UBQLN2:UBA. The error for the unrestrained simulation is 0.62182 and 0.22287 for mUb and UBQLN2:UBA, respectively

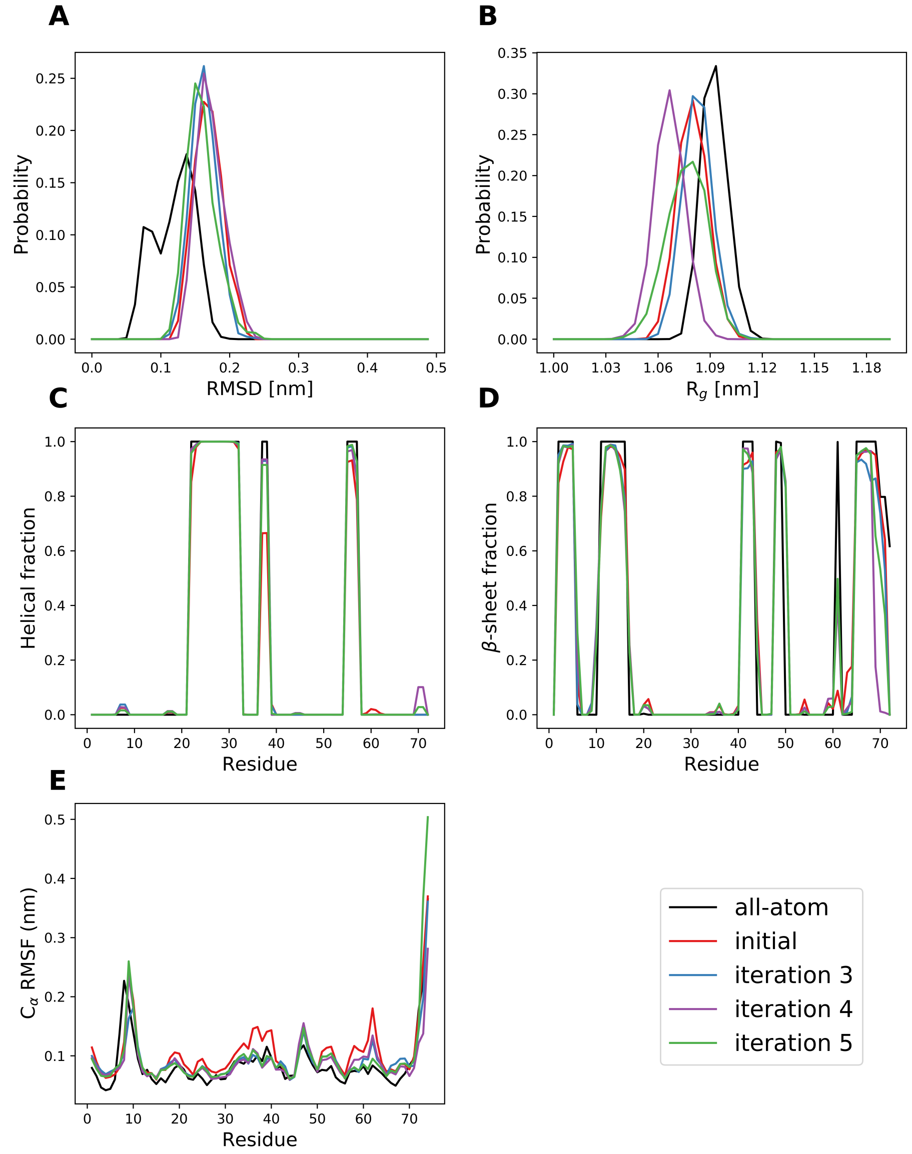

Figure 18: Results for chosen observables for an AA simulation of mUb (black), the chosen initial simulation (red) and iteration 3, 4, and 5 (blue, purple, green, respectively).

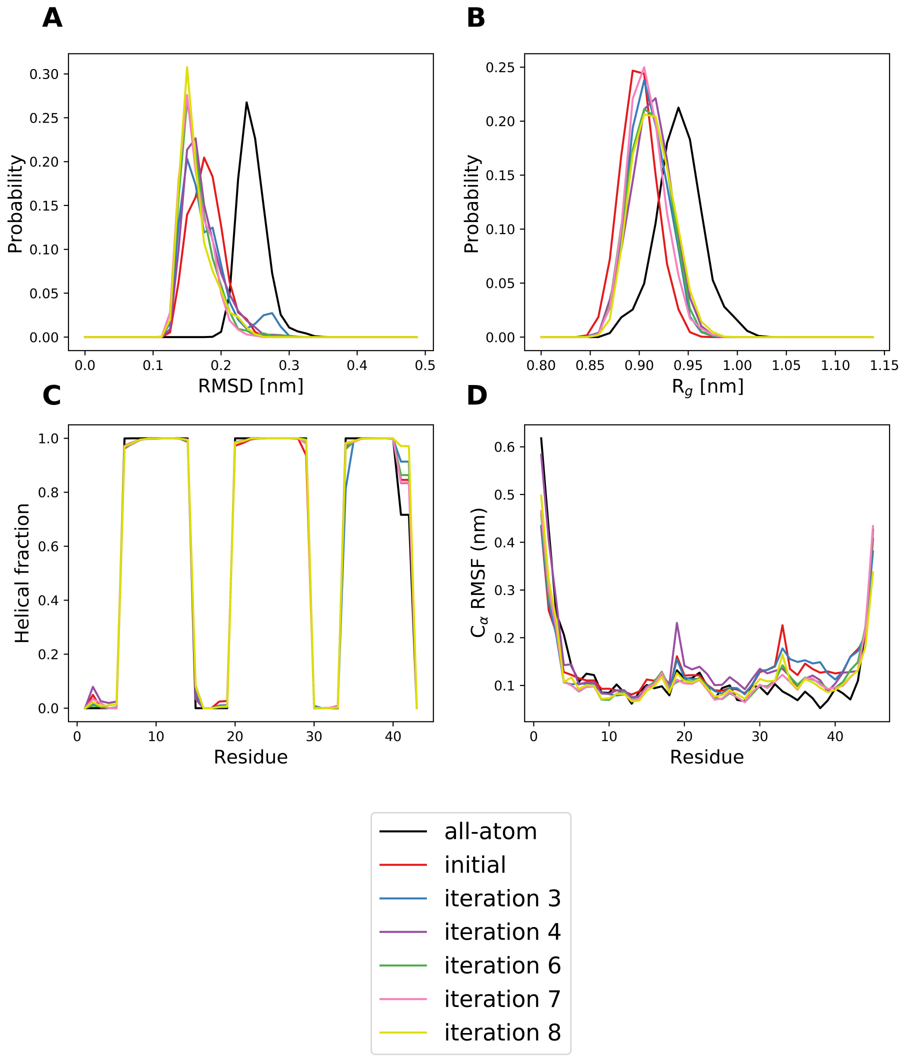

Figure 19: Results for chosen observables for an AA simulation of UBQLN2:UBA (black), the chosen initial simulation (red) and iteration 3, 4, 6, 7, 8 (blue, purple, green, pink, respectively).

Umbrella sampling

The following equations demonstrate the calculation of the angular integral that represents the contribution of the orientational restraints in the bulk:

$$\mathrm{e}^{-\beta G_o^{bulk}} = \frac{1}{8\pi^2}\int_0^{\pi} sin(\Theta)\, d\Theta \int_0^{2\pi} d\Phi \int_0^{2\pi} d\Psi \mathrm{e}^{-\beta u_o(\Theta,\Phi,\Psi} \tag{26a}$$

$$= \frac{1}{8\pi^2}\int_0^{\pi} sin(\Theta)\, d\Theta \mathrm{e}^{-\beta(0.5*418.4)*(\Theta-93.55211*\pi/180)^2} \tag{26b}$$

$$\times \frac{1}{8\pi^2}\int_0^{2\pi} sin(\Phi)\, d\Phi \mathrm{e}^{-\beta(0.5*418.4)*(\Phi-91.88719*\pi/180)^2} \tag{26c}$$

$$\times \frac{1}{8\pi^2}\int_0^{2\pi} sin(\Psi)\, d\Psi \mathrm{e}^{-\beta(0.5*418.4)*(\Psi-4.490039*\pi/180)^2} \tag{26d}$$

$$= \frac{1}{8\pi^2} \times 0.192594 \times 0.19354 \times 0.163529 = 0.00752 \tag{26e}$$

$$\rightarrow G_0^{bulk} = 12.1978 \text{ kJ/mol} \tag{26f}$$

The positional restraints $u_a(\theta, \phi)$ were not applied to the studied system. Therefore, the force constant k for this restraint was set to 0 for the calculation of S^*.

$$S^* = (r^*)^2 \int_0^{2\pi} sin(\theta)\, d\theta \int d\phi\ \mathrm{e}^{-\beta\ u_a(\theta,\phi)} \tag{27a}$$

$$= (r^*)^2 \int_0^{2\pi} sin(\theta)\, d\theta \mathrm{e}^{-\beta\ 0.5*k(\theta-\theta_0)^2} \tag{27b}$$

$$\times \int d\phi\ \mathrm{e}^{-\beta\ 0.5*k(\phi-\phi_0)^2} \tag{27c}$$

$$= (r^*)^2 \int_0^{2\pi} sin(\theta)\, d\theta \int d\phi \tag{27d}$$

$$= (r^*)^2 * 4\pi \tag{27e}$$

S^*I^* was calculated using three different reference points at large distance ($r_1^* = 32.48$ Å, $r_2^* = 34.71$ Å, $r_3^* = 36.56$ Å). This allowed to estimate to which extent the calculated binding affinity depends on the chosen r^*.

$$S_1^* = (r_1^*)^2 * 4\pi = (32.48\text{Å})^2 \times 4\pi = 1.5136 \times 10^4\ \text{Å}^2 \tag{28a}$$

$$S_1^* = (r_2^*)^2 * 4\pi = (34.71\text{Å})^2 \times 4\pi = 1.3260 \times 10^4\ \text{Å}^2 \tag{28b}$$

$$S_1^* = (r_3^*)^2 * 4\pi = (36.56\text{Å})^2 \times 4\pi = 1.6794 \times 10^4\ \text{Å}^2 \tag{28c}$$

$$I_1^* = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-W(r_1^*]} = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-W(32.48)]} = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-81.9821]} = 4.5932\times10^{14}\ \text{Å} \tag{29a}$$

$$I_2^* = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-W(r_2^*)]} = \int_{site} dr\ \mathrm{e}^{-\beta[(r)-W(34.71)]} = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-81.8985]} = 4.4419\times10^{14}\ \text{Å} \tag{29b}$$

$$I_3^* = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-W(r_3^*)]} = \int_{site} dr\ \mathrm{e}^{\beta[W(r)-W(36.56)]} = \int_{site} dr\ \mathrm{e}^{-\beta[W(r)-81.8844]} = 4.4417\times10^{14}\ \text{Å} \tag{29c}$$

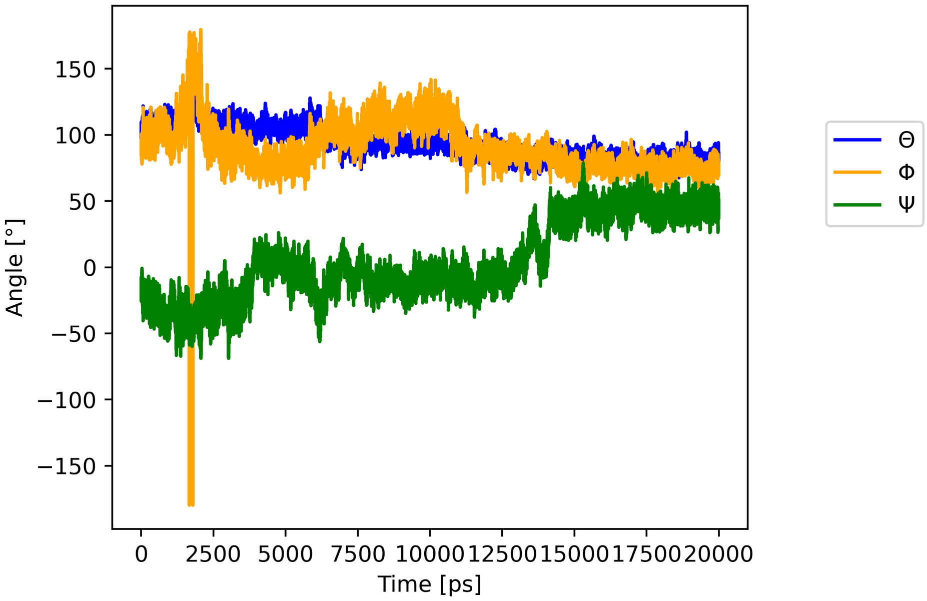

Figure 20: This figure shows the Euler angles over time in the bound state without any restraints.

Window	Average COM distance [nm]	Autocorrel. WHAM [ps]	Autocorrel. MBAR [ps]
1	2.064648 ± 0.024970	12542.1	1240.909415
2	2.081044 ± 0.027179	10349.7	1738.92418
3	2.109755 ± 0.028723	5413.79	2961.88233
4	2.144389 ± 0.028562	4561.47	1567.4221
5	2.201687 ± 0.031048	3621.23	1878.278815
6	2.246470 ± 0.030592	36.5211	1951.438370
7	2.295131 ± 0.029837	512.365	2031.699410
8	2.335651 ± 0.029661	744.982	1635.983905
9	2.382731 ± 0.031890	4966.67	1929.834835
10	2.430102 ± 0.034641	9135.88	29157.775965
11	2.499737 ± 0.036830	421.301	12344.838315
12	2.559072 ± 0.034248	9404.83	9451.734000
13	2.620265 ± 0.029406	8233.17	1046.009265
14	2.661709 ± 0.029624	3227.53	919.074140
15	2.709436 ± 0.029894	6519.61	634.131300
16	2.776913 ± 0.030555	7213.61	903.503780
17	2.811083 ± 0.031838	6022.42	2139.451640
18	2.867627 ± 0.032738	8538.29	2529.152440
19	2.932772 ± 0.032635	12157.8	1935.316330
20	2.973402 ± 0.033888	5205.36	2902.467210
21	3.036323 ± 0.032860	7946.13	1215.971010
22	3.095407 ± 0.032410	13666.4	912.491915
23	3.141835 ± 0.031848	11599.2	627.428580
24	3.192327 ± 0.032272	12368.6	506.308560
25	3.242855 ± 0.032085	22.964	523.585045
26	3.289494 ± 0.031472	12588.7	789.518485
27	3.346963 ± 0.031023	12576.9	586.507705
28	3.394042 ± 0.031068	14393.0	556.408110
29	3.438397 ± 0.030957	12840.8	558.255685
30	3.499614 ± 0.030875	19.8221	526.234000
31	3.550995 ± 0.030918	11445.8	551.539500
32	3.596527 ± 0.030544	15175.3	510.345690
33	3.649396 ± 0.030742	14941.3	527.721335
34	3.701554 ± 0.030748	15977.2	536.103095

Table 4: This table shows the average COM distance with its standard deviation for each window and the autocorrelation times for WHAM and MBAR.

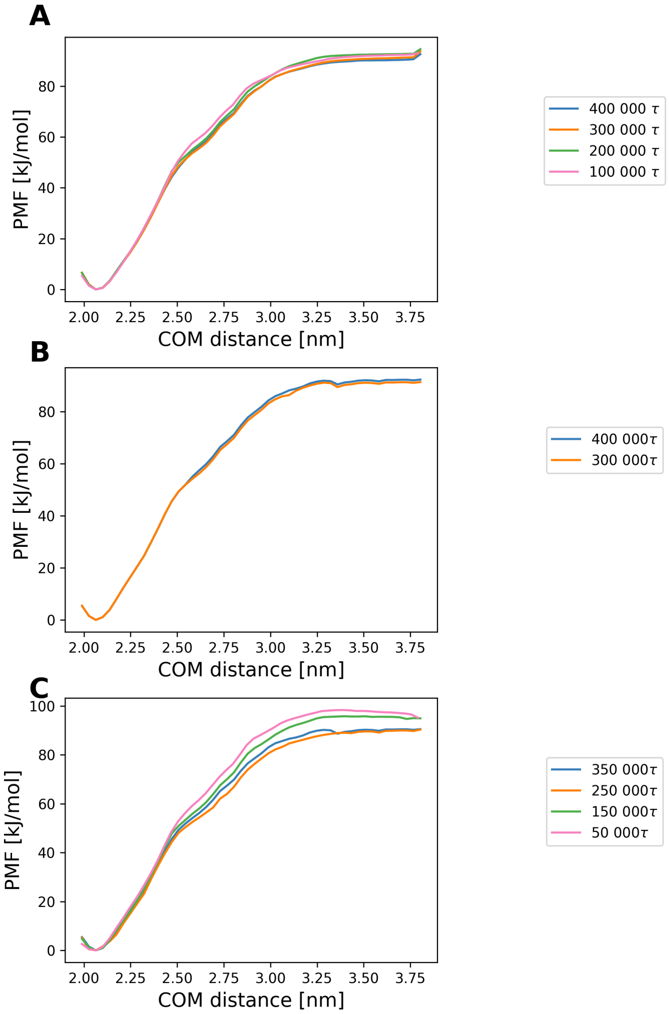

Figure 21: In **A** the PMF was calculated using MBAR for 400 000, 300 000, 200 000 and 100 000 τ. In **B** the WHAM calculations for 400 000 and 300 000 τ including the histograms of the windows that were re-run can be seen. **C** shows the resulting PMF when the first 50 000 τ of each part is removed as equilibration.

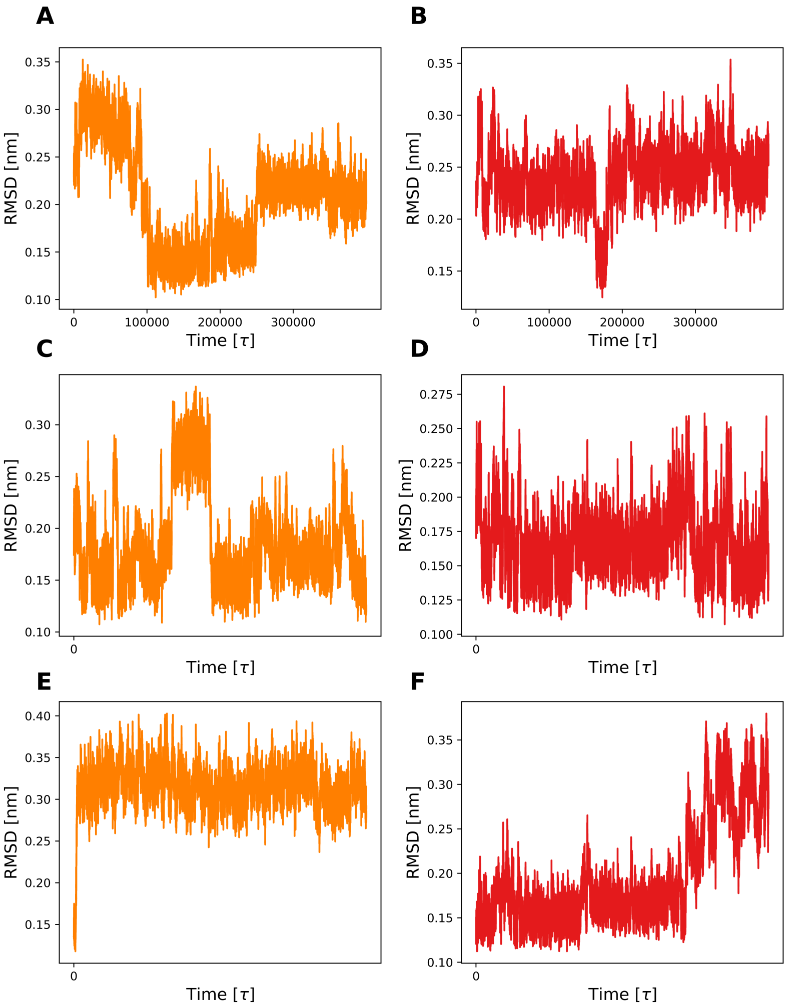

Figure 22: RMSD over time for the first (orange, left side) and the second (red, right side) run. The umbrella sampling was re-run for window 12 with an average COM distance of 2.56 nm (**A** + **B**), window 24 with an average COM distance of 3.19 nm (**C** + **D**) and window 25 with an average COM distance of 3.24 nm (**E** + **F**).

Declaration of independence

I hereby declare that I have written the present thesis independently and without use of other than the indicated means. I also declare that to the best of my knowledge all passages taken from published and unpublished sources have been referenced. The thesis has not been submitted for evaluation to any other examining authority nor has it been published in any form.

Mainz, June 16, 2021

Hanne Zillmer