FREIE UNIVERSITÄT BERLIN

A Molecular Dynamics Study of the Site-dependent Interaction of a Polyglutamine Fibril with an Attached Biotinylated Residue

by

Aniruddha Bapat

A thesis submitted in partial fulfillment for the
degree of Master's in Physics

in the
Department of Physics

June 2017

Declaration of Authorship

I, Aniruddha Bapat, declare that this thesis titled, 'A Molecular Dynamics Study of the Site-dependent Interaction of a Polyglutamine Fibril with an Attached Biotinylated Residue' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: 24/06/2017

Acknowledgements

I sincerely thank Dr. Markus Miettinen for his continued mentorship and stimulating discussions throughout this project. This project began as a quantum computing student's curiosity towards biophysics and ended as a Master's thesis project. If improbable reactions need enzymes to succeed, then this somewhat surprising journey certainly needed a Markus.

Scientifically, I had access many other resources, both at FU Berlin and outside. I thank Dr. Luca Monticelli for his invaluable help with force field parametrization, and for being a wonderful host in Lyon. I am also grateful to Jan Daldrop for sharing his initial structure files. Now, I can certainly imagine how painstaking it must have been to generate them to perfection.

Finally, I am grateful to Prof. Petra Imhof for her supervision over the Master's thesis, and I also thank all the funding and supporting agencies: Nanoscale group at FU Berlin, IBCP (Lyon) and the Physics department at FU Berlin.

My stay at Berlin has been memorable, if at times challenging in unexpected ways. In my time there, I made some close friends, and either directly or indirectly, and whether they know it or not, they have contributed to this thesis. I thank them with every inch of my heart. I cannot possibly end this section without mentioning one group which goes by many names, only one of which is the Core Tex Crew. I spent so much time with these people that I can't really explain why I am thanking them. But they were my family in Berlin, and I could not have made it without them. Carlos, Heather, Julio, and Claire, I love you and you know it!

Contents

List of Figures

List of Tables

Glossary

AMBER Assisted Model Building with Energy Refinement. 31, 37–40, 43, 49, 51, 53, 55, 58, 64–66

CG Coarse-Graining. 19, 20

CHARMM Chemistry at HARvard Macromolecular Mechanics. 31, 37–40, 43, 49, 51–53, 55, 65, 66

ELISA Enzyme-Linked Immunosorbent Assay. 10

GAFF General Atomic Force Field. 38

GLNp Glutamine-prime. vii, 33, 34, 36–39, 43–49, 51–55, 57–60, 62, 64, 66–68

HPLC High-Performance Liquid Chromatography. 12, 14, 15

LEaP Link, Edit and Parm. 37

MD Molecular Dynamics. 2, 18, 20, 21, 25, 39, 40, 43, 66

PBS Phosphate-Buffered Saline. 12

PMF Potential of Mean Force. 55–58, 66

polyQ Polyglutamine. vi, vii, 1–5, 7, 9–19, 31–33, 36, 40, 47, 48, 66, 67

QM/MM Hybrid Quantum Mechanics/Molecular Mechanics. 19

VMD Visual Molecular Dynamics. 39, 40, 44, 53

Chapter 1

Introduction

The global structure of proteins is sensitively dependent on the constituent peptides, their ordering, any substituted groups, the ambient chemistry as well as temperature and pH and other physical conditions. As a result, proteins often exhibit anomalous behaviours *in vivo*. Since a protein's structure is intricately linked to its function, misfolding may cause pathological functioning of the protein. Indeed, the pathology of many diseases has been ascribed to the misfolding of certain proteins.

Misfolded proteins, or parts thereof, are known to *aggregate*, i.e. associate with each other to form large, regular, insoluble structures which are thermodynamically stable. These structures are termed *amyloid fibrils*. In fact, aggregation and formation of amyloid fibrils is a feature that seems to be inherent to the nature of polypeptide chains [1, 8, 9]. The structure and aggregation process of amyloid fibrils is interesting to study for several reasons.

First, out of diseases related to protein misfolding, many are correlated with the formation of amyloid fibrils [10, 11]. One example that stands out is the CAG triplet disorders, in which proteins are found to have enlarged sections of consecutive glutamines (Q) in their primary structure. The severity of the diseases correlates positively with the length of the repeats. It is also found that the polyQs aggregate and form macroscopic cell inclusions, and it is believed that the aggregates, especially in their early stages, may be pathological. This has prompted both theoretical and experimental studies on the process of polyQ aggregation. As is especially apparent in the case of Alzheimer's disease, these disorders no longer rare (see e.g. [12]), and so the need to understand the process of aggregation is well-motivated.

Secondly, it has been suggested that amyloid fibrils may not always be pathological; on the contrary, they may play various functional roles [13–15]. Finally, the study of aggregation kinetics is theoretically well-motivated. Simple physical models often capture the essential features seen in aggregation, and many phenomenological models now exist to describe the various mechanisms by which polypeptides nucleate and aggregate (see [16] for a review).

Kinetic studies of aggregation rely on knowing the number of growing ends per unit weight of the fibril. While measuring this directly has proved challenging, a group at the University of Pittsburgh demonstrated a viable method ([17]). This technique involves growing the fibril using biotinylated polyglutamine monomers, and then tagging by fluorescent Eu-streptavidin, a protein that binds strongly and specifically to biotin. What is yet unclear is why this procedure yields counts proportional to the number of growing ends, and not proportional to the total weight of the fibrils. In other words, why is the assay selective at all?

In this master's thesis, we address this question using Molecular Dynamics (MD) simulation of a biotinylated glutamine residue attached to a polyQ fibril. First, we present the background in more detail and motivate the project. Then, we discuss the methods adopted, following which we state the research problem. Finally, we present the data gathered and some analysis, and end with a discussion of future directions that are open and interesting.

Chapter 2

Background and Motivation

2.1 Polyglutamine diseases

Over the years, a class of nine neurodegenerative diseases has been shown to be linked to an expanded tract of CAG repeats in the protein-coding region of the affected gene. the sequence CAG codes for the amino acid Glutamine, and therefore the coded proteins contain expanded Glutamine tracts, lending this disease class the name *polyQ diseases*. The following diseases are currently classfied as polyQ:

Disease	**Affected protein**	**PolyQ repeat length**	
		Normal	**Pathological**
SCA1	Ataxin-1	6-39	41-83
SCA2	Ataxin-1	14-32	34-77
SCA6	CACNA1A	4-18	21-30
SCA7	Ataxin-7	7-18	38-200
SCA17	TBP	25-43	45-63
MJD/SCA3	Ataxin-3	12-40	62-86
HD	Huntingtin	6-35	36-121
DRPLA	Atrophin-1	3-38	49-88
SBMA	Androgen receptor	6-36	38-62

TABLE 2.1: The nine CAG repeat disorders. SCA: spinocerebellar ataxia, MJD: Machado-Joseph disease, HD: Huntington's disease, DRPLA: dentatorubropallidoluysian atrophy, SBMA: spinal bulbar muscular atrophy (Kennedy's disease), CACNA1A: calcium channel, voltage-dependent, P/Q type, a 1A subunit, TBP: thymine adenine thymine adenine (TATA) box binding protein. Source: [7]

PolyQ diseases occur with frequencies of about $1 - 10$ per $100,000$. With the exception of SMAX1/SBMA, they are all autosomal dominantly inherited [7]. Therefore, if one of the parents has one affected gene, the child has a 50% chance of being affected, regardless of gender.

While research on these diseases is active, the precise disease mechanism is not well understood. For some of the diseases such as HD, SBMA and SCA3 (see table 2.1), the first step in pathogenesis has been identified as the cleavage of large polyQ-containing fragment from the affected proteins. Proteins with abnormally large polyQ tracts are prone to misfolding, and it is believed that the toxic protein fragments initiate the nucleation and the subsequent aggregation of polyQ proteins.

While the initial nucleation is thermodynamically unfavourable, the subsequent growth of the aggregate occurs is faster. The toxic oligomers are then believed to accumulate in intranuclear inclusions in the patient's cells and disrupt biological functions.

In order to design therapeutic strategies, it is necessary to understand the disease mechanism completely. Based on the current understanding, following solutions have been proposed (in e.g. [18]):

1. Reversing cellular defects:

   (a) Transcription: Particularly in the case of HD, mutant polyQ proteins cause a disruption in the activity of key factors, many of which possess acetyl transferase activity. This strategy aims to combat this by increasing gene expression through histone deacetylase (HDAC) inhibitors. However, these molecules may increase the acetylation of other non-histone proteins.

   (b) Cellular metabolism: This strategy targets transcriptional regulators, which in turn affect the expression of expression mutant genes. For example, over-expression of the regulator PGC-1α may be a potential strategy to suppress mutant Huntingtin (Htt).

2. Targeting polyglutamine proteins

   (a) Gene therapy: Targeting the CAG disorder at the source, i.e., by selectively reducing expression of the affected gene. Use of small interfering RNAs to reduce gene expression has borne success in mouse models. However, it remains a challenge to limit the targeting to the mutant gene. Long term safety is uncertain.

   (b) Proteolysis: If specific proteases that cleave polyQ proteins to generate toxic fragments are identified, they can be targeted using protease inhibitors.

   (c) Protein clearance: Stimulating cellular degradation pathways that preferentially target misfolded proteins.

   (d) Direct targeting of protein aggregation: molecular chaperones that aid in protein refolding and degradation, and/or small molecules that directly interfere with polyQ aggregation. The first strategy is likely to have limiting side effects. Cell-based assays with small molecules have yielded promising results, the same promise is not seen in experiments on mice. This strategy does not stop the initial pathological misfolding, which may still result in toxic oligomers that affect cell function.

(e) Stabilizing native conformation: Influence the equilibrium between the natural and toxic conformations. This can be done by directly targeting the protein with interventions that stabilize the native conformer.

In addition to its relevance in therapeutics, polyQ aggregation kinetics is an interesting area of research in its own right. In the following sections, we discuss its theory and experimental aspects.

2.2 PolyQ aggregation kinetics

There are several mechanisms through which proteins can aggregate [19]. An important class is nucleated growth polymerization, in which the aggregation is initiated by the (thermodynamically unfavourable) formation of a nucleus, followed by an elongation phase which leads to fibril formation. This mechanism is reponsible for amyloid fibril formation in the case of, e.g., Aβ(1-40) protein (Alzheimer's disease) and α-Synuclein (Parkinson's disease) [20]. The nucleus is typically a misfolded, prion-like monomer or oligomer which, due to its instability, has a tendency to either dissociate into stable monomers, or elongate.

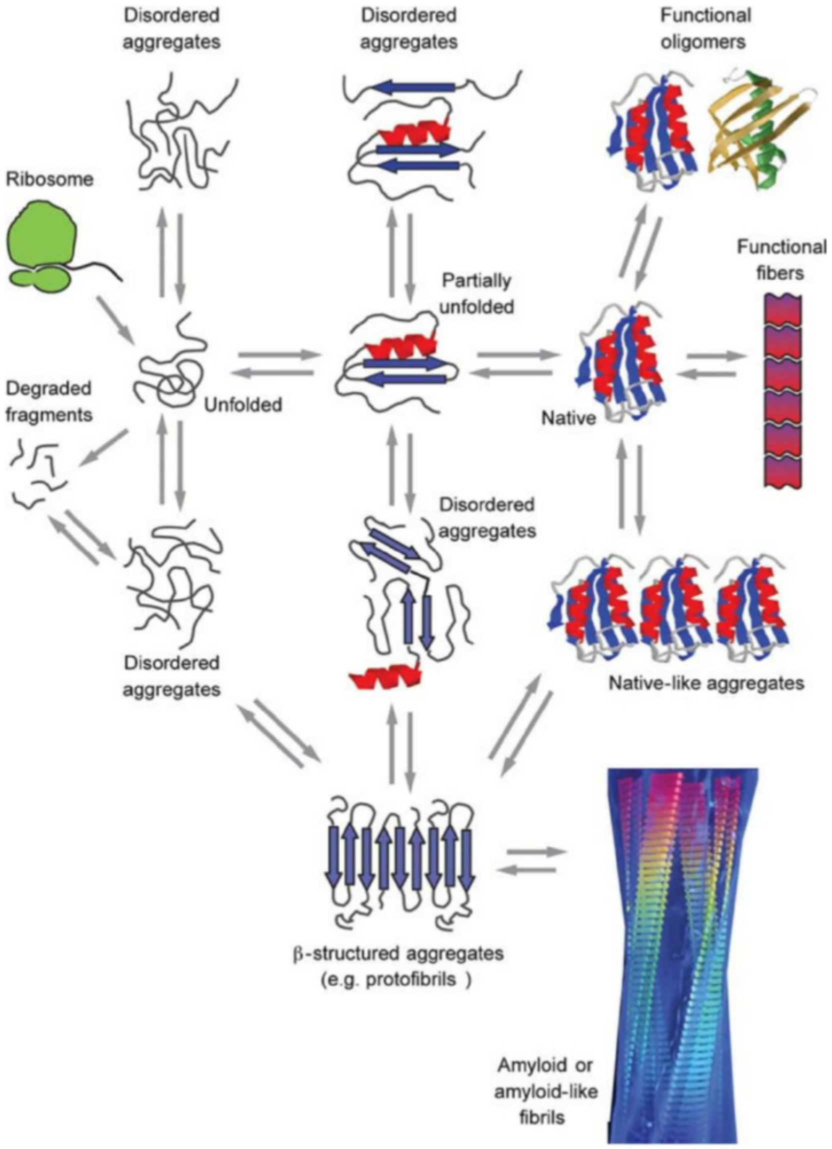

FIGURE 2.1: Various pathways for the formation of amyloid fibrils. Source: [1]

The elongation of amyloid fibrils is commonly described as a two-step kinetic process, the so-called “dock and lock” mechanism. In the first (“docking”) stage, a new monomer from the solution attaches itself to the fibril at an elongation site, also called a growing end. Then, the new monomer undergoes a conformational change which usually mimics the underlying template, and thus “locks” into place. The next monomer can then be incorporated in a similar manner, and the aggregation proceeds. Barring the inital nuclear stage, the forward reaction is more favourable at every step in the aggregation - this is why, even though the nuclei themselves are small and short-lived, the aggregates, once initiated, tend to be macroscopic and very stable.

PolyQ aggregation is known to follow a nucleation-dependent pathway. In fact, polyQ aggregation is particularly straightforward, since polyQ does not exhibit oligomeric or protofibrillar forms that are commonly seen in other amyloid aggregation reactions. Another complicating feature, which is the fracture of aggregates under normal growth conditions leading to secondary nucleation, is not seen in the case of polyQ. Lastly, the nucleus of a polyQ aggregate is known to be monomeric which further simplifies the analysis.

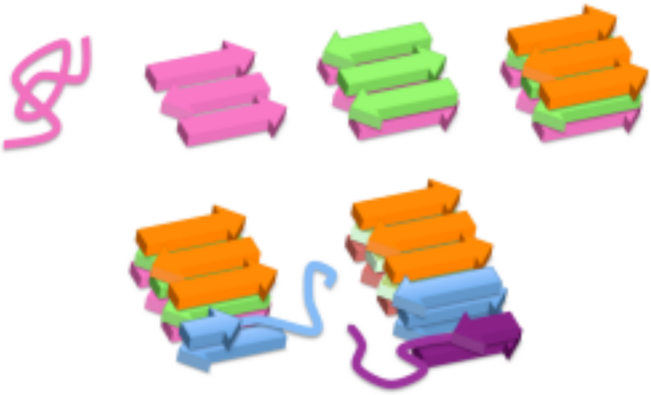

FIGURE 2.2: Nucleation and elongation of polyQ. A polyQ monomer (top left) misfolds into an unstable conformer, a popular candidate for which is the β-hairpin, pictured above (top left-center). It has been suggested that the β-hairpins interdigitate vertically (top right), following which subsequent monomers can "dock and lock" at the growing face, as shown in the bottom row. Source: [2] (figure published as part of poster)

The nucleation growth kinetics of polyQ contains two important quantities, the nucleation equilibrium constant K_{n^*} and the second-order elongation rate constant k_+. The first constant K_{n^*} is crucial in the description of nucleation kinetics. It is however very difficult to measure directly in the lab for two reasons:

1. nucleation is a very rare event, and
2. the nucleus phase is thermodynamically unfavourable and hence very short-lived; nuclei, once formed, either return to the monomeric state or quickly begin aggregation.

As a result, one must resort to indirect calculation of the nucleation constant K_{n^*}. Through kinetic analysis, an auxiliary quantity involving K_{n^*} can be measured.

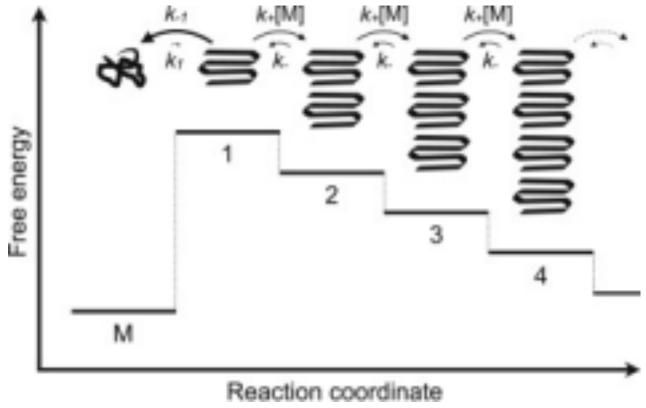

FIGURE 2.3: The typical energy landscape of the nucleation/elongation reactions. ”M” refers to the monomer, which is a stable conformation. Nucleation is the formation of the ”critical nucleus”: a high-energy, unstable conformer of one or more monomers (denoted by 1). Subsequent elongation events (2, 3, 4 etc.) are energetically favourable and lead to aggregation. Source: [2]

When the concentration of critical nuclei is small, the growth rate is determined by the nucei concentration and the rate of elongation of nuclei. As is done in [21], this can be modelled as a first-order rate equation related to crossing an effective energy barrier. So, for a concentration of critical nuclei c^*, nuclear elongation rate J^*, and a polymer concentration c_p, the elongation rate equation is given by

$$\frac{\mathrm{d}c_p}{\mathrm{d}t} = J^* c^* \tag{2.1}$$

Once formed, polymers elongate by aggregation of new monomers at their ends. For a one-dimensional polymer, this rate is then independent of the dimensions (specifically, length) of the polymer. So, a uniform rate J may be assumed for all polymers, and then if the monomer concentration incorporated into polymers is denoted by Δ, we have

$$\frac{\mathrm{d}\Delta}{\mathrm{d}t} = J c_p \tag{2.2}$$

These two rate equations may be combined by integrating equation 2.1 substituting into 2.2. The fully integrated equation describing the nucleation phase then reads

$$\Delta = \frac{1}{2} J J^* c^* t^2 \tag{2.3}$$

Given a critical size n^* (in number of monomers) for the nucleus, the nucleus concentration is given by $c^* = K_{n^*}c^{n^*}$, where as before, K_{n^*} is the equilibrium constant for the monomer-nucleus reaction. Further, the total elongation rates can be expressed in terms of monomer/nuclear concentration as $J^{(*)} = k_+ c^{(*)}$. As before, k_+ is the forward elongation constant. (Note: an important assumption here is that the elongation constant is the same for the nucleus and the aggregate. This is because, as we have already observed, the elongation of the nucleus cannot be directly measured.) Therefore,

$$\Delta = \frac{1}{2}k_+^2 K_{n^*} c^{(n^*+2)} t^2 \tag{2.4}$$

A plot of Δ vs. t^2 will then have a slope of $\frac{1}{2}k_+^2 K_{n^*} c^{(n^*+2)}$, which has a power law dependence on the concentration c. Taking logs,

$$\log(\text{slope}) = (n^* + 2)\log c + \log\left(\frac{1}{2}k_+^2 K_{n^*}\right) \tag{2.5}$$

Therefore, measuring the slope for varying concentrations and plotting $\log(\text{slope})$ vs. $\log(c)$ yields two quantities, the power on c and the log of the pre-factor $\log\left(\frac{1}{2}k_+^2 K_{n^*}\right)$. It was found in [21] that the critical nucleus size n^* (calculated from the slope) is equal to 1.

The pre-factor $\frac{1}{2}k_+^2 K_{n^*}$ contains the desired equilibrium constant K_{n^*} and the elongation constant k_+. In order to solve for K_{n^*}, a value for k_+ must be found. However, a direct measurement of k_+ is also difficult, since it requires the knowledge of the number of productive elongation sites per weight of the polyQ aggregate (denoted as [ends]), and the pseudo-first order elongation rate constant $k*$: $k_+ = k*/[\text{ends}]$ ([22]). A direct measurement of both $k*$ and [ends] was accomplished by Wetzel *et al* in [17]. We discuss this experiment in a later section, 2.5.

2.3 Structure of the nucleus

The first, crucial step in the nucleated polymerization mechanism is the formation of a thermodynamically unstable state of a single monomer or an oligomer, which can either collapse back to stable monomeric states or elongate by addition of other monomers. The added monomers first dock onto the growing fibril and then “lock in” through a change in conformation. The nucleus size at which elongation becomes possible is the critical nucleus n^*, also known as the thermodynamic nucleus.

In the case of polyQ, an important question is, what is the structure of the monomeric nucleus which best supports aggregation? In Miettinen et al [2], it was found through simulation that a beta-hairpin geometry is the most likely candidate out of β-helical and various β geometries, based on the stability of docked monomers. Note that they define the critical nucleus as $1 + n^*$, i.e., the nucleus size at which elongation is more favourable than dissociation. This is called the structural nucleus. The structure of the nucleus and of the added monomers is essential to understanding the secondary structure of the polyQ fibril. This understanding helps us accurately simulate the fibril and its interaction with biotin.

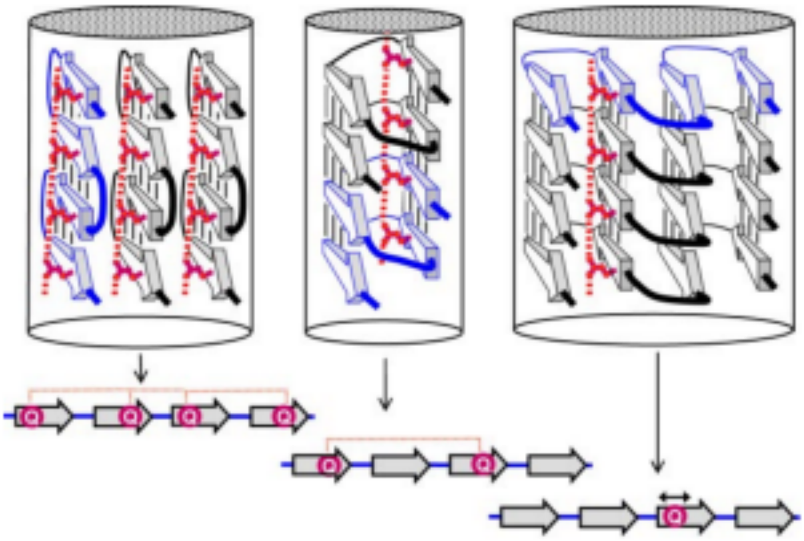

FIGURE 2.4: PolyQ fibrils with three different monomer geometries. From left to right: β-hairpin, β-spiral and β-arc geometries. The growth axis in each fibril is vertical. In each fibril, one glutamine residue has been drawn (red) to illustrate the orientation of the side chains and how they H-bond. The cartoons on the bottom illustrate the loci of the H-bonding glutamines on the monomer. Source: [3]

2.4 Biotin-Streptavidin binding

Biotin and streptavidin (alternatively, avidin) are naturally occuring biomolecules which are known to have a large binding affinity to one another. The dissociation constant for binding is $K_d \sim 10^{-14}$ M, which makes biotin-streptavidin one of the strongest known non-covalent bonds [23]. Moreover, biotin-streptavidin binding is highly specific and stable over a large range of temperature and pH. These features have made this reaction a popular choice in various techniques used in biotechnology.

In biochemical assays, biotin is used to mark molecules of interest, which may be identified and isolated by allowing streptavidin to bind with the attached biotin. The streptavidin may have a fluorescent attachment such as nanogold particles or Europium, or a reporter molecule such as horseradish peroxidase for detection with Enzyme-Linked Immunosorbent Assay (ELISA).

The strength of biotin-streptavidin binding has been used to "anchor" biomolecules - in [24] for example, DNA strands are anchored by Au-biotin-streptavidin-biotin-DNA bonds, and, using flow and electric fields, can be fully extended and imaged. The bond can withstand a force of at least 11 pN.

Lastly, we observe that while avidin has the larger affinity for free biotin, streptavidin holds a number of advantages over avidin in applications, such as

- a higher binding affinity with biotin attached to a substrate
- higher specificity
- lower activity and lower tendency to aggregate (relative to avidin).

2.5 Experimental studies of polyQ aggregation

In this section, we discuss some experimental studies conducted by Ronald Wetzel *et al* on polyQ kinetics. The experiment of most interest is the measurement of the number of growing ends of polyQ by tagging the fibrils with a special biotinylated polyQ monomer, and then binding (fluorescent) Eu-Streptavidin to the biotin incorporated into the fibril.

We start by describing the original experiment conducted in 2005.

The experiment: The aim of the experiment in [17] was to measure not only the pseudo-first order rate constant $k*$ but also the number of elongation sites per aggregate weight, allowing one to calculate the *second order* rate constant $k_+ = k*/[\text{ends}]$. In this section, we provide a specfic description of the experiment in question, [17].

In the experiment, the following synthetic peptide monomers were used:

- Q_{47}: This monomer is used to grow the basic polyQ fibrillar structure *in vitro*.
- Q_{30}: This is a smaller peptide segment which was used to study the elongation kinetics of the fibrils. The length of 30 amino acids is chosen to match the biotinylated residue, which is listed next.
- biotin-PEG-Q_{29} (or BQ_{29}): The biotinylated polyQ residue used to mark the fibrils. This chain length is found to be optimal for elongation of aggregates of a variety of geometries and repeat lengths [25].

Several morphologies are observed for the aggregates grown *in vitro*, for example broad ribbons (~ 50 nm wide), thin filaments (~ 4 nm) and amyloid-like fibrils ([25, 26]). Typically, the aggregates grow to be a few hundreds of nanometers in length. In comparison, a fully extended $Q30$ peptide has a length of ~ 10 nm. By varying experimental conditions (temperature and pH), these varying morphologies can be generated. Six different samples of the Q_{47} aggregate were thus produced [26].

Next, the experiment studied the binding of BQ_{29} to the polyQ aggregate. The main idea here is that the bound biotin can also bind with streptavidin. If one introduces streptavidin which is tagged with a fluorescent marker, then the binding can be measured using fluorescence counts. In order to compute specific binding, the weight of the aggregates was measured using High-Performance Liquid Chromatography (HPLC) on (small) extracted samples, or *aliquots*.

The weight concentration of aggregate used in this experiment was in the range of 190-260 ng/ml. The aggregate was incubated with 0.5-1.0 μM BQ_{29} at $25°$ C.

Then, the aggregate solution sample was incubated with europium-complexed streptavidin for one hour at room temperature and in the dark. We note here that the Eu-streptavidin was diluted 1:1000 from the original solution purchased from PerkinElmer.

Then, time-resolved fluorescence measurements were made on the sample. Using fluorescence counts, it is possible to convert into a concetration of bound Eu-streptavidin using a calibration of seven molecules of europium per streptavidin. In this, one also assumes that, despite being tetrameric, streptavidin binds to only one molecule of attached BQ_{29}. This assumption is not unreasonable, given that the biotin is not free in solution but attached to a large substrate. Non-specific binding was also corrected for using appropriate controls (Q_{30} without biotin).

In order to analyze binding kinetics, binding data has to be collected at various points in time. However, before the binding is measured, the reaction must first be "quenched", i.e., one should be able to turn off the binding of BQ_{29}. Towards this end, the mixture containing the aggregate and BQ_{29} was quenched by adding Q_{30}. The data points were collected at intervals of ~ 15 minutes.

Results: The experiment was in fact carried out in two modes which differed from one another in the method of data collection (see [17]). Without delving into further details, we remark that the first mode provided a qualitative picture, while data from the second mode was used for kinetic calculations. Now, we discuss the results obtained in each mode.

In the first mode, it is found that the amount of biotin bound (in the presence of Q_{30}) increases with time and saturates after about 20 minutes of incubation. The amplitude (~ 10 fmol for unsonicated aggregate in Phosphate-Buffered Saline (PBS)) does not change appreciably after saturation. This behavious is in stark contrast to the various controls.

Firstly, if one does not quench the initial elongation of Q_{47} aggregates with BQ_{29} by adding Q_{30} to the solution, the binding is found to be negligible over time. Therefore, the addition of Q_{30} seems to be crucial to ensure stably bound biotin-polyQ. The paper discusses that any BQ_{29} attached to a growth site must quickly be capped by Q_{30} monomers, otherwise it will not remain stably bound. In the absense of any Q_{30}, none of the bound BQ_{29} may be fixed and may detach from the growth sites during, say, washing of the sample. We add that the "locking" timescale for BQ_{29} may be much larger than the lab timescales of tens of minutes, resulting in low retention of the tagged monomer at growth sites.

In all the other controls, only background levels of BQ_{29} binding were seen. This includes elongation with a biotin-labelled mutant (abbreviated as $PGQ_{29}P^{2,3}$), and cross-binding with biotinyl-A$\beta\,(1-40)$.

In the second mode, the Q_{30} quench is immediately followed by a second incubation of the aggregate, BQ_{29} and Q_{30} mixture at 25° C. However, the authors believe that the fibril growth is not optimally fast at this temperature. Therefore, some bound BQ_{29} may detach during the incubation instead of getting fixed by addition of Q_{30} monomers. Since elongation is strongly favoured at a temperature of 37° C, it is better to carry out the second incubation at this temperature rather than at 25° C, since this would minimize the loss in binding amplitude. This is the motivation for carrying out the second mode of experiment, and why it was believed to be more accurate.

As in the first mode of the experiment, the observed binding of BQ_{29} grows over time and saturates after 20 minutes. However, it is argued that the growth is not due to binding at new sites but multiple BQ_{29} monomers being added to the same sites. In fact, the authors argue that by the end of the first incubation of aggregate with BQ_{29}, it is expected that all productive growth sites are tagged in a one-to-one manner. Therefore, the value which correctly reflects the number of growth sites is the binding amplitude at the moment of quenching with Q_{30}, i.e. at zero time. Therefore, the growth in the first five minutes since quenching is linearly fitted and extrapolated to $t = 0$, and this value was chosen as an indicator for the number of growing ends. Then, the number of bound BQ_{29}, measured through fluorescence counting, was assumed to be equal to the number of growing sites.

In order to compute k_+, the pseuodo-first order elongation rate was measured directly from the time-dependent binding of monomer. Lastly, analysis of the monomer binding rate at different concentrations provided the quantity $1/2K_{n^*}k_+^2$. These together gave a value of $K_{n^*} = 2.6 \times 10^{-9}$ (note: K_{n^*} is dimensionless).

Further observations about biotin-polyQ binding: The group of Ron Wetzel has worked on many experiments involving tagging of polyQ with biotin before and since 2005. The results of the 2005 experiment led to the understanding that biotin-polyQ by itself cannot stably bind to growing fibrils, and that a "chase" of unlabelled polyQ is necessary to fix the bound BQ_{29} in place. Over the years however, a different picture has emerged. In a recent private communication, Dr. Wetzel reported that biotinylated monomers tend not to give any signal when bound far away from the growing end of the aggregate, while only those near the current growing site seem to be observable.

Further evidence that strongly supports this hypothesis is presented in the doctoral thesis of Elizabeth Landrum, [4]. In this work, various quantitative experiments on aggregation kinetics are presented. We discuss some of the relevant experiments and their results below:

- Seeded elongation reactions: A "seed" in this context refers to an existing polyQ fibril with active growth sites, of a known weight and size. This is measured using dynamics light scattering (DLS). Seeded elongation was carried out by coincubation of the pre-formed seed with small polyQ monomers (~ 30 peptides long). By measuring the unbound monomer concentration at various times with the help of a sedimentation assay followed by reverse phase HPLC, the concentration of the amount of peptide bound was monitored. The first 20% of the elongation data was used to compute the pseudo-first order elongation rate constant, $k*$.
- Growing ends reaction: The rate constant $k*$ is an extensive quantity, i.e. it depends on the aggregate weight and is therefore less reflective of the microscopic nature of aggregation. The more useful quantity is the *second-order rate constant* k_+, which is expressed as elongation rate per unit aggregate weight. In order to compute this experimentally, one has to keep track of the number of growing ends while simultaneously carrying out seeded elongation. The growing ends titration experiment is a method developed by the Wetzel group to reliably measure the number of growing ends on the polyQ aggregate. The experiment is very similar to the one described previously, [17], but with one key difference: this experiment lacks the quenching step. That is, unlabelled polyQ monomers are not added to the biotin-polyQ+aggregate incubate. While the 2005 experiment results seem to imply that this step is crucial for polyQ binding, this is in fact not the case - over long timescales (hours), the biotin-polyQ binding is unaffected by the presence of unlabelled polyQ.

- The protocol to measure the amount of polyQ bound is also similar to [17]: fluorescence counts of Eu-streptavidin bound to the biotin are converted directly to number of biotins, which are assumed to be in one-to-one correspondence with the number of growing ends. The counts are measured after biotin binding saturates, which corresponds to all growth sites being bound, it is argued.

- Comparison with [17] raises a basic question. Earlier, it was believed that binding must be measured immediately after the quench, so as to prevent multiple biotins binding on a single growing fibril. Then, why is it that in the more recent experiment, saturation binding is measured? Shouldn't the same problem of multiple binding per site persist? In fact, shouldn't the problem be more serious here because *all* of the monomers are tagged? No. Surprisingly, it is found that Eu-streptavidin binding does not depend on the number of labelled polyQ in the elongated fibrils! Independent of length, only one count is detected per growing end. EM images (see figure 2.5)further show that only the biotins at the very end of a growing fibril bind to streptavidin and give a signal. The interior biotins are rendered unavailable.

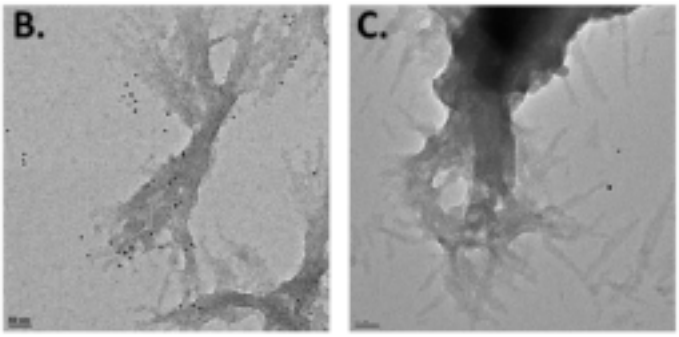

FIGURE 2.5: Electron Microscopy (EM) images of the polyQ fibrils tagged with Au-streptavidin bound to biotin in the BQ_{29} monomers. Left: polyQ fibril with biotinylated monomers incorporated, right: control with no biotinylated monomers in the fibril. The black dots are signals. The signal observed is much lower than the amount of biotin incorporated into the fibrils, and is seen not to come from the bulk of the fibrils but from the ends. Source: [4]

- Could it be that only the end monomer contains biotin? No. By keeping track of the total amount of monomer bound through sedimentation assay/HPLC, one can compare the amount of monomer bound to the amount of monomer detected by fluorescence. It is found that at saturation, the number of monomers bound exceeds the detected biotin *by a factor of* 10^5. This is very significant, as it shows that while biotin is present along the entire fibril, only the end biotin is available for binding.

The final observation raises an interesting question about the mechanics of bioting-polyQ binding: Why is the growing end of polyQ aggregates special? Why do biotin-surface interactions along the length of the fibril (i.e. not on the growth site) make the biotin unavailable to streptavidin with near-certainty?

Chapter 3

Problem Statement

We are now ready to formulate the main research problem. From the experiments conducted by Wetzel *et al*, it is apparent that biotin-streptavidin binding is selective of the surface to which the biotin is attached. In particular, streptavidin has a strong preference for binding with biotin attached to growing ends of the fibril. However, the reason for this selectivity has remained unclear. Experimentally, it is near impossible to "see" how individual biotin-streptavidin interact close to a given surface (either growing or non-growing) of the fibril. On the other hand, any theoretical model would have to capture the surface chemistry of the aggregate and biotin with sufficient accuracy, if any quantitative predictions are to be trusted. A verification of the model would also be necessary.

In light of the challenges above, we propose instead to use a simulation-based approach to gain an understanding into the behaviour of biotin attached to the surface of a polyQ aggregate. The features seen in simulation may then be used to construct simple theoretical models of the biotin-surface interaction.

A computational approach is well-suited to tackle this problem for several reasons.

- Several computational and experimental studies have shown that polyQ fibrils are not amorphous but have a regular structure (at the peptide level). It is therefore possible to study a representative block of the aggregate (with periodic boundary conditions) while preserving the accuracy of the simulation.

- Different “faces” of the aggregate have a few different geometries, all of which can be separately simulated. In particular, growing ends of all polyQ fibrils are structurally similar to each other *and* distinct from other (non-growing) ends of the fibril. Assuming the geometries and the biotin-surface interactions can be represented accurately in simulation, the hypothesis that biotin’s interaction with the surface has a structural dependence on the surface geometry becomes numerically testable.

- Lastly, we note that the entire system can be represented with a collection of no more than 10^{4-5} atoms. An MD simulation with this system size is tractable using standard computational resources.

As mentioned above, it is appealing to think that structural differences between the fibril faces might be responsible for the selective binding of streptavidin to attached biotin. Perhaps biotin (more accurately biotin-PEG-polyQ) interacts with the substrate and becomes unavailable for binding depending on the geometry of the underlying face on which it is attached. While this is not the only possible hypothesis, it is the one we choose to explore, since it is elegant and easier to set up computationally. Due to work done in [27–29] etc., we have an understanding of how the structure of a polyQ fibril might be. The exact biotinylated monomer used in experiments is also known.

Therefore, the goal of this project is to simulate the interaction of a biotinylated glutamine residue attached on either the growing or the non-growing face of a polyQ fibril suspended in water, and to discover the mechanism responsible for the differential binding of biotin to streptavidin when attached to different faces of the fibril.

Chapter 4

Materials and Methods

4.1 Simulation techniques

Since all numerical methods are, to varying degrees, approximations, it is important to pick the method which is accurate for the system and the timescales of interest.

Quantum mechanical (QM) methods are able to simulate the system exactly by solving the relevant Schrödinger equation. However, they are limited to systems of a few molecules due to the high computational cost. Hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods are an interesting compromise, as they try to combine exact QM for a small group of atoms with MM for the larger system.

In the biotin/polyQ system, it is unlikely that a QM/MM level accuracy will be necessary. We are interested in the mechanistic differences between the biotin-polyQ interaction at the two surfaces. On the other hand, this system may have large memory times, so a time sampling on the order of microseconds may be desirable.

All-atom numerical simulations generally face the trade-off of microscopic accuracy vs. sampling timescale. Classical empirical methods are based on calculating a force-field which governs the atom dynamics. Force fields are all-atom potentials which include few-body interaction terms with coefficients that are calculated by fitting against quantum chemistry simulations/calculations or using observables in experiment. Force fields fix the bond configuration, and are thus unable to simulate electron exchange, bond-breaking etc. But by being computationally inexpensive, force fields afford long sampling times.

One technique which can be useful in extending the simulation timescale, especially for systems with a regular and/or large-scale structure, is known as Coarse-Graining (CG). Here, one identifies groups of atoms which are expected to exhibit a collective behaviour and replaces such groups by pseudo-atoms. The coarse-grained (CG) system then consists of such pseudo-atoms that interact with one another. Starting from an all-atom force field, a CG force field must be derived for the pseudo-atoms. Unlike all-atom simulations, CG methods ignore dynamics which occur on length-scales smaller than the group size. However, CG simulations can be carried out with longer time-steps than all-atom, and can as a result be simulated for longer total times.

In evaluating the need to coarse-grain, the main consideration is the system size (10^{4-5} atoms) and the size of biotin, the molecule of interest. We note that the biotin-surface interaction is likely to depend on the exact atom-atom interactions (H-bond, Van der Waals, etc.), and it will be challenging to coarse grain the small biotin molecule while preserving the accuracy of this interaction. Moreover, the total system size is not prohibitively large. Therefore, we choose to carry out all-atom simulations.

For the reasons mentioned above, we choose to use all-atom, classical MD for our simulations. Next, we describe some of the theoretical aspects of MD, and then move on to discussing the details of our MD simulation in section 4.3 onwards.

4.2 Molecular dynamics

4.2.1 Equations of motion

In classical dynamics, the state of a system can be specified by a finite number N of independent coordinates q_i and their time derivatives $\dot{q}_i$, $i = 1, \ldots, N$. The evolution of the state is governed by the system's Lagrangian,

$$\mathcal{L}\left(\{q_i\}, \{\dot{q}_i\}, t\right) = \mathcal{T} - \mathcal{V}$$

where $\mathcal{T}$ and $\mathcal{V}$ are the system's kinetic and potential energy respectively. If $\mathcal{L}$ has no *explicit* time dependence, then one can always express the kinetic energy as a quadratic function of the coordinates,

$$\mathcal{T} = \sum_i \frac{1}{2} m_i \dot{q}_i^2 \tag{4.1}$$

where m_i is a mass term associated with coordinate i. For time-independent Lagrangians it can be shown that the quantity $E = \mathcal{T} + \mathcal{V}$, called the *energy*, is a constant of motion.

If all forces in the system are *conservative*, i.e. if the work done by each force is path-independent, then the potential term is purely a function of the coordinates q_i.

The equations of motion are then given by the Euler-Lagrange formula,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i}\right) = \frac{\partial \mathcal{L}}{\partial q_i} \tag{4.2}$$

For conservative and time-independent forces, the above equation can be expressed as

$$m_i \ddot{q}_i = F_i \tag{4.3}$$

where $F_i = -\partial \mathcal{V} / \partial q_i$ is the force on coordinate i.

In all-atom MD simulations, a system of N atoms is described by the coordinates and velocities of each atom, which for 3-dimensional systems is a set of $6N$ real numbers. Denote the coordinate and velocity vector of the i-th particle by $\vec{r}_i$ and $\vec{v}_i$ $(= \dot{\vec{r}}_i)$ respectively. The spatial configuration of the system is then given by $(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)$ and the velocities by $(\vec{v}_1, \vec{v}_2, \ldots, \vec{v}_N)$.

Typically, molecular processes can be modelled reasonably well by forces with no explicit time or velocity dependence, and therefore we may write the (unconstrained) equation of motion for particle i as

$$m_i \ddot{\vec{r}}_i = \vec{F}_i, \tag{4.4}$$

$$\vec{F}_i = -\nabla_{\vec{r}_i} \mathcal{V} \tag{4.5}$$

The potential function $\mathcal{V}(\{\vec{r}_i\})$ is called the *force field*.

4.2.2 MD integrators

Starting from a set of $6N - C$ initial conditions (where C is the number of constraints), the system configuration at time t can be computed by integrating the equations of motions, 4.4. Since integrating analytically is not possible in most cases, a numerical integration scheme has to be applied. Ordinary differential equations such as 4.4 can be converted into a series of finite difference equations, in which the simulation time is divided into time steps with a spacing of Δt between consecutive steps. The coordinates at the current time step are then given as a function of the coordinates of a small number of previous time steps (typically one or two).

In order to see how various quantities are related, we write out their Taylor series expansions (for the i-th particle),

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i(t)\Delta t + \frac{1}{2m_i}\vec{F}_i(t)\Delta t^2 + O(\Delta t^3) \tag{4.6}$$

$$\vec{v}_i(t + \Delta t) = \vec{v}_i(t) + \frac{1}{m_i}\vec{F}_i(t)\Delta t + O(\Delta t^2) \tag{4.7}$$

Naïvely, we could truncate the Taylor series to obtain the following difference equations:

$$\begin{aligned}\vec{r}_i(t_{n+1}) &= \vec{r}_i(t_n) + \vec{v}_i(t_n)\Delta t + \frac{1}{2m_i}\vec{F}_i(t_n)\Delta t^2 \\ \vec{v}_i(t_{n+1}) &= \vec{v}_i(t_n) + \frac{1}{m_i}\vec{F}_i(t_n)\Delta t\end{aligned}$$

The force $\vec{F}_i(t)$ in this and all future schemes is calculated from the force field $\mathcal{V}$,

$$\vec{F}_i(t) = -\nabla_{\vec{r}_i(t)}\mathcal{V}$$

This integration scheme is not time-reversible, since a sign reversal of Δt does not preserve the equations. It is also not symplectic, so it may not preserve phase-space volume. The discretization error is of the order $O(\Delta t^3)$, which can accumulate over long times. Other schemes can reduce the order of the error. For example, the equations of motion in 4.4 may be discretized directly as follows:

$$\begin{aligned}\vec{F}_i(t_n)/m_i = \ddot{\vec{r}}_i(t_n) &= \frac{\frac{\vec{r}_i(t_{n+1})-\vec{r}_i(t_n)}{\Delta t} - \frac{\vec{r}_i(t_n)-\vec{r}_i(t_{n-1})}{\Delta t}}{\Delta t} \\ &= \frac{\vec{r}_i\left(t_{n+1}\right)+\vec{r}_i\left(t_{n-1}\right)-2\vec{r}_i\left(t_n\right)}{\Delta t^2} \\ \implies \vec{r}_i(t_{n+1}) &= 2\vec{r}_i\left(t_n\right)-\vec{r}_i\left(t_{n-1}\right)+\vec{F}_i(t_n)\Delta t^2/m_i\end{aligned}$$

This method, known as the *Verlet* integrator, has an error of order $O(\Delta t^4)$, a significant improvement over the simple truncation method. The reason for this improvement is that the odd Taylor terms from $\vec{r}_i(t_{n+1})$ and $\vec{r}_i(t_{n-1})$ have opposite sign and therefore cancel out, for a constant time spacing. However, a major limitation of the Verlet method is that velocities are not calculated explicitly; they are however needed in order to compute important quantities such as the pressure and temperature of the system.

This limitation can be overcome by explicitly computing the velocities at each time step as follows

$$\vec{v}_i(t_{n+1}) = \vec{v}(t_n) + \frac{\vec{F}_i(t_n)+\vec{F}_i(t_{n+1})}{2m_i}\Delta t$$

This integrator, known as *Velocity-Verlet*, is time-reversible and symplectic, making it a very robust integration scheme.

Lastly, we introduce a variant of Velocity-Verlet, the *leap-frog* integrator, in which velocities are computed at a half-step time shift from the displacements, as follows:

$$\begin{aligned}\vec{v}_i(t_{n+1/2}) &= \vec{v}(t_{n-1/2}) + \frac{\vec{F}_i(t_n)}{2m_i}\Delta t \\ \vec{r}_i(t_{n+1}) &= \vec{r}(t_n) + \vec{v}(t_{n+1/2})\Delta t\end{aligned}$$

where we used the shorthand $t_{n+1/2} = t_n + \Delta t/2$. Leap-frog has no advantages over Velocity-Verlet; on the contrary, Velocity-Verlet provides velocities and displacements at the same time point, making it more useful than leap-frog.

4.2.3 N,V,T,E control

Systems with many degrees of freedom, such as those frequently encountered in molecular dynamics, have two possible descriptions – microscopic and macroscopic. A microscopic description of the system is a specification of all the independent coordinates (e.g. positions and momenta of all particles) as a function of time. A macroscopic description of the system is given in terms of so-called state functions or thermodynamic variables, which are global properties that are measurable in the laboratory – temperature T, particle number N, Energy E, Volume V, Pressure P etc. Since the timescale of measurement is much larger than the dynamical timescale, the state functions are in fact long-time averages over the evolution of the system. Thus, a function A of the configuration Γ, as measured, is given by ([30])

$$A_{\text{obs}} = \langle A\left(\Gamma(t)\right)\rangle_{\text{time}} = \lim_{t_{\text{obs}}\to\infty} \frac{1}{t_{\text{obs}}} \int\limits_{0}^{t_{\text{obs}}} A\left(\Gamma(t)\right) \mathrm{d}t = \frac{1}{N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} A\left(\Gamma(t_i)\right) \tag{4.8}$$

where in the last step we discretize the integral and express it as an average of A over the N_{obs} time steps.

If the dynamics can sample the phase space of the system fully, i.e. if the evolution is ergodic, then we expect that this time-average approaches an average over the entire phase space of the system. This assumption (known also as the *ergodic hypothesis*) is often applied while studying thermodynamical systems. However, the entire phase space is too large, and so we select a representative subset of configurations which, ideally, samples a certain probability distribution on the phase space. This smaller collection of states is known as an ensemble. For practical purposes, and ensemble is chosen by keeping a certain set of thermodynamic variables constant. Thus, we may have ensembles such as the NVE, NVT or the NPT ensembles.

When performing a simulation in a given thermodynamical ensemble (such as canonical, NVT, or microcanonical, NVE) the corresponding macroscopic variables should not change significantly over the course of the simulation - such that, in the thermodynamic limit ($N \to \infty$), they are perfectly invariant. The volume of the simulation box is a simple function of the box vectors, for instance $V = xyz$ in the case of a rectangular box. During the simulation, the box vector sizes may also be treated as variables, as in the case of an NPT simulation. In order to keep this function constant, a constraint is imposed on x, y, z.

Temperature control is desirable in order to study dynamics as a function of temperature, or to simulate certain laboratory (or physiological) conditions. The temperature of a system of N atoms in three dimensions can be expressed as the time-averaged kinetic energy,

$$\langle \frac{1}{2} m v^2 \rangle = \langle K \rangle_{\text{time}} = \frac{3}{2} k_B T \tag{4.9}$$

During simulations, we use the ensemble average, under the ergodic hypothesis. Then, temperature can be controlled by rescaling all the velocities by a constant factor λ. This scales the temperature by a factor λ^2 due to the quadratic dependence ([31]). Then, the required factor is given by $\lambda = \sqrt{T_{\text{req}}/T_{\text{curr}}}$.

In Berendsen coupling, the system is coupled to an artificial bath which is at the desired temperature. The system temperature relaxes to the bath temperature at a rate proportional to the temperature difference,

$$\frac{\mathrm{d}T(t)}{\mathrm{d}t} = \frac{1}{\tau_T} \left(T_{\text{bath}} - T(t) \right) \tag{4.10}$$

where the temperature coupling time τ_T must be fed as a parameter. The temperature change at each step is implemented by rescaling the velocities by the correct factor.

Lastly, pressure can be controlled in ways similar to temperature. Under Berendsen coupling, the system is coupled to a pressure bath at the required pressure, and the evolution of the system pressure satisfies a relaxation equation similar to the one for T-coupling. The pressure coupling time τ_P must be specified.

In the NVE ensemble, the total energy is constant. In practice, numerical integrators do not conserve energy from one step to the next. However, the leapfrog integrator for example conserves total energy on average, as it is *symplectic*. The kinetic energy and potential energy individually are not constant in a simulation; this would imply that the dynamics is trivial. The energy is exchanged between potential and kinetic, while the sum is constant.

4.2.4 Force fields

In MD, a force field is a set of parameters that fully describe the potential function which appears, e.g., in 4.4. In the following section, we discuss force fields in general for all-atom simulations, and introduce the two main force fields used in this project. All-atom force fields provide a set of *atom types* which are used to model the atoms within any given system. Each atom type is specified by a set of *non-bonded* and *bonded* parameters, for the two types of interactions. In order to derive these parameters, one has to identify a *target dataset*, i.e. observables from analytical experiments (NMR, X-ray diffraction etc.), or from semi-empirical or *ab initio* quantum mechanical calculations. For example, bond lengths and angles may be obtained from vibrational spectra, and independently *ab initio*. This target data then serves as a guide for parameter optimization, i.e., finding parameter values that best fit the target data.

The problem of parameter optimization is typically formalized as follows: we represent the parameter space as an N-dimensional real vector $\mathbf{x} \in \mathbb{R}^N$. For the i-th target data, there is a mapping from the parameters to the data which we denote by the function $f_i^{\text{sim}}(\mathbf{x})$. In other words, any given configuration of parameters x gives an approximation for the i-th observable. In parameter optimization then, one seeks to find the optimal parameter vector $\mathbf{x}$, which provides the best approximation to all the observables. For a collection of n target observables, where the i-th observable has the value (taken from experimental or QM data) f_i^{exp}, the *objective function* F is typically given by

$$F(\mathbf{x}) = \sum_{i=1}^{n} w_i \left(\frac{f_i^{\text{exp}} - f_i^{\text{sim}}(\mathbf{x})}{f_i^{\text{exp}}} \right)^2$$

This objective function has a quadratic form, and minimizing it corresponds to a least-squares fitting of the parameters to the data. The problem of efficiently and accurately optimizing parameters has been actively studied, [32–37]. The main challenge here is that the objective function is a non-trivial function over a large search space ($\mathbb{R}^N$ where N may be on the order of hundreds), and we wish to find the *global* minimum. Local methods such as steepest descent are unlikely to produce the best parameter set, and one has to modify them or use heuristic methods such as Monte Carlo, simulated annealing etc.

Once optimized, the parameters are fed into the force field function, which is a potential energy defined on the configuration space of the atoms. At a high level, the force field may be written as

$$V_{\text{total}} = V_{\text{bonded}} + V_{\text{nonbonded}}$$

Each of the above terms is given in terms of additive energy components as below

$$V_{\text{bonded}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{dihedral}}$$

$$V_{\text{nonbonded}} = V_{\text{coulomb}} + V_{\text{van der Waals}}$$

We discuss the above force field terms in greater detail in sections 4.2.5, 4.2.6 below.

4.2.5 Non-bonded interactions

There are primarily two kinds of non-bonded interactions in molecular systems: Coulomb interactions and Van der Waals (VdW) interactions. Both are long-range interactions with force functions that are given by decaying polynomials. In the case of Coulomb interactions, the potential energy between two charged atoms (with charges q_1, q_2) is given by

$$V_{coul}^{12} = \frac{1}{4\pi\epsilon_0\epsilon}\frac{q_1 q_2}{r_{12}} \tag{4.11}$$

where ϵ is the relative permittivity (or dielectric constant) of the medium, and ϵ_0 is the vacuum permittivity. On the other hand, VdW interactions are typically modelled by the *Lennard-Jones potential* as below,

$$V_{LJ} = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{4.12}$$

where the parameters ε and σ are the well depth and the collision diameter respectively. Out of the two terms above, second term which goes as r^{-6} corresponds to the long-range attractive VdW interactions, while the first term is a r^{-12} repulsion which models the Pauli exclusion principle on overlapping electronic orbitals of the interacting atoms.

VdW interactions originate due to correlations between the polarization of the interacting molecules, charged or neutral. In theory, both kinds of interactions have infinite range and therefore any of the $O(N^2)$ pairs of atoms in a simulation will have non-bonded interaction terms. However, interactions become weak at large distances, so cutoff schemes may be applied to reduce computational overhead.

Another important observation is that the set of neighbours of a given atom does changes slowly over the course of a typical simulation. Therefore, it is computationally favourable to maintain a list of neighbours within a cutoff distance for each atom (the so-called Verlet list). The non-bonded terms of only the atoms in the list are calculated. At a specified frequency, the neighbour list is updated. The list cutoff distance and list update frequency are important parameters that depend on the resources available for computation. They are also inter-dependent: a large cutoff implies that the neighbour list can be updated with a lower frequency.

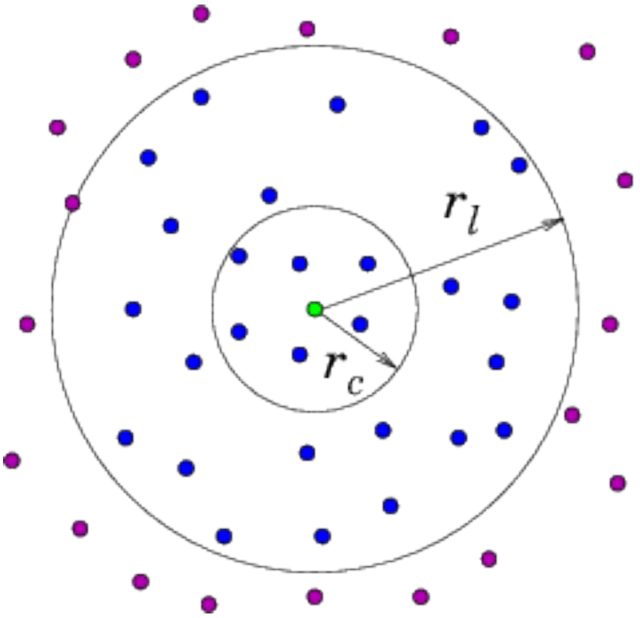

FIGURE 4.1: A depiction of the Verlet list. Source: [5]

The Verlet list is an atom-based method. While overall interaction energies are small at distances near the cutoff, they may consist of a sum of large individual interaction terms. The presence of an atom-based cutoff can then isolate some of these large terms, leading to unphysical energy fluctuations during the simulation. The group-based cutoff can be used to resolve this issue. The system is divided into groups which are units containing a small number of atoms (for example, a water molecule), and cutoffs are applied to entire groups and not to individual atoms within the group. This is especially beneficial if groups are charge-neutral, since the leading charged interaction terms are dipole-dipole, which decay as r^{-3}. Therefore, a group crossing the cutoff does not generate a large fluctuation as before.

While cutoffs are very useful computationally, they come at a physical cost. An abrupt drop to zero of the potential energy is unrealistic, and results in an infinite force term at the cutoff. One way to ensure continuity is to use a constant shift, $V'(r) = V(r) - V(r_c)$ below the cutoff distance r_c. A more sophisticated method is to add a linear term which ensures that the first derivative (i.e. force) is also zero at the cutoff. However, this changes thermodynamic quantities such as potential energy, and it is difficult to correct for this error. Another method is the $switch$ method. A switch function $S(r)$, which decays from 1 to 0 between the so-called cut-on and cutoff distances and is constant everywhere else, is used. The potential $V'(r) = V(r)S(r)$ gets “switched-off” smoothly between the cuton and cutoff distances.

The cut-on and cutoff values must be chosen to accurately capture the dynamics. A cutoff that is too small may ignore important long-range terms, while a very large cutoff could be computationally too expensive.

While the use of cutoffs is usually justified for a fast-decaying Van der Waals potential, the Coulomb potential must be treated more carefully. This is because the Coulomb potential decays with a power that is smaller than the system dimension (see [31]). Methods such as the Ewald summation, which considers the infinite images of charged atoms and sums their contribution in Fourier space, can be used.

4.2.6 Bonded parameters

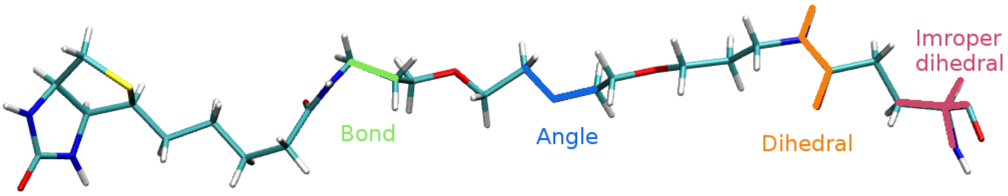

FIGURE 4.2: An illustration showing the types of bonded terms commonly found in force fields.

In biomolecular systems, atoms are covalently bonded to each other via electron cloud sharing. While inherently quantum mechanical, these interactions can be well-modelled by classical terms which are simple functions of the atom positions ([38–40] etc.). Moreover, while the complete energy function depends on the position of all atoms simultaneously, it often suffices to consider only up to four-body interaction terms. Higher orders add to the computational expense but may not significantly improve the accuracy.

1. The first kind of term is *bond stretching* which is typically a simple harmonic function centered around the bond length. For two atoms connected with a bond, with indices i, j:

$$V_{bond}(\vec{r}_{ij}) = \frac{1}{2}k^{b}_{ij}(|\vec{r}_{ij}| - b_{ij})^2 \tag{4.13}$$

2. An angle formed by three consecutive bonded atoms (with indices i, j, k) may undergo vibrations which is capture by the *angle bending* term.

$$V_{angle}(\theta_{ijk}) = \frac{1}{2}k^{\theta}_{ijk}(\theta_{ijk} - \theta^{0}_{ijk})^2 \tag{4.14}$$

3. The first of two four-body terms is the *dihedral* function. This term models the change in energy when the central bond in four consecutive atoms rotates. If the four atoms are $A-B-C-D$, then the rotation of $B-C$ causes a change in the relative alignment of atoms A and D. The parametrization

of dihedral terms is often a crucial task, since the structural dynamics of molecules depends very sensitively on the dihedral parameter values. As the dihedral angle is varied, there may be large energy variations coming from the long-ranged Coulomb interaction (between atoms A and D in the previous example). Dihedral parametrization must therefore be compatible with the assignment of partial charges to the atoms.

Dihedral terms are modelled by sums of sinusoidal functions of the torsion angle. They are typically expressed as sum over the *order* n, which is the frequency of individual sinusoids, e.g.,

$$V_{dihedral}(\phi_{ijkl}) = \sum_n \frac{V_n}{2}\left(1 + \cos\left(n\phi_{ijkl} + \phi_n\right)\right) \tag{4.15}$$

4. Lastly, we have the *improper dihedrals*, which model the out of plane motion of one atom with respect to the other three atoms. The plane of reference is uniquely defined by the three latter atoms (jkl). For an out-of-plane displacement of d, the improper dihedral term is given by

$$V_{improper}(d) = \frac{1}{2}kd^2 \tag{4.16}$$

4.2.7 Boundary conditions

Due to limited computational resources, one cannot simulate a macroscopic system atomistically. Instead, we only simulate a sample which is assumed to be representative of the bulk of the system. Typically, sample sizes are within 10^6 atoms, which is a very small number compared to one mole, or 10^{23}. In order to correctly reproduce bulk properties, the surface effects must be minimized. However, even in a sample of 10^6 atoms, a significant number of atoms ($\sim 10^4$) is at the surface. The presence of a sharp boundary can affect the results, even if the object of interest lies far within the boundaries.

Surface effects can be overcome by using *periodic boundary conditions*. The box is replicated along the box vectors to create an infinite lattice throughout space. The boundaries are then no longer between the system and vacuum but between the system and its periodic image. For any given atom, the motion of any of its images matches the motion of the atom exactly, up to a lateral shift. Atoms are allowed to cross image boundaries, in which case they “reappear” from the opposite wall of the box. Thus, there are no real boundaries in the system and the choice of the box becomes arbitrary (as long as its dimensions and orientation are preserved).

The periodic system is infinite. In order to avoid calculating an infinite number of interaction terms, or double counting interactions due to images, a *minimum image convention* is implemented. In this, an atom only interacts with the closest image of another atom.

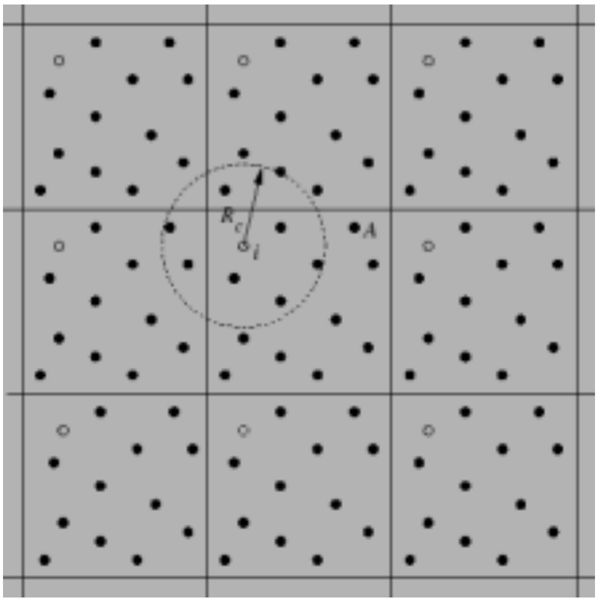

FIGURE 4.3: The minimum image convention. Source: [6]

Periodic boundary conditions can become unphysical when the system itself is non-periodic. A common example of this is a biomolecule in a solvent. While the solvent medium has translational symmetry (although for a discrete solvent, this is only approximately true), the molecule itself is not periodic in space. Any interaction between the molecule and its images is then unphysical, and must be avoided. In such cases, the dimensions of the box must be large enough so that the non-periodic components do not see their images. In other systems which are partially (lipid bilayers) or fully (3-d crystals) periodic, periodic boundary conditions are very helpful.

4.3 Setup

In this part we discuss the computational setup of the system. We simulated using two force fields Chemistry at HARvard Macromolecular Mechanics (CHARMM) 36 [39] and Assisted Model Building with Energy Refinement (AMBER) 14 [38]. For each force field we simulated two different geometries which we refer to as *front* (corresponding to the growing face) and *top* (non-growing face).

4.3.1 Polyglutamine sheet

Amyloid fibrils are known to be rich in β structure, with certain recurring motifs such as β-hairpins, arches, solenoids [3] forming the building blocks of the aggregate. The arrangement that describes a large class of amyloid fibrils is the stacked cross-β sheet geometry. Here, the β-strands (shaped like either hairpins, arches etc.) lie orthogonal to the fibril axis and form a β-sheet via H-bonds that run parallel to the fibril axis. Sheets are stacked along an axis perpendicular to the fibril axis, and the stacking is facilitated by interdigitating side-chains between the sheets. β-chains within a sheet may be aligned in a parallel (all chains oriented along the same direction), or anti-parallel (alternating orientations) manner. PolyQ fibrils are characterized by β-hairpin strands forming anti-parallel β-sheet [2, 27, 41].

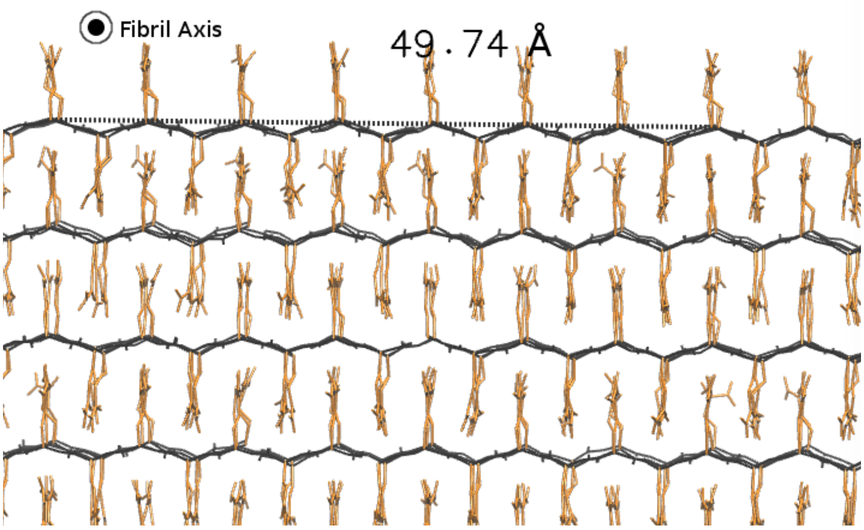

FIGURE 4.4: Interdigitation of side chains in the stacked β-geometry.

The full polyQ fibrillar aggregate (in particular the polyQ stretch thereof) therefore has a stacked beta-hairpin geometry, where each hairpin is a folded polyQ strand.

The β-sheet considered here is a slice across all the stacked hairpins of only the topmost β-chain in each hairpin (see figure 4.5). This system is smaller and hence easier to simulate, and it may capture all the relevant interactions since the biotin will be attached to the topmost layer of the aggregate. Moreover, we have periodic boundary conditions (PBC) with bonding enabled across the boundaries, so that we effectively simulate a stack of quasi-infinite sheets of polyQ immersed in solvent. PBC are also turned on in the direction perpendicular to the sheet, so that there are an infinite number of sheet stacks separated by a water column. However, the box height was chosen to be large, so that neither the sheets nor the fully extended biotinylated GLN residue could interact with the vertical image. Such an interaction would be unphysical.

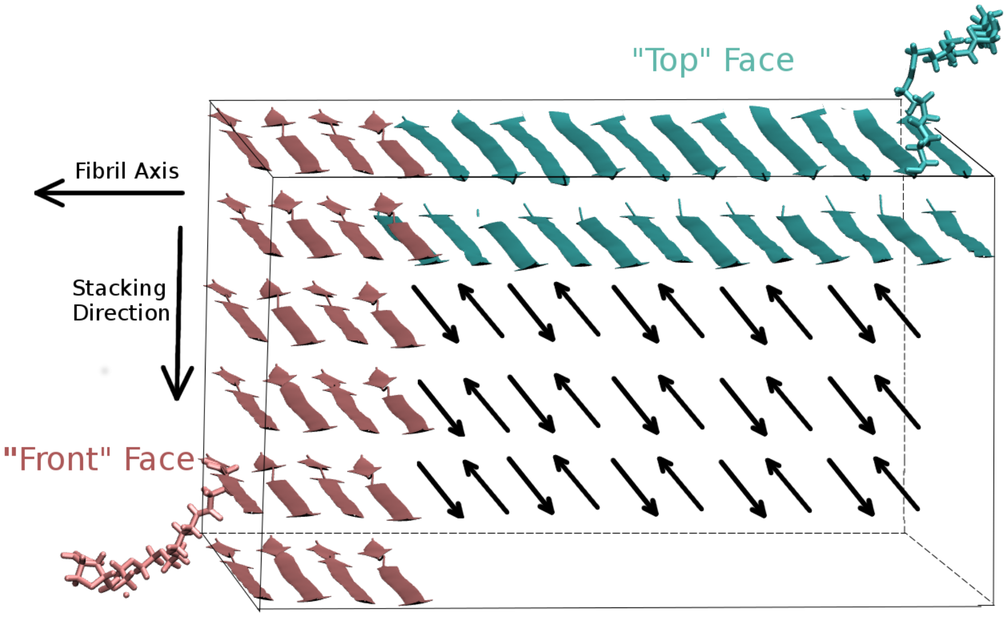

FIGURE 4.5: An illustration of the two simulated systems, top and front. In each system we have a polyQ fibril with antiparallel β chains (red/cyan) in the horizontal direction. The sidechains interdigite in the stacking direction, forming “steric zippers”. The two systems were simulated separately; here they are positioned next to one another only to illustrate their relative orientation. In each case, periodic boundary conditions apply in the plane of the sheet.

4.3.2 Water model

Force fields are equipped with standard libraries for the simulation of water molecules in the solvent medium. Models range in complexity from *implicit solvent*, where the solvent is simply taken to be a continuous medium with a modified dielectric constant, to multi-site water models where, in addition to H and O atoms, other virtual sites are added in order to more accurately simulate the electronic distribution in the molecule. We refer the reader to the following review [42] for details on water models. In our simulations, the 3-point model TIP3P [43–45] was used. This model is fast and captures the essential mechanical features of water (three atoms, correct bond lengths, HOH angle and energies).

4.3.3 Biotinylated glutamine (GLNp)

We denote the biotinylated glutamine residue used in the experiment [17] by or GLNp. GLNp consists of three main components, a glutamine (GLN) whose backbone is free to form peptide bonds, a polyethyleneglycol (PEG) chain (with a few extra Carbons or Nitrogen on the ends) and finally, a biotin molecule. The components are connected via peptide bonds: GLN connects to PEG via the side chain N, while PEG links to the Carboxyl "tail" of biotin. See figure 4.6 below for the structure of GLNp.

The residue is approximately 2 nm in length when fully extended. The C-C dihedrals rotate freely at room temperature, but the linking peptide bonds are initially fixed to the trans- configuration (they stay this way due to the high energy barrier associated with their rotation).

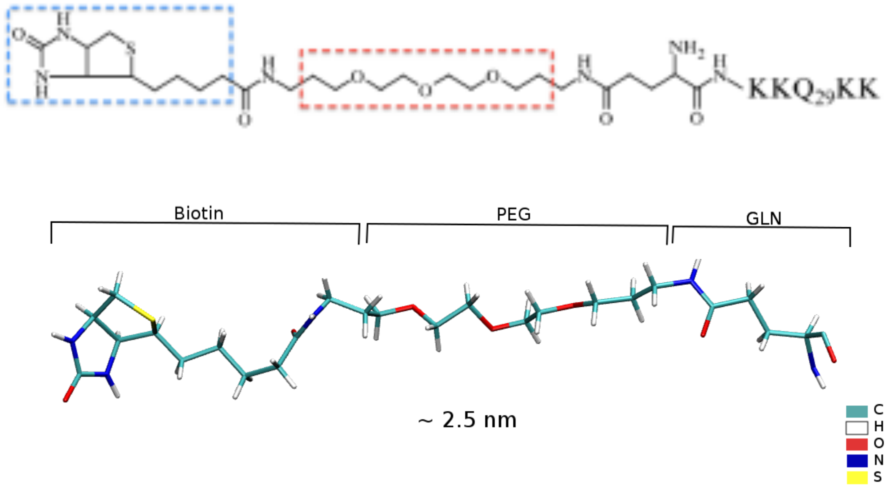

FIGURE 4.6: The schematic structure of a biotinylated polyQ monomer (top), and a molecular view of the GLNp residue used in our simulations (bottom). The blue box demarcates the biotin and the red box the PEG linker. Between PEG and $KKQ_{29}KK$ we have a single Q residue. The biotin-PEG-Q is thus a modified residue (referred to as GLNp) which is attached to the polyQ fibril. Note that Lysine (K) residues are commonly added to the N- and C- termini for *in vitro* studies, since they help solvation of the polypeptides [4]. They were however omitted in our simulations.

4.3.4 Box

For both top and front configurations, the box vectors are determined by the periodic geometry of the sheet in the two planar dimensions and the fully extended length of the the GLNp in the perpendicular dimension.

Top face: In the plane, each β-strand bonds to its periodic image in the direction parallel to the β-strands. This forms a sheet of infinite β-strands aligned anti-parallelly. Neighbouring strands are within H-bonding distance, so that there for each pair of neighbouring residues, there are 2 H-bonds between the side chains and two at the backbone. The β-strand at the edge of the box H-bonds with the periodic image of the strand at the opposite edge.

In the direction perpendicular to the sheet, the box height is determined by the extended length of GLNp plus the thickness of the sheet. This total length is approximately 4 nm in our system, so with a cushion distance of 1 nm, the total box height was set to approximately 5 nm.

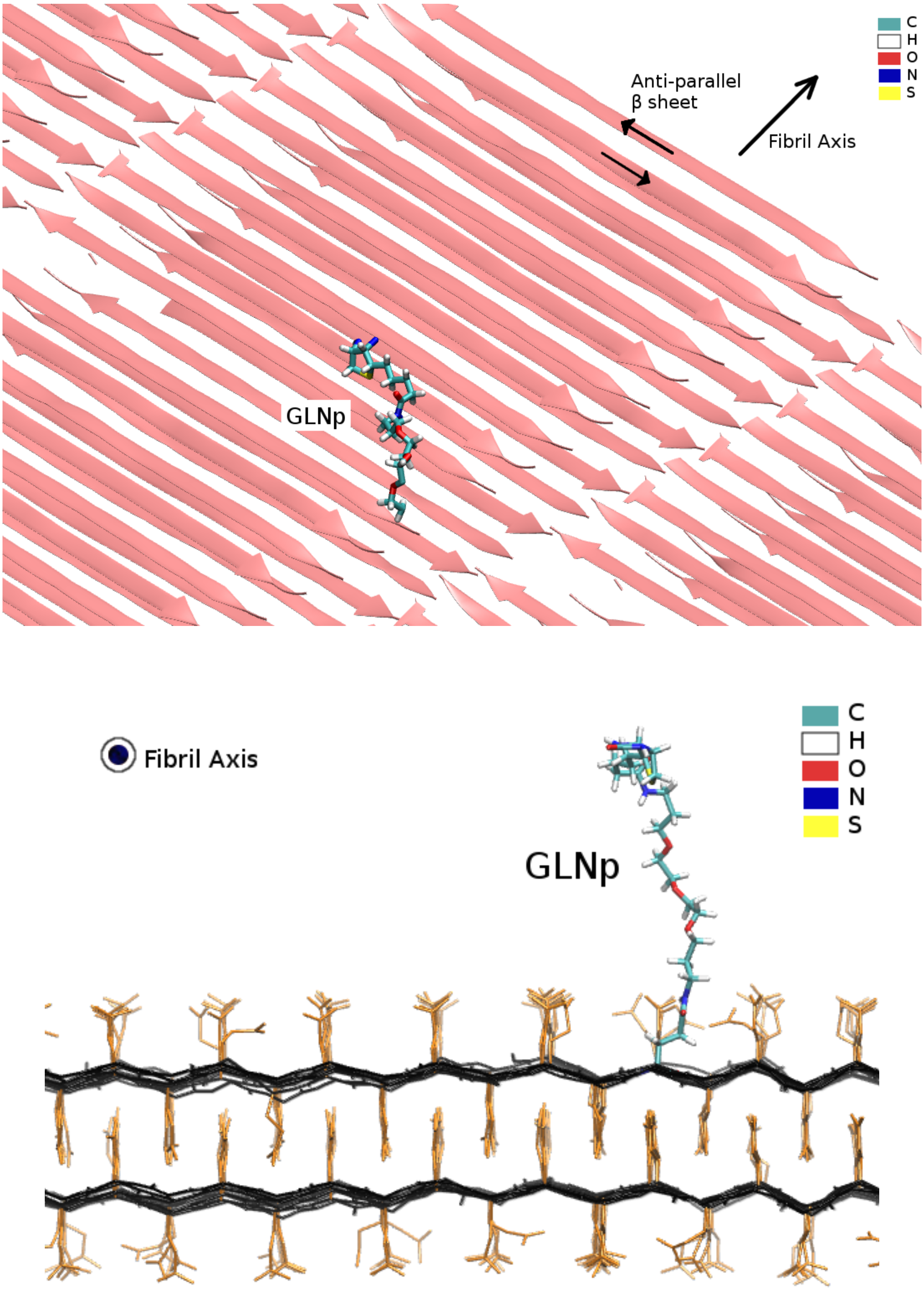

FIGURE 4.7: Top face. On top, we see the anti-parallel alignment of β-strands along the top surface, with the GLNp residue attached. A side view of the same geometry is presented in the bottom graphic, in which side chains are coloured golden while the backbone is coloured black.

Front face: In the front configuration, the sheets are arranged such that the side chains (which point perpendicular to the sheet face) of neighbouring layers interdigitate. The box vector parallel to the side chains must be such that the sheet close to the upper edge meshes with the periodic image of the sheet near the lower edge. In the planar direction parallel to the strands, the box vector must again allow peptide bonding across the box edge. Finally, in the direction perpendicular to the "front face" of the polyQ system, the total box length must exceed the extended GLNp length plus the sheet thickness. Again, a length of 5 nm suffices.

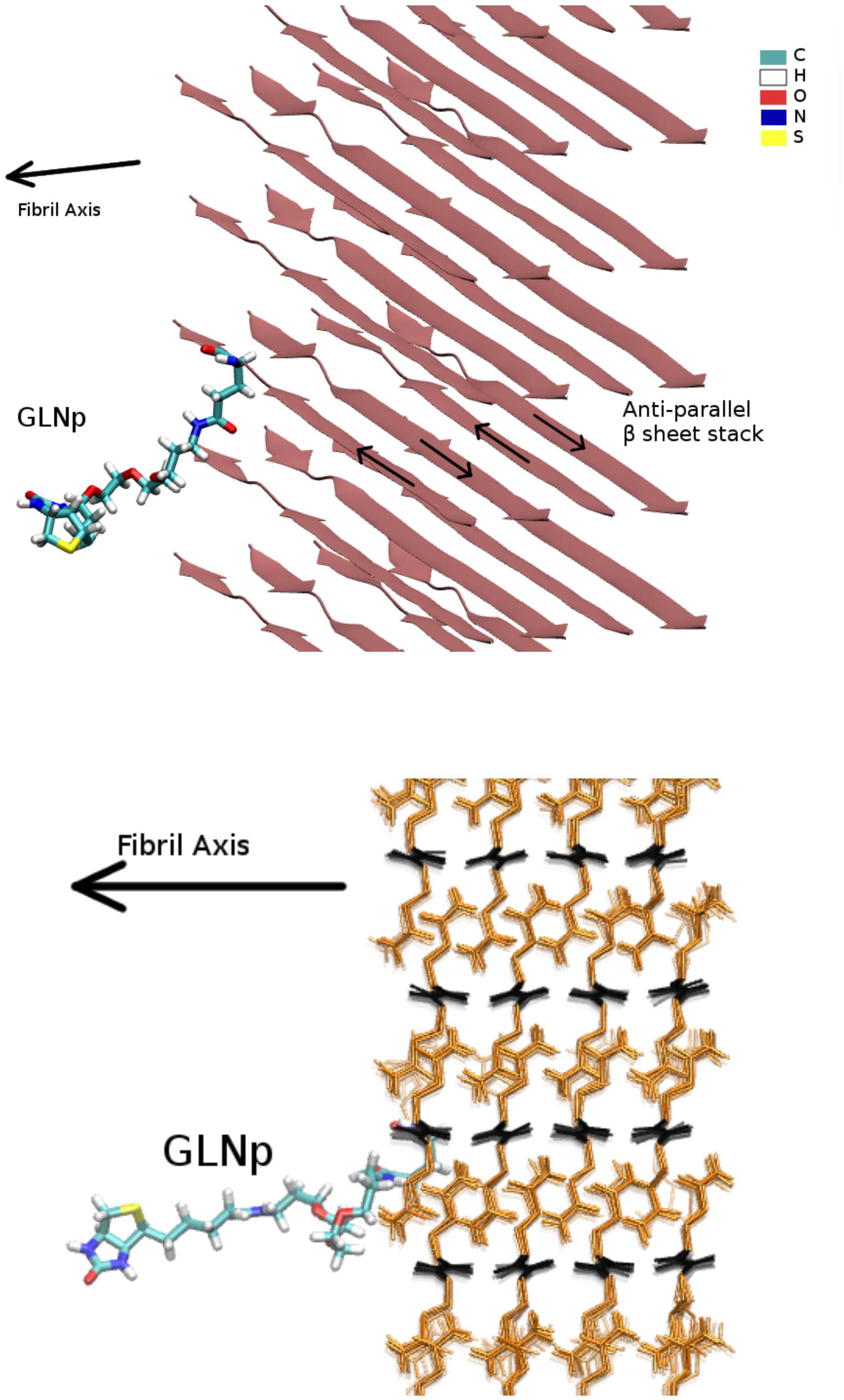

FIGURE 4.8: Front face. On top, we see (a slice of) the anti-parallel β-sheets stacked vertically, with the GLNp residue attached to what we call the "front" face. A side view of the same geometry is presented in the bottom graphic, in which side chains are coloured golden while the backbone is coloured black. The side chains are seen to interdigitate along the stacking direction (top to bottom).

4.4 Parametrization of GLNp

GLNp consists of three main components, a GLN whose backbone is free to form peptide bonds, a PEG chain (with a few extra Carbons or Nitrogen on the ends) and finally, a biotin molecule (see figure 4.6). The components are connected via peptide bonds: GLN connects to PEG via the side chain N, while PEG links to the Carboxyl "tail" of biotin.

Each force field (CHARMM and AMBER) has pre-existing topologies for glutamine. However, the GLNp molecule topology does not exist and has to be created, either from semi-empirical calculations or using the simpler and faster approach of assigning parameters by analogy. In the latter method, the atom types are generated by fitting parts of the molecule to an existing database of chemical groups, and choosing the assignment that shows the biggest match with existing types. For both CHARMM and AMBER, there are freely available programs that assign parameters this way. We discuss them later in this section.

For each force field, we have tried to find a procedure which takes as input a structure file of the molecule, containing a list of atom names and bonds, and returns a Gromacs itp (or top) file corresponding to the molecule and described following the force field. The idea behind this is that the input is a minimal specification of the molecule, while the output is a file that can be used directly to run simulations involving the molecule. Everything in between runs through scripts.

4.4.1 Charge and atom type assignment in AMBER

AMBER makes available on its website (http://ambermd.org/) a freely downloadable collection of tools called AmberTools. It contains packages for a variety of applications, but the ones of interest to us are Link, Edit and Parm (LEaP) and antechamber. In fact, these packages too have broad functions: LEaP is a molecule building program in which one can also make AMBER topologies, while antechamber provides the means to convert between various file types.

Perhaps the most common use of antechamber is to assign charges and atom types and prepare force field descriptor files. These files can then be read by

LEaP to generate AMBER topology and coordinate file (typically .prmtop and .inpcrd respectively).

Antechamber is run with a command of the following form: antechamber -i [INPUT] -fi [INPUT FORMAT] -o [INPUT] -fo [OUTPUT FORMAT] -[OTHER OPTIONS] In our case, the input format is a pdb file, while the output may be a mol2 file. In other options, we must specify: the method to assign charges (-c [METHOD]), and the atom types to use (-at [TYPES]).

Antechamber provides various options to assign charges - RESP, CM1, ESP (Kollman), Gasteiger, AM1-BCC, CM2 and Mulliken. Some of these expect a specific form of input. For example RESP requires an output file from the program Gaussian as input. We use AM1-BCC, which is parametrized to reproduce HF/6-31G* RESP charges, and is also recommended as a fast and efficient method in the AMBER documentation http://ambermd.org/doc12/Amber14.pdf.

The atom types option fixes the naming convention for assigning atom types, so that all atoms are mutually compatible. The documentation claims that General Atomic Force Field (GAFF) and AMBER types are compatible with each other, allowing one to mix the two, and it further recommends GAFF for non-standard residues. Indeed, we use GAFF types for GLNp.

4.4.2 Charge and atom type assignment in CHARMM

CGenFF The CHARMM General Force Field (CGenFF) is program developed primarily by the MacKerell Computational research laboratory at the University of Maryland School of Pharmacy, and the Center for Computational Sciences at the University of Kentucky. It is part of a larger ParamChem project, which additionally includes contributions from the National Center for Supercomputing Applications and the University of Florida. The URL can be found here: https://cgenff.paramchem.org/, with more detailed information available at the MacKerell lab website: http://mackerell.umaryland.edu/~kenno/cgenff/program.php.

CGenFF is a program designed to perform a charge and atom type assignment for the CHARMM force field in a fully automated manner. Complete parametrization of new molecules is often challenging and nuanced due to the complexity of the QM calculations and the consideration of many likely scenarios to find a “best fit” for the molecule. However, it can be very helpful to obtain an initial assignment which may be modified to fit experimental data or *ab initio* (or other) calculations. CGenFF provides precisely this functionality for CHARMM by performing a type and partial charge assignment on the input molecule by *analogy*. That is, CGenFF compares parts of the molecule to a (growing) database of fully parametrized functional groups and finds a maximally satisfactory parameter assignment. In addition, CGenFF provides a penalty score for each atom parametrized, which is a rough indicator of the accuracy of the assignment. According to the MacKerell lab website, a score between 10 and 50 indicate that basic validation is recommended while scores higher than 50 usually imply the need for additional optimization.

The input to CGenFF consists of a mol2 file (or a pdb file with complete bond information). While the input structure need not be geometry optimized, all Hydrogens should be present and each bond should have the correct order. CGenFF returns a CHARMM stream file (extension .str) which contains the parameter assignment and also the penalty scores for each atom. Since the atom types are assigned from existing CHARMM types, many bonded parameters also exist by default in CHARMM. The str file lists any additional bonded parameters to be included.

cgenff_charmm2gmx.py This free program allows the user to produce gromacs topology and structure files from a CHARMM stream file (containing parameters and topologies) input.

4.4.3 Other programs used

Various existing software was used for the preparation of GLNp topologies in both AMBER and CHARMM. In this section, we give an introduction to each tool.

Visual Molecular Dynamics (VMD) Visual Molecular Dynamics (VMD) is a free molecular viewing and analysis software developed by the Beckman Institute of Advance Science and Technology at the University of Illinois at Urbana-Champaign. VMD provides an intuitive GUI in which one can load multiple structure files (as pdb, gro etc.) at a time and apply spatial transformations (translation, rotation, scaling etc.) with mouse commands. MD trajectories can be loaded and viewed as movies in VMD. The URL to the software's website can be found here: http://www.ks.uiuc.edu/Research/vmd/.

VMD provides some extra tools or plugins for various editing and analysis. A plugin called "molefacture" may be used to move atoms individually, modify bonds and add or delete atoms.

UCSF Chimera Like VMD, Chimera is a freely downloadable software for molecule viewing and analysis. It is made available by the Resource for Biocomputing, Visualization, and Informatics (RBVI) at the University of California San Francisco (UCSF), with funding from the National Institute of Health. The URL to the software's website can be found here: https://www.cgl.ucsf.edu/chimera/.

With Chimera one can view molecular structure files in various formats, in particular pdb, gro and mol2. One may also load MD trajectories into Chimera. The main use of this software for our project was in structure editing, which can be done in a very user-friendly and interactive way. Another advantage of Chimera was the ability to store structure files as .pdb or .mol2 with bond information, a feature that is lacking in the other main molecule viewer used, VMD.

Lastly, in Chimera one may also assign Hydrogen bonds and partial charges to a molecule. However, we did not use these features. Instead, the initial structure had the correct number of H atoms per site and partial charge assignment was done for the CHARMM (AMBER) topology was done through CGENFF (antechamber), and later modified by hand.

4.5 Pre-production

We discuss some of the steps (including some trials and errors) in ensuring that the polyQ sheet is stably prepared for long production runs. The basic procedure is sketched below:

- Energy minimization (steepest descent). Put restraints on Cα so that the structure remains planar due to H-bonding. A magnitude of 100000 in each direction is sufficient.
- Vacuum relaxation. At this point the position restraints are lifted and the system is allowed to relax in vacuum with bond constraints. The structure remains intact.
- Solvation. We solvate “gently” (2 ps with 0.2 fs time step), turning the Cα restraints on. No pressure coupling, 2 LINCS iterations. The structure remains intact.
- (Small) production. Now the restraints are again lifted, and the solvated system is run using the “md” (i.e. leapfrog) integrator for 2 ns with a time step of 2 fs. The sheet remains intact, although it does develop some standing modes of low amplitude and frequency.

Initial structure The initial structure was obtained by replicating a single Q20 β-chain obtained from Jan Daldrop. 12 β-chains were placed laterally such that each chain was antiparallel to the next chain. Moreover, the alignment was such that the appropriate O and H atoms of consecutive chains are aligned and close enough to form a H-bond (i.e. < 4 Å apart). The chains were then all in the x-z plane, with the z-axis parallel to the chain backbone. The x-z dimensions were (approximately) 6 nm$\times 6$ nm, while the y dimension was set to 3 nm since this is a sufficient distance to avoid interactions between the sheet and its periodic images.

Vacuum relaxation Early attempts to relax the single layer sheet in vacuum or water without any position restraints resulted in the following irregularities in the structure:

1. The β-chains have a tendency to drift away from one another, creating one or more “rifts” in the sheet across which the H-bonding is too weak.

2. The individual strands may also develop twists independent from one another, breaking the alignment necessary for H-bonding.

3. Lastly, the sheet tends to develop a standing wave out of the plane.

These are undesirable features since the top layer of the fibrillar aggregate is expected to be a flat sheet with tight spacing between neighboring chains. Therefore, we first ensure that the box dimensions (especially the x-dimension) are tight. [should match dimensions in [27]]. Then, we relax the sheet in vacuum with large position restraint forces on the Cα atoms (100000 kJ/mol in each axis). This ensures that the backbones are fixed while the side chains are allowed to relax with respect to each other. Then, we relax again without restraints on Cα. Finally, the sheet structure is intact, with the correct spacing between chains and the side chains aligned vertically out of the plane.

Solvation Once the structure is relaxed in vacuum, the solvent molecules are added. We use TIP3P water. With a time step of 0.2 fs, the sheet+solvent system is run for 2 ps. The Cα restraints are kept on, pressure coupling off, and a total of two LINCS iterations.

Production This is a 2 ns simulation. The β-sheet remains intact throughout this simulation. However, the sheet does not stay flat but stabilizes into a standing wave. With a single periodic sheet this issue may be unavoidable because the β-sheet geometry is expected to have non-zero curvature and the periodic boundary conditions allow for standing waves of certain frequencies.

In order to get a flat geometry, multiple interdigitated layers of the β-hairpin may be required. Given computational resource limitations, we would like to use the minimum number of layers required to simulate the appearance of a real fibril. So, We perform the above production run on a two-layer beta-sheet. If the interdigitated side chains between the sheets provide sufficient rigidity to the system, using two layers will suffice for the main runs.

4.6 Computational resources

We were granted computing time by the HLRN cluster for a total of three quartals (each quartal is a three-month period). In each quartal, we have been allotted 20 kNPL. In total, this allows us to run eight sets of 1 μs on each of the four test cases (a total of 32 μs).

The general-purpose parallel processing system on HLRN is called mpp or **m**assively **p**arallel **p**rocessing system. There are other specialized systems such as smp or **s**ymmetric **m**ulti-**p**rocessing system, which is intended for jobs with high memory and local disk I/O demands. When submitting a job to the cluster, one may request nodes with a certain *node feature*: mpp, ccm, smp etc. based on the job at hand. Alternatively, a *node class* can also be requested Each mpp node consists of 24 individual cores, and there is a maximum of 256 mpp nodes. When one submits a job requesting mpp nodes, a minimum of 4 nodes must be requested. However, in order for Gromacs to parallelize over $4 \times 24 = 96$ cores, the system must have a certain minimum size. Unfortunately, the water box is too small and so we always submit two systems per 4 nodes. This is well within the 75% scaling threshold (which we computed from trial runs), which allows for up to four nodes for a single job.

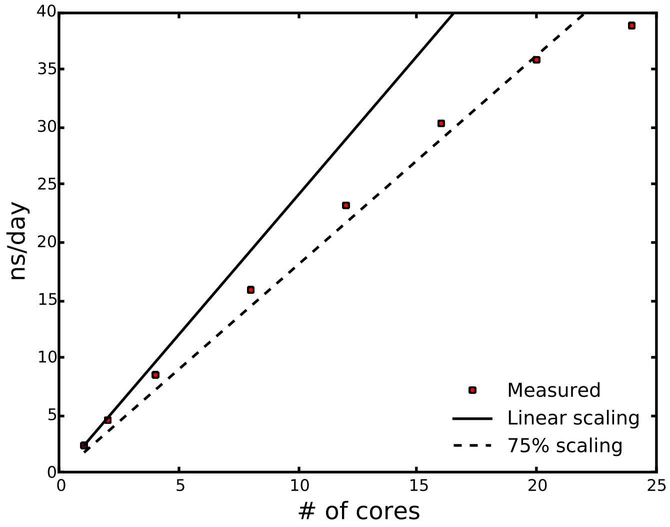

FIGURE 4.9: Scaling of the runtime as a function of the number of cores used. The scaling is sublinear in general, but we generally regard the optimal number of cores to be the value at which we have 75% scaling.

Chapter 5

Results and Discussion

In MD, trajectories are given in the form of large data files which contain information about every atom in the simulation over time with some specified rate of sampling. This information can be selectively extracted and several quantities of interest can be monitored over space and the time duration of the trajectory. Our data was analyzed in several ways, which we discuss in the sections below.

5.1 Trajectories

5.1.1 Maximal sampling

We ran MD on four systems of interest - corresponding to all combinations of geometry and force field, $\{\text{top, front}\} \times \{\text{CHARMM,AMBER}\}$. Recall that “top” here refers to the non-growing faces of the fibril while “front” refers to the growing face. For each case, the computational resources from HLRN allowed us to run eight MD simulations of 1 μs each. Ideally, it should be the case that the eight trajectories are mutually uncorrelated, and explore independent regions of configuration space. This would hypothetically correspond to an eight-fold enhancement in the sampling. However, this is challenging to achieve in practice, since one has to certify that the eight initial configurations (and their initial velocities) are all mutually independent.

One solution may be to generate configurations uniformly at random, and then set up eight random configurations of GLNp and the sheet. The velocities may be randomly generated from the Maxwell distribution. However, this is tedious in practice since each configuration has to be generated by hand.

On the other hand, one could run a small trajectory starting from an initial configuration and pick time points at random - would the configurations at these time points be random? We argue that this is unlikely to be the case, since this system is expected to have a long memory time, possibly on the order of 10^2 ns. Therefore, there is the danger that the eight trajectories would be correlated with one another, up to a time shift.

Is it possible to generate varied trajectories starting from the same initial configuration on all runs? This would overcome the practical issue of generating eight uncorrelated configurations, and would also make the runs more comparable, due to the identical initial state. As an extreme example, consider what happens when all eight initial configurations have identical positions and velocities. In this case, each trajectory will be exactly identical have we do not gain any enhancement in sampling. However, if the velocities are randomly generated, then the configurations have different instantaneous trajectories at $t = 0$. In order for these trajectories to diverge, we should choose an initial configuration which is as close to an unstable equilibrium (or a "crest") in configuration space as possible. Moreover, the configuration space near this point should look isotropic - in other words, there should be no preferred direction. This way, we can expect that random initial velocities will lead to maximally divergent final trajectories, leading to enhanced sampling.

With this in mind, we choose an initial configuration where the GLNp residue is fully extended and perpendicular to the sheet. In this configuration, GLNp is extended furthest from the sheet, and so we may expect that the local energy landscape is least affected by the interaction of the GLNp with the sheet. Moreover, it is unlikely that this configuration is stable, since all previous trial runs have demonstrated that the residue tends to "fall" onto the sheet along some azimuthal direction. For each run, we started with the same fully extended configuration, assigned initial velocities randomly from the Maxwell distribution, and then relaxed the system in NVT and NPT.

5.1.2 Observations and discussion

The relaxed systems were evolved for 1 μs per run. Various interesting features emerge from the trajectory of GLNp atop the sheet, which we will discuss in this section. The trajectories may be visualized using a program such as VMD. In Figs. 5.1 and 5.2 below, snapshots of the attached GLNp residue have been shown for both the front and top faces at various times in the trajectory. The times have been chosen to span the timescales sampled in this simulation, from 1 ns up to 1000 ns (i.e. 1 μs). While these figures do not represent all trajectories, they reflect features that are typical to all other trajectories. We refer the reader to these figures for the rest of the discussion.

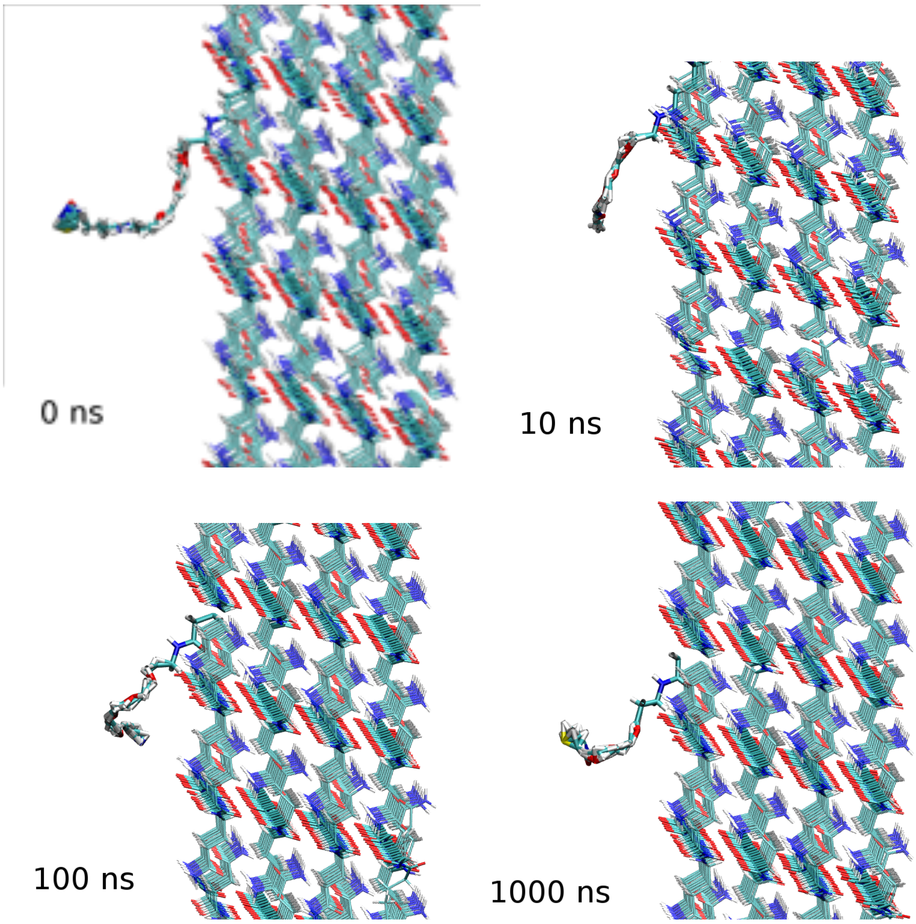

FIGURE 5.1: GLNp configuration at various times when attached to the "front" face of the fibril.

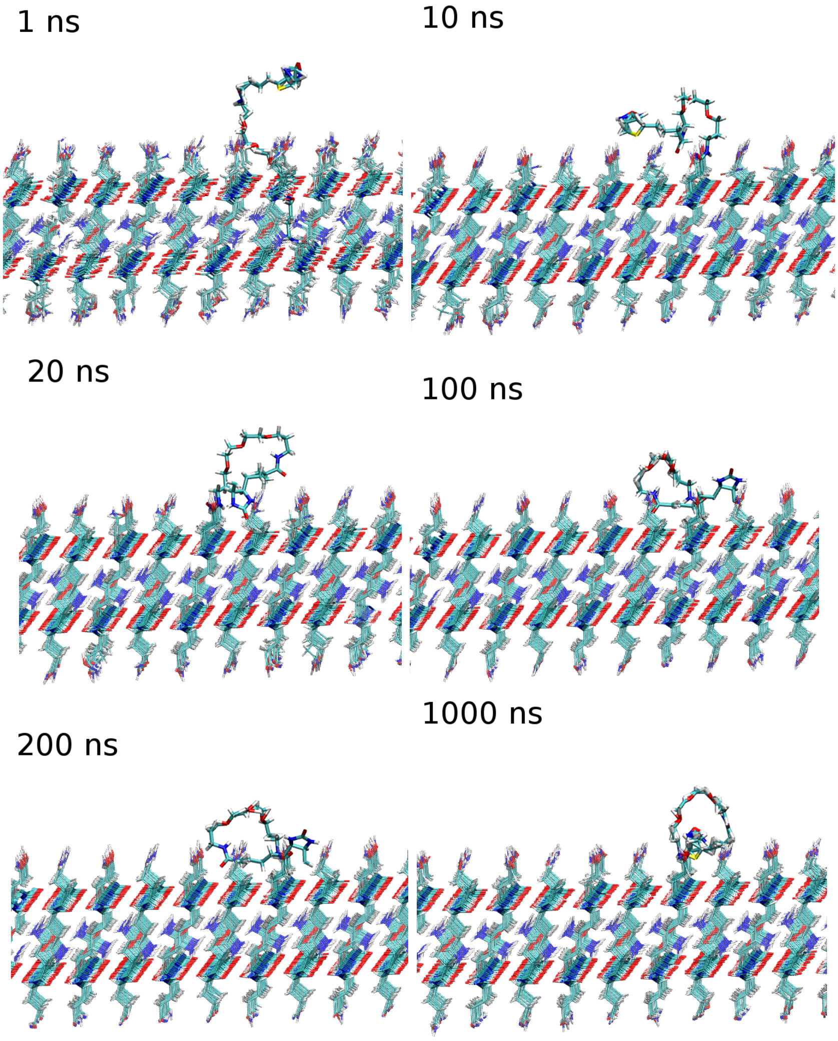

FIGURE 5.2: GLNp configuration at various times when attached to the "top" face of the fibril.

The GLNp head group (consisting of two adjoining pentacyclic rings) has a strong affinity to the surface of the sheet due to the high availability of H-bonding sites. This is seen in both the front and top (5.1 and 5.2 resp.), where the biotin head group has a tendency to remain close to the sheet surface (as early as 10 ns for the top surface). However, the configurations of GLNp are different when attached to the top and front surfaces. On the front face, the GLNp tends to "lay flat" against the surface such that it the residue is extended and parallel to the front face.

On the top however, the situation looks different. we see that the biotin head has a tendency to get embedded *between* the GLN sidechains emerging vertically from the surface. The head is seen to lock in these positions for over ~ tens of ns at a time, and is H-bonded with most of the neighbouring sidechains. It is also sterically locked in place, which adds further stability to these configurations. Due to the rotational degrees of freedom available to the single C-C bonds along the GLNp chain, the GLNp head is able to embed itself into the sheet at any point within the fully extended radius of GLNp (~ 2.5 nm). Thus, the GLNp head tends to have periods of free evolution in the solvent, then "finds" a pocket between side chains on the surface, and then spends a long time embedded in the same pocket. The rest of the GLNp chain explores available "loop" configurations during such periods when the head is locked.

This embedding mechanism was not seen for any of the frontal trajectories, and as such, seems to be a highly unlikely configuration. We may attribute this to the face that the front surface is more closely packed than the top face. This is because the front face is made up of interdigitated, stacked β-chains. The β-chains are also H-bonded (anti-parallelly) with interior chains; moreover, the side chains are parallel to the surface. Due to this rigid structure, the biotin head is perhaps unable to become embedded into the surface. On the top face however, the surface side chains are perpendicular to the sheet. They possess more freedom than the interdigitated side chains - quite like hairs on a surface. Thus, the biotin head has a lower barrier to sampling the embedded states.

Purely from the trajectories, one can guess that biotin attached to growing faces (i.e. the "front" face in the simulations) will be more available for binding than biotin on non-growing (i.e. "top") faces. The observations on the trajectories also support the thinking that the preferential binding seen in experiment, [4, 17], comes about due to the geometric differences between the surfaces of the nearly crystalline polyQ fibrils. Next, we present some quantitative analysis on these differences.

5.2 Density profiles

In the earlier section, we observed that the GLNp head group behaves differently on the top and front faces. In particular, it has a tendency to become embedded between side chains on the top face frequently and for long time periods. Therefore, we expect that the head group is on average closer to the sheet in top configurations than in front configurations. The information most relevant to us is then the spatial distribution of the GLNp atoms over the sheet surface. For a well-sampled run, the density profile along any reaction coordinate samples the free energy landscape along the reaction coordinate. We can compute the free energy profile from the density profile, up to an additive constant.

In experiments, it is the double-ring head group of biotin that forms a complex with the tagged streptavidin. Therefore, we are interested in probing the availability of the head group at the front and the top face. Since streptravidin is a large molecule, we expect that it would not be able to bind to the biotin heads that are close to the polyQ fibril surface.

We selected the biotin oxygen (see Fig. 4.6) as the atom representing the head. There are a few reasons for this choice:

- The oxygen atom is a reactive center of biotin and is involved in streptavidin binding (along with the neighbouring N atoms and the S atom on the opposing ring),
- The oxygen is the most “exterior” heavy atom in the ring, extending out through a carbonyl atoms, and
- It is more intuitive to track a single atom than the center of mass of a group of atoms.

Therefore, we extract the coordinates of the biotin oxygen and also the backbone Cα of the GLNp, which is taken as a reference point for measuring the end-to-end vector (see Fig. 5.3 below). It is reasonable to consider a single Cα as the reference, since the underlying sheet is very stable, and backbone atoms have very little motion.

Further, we write out the box vectors at each time point, since the box dimensions do not remain constant throughout the simulation. Since the box vectors determine the spatial periodicity, the knowledge of box dimensions at each time point allows us to accurately correct for any boundary crossovers. Subtracting the coordinates of the Cα from the O (after correcting for boundary crossovers), we derive spatial profiles of the O atom with respect to the underlying sheet.

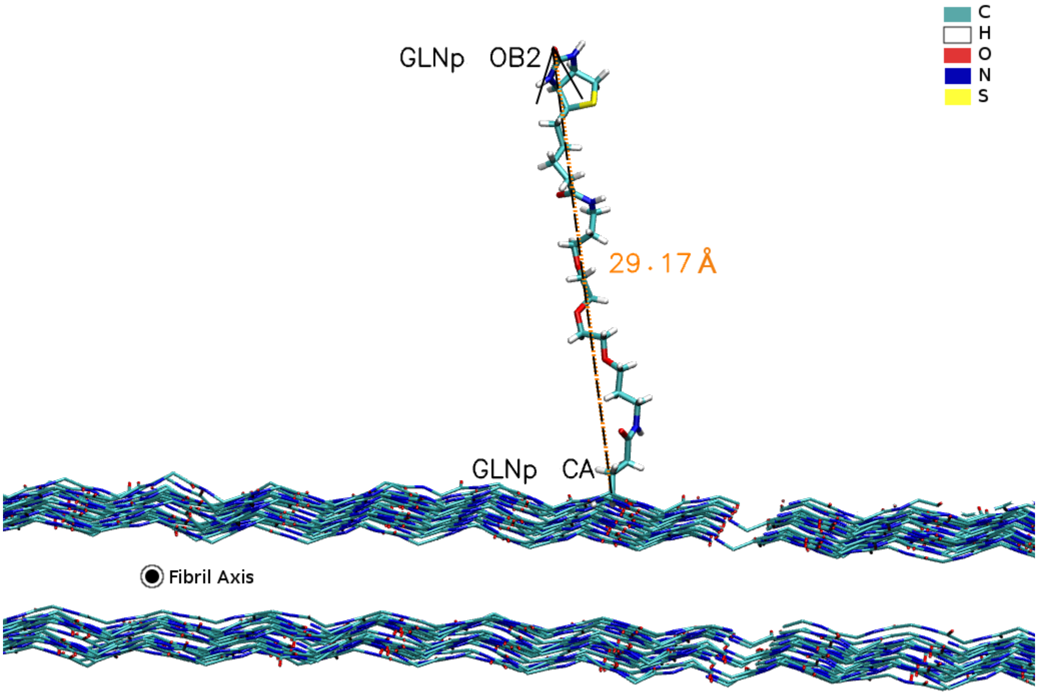

FIGURE 5.3: An illustration of the end-to-end vector of GLNp

First (and perhaps most important) is the density profile of the O atom vs. displacement perpendicular to the sheet (which we henceforth denote as y). The distance of the biotin head from the sheet is an important factor in the availability to streptavidin, which is a large complex with a diameter of ~ 5 nm [46]. Further, we note that for the top face, the side chains extend about 0.45 nm above the backbone in the y-direction. For a well-sampled trajectory, the y-profile is very telling of the interaction between biotin and the sheet surface. In Fig. 5.4 below, we present the aggregate density profiles for all four cases (front/top and AMBER/CHARMM). For each case, the profile shown is an average over all eight independent runs, and the error bars are taken to be the point-wise standard deviations.

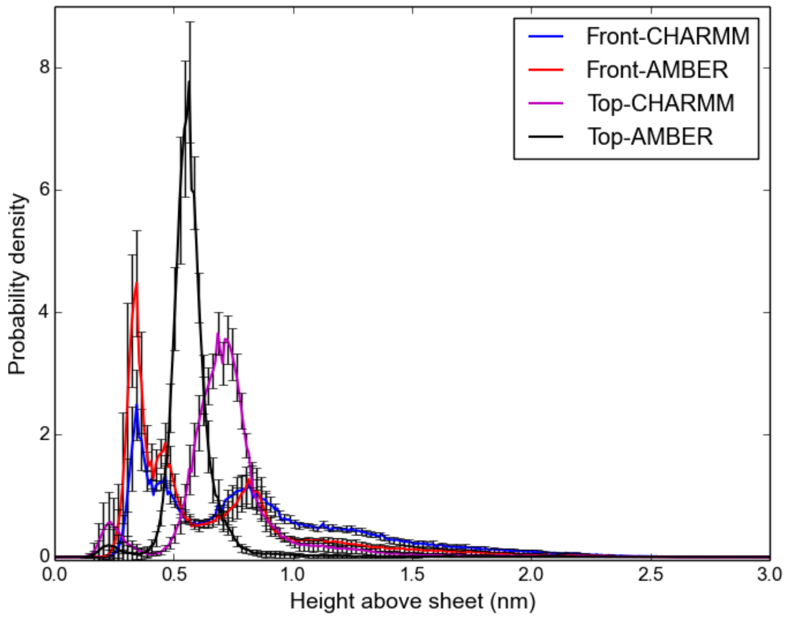

FIGURE 5.4: Plots of (normalized) density vs. vertical displacement of the biotin head group above the sheet.

Additionally, we may look at the *cumulative* probability as a function of height above the sheet. In other words, the probability that the head lies below a certain height y, as a function of y. The aggregate profiles for the four cases (over all eight individual runs in each case) is given in Fig. 5.5 below.

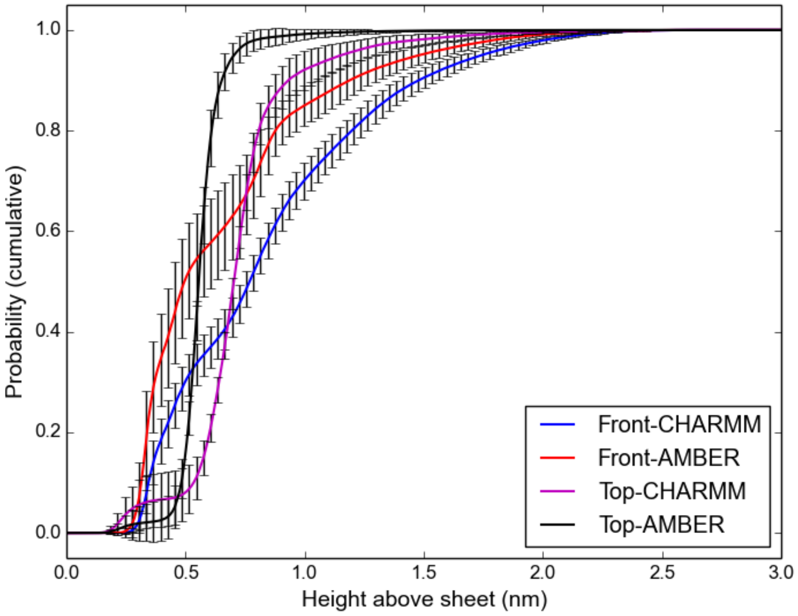

FIGURE 5.5: Plots of (normalized) cumulative density vs. vertical displacement of the biotin head group above the sheet.

The cumulative density profiles provide a sense for how sharply the density of the biotin head group drops off with height. In the case of the top geometries, there is effectively a cutoff at ~ 0.5 nm above the sheet (i.e. backbone CA). This is interesting, since the height of the GLN side chains above the sheet is approximately 0.45 nm (see Fig. 5.12). Thus, *the GLNp head is primarily to be found within a height which is comparable to the height of the side chains on the top face.*

Next, we discuss each of the four cases in greater detail, commenting on similarities and differences between them. Since two force fields have been used to simulate the data, our strategy will be to identify common features seen in the corresponding CHARMM and AMBER simulations. We argue that similar features are likely to correspond to the intrinsic features of our system. Differences in the two force fields are also interesting, and could shed light on numerical differences between the force fields, inaccuracies, or other artifacts.

5.2.1 Front-CHARMM

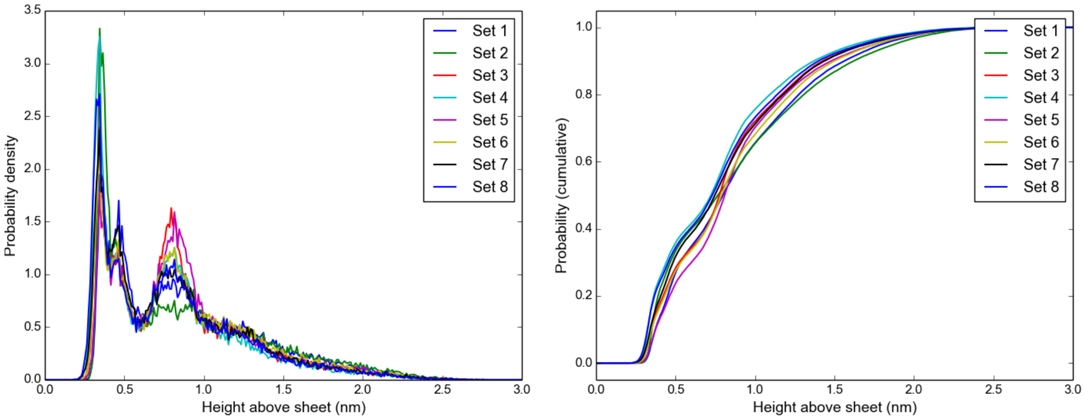

FIGURE 5.6: Density and cumulative density profiles for the Front-CHARMM case.

The Front-CHARMM profiles in Figs. 5.1 and 5.2 (colored blue) show that the GLNp head is primarily found between separations of 0.3 to 1 nm from the sheet surface. Note that his separation is measured from the backbone CA.

In Fig. 5.6 above, we also show the individual density profiles for the eight runs. The sets show, to varying degrees, that the biotin head has a preference for three heights above the sheet: 0.3, 0.5 and 0.8 nm. We may guess that the first height corresponds to H-bonding distance from the sheet. It is likely that the other two maxima (since they are seen in each independent run) each correspond to the same class of metastable configurations. Note that the frequency drops to zero at around 2.5 nm as this is the fully extended length of GLNp.

5.2.2 Front-AMBER

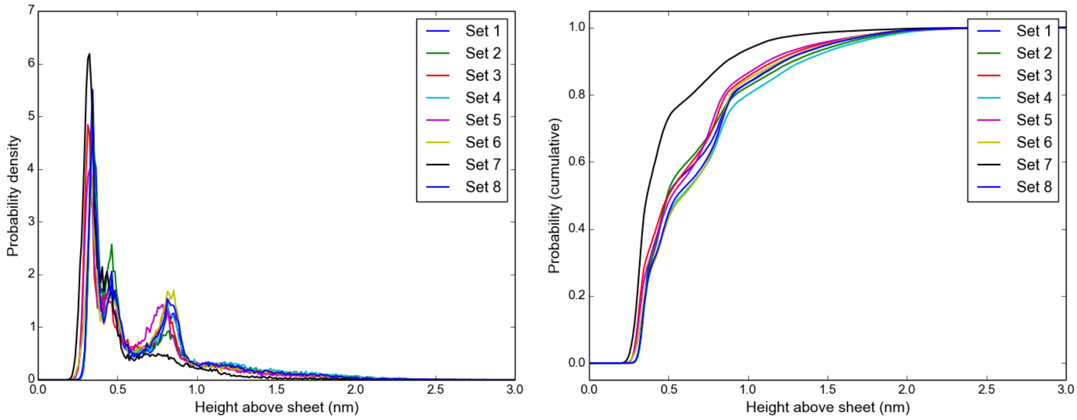

FIGURE 5.7: Density and cumulative density profiles for the Front-AMBER case.

In the above figure (Fig. 5.7, we show individual density profiles for the AMBER runs (front configuration). Now, we may begin to compare the results of the two force fields used in this analysis.

The first similarity between the CHARMM (5.6) and AMBER (5.7) profiles is the presence of three maxima which agree in their locations. Moreover, they agree qualitatively: the maxima at 0.3 nm are sharp and the tallest, followed by the 0.5 nm maximum and then the 0.8 nm peak (which is also the broadest in both force fields). Therefore, we may conclude that the three peaks correspond to a physical feature in the system, namely the existence of metastable states close to the sheet.

Now, we point out some differences between the CHARMM and AMBER profile. First, the density in AMBER is, on the whole, shifted closer to the sheet as compared to the density in CHARMM. This is readily apparent when one compares the tails of the density distributions (in the range of 2-2.5 nm). This is a difference that will be recur, and it is likely attributable to the differences in force field parametrization (especially non-bonded parameters). Essentially, this difference leads to a greater affinity between the biotin and the sheet in the case of AMBER. Secondly, the peaks in AMBER are also seen to be sharper than the CHARMM peaks. One possible way for this difference to arise is that in AMBER, energy minima are deeper than in CHARMM. Since the temperature (and hence energy scale $k_B T$) is the same for simulations in both force fields, the AMBER energy minima could be more restrictive than the CHARMM mimima, leading to sharper features.

In CHARMM, all eight trajectories are qualitatively similar. Since we have chosen the initial configurations in a way that minimizes correlations between trajectories, we may argue that the similarities in the features point to intrinsic features of the biotin-surface interaction. We may also conclude that the time sampling exceeds the memory time of the system, which allows us to make general conclusions about the free energy landscape. In the current front-AMBER set of runs, all but one of the runs are "similar". Set 7 (coloured black in the figures) is an outlier: it effectively has only one peak at 0.3 nm. Upon viewing the trajectory in VMD, it was seen that in this trajectory GLNp spent a many periods of time in one configuration. While this configuration was typical of other front runs too, it was not as dominant in any of the sets except set 7. We believe that this behaviour is not a numerical error. In fact, this deviation could indicate that the memory time of the system is not significantly shorter than the total run time of 1 μs, and that greater sampling times may be needed in order to see better agreement between the independent runs.

5.2.3 Top-CHARMM

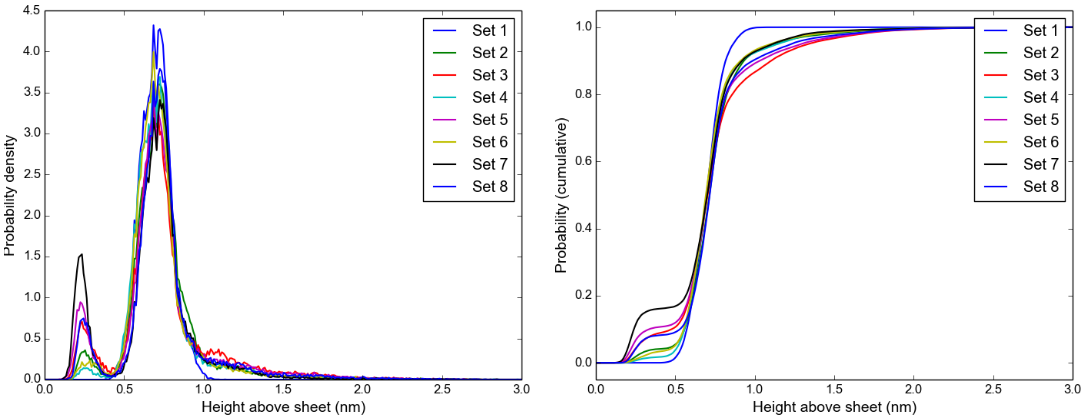

FIGURE 5.8: Density and cumulative density profiles for the Top-CHARMM case.

Now, we turn to simulations of GLNp on the top face. Here, we first comment that while the height above sheet has been measured from the backbone CA, the top face also has side chains which extend to about 0.45 nm above the sheet on average. This should be kept in mind when analyzing the density profiles. For instance, when the head group is below 0.45 nm, it is effectively embedded in the sheet.

In Fig. 5.8, we show the individual density profiles, like we did for the front runs. In the top runs, we see that there are two peaks, the smaller one at 0.25 nm and a larger, broad peak around 0.75 nm. Note that the side chain height is approximately 0.5 nm so the two peaks could correspond to optimal H-bonding distances from the backbone (small peak) or the side chain (large peak). Further, the smaller peak corresponds to "embedded" configurations, i.e. ones where the biotin head group lies between side chains.

Finally, we remark that, as compared to the front runs, the top runs have greater disagreement among themselves. This is apparent when the cumulative density profiles are compared for the top and front runs. One reason for this could be that the top face has many stable, embedded configurations that are energetically less accessible (due to steric hindrance etc.), and therefore require greater sampling. Indeed, the greatest disagreement between the runs occurs at ~ 0.3 nm (Fig. 5.8, which corresponds to the embedded height of GLNp.

5.2.4 Top-AMBER

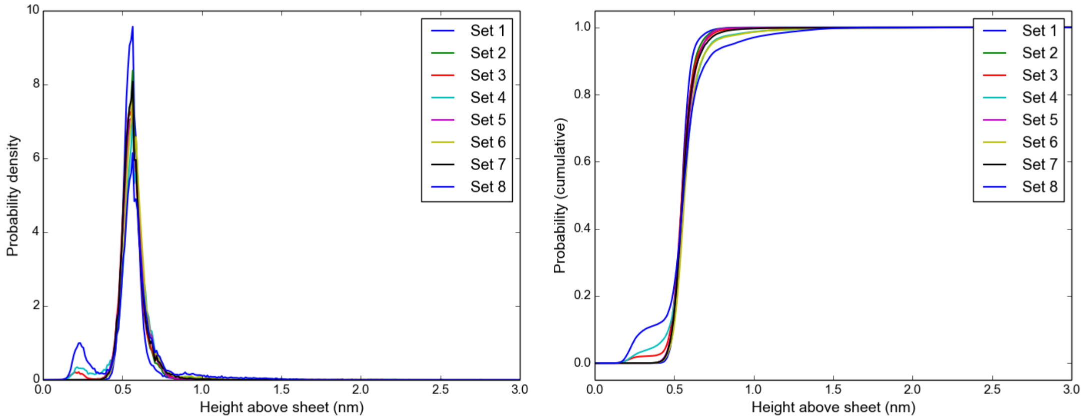

FIGURE 5.9: Density and cumulative density profiles for the Top-AMBER case.

Finally, we look at the Top-AMBER density profiles in Fig. 5.9 above. First, we see that, like top-CHARMM, top-AMBER also has two peaks in the density of

GLNp above the sheet. However, the difference between force fields that were seen earlier in section 5.2.2 reappear: the AMBER densities are shifted closer to the surface, and the peaks are sharper, than in CHARMM.

On the other hand, both AMBER and CHARMM have divergent cumulative density profiles at the first maxima of ~ 0.3 nm. This allows to conclude with greater confidence that more sampling times would improve the sampling of embedded configurations appreciably.

5.3 Potentials of mean force (PMF)

From the above density profiles, we may wish to extract information about the energy landscape as seen by the biotin head group. Perhaps the most straightforward means to accomplish this is to look at the Potential of Mean Force (PMF).

For a system of N coordinates $q_1, q_2, \ldots, q_N$ (e.g. atom locations), a given force field induces a potential function $V(q_1, q_2, \ldots, q_N)$ under which the system evolves. While the complete energy landscape is a non-trivial higher dimensional function, we may be interested in the energy only along one reaction coordinate, say q_1. (For simplicity, we assume the reaction coordinate to be one of the base coordinates). Then, the mean force along this reaction coordinate is an average of the negative gradient of V, taken over all possible configurations of $q_2, \ldots, q_N$.

This mean force is the gradient of some effective potential function - this is called the potential of mean force (PMF). Denoting the PMF by $w(\vec{q})$, we may write

$$-\nabla w(q_1) = \frac{\int\limits_{\vec{q}} e^{-\beta V} \left(-\nabla V(\vec{q})\right) \mathrm{d}q_2 \mathrm{d}q_3 \ldots \mathrm{d}q_N}{\int\limits_{\vec{q}} e^{-\beta V} \mathrm{d}q_2 \mathrm{d}q_3 \ldots \mathrm{d}q_N}$$

In practice, the PMF may be calculated using density distribution along a reaction coordinate as follows.

$$w(q_1) = -k_B T \ln\left(\langle \rho(q_1) \rangle_{\{q_i, i \neq 1\}}\right) + C$$

where ρ is a (normalized) density distribution, and C is an arbitrary constant. Given infinite sampling time, the PMF along the coordinate will be the free energy along the coordinate (up to an arbitrary constant shift). For finite sampling, one has to first ensure that the sampling is sufficient for the PMF to be used as a substitute for free energy.

In our simulations, there was considerable agreement between independent runs of the same geometry and force field (with some exceptions, e.g. see section 5.2.2). This indicates that the trajectories sampled the reaction coordinate (height above the sheet in our case) well, and that a PMF may indeed be representative of the actual energy landscape. Therefore, we generate PMF distributions from the density profiles presented in the previous section. These are given in Fig. 5.10 below.

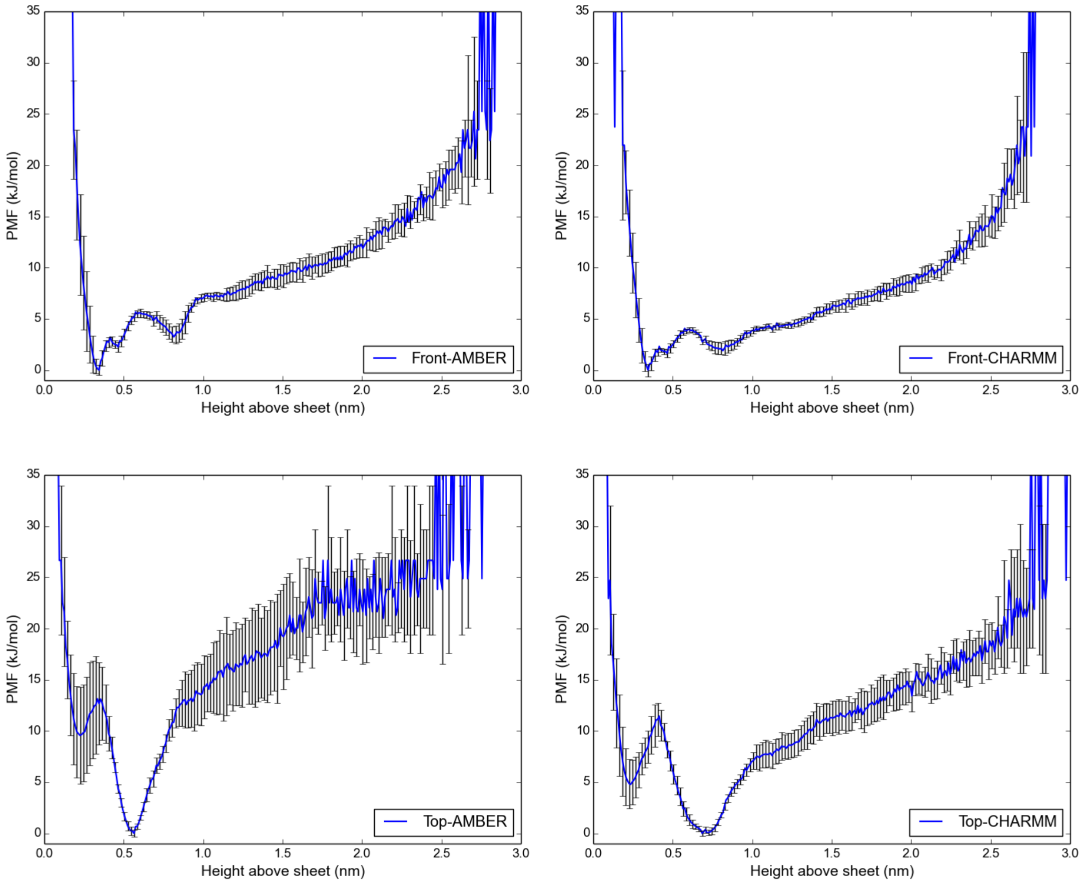

FIGURE 5.10: Potential of mean force, as computed from the density profile of the biotin head group as a function of height above the sheet surface.

First, a comment about sampling: The "trustworthiness" of the PMF at a given point in the reaction coordinate depends on how well the sampling is at that point. This is well-illustrated by the error bars in the PMF: they are large at points where the density is very small. At distances of ~ 2.5 nm from the sheet, the PMF seems to diverge to infinity. The reason for this is clear, since the density goes to zero as the GLNp is fully extended. However, it is unphysical for the PMF to increase to infinity - in practice, there is a large energy barrier corresponding to the binding energy of the GLNp molecule. On the other hand, the system is well-sampled near the sheet, which is the region of interest.

Since the PMF profiles are a direct derivative of the density profiles, we refer the reader to the previous section, 5.2, for a discussion of the features seen in the plots. For example, peaks in the density profile correspond to energy minima in the PMF. Now, we may compare the energy minima by quantitatively in terms of their depth. We must be careful when comparing absolute energy for different runs (especially for runs of different geometry), since the reaction coordinates are different in each case. Still, the relative depth of the analogous energy minima in each case can provide useful information.

Earlier, it was mentioned that AMBER density profiles may have sharper peaks due to deeper energy minima. This is more readily apparent in the PMF distribution. We see that energy wells have depths ranging from $2-3$ kJ/mol to ~ 12 kJ/mol. These are realistic energy scales - for comparison, the H-bonding energy in biomolecular systems tends to be ~ 10 kJ/mol.

We end this section with the remark that, while the PMF distribution does not replicate the exact free energy landscape, it will be sufficient to study kinetics of the biotin-surface interaction, timescales of binding, etc. We will carry out this analysis later in section 5.5.

5.4 Time autocorrelation

Useful information can also be extracted in the form of the time autocorrelation function of the end-to-end vector of GLNp. Mathematically, the time autocorrelation function $R(t)$ may be defined as:

$$R(t) := \lim_{T\to\infty} \frac{1}{T-t} \int\limits_{\tau=0}^{T-t} \vec{v}\,(\tau) \cdot \vec{v}\,(\tau+t)\,\mathrm{d}\tau \tag{5.1}$$

Thus, the autocorrelation function provides information about the memory time of a system. In particular, any time periodicity appears as a peak in the autocorrelation function. For a free residue, one generally sees a fast decay over short times, while over very long times the correlation is expected to be close to zero as the residue samples all of its conformations. Deviation from this behaviour can give interesting insight into the effect of an underlying sheet. For instance, the steady state values of many of the autocorrelation curves in Fig. 5.11 are not at 0 but closer to 0.4. If we simply assume that the e2e vector randomly samples the hemisphere (due to steric hindrance by the sheet), then the average of the dot product between two unit vectors turns out to be $\pi/8 = 0.39$ which in fact matches the data quite well. On the other hand, some curves in Fig. 5.11 are quite peculiar over long times, and give us an idea of the long memory time of this system.

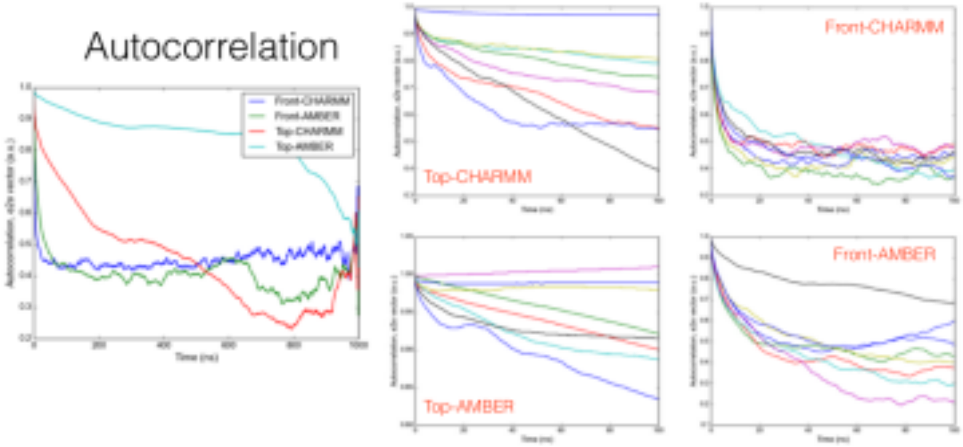

FIGURE 5.11: Time autocorrelation plots of the GLNp end-to-end vector. The figure on the left plots means of the four corresponding figures on the right. On the right we have autocorrelations for each run, categorised by the force field and the geometry of the sheet. The time (on the plots on the right) runs from 0 to 100 ns, while it is 0 to 1000 ns on the left figure.

In the above, we have normalized the autocorrelation functions by the function value at $t = 0$ i.e. we set $R(0) = 1$. We see that for the top face, the correlation times are long, on the orders of many hundreds of ns, while the front face has a memory time under hundred ns.

Finally, we discuss a simple model for analyzing the binding of streptavidin to biotin attached to a fibril surface.

5.5 Markov model

Suppose we are given the trajectory of GLNp attached to the underlying fibril (in either the front or top configuration). Denote the total time of the trajectory by T, and the number of time steps by N. The step size $\mathrm{d}t$ is constant and satisfies $N\mathrm{d}t = T$.

We are interested in the diffusion timescale of biotin-streptavidin binding. Unlike in the free case, the binding availability of biotin is limited by several environmental factors, most significantly the presence of an underlying beta sheet. A large molecule such as streptavidin may not be able to bind when the biotin is too close to the sheet. Therefore, we need to consider not the free binding timescale, but the binding timescale of *attached* biotin with streptavidin. Here, we discuss one way to calculate this.

We make three assumptions about the system:

1. that it is *Markovian*, i.e. the binding rate does not depend on the history of the system, and,
2. that the system obeys a first-order rate equation, and
3. that the reaction is one-way, i.e. bound streptavidin does not unbind.

Suppose we have a some number of attached GLNp molecules A_0 in a single homogeneous solution which contains a large number of streptavidin, so that at any given time the concentration of streptavidin seen by any two GLNp is the same and constant over time. Let $A(t)$ denote the number of free GLNp, i.e. those that have not yet been bound to streptavidin, at time t. Then $A(0) = A_0$, and the binding rate equation may be written as

$$\frac{\mathrm{d}A(t)}{\mathrm{d}t} = -r(c, \mathbf{x}(t))A(t) \tag{5.2}$$

where the rate r is a function of the streptavidin concentration c and the spatial configuration of biotins at time t, which we denote by $\mathbf{x}(t)$. We can make the above equation independent of the number of biotins by dividing both sides by A_0 and observing that $P(t) := A(t)/A_0$ is the *probability* that biotin is free at time t. So,

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = -r(c, \mathbf{x}(t))P(t)$$
$$\implies P(T) = \exp\left(-\int_0^T r(c, \mathbf{x}(t))\mathrm{d}t\right)$$

The binding probability at time t is then given by $Q(t) := 1 - P(t)$.

Ideally, if the function r is specified for any configuration of the biotin molecules with respect to the sheet, then $P(T)$ and hence the binding probability $Q(T)$ can be computed by simple integration. However, this function is complicated in general, so we need to make some simplifications in order to proceed:

1. Since biotin-streptavidin binding takes place primarily at the head group of biotin, we say that the rate r depends only on the "head" coordinate, say by taking the center-of-mass of all head atoms or simply picking one representative. We do the latter, and pick the carbonyl oxygen.

2. As the underlying sheet is planar, it is reasonable to assume that the binding availability depends primarily on the displacement of the head perpendicular to the sheet. Therefore, r now depends on a single parameter, the orthogonal displacement y above the sheet of the carbonyl oxygen.

3. Lastly, we assume that $r(y)$ is a step function: $r(y) = 0$ below a cutoff distance $y < y_c$, and $r = 1/\tau$ for $y > y_c$. Here, τ is the binding timescale of free biotin with streptavidin. Therefore, when the biotin head is "too close" to the surface, it is unavailable for binding, and when it is above the cutoff distance, it binds with streptavidin as if it were free.

After the above simplifications, r can be expressed concisely as

$$r(y) = \tfrac{1}{\tau}\delta_{y>y_c}$$

where the Kronecker delta function δ returns 1 when the condition in subscript is satisfied and 0 otherwise. Now, we compute the binding probability $Q(T)$ for a given trajectory.

$$\begin{aligned} Q(T) = 1 - P(T) &= 1 - \exp\left(-\int_0^T r(y)\mathrm{d}t\right) \\ &\simeq 1 - \exp\left(-\sum_{y=0}^{N} r(y)\mathrm{d}t\right) \\ &= 1 - \exp\left(-N_{>}\frac{\mathrm{d}t}{\tau}\right) \end{aligned}$$

where $N_{>} := \sum\limits_{y=0}^{N} \delta_{y>y_c}$ is the number of frames for which the head group is above cutoff. In the above, we discretized the integral to a sum over time-steps of length dt. It is assumed here that the head does not move appreciably over dt (i.e. y does not change). Assuming a smooth trajectory, this assumption will be justified as long as dt is small.

Let $\tau_{>} := \tau \cdot \frac{N}{N_{>}}$. Substituting in the previous equation,

$$Q(T) = 1 - \exp\left(-\frac{T}{\tau_{>}}\right)$$

where we used $N\mathrm{d}t = T$. Therefore, $\tau_{>}$ defines an *effective timescale* for the binding of attached biotin to streptavidin, for the given trajectory.

Now, we introduce a fourth assumption: the trajectory is well-mixed. In other words, the timescale $\tau_{>}$ does not depend on the particulars of the trajectory (and in particular the initial conditions), and may be used to define an effective rate equation $\frac{\mathrm{d}P}{\mathrm{d}t} = -\frac{1}{\tau_{>}}P$. With this assumption, one can predict binding probabilities over a (usually much longer) timescale S:

$$Q(S) \simeq 1 - \exp\left(-\frac{S}{\tau_{>}}\right) \tag{5.3}$$

Numerically, the timescale $\tau_{>}$ can be calculated from the number of frames for which the oxygen was above the cutoff, $N_{>}$, and the free binding timescale τ of biotin-streptavidin.

We performed this analysis for our trajectories. First, biotin/streptavidin binding timescale τ must be fixed. We compute it as follows: the rate constant for the first binding of biotin to a *free* streptavidin is experimentally known to be [47]

$$k = 3.6 \times 10^6 M^{-1} s^{-1}$$

Since streptavidin has multiple binding sites, multiple binding events are possible. However, we may safely ignore them here since the biotin molecules are not free in solution but attached to the fibril.

There is an additional concern that the binding rate depends on steric factors: due to the attachment of biotin to a large substrate, only one hemisphere is available to streptavidin for binding. Thus, one can expect the true binding rate to be lower by a geometric factor. However, we ignore this factor for the purposes of the calculation.

Finally, the rate k is concentration-dependent: however, using the concentrations used in the motivating experiment by Wetzel *et al* [17], we may derive a rate $\gamma = k \cdot c$, (where c is the streptavidin concentration) with units of s^{-1}. Then, we find

$$\tau = \gamma^{-1} = 5.6 \times 10^4 s \sim 1.5 \text{ hrs}$$

Physically, this timescale indicates that the streptavidin binds to attached biotin in the timescale of hours, which is indeed the case in experiment, [4, 17].

Next, our model needs a cutoff length. This is the height above which we say that the biotin head group has a binding availability of 100%, and below which the biotin is unavailable for binding. Physically, the cutoff length must be less than the total length of GLNp, which is ~ 2.5 nm. An analysis of say the H-bonding of the biotin head group with the sheet as a function of height could yield a value for the cutoff height.

An alternative approach, which does not require a specific cutoff length, is to *sweep* the binding timescale over all possible (and physical) values of the cutoff length. Certainly, a cutoff greater than the GLNp extended length (~ 2.5 nm) is unphysical. On the lower end, we may reasonably demand that the cutoff be greater than zero nm, since the biotin head group is never found below the backbone CA. So, we carry out the above timescale analysis for a range of cutoffs from 0 to 2.5 nm (see figure 5.13.

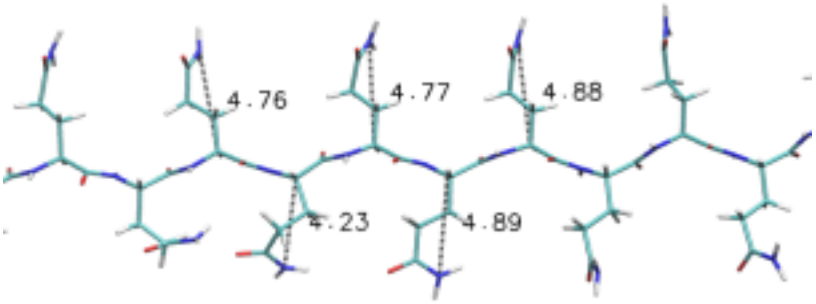

FIGURE 5.12: Typical length of GLN sidechains in the top geometry

So far, the reference point used for measuring distance from sheet has been the backbone CA. However, in this analysis we are interested in dividing the reaction coordinate into “zones of influence”: sheet vs. solvent. For the front face, it is reasonable to use Cα as the reference. However, for the top face, the sheet effectively begins at the extremal points of the GLN side chains, which lie approximately 0.45 nm above the sheet (see Fig. 5.12). Therefore, when deciding the cutoff between the sheet vs. solvent zone of influence, it is more physical to define height with respect to the envelope of the side chain extremal points. In order to simplify this, we introduce a shift of 0.45 nm to the height computed from Cα for all points on the top face. Fig. 5.13 below implicitly contains this relative shift.

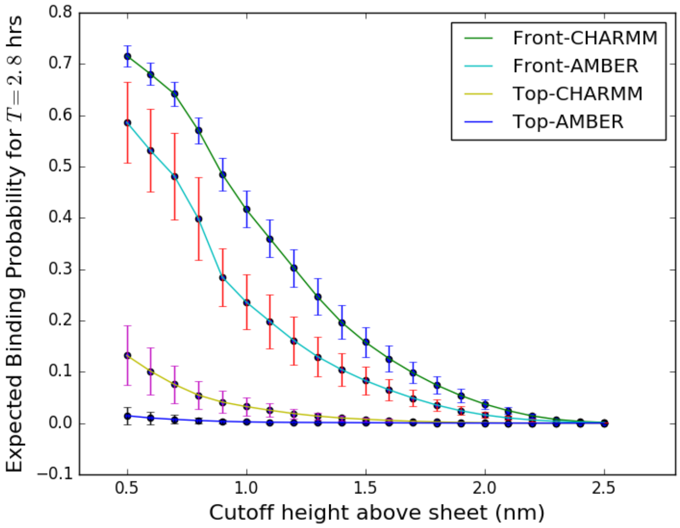

FIGURE 5.13: Probability of biotin-streptavidin binding, scaled up to an experimentally relevant timescale of 2.8 hrs. The cutoff height is measured from the extremal points of the sheet: backbone Cα for the front face, and backbone Cα $+0.45$ nm (i.e. top of GLN side chain) for the top face.

Independent of the value chosen for the cutoff, we see that both force fields predict that biotin-streptavidin binding will be more likely on the front face than on the top face. Consistent with the earlier observation that the GLNp density is shifted closer to the surface in AMBER, we see that binding probabilities are lower in AMBER.

Next, we define the *preferential binding* between the two faces as the log of the ratio of binding probabilities. This form of the expression is motivated by how probabilities are typically related to energy: namely, that the negative exponential of the energy yields the Boltzmann factor, which is a probability weighting. We plot the preferential binding for both force fields in Fig. 5.14 below.

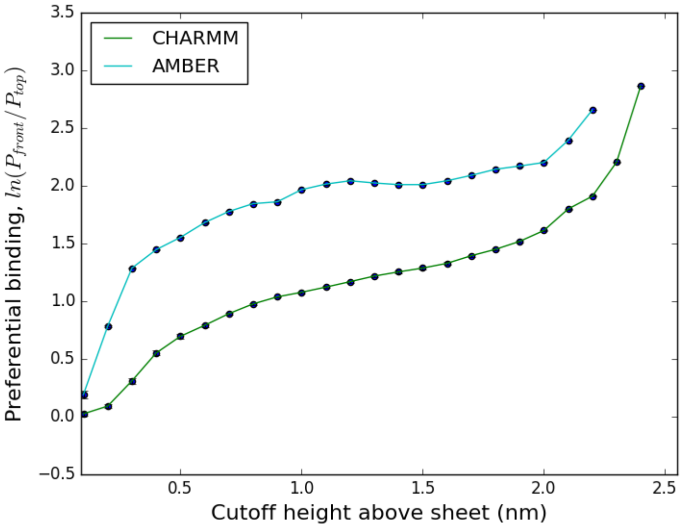

FIGURE 5.14: Preferential binding of biotin to streptavidin on the front face vs. the top face for both force fields, as a function of cutoff height. The preferential binding is given as the logarithm of the ratio of binding probabilities for the two faces. Thus, if the cutoff is taken to be 1 nm, the Markov model predicts that biotin-streptavidin is ten times as likely (for CHARMM) on the front face than on the top face.

The preferential binding makes it clear that for both force fields, biotins on the front face will be more available to bind with streptavidin than biotins on top faces. For reasonable cutoff values of ~ 1 nm, the preference is by a factor of 10 (for CHARMM), i.e. an order of magnitude in probability. For AMBER, the factor is closer to 100.

It is interesting to see that the preferential binding profiles for the two force fields are similar, qualitatively. However, they are off by as much as a factor of 10 for certain cutoff values. This indicates that one must exercise caution when interpreting results from force field simulations as is, since the their technical specifics may result in large discrepancies in the end result.

Therefore, we end this section with noting that using multiple force fields to simulate the same system is advisable, as it provides a basis for distinguishing physical features from numerical artifacts.

Chapter 6

Conclusion

We conclude with a discussion of the results obtained in this thesis project, and their relevance in the context on ongoing experimental work.

The main research problem in this thesis was motivated by a series of experiments conducted by the group of Ron Wetzel, [4, 17]. They observed that biotin attached to a polyQ fibril exhibits preferential binding to streptavidin depending on whether it is attached to a "growing" face of the fibril or not.

In order to probe the reason behind this selective binding mechanism, we argued that a computational approach would be ideal, and set up an MD simulation of the system in question. We simulated two sheet geometries: the growing face (denoted "front") and a non-growing face (denoted "top"). Two force fields, CHARMM36 and AMBER14, were used in parallel so that their results could be compared later. For each of the four cases of interested identified thus, we ran eight independent simulations of 1 μs each. The data collected was then analyzed in several ways.

We calculated the autocorrelation function, and found that for front face trajectories, there is a decay over a timescale of under ~ 100 ns. On the other hand, top face geometries are seen to have a memory over many hundreds of ns. This justifies the use of sampling times of a μs, but we also note that a timescale of ~ 10 μs may be more appropriate for the top face.

Then, we computed density profiles of the GLNp head group as a function of height above the surface. From the density profile in Fig. 5.4, we derived the potential of mean force (PMF), which we argue approaches the free energy profile for well-sampled data. This provided a visual of the energy landscape along the distance perpendicular to the sheet. Finally, we constructed a Markov model which effectively allowed us to predict the binding probability of GLNp to streptavidin over experimentally relevant timescales.

The result of this analysis is that biotin is indeed more available for binding when it is attached to the growing surface, as opposed to a non-growing surface. But moreover, we have gained an insight into *why* this might be the case.

The biotin head group is very reactive, since it contains two pentacyclic rings rich in H-bonding sites. This feature of biotin partly explains why it has such a high binding affinity with streptavidin. When attached to a β-rich substrate such as a polyQ fibril, it is precisely this feature of biotin which appears to play a role in determining its binding availability. In other words, the biotin head group is seen to interact strongly with the GLN substrate, and moreover to varying degrees depending on the geometry of the substrate.

For the front geometry, the interdigitated side chains of the sheet lie parallel to the sheet surface. In this case, the biotin head is found to have multiple metastable states within 1 nm of the sheet surface, one of which is at the H-bonding distance from the sheet. However, in this configuration, biotin is also in contact with waters, and more readily explores configurations where it is extended away from the sheet.

For the top geometry, the side chains at the surface emerge vertically out of the sheet like hairs, and the GLNp head interacts strongly with them. In particular, the GLNp head has a highly stable "embedded" configuration not seen in the front geometry, in which it is found *between* side chains. In this configuration, the head is within H-bonding distance of multiple side chain groups, and is therefore stabilized, despite the sterically unfavourable nature of this arrangement. Finally, the GLNp head has other stable states at ~ 0.8 nm, i.e. ~ 0.3 nm above the side chains. However, it very rarely explores extended configurations and tends to stay in the vicinity of the sheet.

This difference in the behaviour of GLNp with the two faces explains why there may be a difference in binding availability. Since the biotin head group is "stuck" in stable configurations close to the top surface, it may be unavailable for binding to a large molecule such as a streptavidin. We remark that this mechanism seems quite general to fibrils, since the interaction of biotin with the side chains did not depend on the chemical make-up of the residue. It would interesting to check this with experimental studies.

Lastly, we identify open questions and future directions of research. While the central question has been answered, there are many interesting features that remain to be explored.

First, we may also look at the full spatial profile of the O atom. The aim behind this would be to identify metastable states, or regions where all trajectories spend a considerable amount of time. Then, one can prepare a Markov chain model of the identified states. The transition rates between states near the surface and far from surface can provide a timescale for availability of biotin. Of course, ergodicity has to be assumed in this system.

The H-bonding of GLNp with the sheet and the waters would provide additional insight into the biotin-surface interaction. We found that the biotin head ”embeds” itself between neighbouring side chains, so that it is plausible that the stabilizing interaction is H-bonding. Both biotin head group and side chains have multiple H-bonding sites. Still, it is fully plausible that the interaction is Van der Waals in nature. On the other hand, when the biotin head is free in water, it may be stabilized by H-bonding with solvent water. The comparison between biotin-water and biotin-surface H-bonding will be telling.

Finally, it would be very interesting to study how other drug-like molecules interact with fibrils, and whether they exhibit similar geometry-dependent interactions. The understanding of these dynamics might provide some insight into targeted drug-design. Since amyloid fibrils play a role in a variety of diseases, this understanding may prove useful in future therapeutic strategies to combat these often devastating illnesses.

Bibliography

[1] Fabrizio Chiti and Christopher M. Dobson. Protein Misfolding, Functional Amyloid, and Human Disease. *Annual Review of Biochemistry*, 75(1): 333–366, jun 2006. ISSN 0066-4154. doi: 10.1146/annurev.biochem.75.101304.123901. URL http://www.annualreviews.org/doi/10.1146/annurev.biochem.75.101304.123901.

[2] Markus S Miettinen, Luca Monticelli, Praveen Nedumpully-Govindan, Volker Knecht, and Zoya Ignatova. Stable polyglutamine dimers can contain β-hairpins with interdigitated side chains-but not α-helices, β-nanotubes, β-pseudohelices, or steric zippers. *Biophysical journal*, 106 (8):1721–8, apr 2014. ISSN 1542-0086. doi: 10.1016/j.bpj.2014.02.027. URL http://www.ncbi.nlm.nih.gov/pubmed/24739171http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4008795.

[3] A. V. Kajava, U. Baxa, and A. C. Steven. Beta arcades: recurring motifs in naturally occurring and disease-related amyloid fibrils. *The FASEB Journal*, 24(5):1311–1319, may 2010. doi: 10.1096/fj.09-145979. URL http://www.ncbi.nlm.nih.gov/pubmed/20032312.

[4] Elizabeth Christine Landrum. THE KINETICS AND THERMODYNAMICS OF POLYGLUTAMINE AGGREGATION. 2013. URL http://d-scholarship.pitt.edu/20302/1/ETD{_}Landrum{_}Final1.pdf.

[5] Essential Numerical Methods (Course at MIT). http://silas.psfc.mit.edu/22.15/lectures/chap10.html#tth_sEc10.1.1. Accessed: 2017-06-14.

[6] Naveen Kumar Kaliannan. A study on the structure and properties of silica glass and silica nanoparticles via Monte Carlo simulations. Master's thesis, Bergische Universität Wuppertal, 2016.

[7] Hueng-Chuen Fan, Li-Ing Ho, Ching-Shiang Chi, Shyi-Jou Chen, Giia-Sheun Peng, Tzu-Min Chan, Shinn-Zong Lin, and Horng-Jyh Harn. Polyglutamine (PolyQ) Diseases: Genetics to Treatments. *Cell Transplantation*, 23(4):441–458, apr 2014. ISSN 09636897. doi: 10.3727/096368914X678454. URL http://www.ncbi.nlm.nih.gov/pubmed/24816443http://openurl.ingenta.com/content/xref?genre=article{&}issn=0963-6897{&}volume=23{&}issue=4{&}spage=441.

[8] Ronald Wetzel, Shankaramma Shivaprasad, and Angela D. Williams. Plasticity of Amyloid Fibrils. *Biochemistry*, 46(1):1–10, jan 2007. doi: 10.1021/bi0620959. URL http://www.ncbi.nlm.nih.gov/pubmed/17198370http://www.ncbi.nlm.nih.gov/pubmed/17198370http://www.ncbi.nlm.nih.gov/pubmed/17198370.

[9] Fabrizio Chiti and Christopher M Dobson. Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology*, 5 (1):15–22, jan 2009. ISSN 1552-4450. doi: 10.1038/nchembio.131. URL http://www.ncbi.nlm.nih.gov/pubmed/19088715http://www.nature.com/doifinder/10.1038/nchembio.131.

[10] C. M. Dobson. The structural basis of protein folding and its links with human disease. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1406):133–145, feb 2001. ISSN 0962-8436. doi: 10.1098/rstb.2000.0758. URL http://www.ncbi.nlm.nih.gov/pubmed/11260793http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1088418http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2000.0758.

[11] C M Dobson. Protein misfolding, evolution and disease. *Trends in biochemical sciences*, 24(9):329–32, sep 1999. ISSN 0968-0004. URL http://www.ncbi.nlm.nih.gov/pubmed/10470028.

[12] WHO — Dementia: a public health priority. *WHOalzheimer*, 2016. URL http://www.who.int/mental{_}health/publications/dementia{_}report{_}2012/en/.

[13] C. P. J. Maury. The emerging concept of functional amyloid. *Journal of Internal Medicine*, 265(3):329–334, mar 2009. ISSN 09546820. doi: 10.1111/j.1365-2796.2008.02068.x. URL http://doi.wiley.com/10.1111/j.1365-2796.2008.02068.x.

[14] R. Narayanaswamy, M. Levy, M. Tsechansky, G. M. Stovall, J. D. O'Connell, J. Mirrielees, A. D. Ellington, and E. M. Marcotte. Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings of the National Academy of Sciences*, 106(25):10147–10152, jun 2009. ISSN 0027-8424. doi: 10.1073/pnas.0812771106. URL http://www.ncbi.nlm.nih.gov/pubmed/19502427http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2691686http://www.pnas.org/cgi/doi/10.1073/pnas.0812771106.

[15] S. K. Maji, M. H. Perrin, M. R. Sawaya, S. Jessberger, K. Vadodaria, R. A. Rissman, P. S. Singru, K. P. R. Nilsson, R. Simon, D. Schubert, D. Eisenberg, J. Rivier, P. Sawchenko, W. Vale, and R. Riek. Functional Amyloids As Natural Storage of Peptide Hormones in Pituitary Secretory Granules. *Science*, 325(5938):328–332, jul 2009. ISSN 0036-8075. doi: 10.1126/science.1173155. URL http://www.ncbi.nlm.nih.gov/pubmed/19541956http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2865899http://www.sciencemag.org/cgi/doi/10.1126/science.1173155.

[16] Aimee M. Morris, Murielle A. Watzky, and Richard G. Finke. Protein aggregation kinetics, mechanism, and curve-fitting: A review of the literature. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1794(3):375–397, 2009. ISSN 15709639. doi: 10.1016/j.bbapap.2008.10.016. URL http://www.sciencedirect.com/science/article/pii/S1570963908003488.

[17] A. M. Bhattacharyya, A. K. Thakur, and R. Wetzel. Polyglutamine aggregation nucleation: Thermodynamics of a highly unfavorable protein folding reaction. *Proceedings of the National Academy of Sciences*, 102(43):15400–15405, oct 2005. ISSN 0027-8424. doi: 10.1073/pnas.0501651102. URL http://www.ncbi.nlm.nih.gov/pubmed/16230628http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1266079http://www.pnas.org/cgi/doi/10.1073/pnas.0501651102.

[18] J. Shao and M. I. Diamond. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. *Human Molecular Genetics*, 16(R2):R115–R123, jul 2007. ISSN 0964-6906. doi: 10.1093/hmg/ddm213. URL http://www.ncbi.nlm.nih.gov/pubmed/17911155https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddm213.

[19] John S Philo and Tsutomu Arakawa. Mechanisms of protein aggregation. *Current pharmaceutical biotechnology*, 10(4):348–51, jun 2009. ISSN 1873-4316. URL http://www.ncbi.nlm.nih.gov/pubmed/19519409.

[20] Gaetano Invernizzi, Elena Papaleo, Raimon Sabate, and Salvador Ventura. Protein aggregation: Mechanisms and functional consequences. *The International Journal of Biochemistry & Cell Biology*, 44(9):1541–1554, sep 2012. ISSN 13572725. doi: 10.1016/j.biocel.2012.05.023. URL http://www.ncbi.nlm.nih.gov/pubmed/22713792http://linkinghub.elsevier.com/retrieve/pii/S1357272512001896.

[21] F Ferrone. Analysis of protein aggregation kinetics. *Methods in enzymology*, 309:256–74, 1999. ISSN 0076-6879. URL http://www.ncbi.nlm.nih.gov/pubmed/10507029.

[22] H Naiki and F Gejyo. Kinetic analysis of amyloid fibril formation. *Methods in enzymology*, 309:305–18, 1999. ISSN 0076-6879. URL http://www.ncbi.nlm.nih.gov/pubmed/10507032.

[23] N M GREEN. AVIDIN. 1. THE USE OF (14-C)BIOTIN FOR KINETIC STUDIES AND FOR ASSAY. *The Biochemical journal*, 89(3):585–91, dec 1963. ISSN 0264-6021. URL http://www.ncbi.nlm.nih.gov/pubmed/14101979http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1202466.

[24] R M Zimmermann and E C Cox. DNA stretching on functionalized gold surfaces. *Nucleic acids research*, 22(3):492–7, feb 1994. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pubmed/8127690http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC523609.

[25] Songming Chen, Valerie Berthelier, Wen Yang, and Ronald Wetzel. Polyglutamine aggregation behavior in vitro supports a recruitment mechanism of cytotoxicity. *Journal of Molecular Biology*, 311(1):173–182, aug 2001. ISSN 00222836. doi: 10.1006/jmbi.2001.4850. URL http://www.ncbi.nlm.nih.gov/pubmed/11469866http://linkinghub.elsevier.com/retrieve/pii/S0022283601948508.

[26] Songming Chen, Valerie Berthelier, J. Bradley Hamilton, Brian O'Nuallai, and Ronald Wetzel. Amyloid-like Features of Polyglutamine Aggregates and Their Assembly Kinetics. *Biochemistry*, 41(23):7391–7399, jun 2002.

doi: 10.1021/bi011772q. URL http://pubs.acs.org/doi/abs/10.1021/bi011772q.

[27] ‡ and Pawel Sikorski*, † and § Edward Atkins‡. New Model for Crystalline Polyglutamine Assemblies and Their Connection with Amyloid Fibrils. 2004. doi: 10.1021/BM0494388. URL http://pubs.acs.org/doi/abs/10.1021/bm0494388.

[28] Robert Schneider, Miria C. Schumacher, Henrik Mueller, Deepak Nand, Volker Klaukien, Henrike Heise, Dietmar Riedel, Gerhard Wolf, Elmar Behrmann, Stefan Raunser, Ralf Seidel, Martin Engelhard, and Marc Baldus. Structural Characterization of Polyglutamine Fibrils by Solid-State NMR Spectroscopy. *Journal of Molecular Biology*, 412(1): 121–136, sep 2011. ISSN 00222836. doi: 10.1016/j.jmb.2011.06.045. URL http://www.ncbi.nlm.nih.gov/pubmed/21763317http://linkinghub.elsevier.com/retrieve/pii/S0022283611007248.

[29] Lauren E Buchanan, Joshua K Carr, Aaron M Fluitt, Andrew J Hoganson, Sean D Moran, Juan J de Pablo, James L Skinner, and Martin T Zanni. Structural motif of polyglutamine amyloid fibrils discerned with mixed-isotope infrared spectroscopy. *Proceedings of the National Academy of Sciences of the United States of America*, 111(16):5796–801, apr 2014. ISSN 1091-6490. doi: 10.1073/pnas.1401587111. URL http://www.ncbi.nlm.nih.gov/pubmed/24550484http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4000827.

[30] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, 1989. ISBN 0198556454.

[31] Andrew R. Leach. *Molecular modelling : principles and applications*. Prentice Hall, 2001. ISBN 0582382106.

[32] Thijs van Westen, Thijs J. H. Vlugt, and Joachim Gross. Determining Force Field Parameters Using a Physically Based Equation of State. *The Journal of Physical Chemistry B*, 115(24):7872–7880, jun 2011. ISSN 1520-6106. doi: 10.1021/jp2026219. URL http://pubs.acs.org/doi/abs/10.1021/jp2026219.

[33] Marco Hülsmann, Thorsten Köddermann, Jadran Vrabec, and Dirk Reith. GROW: A Gradient-based Optimization Workflow for the Automated Development of Molecular Models.

[34] Almudena García-Sánchez, Conchi O Ania, José B Parra, David Dubbeldam, Thijs J H Vlugt, Rajamani Krishna, and Sofía Calero. Transferable Force Field for Carbon Dioxide Adsorption in Zeolites. doi: 10.1021/jp810871f. URL http://www.acmm.nl/molsim/users/dubbeldam/pdfs/Garcia-Sanchez2009.pdf.

[35] Emeric Bourasseau, Mehalia Haboudou, Anne Boutin, Alain H Fuchs, and Philippe Ungerer. New optimization method for intermolecular potentials: Optimization of a new anisotropic united atoms potential for olefins: Prediction of equilibrium properties. doi: 10.1063/1.1537245. URL http://slapper.apam.columbia.edu/bib/papers/boura{_}jchp03.pdf.

[36] Philippe Ungerer, Christè Le Beauvais, Jé Rô Me Delhommelle, Anne Boutin, Bernard Rousseau, and Alain H Fuchs. Optimization of the anisotropic united atoms intermolecular potential for n - alkanes. doi: 10.1063/1.481116.

[37] Thijs van Westen, Thijs J. H. Vlugt, and Joachim Gross. Determining Force Field Parameters Using a Physically Based Equation of State. *The Journal of Physical Chemistry B*, 115(24):7872–7880, jun 2011. ISSN 1520-6106. doi: 10.1021/jp2026219. URL http://pubs.acs.org/doi/abs/10.1021/jp2026219.

[38] Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the Amber biomolecular simulation package. *WIREs Comput Mol Sci*, 2012. doi: 10.1002/wcms.1121. URL http://casegroup.rutgers.edu/{~}cerutti/amber{_}web/2012{_}09{_}25{_}WIRE{_}CMS{_}AMBER{_}Overview{_}10.1002{_}wcms.1121.pdf.

[39] Jing Huang and Alexander D. MacKerell. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, 34(25):2135–2145, sep 2013. ISSN 01928651. doi: 10.1002/jcc.23354. URL http://www.ncbi.nlm.nih.gov/pubmed/23832629http://www.

pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3800559http://doi.wiley.com/10.1002/jcc.23354.

[40] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):NA–NA, 2009. ISSN 01928651. doi: 10.1002/jcc.21367. URL http://doi.wiley.com/10.1002/jcc.21367.

[41] Karunakar Kar, Cody L. Hoop, Kenneth W. Drombosky, Matthew A. Baker, Ravindra Kodali, Irene Arduini, Patrick C.A. van der Wel, W. Seth Horne, and Ronald Wetzel. β-Hairpin-Mediated Nucleation of Polyglutamine Amyloid Formation. *Journal of Molecular Biology*, 425(7):1183–1197, 2013. doi: 10.1016/j.jmb.2013.01.016. URL http://www.sciencedirect.com/science/article/pii/S0022283613000326.

[42] Bertrand Guillot. A reappraisal of what we have learnt during three decades of computer simulations on water. *Journal of Molecular Liquids*, 101(1-3):219–260, nov 2002. ISSN 01677322. doi: 10.1016/S0167-7322(02)00094-6. URL http://linkinghub.elsevier.com/retrieve/pii/S0167732202000946.

[43] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *J . Chem . Phys . Additional information on J . Chem . Phys . Journal Homepage*, 79(926), 1983. doi: 10.1063/1.445869.

[44] Michael W Mahoney and William L Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. URL http://folding.cnsm.csulb.edu/ffamber/pdfs/mahoney{_}tip5p{_}2000jcp.pdf.

[45] Pekka Mark And and Lennart Nilsson*. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. 2001. doi: 10.1021/JP003020W. URL http://pubs.acs.org/doi/abs/10.1021/jp003020w?journalCode=jpcafh.

[46] W A Hendrickson, A Pähler, J L Smith, Y Satow, E A Merritt, and R P Phizackerley. Crystal structure of core streptavidin determined from multiwavelength anomalous diffraction of synchrotron radiation. *Proceedings of the National Academy of Sciences of the United States of America*, 86(7):2190–4, apr 1989. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pubmed/2928324http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC286877.

[47] Monpichar Srisa-Art, Emily C. Dyson, Andrew J. deMello, and Joshua B. Edel. Monitoring of real-time streptavidinbiotin binding kinetics using droplet microfluidics. *Analytical Chemistry*, 80(18):7063–7067, 2008. doi: 10.1021/ac801199k. URL http://dx.doi.org/10.1021/ac801199k. PMID: 18712935.