Essays on the interplay between glycosaminoglycans and amyloid-β peptides

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Suman Samantray, M.Ing.

aus

Rourkela, Odisha, Indien

Berichter: Univ.-Prof. Dr. rer. nat. Arne Lüchow

Prof. Dr. Birgit Strodel

Tag der mündlichen prüfung: 24. November 2021

DECLARATION

I, Suman Samantray, hereby declare that I wrote the dissertation titled, *Essays on the interplay between glycosaminoglycans and amyloid-β peptides* , without any unauthorized external assistance and used only sources acknowledged in the work. All textual passages which are appropriated verbatim or paraphrased from published and unpublished texts as well as all information obtained from oral sources are duly indicated and listed in accordance with bibliographical rules. In carrying out this research, I complied with the rules of standard scientific practice as formulated in the statutes of RWTH Aachen University, Aachen, Germany to ensure standard scientific practice.

This dissertation includes parts of three original manuscripts published in peer-reviewed journals, one book chapter in press, and one manuscript under review.

Parts of Chapter 2 and 3 are reprinted with permission from *Fatafta, H., Samantray, S., Sayyed-Ahmad, A., Coskuner-Weber, O., Strodel, B. (2021). Molecular simulations of IDPs: from ensemble generation to IDP interactions leading to disorder-to-order transitions. Prog. Mol. Biol. Transl. Sci., 183, 135–185. Copyright ©2021* ***Elsevier*** [1].
In the book chapter, I have contributed to the literature review and writing of the sections of force fields for IDPs and long-time methods for MD simulations. The final draft of the article was reviewed equally by all the authors.

Parts of Chapter 4 are reprinted with permission from **Paul, A., *Samantray, S., Anteghini, M., Khaled, M., Strodel, B. (2021). Thermodynamics and kinetics of the amyloid-β peptide revealed by markov state models based on MD data in agreement with experiment. Chem. Sci., 12(19), 6652-6669. Copyright ©2021* ***Royal Society of Chemistry*** [2] (* *equal authorship*).
In the published article, I analyzed the structural data to check convergence and

performed additional MD simulations. I and Arghadwip Paul (*Research internship student in the Computational Biochemistry Group at IBI-7, FZ Jülich, Germany*) equally contributed to the literature review. The initial analysis of the MD simulation trajectories provided by D. E. Shaw research were performed by Marco Anteghini (*M.Sc. student in the Computational Biochemistry Group at IBI-7, FZ Jülich, Germany*), as part of his M.Sc. thesis. The prediction models were developed with the help of Arghadwip Paul and Mohammed Khaled. The validation of simulation data against experimental data was performed with the help of Arghadwip Paul. The final draft of the article was co-written with Prof. Birgit Strodel and Arghadwip Paul.

Parts of Chapter 5 are reprinted with permission from *Samantray, S., Yin, F., Kav, B., Strodel, B. (2020). Different force fields give rise to different amyloid aggregation pathways in molecular dynamics simulations. J. Chem. Inf. Model., 60(12), 6462–6475. Copyright ©2020* ***American Chemical Society*** [3].
In the published article, my contributions involved performing MD simulations, analyzing simulation data and preparing results, developing computer algorithms, and writing the first draft of manuscript. The MD simulations of the peptide–water interactions were analyzed and co-written with the help of Dr. Batuhan Kav (*PostDoc in the Computational Biochemistry Group at IBI-7, FZ Jülich, Germany*). The final draft of the article was reviewed by Prof. Birgit Strodel.

Parts of Chapter 6 are reprinted with permission from *Samantray, S., Olubiyi, O.O., Strodel, B. (2021). The influences of sulphation, salt type and salt concentration on the structural heterogeneity of glycosaminoglycans. Int. J. Mol. Sci., 22(21), 11529. Copyright ©2021 by the authors. Licensee* ***MDPI, Basel, Switzerland*** [4].
In the manuscript currently under review, my contributions involved designing and performing all the MD simulations, analyzing the simulation data and preparing results, developing computer algorithms, and preparing the draft of manuscript. The literature review was completed with the help of Dr. Olujide O. Olubiyi (*PostDoc in the Computational Biochemistry Group at IBI-7, FZ Jülich, Germany*). The final draft of

the article was reviewed by Prof. Birgit Strodel.

Parts of Chapter 7 are reprinted with permission from *Samantray, S., Strodel, B. (2021). The effects of different glycosaminoglycans on the structure and aggregation of the amyloid-β (16–22) peptide. J. Phys. Chem. B, 125(21), 5511-5525. Copyright ©2021* ***American Chemical Society*** [5].
In the published article, my contributions involved designing and performing all the MD simulations, analyzing the simulation data and preparing results, developing computer algorithms, and writing the draft of manuscript. The final draft of the article was reviewed by Prof. Birgit Strodel.

Suman Samantray
September, 2021
Aachen, Germany

Dedicated to my Parents.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my advisor Prof. Birgit Strodel. It has been an honor to be her student and I appreciate her contribution of time, ideas, and funding to make my Ph.D. experience an unforgettable one. I will be forever grateful for her support during the tough times that I have faced in my Ph.D. journey. This could not have been possible without her insights and guidance.

I would like to acknowledge Prof. Arne Lüchow, for agreeing to be part of my dissertation committee at RWTH Aachen University. I would also like to express my gratitude to the members of the service team in the Graduate School of Aachen Institute of Computational Engineering Science and its Director Dr. Nicole Faber. I am most grateful to Prof. Gunnar Schröeder and his group members Dr. Karunakar and Dr. Benedikt, Dr. Bogdan Barz and his student Soumav Nath, for all the fruitful scientific discussions. I would also like to thank Friedrich Haubensak, Dorothea Stobbe, Tugba Özdal, Volker Söhnitz, Sabine Pick, Sabine Werths and other staff members at IBI-7, FZ Jülich for their constant support.

My time at Jülich and Aachen was made enjoyable in large part due to the many friends that became a part of my life. I am especially grateful to my colleagues - Hiba, Xue, Jennifer, Maryam, Feng, Dr. Orkide, Anna, Alex, Wibke, Mohammed, Subrata, Dr. Batuhan, Dr. Abhishek, Dr. Olujide and all other former members of the Strodel group, for the fun-filled and memorable time. I very much appreciate all the support I received from you all.

A good support system and encouragement are important for surviving and staying sane during Ph.D. A huge thanks to you Anusha, my best friend and wife for always motivating me and being my all-time stress buster. I can always count on you!. Thanks for constantly encouraging me towards achieving my goals and believing in me. Her unconditional love, backing, and constant mental support made this dissertation possible.

Finally, to my family and friends, I am grateful in more ways than I can express for your support and encouragement over the years. Thank you to my parent in-laws for your faith in me.

I dedicate this dissertation to my Parents, whom I owe an immeasurable debt of gratitude for providing me the opportunity to follow my dream.

PUBLICATIONS

- **Samantray, S.**, Yin, F., Kav, B., Strodel, B. (2020). “Different force fields give rise to different amyloid aggregation pathways in molecular dynamics simulations", *J. Chem. Inf. Model.*, 60(12), 6462–6475.

- *Paul, A., ***Samantray, S.**, Anteghini, M., Khaled, M., Strodel, B. (2021). “Thermodynamics and kinetics of the amyloid-β peptide revealed by markov state models based on MD data in agreement with experiment", *Chem. Sci.*, 12(19), 6652-6669. (* *equal authorship*)

- **Samantray, S.**, Strodel, B. (2021). “The effects of different glycosaminoglycans on the structure and aggregation of the amyloid-β (16–22) peptide", *J. Phys. Chem. B*, 125(21), 5511-5525.

- Fatafta, H., **Samantray, S.**, Sayyed-Ahmad, A., Coskuner-Weber, O., Strodel, B. (2021). “Molecular simulations of IDPs: from ensemble generation to IDP interactions leading to disorder-to-order transitions", *Prog. Mol. Biol. Transl. Sci.*, Academic Press, ISBN 978-0-323-85299-9, 183, 135–185.

- **Samantray, S.**, Olubiyi, O.O., Strodel, B. (2021). “The influences of sulphation, salt type, and salt concentration on the structural heterogeneity of glycosaminoglycans", *Int. J. Mol. Sci.*, 22(21), 11529.

Contributions to other publications

- Olubiyi, O.O., **Samantray, S.**, Illig, A.-M. (2022). “Advances in structure-based virtual screening for drug discovery", Academic Press, ISBN 978-0-323-90264-9. (*In Press*)

- **Samantray, S.**, Schumann, W., IIlig, A.-M., Pacheco, M.-C., Paul, A., Barz, B., Strodel, B. (2022). “Molecular dynamics simulations of protein aggregation: pro-

tocols for simulation setup and analysis with markov state models and transition networks", *Methods Mol. Biol.*, Springer US, ISBN 978-1-071-61545-4, 2340. (*In Press*)

- **Samantray, S.**, Cheung, D.L. (2020). "Effect of the air-water interface on the conformation of amyloid beta", *Biointerphases*, 15(6), 061011.

- Deike, S., Rothemund, S., Voigt, B., **Samantray, S.**, Strodel, B., Binder, W.H. (2020). "β-Turn mimetic synthetic peptides as amyloid-β aggregation inhibitors", *Bioorg. Chem.*, 101, 104012.

ABSTRACT

Intrinsically disordered proteins (IDPs), which represent ~40% of the human proteome, play crucial roles in a variety of biological pathways and biomolecular assemblies. Monomeric IDPs such as amyloid-β (Aβ), can aggregate into insoluble, relatively inert, rigid structures called fibrils, but also much more toxic, soluble structures of intermediate size and varying shapes, which are called oligomers. The toxic aggregates of Aβ peptide are implicated in the pathogenesis of Alzheimer's disease (AD). In this thesis, we use in-silico approaches to model Aβ under different physiological and pathological conditions to unravel their effects on the structures and kinetics of the amyloid oligomers. We first highlight the ramifications of molecular mechanics parameters on the structural heterogeneity of Aβ and the aggregation process of various Aβ fragments. Next, we demonstrate how Aβ fragments aggregate in the presence of glycosaminoglycans. To this end, the conformational dynamics of different glycosaminoglycans was first elucidated to understand their behavior in the absence Aβ fragments. The conclusions from these investigations enabled us to identify force fields which predict Aβ structures and dynamics in agreement with experimental observations. From transition networks applied to the aggregation data we deduced the structural transitions during the early and intermediate stages of oligomer formation. Furthermore, we elucidated the intermolecular interactions between Aβ and glycosaminoglycans that transpire towards enhancing, stabilizing, or inhibition of Aβ aggregation behavior.

ZUSAMMENFASSUNG

Intrinsisch ungeordnete Proteine (IDPs), die ~40 % des menschlichen Proteoms ausmachen, spielen in einer Vielzahl von biologischen Stoffwechselwegen und biomolekularen Aggregaten eine entscheidende Rolle. Monomere IDPs wie Amyloid-β (Aβ) können zu unlöslichen, relativ inerten, starren Strukturen, den Fibrillen, aggregieren, aber auch zu viel toxischeren, löslicheren Strukturen mittlerer Größe und unterschiedlicher Form, die als Oligomere bezeichnet werden. Die toxischen Aggregate des Aβ-Peptids sind an der Pathogenese der Alzheimer-Krankheit (AD) beteiligt. In dieser Arbeit verwenden wir In-silico-Ansätze, um Aβ unter verschiedenen physiologischen und pathologischen Bedingungen zu modellieren, um deren Auswirkungen auf die Strukturen und Kinetik der Amyloid-Oligomere zu entschlüsseln. Wir beleuchten zunächst die Auswirkungen molekularmechanischer Parameter auf die strukturelle Heterogenität von Aβ und den Aggregationsprozess verschiedener Aβ-Fragmente. Als nächstes zeigen wir, wie Aβ-Fragmente in Gegenwart von Glykosaminoglykanen aggregieren. Zu diesem Zweck wurde zunächst die Konformationsdynamik verschiedener Glykosaminoglykane aufgeklärt, um ihr Verhalten in Abwesenheit von Aβ-Fragmenten zu verstehen. Die Schlussfolgerungen aus diesen Untersuchungen ermöglichten es uns, Kraftfelder zu identifizieren, die Aβ-Strukturen und -Dynamik in Übereinstimmung mit experimentellen Beobachtungen vorhersagen. Aus Übergangsnetzwerken, die auf die Aggregationsdaten angewendet wurden, leiteten wir die strukturellen Übergänge während der frühen und mittleren Phasen der Oligomerbildung ab. Darüber hinaus haben wir die intermolekularen Wechselwirkungen zwischen Aβ und Glykosaminoglykanen aufgeklärt, die das Aβ-Aggregationsverhalten verstärken, stabilisieren oder hemmen.

CONTENTS

1

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disorder that primarily affects the elderly and is having negative impact on societies across the globe as the overall life expectancy of humans continues to increase. It is evident from previous studies that the amyloid-β peptide (Aβ) plays a central role in the pathogenesis of AD [6, 7]. The Aβ peptides are typically 39-43 amino acid residues in length. The hallmark of the development and progression of AD in living beings is the aggregation of Aβ, where misfolded proteins aggregate either intra- or extracellularly, into insoluble, relatively inert, rigid structures called amyloid fibrils, but also much more toxic, soluble structures of intermediate size, and varying shapes, which are called oligomers [7].The amyloid fibrils possess a unique cross β-sheet quaternary structure, which were initially thought to be the major culprits for AD. But this perception has been reconsidered in the favour of soluble Aβ oligomers [8]. The toxic oligomeric species[9, 10] are lighter and more mobile than the fibrils, allowing them to spread across cell and tissue boundaries.

However, no published simulation studies have considered the full complexity of the biological environment, which plays a crucial role in understanding the initial stages of aggregation. To decode the native aggregation process, it is important to extend the simulations towards *in vivo* conditions. A large part of the *in vivo* environment of Aβ is given by the brain extracellular space (ECS) [11]. The brain ECS is mainly comprised of water, ions and other biomolecules, which form an interlocking mesh of fibrous proteins and glycosaminoglycans (GAGs) [11]. The latter are carbohydrate polymers and mostly attach to extracellular matrix proteins to form proteoglycans. The main types of proteoglycans found within the ECS involve heparan sulphate and chondroitin sulphate. Hyaluronic acid is a non-proteoglycan polysaccharide, which also plays a major structural role in the formation of the brain extra-

cellular matrix. The dominant proteins in the ECS are collagens and elastins. Collagen is fibrillar in nature providing structural support to cells whereas elastin gives elasticity to tissues.

Experimental studies suggested that GAGs play a major role in promoting aggregation, nucleation and fibril formation of Aβ and other amyloid proteins [12–14]. GAGs are mostly polysaccharides abundantly found in the human body, with considerable traces of GAGs found in the diseased tissues of amyloid disorders including Alzheimer's disease as reported clinically [15, 16]. It has been hypothesized that, as typical for catalyzed reactions, GAGs favor aggregation, nucleation and amyloid fibril formation by a mechanism substantially different from that occurring in bulk solution [17]. It is suggested that they can both influence and promote misfolding of polypeptides into pro-amyloidogenic intermediates rich in β-sheets, and also act as a structural template for self-assembly [18, 19]. However, no studies exist, especially no simulation studies that investigated the effects of GAGs on Aβ aggregation at sub-molecular detail, which is the focus of my thesis.

Computer simulations play an important role in modelling the components of the ECS coupled with Aβ protein in explicit solvent at atomistic and coarse grained levels to provide insights into AD [20]. The development of reliable simulation techniques has made computer simulation predictions reflect the experiments adequately. However, the results of an MD simulation can only be as good as the underlying force field. The most common force fields cannot distinguish between aggregation-prone and non-aggregating peptide sequences, producing similar and in most cases too fast aggregation kinetics for all peptides. However, in recent times new force fields specially designed for intrinsically disordered proteins (IDPs) were developed, making it is necessary to test the best suitable force field for peptide aggregation simulations and base future reparameterizations on it.

Here, in this thesis, with the help of computer simulations, we are exploring Aβ conformations and structures that can help deep dive into the research of amyloidogenic peptides, which are otherwise tedious to conduct experimentally. We use molecular dynamics simulations to gain insight at the nanoscopic level to the conformational heterogeneity and aggregation kinetics of the amyloid-β peptide, the effects

of mutagenesis and force fields on the aggregation dynamics of the amyloid-β_{16-22} peptide, the conformational dynamics of glycosaminoglycans under the influence of different salts and salt concentrations, and the role of glycosaminoglycans in the self-assembly of amyloid-β_{16-22} peptides.

This thesis is presented in the following order:

In **chapter 2 and 3**, we introduce the biomolecular systems and the computational approaches. We also explain the prediction models used to describe the folding mechanisms of IDPs.

In **chapter 4**, we explain the importance of force fields for accurate modeling of IDPs such as the Aβ40 peptide and its aqueous environment, and determine when the peptide system attains equilibrium in MD simulations and also provides insights into the structural ensembles of Aβ40.

In **chapter 5**, we extensively discusses the applicability of different force fields to studying peptide aggregation using the Aβ_{16-22} peptide and mutations of it as test case. We investigate the performance of the force fields in modeling the monomeric state, the aggregation into oligomers, and the stability of the aggregation end product, i.e., the fibrillar state.

In **chapter 6**, we investigate the dynamical features of GAGs and successfully unravel conformational fingerprints unique to each GAGs. We further study and provide a quantitative insight into the effect of salt types, salt concentrations as well as effect of sulphation sites on the conformational preferences of GAG chains.

In **chapter 7**, we elucidate the conformational transitions of monomeric and aggregated forms of Aβ_{16-22}, be it in solution or in the presence of GAG chains. Thereafter, we determine the effect of electrostatic interactions between GAGs and Aβ_{16-22} and the sulphation of GAG chains, on the structural motifs of Aβ_{16-22}, thereby understanding the self-assembly kinetics of homo-dimers of Aβ_{16-22}.

2

BIOMOLECULAR SYSTEMS

2.1 DECLARATION

Parts of this chapter were submitted by the author to the *Progress in Molecular Biology and Translational Science* as Fatafta, H., Samantray, S., Sayyed-Ahmad, A., Coskuner-Weber, O., Strodel, B. (2021). Molecular simulations of IDPs: from ensemble generation to IDP interactions leading to disorder-to-order transitions. All the images are reprinted with Copyright ©2021 **Elsevier**.

2.2 INTRINSICALLY DISORDERED PROTEINS

Proteins play a central role as a catalyst to enhancing the rate of chemical reactions in living organisms. The exploration of proteins sequence-structure-function paradigms have led us to believe that, proteins function only when they are folded into their right structures [21]. The central mechanism of molecular biology entails that DNA encodes the genetic information which is transcribed into messenger RNA and then translated into an amino acid sequence that folds into a protein. However, the biophysical mechanisms regulating the folding of an amino acid sequence into well-defined three-dimensional protein structure are still not fully understood [22]. Even though, the sequence-structure paradigm has dominated scientific minds for more than 100 years, the detection of intrinsically dynamic and flexible and biologically active proteins at the end of the 20th century gave scientists access to complete genome sequences [23]. This class of proteins is characterised by inherent flexibility, which enabled scientists to understand that some of the amino acid sequences are not expected to fold into globular protein structures [24]. Coupled with experimental in-

vestigations, the structure-function paradigm was noticed in the case proteins and domains that were incompletely structured or completely disordered in solution, yet remained biologically functional [23]. As a consequence, the concept of “intrinsically disordered proteins" was introduced at the turn of the century [24]. In fact, intrinsically disordered proteins (IDPs), are not rare exceptions and instead represent a unique and a very broad class of proteins [24]. The accepted definition for disorder holds true for protein regions or whole proteins that are biologically active, but are represented by flexible conformational ensembles either at the secondary and/or tertiary structure level [25]. IDPs and intrinsically disordered protein regions (IDPRs) exist as a dynamic conformational ensemble of rapidly interconverting structures resembling “protein clouds" whereby the atom positions and backbone Ramachandran angles vary significantly with time, without obtaining specific equilibrium values [26]. One of the hallmarks in the sequence of IDPs is a marked bias in their amino acid composition, which includes a relatively small proportion of hydrophobic and aromatic residues, but a relatively large proportion of polar and charged amino acid residues [27]. Currently, all the identified IDPs and IDPRs are collected in the DisProt database (https://disprot.org/), which to date contains ≈1,600 non-ambiguous IDPs and ≈3,500 IDPRs [28] and are ever increasing.

2.3 FUNCTIONS OF IDPS

IDPs which represent ≈40% of the human proteome, play crucial roles in a variety of biological pathways and biomolecular assemblies and have been implicated in a large number of human diseases [29]. In the mid-1990s, experimental measurements of regulatory proteins and bioinformatics studies of the genome sequences that were just emerging revealed that disordered regions are very common in eukaryotic proteins [30] and responsible for cellular regulation and signaling. The occurrence of IDPRs of significant size that contain >50 amino acid residues is common in functional proteins [31]. The existence of functional IDPs, such as polypeptide hormones, has been recognized for many years and IDPs were detected in intact cells in nuclear magnetic resonance (NMR) experiments [32]. Functions of IDPs in-

clude transcription and translation regulation, cellular signal transduction, the storage of small molecules, protein phosphorylation, and self-assembly regulation, such as ribosome and bacterial flagellum [33]. IDPRs can function as chaperones for RNA molecules and for other proteins, indicating that IDPRs exert their function by binding to (mis)folded proteins or RNA molecules, thereby providing a function related to recognition elements and/or unfolding and loosening of kinetically trapped folding intermediates [34]. In Figure 1 the different functions of IDPs and IDPRs are summarized.

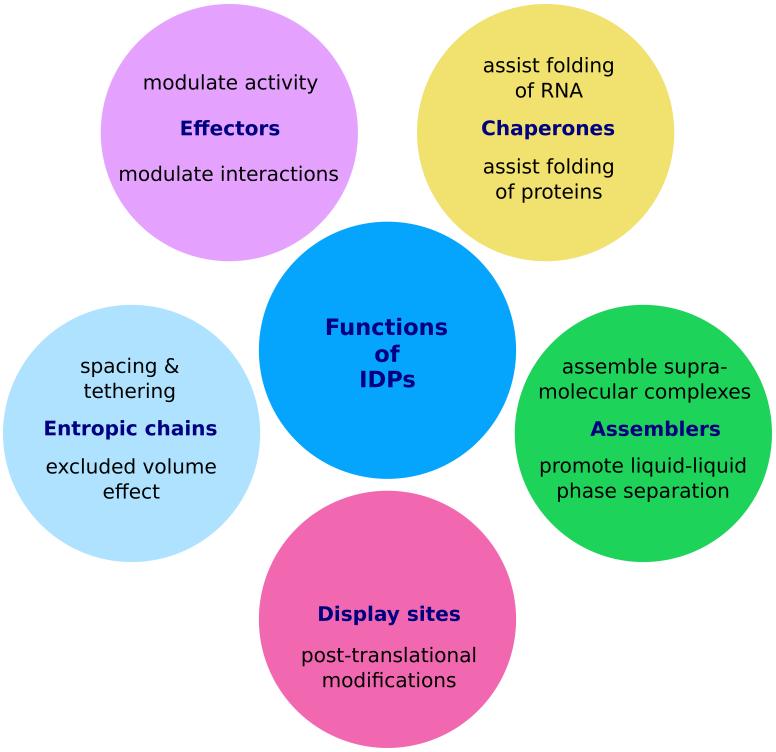

Figure 1: Overview of the different functions of IDPs

2.4 IDPS IN DISEASES

IDPs and IDPRs are prone to engage in promiscuous and unwanted interactions when misexpressed, mismodified, misprocessed and/or dysregulated. Thus they are associated with the development of various diseases [35]. Conformational or protein-folding diseases are divided into two classes. The first class includes errors in the genetic blueprint which leads to misfolded proteins that affect the function as well [36].

Examples of this first class include p53 and specific alterations in diseases such as cystic fibrosis. The second class is through the self-assembly of the misfolded proteins into the formation of multimolecular ordered nanostructures. These self-assembled structures are broadly considered to disrupt cellular function and subsequently damaging the surrounding tissues [36]. Such alterations, known as amyloidosis, are detected in severe diseases, such as Alzheimer's disease, Parkinson's disease (PD), Huntington's disease, and type II diabetes.

2.4.1 *IDPs in amyloid diseases*

Peptide aggregation is the process where individual monomeric IDPs are attracted towards each other inducing a chain reaction which ends in the formation of stable non-covalent aggregates. Amyloid-β (Aβ) and α-synuclein (α-Syn) are examples of two such IDPs that are associated with neurodegenerative diseases, such as Alzheimer's and Parkinson's diseases [37, 38] respectively. Aβ peptides are derived from the amyloid precursor protein (APP) via cleavage by both β- and γ-secretase. Aβ peptides can oligomerize to form flexible soluble oligomers which may exist in several forms [38]. It has been proposed that specifically misfolded Aβ oligomers can induce other Aβ peptides to form misfolded oligomers, which leads to a chain reaction as in a prion infection [37]. These oligomers are toxic to the nerve cells and misfolded Aβ can also induce misfolding in tau protein in Alzheimer's disease [39]. The neuronal protein α-Syn is an IDP that is encoded by the *SNCA* gene in humans and it has 140 amino acid residues [37]. Although the function of α-Syn is not fully understood, studies proposed that it plays a role in restricting the mobility of synaptic vesicles, consequently attenuating synaptic vesicle recycling and neurotransmitter release [37]. However, α-Syn is prone to aggregation, forming insoluble fibrils in pathological conditions, such as Parkinson's disease, dementia with Lewy bodies, or multiple system atrophy. The pathological deposition of the misfolded prion protein (PrP) into its aggregated form causes prion diseases, which are called transmissible spongiform encephalopathies (TSEs) [40]. TSEs range from chronic wasting disease of mule deer and elk to Creutzfeld-Jakob disease in humans [40].

2.4.2 *Amyloid-β peptide*

It is well understood that the Aβ peptide is the key instigator that triggers the progression of Alzheimer's disease (AD) resulting in neurodegeneration via accumulation and aggregation. The Aβ peptide is a self-aggregating peptide and derives its name *amyloid* from its earlier description as a carbohydrate product coined initially by Schleiden and then by Virchow in the mid-19th century [41], which turned out to be wrong, and *beta* due to its secondary structure fingerprints as β-sheets. The peptide was first discovered by Glenner and Wong, as the main component of a vascular amyloid system [42], however it was Kang *et al.* who reported the peptides significance four years later in 1987, via characterization of amyloid plaque deposits in the brain of AD patients [43]. Later it was reported, the distribution of the peptide is not limited to the senile plaques of the brain cells but it is also traced in the cerebrospinal fluid of the central nervous system [44–46]. Aβ is a soluble disordered peptide rcontainingby 36 to 43 amino acid residues. The peptide is derived by proteolytic cleavage of a transmembrane protein called APP catalysed by both β-secretase and γ-secretase enzymes [43, 47, 48]. The most common variants of Aβ peptide are Aβ40 and Aβ42. Their structure are characterised using various biophysical techniques such as NMR spectroscopic experiments [8, 47, 49, 50]. Aβ40 is composed of 40 amino acid residues, with a hydrophilic N-terminus region which are prone to metal binding, and two hydrophobic patches forming the central and C-terminal region as shown in Figure 2. The Aβ42 peptide is composed of 42 amino acid residues, where the C-terminus of Aβ40 is extended by two hydrophobic residues, isoleucine and alanine. The additional two residues reduce the conformational flexibility of the C-terminus region by forming a β-hairpin configuration around residues 37-48 [51]. Aβ40 is found in abundance in the brain, whereas the Aβ42 is more prone to aggregation [52]. Both these variants display a diverse structural heterogeneity as a monomer. Therefore, it is important to study the structural dynamics of the Aβ monomer and identify the conformational motifs of Aβ enhancing oligomerization and subsequently forming aggregates.

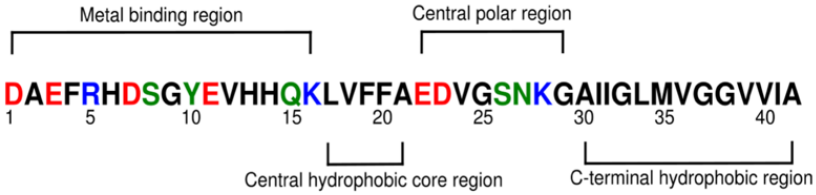

Figure 2: The sequence of Aβ42 can be divided into four regions: the metal binding region, the central hydrophobic core region, the central polar region and the C-terminal hydrophobic region. Residues labelled as red, blue, green and black are negatively charged, positively charged, polar and hydrophobic, respectively (Reproduced with permission from Kepp, 2017. [53] Copyright ©2017 Elsevier).

2.5 GLYCOSAMINOGLYCANS

Carbohydrates, are other key building blocks found in all forms of life. The diverse nature of the carbohydrates is reflected through the combination of distinct building blocks such as monosaccharides or disaccharides, leading to the formation of long linear or branched polysaccharides. Many types of polysaccharide structures exist. Among them, glycosaminoglycans (GAGs) which are considered as biologically active polysaccharides, are found in abundance in mammalian tissues [54]. The GAGs in general are long linear, periodic, anionic hetero-polysaccharide chains, also known as mucopolysaccharides. The GAGs are highly polar and attract water, serving several functions within the body characterised by their molecular structure. Historically, GAGs were believed to posses limited functionality such as cell hydration and structural scaffolding. However, recent studies suggests that GAGs play key role in cell signaling and act as a modulator for a vast amount of biochemical processes, such as regulation of cell growth and proliferation, promotion of cell adhesion, anticoagulation, and wound repair. The GAGs are mostly composed of repeating disaccharide units of a hexuronic acid and a hexosamine. The hexosamine is either a D-glucosamine (GlcNAc) or a D-galactosamine (GalNAc). The hexuronic acid is in most cases, a D-glucuronic acid (GlcUA) or a L-iduronic acid (IdoUA), with the only exception of keratan sulphate where hexuronic acid is a D-galactose (Gal). The main classification of the GAGs is based on the differences of the disaccharide composition per type of GAG, and different sugar content and glycosidic linkage. Apart from differences in disaccharide composition, GAGs can also be classified based on sulphation at various

positions where the sulphation pattern can vary within the GAG sequence. Therefore, the sulphate additions coupled with the carboxyl group of the hexuronic acids result in GAGs being highly negative charged biological macromolecules.

Protein-chaperoned GAGs are also found in large quantities in specific cell structures in the cellular environment, referred to as proteoglycans (PGs). The PGs are large structures consisting of a linear core protein with many GAGs covalently linked to it, exclusively chondroitin sulphate and keratan sulphate. Each GAG is linked to the core protein via a Gal-Gal-xylose sugar linker, where the xylose is attached to a serine residue of the core protein and the GAG is linked to the galactose monosaccharide. PGs are highly diverse with variations occurring both in the type of core protein and in type, size and number of GAGs being linked. Aggrecan, the typical example of a PG, is a supra-molecular assembly composed of a core protein structure encoded by the gene ACAN [55, 56], and covalently attached to several sulphated GAG chains, exclusively chondroitin sulphate (CS) and keratan sulphate (KS). It is a vital component of the cartilage and the brain's extra cellular matrix.

The four primary groups of GAGs are classified based on their core disaccharide units which includes hyaluronic acid, heparin/heparan sulphate, chondroitin sulphate/dermatan sulphate (DS), and keratan sulphate [57].

2.5.1 *Classification of GAGs*

2.5.1.1 *Hyaluronic Acid*

Hyaluronic acid (HA) has the simplest structure among all GAGs and is devoid of any sulphation of the functional groups. HA was first isolated from the vitreous body of the eye, referred as *hyaloid* in Greek language, hence the name hyaluronan [58]. The structure consists of sequentially bound GlcUA linked to GlcNAc via glycosidic bonds [59, 60]. HA is a very long molecule made up of 2,000-25,000 disaccharide repeats with a molecular weight of 10^7 Da [61]. The molecular weight governs the folding mechanism of HA molecules, where HA sequences with higher molecular

weight adopt an anti-parallel β-sheet like tertiary structure, stabilized by specific H-bonds and hydrophobic interactions but such secondary structure are absent at low molecular weight [62]. HA is involved in various biological processes in the extracellular matrix of load-bearing joints. HA has a high binding affinity for water due to the presence of a large number of anionic residues, resulting in the formation of viscous hydrogels even at low solvent concentrations. The absorption of significant amounts of water helps in building a turgor pressure, crucial for connective tissue [63]. Additionally, HA is also involved in cell regulation, cell growth including embryonic development, cell migration, cell proliferation, inflammations, healing processes and tumor growth [64]. HA can be found in various kinds of tissue in vertebrates [65, 66] including the umbilical cord, the synovial fluid in joints, and the rooster combs of animal tissues, which act as major sources for industrial production of HA [63].

2.5.1.2 *Heparan sulphate/Heparin*

Heparin (Hep) was first reported about a century ago [67], however its potency as an anticoagulant was established in the last 30 years [68, 69]. Initially, Hep used to be isolated from liver, *hepatos* in Latin, resulting in the name heparin. The most efficient method to isolate Hep used to be from a polydisperse mixture from animal tissue that is rich in mast cells. Nowadays, heparin is isolated from highly vascularized tissue derived as a side-product from livestock. Heparan sulphate (HS) and Hep share a similar structure containing repeating disaccharide units of GlcNAc and hexuronic acid residues. The hexuronic acid residue GlcUA is seen in HS, while IdoUA is present in Hep. In general, HS has less than one sulphate group per disaccharide, while Hep on an average has three [70]. Consequently, Hep now is regarded as just one of the members within the family of HS [71]. HS/Hep is located at the external surface of cell membranes and also found in the extracellular matrix in the form of PG. The sulphation of the various hydroxyl groups or the amino group present on the glucosamine compound of HS/Hep is tagged with it various process, determining its ability to interact with various proteins, cytokines, and growth factors, and ultimately its bioactive function [54], cell regulation [72], and binding mechanism of HS-based PG helping in cell signalling [73, 74].

2.5.1.3 *Chondroitin sulphate/Dermatan sulphate*

Chondroitin sulphate (CS) and dermatan sulphate (DS) are similar in structural composition to HS and Hep, respectively. Their disaccharide repeat consists of GalNAc and hexuronic acids – GlcUA in CS and IdoUA in DS. CS is the most abundant GAG in the body skeletal muscle and soft tissue. Structurally, CS includes various sulphation patterns, among them two isomers: CS-A (O-sulfo group attached at position 4 of GalNAc residue) and CS-C (O-sulfo group attached at position 6 of GalNAc residue), which are commonly studied. Similarly, there are disulfated CS GAGs which are classified as CS-D (O-sulfo group attached at position 2 of GlcUA and position 6 of GalNAc residue) and CS-E (O-sulfo group attached at position 4 and 6 of GalNAc residue). Currently, the CS isoform which was formerly designated as CS-B is called DS, due to the the presence of IdoUA residues in its disaccharide chains. The structural similarity between HS/Hep and CS/DS enables them to make specific interaction with similar molecules, including growth factors, cytokines, chemokines, adhesion molecules and lipoproteins [75]. CS is predominantly used in the treatment of osteoarthritis and coronary artery disease, whereas DS which is found in abundance in the dermis [76], is used as an anti-coagulant in wound repair, coupled with physiological roles in cardiovascular diseases and carcinogenesis. They are linked to a protein core via a serine residue and a Gal-Gal-xylose sugar linker similar to HS [77].

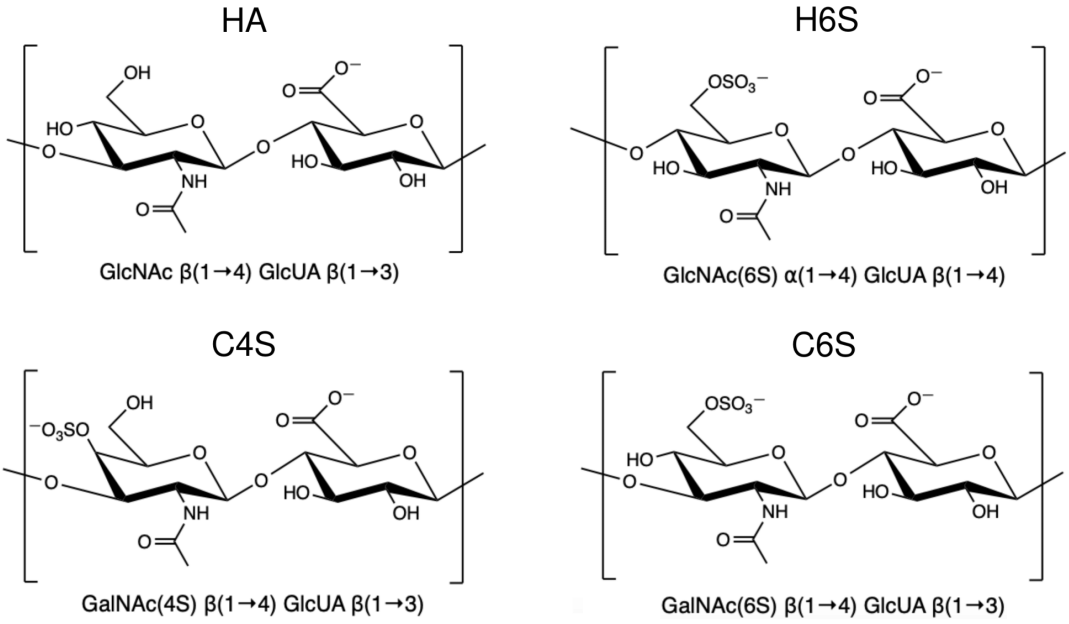

Figure 3: Chemical structure of the GAGs representing the disaccharide unit of the GAGs reported in our work. Heparan-6-sulphate (H6S) is with an HS O-sulfo group attached at position 6 of the GlcNAc residue, Chondroitin-4-sulphate (C4S) is CS with an O-sulfo group attached at position 4 of the GalNAc residue, and Chondroitin-6-sulphate (C6S) is CS with an O-sulfo group attached at position 6 of the GalNAc residue (Reprinted with permission from Samantray and Strodel, 2021. [5] Copyright ©2021 American Chemical Society).

2.5.2 *GAGs in amyloid diseases*

The mammalian GAGs, CS, DS, Hep, HS, and HA are linear polysaccharides and besides HA they all are sulphated to various extent as shown in Figure 3, and in most cases covalently attached to proteins to form PGs. GAGs interaction with numerous compounds in the extracellular space, e.g., signaling molecules like growth factors, morphogenesis, chemokines, proteins and their bioactive fragments, receptors, lipoproteins, and pathogens. These interactions tag them with various biochemical processes starting from embryonic development to extracellular matrix assembly and regulation of cell signaling in various physiological and pathological contexts, such as angiogenesis, cancer, neurodegenerative diseases, and infections [78]. PGs and their constituent GAGs are associated with amyloid deposits and are involved in the amyloidogenic pathway [13, 79–87]. In Alzheimer's disease, plaques are composed of the Aβ peptide and are associated with at least four different PGs, most extensively with

HS [88–90], CS [91], KS [92], and DS [93]. From circular dichroism spectroscopic measurements, it is confirmed that GAGs interaction with the Aβ peptides, Aβ40 and Aβ42 help us to determine their effects on peptide conformation and fibril formation. Monomeric Aβ peptides in trifluoroethanol, when diluted in aqueous buffer causes Aβ to form α-helical structures [80, 94–96]. However, experimental evidences by McLaurin and coworkers [19] suggests that the presence of GAGs, enhances the secondary structure transition of Aβ from random coil structures to rapidly adopting β-sheet conformations. Among the GAGs, C6S induced maximum transitional shift in Aβ from random coil to amyloidogenic β-sheet followed by DS > C4S > HS > KS > Hep. The incubation of preformed Aβ fibrils in the presence of CS results in extensive lateral aggregation into “rope-like" macrofibre structures and subsequently, the formation of well-defined amyloid fibrils takes place indicating an enhanced nucleation of Aβ peptides. The GAGs act at the earliest stage of fibril formation, Aβ nucleation, and are extensively involved in the lateral aggregation of preformed fibrils or nonspecific adhesion to plaques. The elucidation of structure-property relationship based on the interactions between Aβ and the different GAGs, as well as the structural order of GAGs in different cellular conditions are important for the successful development and evaluation of GAG-specific therapeutic interventions for circumventing neurodegenerative diseases [19].

3

COMPUTATIONAL APPROACHES

3.1 DECLARATION

Parts of this chapter were submitted by the author to the *Progress in Molecular Biology and Translational Science* as Fatafta, H., Samantray, S., Sayyed-Ahmad, A., Coskuner-Weber, O., Strodel, B. (2021). Molecular simulations of IDPs: from ensemble generation to IDP interactions leading to disorder-to-order transitions. All the images are reprinted with Copyright ©2021 **Elsevier**.

3.2 INTRODUCTION

Molecular modelling with computational tools is aimed at understanding and predicting macroscopic properties of complex systems (e.g., molecules, proteins, nucleic acids) at atomistic resolution. The complexity of biomolecular systems involves large number of molecules, making it difficult to quantify the properties analytically or by simply using numerical methods to solve it. Molecular dynamics (MD) simulations are a powerful tool for elucidating molecular mechanisms, acting as a bridge between laboratory experiments and computer simulations. The simplest form of MD involves a little more than Newton's second law of motion, irrespective of the fact that it is based on physics, chemistry, and statistical mechanics. The major applications of MD simulations include: understanding the structure and dynamics of proteins, unravelling the folding mechanisms of proteins, protein-protein interactions, docking of molecules and much more.

3.3 STATISTICAL MECHANICS

In principle, the macroscopic properties of a complex molecular system can be divided into two major categories: static equilibrium properties (e.g., temperature, pressure, density) and dynamic equilibrium properties (e.g., diffusion process, phase transition, reversible chemical reactions). In order to compute such macroscopic properties, it is important to generate a statistical ensemble at given conditions to represent all the accessible physical states or *phase space* of a molecular system. The phase space is defined by all possible values of position and momentum variables of the molecular system. To represent the statistical ensemble of a system, we need to account that every degree of freedom of the system is denoted as an axis of a multidimensional space. In a system consisting of N atoms, each possible state corresponds to a specific point in the phase space for which 6N values are required to define the state of that system. In general, a single point in phase space is denoted through three coordinates for position and three components for momentum to describe the state of the system. The motion of the points gives rise to all the accessible system's microstates. The different points in the phase space may belong to the same thermodynamic state. Therefore, the collection of all possible microscopic configurations of a thermodynamic system, that are characterised by an identical thermodynamic state is referred as a statistical ensemble.

Ensembles	Condition	Fixed variables
Microcanonical	Thermally isolated equilibrium	N, V, E
Canonical	Thermal equilibrium with a heat reservoir	N, V, T
Isobaric-isothermal	Isobaric-isothermal	N, p, T
Grand canonical	Grand canonical or Gibbs	μ, V, T

Table 1: Overview of the thermodynamic ensembles. In the microcanonical or NVE ensemble, number of particles N, volume V and total energy E are constant. The canonical or NVT ensemble, is characterized by fixed number of particles N, volume V and temperature T. The Isobaric-isothermal ensemble is characterized by fixed number of particles N, pressure p and temperature T. The Grand canonical ensemble is defined by fixed volume V, temperature T and chemical potential μ.

Determination of the ensemble average of property X is performed by integrating over all possible configurations of the system:

$$< X >= \iint X\left(p^N, r^N\right) \rho\left(p^N, r^N\right) dp^N dr^N \tag{1}$$

where $X\left(p^N, r^N\right)$ is the observable of interest, r is the atomic positions, and p the momenta. The probability density function $\rho\left(p^N, r^N\right)$ of the ensemble is given by:

$$\rho\left(p^N, r^N\right) = \frac{1}{Q} \exp\left[\frac{-H\left(p^N, r^N\right)}{K_b T}\right] \tag{2}$$

where K_b is the Boltzmann factor, T is the temperature, and H is the Hamiltonian. The denominator Q in the expression is called the *partition function*, which can be described as a dimensionless normalizing sum of Boltzmann factors over all microstates of the system:

$$Q = \iint \exp\left[\frac{-H\left(p^N, r^N\right)}{K_b T}\right] dp^N dr^N \tag{3}$$

The partition function denoted in equation 3, is an important variable for calculating various thermodynamics properties. The partition function acts as a bridge between microscopic thermodynamic variables and macroscopic properties. However, in order to overcome the analytical difficulties in solving equation 3, the *ergodic hypothesis* can be involved where the time-average of a certain physical property over long periods of time represents the ensemble-average of the same property.

$$< X >_{\text{ensemble}} =< X >_{\text{time}} \tag{4}$$

The time-average $< X >_{\text{time}}$ can be computed by:

$$< X >_{\text{time}} = \lim_{\tau \to \infty} \int_{t=0}^{\tau} X\left(p^N(t), r^N(t)\right) dt \sim \frac{1}{M} \sum_{t=1}^{M} X\left(p^N, r^N\right) \tag{5}$$

where t is the simulation time, M is the number of time steps in the simulation and $X\left(p^N, r^N\right)$ is the instantaneous value of the calculated property. Therefore, using a specific thermodynamic ensemble from an efficient sampling of microstates sequentially in time, we can compute ensemble-average properties of the system under evaluation [97].

3.4 MOLECULAR DYNAMICS SIMULATIONS

Computer simulations are a powerful tool that can provide solutions to many problems in statistical mechanics, otherwise these problems will be difficult to solve. Simulation predictions can be used to understand experimental results that are difficult to obtain in the laboratory. Therefore, computer simulation can be used as a bridge between model and theoretical prediction and between model and experimental results [98]. MD simulations are a form of computer simulations, used to study systems on the atomic and molecular scale. MD simulations are a valuable method for computing the equilibrium and transport characteristics of many body systems. It acts as a bridge between the length and the timescales of the micro and macro attributes, as shown in the Figure 4. MD simulations provide insight into experiments by revealing hidden details that are difficult to access in the laboratory due to complexity and high cost.

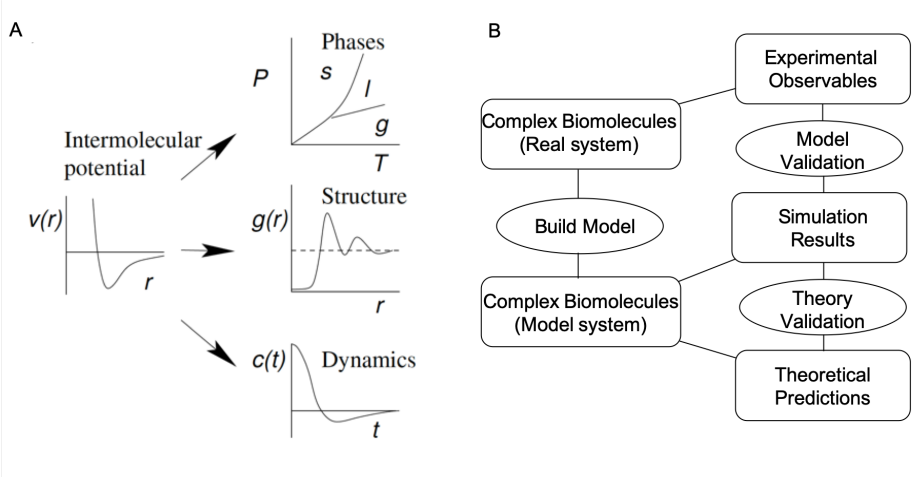

Figure 4: Simulations depicted as a bridge between (**A**) microscopic and macroscopic; (**B**) theory and experiment (Adapted with permission from Allen et al., 2004. [99] Copyright ©2004 by John von Neumann Institute for Computing).

In MD simulations we basically solve the Newton's equation of motion for N number of particles (usually N is very large) interacting through relevant potentials. The equation of motion of the i^{th} particle is given as

$$\mathbf{v}_i = \frac{d\mathbf{r}_i}{dt}, \sum_{j \neq i} \mathbf{F}_{ij} = m_i \frac{d\dot{\mathbf{r}}_i}{dt}, \tag{6}$$

where $\mathbf{r}_i$ and $\mathbf{v}_i$ are the position and velocity of the i^{th} particle and $\mathbf{F}_{ij}$ is the force exerted by particle i on particle j. In order to measure any observable macroscopic thermodynamic property in MD simulations, the corresponding observable has to be first expressed in terms of the position and momentum of the particles in the system. Here, each atom or molecule is given by position $\mathbf{r}_i$ or momentum $\mathbf{p}_i = m_i \mathbf{v}_i$, to propagate the particles in time. For this purpose, we need to calculate the forces $\mathbf{F}_i$ acting on particles, which are usually expressed in terms of potential energy $U(\mathbf{r}^N)$, where $\mathbf{r}^N = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N)$ represents the complete set of 3N coordinates. The Hamiltonian H of the system can be written as

$$H(\mathbf{p}^N, \mathbf{r}^N) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m_i} + U(\mathbf{r}^N), \tag{7}$$

Here $\mathbf{p}^N$ is the union of all momenta, $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$ [100]. Details on the form of the potential energy will be provided later in this chapter. The forces acting on the particles are derived from the potential as,

$$\mathbf{F}_i(\mathbf{r}^N) = -\frac{\partial U(\mathbf{r}^N)}{\partial \mathbf{r}_i}\,. \tag{8}$$

The equations of motion according to the Hamiltonian's equation are

$$\dot{\mathbf{r}}_i = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i}\,, \tag{9}$$

and

$$\dot{\mathbf{p}}_i = -\frac{\partial H}{\partial \mathbf{r}_i} = -\frac{\partial U}{\partial \mathbf{r}_i} = \mathbf{F}_i(\mathbf{r}^N)\,. \tag{10}$$

To integrate the equation of motion, different algorithms have been introduced in MD simulations. The most common and efficient algorithm used so far for integration is the Velocity Verlet algorithm as it is time reversible and conserves the phase and volume space [100, 101]. To derive it, we use the Taylor expansion of the coordinate $\mathbf{r}_i$ of a particle at time $t + \Delta t$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{V}_i(t)\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 + \frac{\Delta t^3}{3!}\dddot{\mathbf{r}}_i + \mathcal{O}(\Delta t^4)\,, \tag{11}$$

where Δt is the MD time step, and the estimated error in the new position is of the order of Δt^4. In the Velocity Verlet algorithm the velocity and position of a given particle is computed at time t and $t + \Delta t$, using the following equations

$$\mathbf{r}_i(t + \Delta t) - \mathbf{r}_i(t - \Delta t) = 2\mathbf{V}_i(t)\Delta t + \mathcal{O}(\Delta t^3)\,. \tag{12}$$

and

$$\mathbf{v}_i(t) = \frac{\mathbf{r}_i(t+\Delta t) - \mathbf{r}_i(t-\Delta t)}{2\Delta t} + \mathcal{O}(\Delta t^2)\,. \tag{13}$$

Here, the estimated error in velocity is $\sim \Delta t^2$, and the temperature, potential energy, and total energy are calculated at each time step. The total energy should be conserved throughout the MD simulation. The old positions and velocities at $t - \Delta t$ can be discarded after the calculations for $t + \Delta t$ have been done, and the new positions and the velocities become the next starting point. The steps are repeated a given number of times in order to reach the desired simulation time frame.

This integration scheme conserves the total potential energy E, so if the total number of particles N and the total volume V also remain unchanged, then the simulation is performed in the microcanonical ensemble NVE as described in Table 1. However, experiments in the laboratories are usually performed at constant temperature and not constant energy. Therefore, we need to reproduce an isothermal ensemble in MD simulations, in which the number of particles, volume, and temperature T is fixed; this ensemble is called canonical ensemble NVT. We can ensure the constant temperature by applying a thermostat to the system. Popular methods to control temperature include the velocity rescaling thermostat, Nosé-Hoover thermostat, Langevin thermostat, Andersen thermostat, and Berendsen thermostat. In addition to a thermostat, a barostat is also necessary to account for laboratory conditions to stabilize the pressure and concurrently the density of the system. Hence, the simulation is performed with an isothermal-isobaric ensemble NpT, wherein the number of particles, pressure, and temperature are kept constant to equilibrate the pressure. The pressure can be controlled using the Berendsen barostat and Parrinello-Rahman barostat.

Periodic boundary conditions (PBC) are used to create (artificial) large-volume systems in MD simulations in which the cubic box is surrounded by the replicas of itself in the entire space to form an infinite lattice. This arrangement allows us to simulate a large phase system with a finite number of particles. We adopt the minimum image convention where each particle interacts with the nearest atom or image in the periodic matrix. During the course of the simulation, when an atom moves in the original box, its periodic image move in the same way in all periodic boxes. This means that

once the atom leaves a box, its image will enter from the opposite side. No walls or surface atoms are present at the boundary of the original box. This box is used for measuring the coordinates of the N particle system. A two-dimensional image of such a periodic system is shown in Figure 5. The number density in the central box (implied to the entire system) is conserved. In addition, we don't need to save the coordinates of the entire system; we only need to save the particle coordinates of the original frame [100]. It is important to note that the imposed artificial periodicity works well with short-range interaction potentials but fails for long-range ones like for charged and dipolar systems. So, it is necessary to make sure the simulation system is neutralised using counter-ions otherwise Ewald summation which is used for computing long-range interactions becomes unphysical.

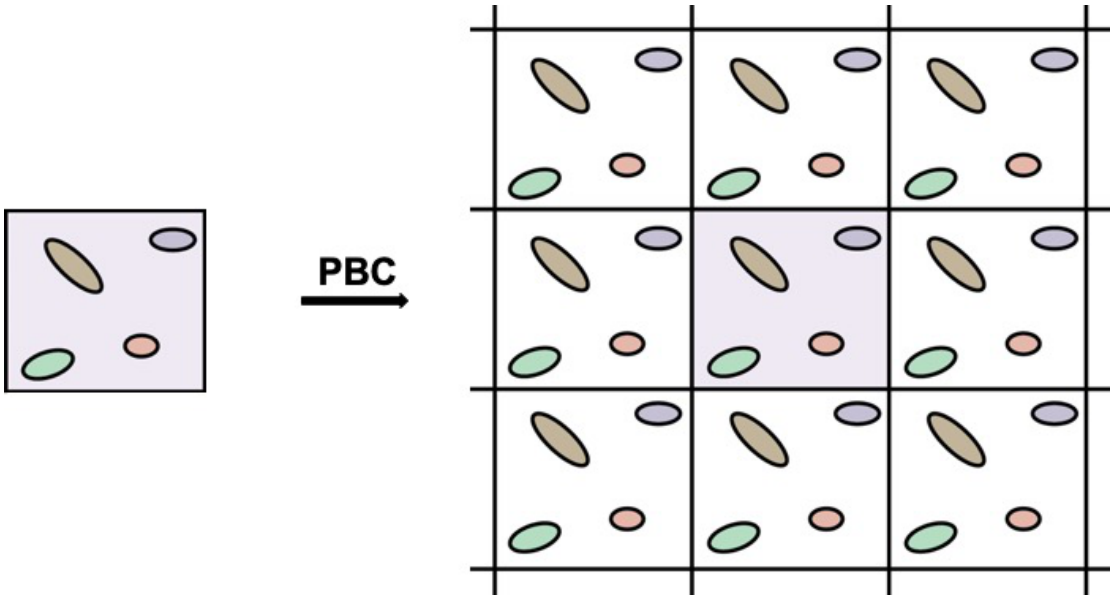

Figure 5: A two dimensional periodic system, the duplicate boxes i.e the image boxes are shown with a white background whereas the original box is shown in with a colored background.

In MD simulations, the forces in the system are computed by pairwise addition of forces on each pair of a molecule. Therefore, computing power in MD simulation will quickly become expensive because it involves a large number of pairwise calculations. So, for a short-range energy function, we can limit the summation by making an approximation. Since the largest contribution to the forces or potential comes from neighbors close to the atom of interest, we usually apply a spherical cutoff, which corresponds to setting the pair potential $V(r)$ to zero for pair distance, greater and

equal to the cutoff distance r_c. The introduction of a spherical cutoff should be a small perturbation. In order to ensure the minimum effect of this, we need to choose a sufficiently large cutoff distance [100].

3.5 MOLECULAR MECHANICS FORCE FIELDS

Biomolecular force fields (FFs) are represented by an energy function that are models to approximately predict the true potential energy surface. Proteins and other biomolecules are predominantly simulated using two ways: either using an atomistic or a coarse-grained FF. A model for the potential energy of a molecular system is composed of mathematical functions and associated constant parameters. The interactions between the particles of the system can be classified into two categories: bonded and non-bonded interactions. As an example, we provide the potential energy function which is used by the various CHARMM FFs [102]:

$$\begin{aligned} U(\mathbf{r}^N) = & \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\ & + \sum_{dihedrals} V_\varphi(1+\cos(n\varphi-\delta)) + \sum_{impropers} k_\omega(\omega-\omega_0)^2 \\ & + \sum_{i,j\,(LJ)} \epsilon_{ij}^{min}\left[\left(\frac{R_{min,ij}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{r_{ij}}\right)^{6}\right] + \sum_{i,j\,(Coul.)} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \\ & + \sum_{Urey-Bradley} k_{UB}(s-s_0)^2 + \sum_{CMAP} U_{CMAP}(\phi,\psi) \end{aligned} \tag{14}$$

Here, $\mathbf{r}^N$ denotes the conformation of the system consisting of N atoms with coordinates $\mathbf{r}^N = (r_{1,x}, r_{1,y}, r_{1,z}, r_{2,x}, r_{2,y}, r_{2,z}, \ldots, r_{N,x}, r_{N,y}, r_{N,z})$. The bonded energy terms describe bond stretching around the equilibrium values b_0 with force constants k_b, angle bending around equilibrium angles θ_0 and with force constants k_θ, torsions around bonds as characterized by the dihedral angles φ, periodicity n, shift δ and energy barrier V_φ, and out-of-plane bending, also called improper torsion, with the minimum at ω_0 and force constant k_ω. The non-bonded interactions contain Lennard-Jones and Coulomb potentials for interacting particles i and j. The Lennard-

Jones (LJ) potential is a 12-6 potential, where the repulsive $1/r^{12}$ term describes the Pauli repulsion at short distances of the interacting particles due to overlapping electron orbitals, and the attractive $1/r^6$ term describes attractions arising from dispersion forces, which are also called van der Waals (vdW) interactions. The distance between the two interacting particles is given by r_{ij}, ϵ_{ij}^{min} is the depth of the potential well, $R_{min,ij}$ is the distance at which the particle-particle LJ potential energy is minimal and can be calculated from the van der Waals radii of the particles i and j. The Coulomb potential models the electrostatic interactions between the partial charges q_i and q_j of atoms i and j with distance r_{ij} between them, where ϵ_0 is the vacuum permittivity. The terms described thus far, which are summarized in Figure 6, are common to all all-atom biomolecular force fields as found in the AMBER [103], CHARMM [102], OPLS-AA [104], and GROMOS [105] packages. Though, depending on the force field, small differences to equation (14) can occur, such as that $\cos(\theta)$ is used for defining the harmonic potential describing angle bending.

3.6 FORCE FIELDS FOR IDPS

Many FFs have been tailored for simulating the structural dynamics of folded proteins, and for a long time it was assumed that they are equally applicable to IDPs. However, as several FF benchmarks revealed, this is not the case [2, 106–112]. It turned out to be difficult to accurately simulate the structural ensembles of IDPs or IDPRs, as they do not fold into a well-defined three-dimensional structure under physiological conditions and instead populate a dynamic conformational ensemble of rapidly interconverting structures. It was found that a good balance between protein-protein and protein-water interaction parameters is key for a good description of IDP ensembles. Small changes to this subtle balance will either produce overly compact IDP structures where protein-protein interactions are overestimated, excessively aggregation-prone proteins due to underestimated protein-water interactions, or, if the latter are overestimated, extremely soluble IDP states that avoid protein-protein contacts. As accurate FFs are needed for the reliable generation of IDP ensembles, numerous FFs have been developed in recent times. They are usually based on exist-

ing FFs and different strategies were applied during their reparameterization, which will be explained in the following. In the CHARMM FFs, two correction terms are added to the potential energy. The Urey-Bradley (UB) term is used to improve the description of angle bending, where s is the distance between the first and third atom that define a bond angle. However, most force fields do not include Urey-Bradley terms, and also in CHARMM FFs no new UB terms were added in the past, since the only advantage of these terms is the better reproduction of subtleties in vibrational spectra. However, the goal of classical MD simulations seldomly is the calculation of infrared spectra. In fact, in most of the MD simulations of proteins the bond lengths are anyhow restrained to their equilibrium values in order to allow an increase in the time step used for the integration of the equations of motions. Moreover, many of the vibrations, especially those involving hydrogen bonds, would require a quantum-mechanical description for proper modeling as classical simulations reach their limit of validity here. The second correction term is called CMAP, which is a grid-based correction and accounts for the correlation between the backbone dihedral angles ϕ and ψ. Unlike the UB term, the CMAP correction has gained in popularity and was included in other FFs too, especially with the aim to improve the modeled ensembles of IDPs (see below).

Over the last decade, several research groups aimed at developing better FFs for IDPs, using equation (14) or a similar equation as starting point. Two major reparameterization strategies have been followed: first, optimization of dihedral angle parameters to provide better descriptions of the tendency of IDPs to adopt random coil conformations; second, strengthening the protein-water interactions to counteract the preference of proteins to collapse into a molten globule state as was seen with many of the common protein force fields. These reparameterization strategies are briefly explained in the next sections, while the reader is referred to a recent review [112] for more details.

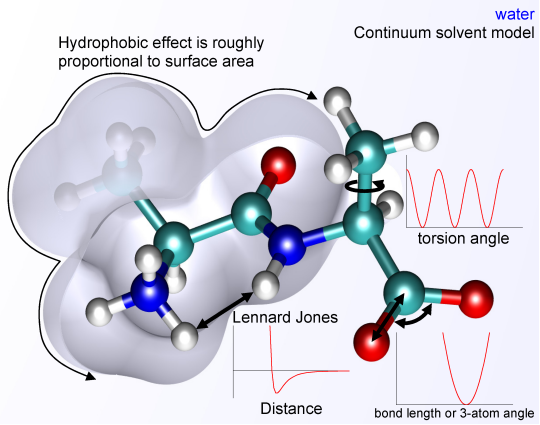

Figure 6: Contributions in all-atom force fields. The interactions between the atoms are divided in bonded and non-bonded interactions. Harmonic potentials are used to describe the vibrations of bonds and bond-angle bending, while periodic functions are needed for modeling the torsion around bonds. The non-bonded interactions are between atoms that are separated by at least three bonds or between atoms of different molecules. They arise from charge-charge interactions as described by the Coulomb potential, and from hydrophobic interactions as well as repulsive interactions if two atoms get too close to each other, which are collectively modeled by the Lennard-Jones potential. The water around a protein or other biomolecules can be modeled explicitly (not shown) using typical water models, such as TIP3P or TIP4P, or using an implicit solvent model. Explicit modeling of the water molecules gives usually better results, especially in the case of IDPs. Reproduced with permission (https://en.wikipedia.org/wiki/Force_field_(chemistry)).

3.6.1 *Optimization of Dihedral Parameters*

In many of the recently developed FFs for IDPs, adjustments of dihedral-angle parameters, especially those of the backbone dihedrals ϕ and ψ, were made [108, 113, 114]. The reason behind this strategy is that many of the common FFs overestimate propensities for α-helix and β-sheet formation in IDPs, as was recently reviewed by Mu et al. [112]. One remedy to resolve this secondary structure bias is to use dihedral-angle data of coil-like fragments in the training sets of FFs [115]. This approach was employed for AMBER03* and AMBER99SB* [115] which are based on AMBER03 [116] and AMBER99SB [117], respectively. However, AMBER03* overestimates and AMBER99SB* underestimates the helical content with respect to their respective predecessing FF. For OPLS-AA/M [118] and OPLS3 [119] a similar refitting approach was followed, including the reparameterization of the side-chain dihedrals, and as training set data from *ab initio* torsional energy scanning of blocked dipeptides was used. However, their ability to produce good IDP ensembles beyond proline

dipeptides and glycine tripeptides remains to be shown. In the CHARMM family of FFs, CHARMM22* [120] also a resulted from refitting dihedral angles based on CHARMM22 [121]. In our analyses of various FFs with regard to their ability to produce Aβ ensembles in agreement with the data obtained from NMR and fluorscence spectroscopy, CHARMM22* yielded acceptable results [2, 122]; in our earlier study of these two [122] it was even the best-performing FF. In the latter study, AMBER99SB-UCB turned out to be by far best suited for modeling Aβ [2]. This FF is based on AMBER99SB/TIP4P-Ew and includes modified backbone torsion parameters [123] and optimized protein–solvent Lennard-Jones parameters [124]. A special reparamterization approach for the dihedral angles was adopted in AMBER99SB-UCB, since it was limited to ϕ, which defines the torsion about the C-N-Cα-Cβ atoms, in order to only shift the equilibrium between the β and PPII states, but leave the α-helical state unaffected [123]. The mentioned FFs employed a universal refitting strategy for dihedral parameters. An alternative approach is to optimize these parameters in a residue-specific manner. This approach was adopted for RSFF1 [125], which is based on OPLS/AA, and RSFF2 [126], which derives from AMBER ff99SB. In both cases the dihedral distributions from a protein coil library were used as the training set.

3.6.2 *Adding CMAP Corrections*

The majority of the CHARMM protein FFs that are based on CHARMM22* include CMAP corrections [127, 128]. The CMAP residue-specific correction for backbone dihedral parameters is a grid-based energy correction map dependent on the (ϕ, ψ) distribution of the backbone dihedrals of the protein residues. The two-dimensional (ϕ, ψ) angle distribution per residue is evenly divided into 24×24 bins with a 15° step size between neighboring bins. The dihedral free energy for bin *i* is given by

$$\Delta G_i^{sim} = RT \ln \left(\frac{N_i}{N_{max}} \right) \tag{15}$$

where N_i is the number a dihedral angles falling into the bin in question and N_{max} refers to the total number of dihedral data in the sampling. The CMAP correction for

each bin is provided as difference of the experimental database value (ΔG_i^{exp}) and the current simulation value from equation 15:

$$U_i^{CMAP} = \Delta G_i^{exp} - \Delta G_i^{sim} \tag{16}$$

Thus, the energy added is the larger, the larger the deviation from the database value is. Since a step size of 15° yields only 576 bins, nearest-neighbour [129] or cubic interpolation [130] is applied to generate a continuous energy-correction surface so that U_{CMAP} can be determined for any conformation and allow the computation of forces.

The CMAP method was first applied in CHARMM22/CMAP[128], which is also known as CHARMM27 and is based on CHARMM22 [121]. However, CHARMM27 did not produce convincing results for IDPs as demonstrated for α-synuclein, where the helical conformation was overestimated, and generally fails to generate a stable hairpin structure. Therefore, the CMAP approach was revised for CHARMM36 [131]. This newer FF performs generally better for IDPs, however, left-handed helices tend to be overpopulated [106]. To overcome this and other shortcomings, CHARMM36m was developed and claimed to be particularly suited for IDPs [113]. However, the FF benchmark by Robustelli et al. showed that, while CHARMM36m performs well for folded proteins, for many of the IDPs it does not produce convincing results [108]. This excludes Aβ, for which CHARMM36m produced acceptable results [2], which are better than those obtained with AMBER99SB-disp developed by Robustelli et al. [108]. Chen and coworkers picked up the CMAP idea and implemented it into various FFs with the aim to improve the modeling of IDPs. They augmented this approach by deriving CMAP potentials for all 20 standard amino acids, instead of applying a universal CMAP correction in equation (16). They implemented their CMAP corrections into various FFs, yielding AMBER14IDPS [132], CHARMM36IDPS [133, 134], and OPLSIDPSFF [135]. Another extension of the CMAP idea is the CMAP energy correction map based on a three-dimensional distribution of dihedrals that includes side-chain dihedrals, which was implemented into RSFF2 [126] and yielded RSFF2C [136]. This FF was demonstrated to provide good models for IDPs, IDPRs, and also folded proteins [136].

3.6.3 *Refining Protein–Water Interactions*

Between protein and water, electrostatic and vdW interactions can occur. The short-range vdW interactions in particular influence the compactness of the simulated IDP conformations, which can be assessed by the radius of gyration (R_g) or the end-to-end distance and be compared to the corresponding values obtained from SAXS or FRET experiments [137, 138]. Since the observation was that the common FFs in general produce too compact IDP conformations, one strategy for the development of IDP-appropriate FFs is to increase the vdW interactions between protein and water. The first approach along this line was realized by Best et al. who uniformly scaled the LJ interactions between protein and water by a factor of 1.1 in AMBER03, leading to AMBER03WS, which recovered the correct dimensions of IDPs or unfolded proteins in their simulations [139]. Shaw and coworkers approached the problem by increasing the ϵ_O value of the oxygen atom of water in the TIP4P water model, yielding TIP4P-D where the 'D' stands for dispersion as the authors aimed at providing a correct description of the water dispersion interactions. However, while this water model indeed improved the R_g values of some simulated IDPs, it also caused some α-helices to unfold and overestimated the R_g of several longer IDPs [140]. The recently developed AMBER99SB-disp force field is based on the TIP4P-D water model and increased the dispersion interactions of water even further, in addition to refining the backbone dihedral potentials and LJ interactions between backbone carbonyl oxygen atoms and backbone amide hydrogen atoms [108]. Along with CHARMM36m an alternative TIP3P water model was published [113]. Opposite to TIP4P-D where ϵ_O of water was modified, in CHARMM36m MacKerell and coworkers changed the LJ well depth parameter ϵ_H of the water hydrogen atoms while the oxygen LJ parameters and the water–water interactions were maintained. The rationale behind altering the ϵ_H and not the ϵ_O value is that by changing the water oxygen atom LJ parameters one would affect its effective size based on the repulsive r^{-12} term. Since the water hydrogen atoms have a very small vdW radius, their repulsive LJ term is basically unaffected when modifying their ϵ_H parameter. For the IDPs CHARMM36mW – the name that our group decided to give to CHARMM36m with the modified water

model [141] – performed better than CHARMM36m in the study by MacKerell and coworkers. Head-Gordon and coworkers followed a different approach and adjusted the LJ parameters of the amino acids on atom type basis with the aim to reproduce experimental solvation free energies of a diverse set of 34 molecules, instead of uniformly scaling the vdW interactions between protein and water [124]. This modification along with the one affecting the backbone torsion parameters [123] (see above) resulted in AMBER99SB-UCB. In our recent study, this FF turned out to be the most suitable one for Aβ [2], while for peptide aggregation we identified CHARMM36mW as the best FF [141]. In fact, all force fields that modified the LJ parameters of the oxygen atom of water, but also AMBER99SB-UCB inhibited the process of aggregation of $A\beta(16-22)$, which is in disagreement with experiment [111, 141, 142].

3.7 DIMENSIONALITY REDUCTION TECHNIQUES

In MD simulations, millions of protein configurations are stochastically explored. To generate such huge number of configurations, hundreds/thousands of nanoseconds of molecular simulation need to be performed. In these computer simulations huge amount of MD data in the form of motions of the atoms is generated, which requires various data mining tools to reduce its size. The most important task during data mining is dimensionality reduction. The increase of simulation data increases the number of dimensionalities that were explored in the simulation, therefore classification becomes significantly more complex and time intensive. While larger number of dimensions allows to understand the biomolecular dynamics better, not all of these dimensions/features are crucial for determining the similarity among them [143]. Since, large amounts of the dimensional space is only sparsely populated, features that occur only once or twice do not contribute to the clustering because they do not impact the entire data space significantly. Therefore, the data space can be reduced drastically while maintaining a high number of distinctive features. The major challenge is to find an appropriate dimensionality reduction method, that reduces the input feature space to the smallest but best possible number of dimensions. The best way to increase the performance of clustering, can be achieved either by employing

supervised or unsupervised techniques. The simple implementation is that for each vector with length n dimensions, eliminate i needless features in order to optimize the minimum set of attributes required to match the probability distribution of the clusters as close as possible to the original distribution [144, 145]. The new vector space then consists of $k = n - i$ dimensions and reduces the space from R^n to R^k.

Various supervised or unsupervised approaches exist for mapping the higher dimensional space to a lower dimensional space depending on the input information. Supervised methods can be applied to datasets where the data is already classified or labelled, and the reduction can learn relevant features from the labelled data. A well-known algorithm is for instance Linear Discriminant Analysis (LDA). Concurrently, unsupervised techniques are clustering based and are exploratory in nature. The attempt is to reduce the dimensionality by retaining as much information as possible without any ground truth labels provided and uncover patterns that were unknown previously. Two most popular unsupervised approaches are Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) [145]. Although various dimensionality reduction methods exist, PCA is one of the most popular ones that is applied to MD simulations [146].

Principal Component Analysis

PCA is a linear dimensionality reduction method, first presented by Pearson [147]. Its goal is to find k dimensions that best represent the dataset and concurrently maintain high variance in the data. PCA projects m data points $x_1, x_2, ..., x_m \in R^n$ to a lower dimensional space R^k by performing the following steps sequentially. (1) First, the data is normalized to have a mean of 0 and a standard deviation of 1. (2) Secondly, the covariance matrix is computed using equation 17 [148, 149],

$$C = \frac{1}{m} \sum_{i=1}^{n} (x_i - \overline{x}) (x_i - \overline{x})^T \tag{17}$$

where T is the transposed vector (that is $n \times 1 \rightarrow 1 \times n$). (3) Thirdly, compute the k orthogonal eigenvectors of $\Sigma u_1, \ldots, u_k \in \mathbb{R}^n$ from the covariance matrix and the corresponding eigenvalues. (4) Lastly, sort the eigenvectors in ascending order of the eigenvalue and determine k to maximize the variance in the remaining k-dimensional space. Figure 7 describes the method to find the maximum variance in a data set.

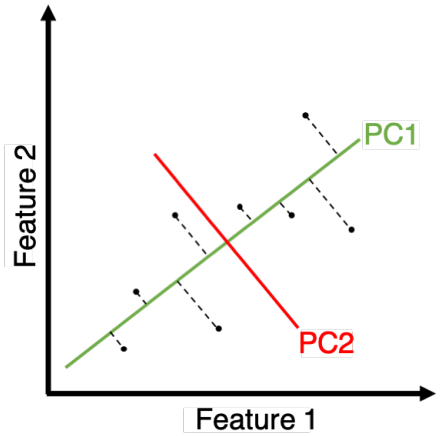

Figure 7: An illustration to determine the maximum variance among the data points in PCA. The first principal component (PC1) represents the axis with the highest variance. The second principal component (PC2) is orthogonal to PC1.

In biomolecular simulations, we use PCA to to determine a small set of collective vectors with the largest contribution to the mean square fluctuations (MSF) of the atomic position of the biomolecule. Using the 3N cartesian coordinates $r_i(t)$ (i=1, ..., 3N), the elements of the covariance matrix of the atomic positions are computed after the translation and rotation motions of the biomolecule are removed from the simulation generated structures by superimposing on the reference structure.

3.8 CLUSTERING ALGORITHMS

The clustering process involves grouping a set of data points into clusters, i.e. collections of data points bearing no pre-defined class labels. The main idea is to assign similar objects to similar clusters, while dissimilar objects do not share the same cluster [145]. Clustering in general, is an unsupervised method since there are no ground truth labels available. Therefore, the number of clusters can highly vary depending on the data, while this number is most likely is unknown in the beginning. Moreover,

groups can either be exclusive, were each data point belongs to exactly one cluster, or hierarchical, such that the top level gives a rough division while each group is refined further [150]. A classification of clustering types and approaches is provided in Figure 8.

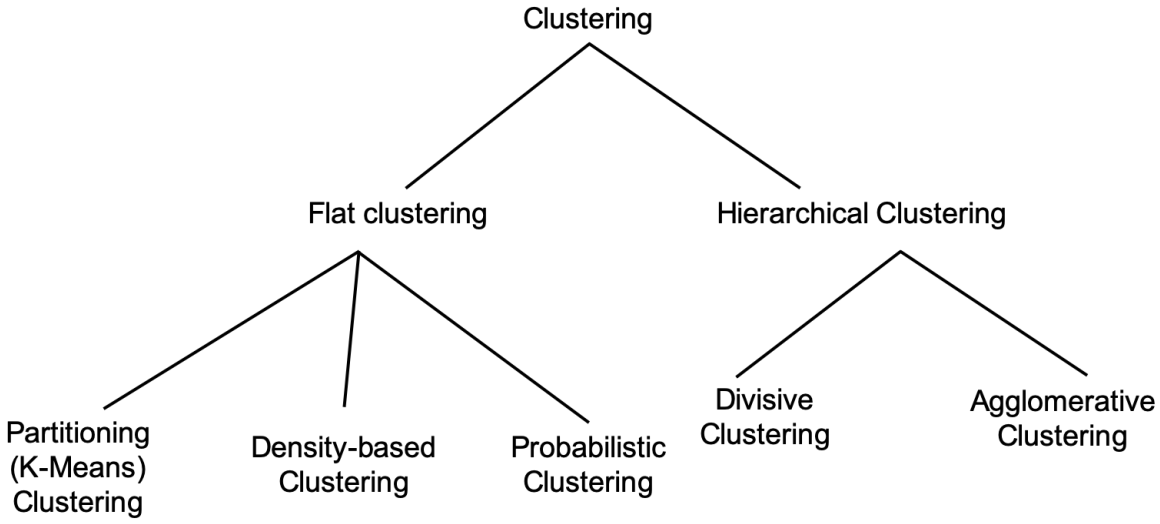

Figure 8: Classification of clustering methods.

3.8.1 *K-means algorithm*

One of the most widely used clustering algorithms, is the K-Means algorithm, developed by Lloyd in 1982 [151]. It has widespread usage due to its simplicity and effectiveness. In K-Means clustering, the data points are partitioned iteratively into k clusters using centroids as representative centre points. The centroid in most cases, is the mean value of the data points within one cluster and is calculated in such a way that the total squared distance from all points to their cluster centres remains minimum [150]. Thereafter, we define the number of clusters k in advance, avoiding the risk of the algorithm converging to a local maximum value [152].

In the K-Means algorithm, following steps are performed: (1) Firstly, k data points are arbitrarily selected as cluster centroids. (2) Secondly, the data points closest to these centroids are assigned to be in the same cluster. (3) The cluster centroids are recalculated based on the newly added data points such that they represent the new

centre of the cluster. The steps (2) and (3) are repeated iteratively until convergence is reached. However, a major drawback of K-Means clustering is that the final clusters are sensitive to the initially selected cluster centroids and shape of the clusters.

3.8.2 *Hierarchical DBSCAN algorithm*

HDBSCAN stands for Hierarchical Density-Based Spatial Clustering of Applications with Noise [153]. It uses a density-based approach which makes few implicit assumptions about the clusters [154]. The basic idea of density-based clustering is that a data space does not only consist of a certain number of data point clusters but also contains a number of outliers, i.e. highly dense regions separated by sparse regions.

The DBSCAN algorithm is defined by two important parameters, the ϵ-radius and $MinPts$ in order to determine the density at a data point q and dense region respectively. The density at a certain data point q is measured by the number of points within a circle of ϵ-radius, likewise a dense region is defined by the minimum number of points ($MinPts$) encircled in the ϵ-radius. The ϵ-radius defines $N_\epsilon(q)$ for the neighbourhood of point q in equation 18,

$$N_\epsilon(q) = \{p \in D \mid dist(p, q) \leqslant \epsilon\} \tag{18}$$

where D represents the dataset. The step-by-step protocol followed in the algorithm is: (1) First, fix a value for $MinPts$ which ranges between [1, (2d-1)], where d is the dimension of the dataset. (2) Secondly, compute the k-distance for all points p in the dataset D with $k = minPts$ and create a k-distance plot showing the k-distances of all objects in decreasing order. (3) Lastly, select the border point (i.e., a point which has fewer than $MinPts$ within ϵ, but is in the ϵ-radius neighborhood of a core point), o from the $MinPts$-distance plot and set to $MinPts$-distance(o) [154, 155]. After selecting the parameters, a point q is taken as the core point of initialisation, (i.e., a point which has more than a specified number of points $MinPts$ within ϵ-radius). A point p in the data space is directly density-reachable from q if $p \in N_\epsilon(q)$ and q

is core object w.r.t. ϵ and MinPts, i.e. $\left|N_{\epsilon(q)}\right| >= \text{MinPts}$, so if there is a chain of points $p_1, \ldots, p_n$ with $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i. Both the points p and q are density-connected if there is a point o which is density-reachable from both points.

The algorithm starts with an initialisation point q and retrieves all points density-reachable from q with respect to ϵ and MinPts. So, if q is marked as a core point, the algorithm defines a cluster as a set of points which are density-reachable from q. In other words, if this point contains MinPts within ϵ-radius neighborhood, cluster formation starts, otherwise the point is labeled as noise point (i.e., a point which is neither a core point nor a border point), and can then become part of another cluster. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database until each point is labeled. The concepts of density-reachable and density-connected are explained in Figure 9,

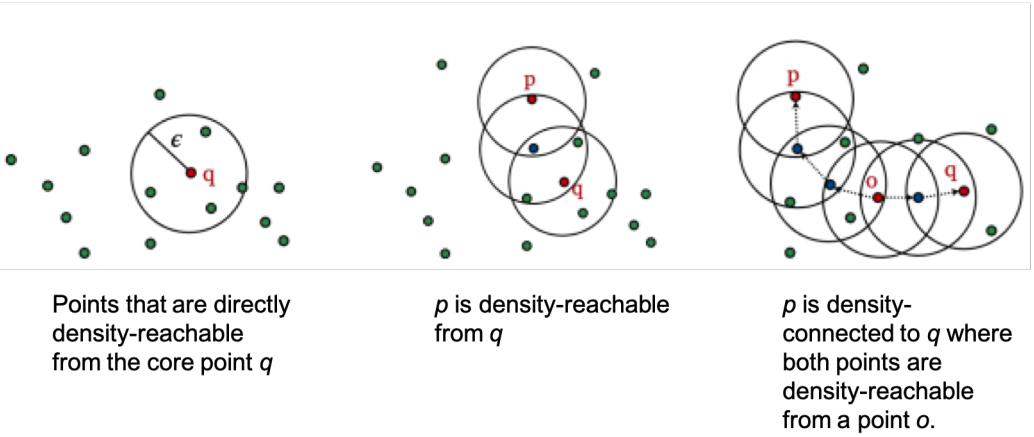

Figure 9: An illustration of the DBSCAN algorithm (Reproduced with permission from Ester et al., 1996. [154] Copyright ©1996, AAAI).

One of the major advantages of density-based methods is the flexibility of data in the n-dimensional space. In this space, clusters can have any shape and are not restricted to convex shapes. However, DBSCAN fails to identify clusters of varying densities. Therefore, the combination of DBSCAN with a hierarchical clustering, which yields HDBSCAN, was developed. Here, whereby performing DBSCAN is performed for varying ϵ values, helping to define clusters with stable ϵ.

3.8.3 *Geometric clustering*

Contrary to the hard-clustering methods of K-Means and DBSCAN, geometric clustering approaches are also widely used in MD simulations. The most commonly used geometric clustering algorithm is the GROMOS clustering algorithm developed by Daura and coworkers [156] and implemented in the GROMACS software package [157]. The GROMOS algorithm is a heuristic approach were each iteration tries to find the largest cluster given a specific root mean square deviation (RMSD) cutoff size until all configurations or snapshots sampled in a MD simulation belongs to a cluster. This produces clusters where all members of a cluster are within the specified RMSD cutoff of the centroid model. All the clusters generated are mutually exclusive, so a configuration can only be a member of a single cluster.

3.9 KINETIC MODELS

A major advance in the realm of alternative methods to analyze MD data has been made by the proposal of transition matrices and discretization of the phase space. This approach improves the MD sampling efficiency by widening its focus to include, in addition to identifying the energy minima, the simulation and definition of informative transition pathways. The transition process between substates is often described by a memoryless master equation,

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{K}\mathbf{p}(t) \tag{19}$$

where $\mathbf{p}(t)$ is a column vector contains the probability of finding the system in each of its m states at time t, $\mathbf{K}$ is the rate matrix with its elements K_{ij} being the rate constants of transitions from state i to state j. In a Marokovian process, the system dynamics can be described by a discrete time transition matrix $\mathbf{T}(\tau)$ as explained below.

3.9.1 *Markov State Model*

Markov state models (MSMs) derived from MD data are kinetic models used to define the kinetically relevant states in the configuration space and the transitions (memoryless jumps) between them [158–160]. In other words, MSMs transform the MD trajectory from structures over time to a human readable network of macrostates that enables extracting the hidden kinetics in the high dimensional MD simulation data. The output of MSMs is Markovian, i.e., the transition to the next state is only affected by the current state and not the past history, thus it is memoryless.

In a Markovian process, the system dynamics can be described by a discrete time transition matrix $\mathbf{T}(\tau)$ with its entries T_{ij} representing the probability of finding the system in state j at time $t+\tau$ knowing that it was in state i at time t. In accordance with equation (20) this can be expressed as

$$\mathbf{p}((k+1)\tau) = \boldsymbol{T}(\tau)\mathbf{p}(k\tau) \tag{20}$$

where τ is the lag time selected by the user to determine $\mathbf{T}(\tau)$. Both equation (19) and (20) give equivalent result at $t = k\tau$, and are related to each other by $\mathbf{T}(\tau) = \exp(\tau\mathbf{K})$. The focus of MSMs is on the transition matrix $\mathbf{T}(\tau)$, which describes the transitions between substates, with its eigenvectors $\mathbf{u}_i$ and corresponding eigenvalues λ_i. They are a measure for the relaxation time of a process described by the eigenvector in question, i.e., $\lambda_i \to 0$ is the fast mode corresponding to fast thermal fluctuations, while $\lambda_i = 1$ represents the stationary distribution. The implied timescale t_i^* of a transition mode i is given by

$$t_i^* = -\frac{\tau}{\ln \lambda_i} \tag{21}$$

For equilibrium molecular dynamics, all λ_i are positive, and there exists a unique equilibrium (or stationary) distribution π that fulfills detailed balance:

$$\pi_i T_{ij} = \pi_j T_{ji} \tag{22}$$

The first step in constructing MSMs is to discretize the configuration space based on a suitable distance metric selected according to the scientific question of interest. For example, in studies interested in conformational dynamics or protein folding, the backbone atom coordinates can be used to define the conformations. This is followed by clustering, such as k-means clustering to define the microstates. Even though the idea is to identify the kinetically stable states, initially geometrical clustering is used to define microstates based on which the subsequent kinetic partitioning of the configuration space is performed. The transition matrix described above defines the kinetic relationship between these microstates. The quality of the resulting MSM depends on the chosen the lag time. It should be long enough so that the system is memoryless (Markovian), while it should be short enough to resolve the different dynamic processes. Whether the models satisfies Markovianity can be assessed by the Chapman-Kolmogorov test [161]:

$$\mathbf{T}(k\tau) = \mathbf{T}^k(\tau) \tag{23}$$

Finally, the MSM can be further optimized by coarse-graining it, for which macrostates are defined and the microstates assigned to them. This yields so-called Hidden Markov Models.

A number of software packages are available to construct MSMs from MD trajectories of biomolecular systems and validate the resulting models. The most popular one are PyEMMA [162] and MSMBuilder [163]. Recent applications of MSM-based methods in the realm of IDPs have shed light on the mechanisms of fibril formation for amyloid-β (Aβ), human islet amyloid polypeptide (hIAPP), and other intrinsically disordered peptides [164, 165]. It further gives insight into the thermodynamics and kinetics of IDPs, such as the amyloid-β peptide [2, 166], and are useful in studying the effect of post-translational modification (e.g., phosphorylated kinase [167]) on the conformational kinetics and structural ensembles of IDPs.

3.9.2 *Transition Network Model*

In contrast to MSMs, transition networks (TNs) rely only on geometry to define the transition pathways [168]. In this approach, the configuration space is discretized based on clustering using certain descriptors (f_i) that map the geometrical conformations $\mathbf{q}_t$ to a state $S(t)$ which is a composition of different f_i:

$$S(t) = [f_1(\mathbf{q}_t), f_2(\mathbf{q}_t), \ldots, f_n(\mathbf{q}_t)] \tag{24}$$

The conformations of an MD trajectory are mapped to the time sequence of states $S(t)$, which are then used for building the TNs. The key aim of f_i is to encode the process under investigation the best, while avoiding network complexity. These descriptors can be measures that give information about the protein secondary structure, the number of electrostatics or hydrophobic contacts, or molecular shape. Importantly, the selected descriptors should be limited to a combination of f_i with minimum correlation among them and maximum sensitivity to the data to avoid increasing the complexity of TNs. For example, if two descriptors are strongly affecting each other (i.e., highly correlated), using both of them will increase the complexity of the generated TN without a major information gain. Thus, using only one of them simplifies the resulting TN and, at the same time, not too much information will be lost. The descriptor sensitivity can be inferred from whether the theoretically possible values of f_i are indeed sampled (which is good) or only limited to a small range (which would be insufficient).

After discretizing the configuration according to equation (24), the transition matrix can be built by counting the number of transitions between different sates $S(t)$ using a certain time lag (spacing) between MD frames. The spacing between successive MD frames should be chosen such that it is not too large in order not to lose important transitions, but also not too small as larger conformational transitions require a certain amount of time. However, unlike to MSM building, there is no quantitative measure such as the convergence of the implied time scales in MSMs that allows to assess whether the selected lag time for TN building is a good choice.

4

STRUCTURE AND DYNAMICS OF AMYLOID-β PEPTIDE

4.1 DECLARATION

Parts of this chapter were submitted by the author to the *Chemical Science* as *Paul, A., *Samantray, S., Anteghini, M., Khaled, M., Strodel, B. (2021). Thermodynamics and kinetics of the amyloid-β peptide revealed by markov state models based on MD data in agreement with experiment (* *equal authorship*). All the images are reprinted with Copyright ©2021 **Royal Society of Chemistry**.

4.2 MOTIVATION

It has been 15 years since the first all-atom MD simulation of full-length Aβ in solution was performed [169]. Since then, hundreds of simulation studies involving Aβ40 or Aβ42 have been published. During the first ~10 years of these studies, research groups relied on the common protein and water FFs as was usual practice for the simulation of folded proteins, which were the more frequent object of simulation studies at that time. However, as time elapsed, very different results regarding the structural preferences of Aβ were obtained depending on which of the FFs was used. In fact, as far back as 2012 when we had first reported our first simulation study of Aβ in solution we concluded that 'proper benchmarking of the protein FFs for unfolded and intrinsically disordered proteins' was needed [170].

Over the years, various FF benchmarks for Aβ and IDPs in general have been performed [122, 171–174]. Depending on the FFs tested, the simulation technique employed (i.e., standard MD versus enhanced-sampling MD, like replica exchange MD), the length of the simulations, and the experimental data used for validation, different

FFs were identified as the most suited ones. For instance, García et al. [171, 172] found OPLS-AA [175, 176] with the TIP3P water model [177] and AMBER99SB [178] with the TIP4P-Ew water model [179] to be the best FFs for Aβ42, while our own benchmark, that included AMBER99SB [178], AMBER99SB*-ILDN [180, 181], AMBER99SB-ILDN-NMR [182], and CHARMM22*(C22*) [183] combined with TIP4P-Ew [179] as well as OPLS-AA/TIP3P [175–177], identified C22* as the best FF [122]. However, more recent benchmarks revealed that the common FFs from the AMBER, GROMOS, OPLS, or CHARMM family in combination with standard three- or four-point water models produce conformational ensembles for IDPs that are too compact and too biased towards folded structures [106, 140]. This conclusion can be explained with the parametrization strategy underlying the standard FFs, which aimed at producing the correct structure for folded proteins [184], while IDPs do not adopt a well-defined equilibrium structure in solution, instead sampling an ensemble of fully and/or partially disordered structures.

Recent studies include modification the existing FF parameters such that they capture the structural diversity and flexibility of IDPs, producing less folded and more expanded IDP conformations. These FF modifications include strengthening the water-protein London dispersion interactions [139, 140], refining the protein backbone parameters to create more expanded structures or reducing the tendency for certain ordered conformations [185], and/or altering the salt-bridge interactions [185]. In a recent effort, D. E. Shaw Research used AMBER99SB*-ILDN [180, 181] combined with TIP4P-D water [140] as a starting point and reparametrized torsion parameters and the protein–water vdW interaction terms with the aim to develop a FF that provides an accurate model for both folded proteins and IDPs [108]. The performance of the resulting FF, called A99SB-disp, was tested for a large benchmark set of 21 experimentally well-characterized proteins and peptides, including folded proteins, fast-folding proteins, weakly structured peptides, disordered proteins with some residual secondary structure, and disordered proteins with almost no detectable secondary structure. In addition, they assessed the accuracy of six other FF/water combinations. Two of these combinations belong to the older FFs that were developed for folded proteins, A99SB*-ILDN/TIP3P ('A' standing for AMBER) [180, 181] and

C22*/TIP3P ('C' standing for CHARMM) [183] and three combinations specifically designed for IDPs: A03ws containing empirically optimized solute–solvent dispersion interactions [139], A99SB-ILDN/TIP4P-D with increased water dispersion interactions [140], and C36m with refined backbone potentials (which by default is used with CHARMM-modified TIP3P) [185]. The seventh FF in the benchmark is A99SB-UCB, which is based on A99SB/TIP4P-Ew [178, 179] and includes modified backbone torsion parameters [124] and optimized protein–solvent Lennard-Jones (LJ) parameters [123] proposed by Head-Gordon and co-workers. For each of the proteins or peptides in their test set and FFs considered, Robustelli et al. performed 30-μs MD simulations and compared the MD-generated ensembles against a number of experimental data mainly derived from nuclear magnetic resonance (NMR) spectroscopy, and, if available, also from small angle X-ray scattering (SAXS) and fluorescence resonance energy transfer (FRET) [108]. They concluded that, taking all tested proteins and peptides into consideration, A99SB-disp is the best-performing FF. One of the peptides included in their test set is Aβ40. Based on Figure 2 of ref [108], A99SB-disp is, together with C22*/TIP3P, the second best choice following C36m for modeling Aβ40. Interestingly, the performance of A99SB-UCB is not shown in this figure. Though for Aβ40 it was concluded that 'A99SB-UCB produced excellent agreement with experimental NMR measurements [108].

A question that was not addressed by the study of Robustelli et al. is how good the different FFs are able to reproduce the kinetics of the conformational transitions of the different proteins and peptides. From a FRET study of Aβ40 and Aβ42 it was found that both peptides do not exhibit conformational dynamics exceeding 1 μs [186], which also agrees with the findings from fluorescence measurements using the method of Trp-Cys contact quenching [187]. Given the simulation length of 30 μs of the MD data generated by D. E. Shaw Research, we use their simulation data provided by them to assess the kinetics of Aβ40 as sampled by the different FFs. One of our goals is to determine how much MD sampling is needed to reach convergence with standard MD simulations applied to Aβ. To this end, we evaluate the convergence of intrinsic structural quantities as well as of NMR observables calculated from the MD data. Moreover, we generate MSMs, which, in addition to providing conver-

gence checks, also elucidate the kinetically stable states of Aβ and the transitions between them. This analysis reveals that the length of an MD simulation required for obtaining equilibrated results for Aβ depends on the FF used for modeling the peptide, but usually requires at least 20–30 μs or more. Another finding is that only two of the seven FFs under consideration are able to reproduce both the structural and kinetic data available from experiments of Aβ40, which are A99SB-UCB, which performs by far the best, and A99SB-ILDN/TIP4P-D. We thus conclude that it has now become reality to predict the thermodynamics and kinetics of the amyloid-β peptide based on tens of microsecond of MD data using A99SB-UCB as FF.

4.3 MODEL AND SIMULATION DETAILS

4.3.0.1 *MD trajectories*

The 30 μs MD trajectories were generated by Robustelli et al. [108] and kindly provided by the authors. All MD simulations were initiated from an Aβ40 structure similar to that found in PDB entry 1BA6 [188], which was placed in a cubic box with edge length of 6 nm, and run for 30 μs using following FFs: A03ws, A99SB-ILDN/TIP4P-D, A99SB-UCB, A99SB-disp, C22*/TIP3P, and C36m. The simulations were performed at pH 7 (i.e., the His residues were modeled as neutral) with 50 mM NaCl added and a simulation temperature of 300 K. The subsequent analyses was applied to each of the seven trajectories.

For convergence checks some additional simulations were performed. For A99SB-disp and C36m the simulations were extended to 35 μs. To this end, we extracted the last snapshot of the corresponding 30 μs MD simulation and used it as starting structure for the additional 5 μs MD simulation. As MD software we employed GROMACS version 2018.3 [157] in combination with either A99SB-disp [108] or C36m [185]. The external conditions were chosen as in the original simulation: 300 K, 50 mM NaCl, pH 7. The peptide was placed in a cubic simulation box with an edge length of 6.0 nm – as chosen by Robustelli et al. [108] – and solvated with the corresponding water model and NaCl added, making sure to also neutralize the system.

Before the production MD run was started, the energy was minimized using the steepest descent algorithm, followed by equilibration, first in the NVT ensemble with position restraints on the non-hydrogen atoms of Aβ40, afterwards in the NpT ensemble without position restraints. The 5 μs production runs followed, which were realized in the NpT ensemble using a velocity rescaling thermostat with canonical sampling [189] with a 0.1 ps time constant for coupling and a Parrinello-Rahman barostat [190] with a relaxation time of 2 ps. All bonds involving hydrogen atoms were restrained using the LINCS algorithm [191], which enabled a time step of 2 fs for the integration of the equations of motion. The electrostatic and van der Waals interactions were calculated using the particle mesh Ewald method [192] in conjunction with periodic boundary conditions and a real-space truncation at 1.2 nm.

4.3.0.2 *Structural analyses*

The trajectories were analyzed using a combination of standard Gromacs-2016.4 package tools [193, 194], custom written Tcl scripts in VMD [195], and Python scripts using the MDAnalysis [196] and MDTraj [197] libraries. As the trajectory files from the Desmond MD package [198] are in DCD file format, there was a need for conversion into Gromacs-compatible TRR format. After this, protein conformations were clustered using the clustering algorithm of Daura et al. [156] as implemented in Gromacs with a RMSD cutoff of 1.0 nm. To assign secondary structure elements to the protein conformations, the STRIDE algorithm [199] was employed. Inter-residue contact maps were constructed by calculating the fraction of structures in which the residue pairs were having at least one pair of atoms within 0.4 nm of each other.

4.3.0.3 *Construction of Markov state models*

The underlying kinetics of the systems are captured by constructing MSMs from the MD simulation data using the PyEMMA library in Python [200]. The first step towards building an MSM is to choose a suitable distance metric, called feature, for defining the conformational space of the molecule. Here, we describe the conformations in terms of the distances between the C_α atoms. This feature was selected based on VAMP-2 scores, where VAMP stands for Variational Approach for Markov

Process [201]. This score is part of the VAMP scores family, which represents a set of score functions that can be used to find optimal feature mappings and optimal Markovian models of the dynamics from time series data. In order to choose a subset of relevant features for our model construction, we considered three different features: C_α distances, minimum distance between residues, and backbone torsion angles. In order to evaluate which feature is the best and to avoid overfitting, a cross-validation was performed, comparing the VAMP-2 scores of each of the three features computed for three subtrajectories of 10 µs length per FF. From this analysis, the C_α distances emerged as the most suitable feature (Figure 10).

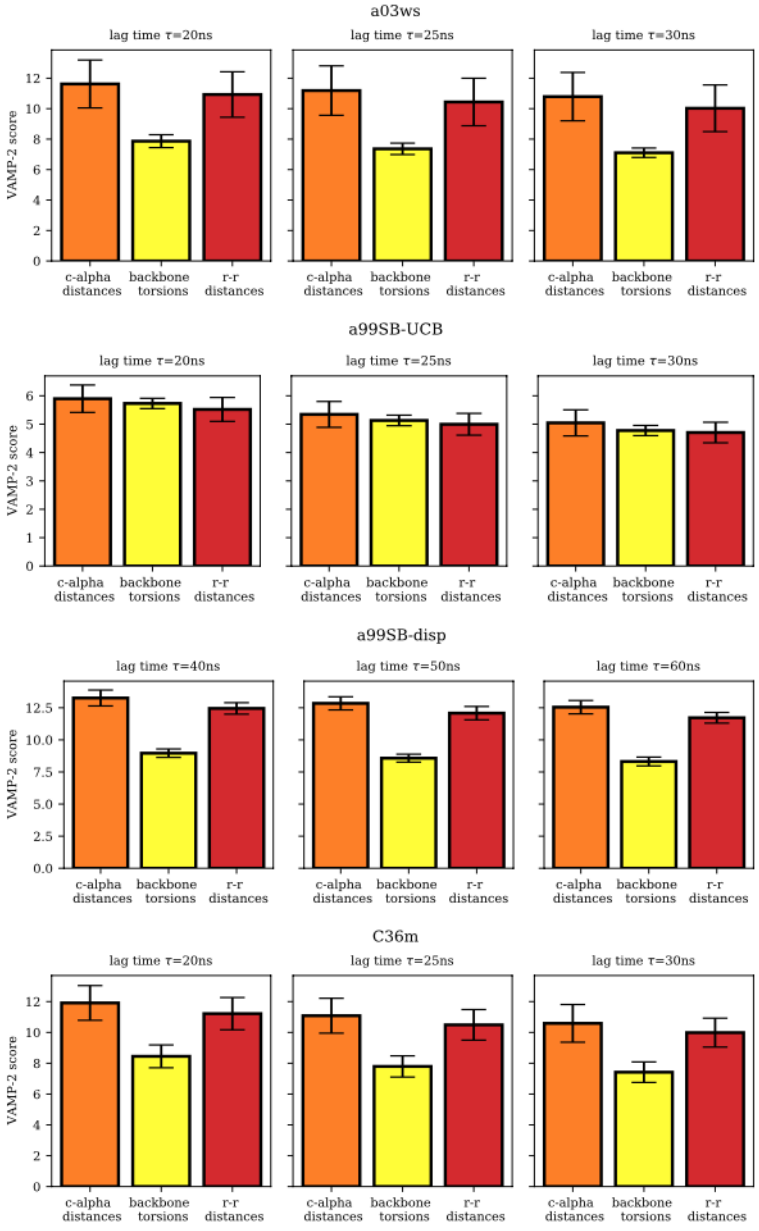

Figure 10: VAMP-2 scores for four FFs (A03ws, A99SB-UCB, A99SB-disp, C36m), three different features (C_α distances, backbone torsion angles, residue–residue distances), and three different lag times. The choice of these lag times was based on the lag times used for building the final MSMs: $\tau = 25$ ns for A03ws, A99SB-UCB, and C36m and $\tau = 50$ ns for A99SB-disp. The scores are the highest for the C_α distances in all the cases (the higher, the better), while they are the smallest for the backbone torsion angles for almost all combinations. We thus chose the C_α distances as the feature for the MSM analysis.

Next, we reduced the dimension of the space from 703 interatomic distances to 2 collective coordinates by applying time-lagged independent component analysis (TICA), a dimensionality reduction technique that identifies the slowest modes in the feature space by maximizing the autocorrelation of the reduced coordinates [202],

and hence is preferred for MSM construction over the more commonly used PCA, which does not take into account any kinetic information. A density based clustering technique, HDBSCAN [203] is then applied to decompose the reduced conformation space into a set of disjoint discrete states and define the trajectory as a sequence of transitions between these states. An MSM can next be built from this discrete trajectory by counting the transitions between the states at a specified lag time (chosen as 100 ns in this chapter), constructing a matrix of the transition counts and normalizing it by the total number of transitions emanating from each state to obtain the transition probability matrix. In the case of A03ws, the resulting MSM is further coarse-grained into a hidden Markov model (HMM) using the robust Perron cluster analysis (PCCA+) [204], which is a fuzzy version of the spectral algorithm for partitioning graphs that assigns each microstate a probability of belonging to a metastable macrostate [205]. For the other FFs this coarse-graining step was not required as the MSMs and HMMs turned out to be identical. Finally, whether the final models satisfy the Markovian assumption is verified with a Chapman-Kolmogorov test [206].

4.3.0.4 *Calculation of experimental observables*

The NMR chemical shifts of the protein backbone atoms were calculated using the SPARTA+ [207] software package. Dihedral angles (ϕ , ψ) were calculated for each Aβ residue from the MD trajectories and converted into residue-specific backbone scalar $^{3}J_{HNH\alpha}$ coupling constants using the Karplus equation [208] with the coefficient values $A = 7.97$ Hz, $B = -1.26$ Hz, and $C = 0.63$ Hz from Vögeli et al.[209]. The simulated and experimentally derived coupling constants were compared by calculating χ^2 using equation (25), including the error term $\Delta = 0.42$ Hz [139, 186]:

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(\langle J_i\rangle_{sim} - \langle J_i\rangle_{exp}\right)^2}{\Delta^2} \tag{25}$$

Here, J_i represents the J-coupling constant for the i-th residue, N is the total number of residues for which the experimental data are available, the subscripts 'sim' and

'exp' correspond to the simulated and experimental data respectively, and $\langle\cdot\rangle$ denotes the ensemble average.

Time-series of the end-to-end distance R_{ee} were calculated as the distance between the C_α atoms of the N- and C-termini of Aβ40 using the MDTraj library [197] in Python. From this, the FRET efficiency is calculated as

$$E_{FRET}(t) = \frac{1}{1 + (R'_{ee}(t)/R_0)^6} \tag{26}$$

where the Förster radius $R_0 = 5.2\,\text{nm}$ for the dye pair of Alexa 488 and 647 was used [186, 210], and R'_{ee} is calculated by scaling up R_{ee} to account for the effects of the experimental dyes by treating them as 12 extra amino acid residues and assuming a Gaussian scaling exponent [186],

$$R'_{ee}(t) = R_{ee}(t)\left(\frac{N+12}{N}\right)^{0.5} \tag{27}$$

where $N = 40$ is the number of residues in the peptide under study.

4.3.0.5 *Bayesian reweighting of trajectories by using experimental data*

The Bayesian/maximum entropy (BME) technique [211] was used to reweight the trajectories and obtain a refined conformational ensemble consistent with selected experimental data, thereby compensating for the discrepancies between the experimental and calculated observables which arise from inaccuracies in the FFs. Here, we considered the J-coupling data to obtain the optimized set of weights for the A99SB-UCB and C36m trajectories.

4.4 THERMODYNAMICS AND KINETICS OF Aβ40 PEPTIDE

We used the 30 μs MD data of Aβ40 from Robustelli et al. [108] to i) assess the convergence of these trajectories, ii) determine the agreement of the simulated Aβ40 ensembles with spectroscopic data, and iii) derive the thermodynamics and kinetics of this peptide. The convergence was tested for various structural properties that are usually calculated from MD trajectories, including the structural RMSD, clustering

analysis, radius of gyration (R_{gyr}), and the secondary structure (section 4.4.1). Another kind of convergence check is provided by Markov state models (section 4.4.2), which is based on kinetically clustering the MD data. Section 4.5.1 contains the calculation of NMR spectroscopic and FRET observables which allows us to compare the MD generated structural ensembles with experimental findings and to also assess the convergence of these spectroscopic quantities. In section 4.5.2 we evaluate the kinetics of Aβ40 and compare the MD results to experimental observations. The structural ensemble of Aβ40 in agreement with the thermodynamic and kinetic data derived from experiments is examined in the Discussion following thereafter.

4.4.1 *Check of Structural Convergence from MD trajectories*

4.4.1.1 *RMSD*

As commonly done with MD data, we calculated the C_α-RMSD of the 30 μs MD trajectories with respect to the starting structure of these simulations. From the time evolution of the RMSD shown in Figure 11 one can see that the Aβ40 conformations quickly move away from the initial conformation, reaching values of 1 to 2 nm within a few nanoseconds, and never return to the starting conformation. This is understandable as the MD simulations were initiated from a helical Aβ40 structure determined in a water-SDS micelle medium [188], which is not preferred in water. One can further observe that within the 30 μs of sampling the RMSD does not considerably further increase but strongly fluctuates between ≈1 and 2 nm. Because of the RMSD fluctuations around an average value, one might easily but incorrectly be tempted to conclude that the simulations converged within a few nanoseconds. As our analyses will show in the following sections, this would have been a wrong conclusion. In fact, for Aβ40 and by extension other IDPs, RMSD values happen to be useless for judging whether an MD simulation has reached convergence.

4.4.1.2 *Number of clusters*

Next, we calculated the number of conformational Aβ40 clusters using the clustering method of Daura et al., which involves the calculation of the RMSDs between all possible MD snapshot pairs – and not only with respect to the MD starting structure as done above – and the identification of similar structures within a certain RMSD cutoff [156]. Given the large flexibility of Aβ40 as visible from the RMSD fluctuations in Figure 11, a cutoff of 1.0 nm was chosen. The results in Figure 11 show that for almost all FFs more than 10 μs of MD sampling is needed before the curves for this quantity converge, which implies that from this time onward all conformations sampled can be assigned to an already identified cluster. But for several of the FFs even beyond 20 μs new conformations are still sampled. Only with A03ws, A99SB*-ILDN/TIP3P and C22*/TIP3P no new conformations were found beyond ~10 μs of MD sampling. Another difference between these and the other FFs is that the total number of clusters is considerably smaller. With A99SB*-ILDN/TIP3P, less than 5 conformational clusters were identified, whereas with C36m more than 30 clusters were sampled. Thus, the different FFs predict different degrees of Aβ40 structural flexibility and the FFs associated with higher conformational diversity were generally observed to require simulation times longer than 20 μs for attaining convergence. It should be noted that due to the use of a relatively large RMSD cutoff, the different clusters considerably vary from each other structurally. Put differently, this implies that transitions from one cluster to another involve non-negligible conformational changes.

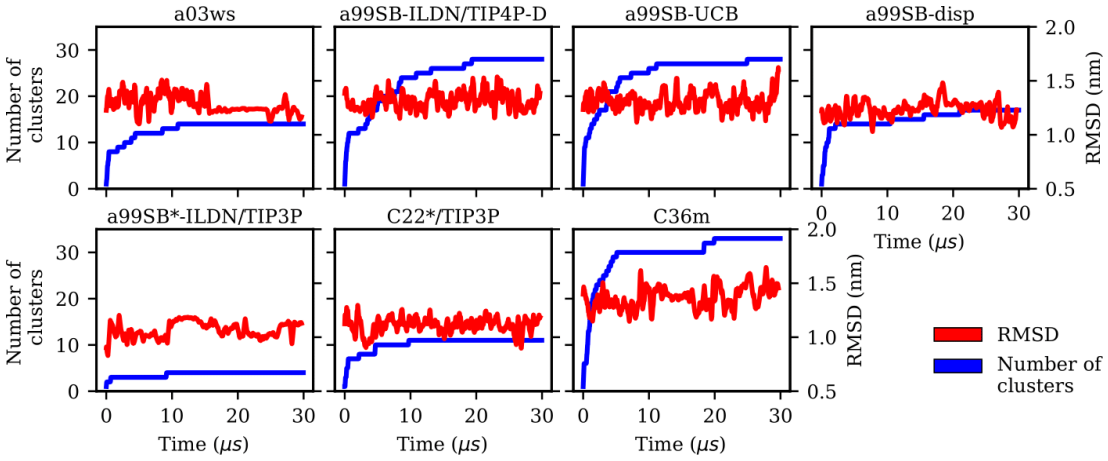

Figure 11: Evolution of the C_α-RMSD with respect to the Aβ40 starting structure of the MD simulations (red, right y-axis) and the number of conformational clusters (blue, left y-axis) for the different FFs (labels on the top of the panels).

4.4.1.3 *Radius of gyration*

The assessment of R_{gyr} (Figure 12) reveals that this quantity equilibrates quickly and in most cases did not considerably change after 10 μs. This applies especially to A99SB-ILDN/TIP4P-D and A99SB-UCB for which more than 20 μs of simulation time is needed for the number of sampled clusters to converge; the two corresponding R_{gyr} distributions however do not change after a simulation time of 10 μs. It is only with A03ws that a considerable change in the R_{gyr} distribution occurs after 10 μs. It changes from a broad distribution with a maximum value of ~1.5 nm to a predominantly narrow distribution with a distinct peak at around ~1.1. Thus, a considerable conformational transition must have happened shortly after 10 μs, leading to a conformation considerably different and more compact to all previously sampled conformations. Figure 11 shows that once this structure was identified, no further new structures were sampled as the number of clusters did not rise after ~11 μs in the case of A03ws. Comparison of the R_{gyr} distributions obtained for the different FFs reveals that quite different structural ensembles are produced: only compact structures with A99SB*-ILDN/TIP3P, compact and also elongated structures with A03ws, A99SB-disp, C22*/TIP3P and C36m, and mostly elongated structures with A99SB-ILDN/TIP4P-D and A99SB-UCB.

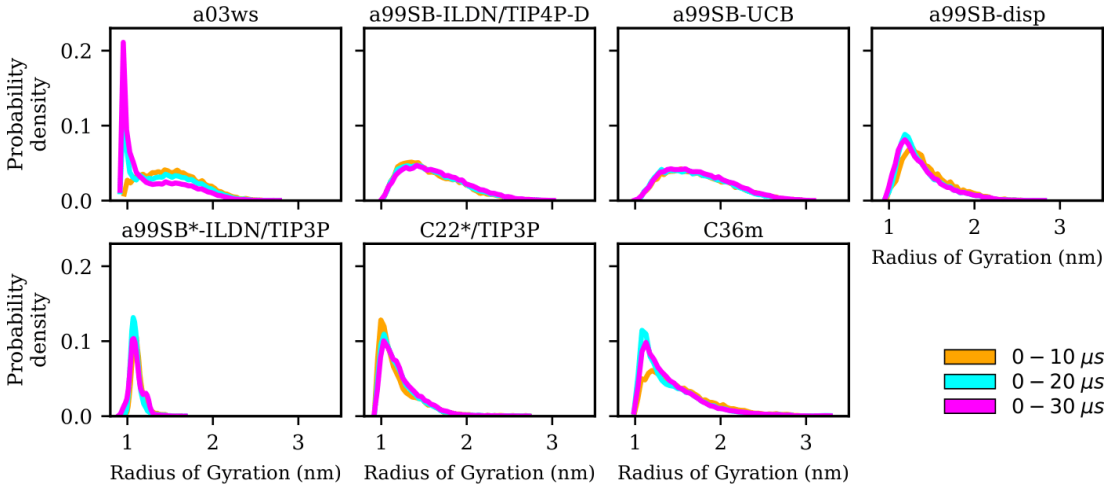

Figure 12: Distribution of the radius of gyration R_{gyr} for increasing trajectory lengths (0–10 μs: yellow, 0–20 μs: cyan, 0–30 μs: magenta) for the different FFs (labels on the top of the panels).

4.4.1.4 *Secondary structure*

Further information on the structural preferences obtained for the different FFs is available by analyzing the secondary structure. Figure 14 shows the evolution of the amounts of α-helix and β-sheet found in Aβ40, while in Figure 13 the time averages for the secondary structure elements can be seen. After 10 μs these time averages are largely converged apart for A03ws. In the latter case, a gradual increase in helix is observed till the end of the simulation, whereas for the other FFs a slight increase in β-sheet is seen for sampling times above 10 μs. The propensity for β-sheet formation differs with the FFs: A99SB-ILDN/TIP4P-D and A99SB-UCB predict a low amount of β-sheet close to zero, a medium amount of β-sheet is found for A03ws, A99SB-disp, C22*/TIP3P and C36m with values of ~10–15%, while A99SB*-ILDN/TIP3P generates a conformational ensemble with more than 20% of the Aβ40 residues adopting a β-sheet structure. The tendency of Aβ40 to adopt helical structures is close to zero for A99SB-ILDN/TIP4P-D, A99SB-UCB, and C36m, whereas with A03ws almost 20% helical content is observed at the end of the simulations. These differences in secondary structure preferences correlate well with the different R_{gyr} distributions. For example, a high amount of helix and/or β-sheet lead to compact structures as observed for A03ws and A99SB*-ILDN/TIP3P, whereas low amounts of helix and β-sheet imply elongated structures as is the case for A99SB-ILDN/TIP4P-D and A99SB-UCB.

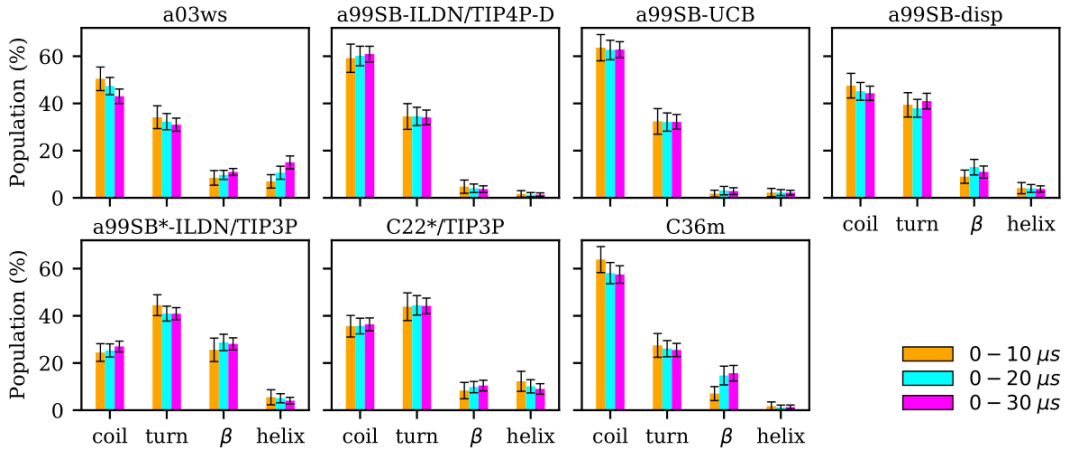

Figure 13: Time-averaged secondary structures coil, turn, β-sheet, and α-helix for increasing trajectory lengths (0–10 μs: yellow, 0–20 μs: cyan, 0–30 μs: magenta) along with standard errors (shown as black bars) for the different FFs (FF labels on the top of the panels).

It is interesting to not only assess time averages for the secondary structure but also its evolution. Figure 14 reveals that within 10 μs – the time window chosen for averaging – considerable changes in secondary structure can occur. This applies to all FFs yet to different extents. The smallest changes occur for A99SB-ILDN/TIP4P-D and A99SB-UCB, i.e., the two FFs which generally predict a low tendency for Aβ40 to adopt a helical or a β-sheet conformation. Nonetheless, also for these two FFs rare β-sheet formation is observed, requiring simulation times abobe 10 μs, e.g., at ~12 and 20 μs in the case of A99SB-UCB. Another extreme is C36m for which huge changes in the amount of β-sheet are observed after 10 μs. It is also interesting to correlate the secondary structure changes to the number of clusters, revealing that even within the same cluster considerable secondary structure changes can occur. This is best seen for the C36m simulation at ~5–20 μs. During this period, the number of clusters is constant, while the amount of β-sheet formed within Aβ40 first varies widely between 0 and 20%, then increases to more than 40% between 13 and 14 μs, followed by a decrease until no more β-sheet is present at around 18 μs. Thus, the sole analysis of time-averaged secondary structures would have been misleading and should always be augmented by further structural analysis in the case of an IDP as flexible as Aβ.

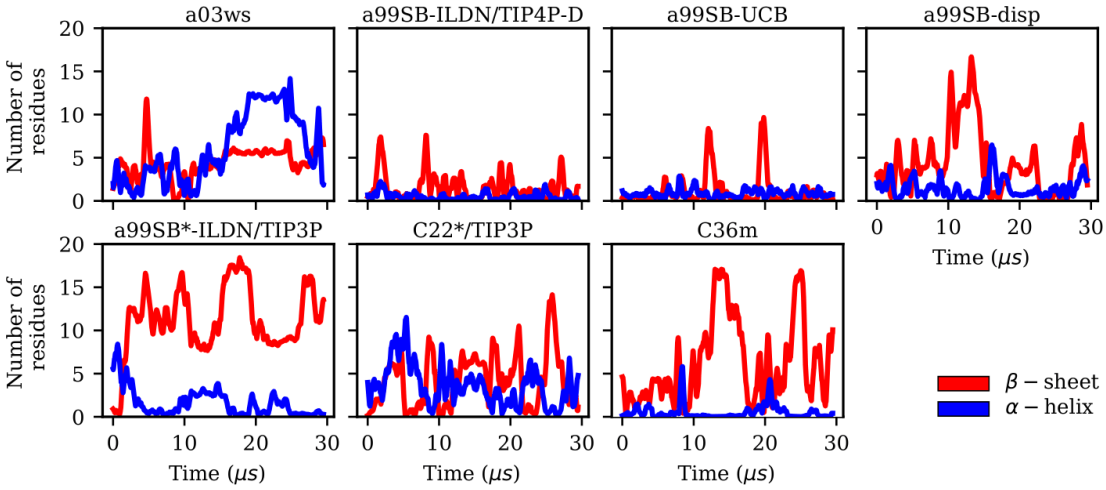

Figure 14: Evolution of the secondary structures β-sheet (red) and α-helix (blue) in terms of the number of Aβ40 residues adopting these structures for the different FFs (FF labels on the top of the panels).

4.4.2 *Convergence checks of MD Simulations using Markov State Models*

4.4.2.1 *Sample density in TIC space*

Another test to determine whether or not the MD data has converged is possible by calculating an MSM; this usually requires the performance of a Chapman-Kolmogorov test for the level of agreement between the MSM predicted dynamics and the actual protein dynamics. Several steps in the construction of an MSM involve dimension reduction and often implies TICA as used in this study. The projection of the MD data along the first two time-independent components (TICs) can be seen in Figure 22. To assess the evolution of the Aβ40 conformations in TIC space, we projected the data from 10 μs time windows. For each of the FFs one can see that different conformational spaces are sampled for the different time windows. Nonetheless, for all FFs but A99SB*-ILDN/TIP3P the explored spaces partially overlap. To verify that no no new structures are sampled for longer simulation times, we extended the simulations for another 5 μs for A99SB-disp and C36m. The projection in TIC space (Figure 15) shows that indeed no new conformations are acquired.

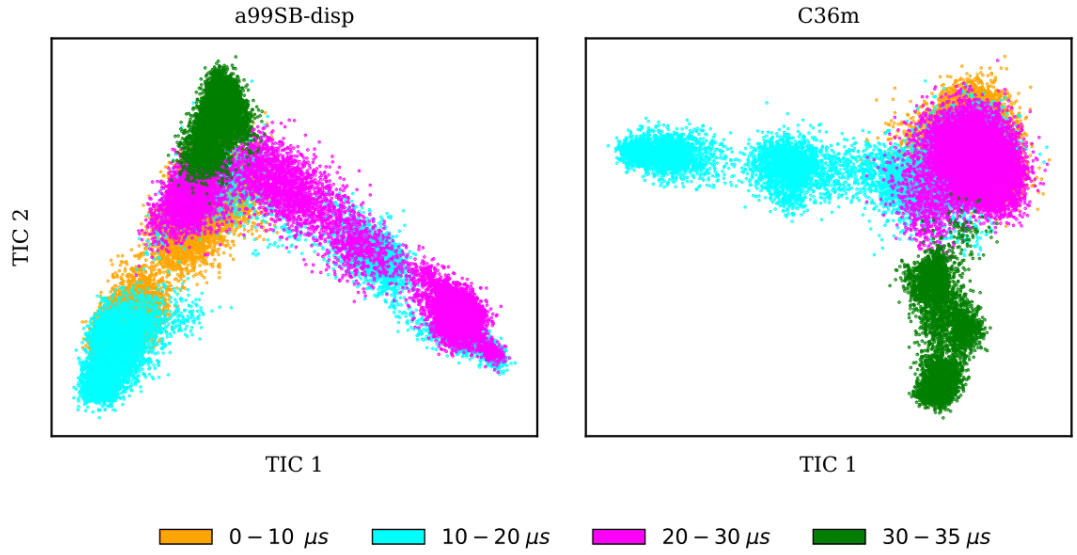

Figure 15: Sample densities for different time windows of the trajectories (0–10 μs: yellow, 10–20 μs: cyan, 20–30 μs: magenta, 30–35 μs: green) projected along the first two TICA components for A99SB-disp (left) and C36m (right).

Taking C36m as an example for a detailed analysis of the sample density, one finds that within the first 10 μs Aβ40 largely remained within the same region of the TIC space. Between 10 and 20 μs it explored new conformations (along TIC 1), which, as revealed by the analysis of the secondary structure, resulted from first a build-up of a β-sheet conformation, followed by its destruction. It should be emphasized that TICA identifies the slowest modes in the feature space and not, unlike PCA, the modes of largest motions. TICA thus confirms the result from Figure 14 that the formation and also disassembly of β-sheets is a slow process in Aβ40, requiring several μs of MD sampling (not only with C36m, but also with.e.g., A99SB-UCB and A99SB-disp). Figure 22 further shows that after 20 μs of MD with C36m a novel conformational transition, evolving along TIC 2, is explored. Comparison with Figure 14 reveals that this process again involves β-sheet formation, followed by its disassembly. In the additional 5-μs trajectory, Aβ40 remained in the energy well corresponding to this β-sheet structure (Figure 15).

4.4.2.2 *Markov state models*

The final MSMs are shown in Figure 23. In the case of A99SB*-ILDN/TIP3P, the convergence of the MD data was not sufficient to allow for the construction of an

MSM. Thus, only for the other six FFs Markov state models are shown. The MSMs for A99SB-ILDN/TIP4P-D and A99SB-UCB are similar to each other as are their previously discussed observables. These two MSMs are dominated by a single state with a population of 95% (state 3 in both MSMs) and two further low-populated states. To characterize the MSM states we calculated the inter-residue contacts for all MD frames assigned to each of the states and averaged the contacts per state (Figures 16–21). The contact maps of states 3 for these two FFs (Figures 17 and 18) involve almost no contacts between residues which are not in neighborhood of each other in the primary structure, i.e., these conformations are mainly elongated structures with large R_{gyr} values and with low amounts of helix and β-sheet.

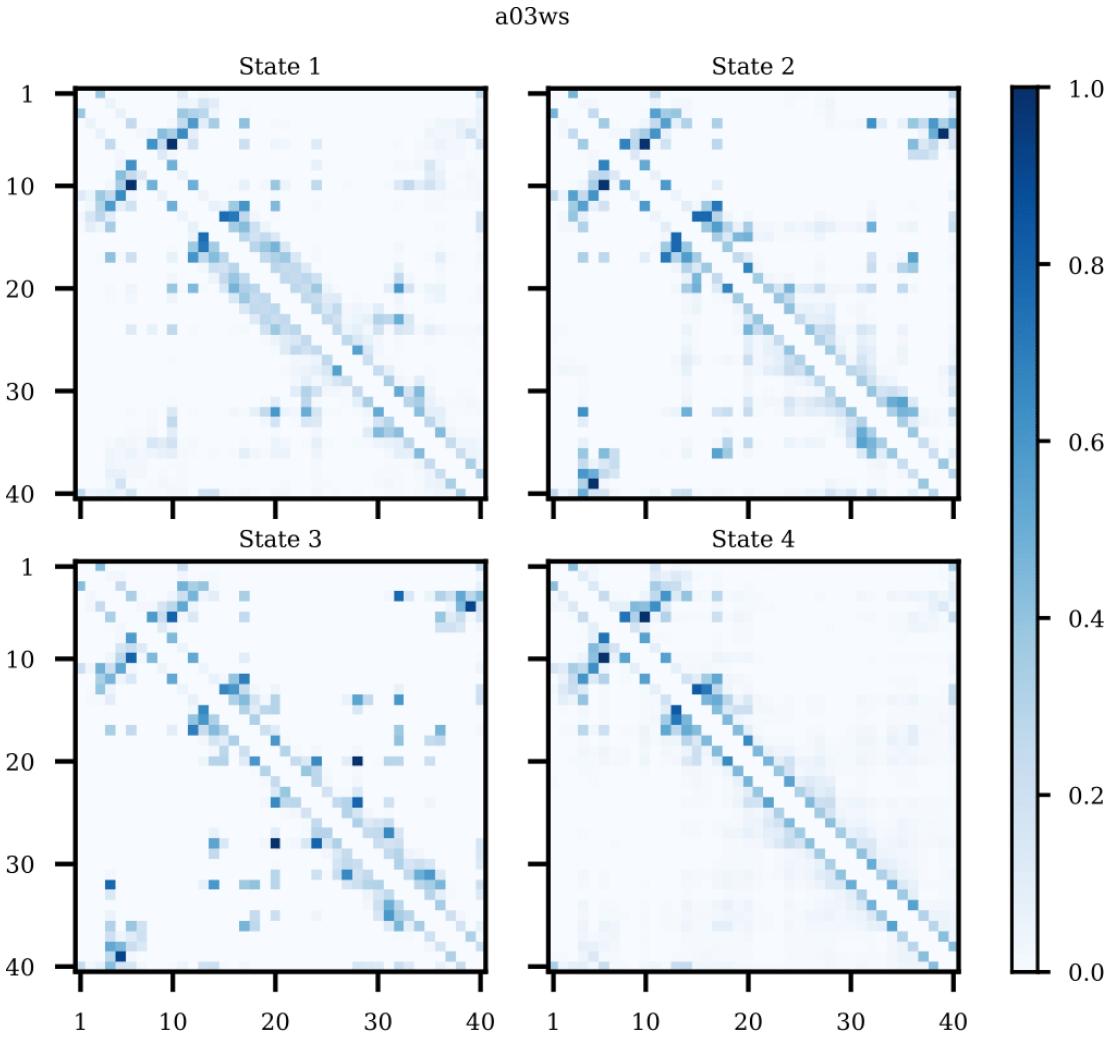

Figure 16: Normalized contacts for the coarse-grained MSM states obtained from the simulation with A03ws.

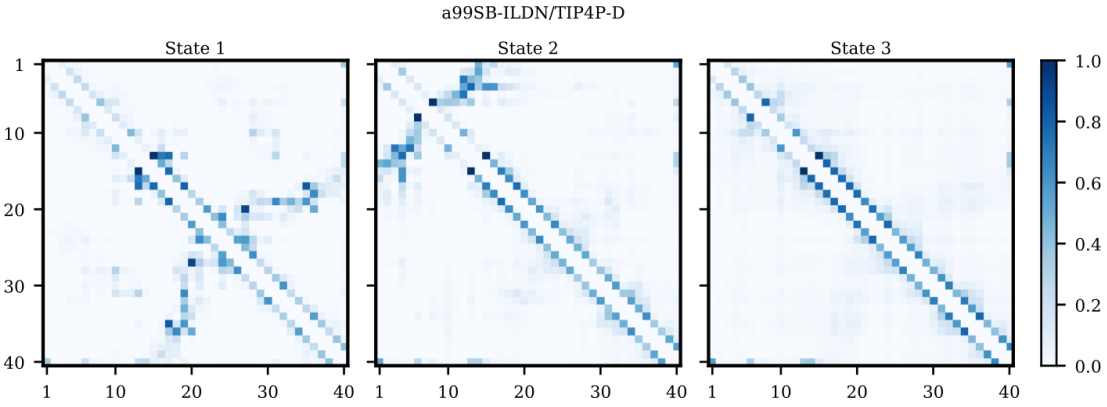

Figure 17: Normalized contacts for the coarse-grained MSM states obtained from the simulation with A99SB-ILDN/TIP4P-D.

The MSM for C36m also involves a dominant state corresponding to a stretched structure with no noteworthy contacts between distant residues (state 3, 81% population). However, with C36m also more compact structures with β-sheets are adopted, yielding MSM states 1, 2 and 4 (19% population in total). These three states exhibit a similar antiparallel β-sheet (Figure 21). In states 1 and 4 this involves a β-hairpin centered at residues G25/S26. The β-sheet in state 4 is on-pathway to that of state 1 (coming from state 3) but is shorter than the fully-formed β-sheet of state 4. The latter involves up to 8 or 9 residues per strand and thus reaches up to residue 15 towards the N-terminus and residue 35 towards the C-terminus. This corresponds to the maximum count of 16 to 17 residues that adopt a β-sheet conformation between 13 and 14 μs of the MD simulation (Figure 14). Another characteristic of the β-sheet in states 1 and 4 is the salt bridge present between D23 an K28, giving rise to the strongest inter-residue contact in either state. While the β-sheet characteristic of state 2 is similar to that of states 1 and 4, it misses the D23–K28 salt bridge, leading a less pronounced turn around G25/S26 and weaker contacts in the adjacent residues.

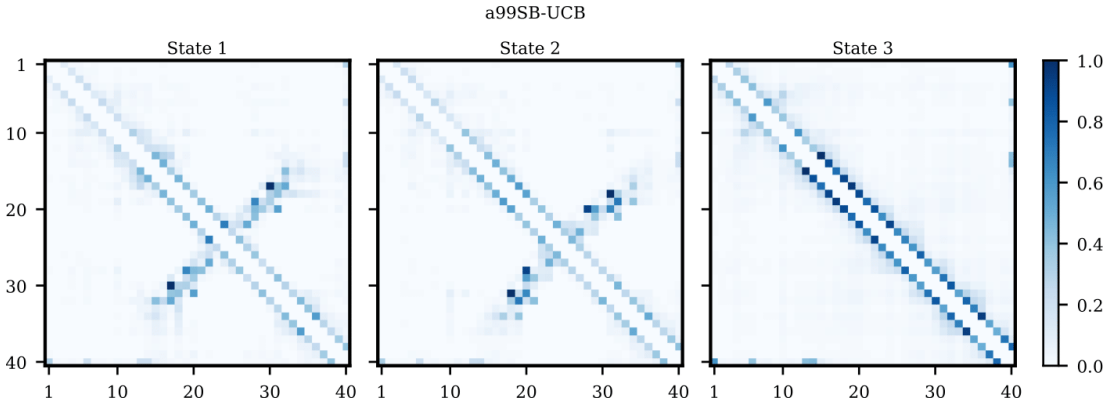

Figure 18: Normalized contacts for the coarse-grained MSM states obtained from the simulation with A99SB-UCB.

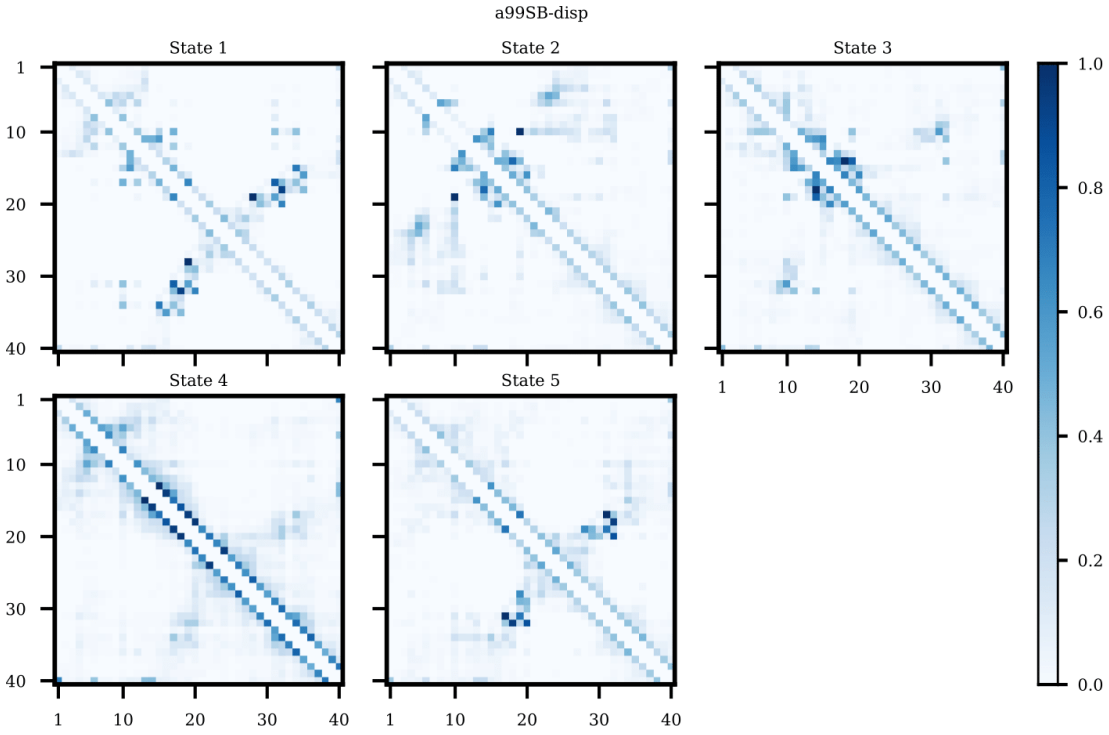

Figure 19: Normalized contacts for the coarse-grained MSM states obtained from the simulation with A99SB-disp.

Similar β-sheet formation is observed with C22*/TIP3P and A99SB-disp, yet the β-sheets are less pronounced (A99SB-disp, Figure 19) and may also occur in the N-terminal half of the peptide (C22*/TIP3P, Figure 20). Furthermore, the MSMs for these two FFs do not feature a state representing an elongated Aβ40 structure. Instead, more structure formation is seen which also involves helical elements seen in the C-terminal half of the peptide in states 2 and 3 of the MSM obtained with C22*/TIP3P. These two states, however, are only characterized by a low population

(2 and 4%, respectively), while the most populated state observed with this FF is associated with a low interpeptide contacts probability (state 4, 76% population) close to an elongated structure. The four states of the MSM for A03ws all feature a β-sheet in the first 10 N-terminal residues, but differ from each other in their contacts in the rest of the peptide (Figure 16). Helices of various lengths and locations along the sequence are presents in states 1–3, while the most populated state 4 (64% population) is without noteworthy inter-residue contacts beyond residue 10, thus representing a mainly elongated structure.

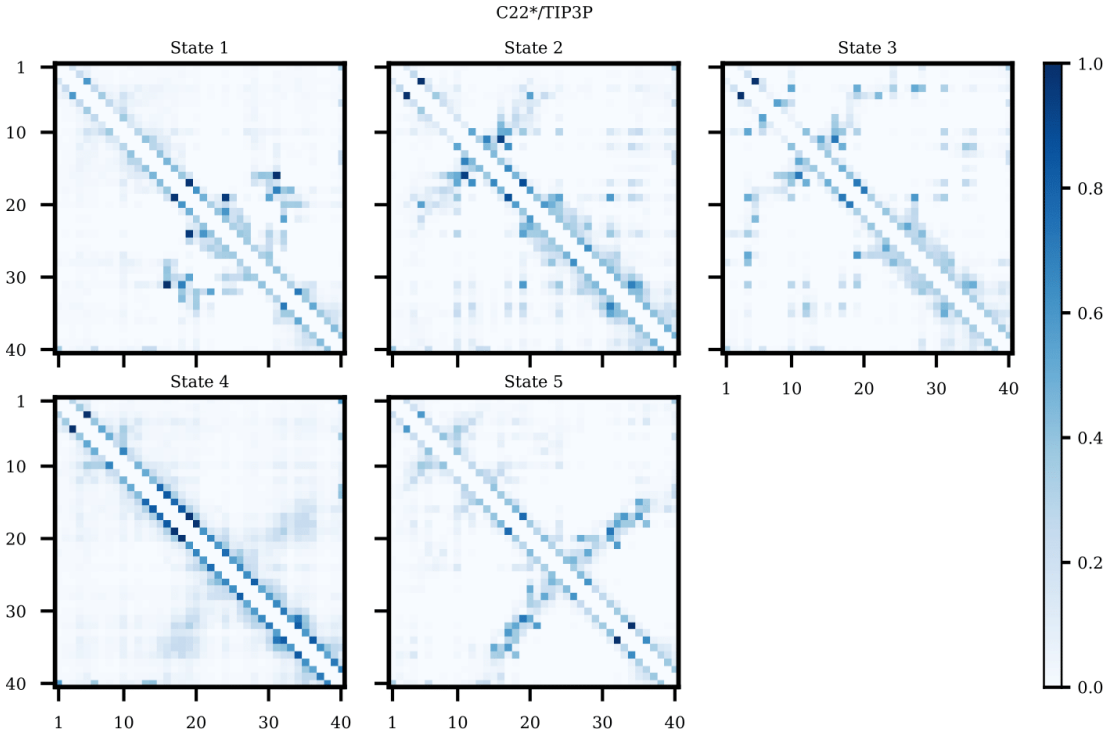

Figure 20: Normalized contacts for the coarse-grained MSM states obtained from the simulation with C22*/TIP3P.

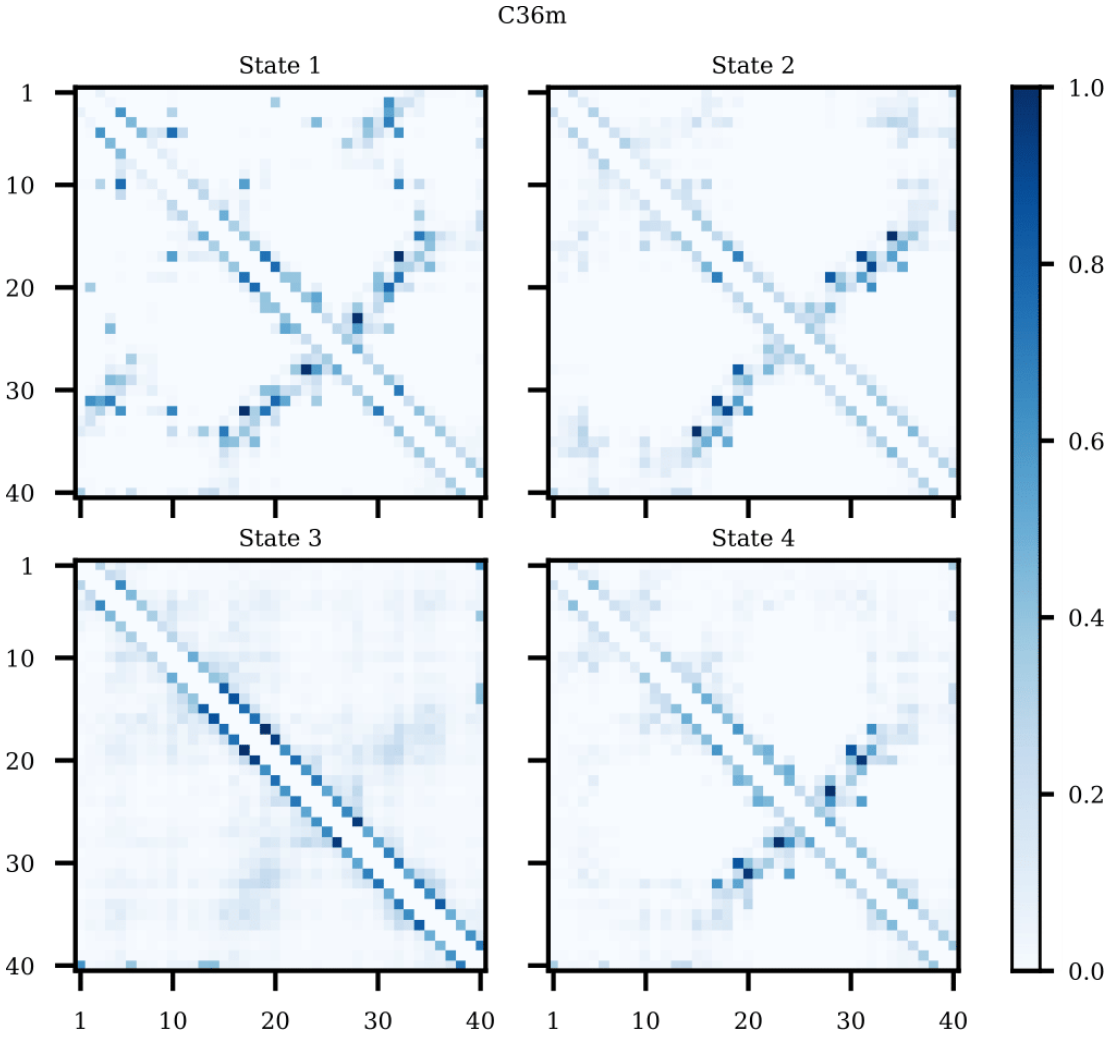

Figure 21: Normalized contacts for the coarse-grained MSM states obtained from the simulation with C36m.

Comparison between Figures 22 and 23 (the MSMs from Figure 23 can be superimposed onto the sample density in Figure 22 as the same TIC space is used for their projection) show that several of the metastable MSM states were only sampled in the last 10 µs. This holds, for instance, true for state 1 of the MSM with A03ws and also state 2 of the MSM with C36m. The calculation of the mean first passage times (MFPTs) for the transitions between the metastable states confirms that the slowest transitions require more than 20 µs for them to occur. In some cases the MFPTs are even predicted to be larger than 30 µs. Such slow transitions are found in all of the MSMs shown in Figure 23. As this chapter aims to provide an answer to the question how much MD sampling is needed before the equilibrium distribution of Aβ is reached, based on the MFPTs the answer would be that at least 30 µs are required.

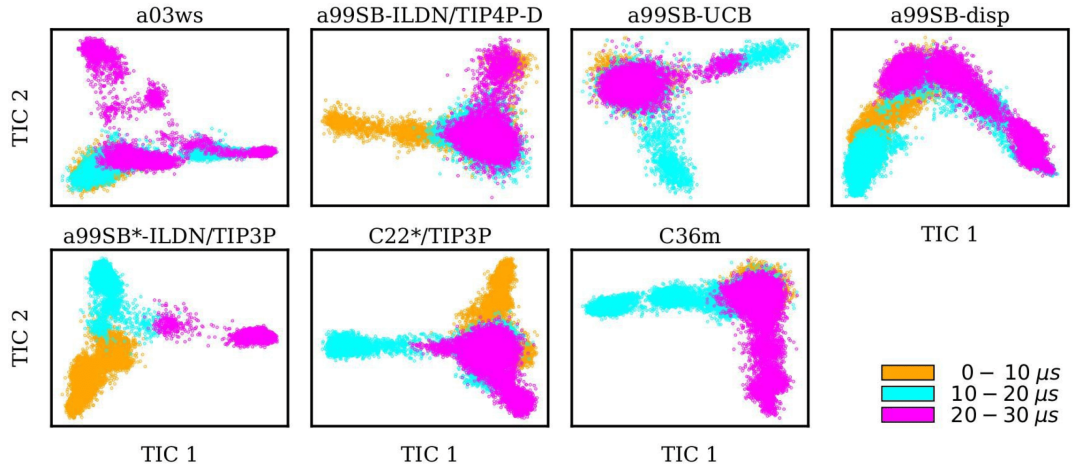

Figure 22: Sample densities for different time windows of the trajectories (0–10 μs: yellow, 10–20 μs: cyan, 20–30 μs: magenta) projected along the first two TICA coordinates (TICs) for the different FFs (labels on the top of the panels). It should be noted that the TICs are different between the FFs.

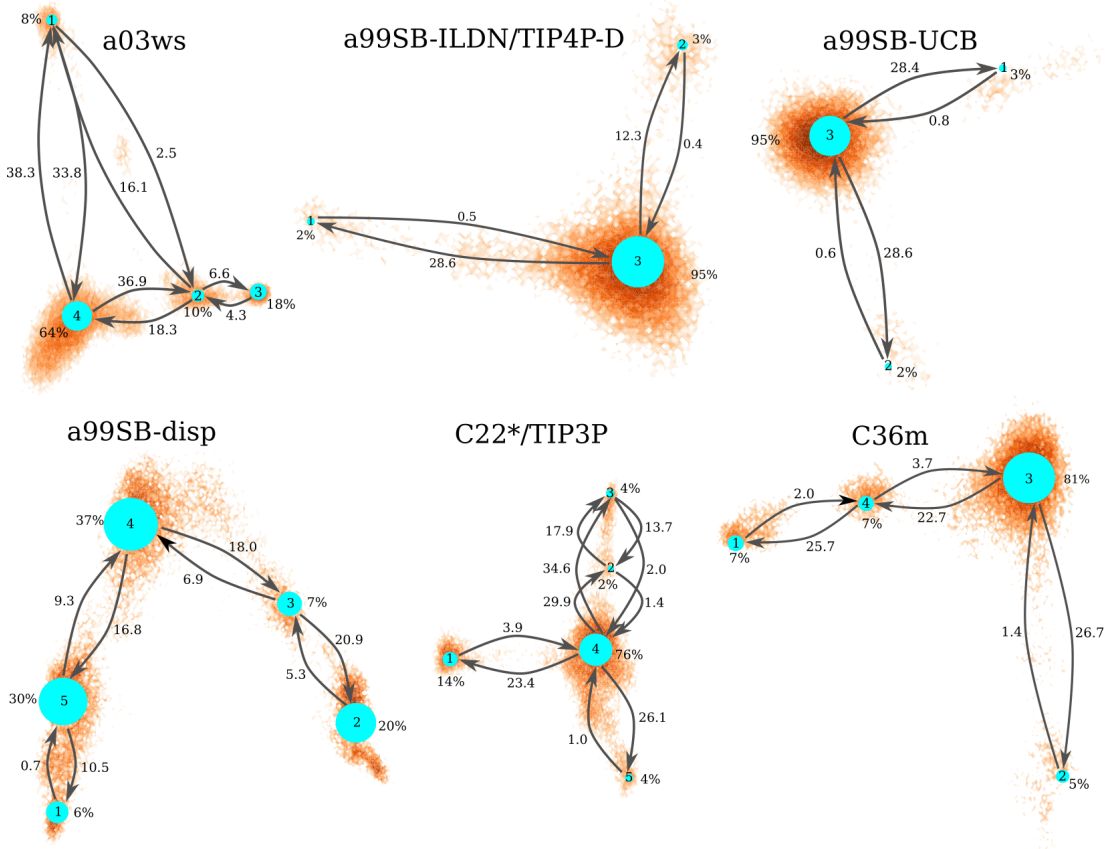

Figure 23: Markov state models for the different FFs (labels above the networks). The size of the network nodes reflects the population of the underlying state (given in % next to the nodes), whereas the thickness of the arrows corresponds to the transition probability. The mean first passage times (in μs) are written next to the arrows. The orientation of the MSMs corresponds to the projection of the MD data along the first two TICs in Figure 22.

4.5 INTEGRATING DATA FROM MD SIMULATIONS AND SPECTROSCOPIC EXPERIMENTS

4.5.1 *Construction of Aβ40 ensembles*

In order to validate the simulation results, we compare the structural ensembles obtained for Aβ40 with the corresponding information deduced from spectroscopic data, which is NMR chemical shifts and J-couplings [212] as well as FRET efficiencies [186]. The FRET experiments were performed at almost identical external condi-

tions as the simulations, which are 297 K as temperature and 50 mM ionic strength in the FRET experiments as opposed to 300 K and also 50 mM ionic strength in the simulations. The NMR experiments, on the other hand, were conducted at 277 K and an ionic strength of 20 mM. Such changes in external conditions are expected to slightly affect the conformational ensemble of Aβ40. A perfect agreement between the results derived from NMR spectroscopy and simulations can thus not be expected.

4.5.1.1 *Chemical shifts*

We calculated the NMR chemical shifts of the carbonyl C_α and C' atoms for all MD generated conformations and provide the averages over the 30-μs MD simulations in Figure 24. The agreement between the measured and calculated C_α chemical shifts is generally good for all FFs as judged by the RMSD between these data sets (Table 2). The smallest RMSD is found for A99SB-ILDN/TIP4P-D with a value of 0.59 ppm, while the largest values of 0.74 and 0.76 ppm result from A03ws and A99SB*-ILDN/TIP3P, respectively. A similar picture emerges if one compares the C'-chemical shifts. By far the largest deviation is observed for A03ws with a value of 1.03 ppm, whereas A99SB-UCB and A99SB-disp lead to the smallest RMSDs of 0.93 ppm. A first conclusion is that A03ws and A99SB*-ILDN/TIP3P are the least agreeable with the employed NMR chemical shifts data.

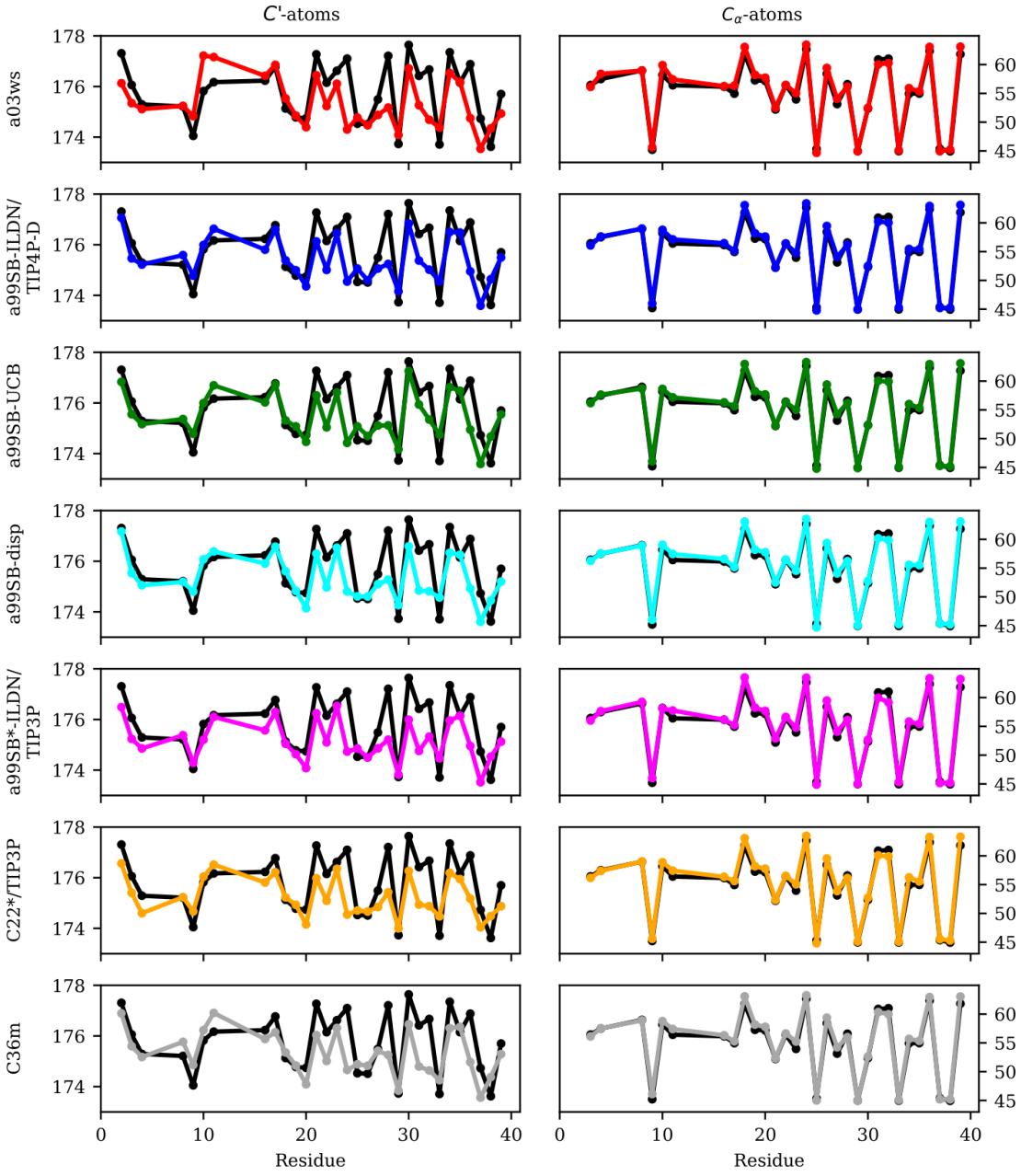

Figure 24: Experimental (black) and calculated (colored) NMR chemical shifts for the C′ atoms (left) and C_α atoms (right) for the different FFs (indicated on the left of each row).

The comparison of the chemical shifts for each residue (Figure 25) shows that the larger deviations in comparison to the other FFs result from chemical shifts on the N-terminal side. For A03ws, the calculated C_α- and C′-chemical shifts are higher than the experimental values, indicating an overestimation for α-helix formation in this simulation [213]. With A99SB*-ILDN/TIP3P, on the other hand, the chemical shifts are smaller than the experimental counterparts, suggesting a bias toward β-

sheets [213]. It should be mentioned, that these interpretations can only indicate tendencies, because for SPARTA+ the intrinsic RMS deviations between the predicted and experimental chemical shifts are 1.07 ppm for C′ and 0.92 ppm for C_α [207]. They are thus in the range or even slightly larger than the RMSDs between the experimental and simulated chemical shifts for Aβ40. Nonetheless, the conclusions drawn for A03ws and A99SB*-ILDN/TIP3P are in accordance with those resulting from the analysis of the secondary structure (Figure 13).

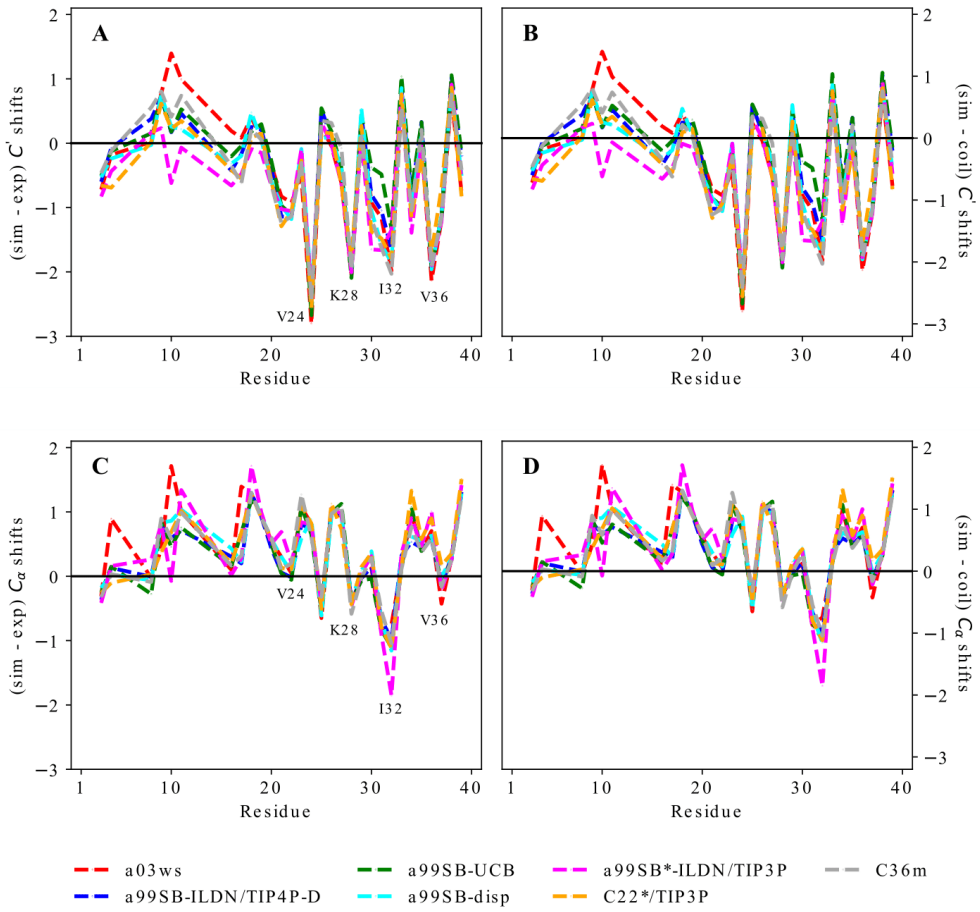

Figure 25: Difference between (A and C) the calculated and experimental and (B and D) the calculated and random coil NMR chemical shifts (i.e., the secondary chemical shifts) for the C′ and C_α atoms (top and bottom, respectively) of Aβ40 residues for the different FFs (see color key). The residues with the large deviations between simulation and experiment and which are discussed in detail in the text are labeled in panels A and C.

Bax and co-workers concluded from their NMR studies that both Aβ40 and Aβ42 predominantly sample random coil (RC) structures [212]. To better estimate how much the simulated ensembles deviate from random coil, we calculated the RMSDs between the simulated and RC chemical shifts using a data set derived for chemical shifts of IDPs [214, 215], which was also employed in ref [212] (Table 2). The RMSD rankings with respect to the experimental chemical shifts and with respect to the RC values are almost identical. In both cases the largest deviations are found for A03ws and A99SB*-ILDN/TIP3P. Closest to RC structures are the structural ensembles generated with A99SB-ILDN/TIP4P-D, A99SB-UCB and A99SB-disp, followed

by C22*/TIP3P. These findings agree with our conclusions drawn from the MSMs in conjunction with an inter-residue contact analysis for the resulting states. Interestingly, C36m, which was purposefully developed for IDPs [185], leads to less RC in the structures of Aβ40, which was also observed in the contact maps of the MSM states as three of the four MSM states feature rather high β-sheet formation (Figure 21). The analysis of the deviations of the simulated chemical shifts from the experimental and RC ones (yielding the secondary chemical shifts) on a per-residue basis reveals a very similar pattern (Figure 25). In both cases and for all FFs the largest deviations are found for V24, K28, I32 and V36. For these residues, the measured and RC chemical shifts are considerably larger than the calculated ones. Comparison of the measured and RC C′-chemical shifts revealed deviations opposite in sign (but smaller in absolute terms) for these four residues (see Figure 1 in ref [212]). The experimental values suggest that these residues have a tendency to adopt a helical conformation. Under consideration of all of the up to 12 measured NMR parameters this tendency was confirmed for V24 and K28 by the MERA program (Maximum Entropy Ramachandran map Analysis from NMR data) [216, 217] that was developed by the Bax lab and applied to the Aβ40/42 NMR data [212]. For I32 the measured secondary chemical shift is small, while for V36 a preference for RC is found if all other NMR data measured for that residue are considered. For the region A30–I32 all the predicted C′-chemical shifts are smaller than the experimental values for all FFs. The smallest deviation is found with A99SB-UCB, which also explains the overall smaller RMSD from experiment found for this FF. Comparison with the RC values shows that with A99SB-UCB this region is in a RC state, which is in line with the experimental findings, while the other FFs sample to some degree a β-conformation here.

Another region that needs attention is V17–A21 forming the central hydrophobic core (CHC), which by many studies is proposed to adopt a β-conformation and play a central role during amyloid aggregation of Aβ [49]. There is almost no deviation between simulation and experiment for V18 and F19, while for the other residues and also the neighboring K16 and E22 the simulations (apart from the one with A03ws) lead to smaller C′-chemical shifts than in experiment. The predicted values are also smaller than the RC chemical shifts, indicating that this region tends to adopt a β-

conformation in the simulations, which was also seen in several of the MSM states for most of the FFs. The NMR data, on the other hand, indicate β-strand formation only for V18–F20, which follows from the measured secondary chemical shifts and also from the MERA anaylsis that considers the other NMR data [212]. However, Bax and co-workers excluded β-sheet formation for that region as from their NMR data no matching set of residues with which to pair these residues in a β-sheet could be identified. In the simulations these residues usually pair with I30–L34 (Figures 16–21), which, as already discussed above, also tend to be in a β-conformation.

Quantity	FFs						
	A03ws	A99SB-ILDN/TIP4P-D	A99SB-UCB	A99SB-disp	A99SB*-ILDN/TIP3P	C22*/TIP3P	C36m
$RMSD_{exp}$ C_α-shift (ppm)	0.738	0.589	0.641	0.631	0.759	0.688	0.634
$RMSD_{RC}$ C_α-shift (ppm)[#]	0.678	0.555	0.593	0.580	0.706	0.627	0.572
$RMSD_{exp}$ C'-shift (ppm)	1.025	0.964	0.927	0.929	0.984	0.961	0.978
$RMSD_{RC}$ C'-shift (ppm)[#]	1.211	1.077	1.041	1.117	1.209	1.188	1.177
χ^2 J-coupling	7.03	3.28	1.94	4.50	6.33	2.50	3.05
$\langle R_{ee} \rangle \pm$ S.D. (nm)[†]	2.64 ± 1.60	3.97 ± 1.53	4.22 ± 1.68	3.17 ± 1.25	2.22 ± 0.82	2.67 ± 1.14	3.09 ± 1.54
$\langle E_{FRET} \rangle \pm$ S.D.	0.83 ± 0.26	0.65 ± 0.31	0.59 ± 0.33	0.81 ± 0.23	0.96 ± 0.06	0.88 ± 0.17	0.80 ± 0.27
$\langle R_{gyr} \rangle \pm$ S.D. (nm)	1.31 ± 0.30	1.65 ± 0.30	1.76 ± 0.40	1.38 ± 0.30	1.11 ± 0.08	1.24 ± 0.20	1.45 ± 0.30
$\langle R_{hyd} \rangle \pm$ S.D. (nm)	1.62 ± 0.16	1.77 ± 0.13	1.80 ± 0.13	1.66 ± 0.11	1.53 ± 0.05	1.60 ± 0.11	1.69 ± 0.13

[#] For the calculation of the RMSD with respect to the RC chemical shifts, the RC values and correction factors determined in refs [214, 215] for IDPs at 300 K and pH 7 were used. [†] S.D. = standard deviation.

Table 2: Simulated properties of Aβ40 sampled with different FFs

4.5.1.2 *J-couplings*

Bax and co-workers did not only record chemical shifts for Aβ40 and Aβ42, but also three-bond J-couplings, including $^3J_{HNH\alpha}$ couplings [212] that are related to the backbone torsion angles ϕ by the empirically parametrized Karplus equation. We use

these experimental values to further validate the Aβ40 ensembles generated by MD simulations with different FFs. The comparison between simulation and experiment for the C′ and C_α chemical shifts led to the conclusion that A99SB-ILDN/TIP4P-D, A99SB-UCB and A99SB-disp produce Aβ40 structural ensembles best in agreement with the NMR chemical shift data, while A03ws and A99SB*-ILDN/TIP3P fail to do so. A closer inspection of the C′ chemical shifts revealed that A99SB-UCB is in particular able to reproduce Aβ40's tendency to adopt a RC state. Comparison for the $^3J_{HNH\alpha}$ couplings confirms that this FF is superior to the other FFs in modeling Aβ40 as a χ^2 value of only 1.94 is obtained (Table 1). The χ^2 values for A03ws and A99SB*-ILDN/TIP3P of 7.03 and 6.33, respectively, also confirm the previous conclusion, which is that these two FFs do not yield good structural ensembles for Aβ40.

For the remaining four FFs the assessment is somewhat more complicated. The usage of A99SB-ILDN/TIP4P-D can be recommended as χ^2 for the $^3J_{HNH\alpha}$ couplings is also still considerably small with a value of 3.28; though A99SB-UCB performs clearly better. Interestingly, A99SB-disp, which was developed for IDPs (but also folded proteins) and claimed to be one of the best FFs for Aβ40 [108], does not yield $^3J_{HNH\alpha}$ couplings in agreement with experiment. The χ^2 value is 4.50 and comparison of the J-couplings on a per residue basis (Figure 28) shows that for most residues these values are smaller than the experimental J-couplings, only reaching values of ~6 Hz or less. An exemption to this is the region K16–E22 where the CHC residues V18–F20 in particular demonstrate a tendency to adopt β-conformation, which is also correctly reflected by the J-couplings. Here, a very good agreement with the experimental values is obtained for A99SB-disp. For the other residues, on the other hand, the average ϕ values must be too small. To validate this conclusion we analyzed the Ramachandran angles of I31 and I32 as sampled in the MD simulations in detail (Figures 26 and 27). From experiment, $^3J_{HNH\alpha}$ couplings above 7 Hz were obtained, indicating a considerable population of extended structures, which is supported by MERA [212]. Figure 26 shows that with A99SB-disp only few structures with $\phi < -90°$ were sampled for I31 and I32. Instead, polyproline II (PPII) conformations were preferentially adopted, which explain the low J-couplings. With A99SB-ILDN/TIP4P-D and A99SB-UCB, on the other hand, which both led to a good agreement for the J-couplings of

I31 and I32, a considerable amount of conformations with $\phi < -120°$ were sampled, which includes extended conformations from the β-basin as well from the type I β-turn region. Interestingly, with these two FFs the largest variability of conformations was adopted. Basically all allowed regions of the Ramachandran space, including the α_L region were sampled, which agrees with the notion that random coil is not a particular structure but a fast fluctuation between all possible ϕ/ ψ combinations.

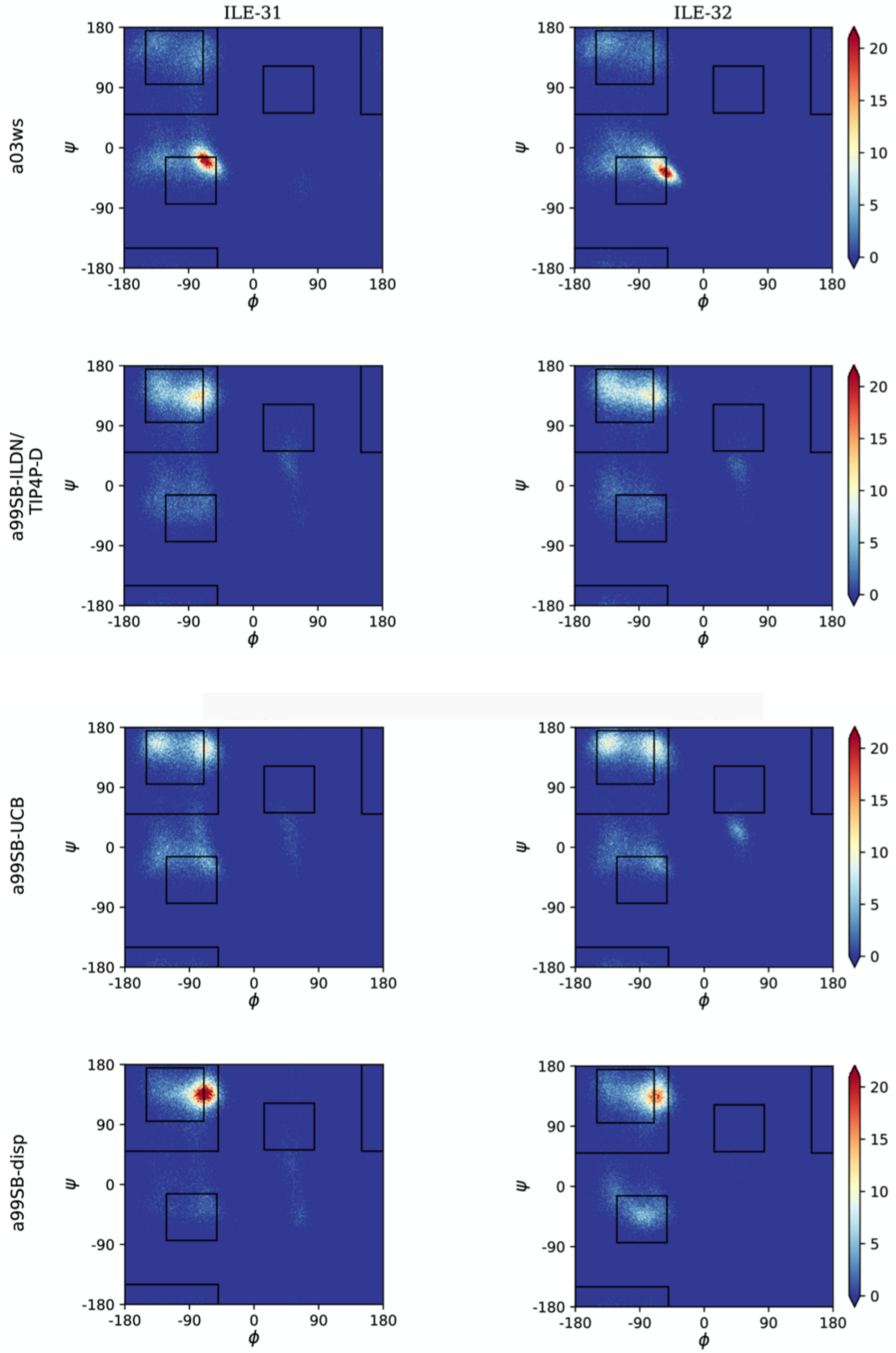

Figure 26: Ramachandran plots of I31 and I32 obtained from the simulation with A03ws, A99SB-ILDN/TIP4P-D, A99SB-UCB, and A99SB-disp (FF labels on the left of the panels).

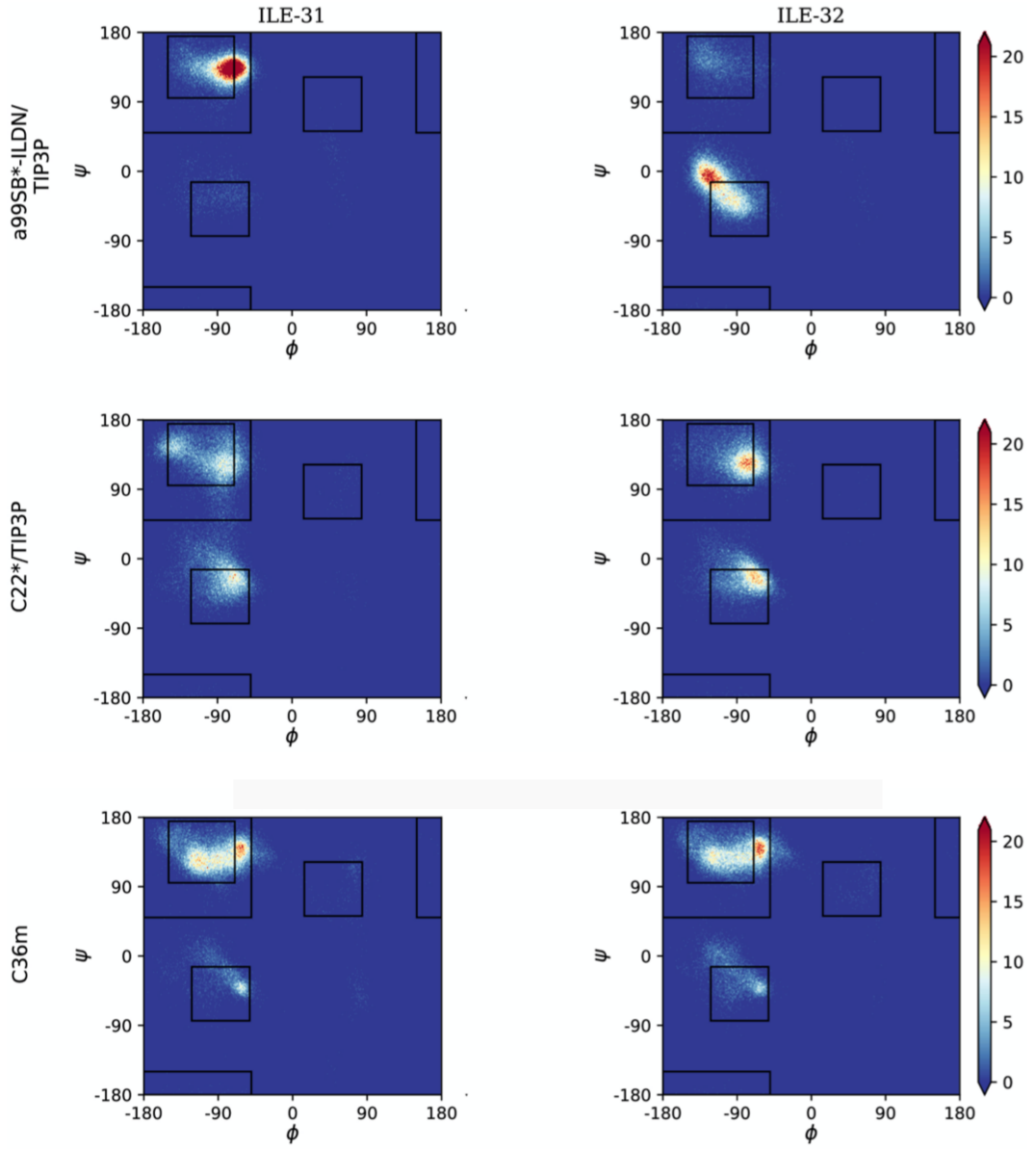

Figure 27: Ramachandran plots of I31 and I32 obtained from the simulation with A99SB*-ILDN/TIP3P, C22*/TIP3P, and C36m (FF labels on the left of the panels).

For the two Charmm FFs under consideration, the situation is contrary to that of A99SB-disp. For C22*/TIP3P and C36m the agreement with NMR chemical shifts is largely insufficient, while the χ^2 values for the $^3J_{HNH\alpha}$ couplings are satisfactory (2.50 and 3.05, respectively). Interestingly, the FF not improved for IDPs, i.e., C22*/TIP3P performs somewhat better for both chemical shifts and J-couplings compared to C36m. However, the disagreement between both FFs is limited to a few residues, such as D7, M35 and V36 where C36m performs worse for the J-couplings, while for most of the remaining residues similar NMR observables are predicted. For M35

and V36, C36m sampled a high amount of PPII and α_R structures, for M35 also α_L based on the corresponding Ramachandran plots, leading to J-couplings below those found from experiment, which however, as Figure 28 shows, increased for sampling times above 10 µs. In general, we observed that it was only beyond 10 µs simulation times that MD convergence for the J-couplings was obtained. This conclusion excludes A99SB-ILDN/TIP4P-D and A99SB-UCB, for which converged J-couplings were already obtained within 10 µs.

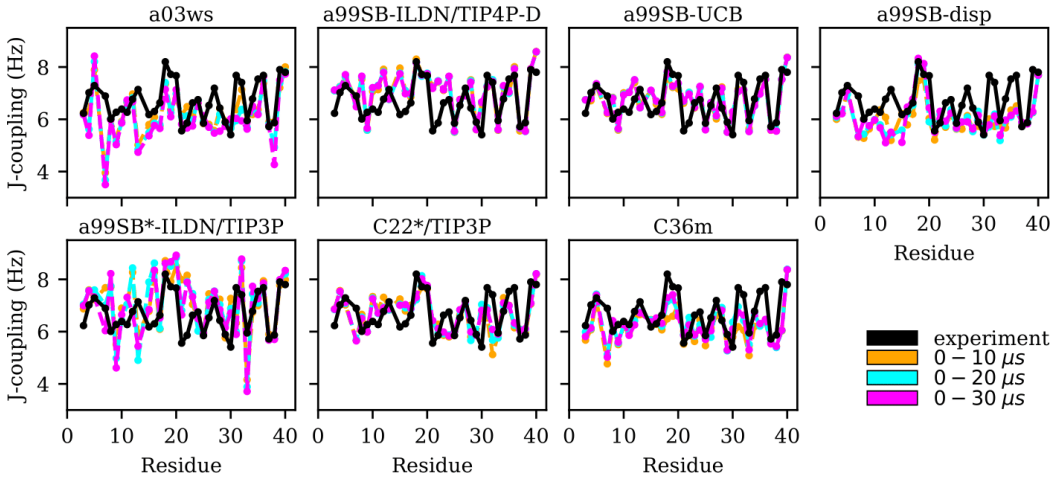

Figure 28: Experimental (black) and simulated (0–10 µs: yellow, 0–20 µs: cyan, 0–30 µs: magenta) $^3J_{HNH\alpha}$ couplings for each Aβ40 residue for the different FFs (labels on the top of the panels).

4.5.1.3 *End-to-end distance and hydrodynamic radius*

Another experimental observable that is available for Aβ40 was derived from 2D FRET data that have been reported by Meng et al. [186]. They determined the end-to-end distance, R_{ee}, of Aβ40 from the analysis of the FRET efficiency between the donor Alexa 488 and the acceptor Alexa 647, which were attached at the termini. To this end, an unnatural amino acid, 4-acetylphenylalanine and a cysteine residue were first introduced at the N- and C-terminus of Aβ40, respectively. An average FRET efficiency of ~0.6 was obtained. According to equation (26) this corresponds to a distance of 4.85 nm between donor and acceptor, called R'_{ee} here. To account for the size of the fluorophores, equation (27) is applied, yielding an average end-to-end distance of ~4.3 nm for Aβ40. The values for R_{ee} and the FRET efficiencies in Table 2 show that A99SB-UCB performs very well and A99SB-ILDN/TIP4P-D does well in

reproducing these observables. While A99SB-disp and C36m are next in performance, they however clearly underestimate R_{ee}, leading to overestimated FRET efficiencies of 0.81 and 0.80, respectively.

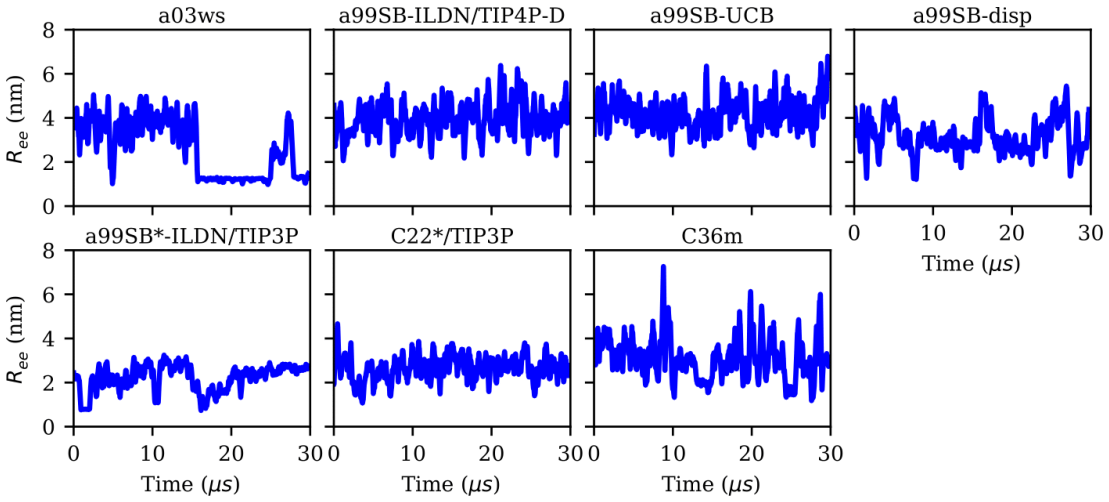

Figure 29: Evolution the end-to-end distance R_{ee} for the different FFs (FF labels on the top of the panels).

Even more compact Aβ40 structures are produced by C22*/TIP3P and a99SB*-ILDN/TIP3, with the latter FF leading to average FRET efficiencies of nearly 1 ($E_{FRET} = 0.96$). For A03ws the situation is somewhat more complicated. The distributions of R_{ee} averaged over different times (Figure 32) shows that for most FFs this quantity converged within 10 µs. Exceptions are A03ws and A99SB*-ILDN/TIP3P. With the former FF, the end-to-end distance became smaller with time, with a pronounced peak that developed at R_{ee} ~1.2 nm. With A99SB*-ILDN/TIP3P, on the other hand, the peak at R_{ee} ~0.9 nm became less important with time, while more extended structures were adopted. The evolution of R_{ee} over the whole trajectory (Figure 32) reveals that with A03ws large fluctuations with respect to the end-to-end distance occurred, involving the formation of a compact conformation with especially low R_{ee} values in which Aβ40 was trapped between 16 and 25 µs. Representative conformations from this trapped state are shown in Figure 30. They are characterized by the presence of helices along the sequence from residue Y10 onward, corresponding to MSM states 1, 2 or 3. This finding is surprising as the increase in vdW interactions between protein and water as done in the development of A03ws was thought to avoid such overly compact protein states [139], yet the current 30 µs MD simulation shows that the he-

lical propensity present in the FF predecessor A03 [218] overrules this modification if simulated long enough. With the other FFs such trapping is not seen; instead, fast fluctuations in R_{ee} are sampled, with C36m showing the largest and C22*/TIP3P and A99SB*-ILDN/TIP3P the smallest fluctuations.

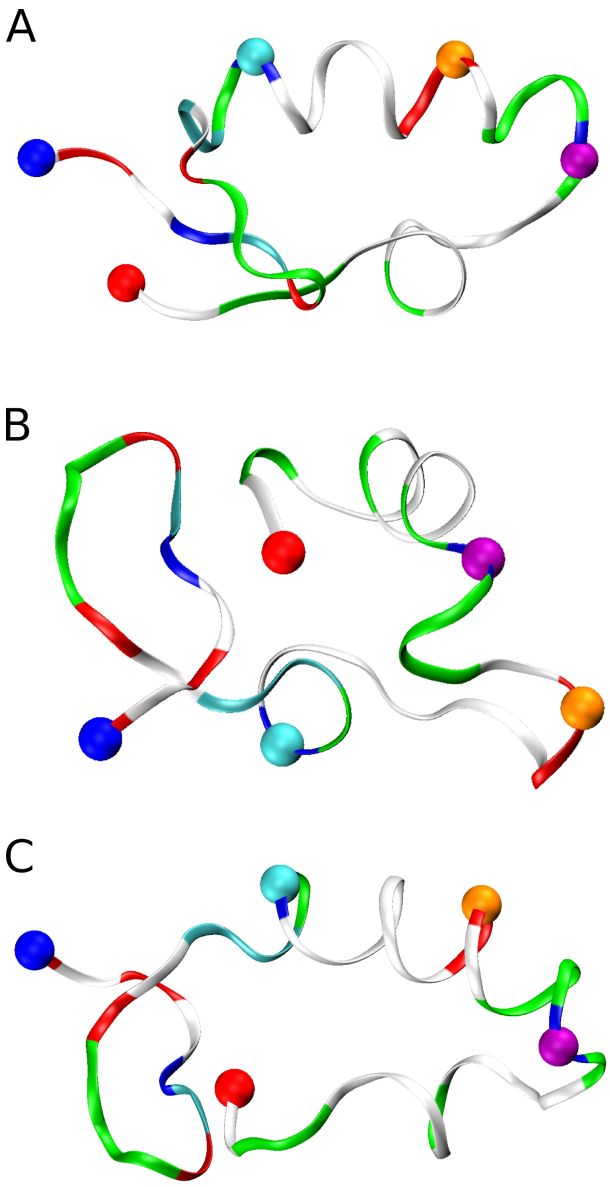

Figure 30: Compact Aβ40 structures sampled with A03ws between 16 and 25 μs. These conformations exhibit a high propenisty for helix formation in different parts along the sequence: (A) between residues K16 and K28 (as present in MSM state 1), (B) between residues G29 and M35 (as present in MSM states 3), (C) between residues K16 to K28 and G29 to M35 (as present in MSM state 2). Aβ40 is shown as band and colored according to amino acid residue type (basic: blue, acidic: red, histidine: cyan, polar: green, hydrophobic: white). Following residues are indicated by spheres: N-terminus (blue), K16 (cyan), D23 (orange), K28 (mauve), C-terminus (red).

While the agreement between average FRET efficiencies determined experimentally and those derived from the simulations with A99SB-ILDN/TIP4P-D and A99SB-UCB is very good, the same cannot be said for the distribution of the FRET efficiencies. In experiments this distribution assumes a Gaussian shape around the average value [186], while we find highly skewed distributions with the maximum for FRET efficiencies being close to one (Figure 31). The same observation was made in other simulation studies where, as observed here with A99SB-ILDN/TIP4P-D and A99SB-UCB, the agreement with the average E_{FRET} value was associated with a considerable amount of structures with R_{ee} values above 6 nm [186, 210, 219]. Further simulations are needed to identify the source of this disagreement. There, effects of the FRET labels including the extra amino acids added at the termini of Aβ should be explicitly considered as well as the orientation of donor and acceptor with respect to each other be accounted for when determining E_{FRET}. Another possible source for the discrepancies between simulations and experiment could lie in the size of the cubic simulation boxes with an edge length of 6 nm used in the simulations analyzed here and in those generated by Head-Gordon et al. [210], while Best and co-workers set up an even smaller box with only 5.5 nm edge length [186]. Comparison with R_{ee} shows that Aβ40 can reach beyond these box dimensions if fully extended, which may limit further expansion of the peptide. Thus, in future simulations of Aβ larger boxes should be employed. However, the current results are not expected to be much influenced by the box sizes as in recent coarse-grained simulations of Aβ with implicit solvent, which do not involve simulation boxes, very similar E_{FRET} distributions as in the explicit-solvent all-atom simulations were obtained [219].

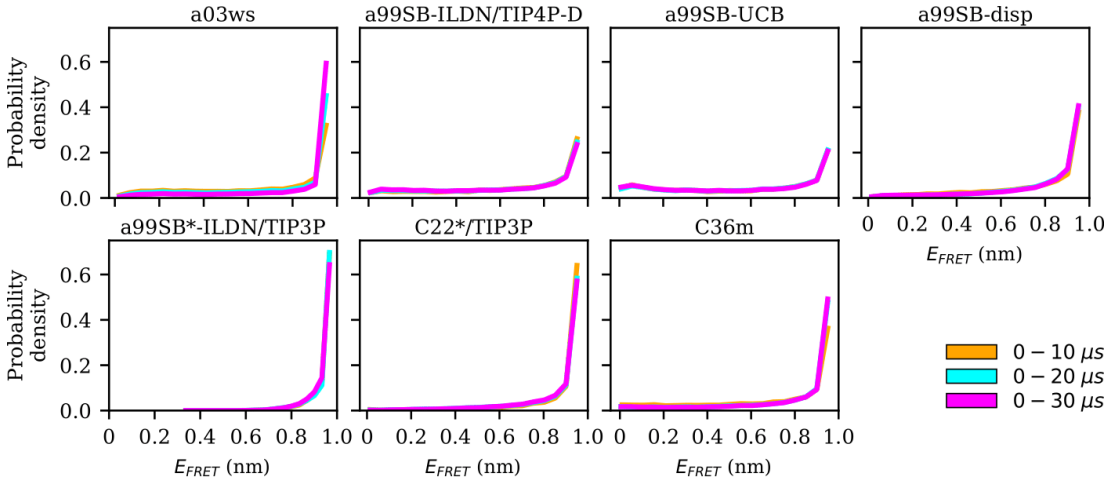

Figure 31: Distribution of the FRET efficiency E_{FRET} for increasing trajectory lengths (0–10 μs: yellow, 0–20 μs: cyan, 0–30 μs: magenta) for the different FFs (FF labels on the top of the panels).

Another quantity closely related to R_{ee} is the hydrodynamic radius, R_{hyd}, which was determined as 1.6 nm for Aβ40 at 298 K using size exclusion chromatography (SEC) and NMR diffusion experiments [220]. In general, the hydrodynamic radius is closely related to the radius of gyration. For IDPs, a relationship between these two quantities was derived that explicitly takes the chain-length dependency into account [221]. Using equation (7) of ref [221], we calculated the average R_{hyd} value for each FF using the R_{gyr} values determined for the corresponding MD snapshots. The results in Table 2 show that the experimental R_{hyd} value of 1.6 nm is best reproduced by A03ws and C22*/TIP3P, followed by A99SB-disp and C36m. With A99SB*-ILDN/TIP3P the hydrodynamic radius is underestimated, while A99SB-ILDN/TIP4P-D and A99SB-UCB lead to R_{hyd} values clearly above the experimental prediction. These observations are mostly in contrast to the conclusions drawn from the calculation of E_{FRET} since for R_{ee} the best agreement was identified for A99SB-ILDN/TIP4P-D and A99SB-UCB. Only for A99SB*-ILDN/TIP3P both R_{ee} and R_{hyd} agree with each other, both being too small compared to experiment as a result of too compact Aβ40 structures being sampled with this FF. Since for all other quantities A99SB-ILDN/TIP4P-D and A99SB-UCB produced the best agreement with experiment, we decided to disregard the assessment based on R_{hyd}, especially since it contradicts the R_{ee} results.

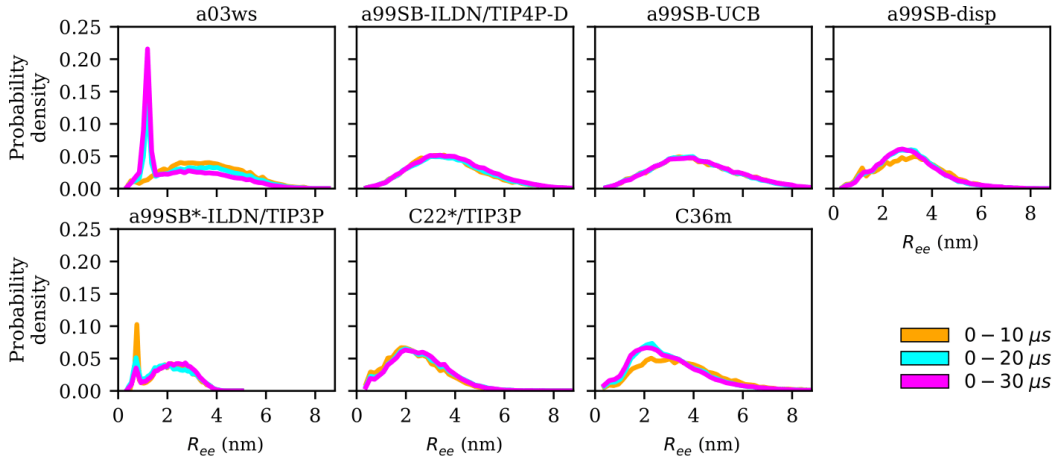

Figure 32: Distribution of the end-to-end distance R_{ee} for increasing trajectory lengths (0–10 μs: yellow, 0–20 μs: cyan, 0–30 μs: magenta) for the different FFs (labels on the top of the panels).

4.5.2 *Prediction of Aβ40 kinetics*

The kinetic analysis of the MD data using MSMs allows to further assess the accuracy of the simulation data based on time scales that were reported for Aβ motions. From the FRET study mentioned above [186] it was concluded that Aβ40 and Aβ42 exhibit no conformational dynamics exceeding 1 μs and that the end-to-end distance relaxation time is ~35 ns, which was determined by nanosecond fluorescence correlation spectroscopy. The upper limit of 1 μs for internal motions agrees with the findings from fluorescence measurements using the method of Trp-Cys contact quenching, which revealed a time scale of ~1 μs for intramolecular reorientation or diffusion for Aβ40 [187]. With NMR spin relaxation measurements the faster motions of Aβ40 were studied, from which a timescale of ~5 ns was determined for segmental motions, which can reach up to ~10 ns for selected residues [222]. The focus of MSMs is the identification of slow, memoryless motions. Thus, for the current comparison of the Aβ40 kinetics as determined by experiment and MD simulations, the upper limit of 1 μs for the slowest motion is of interest to us. All FFs with conformational transitions exceeding this time scale can be rendered as inadequate for modeling the kinetics of Aβ. Here it should be emphasized that the MFPTs discussed above refer to well-defined transitions between specified states, which are not the same as the relaxation times probed by the different spectroscopic techniques. For this, the implied

time scales (ITSs) underlying the constructed MSMs are better suited. The implied time scales reflect how quickly any initial state vector converges towards the equilibrium state vector in an MSM and are thus comparable to relaxation times monitored experimentally. The MFPTs, on the other hand, indicate the time it takes to transition from one equilibrium state vector to another one. This can become considerably larger than the ITSs, especially for transitions into equilibrium states with very small probabilities, which can be seen from the MFPTs in Figure 23. Thus, we concentrate on the ITSs here.

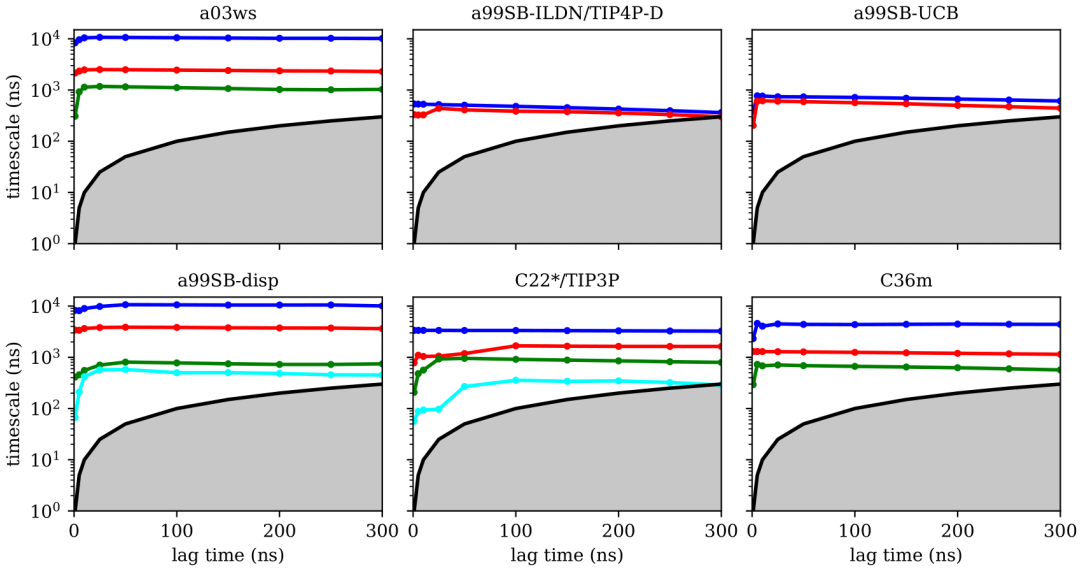

Figure 33: ITSs of the slowest processes (colored lines) obtained for different MSMs at different lag times (dots on colored lines) calculated from the MD trajectories using different FFs (labels on the top of the panels).

The plots of the ITSs against the lag times of the MSMs (Figure 33) show that in the case of A03ws, A99SB-disp, C22*/TIP3P and C36m they clearly exceed 1 µs. Some of these FFs even reach time scales for the slowest motion of more than 10 µs (A03ws and A99SB-disp). It should be noted that for A99SB*-ILDN/TIP3P no results for the implied time scales can be shown as with this FF the slowest dynamics of Aβ40 reached the length of the MD trajectory, i.e., 30 µs. Thus, A99SB-ILDN/TIP4P-D and A99SB-UCB are the only two FFs which agree with the experimental finding that the slowest intramolecular Aβ40 dynamics takes places within 1 µs. The exact ITS is 500 ns for A99SB-ILDN/TIP4P-D and 700 ns for A99SB-UCB.

4.6 DISCUSSION

4.6.1 *Which FFs are suitable for modeling Aβ?*

Based on the detailed comparison with NMR and FRET data we can conclude that A99SB-UCB produced an Aβ40 conformational ensemble that is best in agreement with experiment. The second best FF is A99SB-ILDN/TIP4P-D. Like A99SB-UCB it produces extended Aβ40 conformations in agreement with FRET. Also the NMR chemical shift data are of similar quality, yet the $^{3}J_{HNH\alpha}$ couplings are less in agreement. The performance of the two CHARMM FFs, C22*/TIP3P and C36m, is similar to each other. While the older of these two FFs performs better in reproducing the NMR data, the IDP corrected FF yields, on average, less compact Aβ40 conformations. However, also with C36m overly compact structures with too much β-sheet content are still sampled. The overall still acceptable agreement with experimental findings is realized by the extended structures that are temporarily adopted with this FF. It would be interesting to test C36m with a water model with more favorable LJ interactions between protein and water. This approach improved the performance of C36m for some of the IDPs that were studied by the developers of C36m [185]. It should be noted that in the previous FF benchmark study for Aβ42 we had identified C22*/TIP4P-Ew as the best FF for Aβ42 [122]. However, this benchmark did not include any of the FFs recently developed for IDPs since it was performed prior to their development. Nonetheless, even though C36m was explicitly developed for IDPs (by refining backbone parameters) [185], it does not lead to a better performance than the standard C22*/TIP3P FF.

Another FF that was developed for IDPs (but also for folded proteins) is A99SB-disp [108]. Its developers claimed that it is one of the best FFs for Aβ40. However, our analysis revealed that it fails to produce Aβ40 structures in agreement with $^{3}J_{HNH\alpha}$ coupling data determined by NMR spectroscopy. The underlying reasons for this is that A99SB-disp shows a preference for PPII conformations, which consequentially leads to an underestimation of the $^{3}J_{HNH\alpha}$ couplings for many of the Aβ40 residues. Therefore, the use of this FF for modeling Aβ40 is not recommended. Least suitable

for modeling Aβ are A03ws and A99SB*-ILDN/TIP3P, which should not be applied to Aβ. After 16 μs of MD with A03ws, Aβ40 became trapped in a highly compact, helical state (Figure 30), which is unsupported by any available experimental data. With A99SB*-ILDN/TIP3P as well, overly compact conformations are found, which in this case result from excessive β-sheet structures.

One question however remains: can one understand the basis for the superior performance observed with A99SB-ILDN/TIP4P-D and A99SB-UCB compared to the other FFs? It is proper to recognize that the influence of modifications of the FFs not covered in the current work can only be estimated. For instance, we cannot say for sure whether adjusting the ψ dihedral parameters in a99SB* (that is, compared to a99SB) played a role in the performance of A99SB*-ILDN/TIP3P since we neither have a 30 μs simulation with a99SB-ILDN/TIP3P nor with a99SB*-ILDN/TIP4P-D as references. However, based on the results from our previous benchmark [122], where A99SB/TIP4P-Ew and A99SB*-ILDN/TIP4P-Ew produced almost identical results for Aβ42 (i.e., the adjustment of dihedral parameters made no difference), the conclusion is that the main reason for A99SB*-ILDN/TIP3P being an inadequate FF choice for Aβ is the water model. And the same holds true for the good performance of A99SB-ILDN/TIP4P-D and A99SB-UCB. The latter FF is based on a99SB/TIP4P-Ew with a modified C-N-C_α-C_β dihedral angle (ϕ') potential, which led to improved conformational ensembles for a number of model peptides [124], and optimized LJ interactions between protein and water, that yielded better agreement with experimental solvation free energies for 47 small molecules that incorporated all of the chemical functionalities of standard protein side chains and backbone groups [123]. The conclusion that the revised protein–water interactions are a key ingredient is further supported by the fact that a99SB/TIP4P-Ew produced too compact and too much structured Aβ42 conformations in our previous benchmark [122].

The same arguments apply to A99SB-ILDN/TIP4P-D. The water model TIP4P-D is based on TIP4P/2005, but compared to that features a significantly higher dispersion coefficient C6, which, like in A99SB-UCB, also leads to increased protein–water vdW interactions, producing IDP ensembles better in agreement with experimental data than the original water model [140]. However, the detailed comparison between

A99SB-ILDN/TIP4P-D and A99SB-UCB shows that the better approach is the adjustment of the LJ parameters on atom type basis with the aim to reproduce experimental solvation free energies of a diverse set of molecules [123] instead of uniformly scaling the vdW interactions between protein and water [140]. This conclusion is further supported by the not convincing performance of A99SB-disp, for which also the protein–water vdW interactions were increased in an amino-acid unspecific fashion [108]. In addition, the ϕ and ψ parameters of all amino acids except glycine and proline were modified during the development of A99SB-disp, which together with the increased protein–water interactions led to the overestimation of the PPII propensity.

In summary, the recommendation is to use A99SB-UCB when simulating Aβ, a FF that was carefully optimized on atom-type basis by Head-Gordon and co-workers [123, 124]. If one wishes to further improve its performance for Aβ, special attention should be devoted to the sequence ^{9}GYEVHH^{14} as here the deviation from the experimental J-couplings is the largest. Though, it should be mentioned that the comparison between the MD ensembles discussed here and the NMR results has certain limitations, as the former were generated at 300 K while the latter were obtained at 277 K. However, it would probably not help to repeat the MD simulations at 277 K as a recent MD simulation study of histatin 5 revealed that for this IDP the current FFs fail to capture the temperature dependence of IDP structures, i.e., the increase in folding upon temperature increase cannot be modeled [223]. The conformational ensemble of histatin 5 produced at room temperature looks almost identical to that obtained at 283 K. Therefore, with respect to our study one can speculate that the FFs that produce extended structures of Aβ40 at 300 K will probably also yield extended structures at 277 K, while A03ws and A99SB*-ILDN/TIP3P are likely to still produce collapsed and partially folded structures. Nonetheless, this conjecture needs to be proven in future by investigating the temperature dependence of the structural ensemble of Aβ – and those of further IDPs – in MD simulations. Moreover, the origin of the FF deficiency of not being able to reproduce this temperature dependence needs to be elucidated and corrected. A possible explanation may be that some of the FFs were reparameterized based on NMR observables obtained at temperatures below room temperature, thus yielding structural ensembles corresponding to low temperatures even though

the MD simulations were performed at room temperature. Despite the limitation that the NMR study and MD simulations of Aβ40 were performed at considerably different temperatures, it should be emphasized that A99SB-ILDN/TIP4P-D and A99SB-UCB produce Aβ40 ensembles not only in agreement with NMR results, but are also in accord with those of two different fluorescence studies which were performed at 297 K and 310 K, respectively. This adds confidence to our conclusion that these two FFs are superior to the others for modeling Aβ.

4.6.2 *What does Aβ40 really look like?*

An obvious question of course is what structures Aβ40 assumes in the MD simulations. As most of the FFs under consideration produced structural ensembles that are not in agreement with experiment, we limit the discussion of structures to A99SB-UCB as the best FF and C36m as one of the better FFs serving as comparison to A99SB-UCB. Given the multitude of structures hidden in an MD trajectory, it is impossible to select one structure and claim it to be the most representative one. Moreover, the aim should be to select structures that are best in agreement with experimental findings. To reach this goal we reweighted the conformations sampled by the MD simulations with A99SB-UCB and C36m using the BME technique to reduce the discrepancies between the calculated and experimental $^3J_{HNH\alpha}$ couplings [211]. The regularization parameter θ in the BME approach, which balances the trust in data versus the amount of simulation data to be kept [224], was chosen as 10 (Figure 34). We observed a reduction in χ^2 from 1.94 to 0.13 for A99SB-UCB and from 3.05 to 0.16 for C36m while retaining a significant fraction of the MD frames (Figure 35). The 50 structures with highest and the 50 ones with lowest weights, i.e., the structures best and worst in agreement with experiment were investigated in more detail. To this end, we calculated R_{gyr}, R_{ee}, the distance between the C_α atoms of residues K16 and D23 as well as D23 and K28.

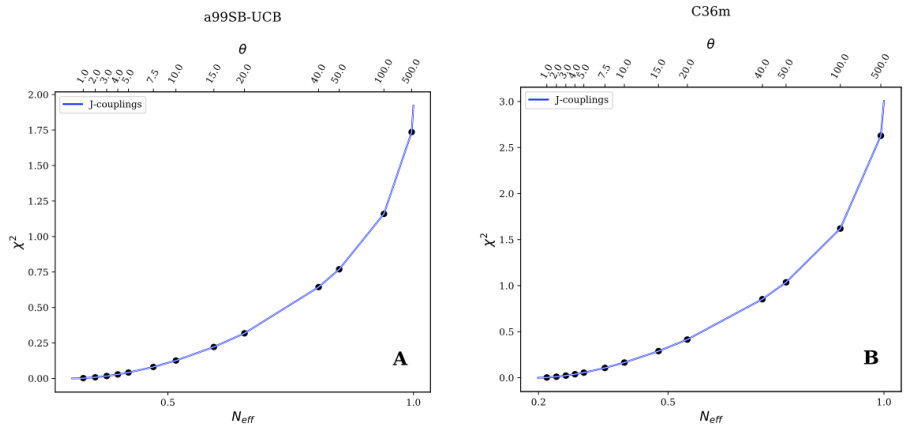

Figure 34: χ^2 versus the effective fraction of frames (N_{eff}) after reweighting the trajectory frames of the simulation with (A) A99SB-UCB and (B) C36m using the maximum entropy principle with different values of the parameter θ. In the BME procedure, one has to choose the parameter θ in such a way that the discrepancy between the experimental and simulation data is reduced (i.e., achieving a low χ^2 value) while minimally perturbing the initial set of weights of the trajectory, which corresponds to retaining a large effective fraction of the MD frames (N_{eff}). This is done by inspecting the χ^2 values vs. N_{eff}, which shows that when $\theta \to 0$, we have $\chi^2 \to 0$ but at the same time N_{eff} becomes very small ($N_{eff} \to 0$). On the other hand, for large θ, we see that χ^2 approaches the initial value (before reweighting) as the weights are least perturbed and as such N_{eff} is also close to 1. A trade-off between these two limits can be found by choosing different θ values, starting from a large one and reducing it until further decrease in θ does not result in a significant decrease in the χ^2 value. Following this, we chose $\theta = 10$ for both FFs.

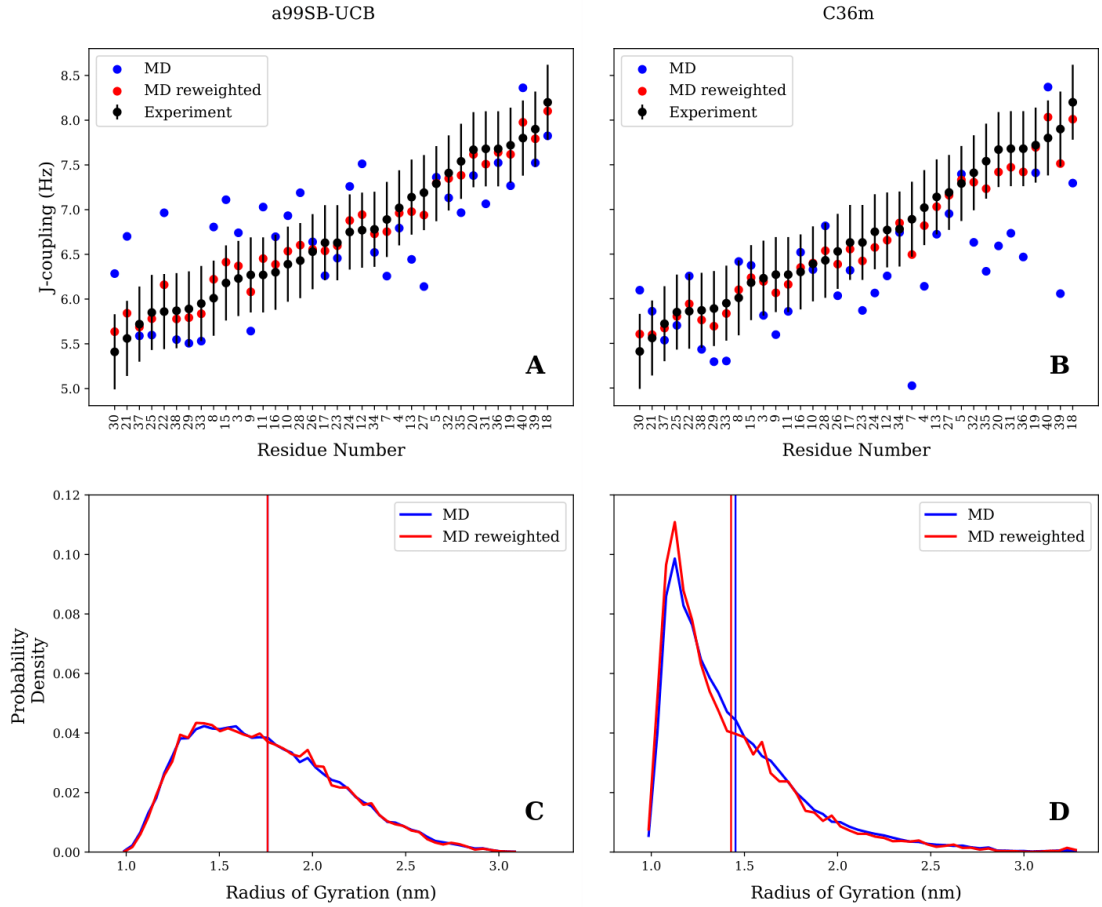

Figure 35: Reweighting of the trajectory frames using the maximum entropy principle to optimize the J-couplings obtained with the MD trajectory with A99SB-UCB (left) and C36m (right). (Top) The black dots indicate the experimental J-couplings for the individual Aβ40 residues (sorted in increasing J-coupling order), blue and red dots indicate the calculated J-couplings before and after, respectively, reweighting. (Bottom) Distribution of the radius of gyration before (blue) and after (red) reweighting the MD frames. The vertical lines indicate the corresponding R_{gyr} average.

For A99SB-UCB, the averages of these values for the high- and low-weight structures are very similar; they are within 0.2 nm of each other, and the R_{gyr} and R_{ee} values are also close to the average values of the whole trajectory. This can also be seen from the distribution of the R_{gyr} values, which did not change much after reweighting the MD frames (Figure 35C). This indicates that large-scale motions are not responsible for the discrepancies from the NMR data of the low-weight structures, which is confirmed by Figure 36 showing the two structures with highest and the two with lowest weights for A99SB-UCB. At first glance, the high- and low-weight structures may look highly similar since they are both generally extended and disordered. However, a closer inspection reveals certain important differences. For instance, the

two low-weight structures feature a F19/F20 turn not present in the two high-weight counterparts. In addition, the N-terminal regions of both groups are different: this, we believe is significant especially since the worst performance for A99SB-UCB was obtained for ^{9}GYEVHH^{14}. In the two high-weight structures the N-terminal sequence up to residue K16 is rather extended, while it is more collapsed in the two low-weight structures. In one of them (Figure 36D) even a helix formed between Y10 and K16, while the overall appearance of this conformation is overly compact.

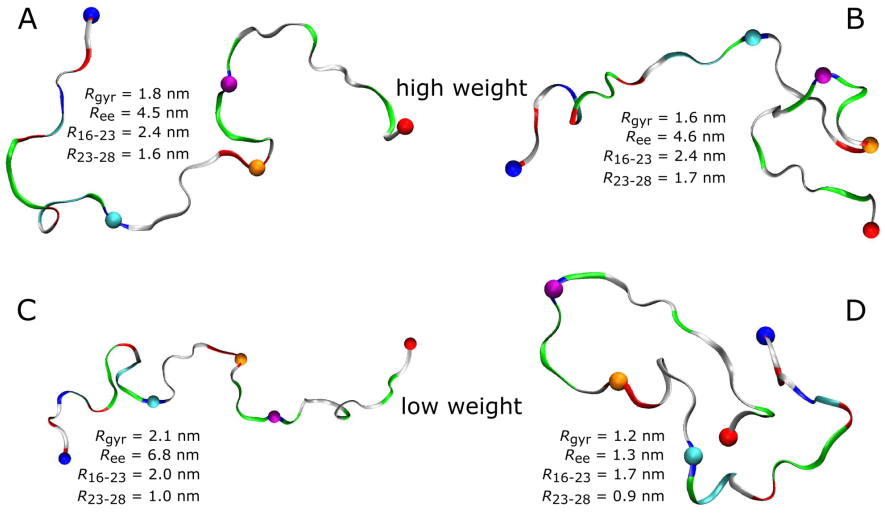

Figure 36: High-weight (A and B) and low-weight structures (C and D) determined by reweighting the A99SB-UCB trajectory using the Bayesian/maximum entropy technique [211]. Aβ40 is shown as band and colored according to amino acid residue type (basic: blue, acidic: red, histidine: cyan, polar: green, hydrophobic: white). Following residues are indicated by spheres: N-terminus (blue), K16 (cyan), D23 (orange), K28 (mauve), C-terminus (red). The structures are characterized in terms of R_{ee}, R_{gyr}, the K16–D23 distance (R_{16-23}), and the D23–K28 distance (R_{23-28}).

The inspection of the corresponding structures for C36m (Figure 37) reveals a similar structural difference for the sequence ^{16}KLVFFAE^{23} as seen for A99SB-UCB. While in the high-weight structures this sequence is rather extended, it exhibits a turn at F19/F20 in the low-weight structures. Thus, based on the J-couplings, a turn in that region should not be sampled. Apart from this turn, visual inspection failed to identify further major differences. In fact, the conformations in Figure 37A and B (high-weight structures) and those in Figure 37C and D (low-weight structures) look rather similar. As for A99SB-UCB it is observed that reweighting the structures does

not change the distribution of R_{gyr} values much. Nonetheless, a small change of R_{gyr} to a lower average value is observed for C36m. This agrees with the finding that the 50 high-weight structures obtained with C36m are generally more compact (average $R_{ee} = 2.79$ nm) than their low-weight counterparts (average $R_{ee} = 3.65$ nm). This is accompanied by a smaller distance between D23 and K28 (0.97 nm vs. 1.24 nm) but a larger K16–D23 distance (2.11 nm vs. 1.90 nm). The conclusion is that, when a FF is not able to provide a generally satisfactory structural ensemble (as in the case of C36m), it might not be possible to optimize for different experimental observables at once; here, agreement with the J-couplings was optimized, which led to a larger disagreement for R_{ee}.

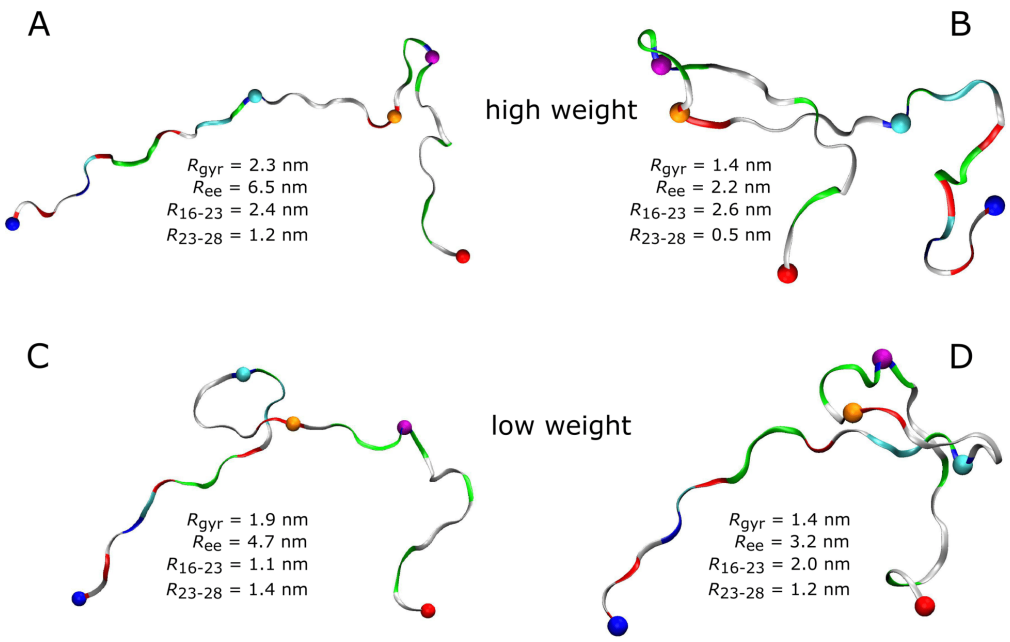

Figure 37: High-weight (A and B) and low-weight structures (C and D) determined by reweighting the C36m trajectory using the Bayesian/maximum entropy technique. Aβ40 is shown as band and colored according to amino acid residue type (basic: blue, acidic: red, histidine: cyan, polar: green, hydrophobic: white). Following residues are indicated by spheres: N-terminus (blue), K16 (cyan), D23 (orange), K28 (mauve), C-terminus (red). The structures were characterized in terms of R_{ee}, R_{gyr}, the K16–D23 distance (R_{16-23}), and the D23–K28 distance (R_{23-28}).

If one wants to gain an in-depth understanding of the structural features sampled in the low-weight structures that are in disagreement with the various available experimental observables, it would be necessary to determine these observables for the

low-weight structures (and also high-weight structures as reference) per residue and then correlate these with the structural preferences as (for instance) measured by the dihedral distribution in Ramachandran space. Based on this, suggestions for possible FF optimizations could be made. An alternative approach would be to employ an automatic reparametrization based on a Bayesian formalism that takes the available experimental data into account [211]. However, considering that the recently attained excellent agreement between a9SB-UCB-simulated Aβ ensemble and experimental data has taken so long to achieve by the simulation research community, it is important that extreme care is taken in performing any reoptimization attempt. The goal should be to make as little changes as necessary in order to avoid losing of what has been achieved. A classical example is shown in the case of A99SB-disp where a FF optimization procedure that was too general and/or too extensive destroyed what had already been gained in A99SB-ILDN/TIP4P-D which for Aβ40 clearly performs better than its optimized A99SB-disp descendant.

4.6.3 *How long should Aβ be simulated?*

The kinetic analysis of the seven 30 μs MD simulation does not only allow us to conclude which of the FFs provides the best structural ensemble of Aβ, but also how long standard MD simulations of Aβ should be if one aims for converged results of the structural ensemble. To answer this question we analyzed the convergence behavior of different quantities, such as the RMSD, number of clusters, secondary structure, R_{gyr} and also R_{ee}, and calculated MSMs based on TICA for dimensionality reduction. The RMSD was identified as a useless measure to assess convergence for the simulation of an IDP like Aβ. If one aims for converged results for R_{gyr} and R_{ee} it seems that for many of the FFs less than 10 μs of MD sampling is sufficient. A time limit below 10 μs has not been provided here since this was not analyzed in the present work. However, from FRET experiments of Aβ it was concluded that conformational dynamics leading to changes of the end-to-end distance does not exceed 1 μs [186]. Thus, it might be that for this quantity a sampling time of 1 μs might even be sufficient. Yet with A03ws and A99SB-disp*-ILDN/TIP3P large changes in both R_{ee} and

R_{gyr} were still observed after 10 µs of MD. However, as both FFs were already identified as not suitable for modeling the structural ensemble of Aβ, in the following discussion we will ignore them. Though it is interesting to note that FFs, which fail to provide an acceptable description of the metastable states, also fail to reproduce the peptide kinetics. This is understandable as conformations in disagreement with experiment are also expected to be connected by transition states different from reality. Moreover, the Aβ conformations favored by these two FFs are overly stable in terms of intrapeptide contacts, which explains the slow kinetics generated with them.

If one uses the secondary structure as a measure for convergence one also finds that 10 µs seem sufficient as the averages of the different structural elements did not change much beyond that time. However, the evolution of the secondary structure shows that for all FFs considerable changes in secondary structure can occur beyond 10 µs (which do not much affect the average values). Thus, the recommendation is to simulate Aβ for at least 10 µs, if possible longer. The same recommendation is made based on the cluster and MSM analyses. The number of clusters converged only beyond a sampling time of 10 µs, and for the best FF, A99SB-UCB, this was achieved only after 20 µs as a result of rare β-sheet formation. This observation agrees with those derived from the MSM analysis. Here, the MFPTs suggest that simulation times of $\gtrsim$ 30 µs are even better as the transitions to rare states require sampling times of that length. As a final check we analyzed the convergence of the time-averaged C_α–C_α distances (Figure 38), an idea originating from testing for ergodicity in supercooled liquids [225] and later applied to Aβ [226]. Depending on how much C_α–C_α distance fluctuations one allows before one considers a simulation to be converged, one again finds that more than 20 or even 30 µs are needed for obtaining stable C_α–C_α distance averages.

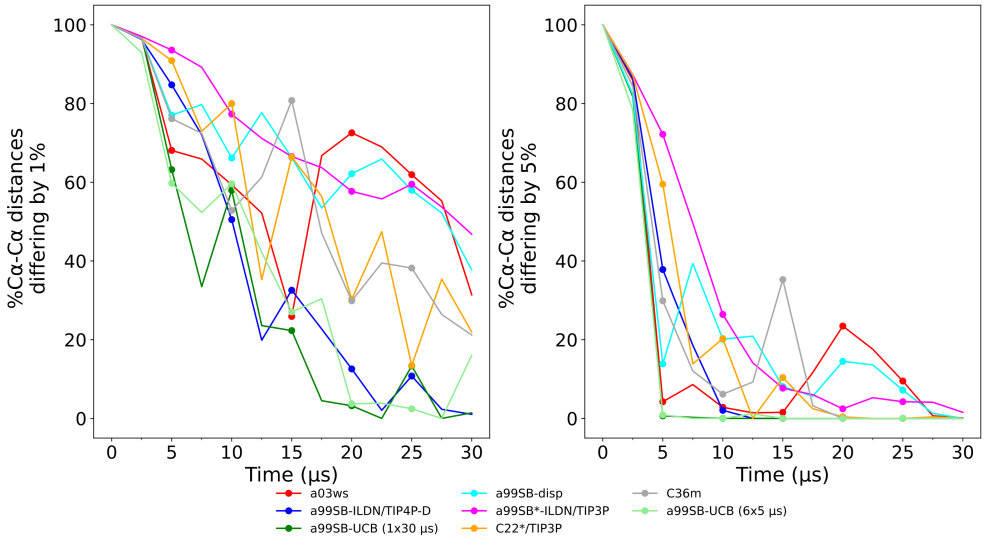

Figure 38: Measure of convergence based on time-averaged C_α–C_α distances. Averages were computed over increasing time window lengths in 2.5 μs increments. The percent of average distances that differ by more than 1% (left) or 5% (right) with respect to the preceding time interval are shown. If one applies the stricter 1% criterion, only A99SB-ILDN/TIP4P-D and A99SB-UCB yield converged C_α–C_α distances within 20–30 μs, which agrees to the conclusions drawn from the various other convergence tests. The second best set of force fields in terms of convergence of C_α–C_α distances is given by C22*/TIP3P and C36m, while the poorest convergence is observed for a03ws, a99SB-disp, and a99SB*-ILDN/TIP3P. If one applies the more generous 5% criterion, all force fields yield converged C_α–C_α distances, yet the order of the force fields with respect to convergence does not change. For instance, considerable motions are still observed between 20 and 25 μs for A03ws and A99SB-disp.

Since the current simulations are limited to 30 μs, one might of course wonder whether completely new conformations would be sampled if the simulations were extended. Another valid question is whether simulating multiple shorter simulations would provide a better approach than sampling one long simulation. To answer these questions we compared our results to those obtained from running 5,119 trajectories between 9.75 and 90.5 ns in length with an aggregated simulation time of 315 μs for Aβ42 modeled by C22*/TIP3P [166]. The resulting MSM identified four states with similar populations and Aβ structures as obtained from our analysis, while several of the MFPTs are lower than those determined here (Figure 39).

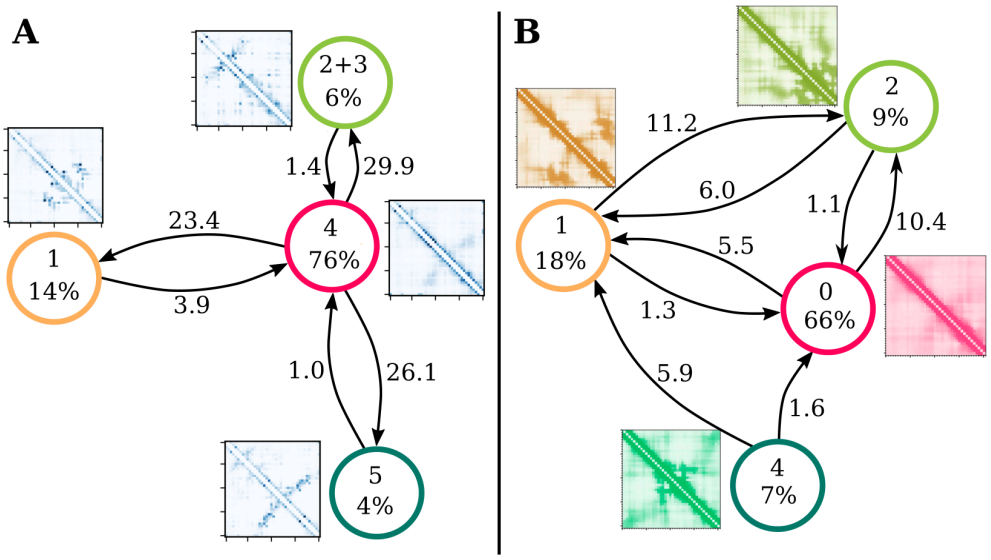

Figure 39: Comparison of (A) the MSM of Aβ40 determined in this chapter from the 30 µs MD simulation in conjunction with the C22*/TIP3P force field with (B) the MSM of Aβ42 derived from 5,119 trajectories between 9.75 and 90.5 ns in length with an aggregated simulation time of 315 µs, also based on C22*/TIP3P [1]. The numbers used for the states correspond to those used in this chapter and in [1], respectively. States 2 and 3 of the MSM for Aβ40 were combined into one state as they harbor very similar structures (see Figure 20). Both MSMs are dominated by the central state corresponding to extended Aβ structures, i.e., state 4 and 0 with a population of 76% and 66%, respectively. These two states have further in common that they tend to form a β-hairpin in the C-terminal half of the peptide. The second most populated state in either MSM is state 1, which is located left to the central state in the MSM, i.e., the transition from the central to this state involves a motion along TIC 1 in negative direction. The corresponding contact maps show that both states are characterized by interactions between residues of the C-terminal half. The states that emerge when going from the central states along TIC 2 in positive direction are the most different from each other when comparing (A) and (B). States 2 + 3 of Aβ40 are characterized by an N-terminal β-hairpin, while state 2 of Aβ42 is dominated by intense contacts limited to residues 30–42. State 5 of Aβ40 and state 4 of Aβ42, which are obtained by transitions from the central state along TIC 2 in negative direction, are rather similar again, with a β-hairpin in the C-terminal half being the dominating structural pattern. Differences between both MSMs are observed for some of the transitions and the accompanying MFPTs (in µs). Many of the MFPTs determined in this chapter are larger than those determined for Aβ42. This is probably caused by the different lag times used for constructing the MSMs: 100 ns in (A) and 12.5 ns in (B) [1]. For other proteins it was shown that lag times $\lesssim$ 100 ns tend to underestimate the MFPTs [2]. Other differences are that some of the transitions sampled for Aβ42, namely $4 \rightarrow 1$ and $1 \rightleftharpoons 2$ were not observed for Aβ40, while for Aβ42 the transition from the central state to state 4 was not sampled. These differences might arise from the fact that two different peptides are compared here, Aβ40 vs. Aβ42, or from the different sampling approaches, i.e., one long simulation vs. thousands of short simulations. Nonetheless, the overall conclusion is that both MSMs are rather similar to each other, suggesting that the 30 µs MD simulation of Aβ40 has reached convergence.

A likely explanation for this discrepancy is the lag time of only 12.5 ns used by Löhr et al. [166]; for a number of other proteins it was shown that lag times $\lesssim$ 100 ns

lead to an underestimation of the MFPTs derived from MSMs [227]. Nonetheless, the comparison with the results obtained by Löhr et al. suggests that it is sufficient if one simulates monomeric Aβ for 30 μs, even when kinetically slow FFs like C22*/TIP3P and C36m are being employed.

4.7 CONCLUSIONS

We analyzed the 30 μs MD simulations that were acquired for Aβ40 using seven different FF/water-model combinations by D. E. Shaw Research [108]. One aim of our analysis was to assess how much of MD time is needed for obtaining fully converged MD ensembles for this peptide. Our analysis of the evolution of different structural quantities as well as Markov state models calculated from the MD data showed that the answer to this question partly depends on the FF used as the different FFs produced different kinetics for Aβ. Since only the FFs that are in agreement with experimental results should be employed, we calculated various NMR and FRET observables from the MD trajectories and compared the resulting values to the experimental ones [186, 212]. This comparison revealed that the best FF field for Aβ is A99SB-UCB, which is based on a99SB/TIP4P-Ew and includes LJ and dihedral modifications implemented by Head-Gordon and co-workers [123, 124]. The second best performance is found for A99SB-ILDN/TIP4P-D, which can also be recommended for modeling Aβ, while all other FFs showed severe failures in reproducing at least one, in most cases more sets of experimental quantities. Usage of A03ws and A99SB*-ILDN/TIP3P for Aβ simulations is clearly discouraged as they produce too much folded Aβ conformations, with too much helix in the case of A03ws and too much β-sheet with A99SB*-ILDN/TIP3P. The other three FFs under study, A99SB-disp, C22*/TIP3P, and C36m, produced acceptable results for Aβ, but considering that there are two FFs that clearly perform better, the recommendation is to use these.

The MSMs resulting from the simulations with A99SB-UCB and A99SB-ILDN/TIP4P-D are both dominated by one state that harbors extended Aβ40 structures with little to none inter-residue contacts beyond direct neighbor contacts. Two further metastable, yet low-populated states (total population of 5%) are identified with both

FFs, which involve β-hairpin formation. With A99SB-UCB the β-hairpins in both these states are centered at residues V24/G25 and involve contacts between F19/F20 with I31/I32. With A99SB-ILDN/TIP4P-D such β-hairpin is also sampled in one of the MSM states, while the other one contains a β-hairpin in the N-terminal region. Transitions to these low-populated states are rare events and most of them are thus associated with MFPTs reaching ≈30 μs or even more. The conclusion therefore is that at least 30 μs of MD sampling is needed to obtain converged trajectories producing the equilibrium distribution of Aβ conformations. This conclusion is supported by the analysis of the convergence of other structural quantities, such as the number of conformational clusters sampled.

Unlike the MFPTs to metastable states, the ITSs derived from an MSM are a measure for intrapeptide reorientations. For these motions, an upper limit of 1 μs was predicted from different fluorescent spectroscopic measurements [186, 187]. Only for A99SB-UCB and A99SB-ILDN/TIP4P-D the slowest implied time scales are below 1 μs, while with all other FFs the kinetics of Aβ40 is predicted to be much slower, in the cases of A03ws, A99SB-disp and A99SB*-ILDN/TIP3P the slowest intrapeptide motions even reach time scales beyond 10 μs. Thus, also in terms of kinetics A99SB-UCB and A99SB-ILDN/TIP4P-D are the only two FFs in agreement with experiment. Nonetheless, even though the thermodynamics and kinetics of Aβ is modeled well with these two FFs, there is still further room for improvement. For instance, while A99SB-UCB is very good in reproducing NMR values for the C-terminal side of the peptide, this is less so the case for the region G9–H13. Thus, further FF improvements for Aβ should focus on this region, while keeping the level of quality for the rest of the peptide. Overall, a major step forward in terms of FF quality for Aβ has been reached with A99SB-UCB and also A99SB-ILDN/TIP4P-D. It will be interesting to see how the kinetics of Aβ oligomer formation and the resulting structures will look like when simulated with these FFs, i.e., when the aggregation process is initiated from extended, disordered, and not partly folded states as was the case due to FF bias in previous Aβ aggregation simulations.

5

APPLICABILITY OF FORCE FIELDS IN AGGREGATION PATHWAYS OF AMYLOIDOGENIC PEPTIDES

Parts of this chapter were submitted by the author to the *Journal of Chemical Information and Modelling* as Samantray, S., Yin, F., Kav, B., Strodel, B. (2020). Different force fields give rise to different amyloid aggregation pathways in molecular dynamics simulations.

5.1 MOTIVATION

Decoding the conformational dynamics of IDPs is a critical step in understanding the process of protein aggregation and fibrillation [228]. However, the structural flexibility and high aggregation propensity impede experimental investigations to capture the dynamics of IDPs at atomic level [229]. An alternative approach to experiments is provided by molecular simulations, which allow for the necessary temporal and spatial resolution to follow the motions of IDPs [49, 230]. The recent years have seen the development of FFs that allow the reliable modeling of IDPs using MD simulations [108, 139, 140, 185, 230]. The flexible nature of IDPs necessitated the FFs originally developed for folded proteins to be revised in order to accurately characterize the unfolded protein state.

However, while numerous FFs have been developed and benchmarked for IDPs (see ref [231] and references therein), it remains to be shown that they can also capture their aggregation behavior correctly. Our group compared the performance of several FFs for the formation of hexamers of the $A\beta_{16-22}$ peptide, which is the sequence ^{16}KLVVFAE^{22} of Aβ, and mutants of this peptide [232]. This benchmark included

FFs from AMBER, CHARMM, GROMOS, and OPLS. One of the main conclusions was that G54A7 [233] and OPLS-AA [175, 176] overstabilize protein–protein interactions, leading to an overestimation of the aggregation speed and an inhibition of protein-aggregate dissociation. Thereafter, Derreumaux and coworkers investigated the protein aggregation behavior for the $A\beta_{16-22}$ dimer using 17 different FFs in combination with conventional MD simulations [109], following up their previous work where they had employed replica exchange MD to study dimers and trimers of $A\beta_{16-22}$ [234]. They concluded that FFs from CHARMM with updated CMAP correction [128, 235] such as C22* [183], C36 [236], and C36m [185] ('C' stands for CHARMM), along with FFs based on AMBER99 [237] ('A' stands for AMBER) with modified torsional parameters for the backbone and for the four amino acids Ile, Leu, Asp, and Asn like A99SB-ILDN [238] and A14SB [239] are best suited for studying amyloid aggregation. From these FFs, C36m is the only one developed for IDPs which was realized by refining backbone potentials in order to model the preference of IDPs to adopt extended structures [185]. Recently, researchers at D. E. Shaw Research developed an FF aimed at being applicable to both folded and unfolded proteins leading to A99SB-disp [108]. To reach this goal, Robustelli at al. used A99SB*-ILDNQ [238] in combination with the TIP4P-D water model [140] as starting point and introduced modifications to the backbone torsional potential and enhanced the interaction potential between the backbone carbonyl oxygens and backbone amide hydrogen atoms, which is thought to increase the overall stability of extended conformations as present in β-sheets [108]. In addition, they revised the water model by increasing the C6 term determining the attractive part of the Lennard-Jones interactions, which is expected to avoid hydrophobic collapse of the proteins. An increased protein–water interaction usually leads to a stabilization of extended conformations [106, 140].

The tendency of Aβ towards aggregation has been proposed to result to a large extent from its hydrophobic core region (residues 17–21). While the $A\beta_{16-22}$ segment is not sufficient to understand the aggregation of full-length Aβ, since the latter involves 5–6 times more residues compared to the former, which not only increases the conformational space but also influences the aggregation behavior, this short peptide is nonetheless an attractive model for studying amyloid aggregation. First, $A\beta_{16-22}$

is able to form fibrils itself, which are characterized by an antiparallel ordering of the peptides forming the β-sheets [240]. Second, given its small size it is ideally suited for exploring the thermodynamics and kinetics of its aggregation using experiments [241–243] and also MD simulations. The rigorous mutation study by Senguen et al. showed that π–π interactions do not play an important role during $A\beta_{16-22}$ aggregation [241]. Instead, the hydrophobicity of the amino acids in the region 17–21 seems to be the dominant factor determining the aggregation speed of $A\beta_{16-22}$ and mutants of it, while electrostatic contacts between K16 and E22 provide further stability of the $A\beta_{16-22}$ aggregates and ensure proper orientation in antiparallel β-sheets [244]. In the past two decades, numerous coarse-grained and all-atom MD simulations have been performed for studying the aggregation of this peptide into small oligomers [109, 232, 244–249]. Nguyen and coworkers have compiled a comprehensive review of the simulation studies of $A\beta_{16-22}$ including both all-atom and coarse-grained peptide models [250].

However, a valid question is – considering the vastly different results that were obtained from FF benchmarks of Aβ [122, 172–174] and other IDPs [106] – to what extent the simulation results of $A\beta_{16-22}$ aggregation are affected by the FFs used in these studies. This question is addressed in this chapter, aiming to elucidate the FFs that are suited to study amyloid aggregation using the heptapeptide $A\beta_{16-22}$ and two of its mutants as a test cases. Experiments by Senguen et al. showed that the single mutant F19L $A\beta_{16-22}$ (m1) forms fibrils faster than wild-type $A\beta_{16-22}$ (*wt*), while the double mutant F19V/F20V $A\beta_{16-22}$ (m2) does no aggregate at all [241]. These findings are difficult to explain based on hydrophobicity alone as Phe, Leu and Val are of very similar hydrophobic nature. Since Senguen et al. derived their results from sedimentation assays and dynamic light scattering that followed the monomer concentration of each of these peptide sequences until fibrils were formed [241], they allow us to compare our simulation results on small fibrillar oligomers with the experimental observations [232]. In the first part of this chapter, we elucidate the structural transitions of the three peptide sequences at the monomeric level since it is known that the monomer state has an effect or even controls the aggregation process [219, 251–253]. Secondly, we investigate the formation of hexamers by these peptides and establish

links between the monomer configuration and the oligomer state. To complement this view, we finally check the stability of a preformed steric zipper involving twelve copies of either *wt*, m1, or m2, providing a conclusive evidence regarding which of the FFs under study is best suited to study amyloid aggregation using MD simulations. The FFs included in our test set are A99SB-disp, C36m and C36mW (which is based on C36m but includes more favorable vdW interactions between protein and water) recently developed for IDPs as well as the older force fields G54A7 and OPLS-AA, which are already known to overstabilize protein–protein interactions [232] but serve as a useful reference here.

5.2 MODEL AND SIMULATION DETAILS

To understand the influence of the different FFs on the kinetics and structures of amyloid aggregation, we simulated residues 16–22 of the amyloid-β peptide, considering the wild-type (*wt*) and the two mutant F19L (m1) and F19V/F20V (m2) sequences. We capped the N- and C-termini of the peptides with acetyl (ACE) and N-methlyamide (NME) groups, respectively, to mimic the experimental conditions [241]. Each peptide was simulated using five different force fields with their respective water models: A99SB-*disp* [108] with modified TIP4P-D [140] (A99-d), C36m [185] with TIP3P [254] (C36m) and with increased protein–water interactions [185] (C36mW), G54a7 [233] with SPC [255] (G54a7), and OPLS-AA [175, 176] with TIP4P [177] (OPLS).

5.2.1 *Monomer of $A\beta_{16-22}$ and its mutants*

We simulated the monomeric peptides starting from extended states and solvated them with a cubic water box. We set the minimum distance between the peptides and the edges of the water box to 1.2 nm. Only for the translational diffusion simulations, we placed the monomers 1.7 nm away from the box edges, effectively doubling the simulation box volume. We assigned the ionization states of lysine and glutamic acid at pH 7 to be protonated and deprotonated, respectively, resulting in electrostatically neutral peptides. In all simulations, we added Na^+ and Cl^- ions to reach

an NaCl concentration of 150 mM. We minimized each system using the steepest descent algorithm, followed by equilibration, first with a 10 ps run in the NVT ensemble while constraining the heavy peptide atoms to their initial positions, afterwards with a 10 ps run in the NpT ensemble without position constraints. For the production runs we simulated each system for 500 ns in the NpT ensemble with $T = 298$ K and $p = 1$ bar. We obtained three independent trajectories for each system starting from the same initial structures but applying independent minimization and equilibration procedures. For the translational diffusion simulations we collected three 400 ns long trajectories for each system. Throughout all simulations we constrained all bond lengths using the LINCS algorithm [191]. The electrostatic and van der Waals interactions were calculated using the particle mesh Ewald (PME) method [192] and the real-space components truncated at 1.2 nm. We controlled the temperature and pressure using a velocity rescaling algorithm [189] with a relaxation time of 0.1 ps and a Parrinello-Rahman barostat [256] with a relaxation time of 2 ps, respectively. For the simulations with G54a7 and OPLS we employed virtual sites for the nonpolar hydrogen atoms [257] allowing for an integration time step of 4 fs, while a time step of 2 fs was used for the remaining FFs.

5.2.2 *Hexamer of $A\beta_{16-22}$ and its mutants*

We introduced six peptides into a cubic box with 10 nm edge length, corresponding to a peptide concentration of 10 mM, to study the oligomer formation for each of the three peptides and FFs. The initial configurations for these simulations were generated with the software PACKMOL [258] using the most populated peptide structures identified in the monomer simulations. We positioned the monomers in such a way that none of the monomer–monomer distances with respect to any atom pair was smaller than 0.4 nm or greater than 1 nm. The simulations were set up using the same procedure as described for the monomer systems in section 5.2.1. For A99-d, C36m, and C36mW we obtained three independent trajectories of 1 µs length each, while for G54a7 and OPLS we used the existing data from previous simulations performed

within our group [232], which contain five production runs in the NpT ensemble for 300 ns each.

5.2.3 *Steric zipper models of of $A\beta_{16-22}$ and its mutants*

In order to corroborate the results from the oligomer formation simulations, we tested the stability of preformed minifibrils composed of twelve peptides stacked in two layers with six peptides forming an antiparallel β-sheet in each layer (Figure 40). This arrangement is also called steric zipper [259]. These minifibrils were generated using the microcrystal structure of KLVFFA as determined by X-ray crystallography (PDB code 3OW9) [260] as starting point. After adding the E22 residue as well as the terminal capping groups ACE and NME to each of the twelve peptide chains, which was accomplished with PyMOL [261], the peptides had to be aligned again so that the terminal residues K16 and E22 were next to each other in the antiparallel β-sheet and above each other in the double layer. In the case of the mutants m1 and m2 the mutations F19L and F20V/F20V were introduced too. We placed each of the minifibrils in the center of a cubic water box of size 15 nm in each spatial dimension. We added Na^+ and Cl^- to adjust the salt concentration at 150 mM. We performed the minimization, equilibration, and production runs with the same simulation parameters as described in section 5.2.1. For each system three production runs in the NpT ensemble and of 300 ns length were carried out. These simulations testing the stability of the steric zipper conformation were performed for A99-d, C36m, and C36mW.

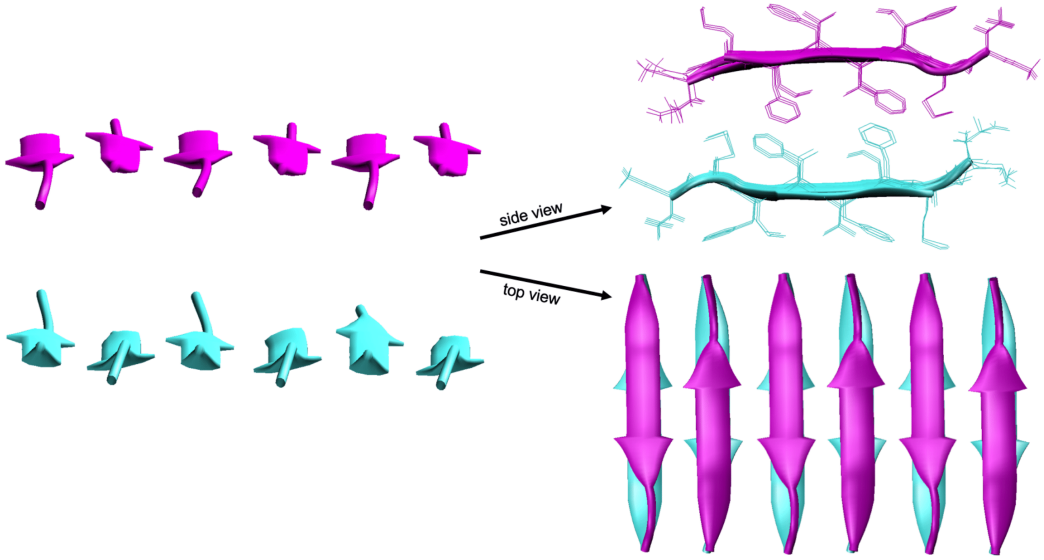

Figure 40: The initial structure of the minifibril.

5.2.4 *Biophysical Characterization Methods*

5.2.4.1 *Structural characterization*

The simulations were analyzed using a combination of standard GROMACS tools, VMD [195], and in-house Python scripts [3] invoking the MDAnalysis [196] and MDTraj [262] libraries. We determined the representative monomer structures of the peptides using the GROMOS clustering algorithm of Daura et al. [156] with a cutoff of the RMSD of 0.2 nm. For the intra- and interpeptide contacts, we considered two residues to be in contact if the distance between any pair of atoms from residue *a* and residue *b* is 0.4 nm or less. Based on this distance cutoff also the size of the oligomers was determined. The nonbonded interaction energies consisting of van der Waals (vdW) and electrostatic interactions were calculated using the *rerun* option of GROMACS *mdrun* for all intra- and interpeptide residue–residue pairs. In order to calculate the interaction energies between peptides, we extracted the peptide pairs present in the simulations using the distance criterion of 0.4 nm or less between the two peptides. Thus, peptide pairs as present in dimers and higher-order oligomers up to hexamers are considered for this analysis. For the characterization of the steric zippers we calculated the number of hydrogen bonds (H-bonds) formed between the

peptides within a β-sheet as well as the nematic order parameter, S_2. An H-bond is defined to be formed if the donor–acceptor distance is less than 0.35 nm and the donor-H-acceptor angle is less than 30°. The nematic order parameter,

$$S_2 = \left\langle \frac{3}{2}\cos^2\theta - \frac{1}{2} \right\rangle , \tag{28}$$

where the angle brackets refer to time and ensemble averaging and θ is the angle between the vector pointing from the N-terminus to the C-terminus of one peptide and the same kind of vector of another peptide in the system, was calculated with MDTraj. It describes the orientational order of a system with values ranging between 0 and 1 corresponding to isotropic and anisotropic systems, respectively.

5.2.4.2 *Transition networks*

In order to visualize the aggregation pathways, we constructed transition networks to elucidate the important intermediate stages [232, 248, 263, 264]. We define the states in the TNs in terms of the oligomer size, which ranges from 1 for monomers to 6 for hexamers, and the β-sheet content divided into five ranges with 0–20%, 20–40% etc. In the resulting network models, which were plotted with Gephi [265], the nodes represent the intermediate aggregation states and the edges the transitions between these states. The size of the nodes are proportional to the population of the states and the edge thickness to the mass flux between the connected states. We calculated the β-sheet content based on the dihedral angles along the peptide backbone. A β-sheet is assumed to be formed when the ϕ and ψ values are located within the polygon with the vertices at (−180°, 180°), (−180°, 126°), (−162°, 126°), (−162°, 108°), (−144°, 108°), (−144°, 90°), (−50°, 90°), and (−50°, 180°) [244, 248]. This definition allows to also assign a monomer to the β-state, which would not be possible if one used, as commonly done, the H-bond pattern between the peptides instead.

5.2.4.3 *Translational diffusion*

In order to accurately compute the translational diffusion constants of the $A\beta_{16-22}$ peptide and its mutants, we extended the trajectories of the monomers to 400 ns in

the NpT ensemble. It is important to note that the velocity-rescaling thermostat used in our simulations [189] has been shown to produce transport properties indistinguishable from the NVE ensemble [266, 267]. Therefore, we did not run additional simulations in the NVE or NVT ensemble. We used the mean-squared displacement (MSD) and the Einstein relation in three dimensions to calculate the translational diffusion constants of the peptides,

$$D_t = \frac{1}{2d} \lim_{t\to\infty} \frac{d}{dt} \langle [r(t) - r(0)]^2 \rangle \,, \tag{29}$$

where D_t is the diffusion constant, d is the system dimension, $r(0)$, $r(t)$ are the particle positions at times $t = 0$ and t, and the term with angle brackets represents the MSD. For normal diffusion, the MSD grows linearly with sufficiently large values of time t. As a result, the three-dimensional translational diffusion constant D_t is equal to one sixth of the slope of the linear region of the MSD-vs-time curve. In order to obtain the MSD of the peptides, we divided each trajectory to 10 ns time windows with 5 ns lag time between two consecutive windows. This yields 80 subtrajectories per 400 ns trajectory, from which the MSD of the entire peptide was obtained by averaging over all atoms and subtrajectories. The three-dimensional translational diffusion constant D_t was calculated from the slope of the subtrajectory- and atom-averaged MSD curves by fitting a linear curve between 2 and 8 ns. The final translational diffusion constant D_t per system (i.e., per peptide and FF) and the associated standard error is reported as the average over three trajectories.

5.3 RESULTS AND DISCUSSIONS

5.3.1 *Structural characterization of the $A\beta_{16-22}$ monomer*

When studying oligomer formation, it is important to separate the effects coming from the monomeric structure from those originating from interpeptide interactions that drive the aggregation process. In this section, we assess how the different FFs af-

fect the monomeric ensemble of Aβ_{16-22} based on RMSD-based geometric clustering as well as intrapeptide contact maps and interaction energies.

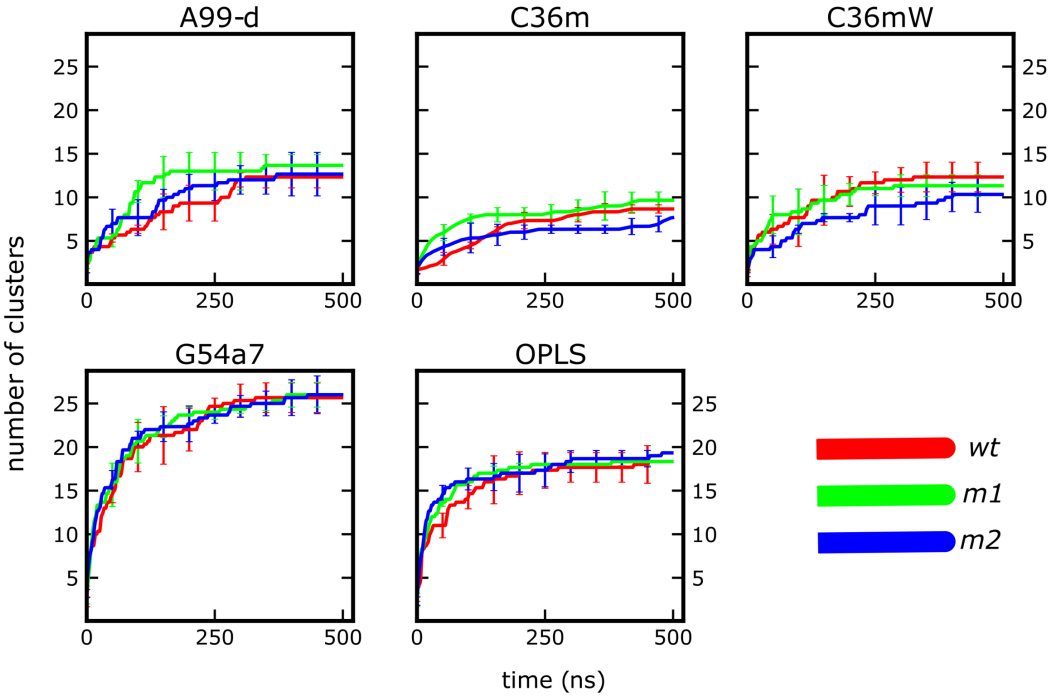

Figure 41: The number of structural clusters as a function of time for *wt* (red), m1 (green), and m2 (blue) from the simulations of the monomers using A99-d, C36m, C36mW, G54a7, and OPLS. The averages over three independent simulations per system are shown. The error bars indicate the standard error of the mean. Representative cluster structures are shown in Figure 42.

5.3.1.1 *Structural clustering and characterization*

We performed an RMSD-based cluster analysis to explore the conformational heterogeneity in the ensemble of peptide structures generated by the MD simulations of a single *wt*, m1, and m2 peptide using the different FFs. By clustering structures that are within an RMSD of 0.2 nm, we calculated how the total number of observed clusters changes with time (Figure 41). The plateau region of the number of clusters over time indicates that from ≈300 ns onwards no new structures different from the clusters already identified were sampled, which applies to all FFs. From the study of Aβ_{1-40}, we identified this cluster number criterion as a highly suitable measure for assessing the convergence of unfolded peptides [268]. The number of clusters obtained signifies the unique structures sampled during the MD simulations. Irrespective of the peptide sequence the G54a7 force field yields the largest number of

clusters, followed by OPLS, while C36m yields the lowest number of clusters. By counting the frequency of each cluster throughout the simulation trajectories, we observe that with A99-d, C36m, and C36mW the two dominant clusters are significantly more populated than the counterparts obtained with OPLS and G54a7. For instance, the first two clusters obtained with A99-d, C36m, and C36mW contain >85% of the structures for the *wt* peptide compared to 72% with OPLS and only 63% with G54a7 (the full list is given as Table 3). This observation indicates that with the newer FFs the peptides are less flexible, which can result from increased peptide–water interactions, as introduced in A99-d [108] and C36mW [185], and/or increased torsional barriers implemented to enforce extended peptide structures.

Peptide	**FF**	**Population**	
		1st cluster	**2nd cluster**
wt	**A99-d**	74%	12%
	C36m	75%	13%
	C36mW	89%	8%
	G54a7	40%	23%
	OPLS	53%	19%
m1	**A99-d**	56%	19%
	C36m	64%	19%
	C36mW	72%	16%
	G54a7	49%	10%
	OPLS	43%	17%
m2	**A99-d**	66%	14%
	C36m	69%	15%
	C36mW	87%	8%
	G54a7	53%	16%
	OPLS	41%	19%

Table 3: Population (in %) of the first two clusters obtained from structural clustering of the monomer conformations sampled for wt, m1, and m2.

The preference for extended structures in the simulations with the IDP-corrected FFs becomes visible by the inspection of representative structures of the most populated clusters (Figure 42). For *wt*, A99-d, C36m, and C36mW produce an extended structure, whereas G54a7 results in a π-helix conformation and OPLS leads to a turn structure. A similar structural pattern is observed for m1 and m2, i.e., extended structures with A99-d and C36mW and structures with more intrapeptide interactions with G54a7 and OPLS. As found in previous studies we thus observe that the differences between force fields for the same peptide are larger than the differences between different peptides but using the same FF [106]. However, it should be noted that we do not know the experimental structures of *wt*, m1 and m2; it is thus not clear whether they should possess different monomer structures and which of the FFs provides the better description for their conformational ensembles. If one looks into the details, some small peculiarities are found, such as a slightly reduced flexibility of m2 compared to *wt* and m1 when modeled with C36mW, or that m2 is more extended than the other two peptides in the simulations with G54a7. These findings are in agreement with the higher β-propensity of Val compared to those of Phe and Leu. For instance, for exposed residues Fujiwara et al. assigned a β-propensity of 2.31, 1.50, and 1.18 for Val, Phe, and Leu, respectively [269].

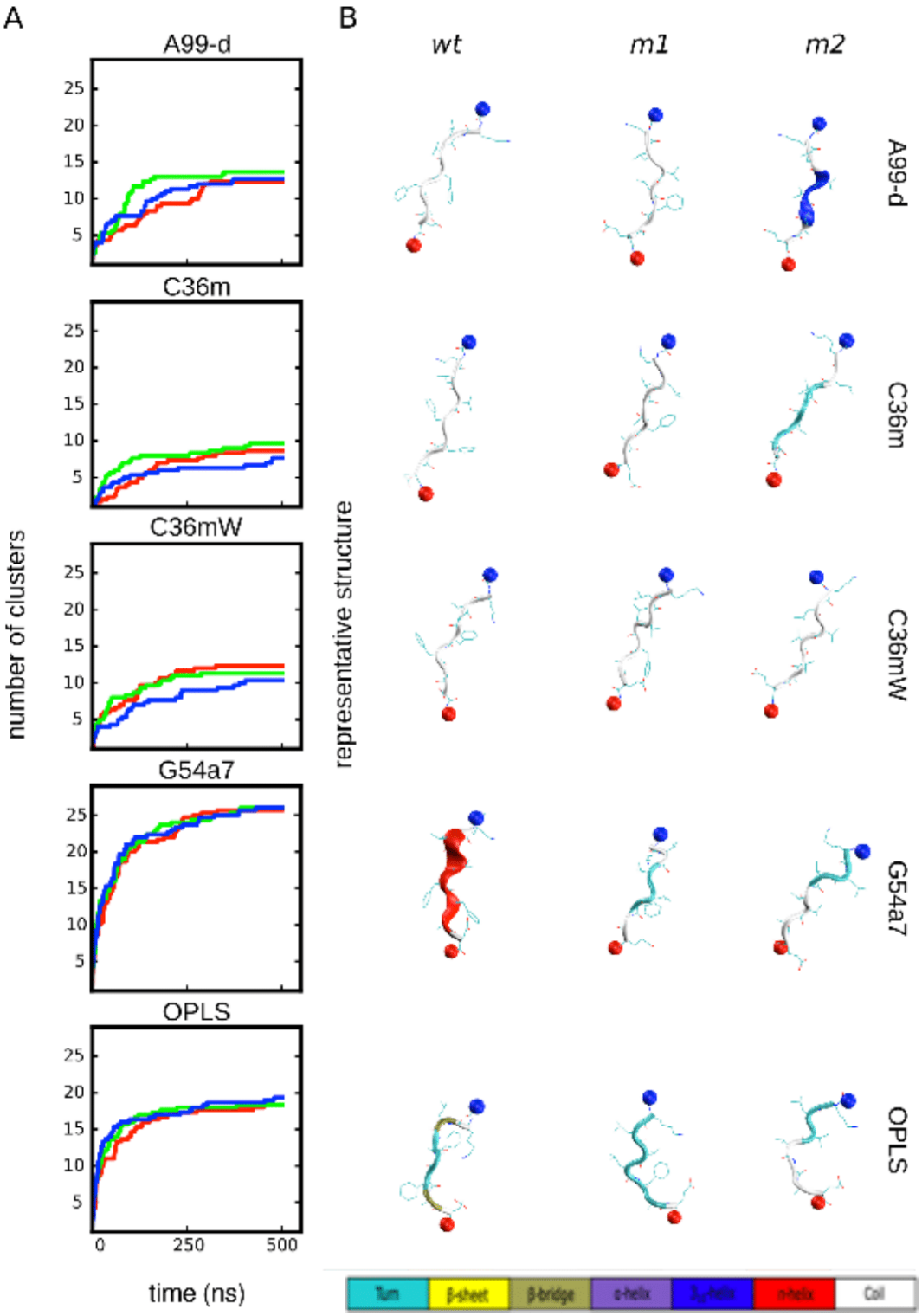

Figure 42: (A) The number of clusters as a function of time for *wt* (red), m1 (green), and m2 (blue) from the simulations of the monomers using A99-d, C36m, C36mW, G54a7, and OPLS. Averages over three simulations per system are shown. (B) Representative peptide structures from the corresponding most populated clusters. The spheres colored in red and blue represent the N- and C-terminus, respectively. The assignment of the secondary structure is according to the color key at the bottom.

5.3.1.2 *Intrapeptide contacts and interaction energies*

To better understand the structural preferences predicted by the different FFs, we characterized the intrapeptide interactions by calculating the frequency of residue–residue contacts (Figure 43). The resulting contacts indicate that among the FFs developed for IDPs, C36mW produces the highest preference for extended structures and A99-d the lowest. A99-d leads in particular to more contacts beyond the first- and second-neighboring residues in m1 and m2, which in the case of m2 is against the propensity of Val to adopt the β-state [269]. C36mW is the FF that best reproduces the preference for extended structures in the order Val (m2) > Phe (*wt*) > Leu (m1). With C36m, on the other hand, this behavior cannot be reproduced, which is visible from the higher presence of intrapeptide contacts including electrostatic contact between K16 and E22 in m2 compared to *wt*. G54a7 and OPLS produce many intrapeptide contacts indicating collapsed monomeric structures, apart from m2 modeled with G54a7.

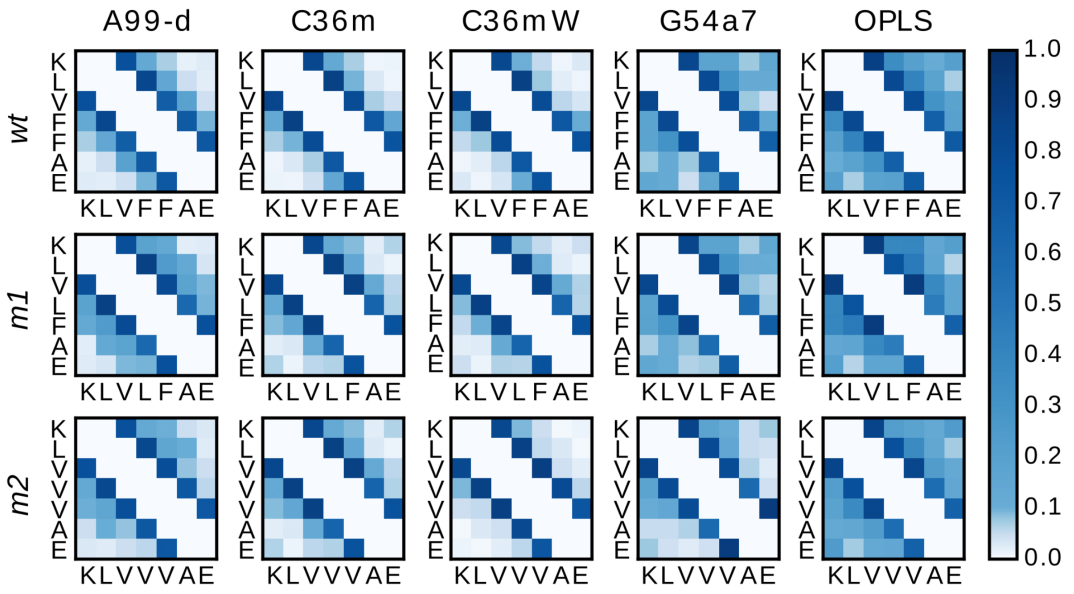

Figure 43: Probability of intrapeptide contacts for monomers of *wt*, m1, and m2 from simulations using A99-d, C36m, C36mW, G54a7, and OPLS. The color code represents the probability of a contact between residues during the MD simulations, averaged over the three runs per system. For the sake of clarity, the diagonal and first off-diagonal elements of the contact maps corresponding to self-contacts and contacts with direct neighbors are not shown.

The intrapeptide contacts can be understood based on the inter-residue interaction energies, which can be dissected into electrostatic and vdW interactions (Figure

S3A and B, respectively). G54a7 and OPLS predict strong electrostatic interactions between K16 and E22, which cause the monomer structures to collapse, entailing several other intrapeptide contacts of vdW nature. An exception is m2 modeled with G54a7 where more extended structures are produced. However, here an electrostatic interaction between K16 and V18 plays a role, which is also present in *wt* and m1 modeled with G54a7, but to a lesser extent. This contact, which is an H-bond between the carbonyl oxygen of K16 and the amide hydrogen of V18, is only present in the simulations with G54a7 and OPLS, but none of the IDP-corrected FFs. Interestingly, with A99-d and C36mW the K16–E22 interaction is also of considerable intensity in m1, which mirrors the medium flexibility of Leu, yet does not lead to completely collapsed structures. This electrostatic attraction must thus be counteracted by solvation energies or torsional parameters of the other residues favoring overall extended structures. For *wt* and m1, none of A99-d, C36m, and C36mW predicts the interaction between the termini to be of (large) relevance as the corresponding interaction energies are close to zero. This suggests that these peptides are generally more extended than m2, which agrees with the contact maps in Figure 43 and can be explained with the high propensity of Phe and Val to adopt a β-conformation. Finally, it can be concluded that for none of the FFs and peptides the vdW interactions play a large role in the structure formation of the monomer.

5.3.2 *Peptide–water interactions*

Peptide structure and dynamics emerge from an interplay between the peptide–peptide and peptide–solvent interactions. To understand the latter, we quantify the first hydration shell of *wt*, m1 and m2 and their translational diffusion.

5.3.2.1 *Hydration shell*

The first hydration shell is defined as the water layer around the peptide in which all the water molecules are directly in contact with the peptide [270]. To fulfill this definition, one needs to i) define the peptide atoms to be used for computing the contacts, and ii) set a cutoff distance between the selected peptide atoms and the

water molecules to define the contacts. There are several procedures available in the literature to obtain the cutoff distance and the peptide atoms [271–275]. Here, we use the non-hydrogen peptide atoms to define the contacts and a single cutoff distance of $r_{cut} = 0.45$ nm between the peptide non-hydrogen atoms and the water oxygens. This particular choice emerged from the analysis of the MD trajectories using different cutoff distances and contact definitions.

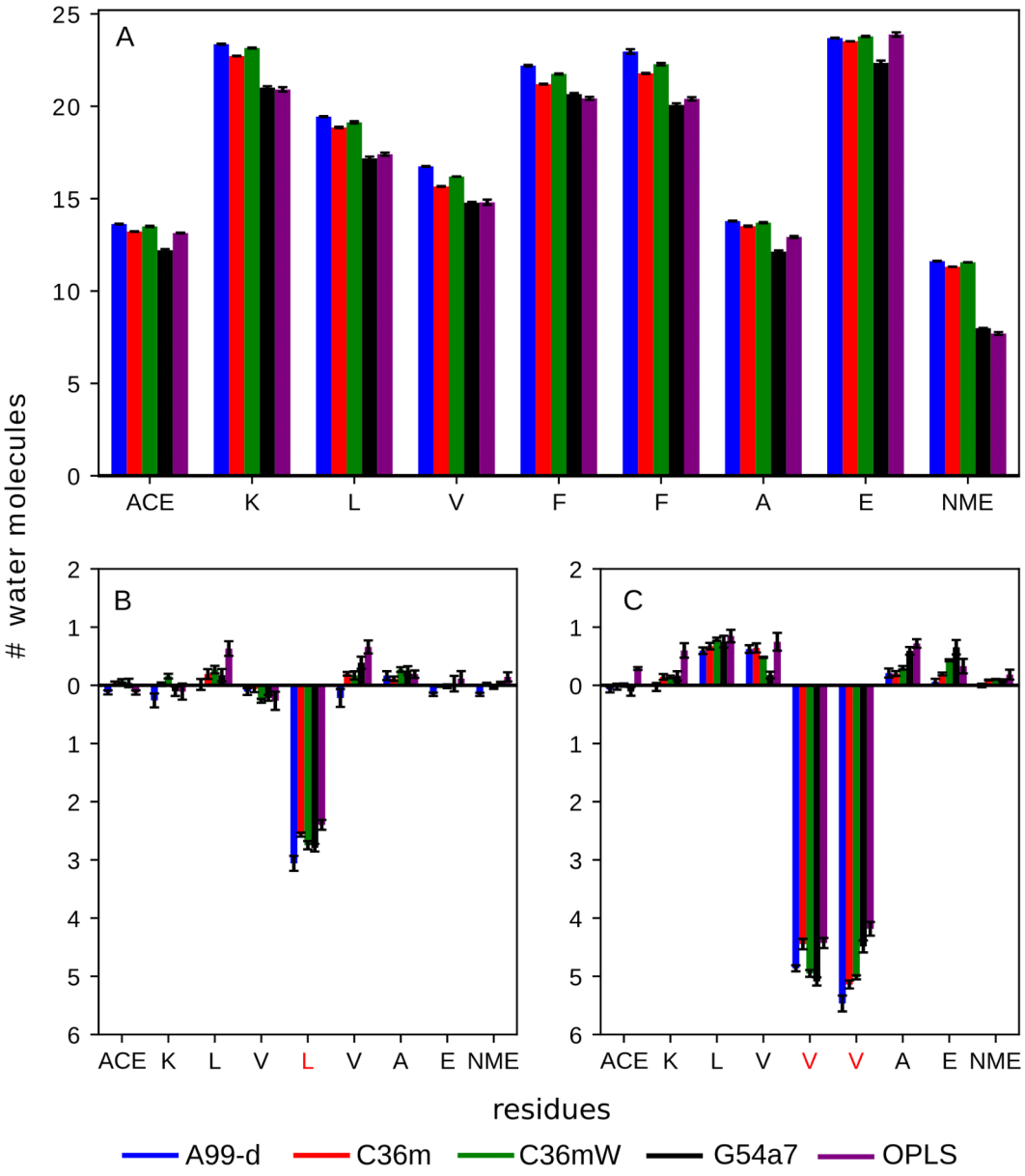

Figure 44: (A) The number of water molecules at a cutoff distance $r_{cut} = 0.45$ nm around each residue for the *wt* peptide. (B) and (C) The difference between the number of water molecules at a cutoff distance $r_{cut} = 0.45$ nm per residue for $m1 - wt$ and $m2 - wt$, respectively. The amino acid sequence for each peptide is given at the x-axis. The red-labeled residue names correspond to the mutations in (C) and (D). Values averaged over three trajectories are shown, along with the standard error of the mean. The color key for the FFs is given at the bottom.

Figure 44A shows the number of water molecules in the first hydration shell on a per-residue basis. The solvation pattern is very similar across the FFs. There are only few notable differences. For *wt* we find that, except for E22, A99-d always yields slightly larger numbers of water molecules per residue than the other FFs, followed by C36mW and C36m. G54a7 and OPLS cause less solvation, especially at the C-terminal capping groups, which is accompanied by a higher K16–E22 contact probability as seen above. Thus, G54a7 and OPLS are more favorable for residue–residue interactions than the the other three FFs where the strength of the protein–water interactions was increased on purpose (A99-d and C36mW) and/or the preference for extended structures was increased. The effect of the mutations in m1 and m2 can be observed in Figures 44B and 44C, respectively. For the m1 variant, the F19L mutation results in less hydration at the mutation site, which can be explained by the smaller volume of the side chain of Leu compared to that of Phe. For the other residues of m1 the differences are small. In the case of m2, both mutation sites F19V and F20V are less hydrated compared to *wt*, which again can be explained with the smaller side-chain size. OPLS predicts for both mutation sites of m2, as also for the mutation in m1, significantly lower differences compared to the other FFs. For the other m2 residues we observe that they are more hydrated compared to *wt* and also have more water molecules in the first hydration shell than the m1 variant. OPLS yields the largest differences for these residues. Thus, with respect to the whole peptides OPLS predicts a more similar solvation for m2 and also m1 compared to *wt*.

5.3.2.2 *Solvent accessible surface area*

One factor that determines the number of water molecules within the first hydration shell is the solvent accessible surface area (SASA) of the solute. We calculated the SASA are for *wt*, m1, and m2 using a standard 0.14 nm probe radius. Our results in Table 4 show that within each FF *wt* has the largest SASA and m2 has the lowest, which can be explained with the different sizes of Phe, Leu, and Val. The differences across the FFs for each peptide positively correlate with differences found for the number of water molecules within the first hydration shell, i.e., G54a7 and OPLS lead to smaller SASA values as a result from the more collapsed structures sampled

with these two FFs. However, the interfacial water area (IWA), which is the ratio of the SASA to the number of waters within the first hydration shell, attains an almost constant value of IWA = 0.096 nm^2 for all peptides and force fields. For globular proteins an IWA value of 0.11 nm^2 has been previously reported [270] suggesting that $A\beta_{16-22}$ has a higher solvent density than the globular proteins.

FF	Variant	SASA (nm^2)	n_{water}	IWA (nm^2)
A99-d	wt	13.52 ± 0.03	140.66 ± 0.24	0.0962 ± 0.0003
	m1	13.21 ± 0.10	137.18 ± 0.83	0.0963 ± 0.0009
	m2	12.68 ± 0.08	132.47 ± 0.67	0.0957 ± 0.0007
C36m	wt	13.37 ± 0.02	138.72 ± 0.10	0.0962 ± 0.0003
	m1	13.19 ± 0.01	136.44 ± 0.05	0.0967 ± 0.0001
	m2	12.63 ± 0.04	131.26 ± 0.30	0.0962 ± 0.0004
C36mW	wt	13.43 ± 0.03	140.98 ± 0.31	0.0953 ± 0.0003
	m1	13.26 ± 0.01	138.81 ± 0.06	0.0955 ± 0.0001
	m2	12.72 ± 0.02	133.76 ± 0.18	0.0951 ± 0.0002
G54a7	wt	12.34 ± 0.09	128.77 ± 0.77	0.0967 ± 0.0009
	m1	12.14 ± 0.04	126.82 ± 0.38	0.0957 ± 0.0004
	m2	11.53 ± 0.02	122.51 ± 0.21	0.0941 ± 0.0002
OPLS	wt	12.27 ± 0.12	127.58 ± 1.08	0.0962 ± 0.0013
	m1	12.14 ± 0.06	126.27 ± 0.51	0.0962 ± 0.0006
	m2	11.72 ± 0.02	122.29 ± 0.17	0.0958 ± 0.0002

Table 4: SASA, number of water molecules, n_{water}, within a cutoff distance of $r_{cut} = 0.45$ nm with respect to the peptide heavy atoms, and the IWA, which is the ratio between the SASA and n_{water}, calculated from the simulations of the monomer of wt, m1, and m2, respectively. Averages and standard errors calculated from three trajectories per system are provided.

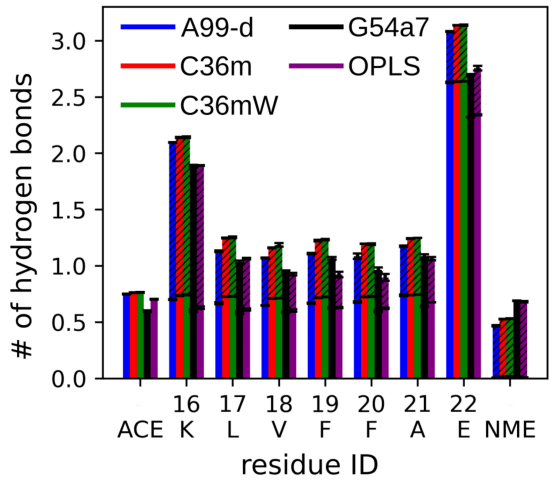

Figure 45: The average number of H-bonds formed between *wt* residues and water. The dashed and non-dashed areas indicate the number of instances where the amino acids act as H-bond donor and acceptor, respectively. Values averaged over three trajectories are shown, along with the standard error (black bars) of the mean. The color key for the FFs is given as part of the plot.

5.3.2.3 *Peptide–water hydrogen bonds*

The analysis of the H-bonds between the peptide and water allows us to assess to what extent specific peptide–water interactions play a role, which serve as a stabilizing factor for the peptide structure and largely influence the peptide dynamics [276, 277]. To this end, we determined the number of H-bonds between peptide and water on a per-residue basis. The results of this analysis are shown for *wt* in Figure 45. Interestingly, the propensity to form H-bonds with water is quite different from the solvation pattern found for the residues. While K16, F19, F20, and E22 are surrounded by the same amount of water molecules, E22 engages in the largest number of H-bonds, followed by K16. These two residues can use both their backbone and side chain for H-bond formation, while this is limited to the backbone for the residues L17–A21. This explains why these residues have similar numbers of H-bonds with water. We further analyzed whether the residues acted as H-bonds acceptor or donor, which is shown as dashed and non-dashed areas, respectively, in Figure 45. The backbone can act as both a donor and an acceptor. Since the side chain of E22 can only act as an acceptor and that of K16 only as a donor, it becomes clear that most H-bonds formed by these residue take place via their side chains. While the different FFs yield the same distribution between H-bond donor and acceptor capabilities for each of the

residues, G54a7 and OPLS lead to a smaller tendency for H-bond formation between peptide and water. This difference is most pronounced for E22, even though OPLS led to the same level of hydration for that residue.

5.3.2.4 *Translational diffusion constants*

The three-dimensional translational diffusion constants for *wt*, m1, and m2 are shown in Table 5, which were calculated from the MSDs (Figure 46). The experimental value for the wild-type $A\beta_{16-22}$ has previously been measured by NMR spectroscopy and has a value of 0.353×10^{-5} cm^2/s [278]. Our results show that none of the FFs can reproduce the experimental three-dimensional translational diffusion constant. A99-d yields considerably smaller diffusion constants than observed in experiment, whereas the other FFs overestimate the diffusion, with C36m leading to the fastest diffusion. However, after correcting the diffusion constants by the viscosity of the solvent models [279], all FFs underestimate the diffusion similarly to A99-d. While there are small differences between the computed diffusion constants for *wt*, m1, and m2, most of them obtained per FF can be considered identical within the standard error. Exceptions are the diffusion constants resulting from the C36mW simulations. However, here no clear trend is observed: with C36m the mutant m2 has the smallest diffusion constant, whereas with C36mW the wild-type peptide has a smaller diffusion constant than its mutants. Since we have no experimental results for the translational diffusion constants for m1 and m2, we are unable to judge the latter findings.

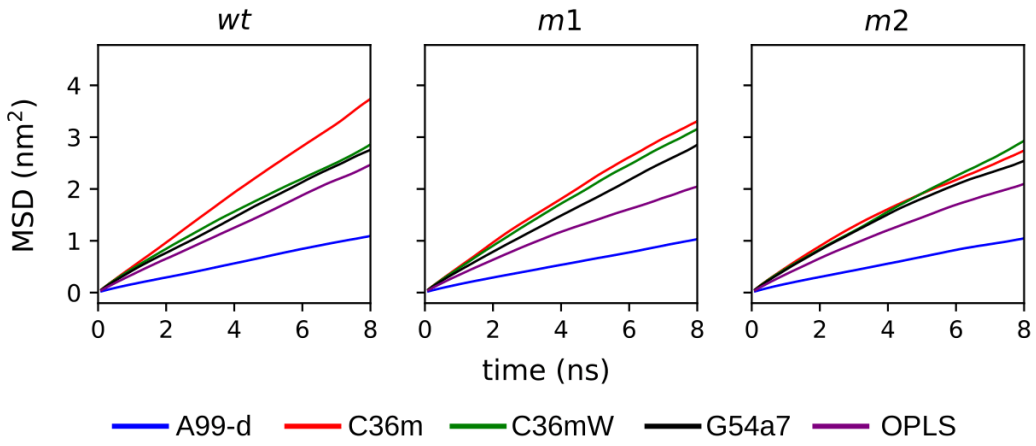

Figure 46: MSD calculated within 10 ns time windows for *wt*, m1, and m2. The results are averaged over three independent trajectories. The slope of each curve is measured from 2 ns to 8 ns to obtain the translational diffusion constants.

	A99-d	C36m	C36mW	G54a7	OPLS
wt	0.21 ± 0.01	0.69 ± 0.04 (0.25±0.02)	0.56 ± 0.01	0.52 ± 0.05 (0.16±0.02)	0.47 ± 0.03 (0.22±0.01)
m1	0.21 ± 0.01	0.66 ± 0.02 (0.24±0.01)	0.61 ± 0.06	0.54 ± 0.05 (0.17±0.02)	0.46 ± 0.04 (0.22±0.02)
m2	0.21 ± 0.01	0.59 ± 0.05 (0.21±0.02)	0.62 ± 0.02	0.53 ± 0.04 (0.16±0.02)	0.43 ± 0.04 (0.20 ±0.02)

Table 5: Three-dimensional translational diffusion constant D_t (mean ± standard error). The values in parentheses are D_t scaled by the correction factor of the solvent modelviscosity (where available). The scaling factors are 2.8 for TIP3P (C36m), 3.27 for SPC (G54a7), and 2.13 for TIP4P(OPLS) [279].

5.3.3 *Characterization of the aggregation process*

5.3.3.1 *Oligomer size*

After studying the conformations of single peptides in solution, we turn our attention to the aggregation properties of $A\beta_{16-22}$ and its mutants represented by the different FFs. We choose to study the aggregation properties by simulating six peptides in a water box. Figure 47 shows the size of the oligomers as a function of time. We observe that the peptides aggregated more quickly attaining the highest oligomer size within

300 ns in the case of G54a7 and OPLS as compared to A99-d, C36m, and C36mW. Despite the intrapeptide interactions observed in the monomer state being of similar strength for A99-d and C36mW, the highest average oligomer size achieved with A99-d was a trimer for *wt* and a dimer for the variants m1 and m2. With C36mW, on the other hand, the highest oligomer size is reached for all peptide systems, albeit for m2 the oligomers are less stable. Both *wt* and m1 aggregate more quickly than m2, while the former two peptides aggregate with a similar speed. C36mW predicts the smallest aggregation tendency for m2. Thus, from the five FFs under study, C36m and C36mW are the only ones which are able to model m2 as slightly less aggregation-prone than *wt* and m1, which is in better agreement with experiments [241]. However, the correct behavior according to experiment would be that m2 does not aggregate at all, while m1 aggregates faster than *wt*. This aggregation ranking cannot be modeled by any of the five FFs and also none of the FFs that were included in our previous benchmark [232]. While G54a7 and OPLS overestimate the aggregation tendency of all three peptides [232], A99-d underestimates it.

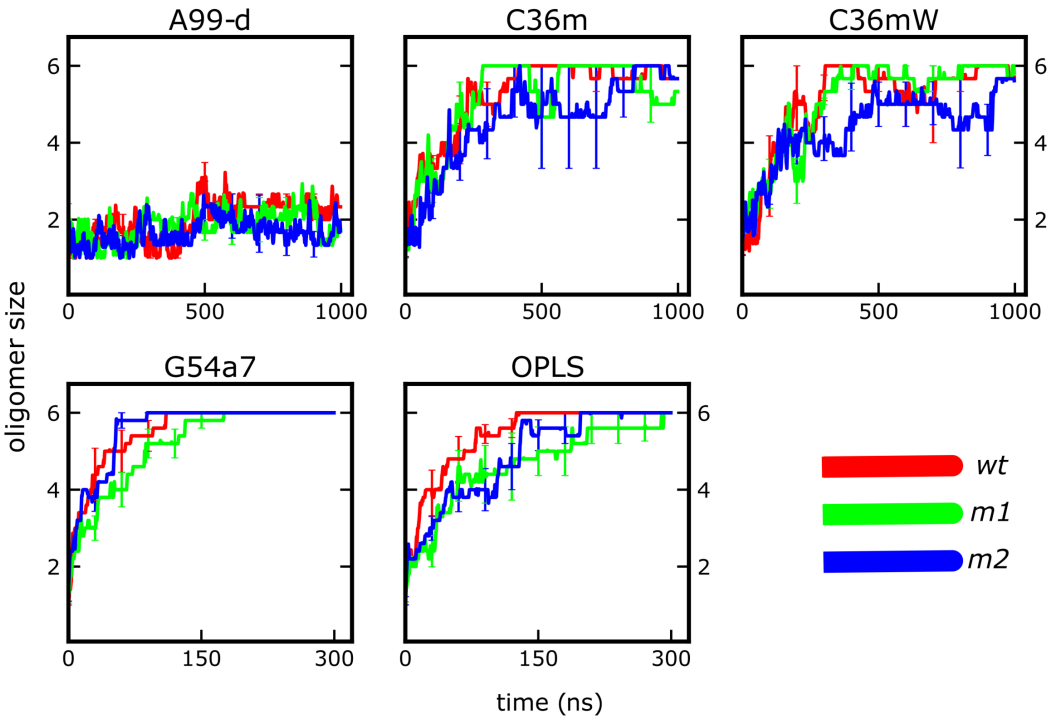

Figure 47: Average oligomer size as a function of time for *wt* (red), m1 (green), and m2 (blue) using A99-d, C36m, C36mW, G54a7, and OPLS. The graphs are averages over three trajectories per system in the case of A99-d and C36mW and five trajectories for G54a7 and OPLS. The different x-axes scales for the top and bottom row are noted. Averages over three (top row) or five (bottom row) independent simulations per system are shown. The error bars indicate the standard error of the mean.

5.3.3.2 *Transition networks and oligomer structures*

For the characterization of the intermediate oligomeric states and the transitions between them, we calculated TNs [3, 248]. As in the preceding study [232], we chose to build the TNs in a two-dimensional space defined by β-strand content (x-axis) and oligomer size (y-axis) (Figure 48). The TNs for A99-d confirm that this FF does not support stable oligomer formation; for all three peptides the most populated state is the extended monomer structure. All oligomer states are only weakly populated with the maximum oligomer size being a pentamer in the case of *wt* and tetramers for m1 and m2. Typical snapshots for oligomers formed with A99-d (Figure 49) show that the peptides in these oligomers are only loosely attached and no β-sheets are formed. As we used the backbone dihedral angles for the definition of the β-state and since with A99-d the peptides have a high preference for extended structures, these oligomers nonetheless populate TN states with high a β-content. In fact, the

three peptides modeled with A99-d only aggregate when being extended, which is understandable as in the collapsed peptide state, K16 and E22 interact with each other within the peptide as seen from the monomer simulations.

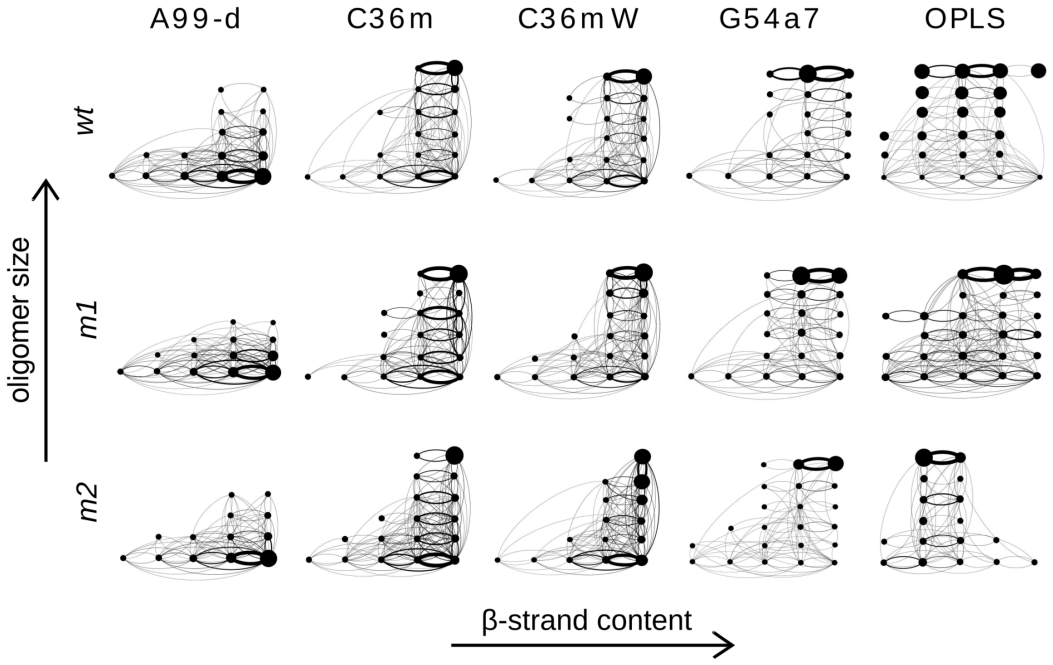

Figure 48: TNs for the aggregation of *wt* (top), m1 (middle), and m2 (bottom) using A99-d, C36m, C36mW, G54a7, and OPLS. The oligomer size (from monomer to hexamer) is given along the vertical axis and the horizontal axis represents the β-strand content (divided into 5 ranges: 0–20%, 20–40% etc.). The size of the nodes is proportional to the population of the state, and the width of the edges is proportional to the mass flux between the states.

The most populated state in the TNs obtained with C36m and C36mW is in each case a hexamer with a high amount of β-sheets. For m2 modeled with C36mW the pentameric state is of almost equal importance, corroborating the finding that with this FF the aggregation is slowed down for this peptide. Typical structures indicate that with C36mW well-developed β-sheets are adopted. They are characterized by a high amount of antiparallel β-sheets resulting from the electrostatic attraction between neighbored K16 and E22 within a sheet. In some cases two sheets with three peptides each form a steric zipper as best seen for *wt* modeled with C36m in Figure 49. Such a double layer can serve as a nucleus for amyloid fibril formation. However, it should be mentioned that the H-bond registry in the β-sheets shown here is mostly not perfect yet. Indeed, previous simulation studies revealed that misregistered β-sheet states are kinetic traps along the amyloid aggregation pathway,

which require sliding motions, so-called "repetition moves" to correct the H-bond pattern in order to reach the perfectly aligned amyloid fibril structure [280–282]. In other cases, such as m2 modeled with C36m barrel formation is observed. With both C36m and C36mW the aggregation pathways involve β-sheet formation from the very beginning, i.e., no disordered aggregates with a low amounts of β-content are formed. This behavior can be understood based on the high tendency of the peptide monomers to adopt extended structures. While the latter also holds true for A99-d, C36m and C36mW provide a better balance between peptide–peptide and peptide–water interactions.

As already seen in the previous study(see ref [232]), G54a7 and OPLS result in the formation of disordered oligomers with lower amounts of β-sheets, which mirrors the behavior of the monomers when modeled with these two FFs. The most extreme case is given by m2 modeled with OPLS. As the representative snapshot of the hexameric state of this system shows, here the peptides mainly adopt a turn conformation as a result of strong intrapeptide K16–E22 attraction (see Figure 42A), inhibiting β-sheet formation in this system. Compared to OPLS, G54a7 provides a better description of the amyloid oligomers formed by $A\beta_{16-22}$ and its mutants, especially for the largest oligomer considered here, which, after it formed, relaxed into β-sheet conformations. However, compared to the structures found for C36mW these β-sheets are less ordered.

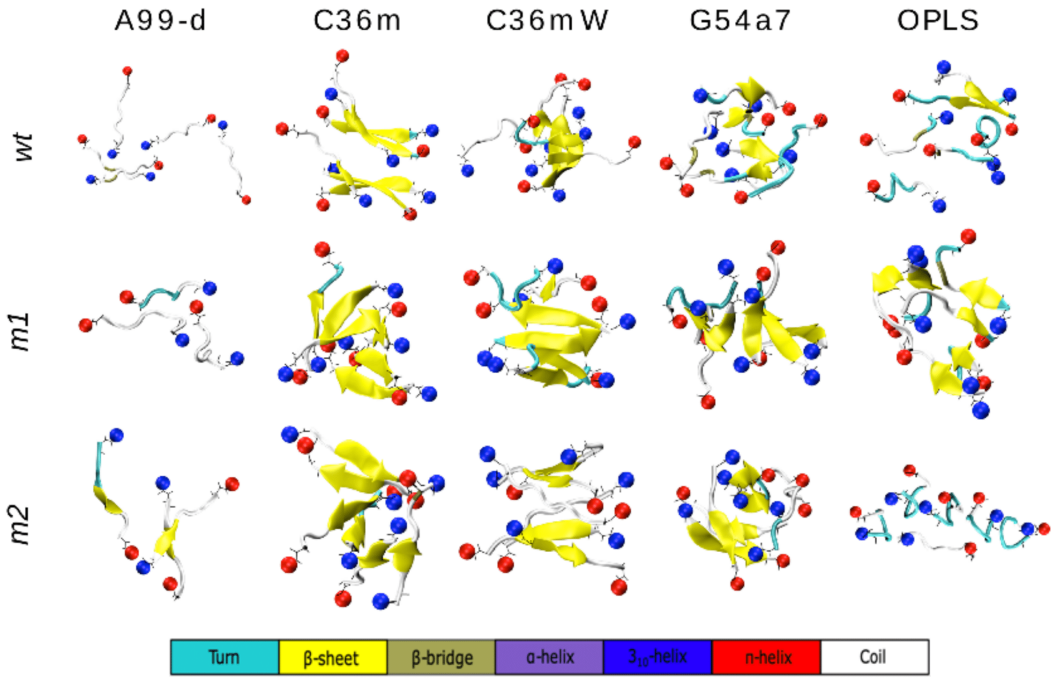

Figure 49: Representative structures of the largest oligomers formed during the MD simulations of oligomer formation of *wt* (top), m1 (middle), and m2 (bottom) modeled with A99-d, C36m, C36mW, G54a7 and OPLS. The spheres colored in red and blue represent the N- and C-terminus, respectively. The assignment of the secondary structure is according to the color key at the bottom.

5.3.3.3 *Interpeptide contacts and interaction energies*

To better understand the driving forces behind the different aggregation patterns observed for the different FFs, we first analyzed the interpeptide contacts present in the oligomers that formed (Figure 50). However, we find that these contacts are too unspecific to allow for an in-depth understanding. Instead, we turned our attention to the nonbonded interaction energies between the peptides on a per residue–residue basis. The resulting energy decomposition is shown for the vdW and electrostatic energy contributions in Figure 51 and Figure 52, respectively, while the sum over all residue–residue interaction energies per system are provided in Table 6.

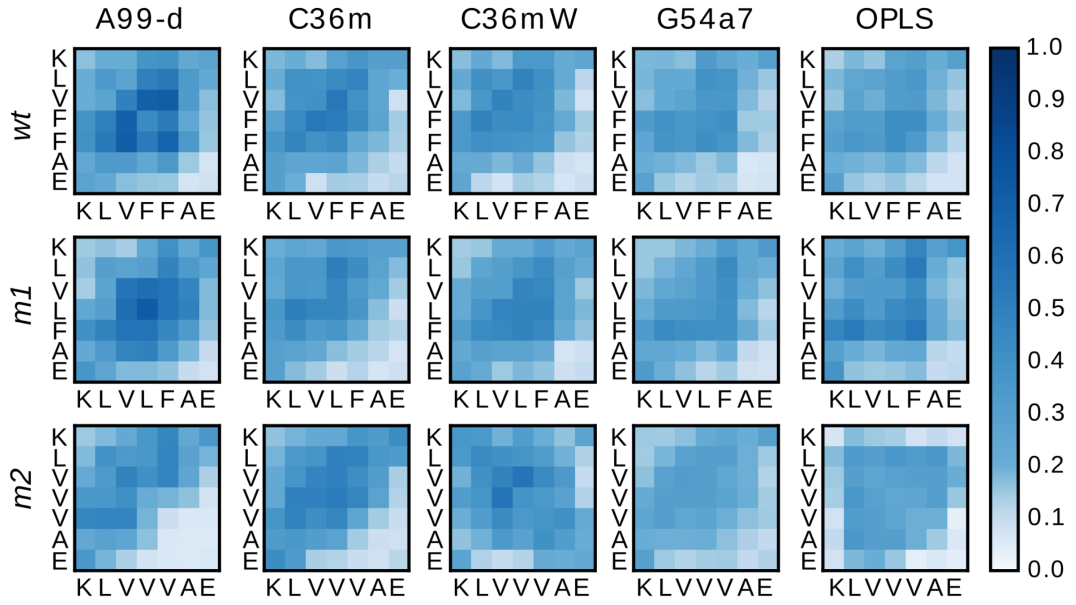

Figure 50: Interpeptide contacts found in the oligomers formed in the simulations of six copies of wt, m1, and m2 using A99-d, C36m, C36mW, G54a7, and OPLS. The color code on the right represents the contact probability between residues averaged over the three simulations per system.

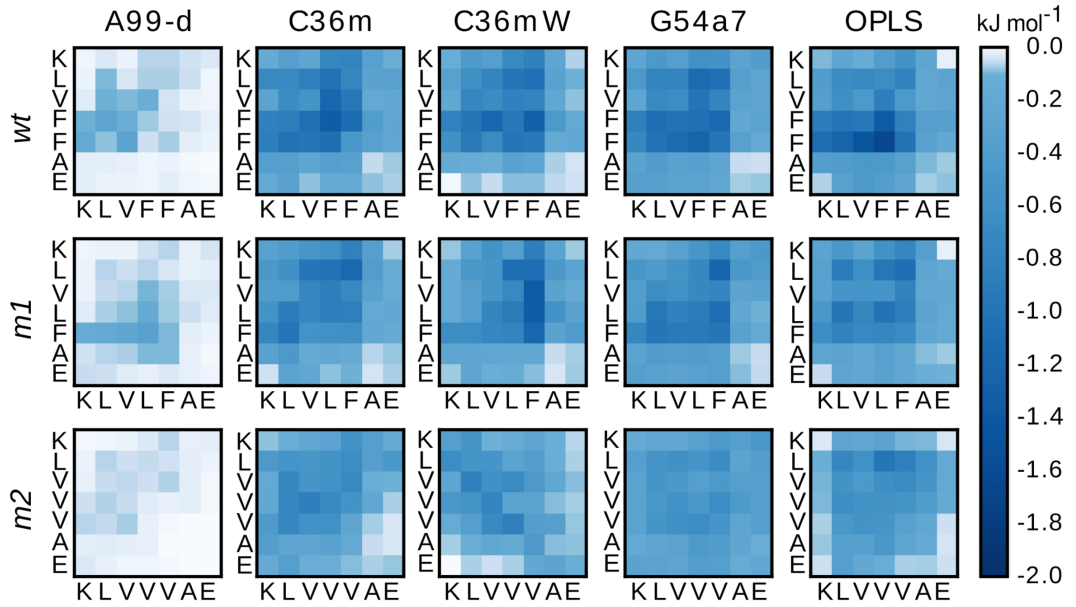

Figure 51: Residue–residue vdW interaction energies between peptides forming oligomers of wt (top), m1 (middle), and m2 (bottom) obtained from simulations using A99-d, C36m, C36mW, G54a7, and OPLS. The interaction energies (in $kJ\,mol^{-1}$) are according to the color key on the right.

For A99-d only weak interaction energies are seen in the interaction maps, and the values in Table 6 show that the level of interaction is an order of magnitude weaker than with the other FFs. This explains why with A99-d the peptides did not aggre-

Peptide	FF	Non-bonded energy term	E (kJ/mol)	$E_{Coul} + E_{vdW}$ (kJ/mol)
wt	**A99-d**	Coul	-4.11 ± 0.29	-9.13
		vdW	-5.02 ± 0.27	
	C36m	Coul	-34.30 ± 3.14	-62.58
		vdW	-28.27 ± 1.04	
	C36mW	Coul	-28.33 ± 3.12	-53.86
		vdW	-25.53 ± 1.37	
	G54a7	Coul	-31.00 ± 1.80	-60.47
		vdW	-29.46 ± 1.05	
	OPLS	Coul	-31.04 ± 2.83	-59.47
		vdW	-28.43 ± 1.07	
m1	**A99-d**	Coul	-4.55 ± 0.29	-9.67
		vdW	-5.13 ± 1.02	
	C36m	Coul	-34.07 ± 3.14	-59.26
		vdW	-25.19 ± 0.98	
	C36mW	coul	-29.42 ± 3.12	-54.35
		vdW	-24.92 ± 1.38	
	G54a7	Coul	-33.96 ± 1.80	-61.37
		vdW	-27.41 ± 0.86	
	OPLS	Coul	-32.60 ± 2.83	-57.46
		vdW	-24.86 ± 1.34	
m2	**A99-d**	Coul	-2.73 ± 0.88	-5.30
		vdW	-2.57 ± 0.81	
	C36m	Coul	-34.25 ± 4.51	-54.65
		vdW	-20.40 ± 2.20	
	C36mW	Coul	-27.13 ± 2.27	-49.41
		vdW	-22.27 ± 1.46	
	G54a7	Coul	-28.21 ± 2.13	-50.75
		vdW	-22.54 ± 1.37	
	OPLS	Coul	-8.05 ± 0.64	-27.55
		vdW	-19.51 ± 0.48	

Table 6: Average electrostatic (Coul) and vdW interaction energies (E), including standard error of the mean, between peptides in oligomers formed by wt, m1, and m2. The sum of both average energy terms is also provided.

gate: the peptide–water interactions are stronger than the interactions between the peptides. With the other FFs strong vdW interactions are observed among residues from the hydrophobic core regions of *wt* and m1, while for m2 they are generally weaker, which is confirmed by the accumulated values in Table 6. For this peptide also the electrostatic interactions are weaker with each of these FFs. In principle, this could translate into a reduced aggregation propensity as seen experimentally [241], which however is not the case, especially not with G54a7 and OPLS. In particular with OPLS the interpeptide interactions for m2 are considerably weaker than those for *wt* and m1, yet the aggregation speed is largely the same for all three peptides. Thus, for OPLS the peptide–water interactions must be too weak so that the peptide–peptide interactions, even when not as pronounced, dominate the aggregation behavior. The analysis of the hydration shell around monomers indeed revealed that solvation is of less importance in G54a7 and OPLS. With C36mW, on the other hand, the increased protein–water interactions result in somewhat weaker interactions between the peptides, yet opposite to A99-d they are still strong enough to enable peptide aggregation.

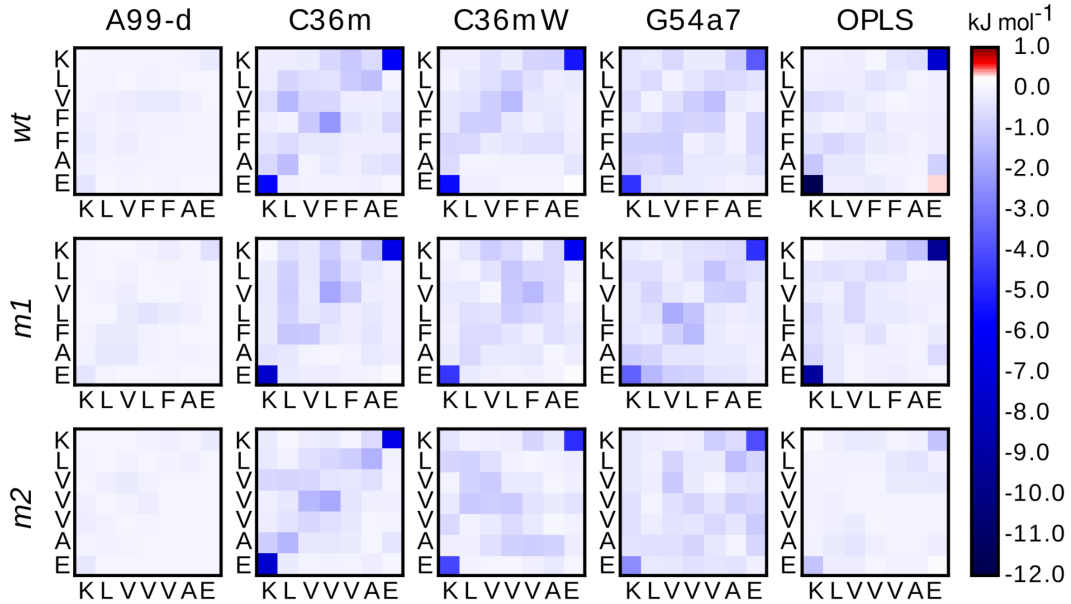

Figure 52: Residue–residue electrostatic interaction energies between peptides forming oligomers of *wt* (top), m1 (middle), and m2 (bottom) obtained from simulations using A99-d, C36m, C36mW, G54a7, and OPLS. The interaction energies (in kJ mol^{-1}) are according to the color key on the right.

In general, all FFs predict very similar total interpeptide interaction energies for *wt* and m1, which explains why both peptides aggregated at similar speeds in the MD simulations. The contribution of the electrostatic interactions to the total interaction potentials is in most cases slightly larger than the vdW contribution. Exceptions are *wt* and m1 modeled with A99-d and m2 simulated with OPLS. The dominant electrostatic interaction is the K16–E22 contact, causing the peptides to arrange themselves in an anti-parallel orientation to each other, which was best seen in the simulations with C36mW. As mentioned above, m2 simulated with OPLS prefers to form a turn structure with intrapeptide K16–E22 salt bridges, which prevents these two residues from interacting with each other across peptides.

5.3.4 *Stability of preformed fibrillar aggregates*

In the previous section, we characterized the self-assembly of six randomly positioned monomers of *wt*, m1, and m2. Here we test whether the various FFs are able to maintain the structure of the aggregation end product, which is a fibril. If this is not possible it follows that one can also not simulate the aggregation into that state. To this end, we set up a minifibril involving two β-sheet layers with six peptides forming an in-register, antiparallel β-sheet in each layer (Figure 40). The stability of this minifibril, also called steric zipper [259], was tested with 3×300-ns MD simulations per system using the recently developed force fields A99-d, C36m, and C36mW. We performed various analyses to test the stability of the simulated minifibril. One of them is the RMSD (Figure S7), which shows that the systems are less stable when simulated with A99-d. The largest RMSD is observed for *wt* modeled with A99-d, which results from one of the twelve peptides leaving the minifibril at $t \approx 200$ ns (Figure 53). However, this occurred during only one of the three simulations for this system. From the RMSD plots for C36m and C36mW one can see that with these two FFs the *wt* minifibril is most stable, while it is least stable for m2.

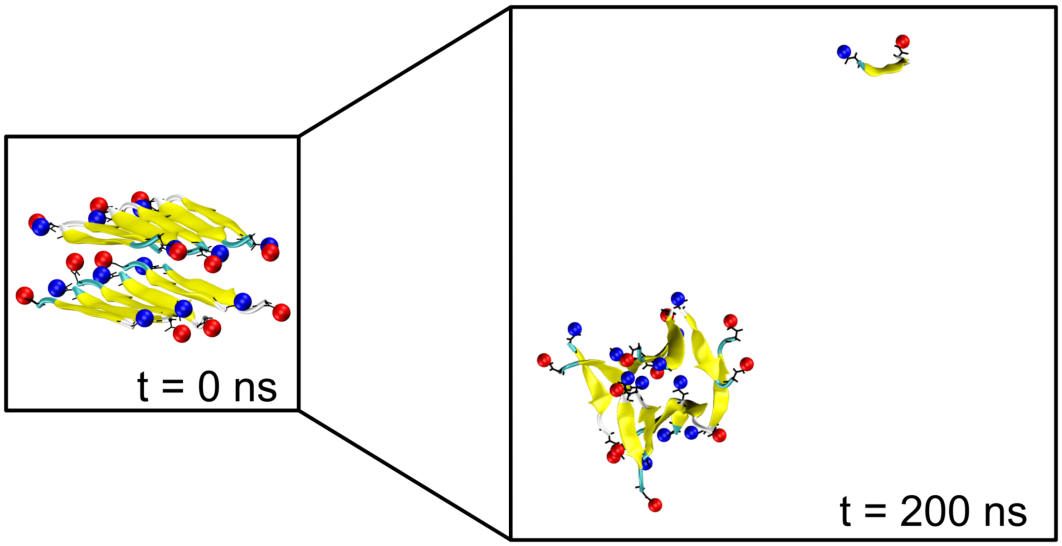

Figure 53: Snapshots from the MD simulation of *wt* using A99-d at t = 0 ns and t = 200 ns. The representative structure at t = 200 ns depicts the detachment of one of the peptides from the fibril.

In order to distinguish whether the RMSD changes arise from a slow disassembly of the fibril structure or from local fluctuations, we calculated the nematic order parameter S_2 defined in equation (28) to describe the orientational order of the peptides with respect to each other (Figure 54). This analysis clearly shows that with A99-d the characteristic arrangement of peptides within a fibril is not supported with A99-d. For all three peptide sequences S_2 drops below 0.5, for m1 and m2 it even reaches values as as low as ≈0.3. The loss of the steric zipper geometry can also be seen in the representative structure shown in Figure 53. With C36m and C36mW the overall fibril structure is generally better retained, especially for *wt* modeled with either FF where $S_2 \approx 0.8$ at the and of the 300-ns simulations. This agrees to the findings of the hexamer simulations, which already revealed steric zipper formation with three peptides per sheet for *wt* (Figure 49). The double mutant m2, on the other hand, appears not to be stable in the fibril state as $S_2 \lesssim 0.6$ at the end of the simulations. The hexamer simulations for m2 revealed a slowed-down aggregation, yet β-sheets were nonetheless formed. The combined picture from both sets of simulations thus suggests that C36m and C36mW model a reticent aggregation of m2 into β-sheets, which however would not develop into a fibril. This only partly agrees with the experimental findings [241], based on which no aggregation at all should be happening. The fibril structure of m1 is found to be stable when modeled with C36mW, yet starts to disintegrate when

simulated with C36m. Thus, C36mW is better suited to reproduce the intricacies of the aggregation of $A\beta_{16-22}$ and its mutants.

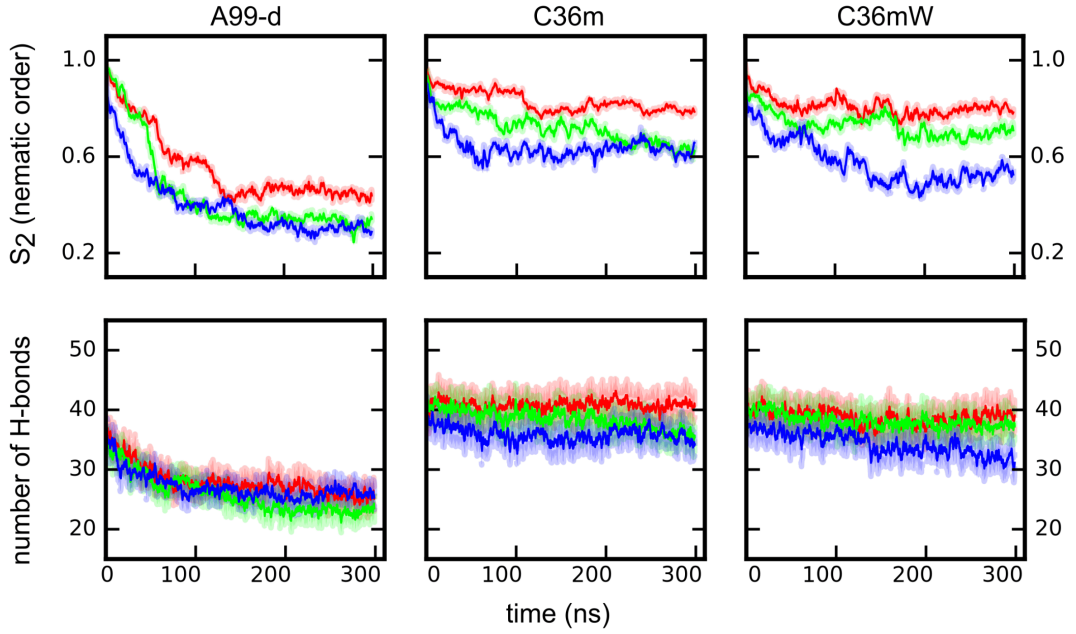

Figure 54: The change in the nematic order parameter (top) and the number of backbone H-bonds (bottom) for the minifibril of *wt* (red), *m1* (green), and *m2* (blue) simulated with A99-d, C36m, and C36mW. The averages over three independent simulations per system are shown. The shaded areas indicate the standard error of the mean.

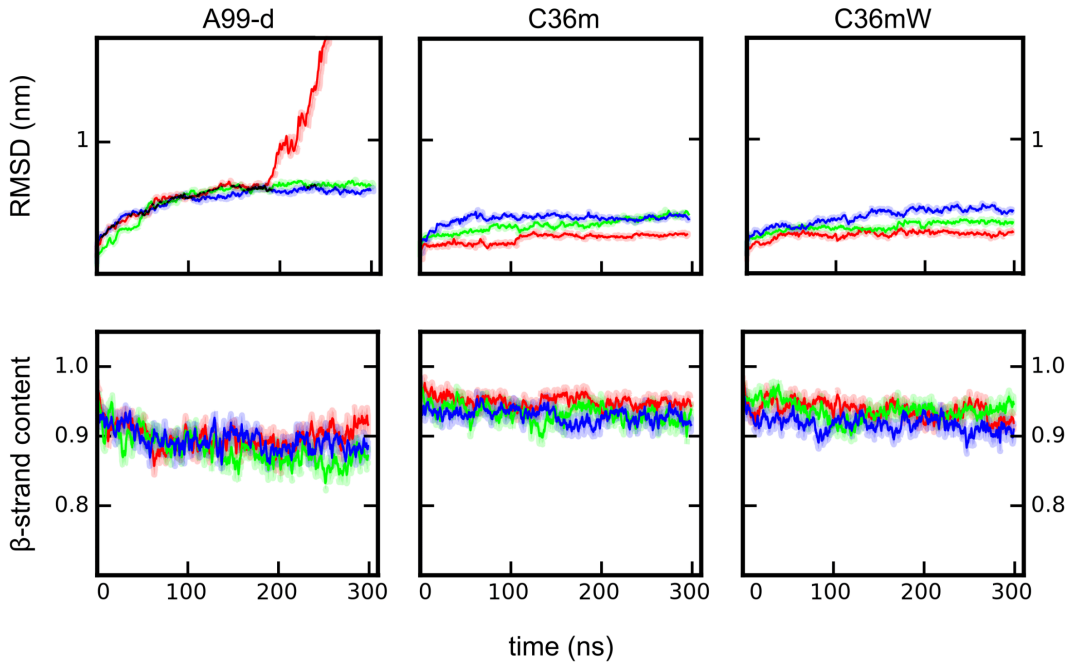

Figure 55: The change in the RMSD (top) and the β-sheet content (bottom) for the minifibril of *wt* (red), *m1* (green), and *m2* (blue) simulated with A99-d, C36m, and C36mW. The averages over three independent simulations per system are shown. The shaded areas indicate the standard error of the mean.

To gain further insight into the origin of the fibril instabilities seen for some of the systems, we determined the β-sheet content as well as the number of H-bonds formed between peptides within the two β-sheets. In Figure 54 the average number of H-bonds per sheet is shown. The maximum number is ≈40, which includes both backbone and side-chain H-bonds. In the simulations with A99-d the number of H-bonds consistently decreased, reaching values between 20 and 30 for all three peptides. The breaking of H-bonds is accompanied by a dissolution of β-sheets, as Figure 55 shows, yet the decline is not as pronounced as for the H-bonds. Taken together, the fibril simulations confirm that with A99-d peptide aggregation cannot be simulated as the peptide–water interactions are considerably more favorable than interpeptide H-bonds. With C36mW, on the other hand, the H-bonds and β-sheets are well conserved for *wt* and with C36mW also for m1, while for m2 the number of H-bonds has dropped to 30–35 at the end of the 300-ns MD simulations. The count of H-bonds for m1 modeled with C36m remained constant until 200 ns after which it started to decline somewhat. Thus, this analysis confirms that from the FFs considered C36mW is best suited for modeling *wt*, m1, and m2, followed by C36m.

5.4 CONCLUSIONS

In this chapter, we examined the applicability of current force fields (FFs) developed for IDPs, namely AMBER99SB-disp [108] (A99-d), CHARMM36m [185] (C36m), and C36m with increased protein–water interactions [185] (C36mW), for studying amyloid aggregation with MD simulations using the $A\beta_{16-22}$ peptide and its two mutants F19L and F19V/F20V (denoted *wt*, m1, and m2 here) as test case. Based on experimental results the order of the aggregation speed should be $m1 > wt \gg m2 \approx 0$ [241]. As amyloid aggregation results from an interplay of monomeric peptide properties as well as the stability of the intermediate oligomers and the final aggregation product, which are amyloid fibrils, we employed a step-wise approach and investigated each of these aspects separately. In order to gain an in-depth understanding of the performance of the FFs we dissected the intra- and interpeptide interaction energies of the monomers and oligomers, and included in our analysis the results of the

simulations obtained with GROMOS54a7 (G54a7) and OPLS-AA (OPLS). For these two FFs we had already demonstrated that they are not suitable for the problem under study, because they overestimate the aggregation process and cannot discriminate between the different aggregation propensities of *wt*, m1, and m2 [232]. By including them in the analysis we wished to learn the origin of their failure, knowledge that in future can be used for FF reparameterization. For the IDP-corrected FFs we can conclude that A99-d is not applicable to the process of amyloid aggregation as this FF does not lead to stable aggregates being formed, whereas C36mW give rise to proper β-sheet formation with the aggregation speed order $wt \gtrsim m1 > m2$. In the following we discuss how these results are affected by the properties at the monomeric and oligomeric level as well as how the components of the FFs give rise to their different behavior.

5.4.0.1 *Different peptide structures with different force fields*

Given the quite small size of the peptides under study, one would expect that the different FFs should not lead to large differences in the peptide structures and dynamics sampled in the MD simulations. However, Figs. 41 and 43 show that this is not the case. The older FFs, G54a7 and OPLS, lead to a considerably larger structural diversity than the IDP-corrected FFs. Nonetheless, this high flexibility at the monomeric level does not hinder the peptides from rapid self-association as can be seen in Figure 47. The larger structural diversity with G54a7 and OPLS is also visible at the oligomeric level, where especially for the smaller oligomers structures other than β-sheets are also sampled. This can be deduced from the TNs in Figure 48. In the simulations with A99-d and C36mW, on the other hand, extended structures prevail both at the monomeric and oligomeric level. This leads to ordered aggregation into β-sheets from the very beginning in the case of C36mW, whereas with A99-d only encounter complexes that promptly dissociate again are formed.

5.4.0.2 *Peptide–water interactions as key determinant for peptide structure and aggregation*

Our analysis of the intra- and interpeptide interaction energies as well as the interactions between peptide and water revealed that the preference of $A\beta_{16-22}$ and its mu-

tants to adopt collapsed structures at the monomeric level and disordered oligomer structures when modeled with G54a7 and OPLS, but extended structures when modeled with A99-d and C36mW is mostly due to the different amounts of peptide–water interactions. Fewer H-bonds between the peptides and water are formed in the simulations with G54a7 and OPLS, which can be seen in Figure 45. This allows for more intra- and interpeptide interactions being formed, which explains both the larger structural diversity as well as the faster aggregation. The IDP-corrected force fields, on the other hand, lead to extended structures as here especially the terminal residues K16 and E22 have a high preference to interact with water via H-bond formation, inhibiting their interaction with each other which would lead to a collapsed peptide structure. However, in the case of A99-d the peptide–water interactions are too much increased, leading to reduced interpeptide interactions and thus inhibiting aggregation. A better balance between peptide–water and peptide–peptide interactions is reached with the other FF with increased vdW interactions between protein and water, i.e., C36mW. While for the remaining FFs the overall strength of the residue–residue interactions is similar, specific interactions nonetheless play a role in the structural preferences at the monomer and oligomer level. For instance with OPLS the electrostatic attraction between K16 and E22 is very strong, causing turn structures in the monomers and fast, electrostatic-driven aggregation.

5.4.0.3 *Too strong peptide–water interactions with A99-d*

In the case of A99-d not only the number, but also the strength of the H-bonds between peptide and water is increased. This can be deduced from the much smaller translational diffusion constant obtained with this FF compared to the four other ones (Table 5). We conclude that this is the main source of the decreased aggregation propensity observed with A99-d. A similar behavior was seen in the simulations with AMBER03WS in a previous study(see ref [232]), a FF where the short-range protein–water pair interactions were increased by a factor of 1.1, while all the water–water and protein–protein parameters were left unchanged [139]. However, it should be mentioned that also the peptide conformation may play a role in the simulations with A99-d. While the peptides prefer extended structures, these are in fact to a large

extent poly-proline II (PPII) conformations. This shortcoming of A99-d was already revealed in our previous chapter [268] and confirmed for $A\beta_{16-22}$ here. It has been proposed that PPII might prevent the formation of β-sheet [283], due to its particular orientation of the amide bonds [284]. Thus, the problems of A99-d are twofold and involve both backbone and protein–water interaction parameters.

5.4.0.4 *Best performance for C36mW*

A generally better performance was obtained with C36m and C36mW. The differences between their results are subtle. While the numbers of H-bonds between peptides and water are identical, with C36mW they are somewhat stronger – judged by the translational diffusion – as a result of increasing the protein–water interactions [185]. This parameterization change slightly increased the solvation of m2 modeled with C36mW compared to C36m (Figure 44), which leads to a decreased aggregation propensity as both the hexamer and the minifibril simulations showed. Thus, only small adaptations of of the protein–water interactions as done in C36mW [185] can make a positive difference. Nonetheless, the results obtained with C36 and C36mW are rather similar, allowing to conclude that C36m is in general suitable for modeling peptide aggregation. This is partly due to the refined backbone torsion parameters leading to more extended peptide structures [185], but also seems to be a general feature of CHARMM force fields since in the previous study C22* was already identified as the best FF for modeling the aggregation of *wt*, m1, and m2. Also in our benchmark of $A\beta_{1-40}$ the performances of C36m and C22* were quite similar [268].

6

STRUCTURAL HETEROGENEITY OF GLYCOSAMINOGLYCANS

Parts of this chapter are submitted by the author to the *International Journal of Molecular Sciences* as Samantray, S., Olubiyi, O.O., Strodel, B. (2021). The influences of sulphation, salt type and salt concentration on the structural heterogeneity of glycosaminoglycans.

6.1 MOTIVATION

The ubiquitous presence of GAGs in the extracellular matrix [285] as much as on cell surfaces points to a primary role in cellular communication and adhesion and other biological processes dependent on these two, including cell recognition of growth factors, chemokines, adhesion molecules and enzymes [57, 286, 287]. While describing the full interactome of GAGs is an ongoing venture, other molecular systems they interact with include serpins involved in anticoagulation; cytokines involved in immune regulation, inflammation and blood development; matrix proteins for structural stability; protein receptors in the control of receptor dimerization; lipoproteins involved in lipid translocation; plaque proteins implicated in neurodegeneration; as well as complement structures in viral pathogens needed for entry into the host cell [288].

An inspection of the GAG interactome indicates their involvement in more than just physiological processes; they have been found to play a role in various pathological conditions. For instance, their interaction with amyloidogenic proteins implicates them as participants in the pathogenesis of amyloid-related neurodegenerative disorders, such as AD, type-2 diabetes, PD and prion diseases [289, 290]. Their involvement

in several biomolecular processes has carved out a niche for GAGs in clinical therapeutics. For example, chondroitin and hyaluronate are used in different pharmaceutical and nutraceutical formulations for the treatment of osteoarthritis [291–294], Hep and its mimetics are used as anticoagulants for treating thrombosis, while KS is incorporated in ophthalmic products for treating certain eye defects and as biomarker for female genital tract carcinoma [295]. Beyond their use in different capacities as replacement therapy, recent studies have suggested additional roles as possible therapeutic agents in inhibiting cell invasion by viruses causing severe acute respiratory syndrome (SARS) diseases [296, 297].

It is crucial to point out that in all their areas of relevance – physiological, pathological or therapeutics –, the participation of GAGs results principally from their unique physicochemical and structural features, including high negative charge, high viscosity and lubricative attributes, unbranched polysaccharide structures, low compressibility, as well as the ability to attract and imbibe large amounts of water. Because GAGs are highly heterogenous, they pose great challenges to research investigation. First variable building units are involved in assembling the polymers whose final lengths can vary widely. GAGs are also differentially sulphated, an event that is influenced by a host of factors including age, physiological conditions, and diseases. Their structural complexity is further extended by attachment to sugar and protein units.

GAGs behave as a hydrophilic polyelectrolyte whereby they form swollen random coils in aqueous environment [298–300]. However, such structural preference of GAGs is modified under the influence of salts differing in type and concentrations. For example, the ion concentrations in solution influence the size and shape of hyaluronic acid (HA) random coils, allowing site-specific HA-ion interactions via the decay of hairpin-like loops [301]. The ionic interactions between the cations and specific sites on the GAG surface are influenced by sulphation of the monosaccharide units of GAGs, such as glucosamine in heparan-6-sulphate (H6S) and galactosamine in chondroitin-4-sulphate (C4S) and chondroitin-6-sulphate (C6S) [302, 303]. The sulphation pattern encoded in the GAG sequences acts as a molecular recognition tool to identify growth factors and allow activation of connected signalling pathways. Apart

from the non-specific electrostatic interactions in complex environments, GAGs also display ion-specific interactions by behaving as a biochemical barrier permeable only to monovalent and divalent ions [304], such as Na^{+}, K^{+}, Ca^{2+}, and Mg^{2+} [305, 306]. A detailed understanding of the sulphate moieties in controlling their intra- and intermolecular interactions as well as their interactions with cations would be key for designing GAG-based therapeutic agents to induce their binding affinity to other bio-macromolecules. Taken together, arriving at a common structural description for a biomolecular system so complex, or even determining their conformational preferences by traditional structure determination methods has been nothing short of daunting.

To characterise the GAGs behaviour in biological processes, it is paramount to develop an understanding of which specific structural patterns they adopt under different physiological conditions and how they do so [307, 308]. In this chapter, we provide crucial insight into the structure and dynamics of various GAGs using MD simulations. Compared to other techniques, MD simulations are an attractive and relatively inexpensive tool for exploring conformational transitions in GAGs at microsecond time scale and with atomic resolution [309–312]. Simulations have been used extensively to study the conformational plasticity of different GAGs including HA [313], chondroitin sulphate (CS) [314], and heparan sulphate (HS) [315]. The structural diversity of GAGs coupled with GAG-protein binding interactions using multiscale modelling approaches is highlighted in recent reviews [316, 317]. By modeling various physiological conditions, we are able to present insight into how changes in these conditions affect the structures and dynamics of HA, H6S, C4S and C6S. A heuristic method is provided to predict the free energy landscape characterising the conformational ensembles of the GAGs.

6.2 MODELS AND SIMULATION DETAILS

6.2.1 *GAG models*

To understand the effects of different salts and salt concentrations on the structure and dynamics of different GAGs, we simulated the following five repeating disaccharide units: HA as a nonsulphated GAG and H6S, C4S, and C6S as sulphated GAGs (Figure 3 on page 13). The repeating disaccharide units for HA involve D-glucosamine (GlcNAc) and D-glucoronic acid (GlcUA) linked via alternating β-(1→4) and β-(1→3) glycosidic bonds (O-linked): -GlcNAc-β(1→4)-GlcUA-β(1→3)-. In H6S, the disaccharide building blocks are identical to those in HA apart from sulphation at position 6 of GlcNAc, and glycosidic linkages with different geometries: -GlcNAc(6S)-α(1→4)-GlcUA-β(1→4)-. C4S and C6S are composed of similar components and glycosidic linkage patterns as HA, except for GlcNAc being replaced by GalNAc. The difference between C4S and C6S exists in the sulphation points on the D-galactosamine (GalNAc) unit, which is at carbon 4 in C4S and position 6 of GalNAc in C6S, leading to -GalNAc(4S)-β(1→4)-GlcUA-β(1→3)- and -β(1→4)-GlcUA-β(1→3)-GalNAc(6S)-, respectively. The disaccharide units and linkages (referred to as $\mathtt{Linkage}_1$ and $\mathtt{Linkage}_2$) are summarized in Table 7.

System	$\mathtt{Linkage}_1$	$\mathtt{Linkage}_2$
HA	(-GlcNAc-β(1→4)-GlcUA-)	(-GlcUA-β(1→3)-GlcNAc-)
H6S	(-GlcNAc(6S)-α(1→4)-GlcUA-)	(-GlcUA-β(1→4)-GlcNAc(6S)-)
C4S	(-GalNAc(4S)-β(1→4)-GlcUA-)	(-GlcUA-β(1→3)-GalNAc(4S)-)
C6S	(-GalNAc(6S)-β(1→4)-GlcUA-)	(-GlcUA-β(1→3)-GalNAc(6S)-)

Table 7: The glycosidic linkage models of the GAG systems.

6.2.2 *Simulation protocols*

We simulated the penta-disaccharide GAG units by placing their extended structures in the simulation box (Figure 56), making sure that the minimum distance between the corresponding GAG structure and any edge or face of the cubic box is at least 1.2 nm. In modelling the GAG molecules, we employed the all-atom force field C36 [318–321], using the CHARMM parameters as available through the 'Glycan Reader & Modeler' module [322–324] of the CHARMM-GUI web server [325]. The resulting simulation box had an edge length of 6.8 nm, which was solvated with water using the TIP3P model [254] and resulted in about 32,000 atoms per system. We designed a set of four simulations for each GAG by employing different salts and salt concentrations. In the first system, Na^+, but no Cl^-, was added to neutralize the system. This system is denoted to have 0 mM NaCl. In the second system, a concentration of 150 mM NaCl was added. Two corresponding KCl systems were created and denoted 0 mM and 150 mM KCl, respectively. The number of ions required to neutralize each GAG system corresponding to their salt concentration and total number of atoms per system are listed in Table 8. Each system was minimized using the steepest descent algorithm, followed by equilibration, first with a 20 ps run in the NVT ensemble with position restraints on the non-hydrogen atoms of the GAGs, afterwards with a 20 ps run in the NpT ensemble without position restraints. For the production runs, we simulated the GAG systems for 1 μs in the NpT ensemble at $T = 300$ K and $p = 1$ bar. Throughout the simulations, we constrained all bonds involving hydrogen atoms using the LINCS algorithm [191]. The electrostatic interactions were calculated using the PME method [192] and their real-space components were truncated at 1.2 nm. The same cutoff was applied for the calculation of the van der Waals interactions. The temperature and pressure were controlled using a Nose-Hoover algorithm with a 1 ps time constant for coupling and a Parrinello-Rahman barostat [190] with a relaxation time of 5 ps, respectively. A time step of 2 fs was used for the integration of the equations of motion for all systems. All MD simulations were realized with GROMACS version 2018.3 [157, 326, 327]. For the analysis of the simulation trajectories, we employed a combination of standard GROMACS tools,

VMD [195] and in-house Python scripts [5] invoking the MDAnalysis [196], and MDTraj [328] libraries. The workflow to perform a dimensionality reduction, clustering based on internal GAG coordinates, and calculation of the free energy profile of the GAGs is available at https://github.com/suman-samantray/GAG-clustering-FES.

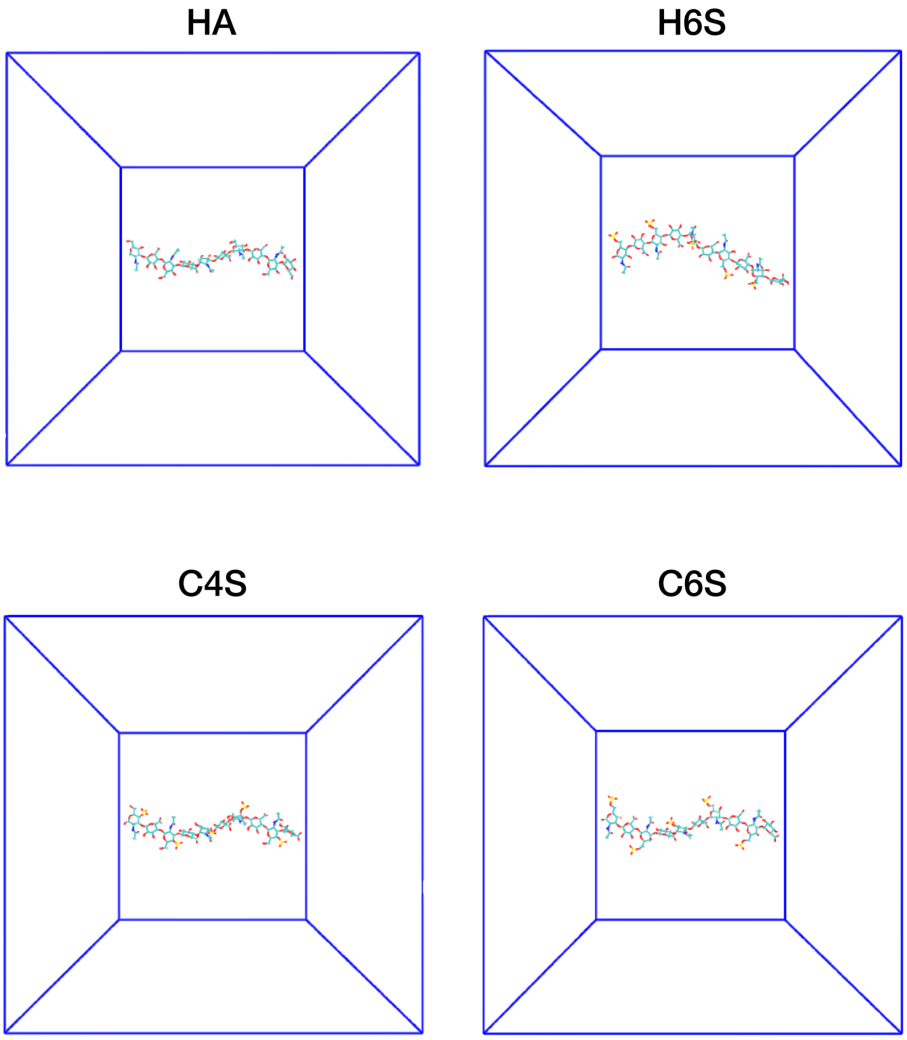

Figure 56: Initial GAG configurations used in the MD simulations.

System	0 mM NaCl	150 mM NaCl	0 mM KCl	150 mM KCl
HA	5 Na^+, 0 Cl^-	35 Na^+, 30 Cl^-	5 K^+, 0 Cl^-	35 K^+, 30 Cl^-
H6S	10 Na^+, 0 Cl^-	40 Na^+, 30 Cl^-	10 K^+, 0 Cl^-	40 K^+, 30 Cl^-
C4S	10 Na^+, 0 Cl^-	40 Na^+, 30 Cl^-	10 K^+, 0 Cl^-	40 K^+, 30 Cl^-
C6S	10 Na^+, 0 Cl^-	40 Na^+, 30 Cl^-	10 K^+, 0 Cl^-	40 K^+, 30 Cl^-

Table 8: Overview of the simulated systems

6.2.3 *Conformational analysis*

To quantify the flexibility of the GAGs, we calculated the root mean square deviation (RMSD) by fitting the whole GAG structure to the initial configurations. To identify the sampled conformations and similar structures within selected RMSD cutoffs, we performed cluster analyses using the GROMOS clustering algorithm from Daura et al.[156]. Cutoff values of 0.3 nm, 0.4 nm, and 0.5 nm were tested. For both RMSD calculation and clustering, all GAG atoms were included in the computations. We also calculated contact formation between the GAGs monosaccharide units. In our analysis, two GAG residues are assumed to be in contact if the distance between any pair of atoms from residue *a* and residue *b* is 0.4 nm or less. The end-to-end distance (R_{ee}) of the GAGs was determined by calculating the distance between the C_4 atom of the starting residue and the C_1 atom of the terminal residue (see Figure 57 for the atom numbering). To characterise the torsional motions of the GAG chains, we calculated the dihedral angles ϕ and ψ across the glycosidic linkage bonds. Their definitions are shown in Figure 57. The ϕ and ψ dihedral angles for each linkage type are combined into a single dihedral offset function [329] (D_{offset}),

$$D_{offset} = \frac{1}{N}\sum_{i=1}^{N}(1+\cos(\phi_i-\phi_{ref}))+(1+\cos(\psi_i-\psi_{ref})) \tag{30}$$

where ϕ_{ref} and ψ_{ref} are the reference dihedral angles as present in the initial, fully extended GAG structures. The number N refers to the number of linkages found in each GAG, which is $N = 5$ for $\mathtt{Linkage}_1$ and $N = 4$ for $\mathtt{Linkage}_2$ for a GAG with five disaccharide units.

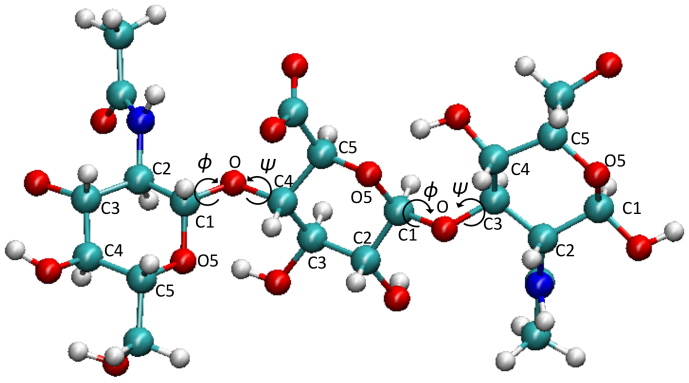

Figure 57: The definition of the dihedral angles ϕ and ψ is shown for HA.

To get a better understanding of the structural preferences of the GAG chains, we further calculated the number of hydrogen bonds (H-bonds) formed between the GAG residues and water molecules (N_{HB}). Here, a hydrogen bond was assumed to be present when the donor–acceptor distance is less than 0.35 nm and the donor-H-acceptor angle is less than 30°. We also characterised the interactions between the GAGs and the cations by calculating the radial distribution functions (RDFs) between Na^+/K^+ and the sulphate oxygen atoms (in GlcNAc of H6S and GalNAc of C4S and C6S) as well as the carboxylate oxygen atoms (in GlcUA of all GAGs). This was done for all five sulphate and five carboxylate groups present in the GAGs after which the five corresponding profiles obtained were averaged.

6.2.4 *Free-energy landscape*

In our effort to describe the collective modes of structural fluctuations observed in GAGs, we performed PCA on the previously calculated observables, specifically the D_{offset} for $\mathtt{Linkage}_1$ and $\mathtt{Linkage}_2$, R_{ee}, and N_{HB}. Based on the first two principal components (PCs), which capture the majority of the variance in our MD data, we

built a 2D free energy landscape of the GAGs conformational space. The free energy was calculated by computing the probability distribution along the PCs, $P(PC)$,

$$\Delta G(PC) = -k_B T \ln\left(\frac{P(PC)}{P_{max}}\right) \tag{31}$$

where k_B is the Boltzmann constant and P_{max} denotes the maximum probability along the selected PC. The segregation of the conformational space into distinguishable cluster states projected onto the first two PCs is accomplished using k-means clustering. To determine the ideal number of clusters, we measured the sum of the squared distances to the nearest cluster center, i.e. the inertia, plotted the inertia for increasing cluster numbers and applied the elbow method to the resulting curve.

6.3 RESULTS AND DISCUSSIONS

6.3.1 *Characterisation of the GAGs structural data*

Polymeric GAG chains in solution can assume a number of configurations that contribute to their entropy. For elucidation of the energetically stable structures of the GAGs, it is important to identify and segregate their conformational ensembles under diverse physiological conditions. In this section, we assess the role of salt ions in inducing conformational heterogeneity in the different GAGs and subsequently determine critical collective variables in order to describe the structural transitions.

6.3.1.1 *RMSD-based conformational clustering*

To determine the flexibility of the GAGs at different salt concentrations, we studied the time evolution of the RMSD of the GAGs (Figure 58). With obtained RMSD variations in the range of 0.4–0.6 nm and regardless of salt type and concentration, we found the GAGs not to be very flexible. To further characterise the structural heterogeneity and also to assess sampling convergence, we performed RMSD-based geometric clustering using different cutoff values: $RMSD_{cutoff}$ = 0.3, 0.4 and 0.5 nm (Figure 60, Figure 59 and Figure 61, respectively). These results for 0.4 nm cutoff

show H6S with the highest number of clusters and as a result the most flexible GAG, followed by C6S, C4S and HA (in this order). It is interesting to note that this trend is unaffected by salt type and concentration. HA, the least flexible of the lot, showed only limited dynamics and stabilises quickly without much sampling of unique configurations. This behaviour contrasts with that of the sulphated GAGs. Sulphation at carbon 6, compared to carbon 4, generally increases the GAGs flexibility but makes H6S and C6S more susceptible to a higher salt concentration. At higher salt concentrations, the flexibility of H6S and C6S was reduced compared to their behaviour at 0 mM salt concentration.

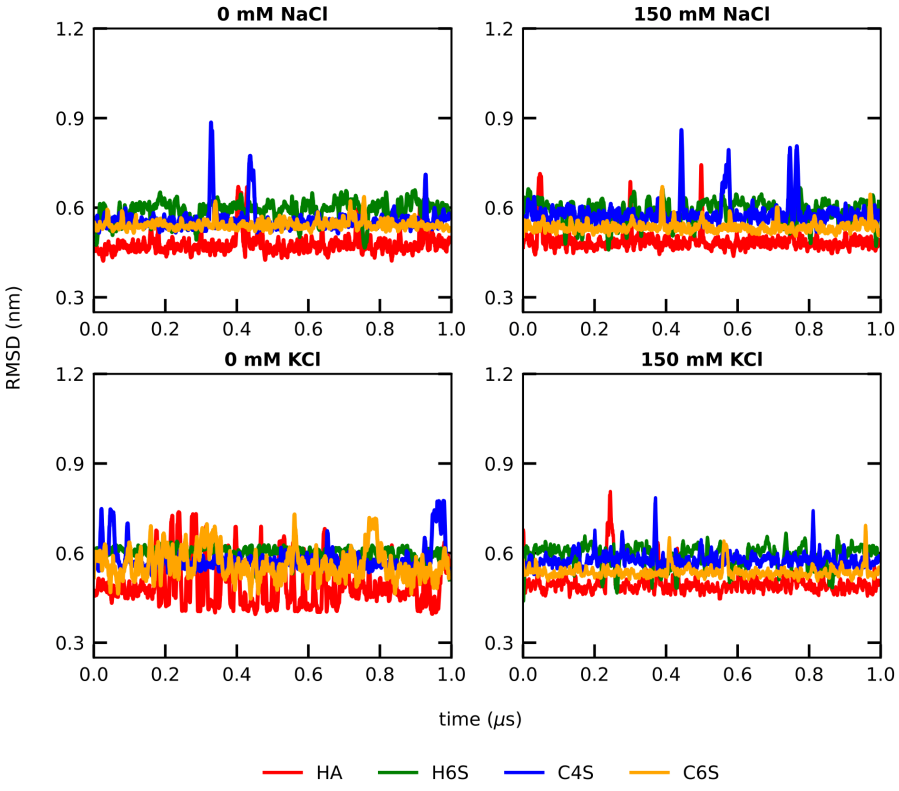

Figure 58: Evolution of the RMSD of the GAGs at 0 mM (left) and 150 mM (right) salt concentrations (see color key at the bottom).

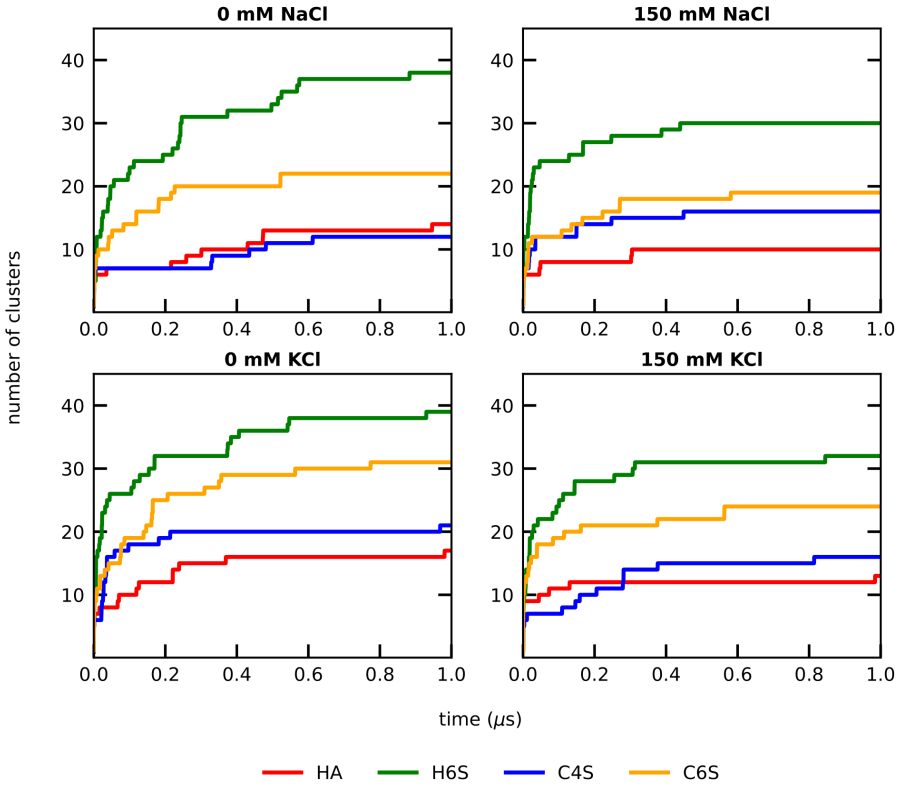

Figure 59: Evolution of the number of conformational clusters at $RMSD_{cutoff}$ = 0.4 nm for the different GAGs (see color key at the bottom) at 0 mM (left) and 150 mM (right) salt concentrations.

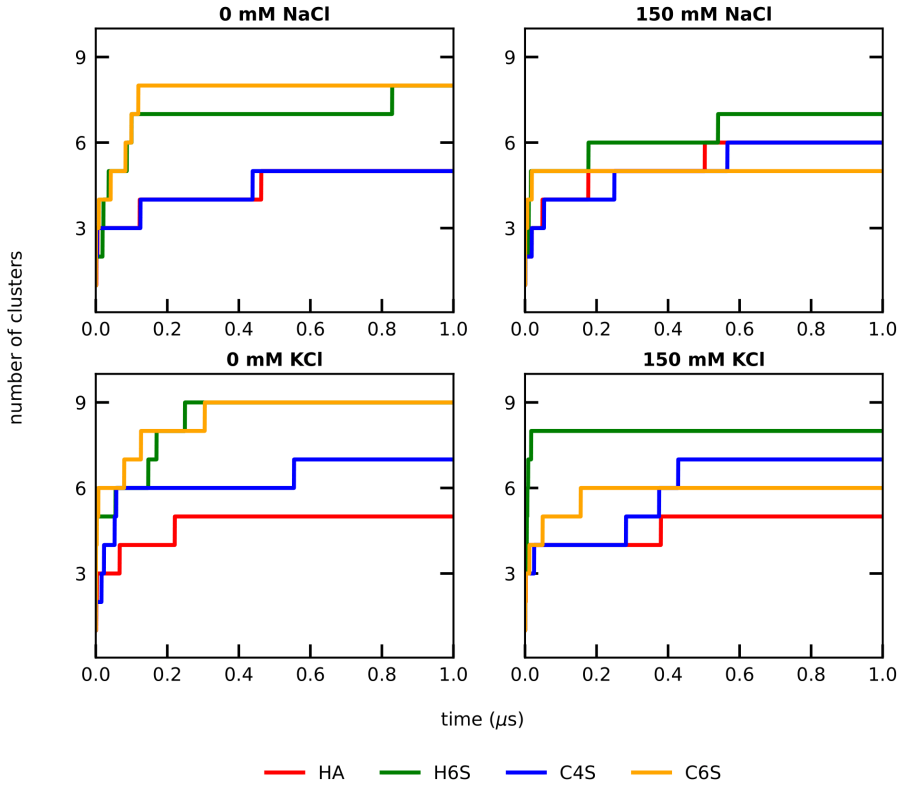

Figure 60: Evolution of the number of conformational clusters at $RMSD_{cutoff}$ = 0.3 nm for the different GAGs (see color key at the bottom) at 0 mM (left) and 150 mM (right) salt concentrations.

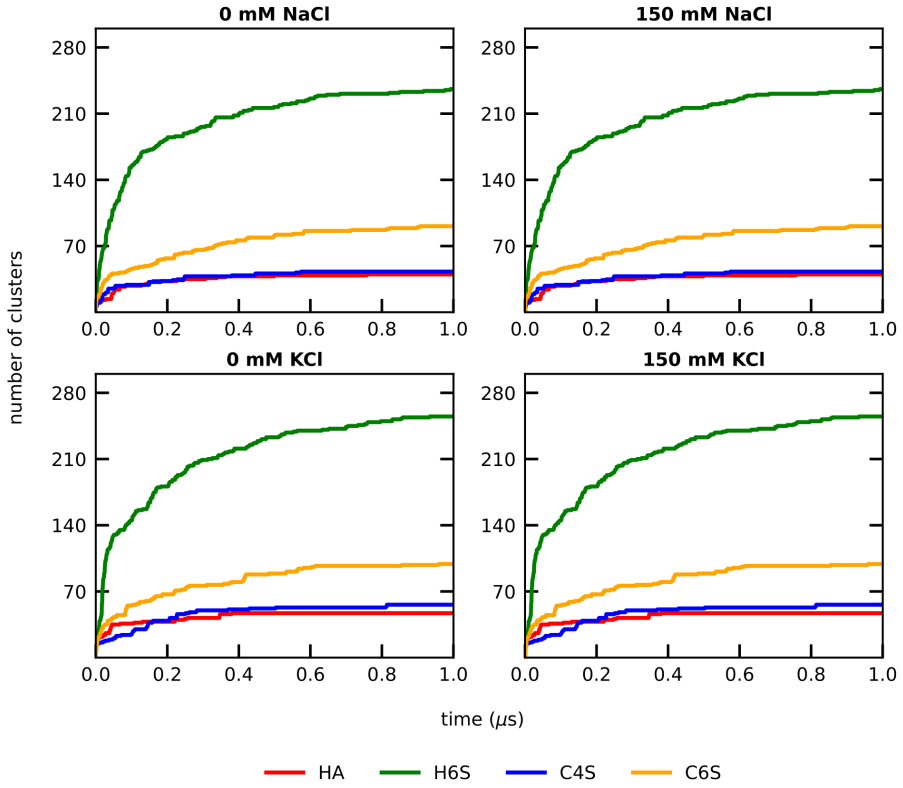

Figure 61: Evolution of the number of conformational clusters at $RMSD_{cutoff}$ = 0.5 nm for the different GAGs (see color key at the bottom) at 0 mM (left) and 150 mM (right) salt concentrations.

The clustering data at $RMSD_{cutoff}$ = 0.4 nm further show that the simulations had reached convergence. For a cutoff value of 0.3 nm, the number of clusters for the H6S system is rather high (Figure 60), as already minor structural changes are recorded here. This indicates that this cutoff value is too small to assess sampling convergence and structural diversity. On increasing the $RMSD_{cutoff}$ to 0.5 nm (Figure **??**, the number of clusters observed is greatly reduced, blanketing the diverse conformational classification of the GAGs.

6.3.1.2 *Intramolecular interactions in GAGs*

In order to elucidate the origin of the flexibility differences observed in the different investigated GAG molecules, we calculated the intramolecular contacts between individual residues (Figure 62). What can be immediately observed is that the interaction map is predominantly dominated by weak contacts. And in the few cases with more pronounced contacts, mainly for H6S and C6S, these interactions are limited to dis-

accharide units in close proximity. For instance, for H6S at 0 mM NaCl, GlcUA of the first disaccharide unit interacts strongly with GlcNAc of the second disaccharide unit and also GlcUA of that disaccharide. From the generally low contact probabilities one can infer that the predominantly sampled structures are extended but with sufficient curvature that allows contacts between neighbouring residues. The observation that GAGs with sulphated moieties at carbon 6 have a tendency to form more contacts suggests that these intramolecular interactions increase their overall flexibility, as reported above. The contacts are stronger at 0 mM salt concentration than at 150 mM NaCl or KCl. For instance, some interactions are present in HA at 0 mM KCl while they are missing at 150 mM KCl. For all simulated systems most contacts are formed at 0 mM KCl, which suggests that Na^+ interacts more strongly with the GAGs thereby preventing the formation of intramolecular GAG contacts.

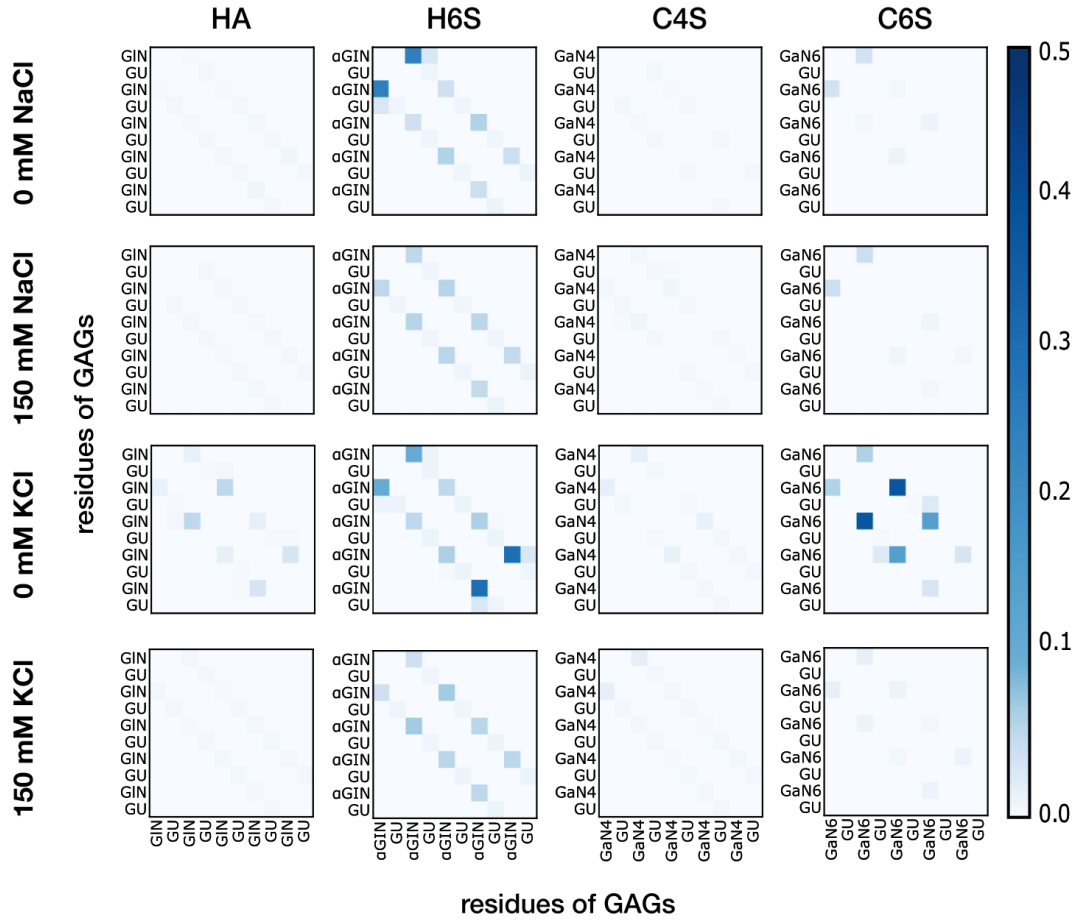

Figure 62: Intramolecular contacts between the residues of the GAGs at 0 mM NaCl, 150 mM NaCl, 0 mM KCl, and 150 mM KCl concentrations (labels left of each panel). The residues of the GAGs are abbreviated with the following codes: GU (GlcUA), GlN (GlcNAc in HA), αGlN (GlcNAc(6S) in H6S), GaN4 (GalNAc(4S)), and GaN6 (GalNAc(6S)). The color code on the right represents the probability of a contact between residues during the MD simulations. For the sake of clarity, the diagonal and first off-diagonal elements of the contact maps corresponding to self-contacts within the same disaccharide unit are not shown.

6.3.1.3 *Shape of the GAGs*

To further characterize the structural preferences of the GAGs, we calculated their R_{ee}. The fully extended GAG molecules involving five disaccharide units have an R_{ee} of about 4.4 nm. The results in Figure 63 reveal that the GAGs simulated at different salt concentrations prefer to remain mostly extended with the main peaks of the R_{ee} distributions hovering around ~4.0 nm. Nonetheless, these distributions also uncover that at 0 mM salt more compact GAG conformations exhibiting smaller R_{ee} values are possible. This is the case for H6S at both 0 mM NaCl and KCl as well as HA, C4S and C6S at 0 mM KCl. These are the same systems for which intramolec-

ular contacts have been identified (Figure 62), which indicates that the interactions between the saccharide units cause the GAGs to deviate from the extended geometry. The R_{ee} distributions for HA, H6S and C4S at the indicated salt concentrations feature a second peak for smaller R_{ee} values, suggesting a switch between extended and more compact GAG conformations. In the case of C6S at 0 mM KCl, on the other hand, the R_{ee} distribution becomes broader and is generally shifted to smaller values. This signals a conformational preference for more compact structures stabilized by contacts between the second and third as well as between the third and fourth disaccharide units. In the case of H6S the strongest inter-disaccharide contacts are formed at the GAG termini, i.e., between the first and second as well as fourth and fifth disaccharide units.

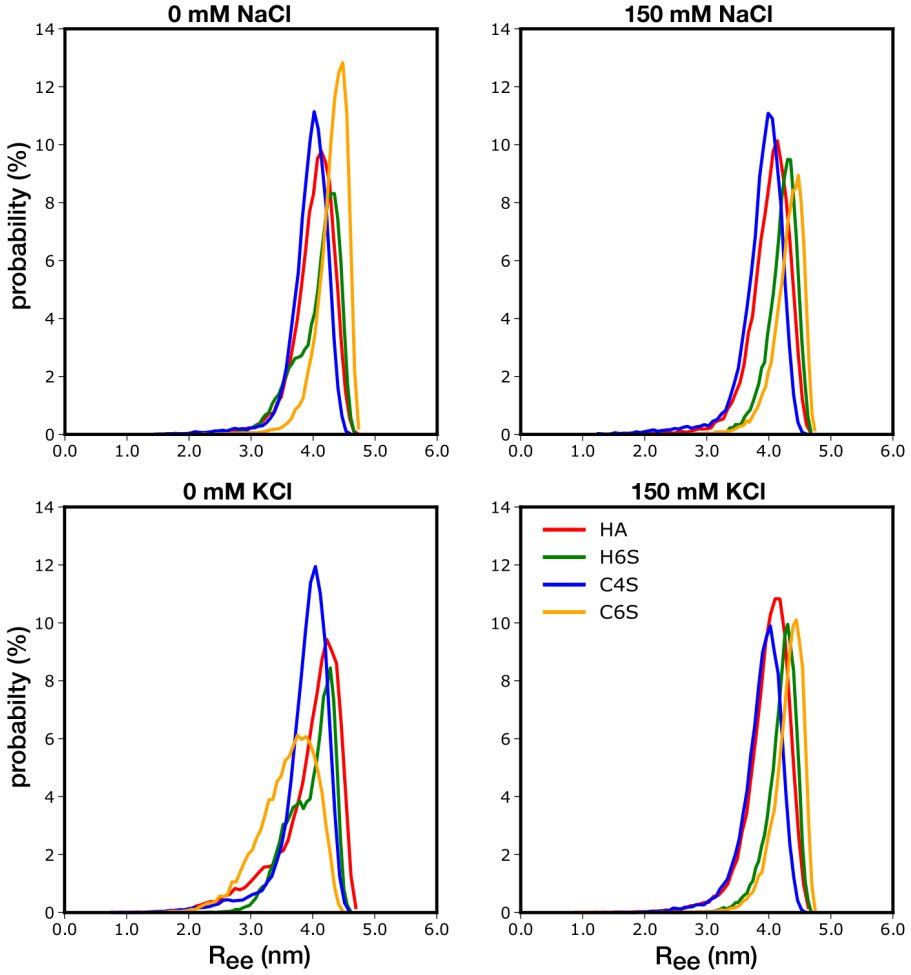

Figure 63: Distribution of the R_{ee} for the different GAGs (see color key) at 0 mM (left) and 150 mM (right) salt concentrations (see color key in the lowest right panel).

The shape of the GAGs is closely coupled to the dihedral angles on either side of the oxygen atoms linking adjacent monosaccharides: these are the dominant degrees of freedom in GAGs since they allow the monosaccharides to adopt different orientations relative to each other. In order to simplify the discussion of these torsional angles, we combined them into a collective offset function for each linkage type according to equation 30. As reference angles we used those for the extended GAG structures. Therefore, $D_{offset} = 2$ corresponds to fully extended conformations and values smaller than two indicate deviation from linearity and thus more compact structures. The results in Figure 64, show that the GAGs experience limited flexibility across $\mathtt{Linkage}_1$. Independent of the salt concentration, salt type and GAG, the D_{offset} values are between 1.6 and 2.0. The smallest values are found for C4S, which agrees to the R_{ee} distributions which also peaked at slightly smaller values compared to the other GAGs at corresponding conditions. However, since we know from the contact maps that no noteworthy intramolecular contacts formed in C4S, we expect this particular GAG to be maximally extended and only slightly curved. We thus conclude that for $\mathtt{Linkage}_1$ only, $D_{offset} < 1.6$ would indicate significant deviation from extended structures. For $\mathtt{Linkage}_2$ such low values are reached for H6S (under all conditions but 0 mM KCl) and C6S (at 0 mM KCl). These are the same systems that feature notable intramolecular contacts leading to more compact conformations. Special attention is needed for HA and H6S at 0 mM KCl where the D_{offset} values are limited to the range 1.6 and 2.0 for $\mathtt{Linkage}_2$, despite the presence of intramolecular interactions. In the case of HA, and also for C6S at 0 mM KCl and H6S under all conditions but 0 mM KCl, a splitting of the D_{offset} distribution into two peaks can be observed. We reckon this to imply that for $\mathtt{Linkage}_2$ this split indicates the formation of more compact GAG conformations. In the case of H6S at 0 mM KCl, however, the D_{offset} function fails to capture the GAG structures deviating from linearity. Here, a more detailed inspection of the ϕ and ψ dihedral angles would be required if one wants to learn more about the individual glycosidic linkage configurations.

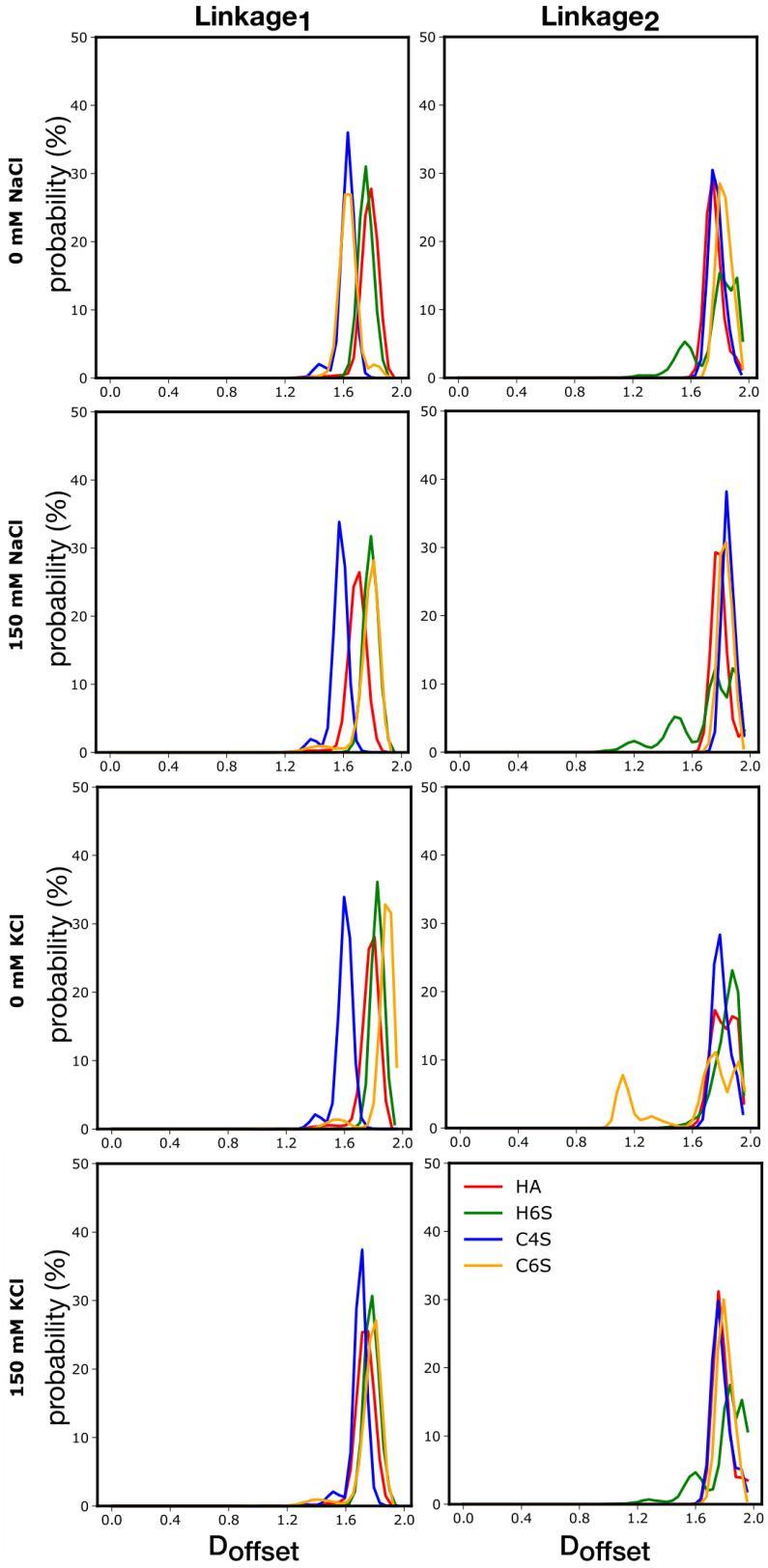

Figure 64: Distribution of the dihedral angle offset function, D_{offset}, for $Linkage_1$ (left) and $Linkage_2$ (right). The different GAGs are shown with different colors (see color key in the lowest right panel). The salt concentrations are shown from top to bottom as 0 mM NaCl, 150 mM NaCl, 0 mM KCl, and 150 mM KCl.

6.3.2 *Characterisation of the GAGs interactions with water and ions*

The structural preferences of the GAGs originate not only from their intrinsic properties, like their chemical composition and sulphation, but also from their interactions with the environment. Here, the GAGs surrounding is provided by the aqueous solvent and ions. To explore their role on the GAGs structures, we analyse the interactions between them here.

6.3.2.1 *GAG-water interactions*

For characterising the GAG-water interactions, we calculated the number of H-bonds formed between GAGs and solvent molecules. In Figure 65 it can be observed that irrespective of the salt concentrations, HA forms the least number of such H-bonds. This can be easily explained with the presence of the sulphate (OSO_3^-) group in the other three GAGs. A comparison between the H-bond numbers reveals that H6S, C4S and C6S tend to build 20–30 H-bonds more with the surrounding water molecules than does HA. Since there are five OSO_3^- in each of H6S, C4S and C6S, one can estimate that each sulphate group gives rise to about 4–5 H-bonds. This is probably slightly overestimated since H-bonds calculated were identified on geometric grounds only. Yet the picture presented serves the purpose to demonstrate the significant effect that OSO_3^- has on GAGs binding of water molecules. The H-bond distributions of H6S, C4S and C6S are very similar, and the GAG-water interactions do not help to explain the structural characteristics of the different GAGs.

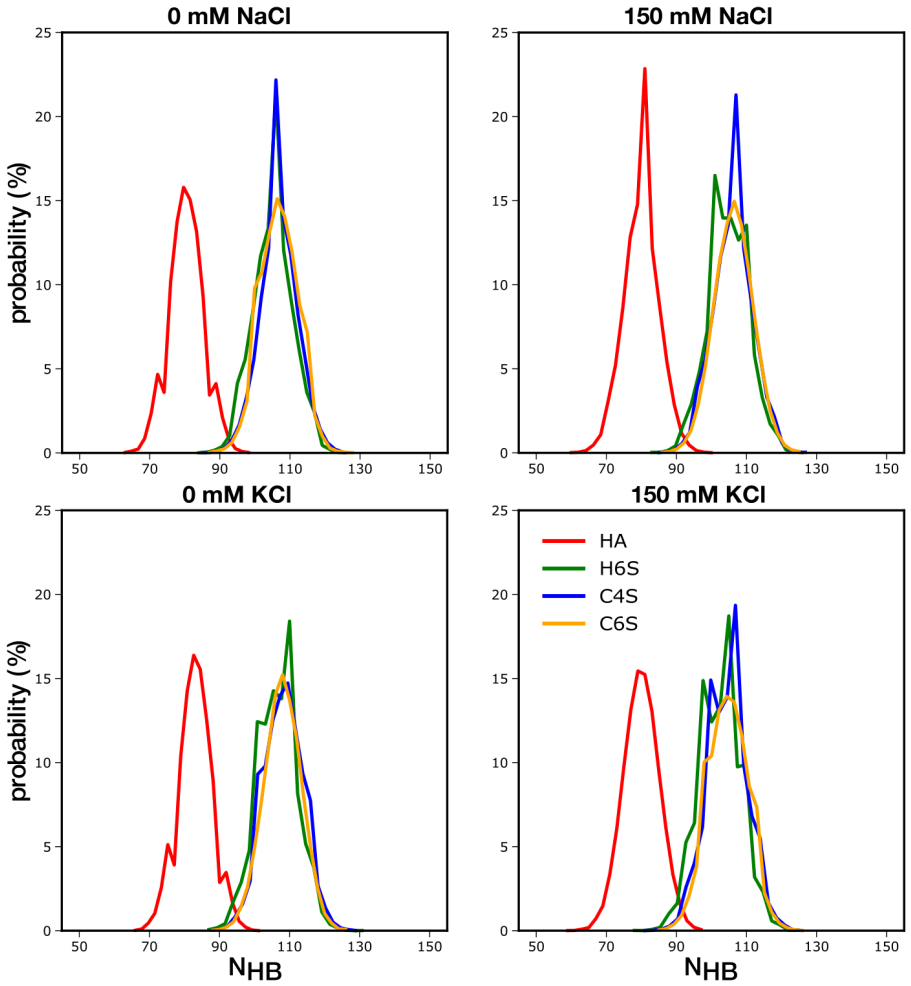

Figure 65: Distribution of the number of H-bonds formed between water molecules and the different GAGs (see color key in the lowest right panel) at 0 mM (left) and 150 mM (right) salt concentrations.

6.3.2.2 *GAG-ion interactions*

We next turn our attention to the interactions between the GAGs and the cations, K^+ and Na^+. Given the negative charge of the GAGs, there are no attractive interactions with the Cl^- ions that need to be considered. As shown by others, cations prefer to localise near the OSO_3^- and COO^- groups of the GAGs (see Figure 66 for the atom naming of these groups), leading to the formation of contact ion pairs and solvent separated ion pairs [330]. These pairs can be identified from the radial distribution functions ($g(r)$), which are shown in Figure 68 for COO^- and in Figure 67 for OSO_3^-. The cation-anion RDFs typically display two peaks: a larger one corresponding to direct cation-anion contacts at distances between 0.2 and 0.3 nm, and a smaller one at a distance of 0.4–0.5 nm for the water separated ion pair.

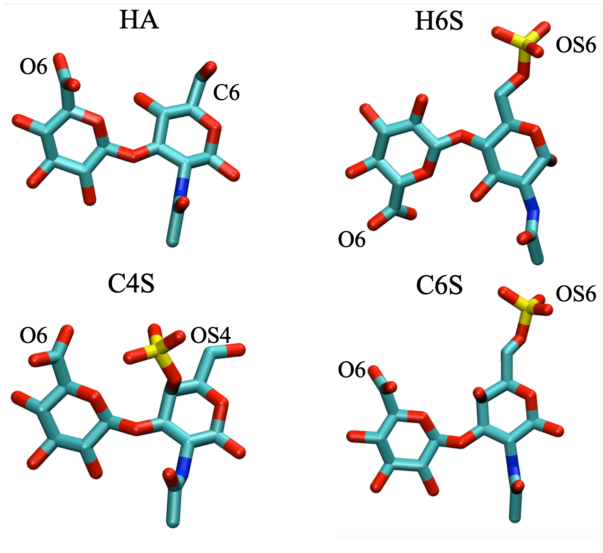

Figure 66: Identification of atoms relevant for the RDF calculations.

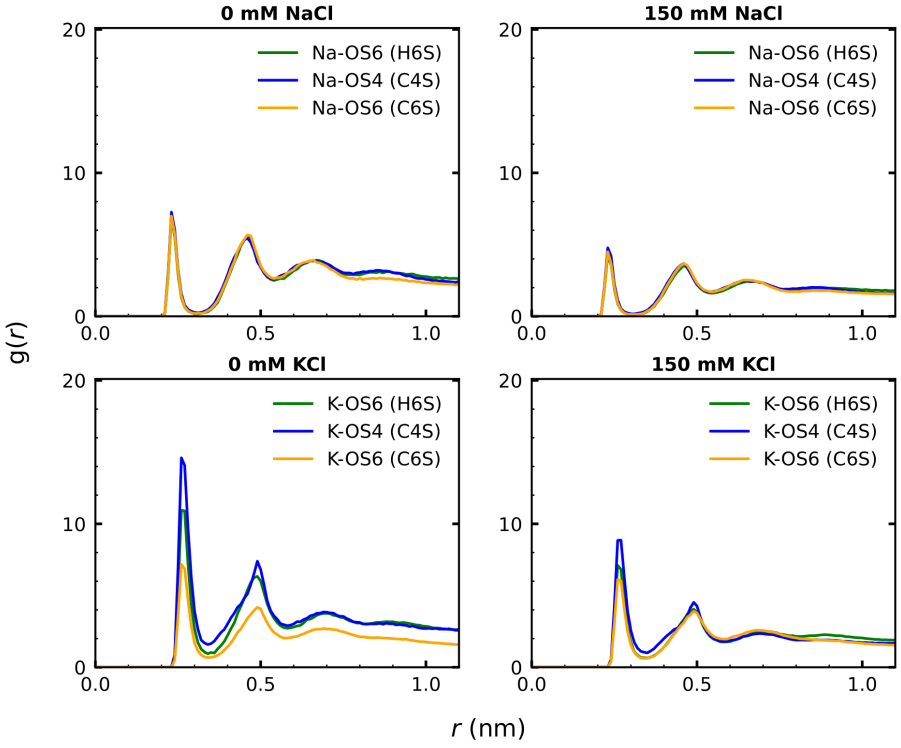

Figure 67: RDF curves representing the interactions between OSO_3^- and the cations Na^+ (top) and K^+ (bottom) at 0 mM (left) and 150 mM (right) salt concentration. The RDF curves are averaged over the three oxygen atoms bound to the sulphur atom (see Figure 66 for the atom naming) and over the five OSO_3^- per GAG.

Comparison of the large peaks for the contact ion pairs uncovers three key differences among the systems. First, the direct cation-anion interaction with K^+ is generally stronger than with Na^+. This difference is especially pronounced for the interac-

tion with COO^-. This finding is somewhat surprising, as the more strongly hydrated cations (here Na^+) usually bind more tightly to carboxylate groups than do the less hydrated cations (here K^+) [331], which can be rationalised with the fact that Na^+ matches better the hydration enthalpy of COO^- [332]. However, inspection of the second peak shows that the water separated Na^+–COO^- pair is stronger than the corresponding K^+–COO^- pair. This observation is understandable if one considers that one cation is often shared between two neighbouring anions [330]. Such situations can be seen in Figure 73A, where Na^+ ions are shared between COO^- and OSO_3^- groups of C6S, with all of the Na^+-anion distances being above 0.4 nm. These are distances of solvent separated ion pairs, which are the preferred arrangement for Na^+ given its smaller size compared to K^+ that decreases the possibility of direct Na^+-anion contacts. For K^+ direct ion contacts are possible despite the cation being shared between COO^- and OSO_3^- (Figure 73B). The second difference is that the carboxylate group interacts more strongly with either cation than the sulphate group; this agrees well with previous findings [330]. Finally, in the absence of Cl^- ions, i.e. in the 0 mM KCl and NaCl systems, the interactions between either cation and either of COO^- and OSO_3^- are stronger than in the presence of Cl^- ions (i.e. in the 150 mM KCl and NaCl systems). This indicates that the presence of the Cl^- ions reduces the affinities of Na^+ and K^+ for the other anions in the system, i.e., COO^- and OSO_3^-.

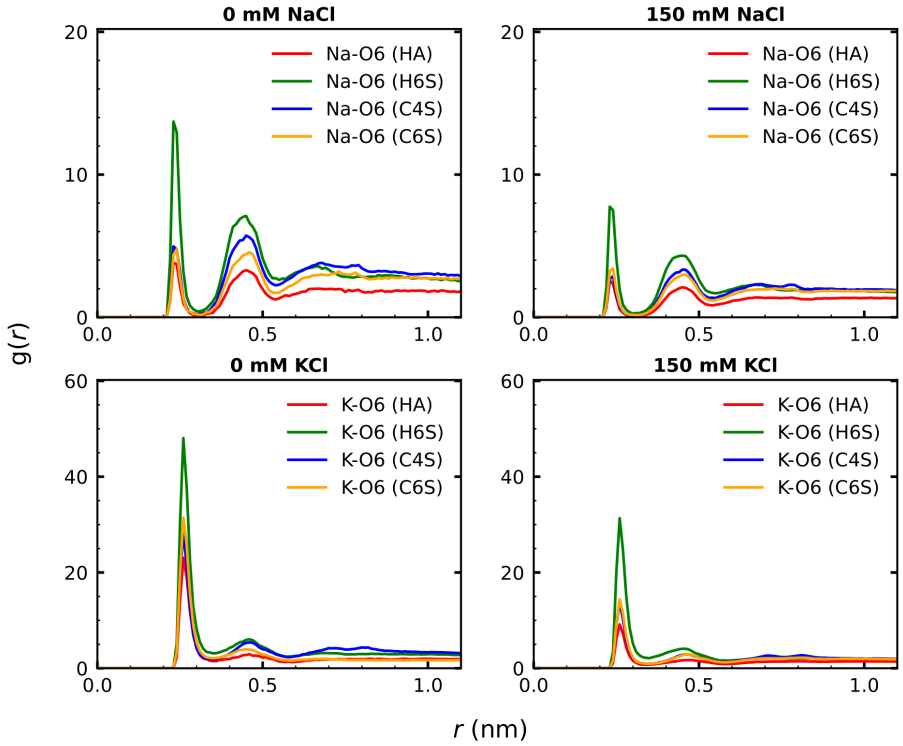

Figure 68: RDF curves representing the interactions between COO^- and the cations Na^+ (top) and K^+ (bottom) at 0 mM (left) and 150 mM (right) salt concentration. The RDF curves are averaged over both oxygen atoms bound to carbon atom 6 (see Figure 66 for the atom naming) of GlcUA and over the five GlcUA units per GAG, in addition to averaging over the simulation time.

The analysis of the interactions of the GAGs with the cations permits to identify their effects on the GAGs structural preferences, especially when focusing on COO^- (Figure 68). The H6S group displays the highest interaction preference for both cations and salt concentrations. It should be recollected that H6S is the same GAG with the highest level of intramolecular contacts. We can thus conclude that the GAG-ion interactions enable more conformational flexibility in the GAGs. This results from a screening of their negative charges from each other, reducing the tendency of the GAGs to adopt extended structures where maximum distances between the COO^- and OSO_3^- groups can be reached. This reasoning is supported by the fact that at 0 mM KCl and NaCl more intramolecular GAG contacts are formed than at 150 mM salt concentration where the GAG-ion interactions are weaker. Another argument in favor of it is provided by the fact that C6S shows both a substantially higher interaction with cations at 0 mM KCl compared to all other salt conditions, and the largest number of intramolecular contacts at 0 mM KCl. The same trend is seen with

HA even though HA has no sulphate group and the effect is therefore reduced. Only for C4S, no profound structural flexibility effect of the GAG-ion interactions could be identified, even though the interactions with Na^+ and K^+ follow similar trends as seen in C6S. Thus, the conformational freedom of a GAG is an intricate interplay between the position of the sulphate group and external conditions.

6.3.3 *Free energy profile based on collective fluctuations*

As we have studied the behaviour of GAGs with a wide variety of calculated structural observables, we now present an approach to disentangle the conformational states by a free-energy method. We start by concatenating four of the previously introduced features, namely D_{offset} for $Linkage_1$ and $Linkage_2$, R_{ee}, and N_{HB}, into one state vector, which represents the structural evolution of the GAGs. As an example, the projection of all MD snapshots sampled for the GAGs at 150 mM NaCl are shown in Figure 69. A PCA is applied to this state vector and the first two PCs are chosen for calculating the free energy according to equation 31. The resulting free energy surfaces for the 150 mM NaCl systems (Figure 71) reveal one main energy basin for each GAG. The width of these basins is somewhat larger for C6S and H6S, which is in agreement with their increased flexibility identified by the clustering analysis (Figure 59). The corresponding projection of all MD snapshots onto the first two PCs can be seen in Figure 72. In order to identify representative structures, we performed a k-means clustering of the structures. For deciding on the numbers of clusters (k) to consider, we plotted the inertias for increasing k (Figure 70) and applied the elbow method to it. From this we found out that four clusters are sufficient in segregating the conformational space, since with $k > 4$ the inertias reduce only mildly. The centroids of the four clusters were projected along the first two PCs and representative GAG structures per cluster centroid were retrieved (Figure 72). In Table 9 the structural characteristics of these centroid structures are summarized.

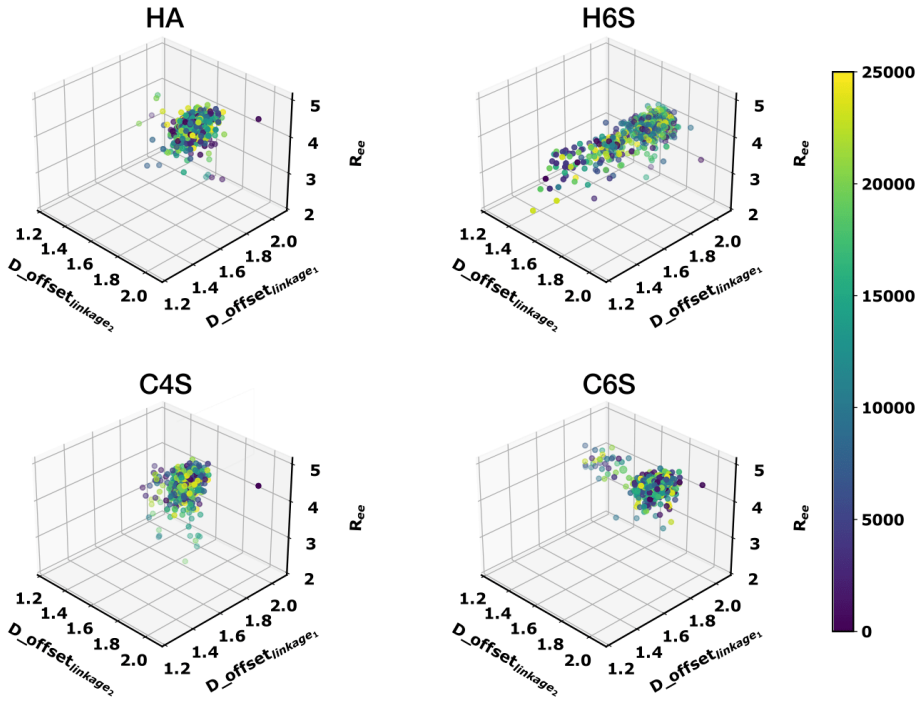

Figure 69: Distribution of the R_{ee} and D_{offset} values for $Linkage_1$ and $Linkage_2$ obtained for the GAG systems at 150mM NaCl. Results for 25,000 MD snapshots are shown and the coloring of the dots reflects the simulation time, with dark blue for $t = 0$ (first snapshot) and yellow for $t = 1$ µs (last snapshot).

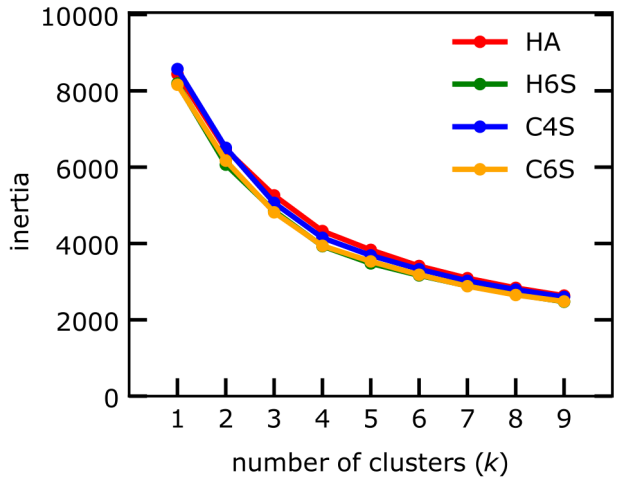

Figure 70: Determination of the optimal number of clusters, k, by plotting the inertia for increasing k. Results are shown for the GAGs simulated in the presence of 150 mM NaCl.

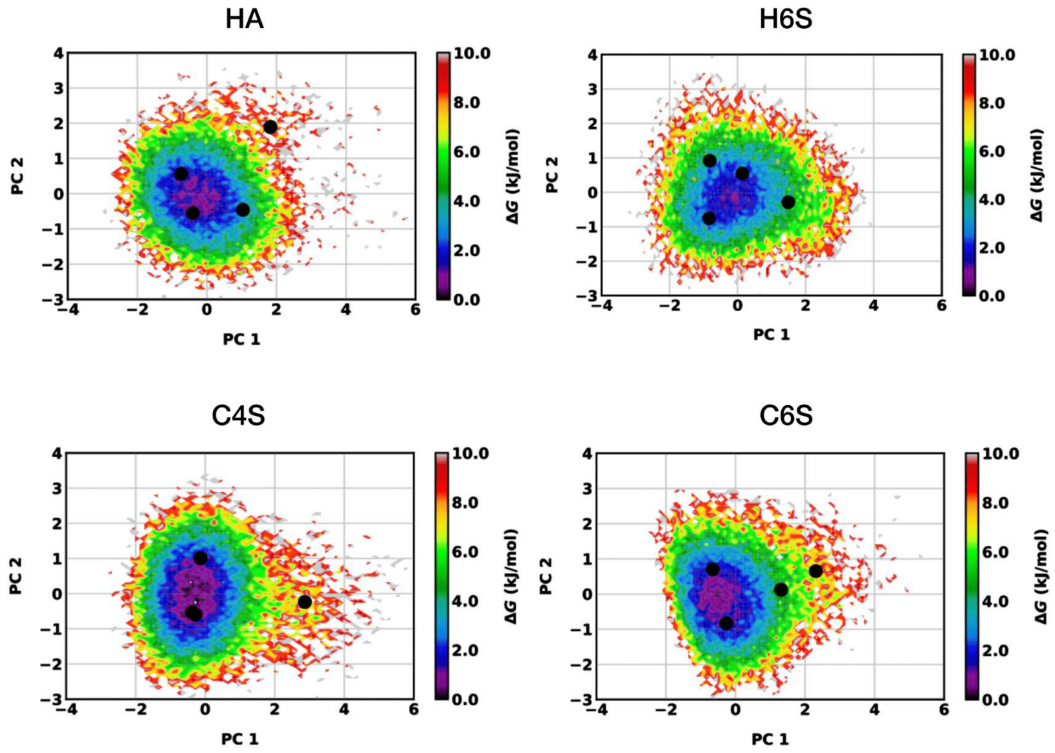

Figure 71: The free energy landscapes projected onto PC1 and PC2 for the GAG systems simulated in the presence of 150mM NaCl.

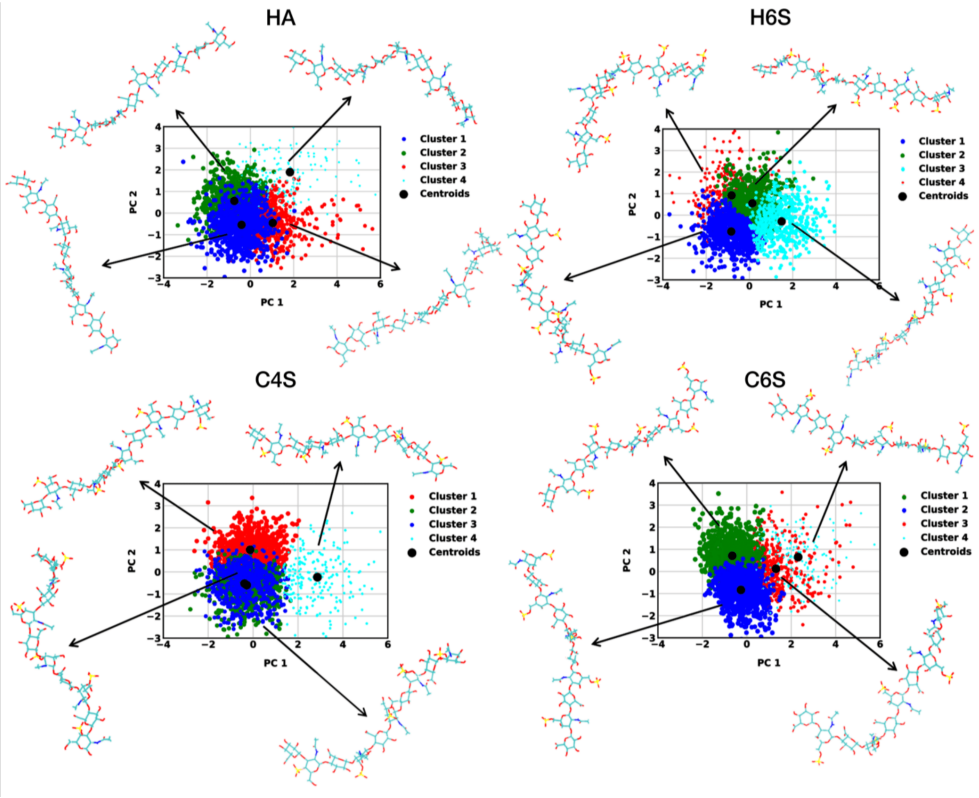

Figure 72: Projection of the MD trajectories obtained at 150 mM NaCl onto the first two PCs for the different GAGs (see label above each panel). The conformations are segregated into four clusters, which are represented by different colours. Structures corresponding to the centroid structures are shown.

Inspection of the cluster structures supports the conclusions drawn from the contact map analysis (Figure 62). First, the GAGs have a high preference to adopt extended conformations, as the structures representing the most populated clusters for each of the GAGs fall into this category. For HA, H6S and C6S the R_{ee} of the first two centroid structures are above 4.2 nm and for C4S above 3.9 nm. Structures with smaller end-to-end distances are sampled with lower probabilities and are included in cluster 4 of all four GAGs. These more compact structures feature kinks across one of their glycosidic linkages. These kinks can lead to perpendicular orientations, especially when they occur after two or three disaccharide units, while at the termini the kinks lead to hairpin like turns. Such kinks can be stabilised by binding a cation between the COO^- and OSO_3^- group facing each other, as seen in Figure 73. Alternatively, the ϕ and ψ angles adopt values such that the neighbouring negative charges point away from each other. An interesting observation noted in Table 9 is that the average N_{HB} value is greater for the highest populated cluster states for all GAGs. This

is understandable considering that these are the most expanded GAG structures that have a larger solvent accessible surface area than more compact structures. This analysis shows that, while the N_{HB} distributions shown in Figure 65 did not help much to gain a deeper understanding of the GAGs conformational preferences, combining this analysis with structural clustering does help to correlate the different structural aspects with each other.

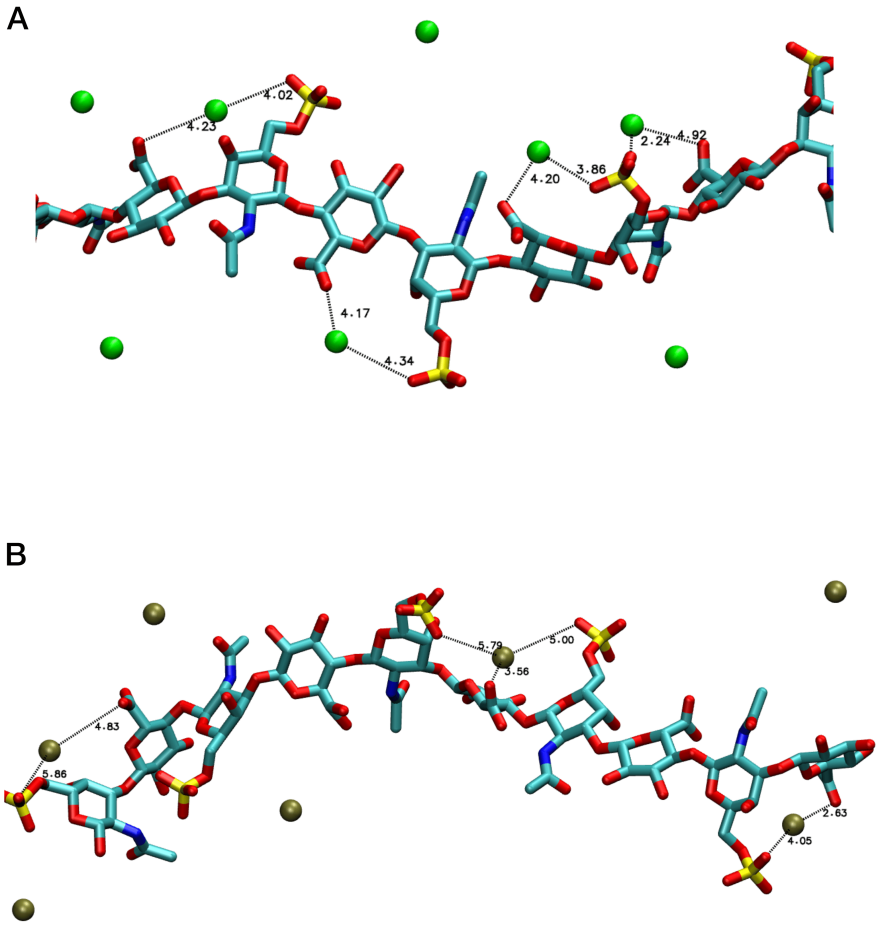

Figure 73: Snapshots illustrating the GAG interactions with Na^+ (top) and K^+ (bottom). C6S is chosen as example here. (Distances shown are in Å).

6.4 CONCLUSIONS

We have investigated the structural dynamics of four GAGs – HA, H6S, C4S, and C6S – using microsecond-long MD simulations and modeling different salt conditions.

System	Structure	%	R_{ee}(nm)	N_{HB}	$Linkage_1$(°)	$Linkage_2$(°)
HA	starting		4.5	83.0	(-132.1, -146.1)	(-93.4, 76.1)
	cluster 1	34.8	4.2±0.0	84.9±0.1	(-74.5±0.2, -128.2±0.2)	(-77.2±0.2, 124.3±0.3)
	cluster 2	34.6	4.2±0.0	76.3±0.1	(-73.8±0.2, -128.7±0.2)	(-78.6±0.2, 124.6±0.3)
	cluster 3	25.0	3.8±0.0	81.5±0.2	(-67.7±0.2, -119.7±0.4)	(-78.9±0.3, 129.3±0.2)
	cluster 4	5.6	3.4±0.1	81.8±0.4	(-71.5±0.5, -123.4±0.8)	(-78.4±0.6, 87.9±1.3)
H6S	starting		4.5	102.00	(-111.6, 87.2)	(45.6, 64.3)
	cluster 1	32.3	4.3±0.0	109.7±0.2	(-70.9±0.2, 123.9±0.2)	(79.4±0.4, 46.9±1.0)
	cluster 2	31.4	4.3±0.0	104.1±0.2	(-76.8±0.2, 116.3±0.2)	(80.8±0.3, 40.4±1.0)
	cluster 3	22.2	4.4±0.0	102.7±0.2	(-73.4±0.2, 119.1±0.3)	(86.0±0.3, -22.4±0.9)
	cluster 4	14.1	3.8±0.0	106.3±0.3	(-73.4±0.3, 120.7±0.4)	(79.6±0.5, 23.1±1.8)
C4S	starting		4.4	108.0	(-139.4, -146.9)	(-102.8, 91.1)
	cluster 1	34.0	3.9±0.0	111.1±0.1	(-67.3±0.2, -121.2±0.2)	(-76.4±0.2, 128.4±0.3)
	cluster 2	30.7	4.0±0.0	101.6±0.1	(-67.6±0.2, -121.6±0.2)	(-78.5±0.2, 125.5±0.3)
	cluster 3	26.6	4.0±0.0	109.0±0.2	(-67.5±0.2, -121.9±0.2)	(-80.2±0.2, 112.7±0.5)
	cluster 4	8.7	3.2±0.0	105.9±0.4	(-52.3±0.7, -119.0±0.3)	(-78.7±0.4, 114.8±1.5)
C6S	starting		4.4	114.0	(-120.8, -156.3)	(-106.6, 87.4)
	cluster 1	38.8	4.4±0.0	111.5±0.1	(-74.3±0.2, -140.6±0.3)	(-72.0±0.2, 118.6±0.4)
	cluster 2	38.0	4.4±0.0	103.3±0.1	(-74.3±0.2, -140.1±0.2)	(-75.4±0.2, 114.0±0.4)
	cluster 3	17.6	3.9±0.0	106.8±0.2	(-69.8±0.3, -133.4±0.4)	(-75.6±0.3, 111.4±0.9)
	cluster 4	5.6	4.3±0.0	106.3±0.5	(-73.7±0.8, -78.9±1.3)	(-75.7±0.5, 117.0±0.9)

Table 9: The values of R_{ee}, N_{HB}, ϕ and ψ pairs for $Linkage_1$ and $Linkage_2$ of the centroid structures of the four clusters (populations are provided above) obtained for the GAGs at 150 mM NaCl. The corresponding values of the initial structures used in the MD simulations are provided too. (mean ± standard error)

This allowed us to asses the influences of GAG composition and glycosidic linkage, sulphation, and external situations on their conformational ensembles. To quantify the GAGs dynamics, different quantities, such as the RMSD, the end-to-end distance, and the dihedral angles across the glycosidic linkages were determined. Moreover, to rationalize the GAGs structural preferences, we analysed their interactions with water and the metal ions present in the systems. Finally, a method was presented that allows to efficiently determine the preferred GAGs structures and locate them in the conformational free energy landscape.

A general conclusion is that the GAGs have an intrinsic preference for extended structures. Deviation from linearity is enabled by sulphation, especially at position 6, and interaction with cations. The latter screen the negative charges that are next

to each other in the GAGs, which allows the COO^- and OSO_3^- groups to get closer to each other, causing kinks and bends in the GAGs. If other anions, such as Cl^- are present in the system, this screening is reduced. As a result, H6S and C6S in the presence of only Na^+ or K^+ (but no Cl^-) are the most flexible, while HA and C_4S in the presence of 150 mM salt are the stiffest. The higher flexibility represents itself in more contacts between the disaccharide units of C6S and especially H6S. The differences between C6S and H6S highlight that also the GAG sequence and glycosidic linkage matter with respect to their conformational flexibility. The metal ions usually adopt a bidentate configuration when binding to the GAGs, where they interact with two anions at the same time. In order to realize such bidentate binding, the small Na^+ must interact with COO^- and OSO_3^- with a water molecule between them. The OSO_3^- groups have a large effect on attracting water to the GAGs, as the comparison of the GAGs-water interactions for HA and the three sulphated GAGs showed.

In summary, we showed that it has become possible to reveal key insight into the structural dynamics of GAGs by means of MD simulations. Detailed knowledge thereof is needed if one aims to correlate GAG sequence with activity.

7

INFLUENCE OF GLYCOSAMINOGLYCANS ON Aβ (16–22) AGGREGATION

Parts of this chapter are submitted by the author to the *Journal of Physical Chemistry B* as Samantray, S., Strodel, B. (2021). The effects of different glycosaminoglycans on the structure and aggregation of the amyloid-β (16–22) peptide. All the images are reprinted with Copyright ©2021 **American Chemical Society**.

7.1 MOTIVATION

Clinical examinations of the Aβ aggregates samples from AD affected patients have revealed a significant presence of charged polyelectrolytes, especially polysaccharides, belonging to GAGs [84, 333, 334]. Concurring to this observation, about 20% of the total volume of an adult's brain is occupied by the extracellular space, which is composed of an interlocking mesh of fibrous proteins, GAGs, PGs, and glycoproteins [335]. Structurally, the GAGs and PGs play a major role in preventing the diffusion of soluble proteins and cell migration inside the brain's extracellular space [285, 287]. Moreover, the GAGs play an active role in enhancing amyloid fibril formation and its stabilization [14, 290, 336].

Many studies have reviewed the biochemical nature of GAGs and their role in neurodegeneration [337]. The hypotheses resulting from these previous studies suggest that the mechanisms involved in GAG-catalyzed fibril formation are different from that occurring in bulk phases [337]. It is thought that GAGs function as a structural template, thereby influencing the structural preference of amyloid forming peptides, causing them to favor β-sheet conformations and subsequently shaping the aggregation patterns. At the later aggregation stages, fibrillation is clearly favoured [19],

thereby scavenging neurotoxic oligomeric precursors [14]. Coupled with this, different GAGs induce different changes to the aggregation pathways, probably because of the presence and position of the sulphate groups [338]. For instance, HS greatly affects the structural transition of Aβ in the early stages of amyloid formation [289]. However, all the hypotheses compiled by Iannuzzi *et al.* [14] fail to provide conclusive mechanistic insights on the role of GAGs in the early stages of fibrillation albeit highlighting the role of GAGs to induce amyloidogenesis in soluble peptides and proteins [14, 290, 296, 339].

The tendency of Aβ to form aggregates in bulk solution is governed to a large extent by its hydrophobic core region (residues 17–21). In the presence of heteropolysaccharides, on the other hand, the aggregation pathways of proteins are mostly guided by electrostatic interactions involving the negative polyelectrolyte charges of the GAGs and positively charged side chains [339, 340]. Apart from the fact that the electrostatic nature of the sulphate groups of GAGs plays a crucial role in the fibrillation process, little is known about the intrinsic details on the mechanism followed during the aggregate formation. To resolve this question, MD simulations at atomistic resolution can be used [49].

In this chapter, we provide a detailed description of the mutual effects of GAGs and the $A\beta_{16-22}$ peptide on their structural heterogeneity, dynamics, and aggregation. $A\beta_{16-22}$ is an attractive model for studying amyloid aggregation since this peptide is able to form fibrils [240]. Moreover, its small size makes $A\beta_{16-22}$ ideal for exploring its aggregation using MD simulations [3, 109, 232, 244–246, 248, 250, 341]. For the GAGs we consider HA, H6S, C4S, and C6S (Figure 3 on page 13). We simulate the interactions between $A\beta_{16-22}$ and the respective GAG at the 1:1 and 2:2 stoichiometry. The 1:1 simulations allow us to obtain a first idea of the interactions between the GAGs and $A\beta_{16-22}$, while the 2:2 simulations provide insight into the effects of the GAGs on the aggregation of $A\beta_{16-22}$, where the aggregation pathways are being elucidated by transition networks [248, 341]. As reference systems the MD simulations of $A\beta_{16-22}$ without GAGs and of monomeric GAGs are also performed and included in this chapter.

7.2 MODELS AND SIMULATION DETAILS

7.2.1 *GAG Models*

To understand the effects of different GAGs on Aβ_{16-22}, we simulated the following five repeating disaccharide units: HA as a nonsulphated GAG and H6S, C4S, and C6S as sulphated GAGs (Figure 3 on page 13). The disaccharide repeating units for HA are composed of GlcNAc and GlcUA linked via alternating β-(1→4) and β-(1→3) glycosidic bonds (O-linked): -GlcNAc-β(1→4)-GlcUA-β(1→3)-. In H6S, the disaccharide repeating units are identical to those in HA apart from the sulphation at position 6 of GlcNAc. Moreover, the first glycosidic linkage has a different geometry, α-(1→4) instead of β-(1→4), which results in -GlcNAc(6S)-α(1→4)-GlcUA-β(1→3)-. C4S and C6S are composed of similar components and glycosidic linkage patterns as HA, except for GlcNAc being replaced by GalNAc. The difference between C4S and C6S exists in the sulphation position, which is at carbon 4 of GalNAc in C4S and at position 6 of GalNAc in C6S, leading to -GalNAc(4S)-β(1→4)-GlcUA-β(1→3)- and -β(1→4)-GlcUA-β(1→3)-GalNAc(6S)-, respectively. The disaccharide units and linkages (referred to as $\mathtt{Linkage}_1$ and $\mathtt{Linkage}_2$) are summarized in Table 7 on page 134. For modeling Aβ_{16-22}, we capped the N- and C-termini of the heptapeptide with acetyl and N-methylamide groups.

7.2.2 *Simulation Protocols*

We simulated the 1:1 systems by placing an extended GAG structure and an Aβ_{16-22} peptide randomly in the simulation box, making sure that their intermolecular distance is at least 1.0 nm for any pair of atoms. The starting configurations are shown in Figure 74.

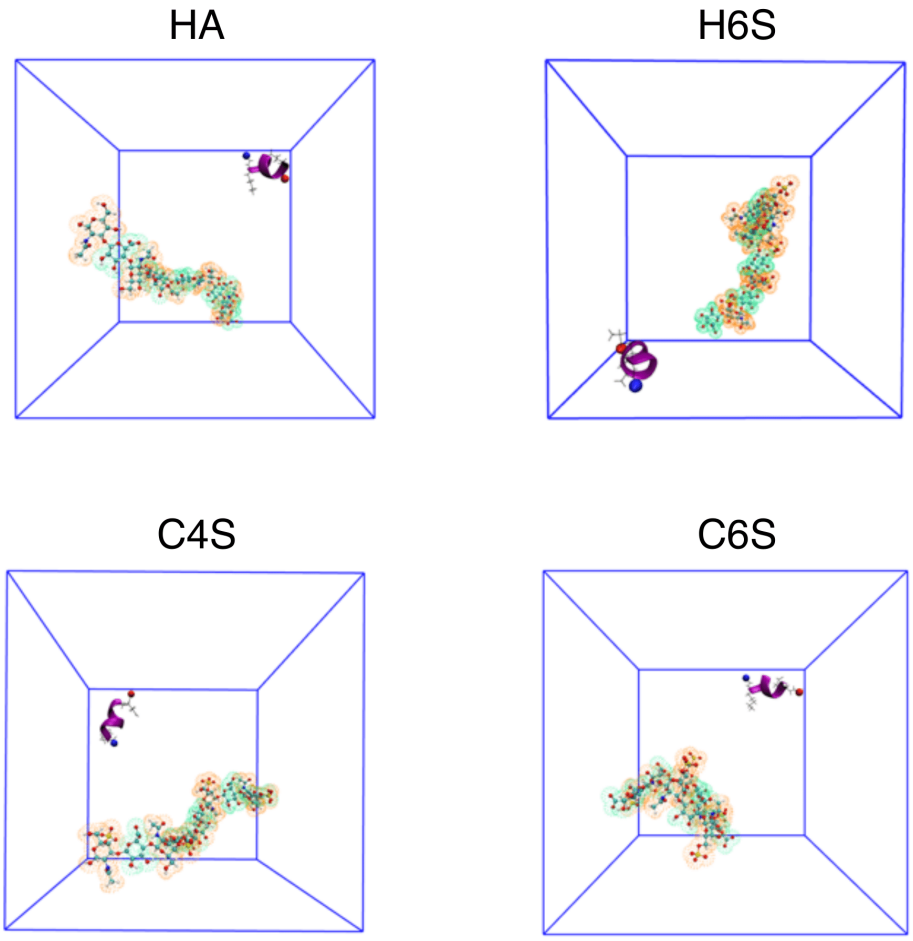

Figure 74: Initial configurations of the 1:1 $A\beta_{16-22}$-GAG systems.

For modeling $A\beta_{16-22}$, we employed the all-atom force field C36m [113], while CHARMM parameters as available through the 'Glycan Reader & Modeler' module [322–324] of the CHARMM-GUI web server [325] were used for the GAGs. The peptide and the respective GAG were randomly placed in a cubic simulation box with an edge length of 4.9 nm and solvated with water using the TIP3P model [254], resulting in ~11,700 atoms in total. To each system, NaCl was added to reach a physiological concentration of about 150 mM and neutralize the simulation system, which corresponds to 15 Na^+ and 10 Cl^- ions in the case of HA and 20 Na^+ and 10 Cl^- ions for C4S, C6S and H6S. Each system was minimized using the steepest descent algorithm, followed by equilibration, first with a 20 ps run in the NVT ensemble while restraining the non-hydrogen atoms of $A\beta_{16-22}$ and GAG atoms to their initial positions, afterwards with a 20 ps run in the NpT ensemble without position restraints. For the production runs, we simulated the 1:1 $A\beta_{16-22}$-GAG systems for 1 µs in

the NpT ensemble with T = 300 K and p = 1 bar. Throughout the simulations, we constrained all bonds involving hydrogen atoms using the LINCS algorithm [191]. The electrostatic and van der Waals interactions were calculated using the particle mesh Ewald (PME) method [192] and the real-space components truncated at 1.2 nm. The temperature and pressure were controlled using a Nose-Hoover algorithm with a 1 ps time constant for coupling and a Parrinello-Rahman barostat [190] with a relaxation time of 5 ps, respectively. A time step of 2 fs was used for the integration of the equations of motion for all systems.

System	Size (atoms)	Box-edge length (nm)	Simulation time (μs)
Reference systems			
$A\beta_{16-22}$ (monomer)	14,392	5.0	1.0
$A\beta_{16-22}$ (dimer)	12,931	5.0	5.0
HA (monomer)	31,939	6.8	1.0
H6S (monomer)	31,902	6.8	1.0
C4S (monomer)	31,890	6.8	1.0
C6S (monomer)	31,875	6.8	1.0
1:1 systems			
$A\beta_{16-22}$:HA	11,723	4.9	1.0
$A\beta_{16-22}$:H6S	11,449	4.9	1.0
$A\beta_{16-22}$:C4S	11,713	4.9	1.0
$A\beta_{16-22}$:C6S	11,668	4.9	1.0
2:2 systems			
$A\beta_{16-22}$:HA	94,870	9.8	5.0
$A\beta_{16-22}$:H6S	94,209	9.8	5.0
$A\beta_{16-22}$:C4S	94,812	9.8	5.0
$A\beta_{16-22}$:C6S	94,758	9.8	5.0

Table 10: Overview of the simulated systems

A similar simulation protocol was used to simulate the 2:2 $A\beta_{16-22}$-GAG systems. An initial 1:1 conformation, in which the $A\beta_{16-22}$ peptide and the GAG were not in contact with each other, was extracted from the preceding simulations and one $A\beta_{16-22}$ peptide and one GAG molecule were randomly added to the corresponding system. Here, it was made sure that the additional peptide and additional GAG molecule were at least 1.0 nm away from each other and also from the two molecules already present in the simulation box. The four starting configurations are shown in Figure 75. The box size was increased to an edge length of 9.8 nm leading to ~94,500 atoms including water. Also, the number of ions had to be increased to still reach an NaCl concentration of 150 mM: 99 Na^+ and 89 Cl^- ions in the case of HA and 108 Na^+ and 88 Cl^- ions for C4S, C6S and H6S. Each 2:2 system was simulated for 5 µs.

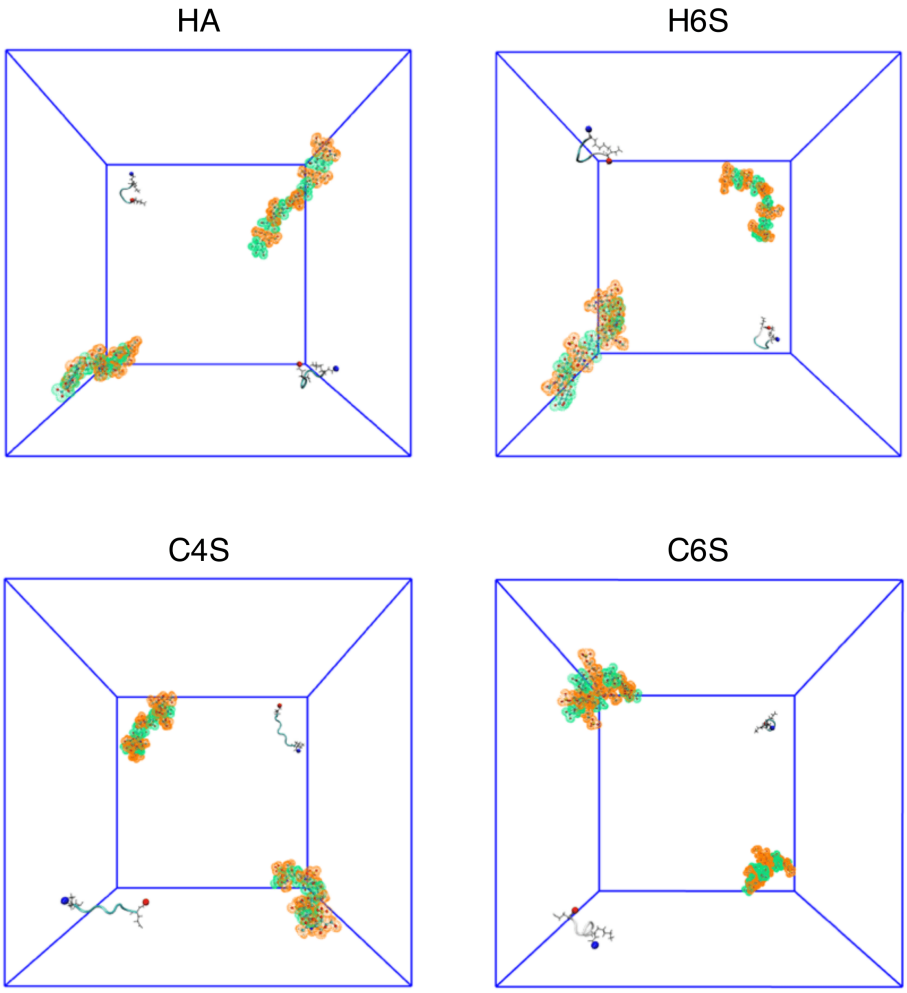

Figure 75: Initial configurations of the 2:2 $A\beta_{16-22}$-GAG systems.

In addition, reference systems, including monomeric and dimeric $A\beta_{16-22}$ as well as the monomeric GAGs were also simulated. All simulations performed in this chapter along with their key settings are listed in Table 10. All simulation were realized with GROMACS version 2018.3 [157, 326, 327]. For the analysis of the simulation trajectories, we employed a combination of standard GROMACS tools, VMD [195] and in-house Python scripts [341, 342] invoking the MDAnalysis [196] and MDTraj [328] libraries.

7.2.3 *Conformational analysis of GAGs*

We determined representative GAG structures using the GROMOS clustering algorithm of Daura et al. [156] with a cutoff for the root-mean-square deviation (RMSD) of 0.2 nm. For characterizing the GAG structures, we calculated the end-to-end distance (R_{ee}) by computing the distance between the C_4 atom of the starting monosaccharide and the C_1 atom of the terminal monosaccharide (see Figure 57 for the numbering of the GAG atoms). Moreover, we also calculated the dihedral angles ϕ and ψ across the glycosidic linkage bonds. Their definitions are shown in Figure 57 on page 138. The ϕ and ψ dihedral angles for each linkage type are combined into a single dihedral offset function [329] (D_{offset}) in equation 30 on page 137. For determining the contacts between GAG molecules, we considered two saccharide residues to be in contact with each other if the distance between any pair of atoms from residue *a* and residue *b* was 0.4 nm or less.

7.2.4 *Structural characterization of $A\beta_{16-22}$*

For the characterization of the $A\beta_{16-22}$ structures in the 1:1 systems, we calculated the secondary structure propensity of the $A\beta_{16-22}$ residues based on their Φ and Ψ angles. Here, an amino acid is assumed to adopt a β-strand conformation if its (Φ, Ψ) pair is within the polygon with vertices at $(-180°, 180°)$, $(-180°, 126°)$, $(-162°, 126°)$, $(-162°, 108°)$, $(-144°, 108°)$, $(-144°, 90°)$, $(-50°, 90°)$, and $(-50°, 180°)$, and in an α-helical conformation if Φ and Ψ are confined to the polygon $(-90°, 0°)$, $(-90°, -54°)$,

$(-72°, -54°)$, $(-72°, -72°)$, $(-36°, -72°)$, $(-36°, -18°)$, $(-54°, -18°)$, and $(-54°, 0°)$. All other angles are counted as random coil [343]. Moreover, we determined the end-to-end distance using the C_α atoms of the first and last residue as well was the radius of gyration of $A\beta_{16-22}$. For the 2:2 systems, we analyzed the H-bonds that formed between the peptides and between $A\beta_{16-22}$ and the GAGs. Here, an H-bond is defined to be present if the donor–acceptor distance is less than 0.35 nm and the donor-H-acceptor angle is less than 30°. The contacts between $A\beta_{16-22}$ and the GAGs were calculated using the same procedure as for the contacts between GAG molecules. To understand the nature of these contacts, the electrostatic and LJ interaction energies between $A\beta_{16-22}$ and the GAGs were calculated using the 'rerun' option of the GROMACS 'mdrun' command. The same was done for the intra- and interpeptide (this only in the 2:2 system) interactions. To elucidate the aggregation pathways of $A\beta_{16-22}$ in the 2:2 systems, TNs were calculated using the Python notebook ATRANET (`https://github.com/strodel-group/ATRANET`) [341] and visualized with Gephi (`https://gephi.org/`) [265]. To characterize the nodes of the TNs, the following descriptors were used: (i) the oligomer size, which can be a monomer or dimer; (ii) the number of interpeptide hydrophobic residue contacts; (iii) the compactness of the aggregate; (iv) the number of residues that are part of a β-sheet (as determined by the STRIDE algorithm [344]).

7.3 RESULTS AND DISCUSSIONS

7.3.1 *Conformational behaviour of GAGs and $A\beta_{16-22}$ and their interactions at 1:1 stoichiometry*

7.3.1.1 *Effects of $A\beta_{16-22}$ on GAGs*

We start the analysis by elucidating whether and how $A\beta_{16-22}$ affects the conformations of the GAGs. To this end, we calculated the distribution of the end-to-end distance, R_{ee}. The fully extended GAG molecules involving five disaccharide units have an R_{ee} of about 4.4 nm. When being simulated without $A\beta_{16-22}$ being present,

Figure 76 shows that the GAGs mostly remain extended, especially H6S and C6S, as the peaks of the R_{ee} distributions are all close to 4.0 nm. In Figure 77A, for each of the four GAGs an elongated conformation is shown. Moreover, their R_{ee} distributions are narrow and have only one peak, which indicates that the GAGs are only mildly flexible as isolated monomers. This, however, changes upon the addition of $A\beta_{16-22}$ to the GAGs. The R_{ee} distributions in Figure 76 become broader due to a higher population of lower R_{ee} values. This is also visible in the most populated GAG structures in the presence of $A\beta_{16-22}$ (Figure 77B) which display curves or kinks in their conformations. The chondroitin sulfates appear to be more affected by the presence of $A\beta_{16-22}$ than the nonsulphated HA and the structurally similar H6S. It should be noted that some of the R_{ee} values almost reached the dimension of the simulation box size, which we had chosen as 4.9 nm per edge. To rule out that our results are affected by artefacts arising from interactions of the solutes with their periodic images, we calculated their minimum distance and found that they are all $>$ 1.2 nm, i.e., above the cutoff chosen for the short-range nonbonded interactions. We can thus conclude that our results do not suffer from such artificial interactions.

Some of the R_{ee} values are even smaller than 3.0 nm, indicating that the GAGs can deviate considerably from the linear conformation by forming loops or coils [338]. Coil structures of the four GAGs corresponding to their minimal R_{ee} values are provided in Figure 77C, along with $A\beta_{16-22}$ in interaction with the GAGs. The C4S and C6S structures deviate the most from the linear conformation and adopt a spherical shape. H6S, on the other hand, forms only small coils at the termini, while the structure of HA represents the transition between the linear and spherical-shape conformations. The structures in Figure 77 already suggest that electrostatic interactions

between Aβ_{16-22} and the GAGs are likely to play role in their binding since K16 is very often orientated toward a sulphate or carboxylate group.

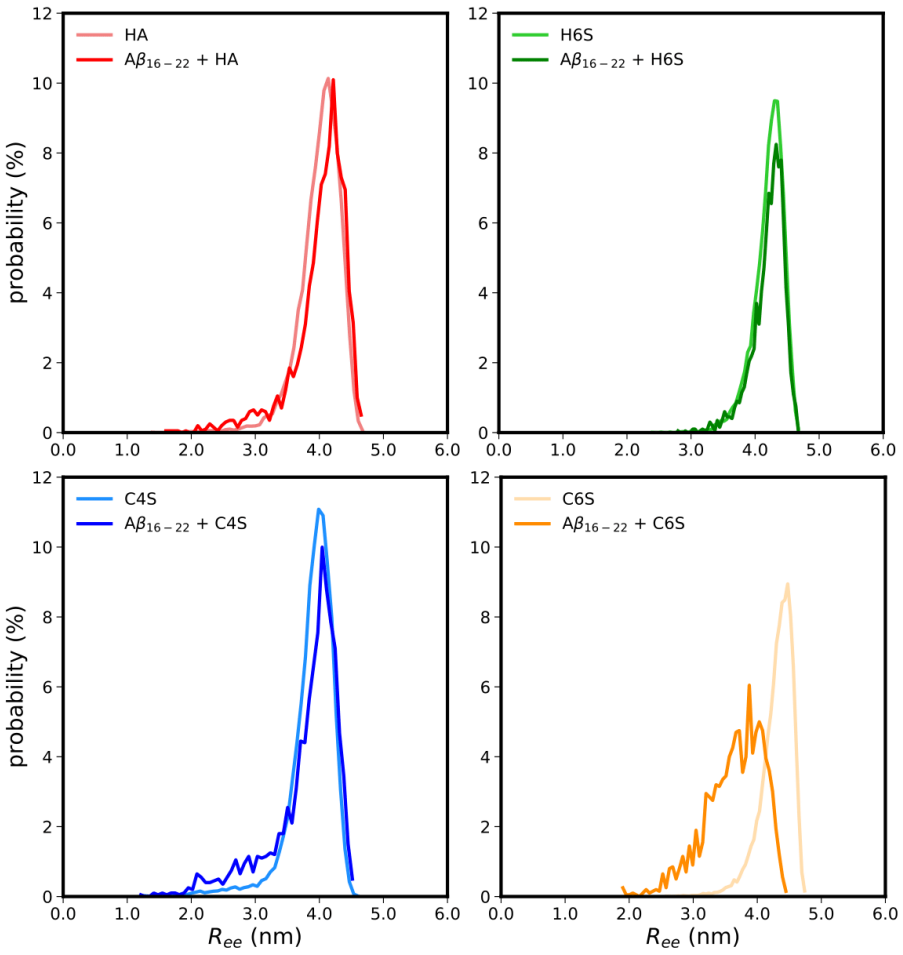

Figure 76: Distribution of the end-to-end distance R_{ee} of the GAGs in the absence (light colors) and presence (darker colors) of Aβ_{16-22} when simulated at a 1:1 stoichiometry. The colors correspond to the GAGs as provided in the legends inside the plots.

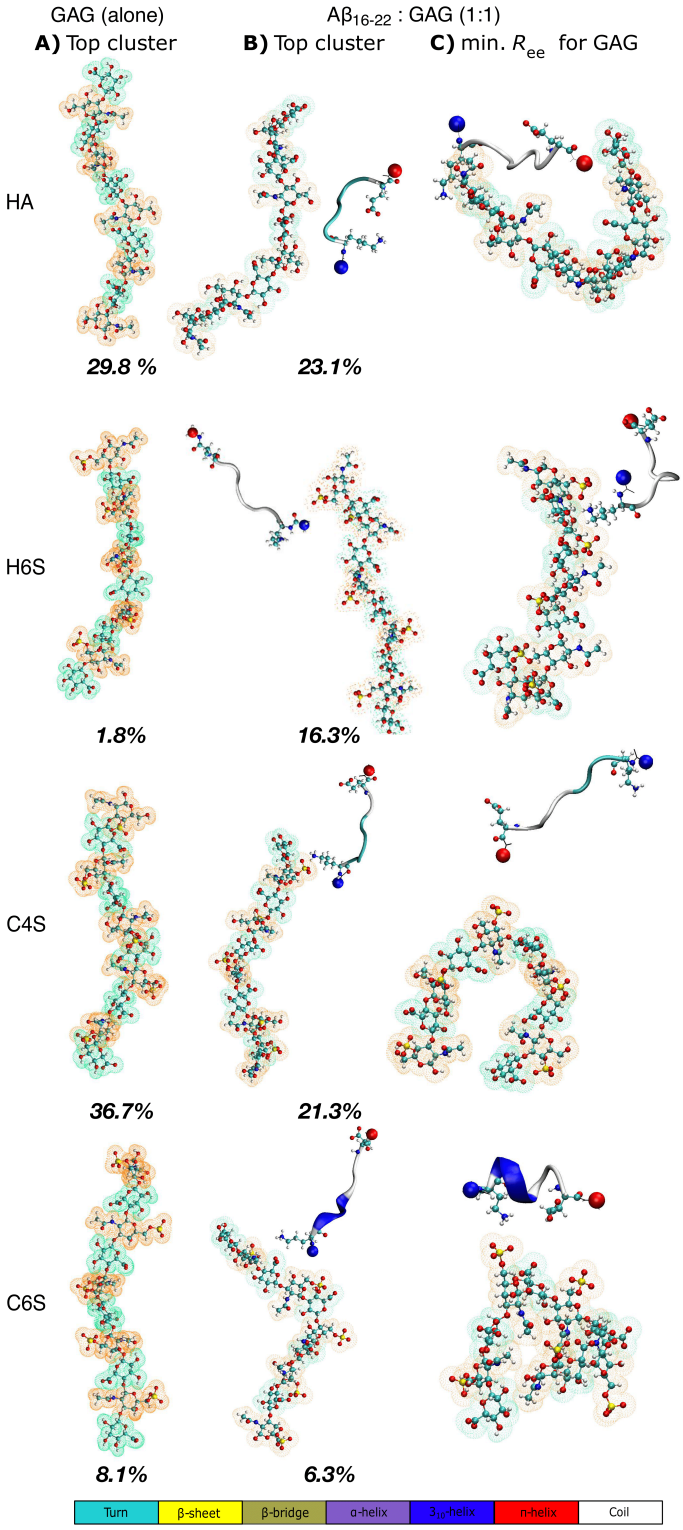

Figure 77: Snapshots of GAG and $A\beta_{16-22}$-GAG (1:1) configurations. (A) The most populated GAG structures without $A\beta_{16-22}$ are extended conformations. (B) Extended GAG conformations, yet with curves or kinks, also prevail in the presence of $A\beta_{16-22}$. The populations of these conformations are given. (C) The interaction with $A\beta_{16-22}$ can further induce coil GAG conformations, leading to small R_{ee} values in the GAGs. The GlcUA residues are highlighted in green for all the GAGs, whereas the orange color reflects the GlcNAc and GalNAc residues. $A\beta_{16-22}$ is represented as cartoon and colored according to its secondary structure (see color key at the bottom). The side chains of K16 and E22 are shown explicitly, and the spheres colored in blue and red highlight the N- and C-terminus, respectively.

To further characterize the GAG structures, we calculated the distribution of the dihedral angles on either side of the oxygen atoms linking adjacent monosaccharides and combined them for each linkage type into an offset function according to equation 30 on page 137. The results in Figure 78 show that without $A\beta_{16-22}$ being present, the GAGs experienced limited flexibility across either of the linkages. The only exception is $Linkage_2$ in H6S, which however is not connected with severe deviations from the linear conformation as the results for R_{ee} revealed. However, the motions around $Linkage_2$ lead to more conformational clusters and a lower probability of the highest populated cluster (provided in Figure 77A). The preference of H6S to remain in an elongated state must thus result from the $\alpha(1\rightarrow4)$ linkage for $Linkage_1$, which is present in H6S and is one of the key differences between this and the other three GAGs. This is supported by the fact that D_{offset} for $Linkage_1$ of H6S hardly changed when $A\beta_{16-22}$ was added to the system, while some differences were recorded for the other three GAGs (Figure 79). Changes in D_{offset} upon addition of $A\beta_{16-22}$ are also visible for $Linkage_2$, especially in the case of C6S. This is in agreement with the observations made for R_{ee} and the spherical-shape conformation that C6S can adopt.

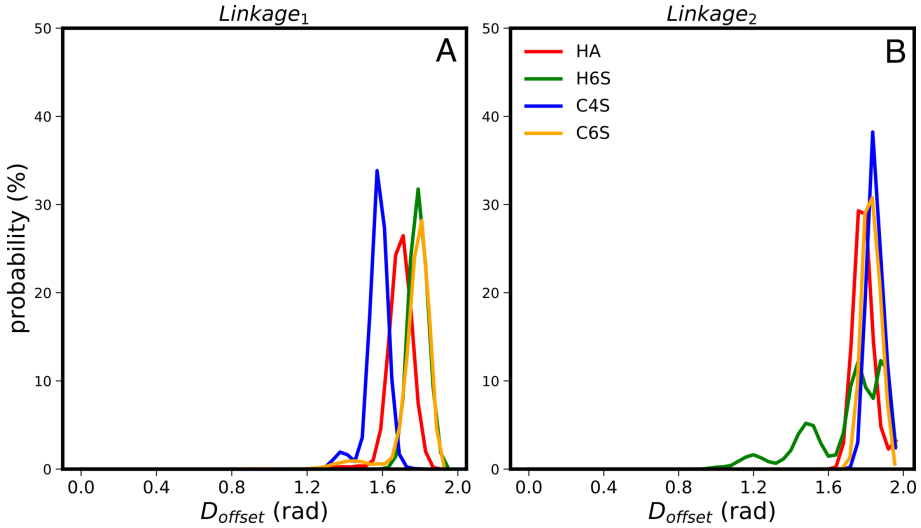

Figure 78: Distribution of the dihedral angles across the two glycosidic linkage types given as collective dihedral offset function (D_{offset}) for the different GAGs in the absence of $A\beta_{16-22}$. The color key for the GAGs is given inside the right panel.

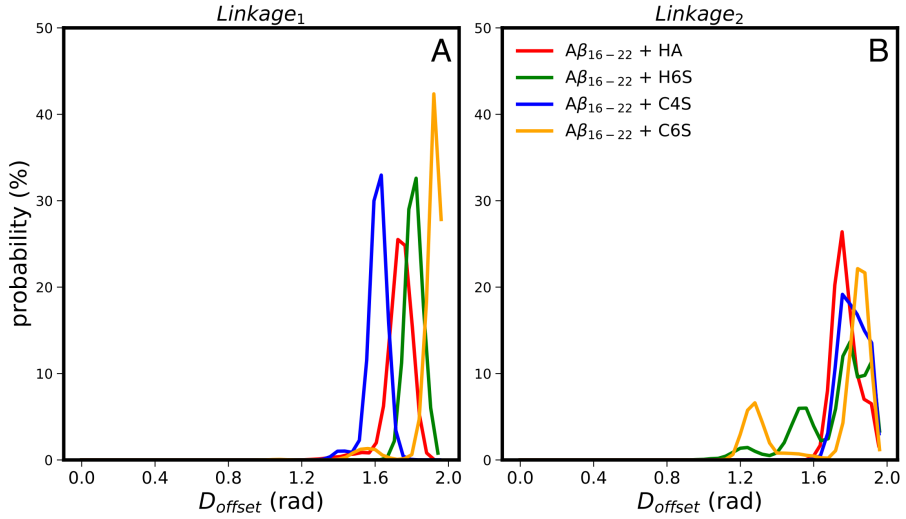

Figure 79: Distribution of the dihedral angles across the two glycosidic linkage types given as collective dihedral offset function (D_{offset}) for the different GAGs in the presence of $A\beta_{16-22}$. The color key for the GAGs is given inside the right panel.

In summary, it can be concluded that $A\beta_{16-22}$ enables the GAGs to abandon the linear conformation and transition into more coil conformations. However, this tendency is small for H6S. Moreover, the various quantities calculated here reveal that none of the GAGs remained in collapsed states; instead, extended conformations are still preferred.

7.3.1.2 *Effects of GAGs on* $A\beta_{16-22}$

In order to test whether the GAGs affect the structural preferences of the $A\beta_{16-22}$ peptide, we calculated R_{ee} and R_g of $A\beta_{16-22}$ in bulk solution and in the presence of GAGs. Noteworthy effects on the end-to-end distance of the peptide are only observed for the C6S system, causing the R_{ee} distribution to move toward smaller values (Figure 80A). The other GAGs, on the other hand, allow $A\beta_{16-22}$ to mainly stay in an extended state. This conclusion is supported by the intrapeptide contacts, which only in the case of C6S show considerable contacts between residues in addition to the contacts resulting from the closeness of residues along the sequence (Figure 81). However, the distribution of the R_g values reveals that $A\beta_{16-22}$ is also influenced by the other GAGs as in the presence of either GAG the radius of gyration is smaller than for $A\beta_{16-22}$ alone (Figure 80B). Nonetheless, also here the change in R_g is most distinct in the case of C6S. The analysis of the secondary structure shows that $A\beta_{16-22}$

adopts less β-strand conformations in the presence of the GAGs, causing the amount of α-helix or random coil to increase (Figure 80C). The amount of β-strand decreases in the order C6S > C4S > HA ≈ H6S. When accompanied by C4S and C6S, $A\beta_{16-22}$ even transiently adopts α-helical structures with probabilities above 10%. It should be mentioned that this affects only isolated residues and not the whole peptide, i.e., the tendency of a residue to be in the α-helical region of the Φ, Ψ-space manifests itself in turn formations. The snapshots for the $A\beta_{16-22}$–C6S complexes in Figure 77B and C illustrate this tendency as $A\beta_{16-22}$ adopts 3_{10}-helical structures here.

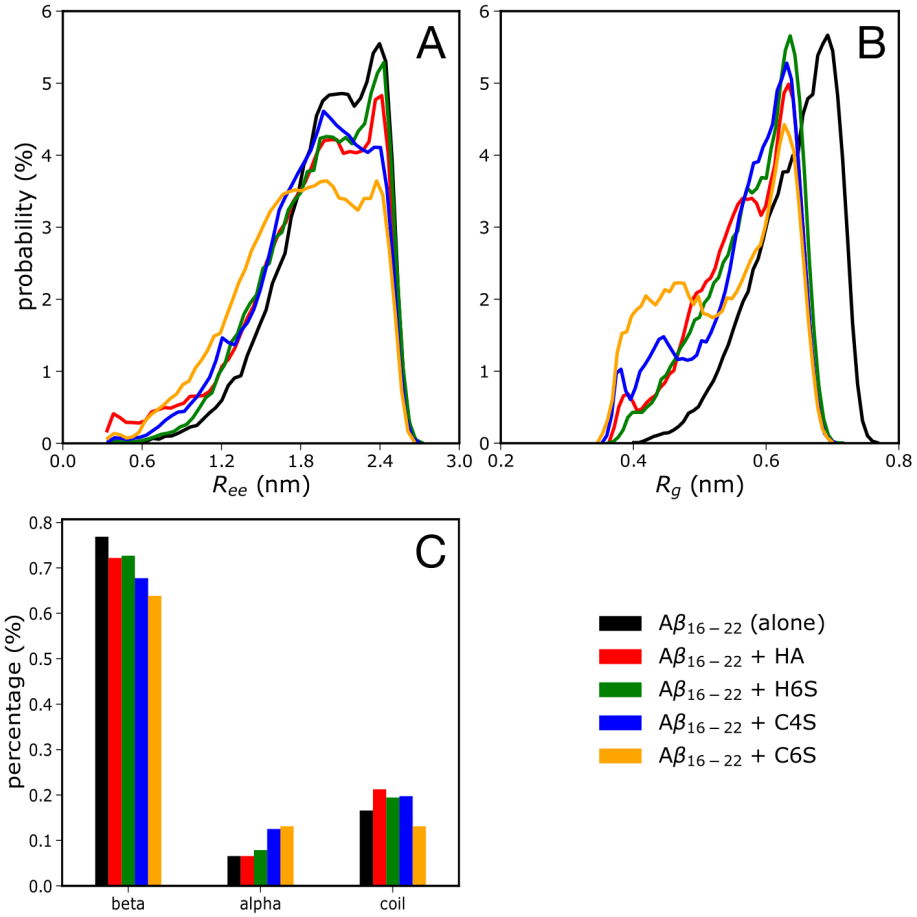

Figure 80: Distribution of (A) the end-to-end distance, R_{ee}, and (B) the radius of gyration, R_g, as well as (C) the time-averaged secondary structures classified into coil, turn, β-sheet, and α-helix for the $A\beta_{16-22}$ monomer alone (black) and in presence of HA (red), H6S (green), C4S (blue), and C6S (yellow) as obtained from the 1:1 simulations.

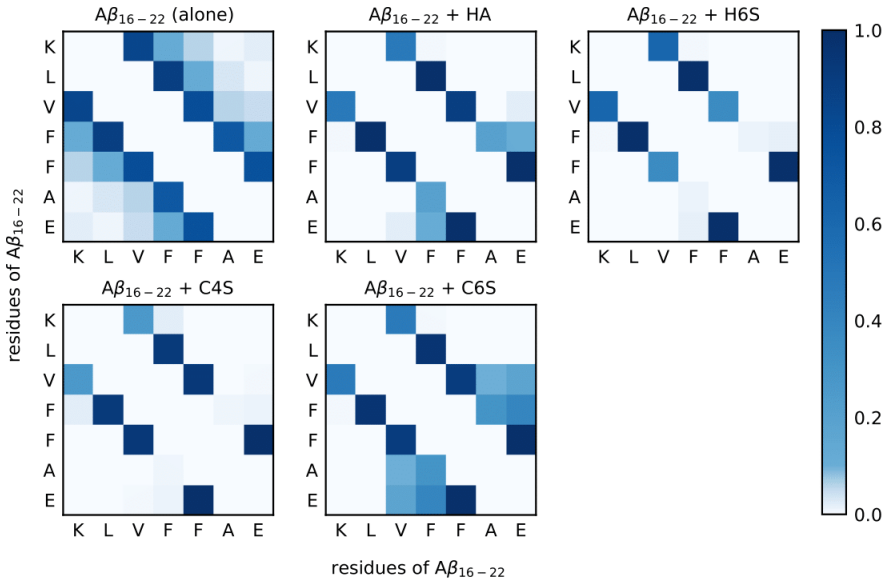

Figure 81: Probability of intrapeptide contacts between the residues of $A\beta_{16-22}$ monomers alone (top left) and in the presence of GAGs. The color code on the right represents the probability of a contact between residues during the MD simulations. For the sake of clarity, the diagonal and first off-diagonal elements of the contact maps corresponding to self-contacts and contacts with direct neighbors are not shown.

7.3.1.3 *Classification of the interactions between $A\beta_{16-22}$ and GAGs*

The mutual effects of $A\beta_{16-22}$ and the GAGs can be rationalized by dissecting their interactions. A first impression of their interplay is obtained from the contact probability maps between the $A\beta_{16-22}$ residues and GAG monosaccharide units (Figure 84). A clear preference for interactions between K16 and the GAGs is observed, where GlcNAc of HA, and the sulphated GlcNAc of H6S and GalNAc of C4S and C6S form slightly more contacts with $A\beta_{16-22}$ than GlcUA does. Given that K16 is positively charged and GlcUA as well as the sulphated residues of H6S, C4S and C6S are negatively charged, electrostatic attraction between $A\beta_{16-22}$ and the GAGs seem obvious. This conjecture is corroborated by the snapshots in Figure 77, where K16 is often found being oriented toward one of the sulphate groups in H6S, C4S and C6S. Since this interaction is not possible with the nonsulphated HA, the $A\beta_{16-22}$–HA interaction is overall weaker. Support is further provided by the decomposition of the intermolecular interaction energies into their electrostatic and LJ contributions (Figures 82 and 83).

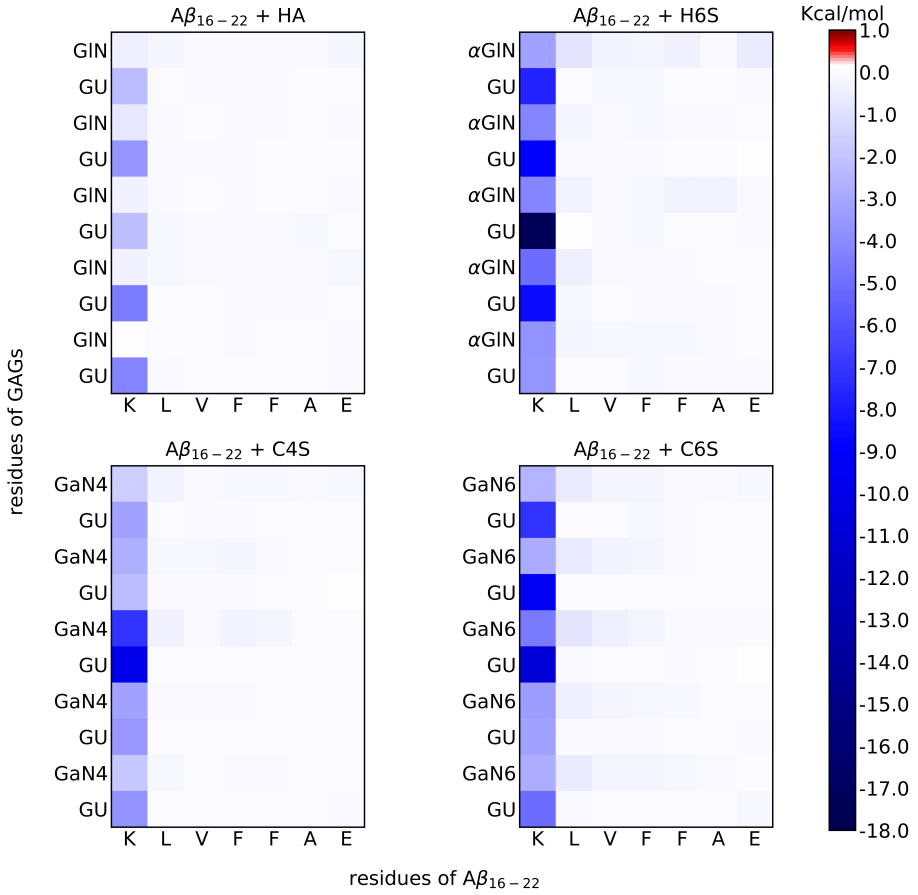

Figure 82: Average electrostatic interaction energies between the residues of $A\beta_{16-22}$ and the monosaccharide units of the different GAGs from the 1:1 simulations. The residues of the GAGs are abbreviated with the following codes: GU (GlcUA), GlN (GlcNAc in HA), αGlN (GlcNAc(6S) in H6S), GaN4 (GalNAc(4S)), and GaN6 (GalNAc(6S)). The interaction energies (kcal mol^{-1}) are according to the color key on the right.

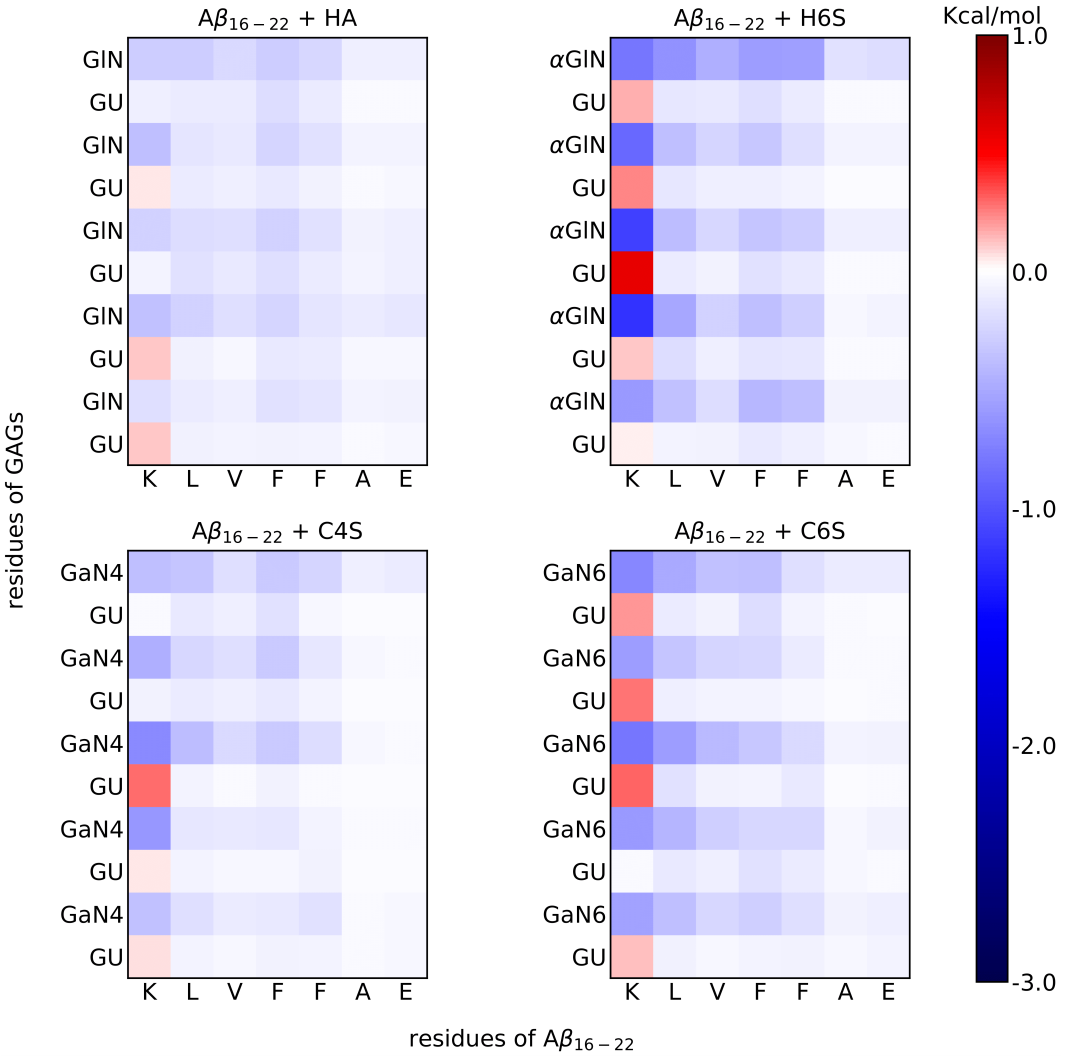

Figure 83: Average LJ interaction energies between the residues of $A\beta_{16-22}$ and the monosaccharide units of the different GAGs from the 1:1 simulations. The residues of the GAGs are abbreviated with the following codes: GU (GlcUA), GlN (GlcNAc in HA), αGlN (GlcNAc(6S) in H6S), GaN_4 (GalNAc(4S)), and GaN6 (GalNAc(6S)). The interaction energies (kcal mol^{-1}) are according to the color key on the right.

For all four GAGs the electrostatic interactions are dominated by the attraction of K16 to the GAGs. However, while the contact probabilities are somewhat higher for GlcNAc and GalNAc compared to GlcUA, the electrostatic attractions are stronger for the latter. This is the case for all four GAGs and can be explained with the negative charge of this monosaccharide unit (Figure 3 on page 13). However, these attractions are partly counteracted by repulsive LJ interactions between K16 and GlcUA, which result from close contacts between K16 and GlcUA. Between K16 and GlcNAc or GalNAc electrostatic attractions are also present, which are reinforced by attractive LJ interactions, which explains the slight preference for these monosaccharides as com-

pared to GlcUA. While the electrostatic interactions are largely limited to K16, the LJ interactions involve also the other residues apart from A21 and E22, especially in the case of H6S and C6S. In the case of the latter, these interactions lead to considerable effects on both the $A\beta_{16-22}$ and GAG structure. The interaction patterns involving HA and C4S are similar to that of C6S but generally weaker, explaining their smaller mutual effects. The interactions between H6S and $A\beta_{16-22}$ are of similar strength as in the case of C6S, but the mutual effects are minor.

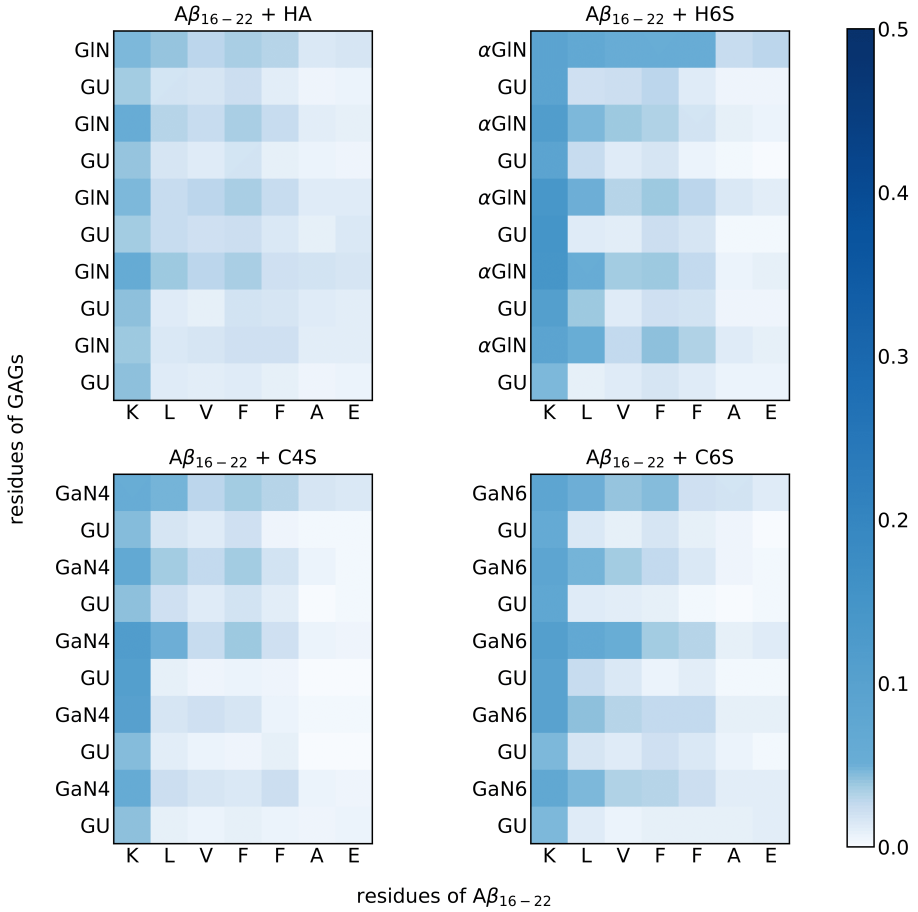

Figure 84: Probability of intermolecular contacts between the residues of $A\beta_{16-22}$ and the monosaccharides of the different GAGs obtained from the 1:1 simulations. The residues of the GAGs are abbreviated with the following codes: GU (GlcUA), GlN (GlcNAc in HA), αGlN (GlcNAc(6S) in H6S), GaN4 (GalNAc(4S)), and GaN6 (GalNAc(6S)). The color code on the right represents the contact probability between residues of GAG and $A\beta_{16-22}$.

7.3.2 *Effects of GAGs on $A\beta_{16-22}$ dimerization and their interactions at 2:2 stoichiometry*

7.3.2.1 *Interpeptide interactions*

We started the analysis of the simulations of the $A\beta_{16-22}$-GAG systems at a 2:2 stoichiometry by elucidating the contacts that occurred between the two $A\beta_{16-22}$ peptides since this gives a first impression of the aggregates that formed (Figure 86). Without GAGs, the two $A\beta_{16-22}$ peptides established many contacts with each other, with no clear interaction pattern being visible. Only a preference for interactions between the hydrophobic core residues V18–A21 can be deduced. This finding is somewhat surprising as from the $A\beta_{16-22}$ fibril structure it is known that the antiparallel arrangement is the preferred one [240]. However, unlike the fibrils, the dimer is still rather mobile and undergoes constant dissociation and reassociation events. When the two monomers associate again, they do so in all kinds of orientations with respect to each other, which explains the presence of different contacts between them. Moreover, the two peptides can even adopt a parallel aligment, which is stabilized by hydrophobic interactions between the hydrophobic residues ^{17}LVFF^{20}, while the like-charged side chains of the terminal residues point into opposite directions to avoid repulsive forces between them (Figure 85).

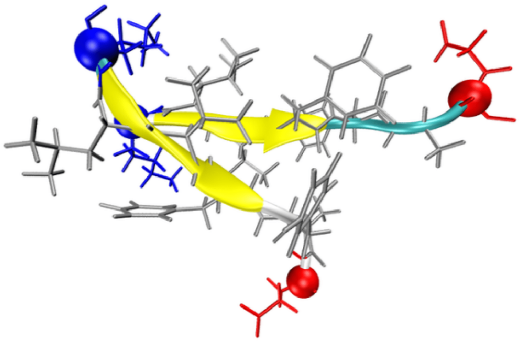

Figure 85: Snapshot of the parallel-aligned $A\beta_{16-22}$ dimer as sampled in the dimerization simulation without any GAG being present. The parallel orientation is stabilized by hydrophobic contacts and avoids destabilization that could arise from Coulomb repulsion by orienting the peptides' charged terminal side chains toward the aqueous solvent.

In the presence of GAG molecules, the interactions between the $A\beta_{16-22}$ peptides are altered. However, the different GAGs exert diverse effects on the aggregation pat-

terns of $A\beta_{16-22}$. The nonsulphated HA seems to attenuate the aggregation process as only few interpeptide contacts formed. In the presence of the sulphated GAGs, on the other hand, the aggregation process becomes more ordered as compared to the aggregation of $A\beta_{16-22}$ without GAGs. This is especially true for C4S, where a well-defined contact pattern of an in-register, antiparallel arrangement appeared. Somewhat less well defined is the contact map for the system involving C6S, yet a preference for a parallel arrangement between the two peptides is clearly visible. In the presence of H6S, the interaction pattern between the two peptides is very similar to the situation in solution, yet generally weaker. Both parallel and antiparallel arrangements seem to be adopted.

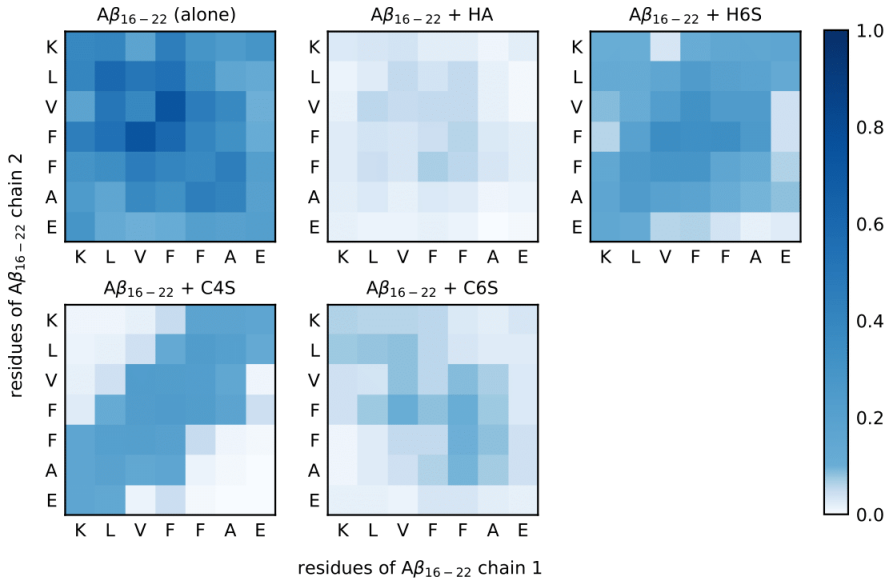

Figure 86: Probability of interpeptide contacts between the residues of the two $A\beta_{16-22}$ peptides as obtained from the simulations of $A\beta_{16-22}$ dimerization without GAGs (upper left panel) or in the presence of GAGs (label on top of each panel) at 2:2 stoichiometry. The color code on the right represents the contact probability between the residues of the two $A\beta_{16-22}$ peptides.

7.3.2.2 *$A\beta_{16-22}$-GAG interactions*

The interpeptide interactions are in competition with interactions occurring between $A\beta_{16-22}$ and the GAGs. Therefore, the latter need to be analyzed too (Figure 87). Interestingly, the contact maps reveal that the interactions between $A\beta_{16-22}$ and the GAGs are generally weak. They are weaker than the interpeptide contacts and also weaker than the contacts that formed between $A\beta_{16-22}$ and the GAGs in the 1:1

systems (Figure 84). It can thus be concluded that the $A\beta_{16-22}$ peptides prefer to interact with each other rather than with the GAGs. The two sulphated GAGs H6S and C6S form somewhat more contacts with $A\beta_{16-22}$ than the other two GAGs, where the nonsulphated GAG HA appears to have the smallest tendency to interact with $A\beta_{16-22}$. This propensity was already visible for the 1:1 systems. Also the $A\beta_{16-22}$-GAG interaction patterns are similar in the 1:1 and 2:2 systems. The positively charged residue K16 is the preferred interaction partner for the GAGs, and it forms slightly more contacts with the monosaccharides GlcNAc and GalNAc as compared to GlcUA. This behavior, which includes also H6S this time, can be explained with the electrostatic and LJ energies between $A\beta_{16-22}$ and the GAGs (Figures 88 and 89). As for the 1:1 systems, the electrostatic interaction between K16 and GlcUA are stronger than for GlcNAc and GalNAc. However, these attractive interactions with GlcUA are partly offset by repulsive LJ interactions, leading to the slight contact preference with GlcNAc and GalNAc. In the interactions with C6S, some other residues than K16 are also involved, giving rise to attractive electrostatic and LJ energies. This indicates a larger contact area between $A\beta_{16-22}$ and this GAG as compared to the other three. It can thus be concluded that the sulphation plays a relevant role in the $A\beta_{16-22}$–GAG interactions. In the case of C4S the sulphation position is concealed near the glycosidic linkage, impeding its interaction with the peptide. Between C6S and H6S the main difference is the α-(1→4) linkage in the latter, which, based on the analysis of the 1:1 simulations, reduces the flexibility of H6S and therefore increases in the 2:2 system the tendency of $A\beta_{16-22}$ and H6S to align next to each other.

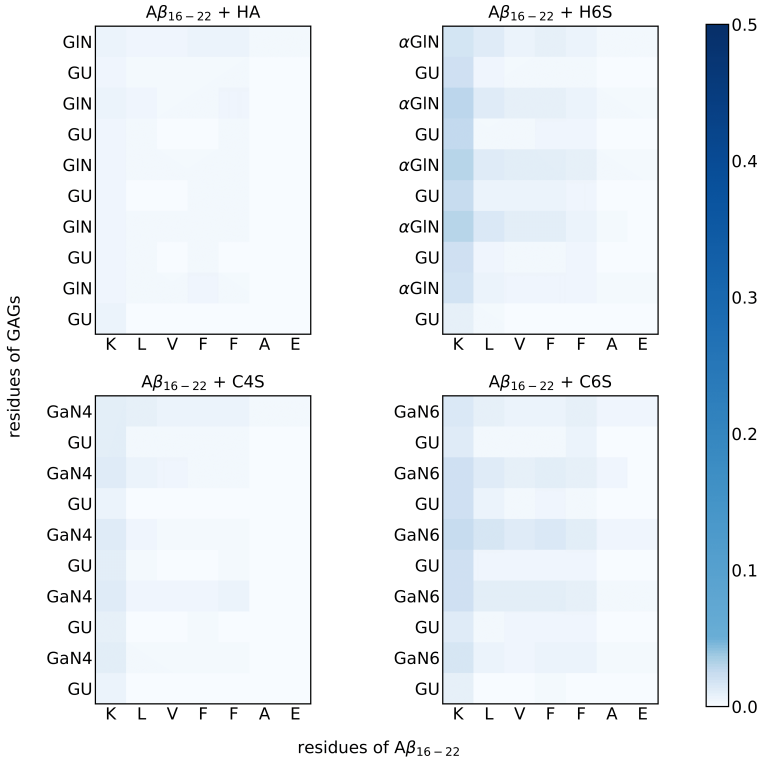

Figure 87: Probability of intermolecular contacts between the residues of $A\beta_{16-22}$ and the monosaccharides of the different GAGs obtained from the 2:2 simulations. The residues of the GAGs are abbreviated with the following codes: GU (GlcUA), GlN (GlcNAc in HA), αGlN (GlcNAc(6S) in H6S), GaN_4 (GalNAc(4S)), and GaN6 (GalNAc(6S)). The color code on the right represents the contact probability between residues of GAG and $A\beta_{16-22}$.

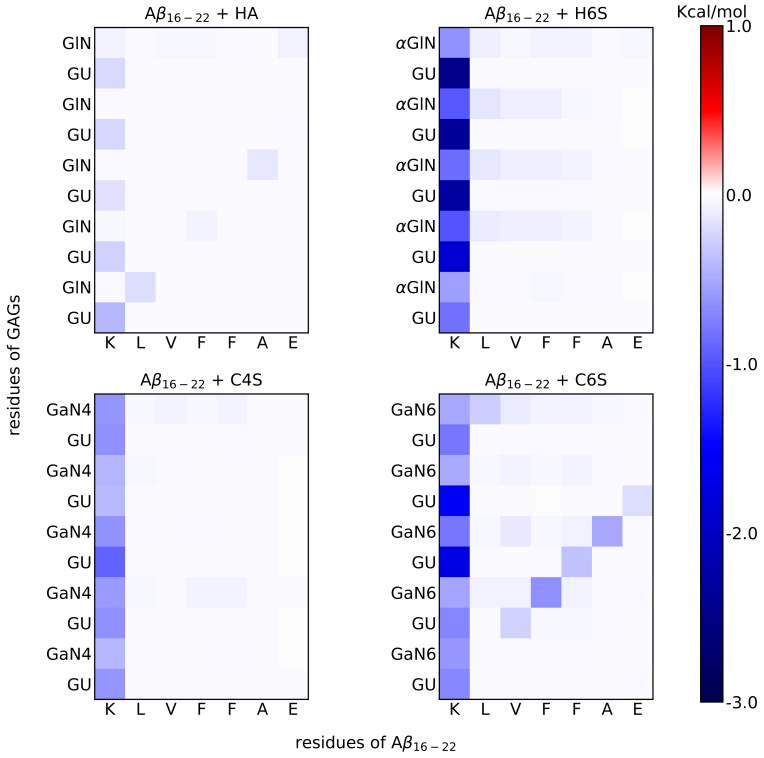

Figure 88: Average electrostatic interaction energies between the residues of $A\beta_{16-22}$ and the monosaccharide units of the different GAGs from the 2:2 simulations (see Figure 87 for the GAG residue abbreviations). The interaction energies ($kcal\,mol^{-1}$) are according to the color key on the right.

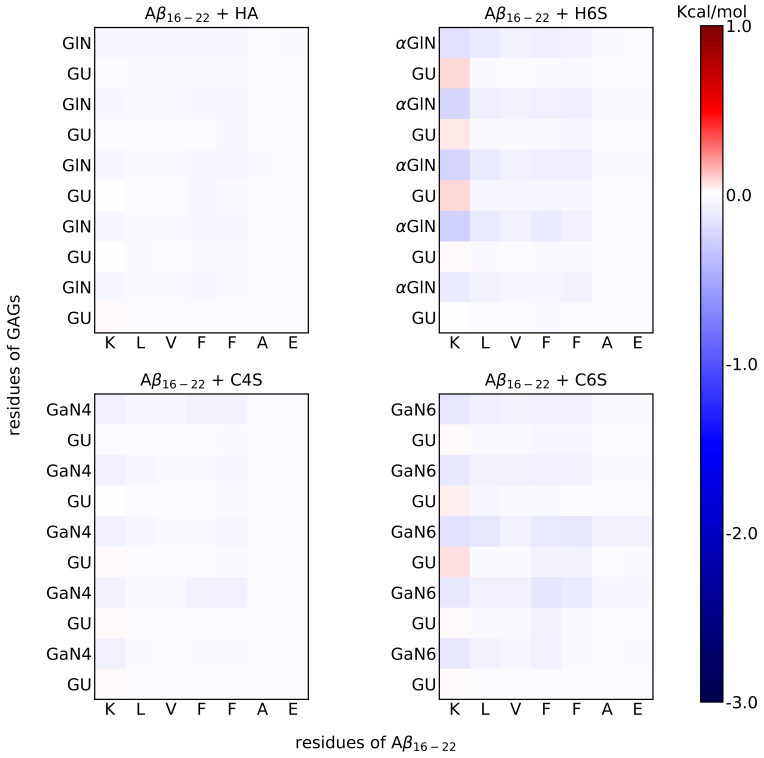

Figure 89: Average LJ interaction energies between the residues of $A\beta_{16-22}$ and the monosaccharide units of the different GAGs from the 2:2 simulations (see Figure 87 for the GAG residue abbreviations). The interaction energies (kcal mol^{-1}) are according to the color key on the right.

7.3.2.3 *Representative $A\beta_{16-22}$-GAG complex structures*

To illustrate the conclusions drawn from the contact analyses, we show and discuss representative conformations from the 2:2 simulations (Figure 90). The first observation is that $A\beta_{16-22}$ and the GAGs are only loosely attached, which agrees to the findings from the contact probability maps. The second observation is that all systems are very dynamic, especially those involving HA and H6S. Therefore, the representative structures have a rather low population. Nonetheless, from inspecting the trajectories we observed recurrent interaction modes. In the case of H6S and C4S, the GAGs and $A\beta_{16-22}$ tend to form a sandwich-like structure, with the antiparallel-aligned $A\beta_{16-22}$ dimer being in the middle and the two GAGs on either side of the dimer. This kind of arrangement is associated with more stable $A\beta_{16-22}$ dimers than found in the 2:2 conformations for HA and C6S (based on the contact probabilities in Figure 86). For C6S, a situation is shown in Figure 90 where each of the $A\beta_{16-22}$ peptides is in interaction with one of the GAG molecules. These interactions are mainly

mediated by K16–sulphate (upper peptide in the figure) and K16–carboxylate (lower peptide in the figure) contacts. This prevents electrostatic repulsion between these two residues when being next to each other, allowing the two peptides to approach each other via their N-termini (as shown in Figure 90) or to adopt parallel orientations. In the case of HA, only one of the $A\beta_{16-22}$ peptides is in contact with one of the HA molecules, while the second peptide is loosely attached to the HA-bound $A\beta_{16-22}$. This situation agrees to the low contact probabilities observed in the corresponding contact maps of the $A\beta_{16-22}$–$A\beta_{16-22}$ and $A\beta_{16-22}$–HA interactions. The differences between HA and H6S are the sulphation of GlcNAc and the α-(1→4) linkage in the latter. Figure 90 reinforces the finding from the 1:1 simulations that H6S, due to its different $\mathtt{Linkage}_1$ geometry, is on average more elongated. This is positively correlated with an increased tendency of $A\beta_{16-22}$ to adopt extended conformations, which was deduced from the secondary-structure analysis of the 1:1 systems and is supported by the representative 2:2 structure for the $A\beta_{16-22}$–H6S system. This effect as well as the preferential interactions with the sulphate groups are missing in the HA system, explaining the differences between the HA and H6S systems. It can be summarized that the sulphate group in H6S, but also that in C4S and in C6S are important for their impact on $A\beta_{16-22}$ aggregation.

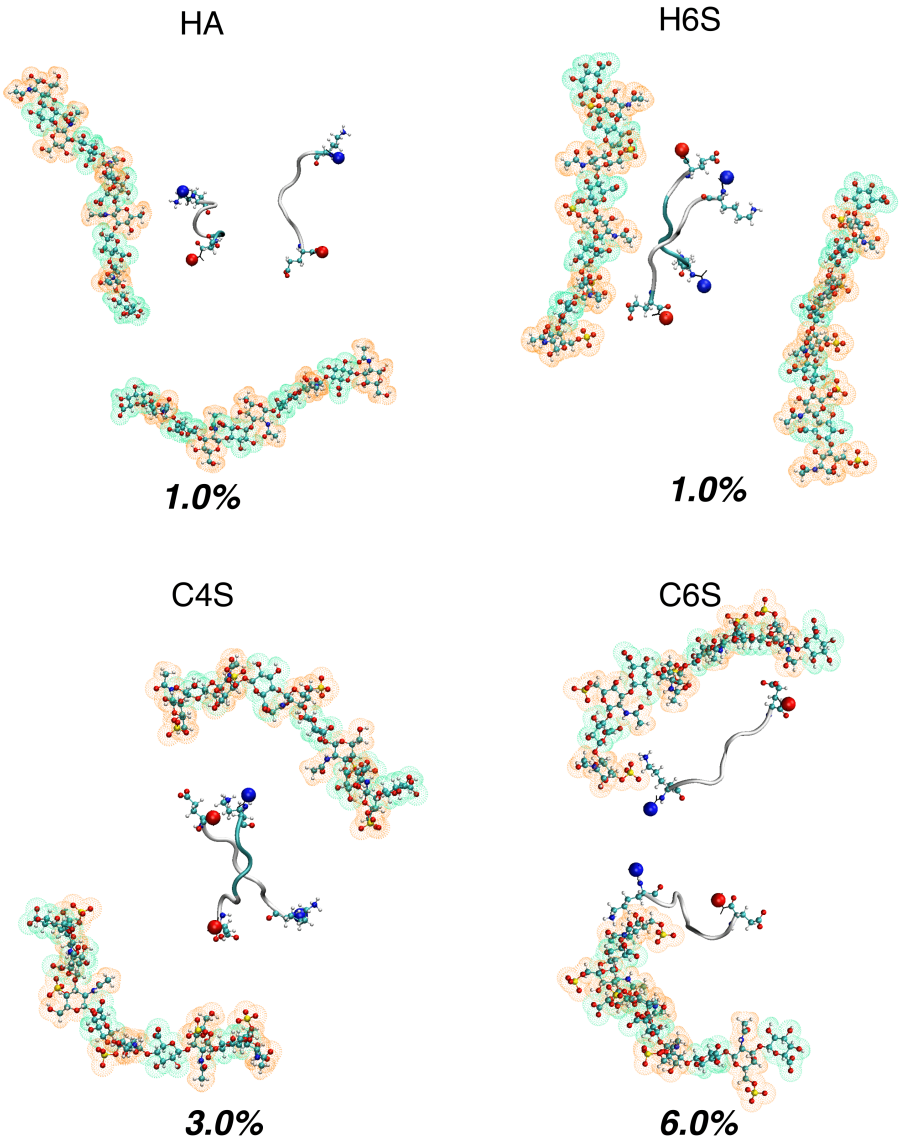

Figure 90: Snapshots showing typical interaction patterns between $A\beta_{16-22}$ and the different GAGs as sampled during the 2:2 simulations. The populations of these structures are provided. See Figure 77 for an explanation of the coloring scheme.

Interestingly, even though the interplay between $A\beta_{16-22}$ and HA is minor, these interactions are sufficient to reduce the self-aggregation of $A\beta_{16-22}$ (Figure 86). In order to understand this effect, we visually inspected the trajectory in detail. A very striking observation is that the whole system is very dynamic. $A\beta_{16-22}$ neither formed stable contacts with HA nor with each other. In fact, throughout the whole 5 µs simulation, it bounced back and forth between HA and the other peptide. Moreover, when HA is moving, $A\beta_{16-22}$ is often pushed away from the HA molecule it was interacting with. It almost never occurred that both peptides interacted with the same HA molecule at the same time, which would have favored aggregation. Instead,

the preferred situations were that i) both peptides are in solution but not close to each other, or ii) one of them was transiently interacting with one of the HA molecules. To illustrate the dynamics of this system, we show some simulation snapshots in Figure 91.

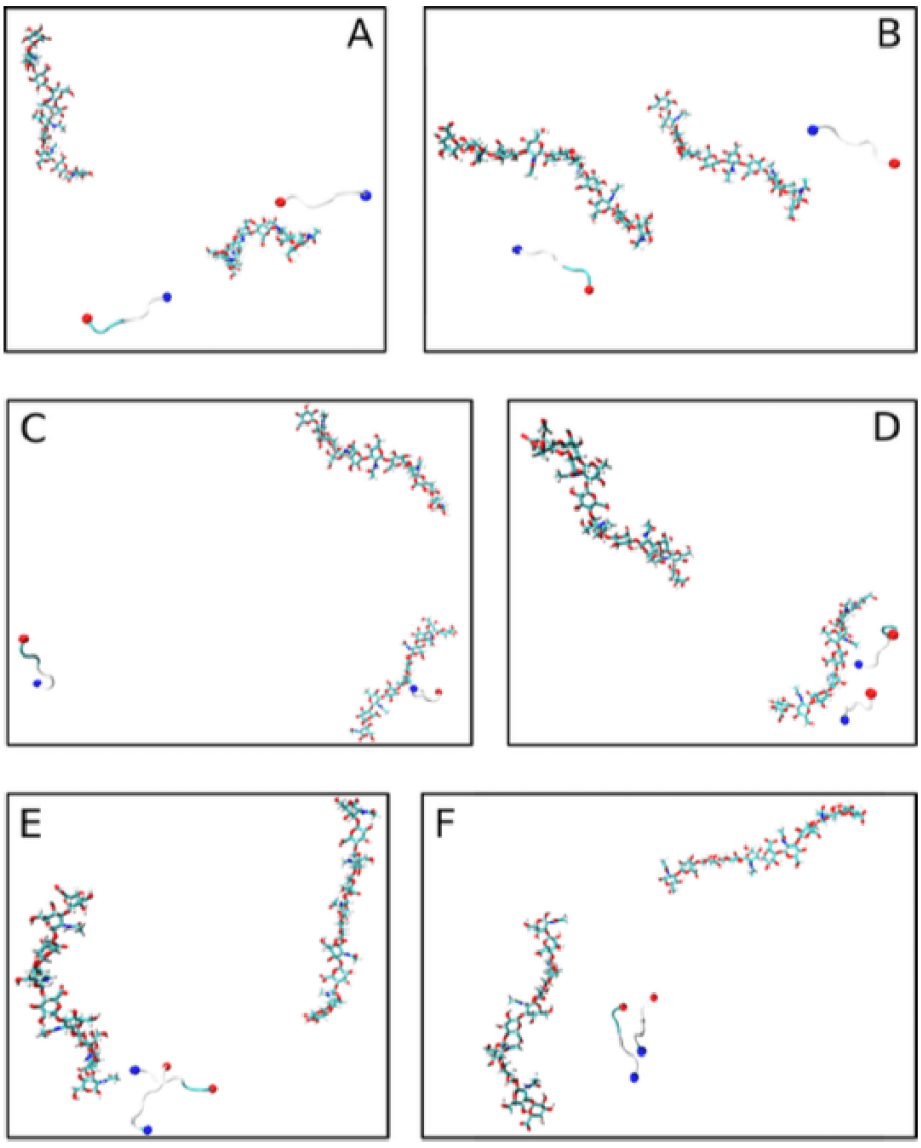

Figure 91: Illustrative snapshots to demonstrate the dynamics of the 2:2 $A\beta_{16-22}$-HA system. (A) and (B) show the most common situation where both peptides neither interact with themselves nor with HA. (C) Another common situation is where $A\beta_{16-22}$ interacts with one of the HA molecules while the other peptide and HA molecule are in solution. (D) A very rare case is that both peptides interact with the same HA molecule. (E) The more typical situation is that one $A\beta_{16-22}$ peptide is in interaction with one HA molecule while the second peptide associated with it. (F) Another dimerization motif that was observed is that the peptides form a dimer while not interacting with any of the HA molecules.

7.3.2.4 *Aggregation pathways*

In order to obtain a better overview of the aggregation pathways and the conformations sampled along with them, we built transition networks (TNs) for each of the 2:2 systems and the $A\beta_{16-22}$ dimerization without a GAG being present. For the calculation of TNs, one needs to define and determine descriptors that are able to provide a clear picture of the process under study [341]. Here, we used the following descriptors: (i) the oligomer size (OS) which can be 1 (monomer) or 2 (dimer); (ii) the number of hydrophobic atom–atom contacts between the two peptides (HC); (iii) the compactness (C), which is defined as the ratio between the lowest and highest moment of inertia, multiplied by 10 and rounded to the nearest integer, leading to a possible value range of $[0, 10]$, where a value of zero corresponds to a stick and a value of ten represents a sphere; (iv) the peptide-averaged number of residues that are part of the β-sheet (β) [341]. This selection of descriptors allows us to assess the physicochemical properties during the intermediate stages of $A\beta_{16-22}$ dimer formation.

7.3.2.5 *TN for $A\beta_{16-22}$ dimerization without GAGs*

The TN for the dimerization of $A\beta_{16-22}$ without any GAG (Figure 92) confirms the conclusion from our earlier results [3, 341]. The formation of antiparallel β-sheets is the dominant motif during this aggregation process. These β-sheets can evolve from a so-called encounter complex which forms as a result of electrostatic attraction between K16 of one of the peptides and E22 of the other peptide (green node). In-register and out-of-register β-sheets emerged. The latter are shifted by only one residue which still allows for salt-bridge formation between the termini (blue, orange and magenta nodes). The formation of only one salt bridge prevails, even though in most of the antiparallel β-sheets two K16–E22 salt bridges are in principle possible. This includes the in-register β-sheet (cyan node) where the side chains of one of the K16–E22 pairs rather point away from each other and interact with surrounding water molecules instead. In addition to the electrostatic interactions between the terminal residues, the in-register, antiparallel β-sheet gains in stability over the out-of-register

geometry by forming more hydrophobic contacts, the number of which grows when going from left to right in the TN representation, as well as H-bonds. Figure 93A shows that on average five backbone H-bonds are present between the two peptides in the antiparallel β-sheet that prevails toward the end of this 5 μs simulation.

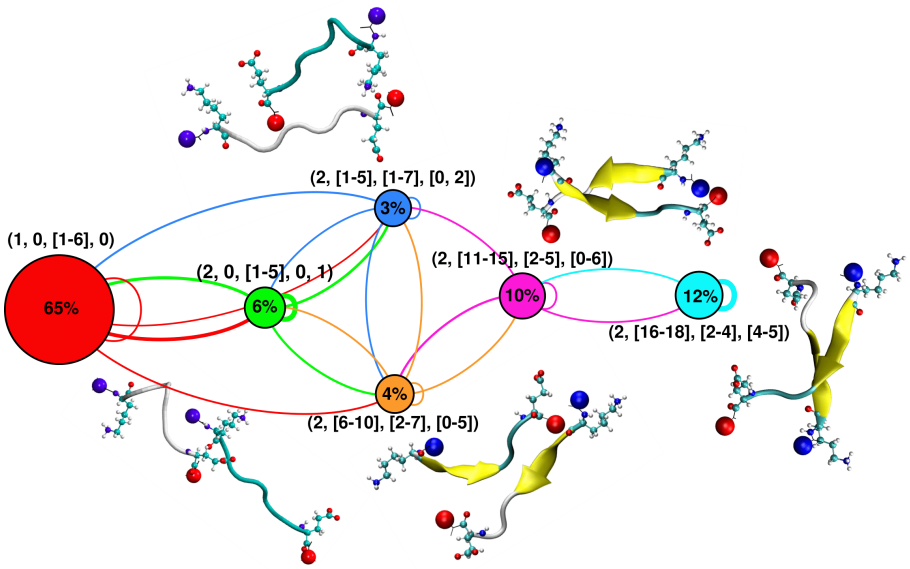

Figure 92: Transition network for the dimerization of $A\beta_{16-22}$ monomers in the absence of GAGs. The nodes are characterized by following descriptors: (OS,HC,C,β), where some of the descriptor values are combined into ranges to reduce the number of nodes. The node color is based on the number of hydrophobic contacts: monomer with $HC = 0$ (red), dimer with $HC = 0$ (green), $1 \leqslant HC \leqslant 5$ (blue), $6 \leqslant HC \leqslant 10$ (orange), $11 \leqslant HC \leqslant 15$ (magenta), $HC > 15$ (cyan). The size of the nodes and thickness of the edges represent their probabilities. In the case of the nodes, the probabilities are also provided as numbers. For the nodes corresponding to dimers, representative snapshots are shown, where the peptides are displayed as cartoon and colored according to their secondary structure (see color key in Figure 77). The spheres colored in blue and red represent the N- and C-terminus, respectively.

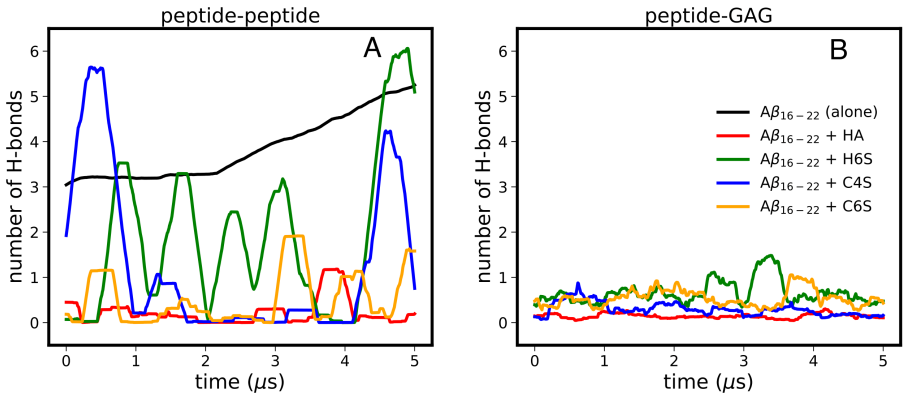

Figure 93: Evolution of the number of (A) interpeptide and (B) peptide–GAG H-bonds. The color key for the systems is given inside the right panel. The graphs are smoothed by showing moving averages. Therefore, the dissociation events in the system without GAGs are not visible in the corresponding graph for interpeptide H-bonds. They would correspond to zero H-bonds as the peptides are present as two monomers when being dissociated.

7.3.2.6 *TN for Aβ$_{16-22}$ dimerization in the presence of HA*

The TN for HA confirms that this GAG disfavors the aggregation of Aβ$_{16-22}$ (Figure 94). The node representing monomers (red) has a population of 92%. Without GAGs, this node has a population of only 65%. However, these numbers should be treated with care as the simulation of the dimerization process without GAGs involved a considerably smaller simulation box (Table 10). The reason for this is that the presence of the GAGs requested larger simulation boxes while the simulations without GAGs were already performed as part of one of our other studies [3]. Since the size of the simulation box is expected to influence the equilibrium between monomeric and dimeric states, it is safer to only compare the structures obtained from the simulations with and without GAGs. Nonetheless, among the different GAG systems, the node populations can still be compared with each other. In the presence of HA, four nodes corresponding to dimers are present. One of them (green) is for the encounter complex where the only contact is between the oppositely charged termini of the two peptides. Two of the nodes (orange and magenta) are for dimers involving β-sheets in a parallel alignment. As explained above, the parallel alignment is supported by the interaction of K16 with the neighboring GAG molecules, preventing the repulsion between the two K16 residues. The fourth dimer conformation (blue node) is rather unstructured, yet with transient β-sheet formation. Overall, in the

presence of HA not only the dimerization, but also the formation of antiparallel β-sheets, i.e., the usual aggregation pattern of $A\beta_{16-22}$ [3, 341] is attenuated. This is further supported by the fact that the numbers of interpeptide hydrophobic contacts and H-bonds (Figure 93A) are lower than in all other systems studied here. Interestingly, also the number of H-bonds between $A\beta_{16-22}$ and HA (Figure 93B) is small – the smallest among the four GAG systems –, underpinning the conclusion that in this 2:2 system all intermolecular interactions are reduced.

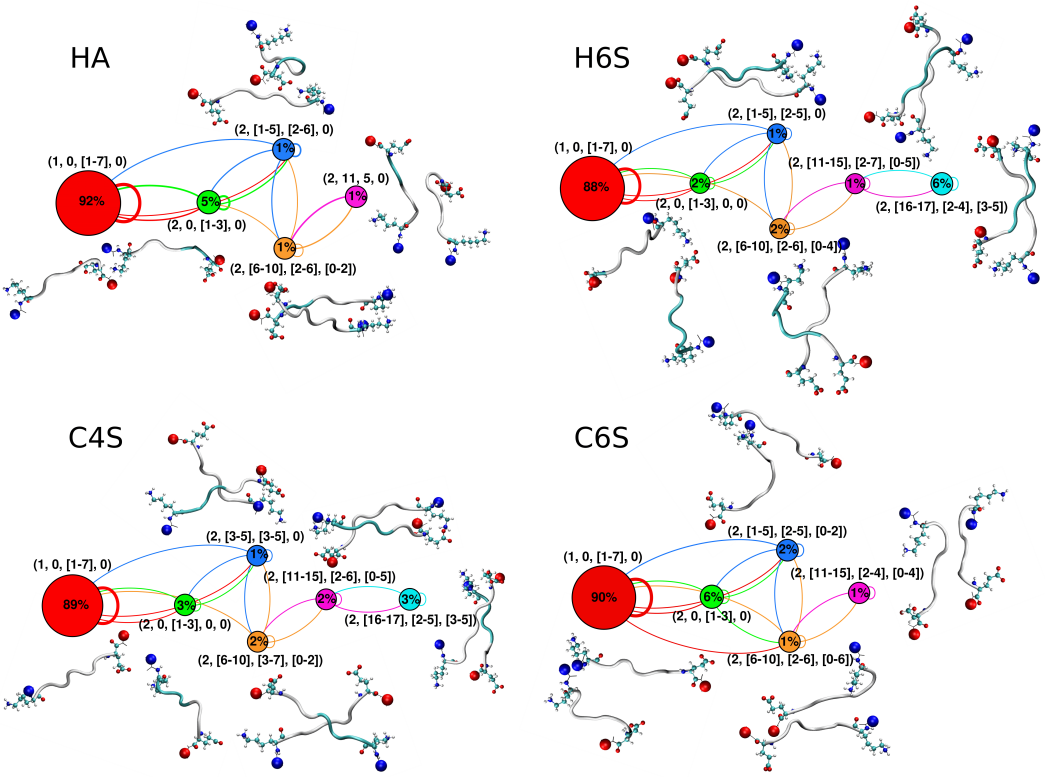

Figure 94: Transition networks for the dimerization of $A\beta_{16-22}$ monomers in the presence of GAGs (see labels). The nodes are characterized by following descriptors: (OS,HC,C,β), where some of the descriptor values are combined into ranges to reduce the number of nodes. For further explanations, see Figure 92.

7.3.2.7 *TN for $A\beta_{16-22}$ dimerization in the presence of H6S*

In the presence of H6S, aggregation is more favored than in the company of HA. Thus, the sheer presence of a sulphate group in a GAG has an influence on $A\beta_{16-22}$ aggregation. H6S encourages the formation of both antiparallel (magenta and cyan node with a combined population of 7%) and parallel alignments (orange and blue node with a combined population of 3%). Despite the perfect alignments, the ten-

dency of β-sheet formation in the dimers with H6S being present is reduced. This can be explained with the competition of building H-bonds between $A\beta_{16-22}$ and H6S or between the two peptides (Figure 93). Nonetheless, at times more than five interpeptide H-bonds are formed, which are, however, only transient and thus less stable than in the $A\beta_{16-22}$ dimer in bulk water.

7.3.2.8 *TN for $A\beta_{16-22}$ dimerization in the presence of C4S*

The contact map revealed that C4S is the GAG which exerts the largest effect on the ordering of the $A\beta_{16-22}$ dimer, leading to the preferential formation of an in-register, antiparallel alignment of the two peptides (Figure 86). The same finding reflects itself in four of the five nodes corresponding to dimers in the TN for the C4S system. Only the green node corresponds to different dimer conformations, namely the encounter complexes, while all other nodes harbor antiparallel structures with increasing amounts of hydrophobic contacts when going from left to right in the TN representation. However, in several of these dimers the amount of β-sheet is quite low, which is supported by the very low number of interpeptide H-bonds (on average, less than one) that formed for $1 < t < 4$ μs (with t being the simulation time, Figure 93A). As for H6S, more such H-bonds are possible ($t < 1$ μs and $t > 4$ μs), yet they are transient. Interestingly, the number of H-bonds that $A\beta_{16-22}$ establishes with C4S is also low.

7.3.2.9 *TN for $A\beta_{16-22}$ dimerization in the presence of C6S*

Like HA, C6S encourages the formation of a parallel alignment between the two $A\beta_{16-22}$ peptides. This is best seen in the corresponding contact map (Figure 86), but also visible in the structures shown along with the TN. Even the representative encounter complex (green node) has the two K16 residues next to each other. This would usually lead to an electrostatic repulsion, which is however avoided due to the synchronous interaction of K16 with C6S. The analysis of the interpeptide electrostatic interactions (Figure 95) supports that the neighboring K16 residues do not give rise to Coulomb repulsion. The energy map is similar in the presence of HA, while for the H6S and C4S systems and especially for the one without GAGs the elec-

trostatic attraction between K16 and E22 is the dominant force. Without GAGs, this interaction has the main ordering effect on the resulting $A\beta_{16-22}$ dimers, while in the presence of GAGs the ordering effect stems from the interaction with the GAGs and the interaction between the hydrophobic-core residues. Also similar to the HA system, the C6S system lacks the node corresponding to HC > 15. Since these are the two systems with the highest tendency to form parallel alignments, the conclusion is that this alignment reduces the hydrophobic contact area. A possible explanation – supported by both the contact probability maps and representative dimer conformations – is that the like charges cause the neighboring termini to point away from each other, which in turn also reduces the contact probability between the directly following L17–L17 and A21–A21 pairings. The presence of C6S, like that of the other GAGs and for the same reasons, reduces the amount of β-sheet in the $A\beta_{16-22}$ dimer. In fact, after HA, C6S leads to the second strongest depletion in interpeptide H-bonds, which are however not compensated by $A\beta_{16-22}$–C6S H-bonds (Figure 93).

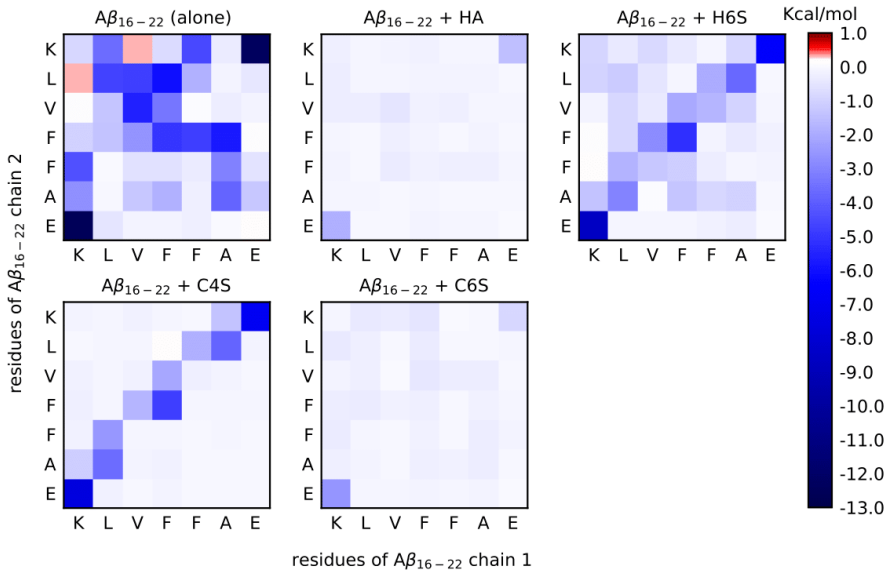

Figure 95: Average interpeptide electrostatic energies between the residues of the two $A\beta_{16-22}$ peptides as obtained from the simulations of $A\beta_{16-22}$ dimerization without GAGs (upper left panel) or in the presence of GAGs (label on top of each panel) at 2:2 stoichiometry. The interaction energies (kcal mol^{-1}) are according to the color key on the right. The red and blue colors denote repulsive and attractive interactions, respectively.

7.4 CONCLUSIONS

In this chapter, we studied the influence of various GAGs on the conformational preference and dimerization of $A\beta_{16-22}$, which contains the hydrophobic core of the Aβ peptide associated with Alzheimer's disease. To this end, MD simulations on the microsecond time scale were performed, modeling the various GAGs and $A\beta_{16-22}$ in a 1:1 and a 2:2 stoichiometry. For the GAGs, we considered HA as a nonsulphated GAG as well as H6S, C4S, and C6S as sulphated GAGs (Figure 3 on page 13). The 1:1 simulations revealed that either GAG and $A\beta_{16-22}$ interact with each other, yet the mutual effects are small. The GAGs become somewhat more flexible in the presence of $A\beta_{16-22}$ and also adopt coil and even a few spherical-shape conformations. The same applies to $A\beta_{16-22}$, whose high preference for extended structures as a monomer [3, 341] is slightly reduced in favor of turns in the presence of the GAGs.

The 2:2 simulations, on the other hand, revealed a profound effect of the GAGs on the dimerization of $A\beta_{16-22}$. In solution, the peptide has a high propensity to aggregate into an in-register, antiparallel β-sheet structure (Figure 92) [3, 341], which is also present in the amyloid fibrils of $A\beta_{16-21}$ [240]. In the presence of the GAGs, the amount of β-sheet present in the dimers is generally lower as a result of the competition between interpeptide and peptide–GAG interactions. The latter, however, does not involve many H-bonds, but are dominated by electrostatic interactions involving K16 and negative charges found in the GAGs. This interaction allows $A\beta_{16-22}$ to aggregate into parallel alignments since it prevents the two lysine residues from interacting with each other. Even though the amount of β-sheet is smaller than for the $A\beta_{16-22}$ dimer in solution, the GAGs have an ordering effect on the resulting dimer structures. In-register parallel or antiparallel alignments between the two peptides are clearly preferred. This can be explained with the reduced influence of the charged residues K16 and E22 in the presence of the GAGs, which strengthens the influence of the van der Waals interactions during the aggregation process. They are strongest if the contact area between the two peptides is maximal, which is the case for perfect alignment.

The effects of the different GAGs on $A\beta_{16-22}$ are generally similar and the differences are subtle. In comparison to HA, the sulphation in H6S, C4S and C6S increases the interaction strength between them and $A\beta_{16-22}$. However, the interaction strength is not always correlated with the effects on $A\beta_{16-22}$. This excludes C6S which forms the strongest contacts with $A\beta_{16-22}$ at both the 1:1 and 2:2 ratio and also causes the most substantial changes in the $A\beta_{16-22}$ monomer and dimer structures. The sulphate group of C6S is well exposed and therefore interacts strongly with K16, causing these effects. HA, on the other hand, has the weakest interactions with $A\beta_{16-22}$, but nonetheless has more pronounced consequences on $A\beta_{16-22}$ as H6S has. Thus, further studies are needed to fully elucidate the interplay between amyloid aggregation and GAG interactions.

In summary, we find that GAGs have a noteworthy effect on the ordering of the peptides during the aggregation process. While we have not observed an increased β-sheet amount in the $A\beta_{16-22}$ dimers – as concluded from experiments of full-length Aβ [19, 289] –, the increased ordering should lead to enhanced β-sheet formation if a dimer that built following the interaction with a GAG molecule is the starting point, i.e., the nucleus for larger aggregates. Our future simulations including GAGs and six or even more $A\beta_{16-22}$ peptides will show whether this hypothesis holds true. Moreover, we will also address how Aβ-42 behaves in the presence of the different GAGs. Our current simulation results support the view that GAGs act as a template for amyloid aggregation [19], and that the aggregation is largely influenced by electrostatic interactions between the negative polyelectrolyte charges of the GAGs and the positively charged side chains [84, 333, 334].

8

CONCLUSIONS AND FUTURE DIRECTIONS

Peptide aggregation is a complex problem. The peptides in this problem lack a ordered three-dimensional structure and are also known as IDPs. The aggregation phenomena of these IDPs into highly structured amyloid fibrils is a cause or associated symptom in diseases such as Alzheimer's disease and 50+ neurodegenerative diseases [229, 345, 346]. A variety of biomolecules can also influence the formation of these aggregates, such as GAGs, although their mechanisms of action are still largely unknown. An intriguing way to circumvent the aggregation process or designing drugs against neurodegenerative diseases is to elucidate the underlying sequence–structure–function relationship altogether for the IDPs coupled with their interaction with other biomolecules. Experimental approaches reveal that toxicity of the transient oligomeric species formed during the aggregation pathway is the key element in the onset of degeneration of brain cells. However, understanding the conformational heterogeneity of the oligomeric species and their preceding monomers via experiments is an arduous process. In this thesis, I used MD simulations to interpret the peculiarities of amino acid sequences in the conformational flexibility and structural plasticity of Aβ, the folding mechanisms of Aβ in the presence of GAGs, and provide a detailed understanding of their aggregation process in different cellular environment. Coupled with it, I also discussed about the quality of FFs which is required to model both folded and non-folded peptide (IDPs).

In Chapter 4, I discussed the accuracy of seven FF/water combinations by validating against the thermodynamic and kinetic data derived from experiments. The detailed comparison with NMR and FRET data, allows us to conclude that A99SB-UCB produced an Aβ40 conformational ensemble that is best in agreement with experiment results. The second best performance was found for A99SB-ILDN/TIP4P-D, which can also be recommended for modeling Aβ. Out of the seven FFs, the usage

of two: A03ws and A99SB*-ILDN/TIP3P is clearly discouraged as they produce too much folded Aβ40 conformations, with too much helix in the case of A03ws and too much β-sheet with A99SB*-ILDN/TIP3P. Concurrently, acceptable results were produced for A99SB-disp, C22*/TIP3P, and C36m. Even though C36m was explicitly developed for IDPs, it does not lead to a better performance than C22*/TIP3P.

In summary, the recommendation is to use A99SB-UCB when simulating Aβ, a FF that was carefully optimized on atom-type basis. To further improve the performance of this FF for Aβ, special attention should be devoted to the sequence ^{9}GYEVHH^{14} as here the deviation from the experimental J-couplings is the largest. However, it should be noted that comparison between the MD ensembles discussed here and the NMR results has certain limitations in the conditions used, as the former were generated at 300 K while the latter were obtained at 277 K. A recent MD simulation study of histatin 5 at 277 K prevented us from performing the simulation again, has revealed that for histatin 5 the current FFs fail to capture the temperature dependence of IDP structures, i.e., the increase in folding upon temperature increase cannot be modeled. The conformational ensemble of histatin 5 produced at room temperature looks almost identical to that obtained at 283 K. Using this analogy of histatin 5, one can speculate that FFs that produce extended structures of Aβ40 at 300 K will most likely also yield extended structures at 277 K, while A03ws and A99SB*-ILDN/TIP3P are likely to still produce collapsed and partially folded structures. The temperature dependence of the conformational ensemble of IDPs still needs to investigated in future and also mechanism of FF development, i.e. experimental conditions used based on which FF are reparameterized.

In Chapter 5, the major conclusion is again in the context of FFs but with respect to peptide aggregation of $A\beta_{16-22}$ and its mutants. The simulation studies included FFs specifically designed for IDPs, such as A99SB-disp and C36m but also older FFs like OPLS-AA and Gromos54A7. In comparison to the older FFs, the new improved FF versions incorporated optimized backbone torsion parameters, addition of CMAP corrections and/or stronger van der Waals interactions between peptide and water. In the case of A99SB-disp (A99-d) the number, and the strength of the H-bonds between the peptide and water is greatly increased which leads to the formation unstable

aggregates. Whereas with C36mW, the introduction of stronger van der Waals interactions between peptide and water in C36m, does not lead to major changes in the aggregation propensity of $A\beta_{16-22}$ and its mutant F19L with the formation of ordered and stable aggregates validating experimental observations [241]. The only exception being the mutation of F19V/F20V in $A\beta_{16-22}$ which leads to decreased propensity in C36mW because of increased solvation characteristics of peptides. The peptide aggregation characteristics of A99SB-disp and A03ws are found to be very similar [232], A03ws is the same FF which was discouraged for modeling Aβ. The A99SB-UCB FF which was identified to be best suitable for modeling of IDPs in Chapter 4 [268], was found to form unstable aggregates as shown in Figure 96. So, in general peptide FFs which were speculated to be accurate for modeling monomeric IDPs are considered unsuitable for studying aggregation kinetics of IDPs.

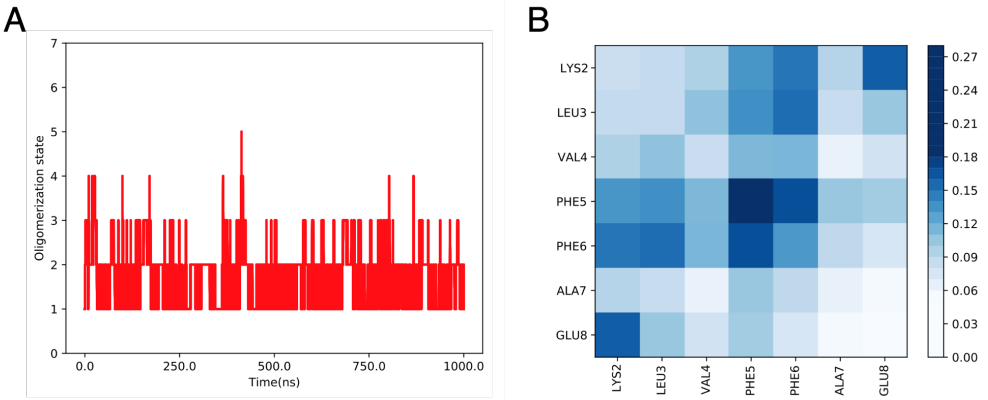

Figure 96: (A) Average oligomer size as a function of time for *wt* (red) using A99SB-UCB. The graphs are averages over one trajectory of *wt* using A99SB-UCB. (B) Interpeptide contacts found in the aggregates formed in the simulation of *wt* using A99SB-UCB.

The C36m and C36mW are found suitable to reproduce experimental aggregation propensities of $A\beta_{16-22}$ and its two mutants. This concludes that these two FFs incorporates somewhat balanced peptide–water interactions. In addition, future FF reparameterisations should still include careful revisions of the van der Waals interactions between peptide and water, and experimental observables, such as the translational diffusion of different IDPs, as well as conformational characteristics of the monomeric state. Nowadays, such information will aid machine learning models that have made significant impact on the analysis of peptide simulations. As noted above, armed with an accurate protein force-field and computational resources, we could simulate the

dynamics of a peptide long enough to sample its equilibrium distribution. However, the problem of extracting the essential information from the simulation, and to validate it to experimental measurements still persists. With the help of unsupervised learning methods can help to extract metastable states from high dimensional simulation data by representing the classical energy function of a system as a function of the atomic coordinates to reproduce experimentally relevant properties.

In Chapter 6, I introduced a different biomolecules known as GAGs. The aim of this chapter was to assess the influence of monosaccharide composition, glycosidic linkage, sulphation sites, salt types, and salt concentrations on the conformational ensembles of GAGs. It was observed that polymeric GAG chains in solution can assume a number of configurations that contribute to their entropy. Heparan-6-sulphate, was considered the most flexible GAG, followed by chondroitin-6-sulphate, chondroitin-4-sulphate and hyaluronic acid. But at higher salt concentrations of 150 mM, the flexibility of heparan-6-sulphate and chondroitin-6-sulphate was reduced compared to their behaviour at 0 mM salt concentration due to the presence of anions. A general conclusion is that the GAGs have an intrinsic preference for extended structures, but such phenomena is inconsistent because of sulphate moieties, especially at position 6, and interaction with cations.

In Chapter 7, I integrated the information from Chapter 5 regarding FF, and conformational characteristics of GAGs from Chapter 6 to understand the influence of different GAGs on the conformational preference and dimerization of $A\beta_{16-22}$. It is revealed that GAGs have a noteworthy effect on the ordering of the peptides during the aggregation process of $A\beta_{16-22}$, but the dimers are bereft of the β-sheet content. But it is assumed that increased ordering of the peptides should lead to enhanced β-sheet formation. The driving force behind the aggregation process is the electrostatic interactions between the negative polyelectrolyte charges of the GAGs and the positively charged side chains of the $A\beta_{16-22}$ peptide. Conversely, GAGs become more flexible in the presence of $A\beta_{16-22}$ and adopt coil and even spherical-shape conformations. It is clearly understood that GAGs play a crucial role in the aggregation process of $A\beta_{16-22}$ and other IDPs [347], so it is of utmost importance to elucidate the underlying mechanisms of physiological and pathological activities of GAGs. A

key help from neural network models, would enable us to model artificial GAGs that can regulate the function of GAGs.

BIBLIOGRAPHY

[1] Hebah Fatafta, Suman Samantray, Abdallah Sayyed-Ahmad, Orkid Coskuner-Weber, and Birgit Strodel. "Molecular simulations of IDPs: From ensemble generation to IDP interactions leading to disorder-to-order transitions." In: *Prog. Mol. Biol. Trans. Sci.* 183 (2021), pp. 135–185.

[2] Arghadwip Paul, Suman Samantray, Marco Anteghini, Mohammed Khaled, and Birgit Strodel. "Thermodynamics and kinetics of the amyloid-β peptide revealed by Markov state models based on MD data in agreement with experiment." In: *Chem. Sci.* 12.19 (2021), pp. 6652–6669.

[3] Suman Samantray, Feng Yin, Batuhan Kav, and Birgit Strodel. "Different Force Fields Give Rise to Different Amyloid Aggregation Pathways in Molecular Dynamics Simulations." In: *J. Chem. Inf. Model.* 60.12 (2020), pp. 6462–6475.

[4] Suman Samantray, Olujide O Olubiyi, and Birgit Strodel. "The Influences of Sulphation, Salt Type, and Salt Concentration on the Structural Heterogeneity of Glycosaminoglycans." In: *Int. J. Mol. Sci.* 22.21 (2021), p. 11529.

[5] Suman Samantray and Birgit Strodel. "The effects of different glycosaminoglycans on the structure and aggregation of the amyloid-β (16–22) peptide." In: *J. Phys. Chem. B* 125.21 (2021), 5511–5525.

[6] Alistair Burns and Steve Iliffe. "Alzheimer's disease." In: *BMJ* 338 (2009).

[7] M Paul Murphy and Harry LeVine III. "Alzheimer's disease and the amyloid-β peptide." In: *J. Alzheimer's Dis.* 19.1 (2010), pp. 311–323.

[8] Luitgard Nagel-Steger, Michael C Owen, and Birgit Strodel. "An account of amyloid oligomers: facts and figures obtained from experiments and simulations." In: *ChemBioChem* 17.8 (2016), pp. 657–676.

[9] Mary P Lambert, AK Barlow, Brett A Chromy, C Edwards, R Freed, M Liosatos, TE Morgan, I Rozovsky, B Trommer, Kirsten L Viola, et al. "Diffusible, nonfibrillar ligands derived from Aβ1–42 are potent central nervous system neurotoxins." In: *Proc. Natl. Acad. Sci. U. S. A.* 95.11 (1998), pp. 6448–6453.

[10] Marina D Kirkitadze, Gal Bitan, and David B Teplow. "Paradigm shifts in Alzheimer's disease and other neurodegenerative disorders: the emerging role of oligomeric assemblies." In: *J. Neurosci. Res.* 69.5 (2002), pp. 567–577.

[11] Padideh Kamali-Zare and Charles Nicholson. "Brain extracellular space: geometry, matrix and physiological importance." In: *Basic Clin. Neurosci.* 4.4 (2013), p. 282.

[12] Javier Dıéaz-Nido, Francisco Wandosell, and Jesús Avila. "Glycosaminoglycans and β-amyloid, prion and tau peptides in neurodegenerative diseases." In: *Peptides* 23.7 (2002), pp. 1323–1332.

[13] E Gruys, A Ultee, and N Upragarin. "Glycosaminoglycans are part of amyloid fibrils: Ultrastructural evidence in avian AA amyloid stained with cuprolinic blue and labeled with immunogold." In: *Amyloid* 13.1 (2006), pp. 13–19.

[14] Clara Iannuzzi, Gaetano Irace, and Ivana Sirangelo. "The effect of glycosaminoglycans (GAGs) on amyloid aggregation and toxicity." In: *Molecules* 20.2 (2015), pp. 2510–2528.

[15] Iain D Young, L Ailles, S Narindrasorasak, R Tan, and R Kisilevsky. "Localization of the basement membrane heparan sulfate proteoglycan in islet amyloid deposits in type II diabetes mellitus." In: *Arch. Pathol. Lab. Med.* 116.9 (1992), pp. 951–954.

[16] AD Snow, TN Wight, D Nochlin, Y Koike, K Kimata, SJ DeArmond, and SB Prusiner. "Immunolocalization of heparan sulfate proteoglycans to the prion protein amyloid plaques of Gerstmann-Straussler syndrome, Creutzfeldt-Jakob disease and scrapie." In: *Lab. Invest. LAB INVEST* 63.5 (1990), pp. 601–611.

[17] Min Zhu, Pierre O Souillac, Cristian Ionescu-Zanetti, Sue A Carter, and Anthony L Fink. "Surface-catalyzed amyloid fibril formation." In: *J. Biol. Chem.* 277.52 (2002), pp. 50914–50922.

[18] John B Ancsin. "Amyloidogenesis: historical and modern observations point to heparan sulfate proteoglycans as a major culprit." In: *Amyloid* 10.2 (2003), pp. 67–79.

[19] Jo Anne McLaurin, Trudy Franklin, Xiaoqiong Zhang, Jianping Deng, and Paul E. Fraser. "Interactions of Alzheimer amyloid-β peptides with glycosaminoglycans: Effects on fibril nucleation and growth." In: *Eur. J. Biochem.* 266.3 (1999), pp. 1101–1110.

[20] Martín Carballo-Pacheco and Birgit Strodel. "Advances in the simulation of protein aggregation at the atomistic scale." In: *J. Phys. Chem. B* 120.12 (2016), pp. 2991–2999.

[21] Christine A Orengo, Annabel E Todd, and Janet M Thornton. "From protein structure to function." In: *Curr. Opin. Struct.* 9.3 (1999), pp. 374–382.

[22] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. "The protein folding problem." In: *Annu. Rev. Biophys.* 37 (2008), pp. 289–316.

[23] Christopher J Oldfield and A Keith Dunker. "Intrinsically disordered proteins and intrinsically disordered protein regions." In: *Annu. Rev. Biochem.* 83 (2014), pp. 553–584.

[24] A Keith Dunker, J David Lawson, Celeste J Brown, Ryan M Williams, Pedro Romero, Jeong S Oh, Christopher J Oldfield, Andrew M Campen, Catherine M Ratliff, Kerry W Hipps, et al. "Intrinsically disordered protein." In: *J. Mol. Graph.* 19.1 (2001), pp. 26–59.

[25] Vladimir N Uversky. "Intrinsically disordered proteins from A to Z." In: *Int. J. Biochem. Cell Biol.* 43.8 (2011), pp. 1090–1103.

[26] Vladimir N Uversky. "A decade and a half of protein intrinsic disorder: biology still waits for physics." In: *Protein Sci.* 22.6 (2013), pp. 693–724.

[27] Johnny Habchi, Peter Tompa, Sonia Longhi, and Vladimir N Uversky. "Introducing protein intrinsic disorder." In: *Chem. Rev.* 114.13 (2014), pp. 6561–6588.

[28] András Hatos et al. "DisProt: intrinsic protein disorder annotation in 2020." In: *Nucleic Acids Res.* 48.D1 (2019), pp. D269–D276.

[29] Walter Basile, Marco Salvatore, Claudio Bassot, and Arne Elofsson. "Why do eukaryotic proteins contain more intrinsically disordered regions?" In: *PLoS Comput. Biol.* 15.7 (2019), e1007186.

[30] Peter E Wright and H Jane Dyson. "Intrinsically disordered proteins in cellular signalling and regulation." In: *Nat. Rev. Mol. Cell Biol.* 16.1 (2015), pp. 18–29.

[31] Vladimir N Uversky. "The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins." In: *Intrinsically Disord. Proteins* 1.1 (2013), e24684.

[32] Simone Kosol, Sara Contreras-Martos, Cesyen Cedeño, and Peter Tompa. "Structural characterization of intrinsically disordered proteins by NMR spectroscopy." In: *Molecules* 18.9 (2013), pp. 10802–10828.

[33] Ursula Jakob, Richard Kriwacki, and Vladimir N Uversky. "Conditionally and transiently disordered proteins: awakening cryptic disorder to regulate protein function." In: *Chem. Rev.* 114.13 (2014), pp. 6779–6805.

[34] Hedi Hegyi and Peter Tompa. "Intrinsically disordered proteins display no preference for chaperone binding in vivo." In: *PLoS Comput. Biol.* 4.3 (2008), e1000017.

[35] Prakash Kulkarni and Vladimir N Uversky. *Intrinsically disordered proteins in chronic diseases*. 2019.

[36] Paola Mendoza-Espinosa, Victor García-González, Abel Moreno, Rolando Castillo, and Jaime Mas-Oliva. "Disorder-to-order conformational transitions in protein structure and its relationship to disease." In: *Mol. Cell. Biochem.* 330.1 (2009), pp. 105–120.

[37] Orkid Coskuner and Vladimir N Uversky. "Intrinsically disordered proteins in various hypotheses on the pathogenesis of Alzheimer's and Parkinson's diseases." In: *Prog. Mol. Biol. Transl. Sci.* 166 (2019), pp. 145–223.

[38] Phuong H. Nguyen et al. "Amyloid Oligomers: A Joint Experimental/Computational Perspective on Alzheimer's Disease, Parkinson's Disease, Type II Diabetes, and Amyotrophic Lateral Sclerosis." In: *Chem. Rev.* 121 (2021), 2545–2647.

[39] Rakez Kayed and Cristian A. Lasagna-Reeves. "Molecular mechanisms of amyloid oligomers toxicity." In: *J Alzheimers Dis* 33 (2013), S67–S78.

[40] Gabor G Kovacs and Herbert Budka. "Prion diseases: from protein to cell pathology." In: *Am. J. Pathol.* 172.3 (2008), pp. 555–565.

[41] N Friedreich and A Kekulé. "Zur amyloidfrage." In: *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin* 16.1 (1859), pp. 50–65.

[42] George G Glenner and Caine W Wong. "Alzheimer's disease: initial report of the purification and characterization of a novel cerebrovascular amyloid protein." In: *Biochem. Biophys. Res. Commun.* 120.3 (1984), pp. 885–890.

[43] Jie Kang, Hans-Georg Lemaire, Axel Unterbeck, J Michael Salbaum, Colin L Masters, Karl-Heinz Grzeschik, Gerd Multhaup, Konrad Beyreuther, and Benno Müller-Hill. "The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor." In: *Nature* 325.6106 (1987), pp. 733–736.

[44] Christian Haass, Edward H Koo, Angela Mellon, Albert Y Hung, and Dennis J Selkoe. "Targeting of cell-surface β-amyloid precursor protein to lysosomes: alternative processing into amyloid-bearing fragments." In: *Nature* 357.6378 (1992), pp. 500–503.

[45] Peter Seubert, Carmen Vigo-Pelfrey, Fred Esch, Michael Lee, Harry Dovey, Dave Davis, Sukanto Sinha, Michael Schiossmacher, Justine Whaley, Cathy Swindlehurst, et al. "Isolation and quantification of soluble Alzheimer's β-peptide from biological fluids." In: *Nature* 359.6393 (1992), pp. 325–327.

[46] Mikio Shoji, Todd E Golde, Jorge Ghiso, Tobun T Cheung, Steven Estus, Lillian M Shaffer, Xiao-Dan Cai, Deborah M McKay, Ron Tintner, Blas Frangione, et al. "Production of the Alzheimer amyloid beta protein by normal proteolytic processing." In: *Science* 258.5079 (1992), pp. 126–129.

[47] Axel Abelein, Jan Pieter Abrahams, Jens Danielsson, Astrid Gräslund, Jüri Jarvet, Jinghui Luo, Ann Tiiman, and Sebastian KTS Wärmländer. "The hairpin conformation of the amyloid β peptide is an important structural motif along the aggregation pathway." In: *J. Biol. Inorg. Chem.* 19.4 (2014), pp. 623–634.

[48] Xiaojuan Sun, Wei-Dong Chen, and Yan-Dong Wang. "β-Amyloid: the key peptide in the pathogenesis of Alzheimer's disease." In: *Front. Pharmacol.* 6 (2015), p. 221.

[49] Jessica Nasica-Labouze and et al. "Amyloid β Protein and Alzheimer's Disease: When Computer Simulations Complement Experimental Studies." In: *Chem. Rev.* 115.9 (2015), pp. 3518–3563.

[50] Matthias Schmidt, Alexis Rohou, Keren Lasker, Jay K Yadav, Cordelia Schiene-Fischer, Marcus Fändrich, and Nikolaus Grigorieff. "Peptide dimer structure in an Aβ (1–42) fibril visualized with cryo-EM." In: *Proc. Natl. Acad. Sci. U. S. A.* 112.38 (2015), pp. 11858–11863.

[51] Guo-fang Chen, Ting-hai Xu, Yan Yan, Yu-ren Zhou, Yi Jiang, Karsten Melcher, and H Eric Xu. "Amyloid beta: structure, biology and structure-based therapeutic development." In: *Acta Pharmacol. Sin.* 38.9 (2017), pp. 1205–1235.

[52] Colin L Masters and Dennis J Selkoe. "Biochemistry of amyloid β-protein and amyloid deposits in Alzheimer disease." In: *Cold Spring Harb. Perspect. Med.* 2.6 (2012), a006262.

[53] Kasper P Kepp. "Alzheimer's disease: How metal ions define β-amyloid function." In: *Coord. Chem. Rev.* 351 (2017). Bioinorganic Chemistry in the European Union, pp. 127–159.

[54] Nikos Afratis, Chrisostomi Gialeli, Dragana Nikitovic, Theodore Tsegenidis, Evgenia Karousou, Achilleas D Theocharis, Mauro S Pavão, George N Tzanakakis, and Nikos K Karamanos. "Glycosaminoglycans: key players in cancer cell biology and treatment." In: *FEBS J.* 279.7 (2012), pp. 1177–1197.

[55] Mariano S Viapiano and Russell T Matthews. "From barriers to bridges: chondroitin sulfate proteoglycans in neuropathology." In: *Trends. Mol. Med.* 12.10 (2006), pp. 488–496.

[56] M Morawski, G Brückner, T Arendt, and RT Matthews. "Aggrecan: beyond cartilage and into the brain." In: *Int. J. Biochem. Cell Biol.* 44.5 (2012), pp. 690–693.

[57] Neha S. Gandhi and Ricardo L. Mancera. "The structure of glycosaminoglycans and their interactions with proteins." In: *Chem. Biol. Drug Des.* 72.6 (2008), pp. 455–482.

[58] Karl Meyer and John W Palmer. "The polysaccharide of the vitreous humor." In: *J. Biol. Chem.* 107.3 (1934), pp. 629–634.

[59] Bernard Weissmann and Karl Meyer. "The structure of hyalobiuronic acid and of hyaluronic acid from umbilical Cord1, 2." In: *J. Am. Chem. Soc.* 76.7 (1954), pp. 1753–1757.

[60] Giancarlo Ghiselli. "Drug-Mediated Regulation of Glycosaminoglycan Biosynthesis." In: *Med. Res. Rev.* 37.5 (2017), pp. 1051–1094.

[61] Barrie Fong Chong, Lars M Blank, Richard Mclaughlin, and Lars K Nielsen. "Microbial hyaluronic acid production." In: *Appl. Microbiol. Biotechnol.* 66.4 (2005), pp. 341–351.

[62] John E Scott and Frank Heatley. "Hyaluronan forms specific stable tertiary structures in aqueous solution: a 13C NMR study." In: *Proc. Natl. Acad. Sci. U.S.A.* 96.9 (1999), pp. 4850–4855.

[63] J Robert E Fraser, Torvard C Laurent, and UBG Laurent. "Hyaluronan: its nature, distribution, functions and turnover." In: *J. Intern. Med.* 242.1 (1997), pp. 27–33.

[64] Larry Sherman, Jonathan Sleeman, Peter Herrlich, and Helmut Ponta. "Hyaluronate receptors: key players in growth, differentiation, migration and tumor progression." In: *Curr. Opin. Cell Biol.* 6.5 (1994), pp. 726–733.

[65] Torvard C Laurent, Ulla BG Laurent, and J Robert E Fraser. "The structure and function of hyaluronan: an overview." In: *Immunol. Cell Biol.* 74.2 (1996), a1–a7.

[66] E George. "Intra-articular hyaluronan treatment for osteoarthritis." In: *Ann. Rheum. Dis.* 57.11 (1998), pp. 637–640.

[67] William Henry Howell and Emmett Holt. "Two new factors in blood coagulation—heparin and pro-antithrombin." In: *Am. J. Physiol.* 47.3 (1918), pp. 328–341.

[68] Robert J. Linhardt. "Heparin: An important drug enters its seventh decade." In: *Chem. and Ind.* 2 (1991), pp. 45–50.

[69] Cui Hao, Hongmei Xu, Lingfan Yu, and Lijuan Zhang. "Heparin: an essential drug for modern medicine." In: *Prog. Mol. Biol. Transl. Sci.* 163 (2019), pp. 1–19.

[70] Robert J Linhardt, Stephen A Ampofo, Jawed Fareed, Debra Hoppensteadt, Judah Folkman, and John B Mulliken. "Isolation and characterization of human heparin." In: *Biochemistry* 31.49 (1992), pp. 12441–12445.

[71] Benito Casu and Ulf Lindahl. "Structure and biological interactions of heparin and heparan sulfate." In: *Adv. Carbohydr. Chem. Biochem.* 57 (2001), pp. 159–206.

[72] Norbert Perrimon and Merton Bernfield. "Specificities of heparan sulphate proteoglycans in developmental processes." In: *Nature* 404.6779 (2000), pp. 725–728.

[73] Deepak Shukla, Jian Liu, Peter Blaiklock, Nicholas W Shworak, Xiaomei Bai, Jeffrey D Esko, Gary H Cohen, Roselyn J Eisenberg, Robert D Rosenberg, and Patricia G Spear. "A novel role for 3-O-sulfated heparan sulfate in herpes simplex virus 1 entry." In: *Cell* 99.1 (1999), pp. 13–22.

[74] Kristian Prydz. "Determinants of glycosaminoglycan (GAG) structure." In: *Biomolecules* 5.3 (2015), pp. 2003–2022.

[75] Athanasia P Asimakopoulou, Achilleas D Theocharis, George N Tzanakakis, and Nikos K Karamanos. "The biological role of chondroitin sulfate in cancer and chondroitin-based anticancer agents." In: *In vivo* 22.3 (2008), pp. 385–389.

[76] Marco Mende, Christin Bednarek, Mirella Wawryszyn, Paul Sauter, Moritz B Biskup, Ute Schepers, and Stefan BraHERE!HERE!se. "Chemical synthesis of glycosaminoglycans." In: *Chem. Rev.* 116.14 (2016), pp. 8193–8255.

[77] Rahul Raman, V Sasisekharan, and Ram Sasisekharan. "Structural insights into biological roles of protein-glycosaminoglycan interactions." In: *Chem. Biol.* 12.3 (2005), pp. 267–277.

[78] Sylvain D Vallet, Olivier Clerc, and Sylvie Ricard-Blum. "Glycosaminoglycan–Protein Interactions: The First Draft of the Glycosaminoglycan Interactome." In: *J. Histochem. Cytochem.* 69.2 (2021), pp. 93–104.

[79] Sten Eriksson, Sabina Janciauskiene, and Lars Lannfelt. "Alpha 1-antichymotrypsin regulates Alzheimer beta-amyloid peptide fibril formation." In: *Proc. Natl. Acad. Sci. U. S. A.* 92.6 (1995), pp. 2313–2317.

[80] Evelyne Terzi, Günter Hölzemann, and Joachim Seelig. "Self-association of β-amyloid peptide (1–40) in solution and binding to lipid membranes." In: *J. Biol. Chem.* 252.5 (1995), pp. 633–642.

[81] Gerardo M Castillo, Catherine Ngo, Joel Cummings, Thomas N Wight, and Alan D Snow. "Perlecan binds to the β-amyloid proteins (Aβ) of Alzheimer's disease, accelerates Aβ fibril formation, and maintains Aβ fibril stability." In: *J. Neurochem.* 69.6 (1997), pp. 2452–2465.

[82] JoAnne McLaurin, Trudy Franklin, Paul E Fraser, and Avijit Chakrabartty. "Structural transitions associated with the interaction of Alzheimer β-amyloid peptides with gangliosides." In: *J. Biol. Chem.* 273.8 (1998), pp. 4506–4515.

[83] Bernd Bohrmann, Lars Tjernberg, Pascal Kuner, Sonia Poli, Bernard Levet-Trafit, Jan NaHERE!HERE!slund, Grayson Richards, Walter Huber, Heinz Do-HERE!HERE!beli, and Christer Nordstedt. "Endogenous proteins controlling amyloid β-peptide polymerization: possible implications for β-amyloid formation in the central nervous system and in peripheral tissues." In: *J. Biol. Chem.* 274.23 (1999), pp. 15990–15995.

[84] Jack Van Horssen, Pieter Wesseling, Lambert P.W.J. Van Den Heuvel, Robert M.W. De Waal, and Marcel M. Verbeek. "Heparan sulphate proteoglycans in Alzheimer's disease and amyloid-related disorders." In: *Lancet Neurol.* 2.8 (2003), pp. 482–492.

[85] Norifumi Watanabe, Wataru Araki, De-Hua Chui, Takao Makifuchi, Yasuo Ihara, and Takeshi Tabira. "Glypican-1 as an Aβ binding HSPG in the human brain: Its localization in DIG domains and possible roles in the pathogenesis of Alzheimer's disease." In: *FASEB J.* 18.9 (2004), pp. 1013–1015.

[86] Thomas Wisniewski, Eduardo M Castano, Adam Golabek, Tikva Vogel, and Blas Frangione. "Acceleration of Alzheimer's fibril formation by apolipoprotein E in vitro." In: *Am. J. Clin. Pathol.* 145.5 (1994), p. 1030.

[87] Scott Webster and Joseph Rogers. "Relative efficacies of amyloid β peptide (Aβ) binding proteins in Aβ aggregation." In: *J. Neurosci. Res.* 46.1 (1996), pp. 58–66.

[88] ALAN D Snow, H Mar, D Nochlin, K_ Kimata, M Kato, S Suzuki, J Hassell, and TN Wight. "The presence of heparan sulfate proteoglycans in the neuritic plaques and congophilic angiopathy in Alzheimer's disease." In: *Am. J. Pathol.* 133.3 (1988), p. 456.

[89] Suree Narindrasorasak, David Lowery, P Gonzalez-DeWhitt, Roger A Poorman, Barry Greenberg, and Robert Kisilevsky. "High affinity interactions between the Alzheimer's beta-amyloid precursor proteins and the basement membrane form of heparan sulfate proteoglycan." In: *J. Biol. Chem.* 266.20 (1991), pp. 12878–12883.

[90] D Van Gool, Guido David, M Lammens, Franz Baro, and René Dom. "Heparan sulfate expression patterns in the amyloid deposits of patients with Alzheimer's and Lewy body type dementia." In: *Dement. Geriatr. Cogn. Disord.* 4.6 (1993), pp. 308–314.

[91] David A Dewitt, Jerry Silver, David R Canning, and George Perry. "Chondroitin sulfate proteoglycans are associated with the lesions of Alzheimer's disease." In: *Exp. Neurol.* 121.2 (1993), pp. 149–152.

[92] Alan D Snow, David Nochlin, Raymond Sekiguchi, and Steven S Carlson. "Identification and immunolocalization of a new class of proteoglycan (keratan sulfate) to the neuritic plaques of Alzheimer's disease." In: *Exp. Neurol.* 138.2 (1996), pp. 305–317.

[93] AD Snow, H Mar, D Nochlin, H Kresse, and TN Wight. "Peripheral distribution of dermatan sulfate proteoglycans (decorin) in amyloid-containing plaques and their presence in neurofibrillary tangles of Alzheimer's disease." In: *J. Histochem. Cytochem.* 40.1 (1992), pp. 105–113.

[94] Colin J Barrow and Michael G Zagorski. "Solution structures of beta peptide and its constituent fragments: relation to amyloid deposition." In: *Science* 253.5016 (1991), pp. 179–182.

[95] Heinrich Sticht, Peter Bayer, Dieter Willbold, Sonja Dames, Caroline Hilbich, Konrad Beyreuther, Rainer W Frank, and Paul Rösch. "Structure of amyloid A4-(1–40)-peptide of Alzheimer's disease." In: *Eur. J. Biochem.* 233.1 (1995), pp. 293–298.

[96] JoAnne McLaurin and Avi Chakrabartty. "Membrane disruption by Alzheimer β-amyloid peptides mediated through specific binding to either phospholipids or gangliosides: implications for neurotoxicity." In: *J. Biol. Chem.* 271.43 (1996), pp. 26482–26489.

[97] Andrew R Leach, Christopher J Cramer, Daan Frenkel, Berend Smit, Ken Dill, Sarina Bromberg, Herman JC Berendsen, MP Allen, and DJ Tildesley. *Molecular modelling: principles and applications*. Pearson-Prentice Hall, 2001.

[98] MP Allen and DJ Tildesley. "Molecular Simulation of Liquids." In: *Clarendon, Oxford* (1987).

[99] Michael P Allen et al. "Introduction to molecular dynamics simulation." In: *Computational soft matter: from synthetic polymers to proteins* 23 (2004), pp. 1–28.

[100] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[101] Loup Verlet. "Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules." In: *Physical review* 159.1 (1967), p. 98.

[102] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations." In: *J. Comput. Chem.* 4.2 (1983), pp. 187–217.

[103] Christopher I. Bayly, Kenneth M. Merz, David M. Ferguson, Wendy D. Cornell, Thomas Fox, James W. Caldwell, Peter A. Kollman, Piotr Cieplak, Ian R. Gould, and David C. Spellmeyer. "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules." In: *J. Am. Chem. Soc.* 117.19 (1995), pp. 5179–5197.

[104] William L. Jorgensen and Julian Tirado-Rives. "The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin." In: *J. Am. Chem. Soc.* 110.6 (1988), pp. 1657–1666.

[105] Walter R.P. Scott, Philippe H. Hünenberger, Ilario G. Tironi, Alan E. Mark, Salomon R. Billeter, Jens Fennen, Andrew E. Torda, Thomas Huber, Peter Krüger, and Wilfred F. Van Gunsteren. "The GROMOS biomolecular simulation program package." In: *J. Phys. Chem. A* 103.19 (1999), pp. 3596–3607.

[106] Sarah Rauscher, Vytautas Gapsys, Michal J. Gajda, Markus Zweckstetter, Bert L. de Groot, and Helmut Grubmüller. "Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment." In: *J. Chem. Theory Comput.* 11 (2015), pp. 5513–5524.

[107] Martín Carballo-Pacheco and Birgit Strodel. "Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins." In: *Protein Sci.* 26.2 (2017), pp. 174–185.

[108] Paul Robustelli, Stefano Piana, and David E. Shaw. "Developing a molecular dynamics force field for both folded and disordered protein states." In: *Proc. Natl. Acad. Sci. U.S.A.* 115.21 (2018), E4758–E4766.

[109] Viet Hoang Man, Xibing He, Philippe Derreumaux, Beihong Ji, Xiang Qun Xie, Phuong H. Nguyen, and Junmei Wang. "Effects of All-Atom Molecular Mechanics Force Fields on Amyloid Peptide Assembly: The Case of Aβ 16-22 Dimer." In: *J. Chem. Theory Comput.* 15.2 (2019), pp. 1440–1452.

[110] Mueed Ur Rahman, Ashfaq Ur Rehman, Hao Liu, and Hai Feng Chen. "Comparison and evaluation of force fields for intrinsically disordered proteins." In: *J. Chem. Info. Model.* 60.10 (2020), pp. 4912–4923.

[111] Birgit Strodel. "Amyloid aggregation simulations: challenges, advances and perspectives." In: *Curr. Opin. Struct. Biol.* 67 (2021), pp. 145–152.

[112] Junxi Mu, Hao Liu, Jian Zhang, Ray Luo, and Hai Feng Chen. "Recent Force Field Strategies for Intrinsically Disordered Proteins." In: *J. Chem. Inf. Model.* 61.3 (2021), pp. 1037–1047.

[113] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. De Groot, Helmut Grubmüller, and Alexander D. MacKerell. "CHARMM36m: An improved force field for folded and intrinsically disordered proteins." In: *Nat. Methods.* 14.1 (2016), pp. 71–73.

[114] Stefano Piana, Paul Robustelli, Dazhi Tan, Songela Chen, and David E. Shaw. "Development of a Force Field for the Simulation of Single-Chain Proteins and Protein-Protein Complexes." In: *J. Chem. Theory Comput.* 16.4 (2020), pp. 2494–2507.

[115] Robert B. Best and Gerhard Hummer. "Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides." In: *J. Phys. Chem. B* 113.26 (2009), pp. 9004–9015.

[116] Yong Duan et al. "A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations." In: *J. Comput. Chem.* 24.16 (2003), pp. 1999–2012.

[117] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. "Comparison of multiple amber force fields and development of improved protein backbone parameters." In: *Proteins* 65.3 (2006), pp. 712–725.

[118] Michael J. Robertson, Julian Tirado-Rives, and William L. Jorgensen. "Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field." In: *J. Chem. Theory Comput.* 11.7 (2015), pp. 3499–3509.

[119] Edward Harder et al. "OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins." In: *J. Chem. Theory Comput.* 12.1 (2016), pp. 281–296.

[120] Stefano Piana, Kresten Lindorff-Larsen, and David E. Shaw. "How robust are protein folding simulations with respect to force field parameterization?" In: *Biophys. J.* 100.9 (2011), pp. L47–L49.

[121] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee

Ha, et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins." In: *J. Phys. Chem. B* 102.18 (1998), pp. 3586–3616.

[122] Martín Carballo-Pacheco and Birgit Strodel. "Comparison of force fields for Alzheimer's A : A case study for intrinsically disordered proteins." In: *Protein Sci.* 26.2 (2017), pp. 174–185.

[123] Paul S. Nerenberg and Teresa Head-Gordon. "Optimizing protein-solvent force fields to reproduce intrinsic conformational preferences of model peptides." In: *J. Chem. Theory Comput.* 7.4 (2011), pp. 1220–1230.

[124] Paul S Nerenberg, Brian Jo, Clare So, Ajay Tripathy, and Teresa Lyn Head-Gordon. "A New Protein and Water Force Field Combination for Reproducing Solvation Free Energies." In: *J. Phys. Chem. B* 116.15 (2012), pp. 4524–4534.

[125] Fan Jiang, Chen-Yang Zhou, and Yun-Dong Wu. "Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L." In: *J. Phys. Chem. B* 118.25 (2014), pp. 6983–6998.

[126] Chen Yang Zhou, Fan Jiang, and Yun Dong Wu. "Residue-specific force field based on protein coil library. RSFF2: Modification of AMBER ff99SB." In: *J. Phys. Chem. B* 119.3 (2015), pp. 1035–1047.

[127] Alexander D. MacKerell, Michael Feig, and Charles L. Brooks. "Improved Treatment of the Protein Backbone in Empirical Force Fields." In: *J. Am. Chem. Soc.* 126.3 (2004), pp. 698–699.

[128] Alexander D. Mackerell, Michael Feig, and Charles L. Brooks. "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulation." In: *J. Comput. Chem.* 25.11 (2004), pp. 1400–1415.

[129] Rukundo Olivier and Cao Hanqiang. "Nearest Neighbor Value Interpolation." In: *Int. J. Adv. Comput. Sci. Appl.* 3.4 (2012).

[130] L. R. Taylor, William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. "Numerical Recipes: The Art of Scientific Computing." In: *J. Anim. Ecol.* 56.1 (1987), p. 374.

[131] Jing Huang and Alexander D. Mackerell. "CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data." In: *J. Comput. Chem.* 34.25 (2013), pp. 2135–2145.

[132] Dong Song, Ray Luo, and Hai Feng Chen. "The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins." In: *J. Chem. Inf. Model.* 57.5 (2017), pp. 1166–1178.

[133] Hao Liu, Dong Song, Hui Lu, Ray Luo, and Hai Feng Chen. "Intrinsically disordered protein-specific force field CHARMM36IDPSFF." In: *Chem. Biol. Drug Des.* 92.4 (2018), pp. 1722–1735.

[134] Hao Liu, Dong Song, Yangpeng Zhang, Sheng Yang, Ray Luo, and Hai Feng Chen. "Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins." In: *Phys. Chem. Chem. Phys.* 21.39 (2019), pp. 21918–21931.

[135] Sheng Yang, Hao Liu, Yangpeng Zhang, Hui Lu, and Haifeng Chen. "Residue-specific force field improving the sample of intrinsically disordered proteins and folded proteins." In: *J. Chem. Inf. Model.* 59.11 (2019), pp. 4793–4805.

[136] Wei Kang, Fan Jiang, and Yun Dong Wu. "Universal Implementation of a Residue-Specific Force Field Based on CMAP Potentials and Free Energy Decomposition." In: *J. Chem. Theory Comput.* 14.8 (2018), pp. 4474–4486.

[137] Jing Huang and Alexander D. MacKerell. "Force field development and simulations of intrinsically disordered proteins." In: *Curr. Opin. Struct. Biol.* 48 (2018), pp. 40–48.

[138] Jianan Chen and Yundong Wu. "Recent development of atomistic force fields and simulations of intrinsically disordered proteins." In: *Sci. Sin. Chim.* 50.10 (2020), pp. 1320–1332.

[139] Robert B. Best, Wenwei Zheng, and Jeetain Mittal. "Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association." In: *J. Chem. Theory Comput.* 10.11 (2014), pp. 5113–5124.

[140] Stefano Piana, Alexander G. Donchev, Paul Robustelli, and David E. Shaw. "Water dispersion interactions strongly influence simulated structural properties of disordered protein states." In: *J. Phys. Chem. B* 119.16 (2015), pp. 5113–5123.

[141] Suman Samantray, Feng Yin, Batuhan Kav, and Birgit Strodel. "Different Force Fields Give Rise to Different Amyloid Aggregation Pathways in Molecular Dynamics Simulations." In: *J. Chem. Info. Model.* 60.12 (2020), pp. 6462–6475.

[142] Martín Carballo-Pacheco, Ahmed E. Ismail, and Birgit Strodel. "On the Applicability of Force Fields to Study the Aggregation of Amyloidogenic Peptides Using Molecular Dynamics Simulations." In: *J. Chem. Theory Comput.* 14.11 (2018), pp. 6063–6075.

[143] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[144] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[145] Tao Li, Mitsunori Ogihara, and George Tzanetakis. *Music data mining*. CRC Press, 2011.

[146] Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. "Dimensionality reduction: a comparative." In: *J. Mach. Learn. Res.* 10.66-71 (2009), p. 13.

[147] Karl Pearson. "LIII. On lines and planes of closest fit to systems of points in space." In: *London, Edinburgh Dublin Philos. Mag. J. Sci.* 2.11 (1901), pp. 559–572.

[148] Cheng Chu, Sang Kyun Kim, Yian Lin, YuanYuan Yu, Gary Bradski, Andrew Y Ng, and Kunle Olukotun. "Map-reduce for machine learning on multicore." In: *Adv. Neural Inf. Process. Syst.* 19 (2007), p. 281.

[149] Cheng Li and Bingyu Wang. *Principal components analysis*. 2014.

[150] Ian H Witten and Eibe Frank. "Data mining: practical machine learning tools and techniques with Java implementations." In: *ACM Sigmod Rec.* 31.1 (2002), pp. 76–77.

[151] Stuart Lloyd. "Least squares quantization in PCM." In: *IEEE Trans. Inf. Theory* 28.2 (1982), pp. 129–137.

[152] Pierre Legendre and H John B Birks. "Comparing clustering and partitioning strategies." In: *AIP Conference Proceedings*. Vol. 1479. AIP, 2012, pp. 782–785.

[153] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." In: *J. Open Source Softw.* 2.11 (2017), p. 205.

[154] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *KDD*. Vol. 96. 34. AAAI, 1996, pp. 226–231.

[155] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." In: *ACM Trans. Database Syst.* 42.3 (2017), pp. 1–21.

[156] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F. van Gunsteren, and Alan E. Mark. "Peptide Folding: When Simulation Meets Experiment." In: *Angew. Chemie Int. Ed.* 38.1/2 (1999), pp. 236–240.

[157] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers." In: *SoftwareX* 1 (2015), pp. 19–25.

[158] John D Chodera and Frank Noé. "Markov state models of biomolecular conformational dynamics." In: *Curr. Opin. Struct. Biol.* 25 (2014), pp. 135–144.

[159] Christian R Schwantes, Robert T McGibbon, and Vijay S Pande. "Perspective: Markov models for long-timescale biomolecular dynamics." In: *J. Chem. Phys.* 141.9 (2014), 09B201_1.

[160] Vijay S Pande, Kyle Beauchamp, and Gregory R Bowman. "Everything you wanted to know about Markov State Models but were afraid to ask." In: *Methods* 52.1 (2010), pp. 99–105.

[161] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. "Markov models of molecular kinetics: Generation and validation." In: *J. Chem. Phys.* 134.17 (2011), p. 174105.

[162] Martin K Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. "PyEMMA 2: A software package for estimation, validation, and analysis of Markov models." In: *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[163] Matthew P Harrigan, Mohammad M Sultan, Carlos X Hernández, Brooke E Husic, Peter Eastman, Christian R Schwantes, Kyle A Beauchamp, Robert T McGibbon, and Vijay S Pande. "MSMBuilder: statistical models for biomolecular dynamics." In: *Biophys. J.* 112.1 (2017), pp. 10–15.

[164] Yu-Shan Lin, Gregory R Bowman, Kyle A Beauchamp, and Vijay S Pande. "Investigating how peptide length and a pathogenic mutation modify the structural ensemble of amyloid beta monomer." In: *Biophys. J.* 102.2 (2012), pp. 315–324.

[165] Qin Qiao, Gregory R Bowman, and Xuhui Huang. "Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation." In: *J. Am. Chem. Soc.* 135.43 (2013), pp. 16092–16101.

[166] Thomas Löhr, Kai Kohlhoff, Gabriella Heller, Carlo Camilloni, and Michele Vendruscolo. "A kinetic ensemble of the Alzheimer's Aβ peptide." In: *Nature Comput. Sci.* 1 (2021), pp. 71–78.

[167] Nathaniel Stanley, Santiago Esteban-Martín, and Gianni De Fabritiis. "Kinetic modulation of a disordered protein domain by phosphorylation." In: *Nat. Commun.* 5.1 (2014), pp. 1–8.

[168] Alexander-Maurice Illig and Birgit Strodel. "Performance of markov state models and transition networks on characterizing amyloid aggregation pathways from md data." In: *J. Chem. Theory Comput.* 16.12 (2020), pp. 7825–7839.

[169] Yechun Xu, Jianhua Shen, Xiaomin Luo, Weiliang Zhu, Kaixian Chen, Jianpeng Ma, and Hualiang Jiang. "Conformational transition of amyloid β-peptide." In: *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005), pp. 5403–5407.

[170] Olujide O. Olubiyi and Birgit Strodel. "Structures of the Amyloid β-Peptides $A\beta_{1-40}$ and $A\beta_{1-42}$ as Influenced by pH and a d-Peptide." In: *J. Phys. Chem. B* 116.10 (2012), pp. 3280–3291.

[171] Nikolaos G. Sgourakis, Yilin Yan, Scott A. McCallum, Chunyu Wang, and Angel E. García. "The Alzheimer's Peptides Aβ40 and 42 Adopt Distinct Conformations in Water: A Combined MD / NMR Study." In: *J. Mol. Biol.* 368.5 (2007), pp. 1448–1457.

[172] Nikolaos G. Sgourakis, Myrna Merced-Serrano, Christos Boutsidis, Petros Drineas, Zheming Du, Chunyu Wang, and Angel E. García. "Atomic-Level Characterization of the Ensemble of the Aβ(1-42) Monomer in Water Using Unbiased Molecular Dynamics Simulations and Spectral Algorithms." In: *J. Mol. Biol.* 405.2 (2011), pp. 570–583.

[173] Arun Kumar Somavarapu and Kasper P. Kepp. "The Dependence of Amyloid-β Dynamics on Protein Force Fields and Water Models." In: *ChemPhysChem* 16.15 (2015), pp. 3278–3289.

[174] Stacey R. Gerben, Justin A. Lemkul, Anne M. Brown, and David R. Bevan. "Comparing Atomistic Molecular Mechanics Force Fields For a Difficult Target: A Case Study on the Alzheimer's Amyloid β-peptide." In: *J. Biomol. Struct. Dyn.* 32.11 (2014), pp. 1817–1832.

[175] William L. Jorgensen and Julian Tirado-Rives. "The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin." In: *J. Am. Chem. Soc.* 110.6 (1988), pp. 1657–1666.

[176] George A. Kaminski, Richard A. Friesner, Julian Tirado-Rives, and William L. Jorgensen. "Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides." In: *J. Phys. Chem. B* 105.28 (2001), pp. 6474–6487.

[177] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffry D. Madura, Roger W. Impey, and Michael L. Klein. "Comparison of Simple Potential Functions for Simulating Liquid Water." In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[178] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. "Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters." In: *Proteins* 65.3 (2006), pp. 712–725.

[179] Hans W. Horn, William C. Swope, Jed W. Pitera, Jeffry D. Madura, Thomas J. Dick, Greg L. Hura, and Teresa Head-Gordon. "Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew." In: *J. Chem. Phys.* 120.20 (2004), pp. 9665–9678.

[180] Robert B. Best and Gerhard Hummer. "Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides." In: *J. Phys. Chem. B* 113.26 (2009), pp. 9004–9015.

[181] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. "Improved side-chain torsion potentials for the Amber ff99SB protein force field." In: *Proteins* 78.8 (2010), pp. 1950–1958.

[182] Da-Wei Li and Rafael Brüschweiler. "NMR-Based Protein Potentials." In: *Angew. Chem., Int. Ed.* 49.38 (2010), pp. 6778–6780.

[183] Stefano Piana, Kresten Lindorff-Larsen, and David E. Shaw. "How Robust are Protein Folding Simulations with Respect to Force Field Parameterization?" In: *Biophys. J.* 100.9 (2011).

[184] Lee Ping Wang, Todd J. Martinez, and Vijay S. Pande. "Building force fields: An automatic, systematic, and reproducible approach." In: *J. Phys. Chem. Lett.* 5.11 (2014), pp. 1885–1891.

[185] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. De Groot, Helmut Grubmüller, and Alexander D. MacKerell. "CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins." In: *Nat. Methods* 14.1 (2017), pp. 71–73.

[186] Fanjie Meng, Mathias Bellaiche, Jae-Yeol Kim, H. Gul Zerze, B. Robert Best, and Hoi Sung Chung. "Highly Disordered Amyloid-β Monomer Probed by

Single-Molecule FRET and MD Simulation." In: *Biophys. J.* 114.4 (2018), pp. 870–884.

[187] Srabasti Acharya, Kinshuk Srivastava, Sureshbabu Nagarajan, and Lisa Lapidus. "Monomer dynamics of the Alzheimer peptides and Kinetic Control of Early Aggregation in Alzheimer's Disease." In: *ChemPhysChem* 17 (2016), pp. 3470–3479.

[188] Andrew A. Watson, David P. Fairlie, and David J. Craik. "Solution Structure of Methionine-Oxidized Amyloid β-Peptide (1–40). Does Oxidation Affect Conformational Switching?," in: *Biochemistry* 37.37 (1998), pp. 12700–12706.

[189] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical Sampling Through Velocity Rescaling." In: *J. Chem. Phys.* 126.1 (2007). arXiv: 0803.4060.

[190] M. Parrinello and A. Rahman. "Polymorphic transitions in single crystals: A new molecular dynamics method." In: *J. Appl. Phys.* 52.12 (1981), pp. 7182–7190.

[191] Berk Hess, Henk Bekker, Herman J.C. Berendsen, and Johannes G.E.M. Fraaije. "LINCS: A Linear Constraint Solver for molecular simulations." In: *J. Comput. Chem.* 18.12 (1997), pp. 1463–1472.

[192] Tom Darden, Darrin York, and Lee Pedersen. "Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems." In: *J. Chem. Phys.* 98.12 (1993), pp. 10089–10092.

[193] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J.C. Berendsen. *GROMACS: Fast, Flexible, and Free*. 2005.

[194] Sander Pronk et al. "GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit." In: *Bioinformatics* 29.7 (2013), pp. 845–854.

[195] William Humphrey, Andrew Dalke, and Klaus Schulten. "VMD: Visual molecular dynamics." In: *J. Mol. Graph.* 14.1 (1996), pp. 33–38.

[196] Naveen Michaud-Agrawal, Elizabeth J. Denning, Thomas B. Woolf, and Oliver Beckstein. "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations." In: *J. Comput. Chem.* 32.10 (2011), pp. 2319–2327.

[197] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories." In: *Biophys. J.* (2015), pp. 1528 –1532.

[198] Kevin J. Bowers et al. "Scalable algorithms for molecular dynamics simulations on commodity clusters." In: *Proc. 2006 ACM/IEEE Conf. Supercomput. SC'06*. 2006, p. 43. ISBN: 0769527000.

[199] Dmitrij Frishman and Patrick Argos. "Knowledge-based protein secondary structure assignment." In: *Proteins* (1995), pp. 566–579.

[200] Martin K. Scherer, Benjamin Trendelkamp-Schroer, Fabian Paul, Guillermo Pérez-Hernández, Moritz Hoffmann, Nuria Plattner, Christoph Wehmeyer, Jan-Hendrik Prinz, and Frank Noé. "PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models." In: *J. Chem. Theory Comput.* 11.11 (2015), pp. 5525–5542.

[201] Martin K. Scherer, Brooke E. Husic, Moritz Hoffmann, Fabian Paul, Hao Wu, and Frank Noé. "Variational selection of features for molecular kinetics." In: *J. Chem. Phys.* 150.19 (2019), p. 194108.

[202] Guillermo Pérez-Hernández, Fabian Paul, Toni Giorgino, Gianni De Fabritiis, and Frank Noé. "Identification of slow molecular order parameters for Markov model construction." In: *J. Chem. Phys.* 139.1 (2013), p. 015102.

[203] Ricardo Campello, Davoud Moulavi, and Joerg Sander. "Density-Based Clustering Based on Hierarchical Density Estimates." In: *Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science*. Vol. 7819. Springer, Apr. 2013, pp. 160–172.

[204] Susanna Kube and Marcus Weber. "A coarse graining method for the identification of transition rates between molecular conformations." In: *J. Chem. Phys.* 126.2 (2007), p. 024103.

[205] Susanna Röblitz and Marcus Weber. "Fuzzy spectral clustering by PCCA+: Application to Markov state models and data classification." In: *Adv. Data Anal. Classi.* 7 (June 2013), pp. 147–179.

[206] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. "Markov models of molecular kinetics: Generation and validation." In: *J. Chem. Phys.* 134.17 (2011), p. 174105.

[207] Yang Shen and Ad Bax. "SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network." In: *J. Biomol. NMR* (2010), pp. 13–22.

[208] Martin Karplus. "Contact electron-spin coupling of nuclear magnetic moments." In: *J. Chem. Phys.* (1959), pp. 11–15.

[209] Beat Vögeli, Jinfa Ying, Alexander Grishaev, and Ad Bax. "Limits on variations in protein backbone dynamics from precise measurements of scalar couplings." In: *J. Am. Chem. Soc.* (2007), pp. 9377–9385.

[210] James Lincoff, Sukanya Sasmal, and Teresa Head-Gordon. "The combined force field-sampling problem in simulations of disordered amyloid-β peptides." In: *J. Chem. Phys.* 150.10 (2019), p. 104108.

[211] Sandro Bottaro, Tone Bengtsen, and Kresten Lindorff-Larsen. "Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach." In: *Structural Bioinformatics: Methods and Protocols*. Ed. by Zoltán Gáspári. New York, NY: Springer US, 2020, pp. 219–240. ISBN: 978-1-0716-0270-6.

[212] Julien Roche, Yang Shen, Jung Ho Lee, Jinfa Ying, and Ad Bax. "Monomeric Aβ1–40 and β1–42 Peptides in Solution Adopt Very Similar Ramachandran Map Distributions That Closely Resemble Random Coil." In: *Biochemistry* 55.5 (2016), pp. 762–775.

[213] D.S. Wishart and D.A. Case. "Use of chemical shifts in macromolecular structure determination." In: *Methods Enzymol.* 338 (2001), pp. 3–34.

[214] Magnus Kjaergaard and Flemming Poulsen. "Sequence correction of random coil chemical shifts: Correlation between neighbor correction factors and changes in the Ramachandran distribution." In: *J. Biomol. NMR* 50 (June 2011), pp. 157–65.

[215] Magnus Kjaergaard, Søren Brander, and Flemming Poulsen. "Random coil chemical shift for intrinsically disordered proteins: Effects of temperature and pH." In: *J. Biomol. NMR* 49 (Feb. 2011), pp. 139–49.

[216] Alexey B. Mantsyzov, Alexander S. Maltsev, Jinfa Ying, Yang Shen, Gerhard Hummer, and Ad Bax. "A maximum entropy approach to the study of residue-specific backbone angle distributions in α-synuclein, an intrinsically disordered protein." In: *Protein Sci.* 23.9 (2014), pp. 1275–1290.

[217] Alexey Mantsyzov, Yang Shen, Jung Lee, Gerhard Hummer, and Ad Bax. "MERA: A webserver for evaluating backbone torsion angle distributions in dynamic and disordered proteins from NMR data." In: *J. Biomol. NMR* 63 (2015), pp. 85–95.

[218] Robert Best, Nicolae-Viorel Buchete, and Gerhard Hummer. "Are Current Molecular Dynamics Force Fields too Helical?" In: *Biophys. J.* 95 (Aug. 2008), pp. L07–9.

[219] Debayan Chakraborty, John Straub, and D. Thirumalai. "Differences in the free energies between the excited states of Aβ40 and Aβ42 monomers encode their aggregation propensities." In: *Proc. Natl. Acad. Sci. U.S.A.* (2020), p. 202002570.

[220] Daniele Granata, Fahimeh Baftizadeh, Johnny Habchi, Celine Galvagnion, Alfonso De Simone, Carlo Camilloni, Alessandro Laio, and Michele Vendruscolo. "The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments." In: *Sci. Rep.* 5 (2015), p. 15449.

[221] Mads Nygaard, Birthe B. Kragelund, Elena Papaleo, and Kresten Lindorff-Larsen. "An Efficient Method for Estimating the Hydrodynamic Radius of Disordered Protein Conformations." In: *Biophys. J.* 113.3 (2017), pp. 550–557.

[222] Nasrollah Rezaei-Ghaleh, Giacomo Parigi, and Markus Zweckstetter. "Reorientational Dynamics of Amyloid-β from NMR Spin Relaxation and Molecular Simulation." In: *J. Phys. Chem. Lett.* 10.12 (2019), pp. 3369–3375.

[223] S. Jephthah, L. Staby, B. B. Kragelund, and M. Skepö. "Temperature Dependence of Intrinsically Disordered Proteins in Simulations: What are We Missing?" In: *J. Chem. Theory Comput.* 15.4 (2019), pp. 2672–2683.

[224] Andreas Haahr Larsen, Yong Wang, Sandro Bottaro, Sergei Grudinin, Lise Arleth, and Kresten Lindorff-Larsen. "Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multidomain proteins in solution." In: *PLoS Comput. Biol.* 16.4 (Apr. 2020), pp. 1–29.

[225] D. Thirumalai and Raymond D. Mountain. "Activated dynamics, loss of ergodicity, and transport in supercooled liquids." In: *Phys. Rev. E* 47 (1993), pp. 479–489.

[226] K. Aurelia Ball, Aaron H. Phillips, Paul S. Nerenberg, Nicolas L. Fawzi, David E. Wemmer, and Teresa Head-Gordon. "Homogeneous and Heterogeneous Tertiary Structure Ensembles of Amyloid-β Peptides." In: *Biochemistry* 50 (2011), pp. 7612–7628.

[227] Ernesto Suárez, Rafal Wiewiora, Chris Wehmeyer, Frank Noé, John Chodera, and Daniel Zuckerman. "What Markov state models can and cannot do: Correlation versus path-based observables in protein folding models." In: *bioRxiv* (Nov. 2020), 00i:10.1101/2020.11.09.374496.

[228] Vladimir Uversky. "Intrinsically Disordered Proteins and Their (Disordered) Proteomes in Neurodegenerative Disorders." In: *Front. Aging Neurosci.* 7 (Mar. 2015), p. 18.

[229] Fabrizio Chiti and Christopher M. Dobson. "Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade." In: *Annu. Rev. Biochem.* 86.1 (2017), pp. 27–68.

[230] Robert B. Best. "Computational and theoretical advances in studies of intrinsically disordered proteins." In: *Curr. Opin. Struct. Biol.* 42 (2017), pp. 147–154.

[231] Jing Huang and Alexander D MacKerell. "Force Field Development and Simulations of Intrinsically Disordered Proteins." In: *Curr. Opin. Struct. Biol.* 48 (2018), pp. 40–48.

[232] Martín Carballo-Pacheco, Ahmed E. Ismail, and Birgit Strodel. "On the Applicability of Force Fields To Study the Aggregation of Amyloidogenic Peptides Using Molecular Dynamics Simulations." In: *J. Chem. Theory Comput.* 14 (2018), pp. 6063–6075.

[233] Nathan Schmid, Andreas P. Eichenberger, Alexandra Choutko, Sereina Riniker, Moritz Winger, Alan E. Mark, and Wilfred F. Van Gunsteren. "Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7." In: *Eur. Biophys. J.* 40.7 (2011), pp. 843–856.

[234] Phuong H. Nguyen, Mai Suan Li, and Philippe Derreumaux. "Effects of all-atom force fields on amyloid oligomerization: Replica exchange molecular dynamics simulations of the Aβ16-22 dimer and trimer." In: *Phys. Chem. Chem. Phys.* 13.20 (2011), pp. 9778–9788.

[235] Matthias Buck, Sabine Bouguet-Bonnet, Richard W. Pastor, and Alexander D. MacKerell. "Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme." In: *Biophys. J.* 90.4 (2006).

[236] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E.M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. "Optimization of the Additive CHARMM All-atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-chain $\chi 1$ and $\chi 2$ Dihedral Angles." In: *J. Chem. Theory Comput.* 8 (2012), pp. 3257–3273.

[237] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. "How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules?" In: *J. Comput. Chem.* 21.12 (2000), p. 1049.

[238] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. "Improved Side-chain Torsion Po-

tentials for the Amber ff99SB Protein Force Field." In: *Proteins* 78.8 (2010), pp. 1950–1958.

[239] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB." In: *J. Chem. Theory Comput.* 11.8 (2015), pp. 3696–3713.

[240] J. J. Balbach, Y. Ishii, O. N. Antzutkin, R. D. Leapman, N. W. Rizzo, F. Dyda, J. Reed, and R. Tycko. "Amyloid fibril formation by Aβ16-22, a seven-residue fragment of the Alzheimer's β-amyloid peptide, and structural characterization by solid state NMR." In: *Biochemistry* 39.45 (2000), pp. 13748–13759.

[241] F. Timur Senguen, Todd M. Doran, Elizabeth A. Anderson, and Bradley L. Nilsson. "Clarifying the influence of core amino acid hydrophobicity, secondary structure propensity, and molecular volume on amyloid-β 16-22 self-assembly." In: *Mol. BioSyst.* 7.2 (2011), pp. 497–510.

[242] Sanjai Kumar Pachahara and Ramakrishnan Nagaraj. "Probing the Role of Aromatic Residues in the Self-assembly of Aβ(16–22) in Fluorinated Alcohols and Their Aqueous Mixtures." In: *Biochem. Biophys. Rep.* 2 (2015), pp. 1–13.

[243] Kai Tao, Jiqian Wang, Peng Zhou, Chengdong Wang, Hai Xu, Xiubo Zhao, and Jian R. Lu. "Self-assembly of Short Aβ(16-22) Peptides: Effect of Terminal Capping and the Role of Electrostatic Interaction." In: *Langmuir* 27.6 (2011), pp. 2723–2730.

[244] Dmitri Klimov and D Thirumalai. "Dissecting the Assembly of $A\beta_{16-22}$ Amyloid Peptides into Antiparallel β Sheets." In: *Structure* 11 (Mar. 2003), pp. 295–307.

[245] S. Gnanakaran, Ruth Nussinov, and Angel E. García. "Atomic-Level Description of Amyloid β-Dimer Formation." In: *J. Am. Chem. Soc.* 128.7 (2006), pp. 2158–2159.

[246] Phuong H. Nguyen, Mai Suan Li, Gerhard Stock, John E. Straub, and D. Thirumalai. "Monomer adds to preformed structured oligomers of Aβ-peptides by

a two-stage dock–lock mechanism." In: *Proc. Natl. Acad. Sci. USA* 104 (2007), pp. 111–116.

[247] Mookyung Cheon, Iksoo Chang, and Carol K. Hall. "Spontaneous Formation of Twisted Aβ 16-22 Fibrils in Large-Scale Molecular-Dynamics Simulations." In: *Biophys. J.* 101.10 (2011), pp. 2493–2501.

[248] Bogdan Barz, David J. Wales, and Birgit Strodel. "A Kinetic Approach to the Sequence–Aggregation Relationship in Disease-Related Protein Assembly." In: *J. Phys. Chem. B* 118.4 (2014), pp. 1003–1011.

[249] Yiming Wang, Samuel J. Bunce, Sheena E. Radford, Andrew J. Wilson, Stefan Auer, and Carol K. Hall. "Thermodynamic Phase Diagram of Amyloid-β (16–22) Peptide." In: *Proc. Natl. Acad. Sci. U.S.A.* 116.6 (2019), pp. 2091–2096.

[250] Phuong H. Nguyen, Fabio Sterpone, and Philippe Derreumaux. "Chapter Nine - Aggregation of disease-related peptides." In: *Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*. Ed. by Birgit Strodel and Bogdan Barz. Vol. 170. Prog. Mol. Biol. Transl. Sci. Academic Press, 2020, pp. 435 –460.

[251] Basir Ahmad, Yujie Chen, and Lisa J. Lapidus. "Aggregation of α-Synuclein is Kinetically Controlled by Intramolecular Diffusion." In: *Proc. Natl. Acad. Sci. U.S.A.* 109.7 (2012), pp. 2336–2341. eprint: http://www.pnas.org/content/109/7/2336.full.pdf.

[252] Lisa J. Lapidus. "Understanding Protein Aggregation from the View of Monomer Dynamics." In: *Mol. BioSyst.* 9 (1 2013), pp. 29–35.

[253] Pavel I. Zhuravlev, Govardhan Reddy, John E. Straub, and D. Thirumalai. "Propensity to Form Amyloid Fibrils Is Encoded as Excitations in the Free Energy Landscape of Monomeric Proteins." In: *J. Mol. Biol.* 426.14 (2014), pp. 2653–2666.

[254] Pekka Mark and Lennart Nilsson. "Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K." In: *J. Phys. Chem. A* 105.43 (2001), pp. 9954–9960.

[255] H. J.C. Berendsen. "Biophysical Applications of Molecular Dynamics." In: *Comput. Phys. Commun.* 44.3 (1987), pp. 233–242.

[256] M. Parrinello and A. Rahman. "Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method." In: *J. Appl. Phys.* 52.12 (1981), pp. 7182–7190.

[257] K. Anton Feenstra, Berk Hess, and Herman J.C. Berendsen. "Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-rich Systems." In: *J. Comput. Chem.* 20.8 (1999), pp. 786–798.

[258] L. Martinez, R. Andrade, E. G. Birgin, and J. M. Martínez. "PACKMOL: A Package for Building Initial Configurations for Molecular Dynamics Simulations." In: *J. Comput. Chem.* 30.13 (2009), pp. 2157–2164.

[259] Michael R. Sawaya et al. "Atomic Structures of Amyloid Cross-β Spines Reveal Varied Steric Zippers." In: *Nature* 447.7143 (2007), pp. 453–457.

[260] Jacques Philippe Colletier, Arthur Laganowsky, Meytal Landau, Minglei Zhao, Angela B. Soriaga, Lukasz Goldschmidt, David Flot, Duilio Cascio, Michael R. Sawaya, and David Eisenberg. "Molecular Basis for Amyloid-β Polymorphism." In: *Proc. Natl. Acad. Sci. U.S.A.* 108.41 (2011), pp. 16938–16943.

[261] W L DeLano. "The PyMOL Molecular Graphics System, Version 1.8." In: *Schrödinger LLC* (Date of access: 12-01-2018), http://www.pymol.org.

[262] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories." In: *Biophys. J.* (2015).

[263] Laura Riccardi, Phuong H. Nguyen, and Gerhard Stock. "Construction of the Free Energy Landscape of Peptide Aggregation from Molecular Dynamics Simulations." In: *J. Chem. Theory Comput.* 8 (2012), pp. 1471–1479.

[264] Bogdan Barz, Qinghua Liao, and Birgit Strodel. "Pathways of Amyloid-β Aggregation Depend on Oligomer Shape." In: *J. Am. Chem. Soc.* 140.1 (2018), pp. 319–327.

[265] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks." In: *International AAAI Conference on Weblogs and Social Media* (2009).

[266] Joseph E. Basconi and Michael R. Shirts. "Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations." In: *J. Chem. Theory Comput.* 9.7 (2013), pp. 2887–2899.

[267] Efrem Braun, Seyed Mohamad Moosavi, and Berend Smit. "Anomalous Effects of Velocity Rescaling Algorithms: The Flying Ice Cube Effect Revisited." In: *J. Chem. Theory Comput.* 14.10 (2018), pp. 5262–5272.

[268] Arghadwip Paul, Suman Samantray, Marco Anteghini, Mohammed Khaled, and Birgit Strodel. "Thermodynamics and kinetics of the amyloid-β peptide revealed by Markov state models based on MD data in agreement with experiment." In: *Chem. Sci.* 12.19 (2021), pp. 6652–6669.

[269] Kazuo Fujiwara, Hiromi Toda, and Masamichi Ikeguchi. "Dependence of α-helical and β-sheet Amino Acid Propensities on the Overall Protein Fold Type." In: *BMC Struct. Biol.* 12 (Aug. 2012), p. 18.

[270] Filip Persson, Pär Söderhjelm, and Bertil Halle. "The Geometry of Protein Hydration." In: *J. Chem. Phys.* 148.21 (2018), p. 215101.

[271] Massimo Marchi, Fabio Sterpone, and Matteo Ceccarelli. "Water Rotational Relaxation and Diffusion in Hydrated Lysozyme." In: *J. Am. Chem. Soc.* 124.23 (2002), pp. 6787–6791.

[272] Fabio Sterpone, Matteo Ceccarelli, and Massimo Marchi. "Dynamics of Hydration in Hen Egg White Lysozyme." In: *J. Mol. Biol.* 311.2 (2001), pp. 409–419.

[273] Nikolai Smolin and Roland Winter. "Molecular Dynamics Simulations of Staphylococcal Nuclease: Properties of Water at the Protein Surface." In: *J. Phys. Chem. B* 108.40 (2004), pp. 15928–15937.

[274] Sudipta K.umar Sinha, Madhurima Jana, Kausik Chakraborty, and Sanjoy Bandyopadhyay. "In silico Studies of the Properties of Water Hydrating a Small Protein." In: *J. Chem. Phys.* 141.22 (2014), 12B602_1.

[275] Filip Persson, Pär Söderhjelm, and Bertil Halle. "The Spatial Range of Protein Hydration." In: *J. Chem. Phys.* 148.21 (2018), p. 215104.

[276] M. Tarek and D. J. Tobias. "Role of Protein-Water Hydrogen Bond Dynamics in the Protein Dynamical Transition." In: *Phys. Rev. Lett.* 88.13 (2002), p. 4.

[277] Sheh Yi Sheu, Dah Yen Yang, H. L. Selzle, and E. W. Schlag. "Energetics of Hydrogen Bonds in Peptides." In: *Proc. Natl. Acad. Sci. U.S.A.* 100.22 (2003), pp. 12683–12687.

[278] Jens Danielsson, Jüri Jarvet, Peter Damberg, and Astrid Gräslund. "Translational Diffusion Measured by PFG-NMR on Full Length and Fragments of the Alzheimer Aβ(1–40) Peptide. Determination of Hydrodynamic Radii of Random Coil Peptides of Varying Length." In: *Magn. Reson. Chem.* 40.13 (2002), S89–S97.

[279] Yijin Mao and Yuwen Zhang. "Thermal Conductivity, Shear Viscosity and Specific Heat of Rigid Water Models." In: *Chem. Phys. Lett.* 542 (July 2012), pp. 37–41.

[280] Sébastien Santini, Guanghong Wei, Normand Mousseau, and Philippe Derreumaux. "Pathway Complexity of Alzheimer's Beta-amyloid Abeta16-22 Peptide Assembly." In: *Structure* 12.7 (2004), 1245—1255.

[281] S. Santini, N. Mousseau, and P. Derreumaux. "In silico assembly of Alzheimer's $A\beta_{16-22}$ peptide into β-sheets." In: *J. Am. Chem. Soc.* 126 (2004), pp. 11509–11516.

[282] Zhiguang Jia, Jeremy D. Schmit, and Jianhan Chen. "Amyloid Assembly is Dominated by Misregistered Kinetic Traps on an Unbiased Energy Landscape." In: *Proc. Natl. Acad. Sci. U.S.A.* 117.19 (2020), pp. 10322–10328.

[283] Gregory Darnell, Joseph P.R.O. Orgel, Reinhard Pahl, and Stephen C. Meredith. "Flanking Polyproline Sequences Inhibit β-Sheet Structure in Polyglutamine Segments by Inducing PPII-like Helix Structure." In: *J. Mol. Biol.* 374.3 (2007), pp. 688 –704.

[284] Michael Kuemin, Sabine Schweizer, Christian Ochsenfeld, and Helma Wennemers. "Effects of Terminal Functional Groups on the Stability of the Polyproline II Structure: A Combined Experimental and Theoretical Study." In: *J. Am. Chem. Soc.* 131 (2009), pp. 15474–15482.

[285] Charles Nicholson and Eva Syková. "Extracellular space structure revealed by diffusion analysis." In: *Trends Neurosci.* 21.5 (1998), pp. 207–215.

[286] Richard L Jackson, Steven J Busch, and Alan D Cardin. "Glycosaminoglycans: molecular properties, protein interactions, and role in physiological processes." In: *Physiol. Rev.* 71.2 (1991), pp. 481–539.

[287] Alexander Dityatev, Melitta Schachner, and Peter Sonderegger. "The dual role of the extracellular matrix in synaptic plasticity and homeostasis." In: *Nat. Rev. Neurosci.* 11.11 (2010), pp. 735–746.

[288] Khushboo Gulati and Krishna Mohan Poluri. "Mechanistic and therapeutic overview of glycosaminoglycans: the unsung heroes of biomolecular signaling." In: *Glycoconj. J.* 33.1 (2016), pp. 1–17.

[289] Neda Motamedi-Shad, Elodie Monsellier, and Fabrizio Chiti. "Amyloid formation by the model protein muscle acylphosphatase is accelerated by heparin and heparan sulphate through a scaffolding-based mechanism." In: *J. Biochem.* 146.6 (2009), pp. 805–814.

[290] Toshio Ariga, Tadashi Miyatake, and Robert K. Yu. "Role of proteoglycans and glycosaminoglycans in the pathogenesis of Alzheimer's disease and related disorders: Amyloidogenesis and therapeutic strategies - A review." In: *J. Neurosci. Res.* 88.11 (2010), pp. 2303–2315.

[291] Antonio Gigante and Leonardo Callegari. "The role of intra-articular hyaluronan (Sinovial®) in the treatment of osteoarthritis." In: *Rheumatol. Int.* 31.4 (2011), pp. 427–444.

[292] Alberto Migliore and Mauro Granata. "Intra-articular use of hyaluronic acid in the treatment of osteoarthritis." In: *Clin. Interv. Aging* 3.2 (2008), p. 365.

[293] Vilím Simanek, Vladimír Kren, Jitka Ulrichová, and Jiří Gallo. "The efficacy of glucosamine and chondroitin sulfate in the treatment of osteoarthritis: are these saccharides drugs or nutraceuticals." In: *Biomed. Pap. Med. Fac. Univ Palacky. Olomouc. Czech. Repub.* 149.1 (2005), pp. 51–6.

[294] S Salvatore, R Heuschkel, S Tomlin, SE Davies, S Edwards, JA Walker-Smith, I French, and SH Murch. "A pilot study of N-acetyl glucosamine, a nutritional substrate for glycosaminoglycan synthesis, in paediatric chronic inflammatory bowel disease." In: *Aliment. Pharmacol. Ther.* 14.12 (2000), pp. 1567–1579.

[295] Vitor H Pomin. "Keratan sulfate: an up-to-date review." In: *Int. J. Biol. Macromol.* 72 (2015), pp. 282–289.

[296] Fuming Zhang, Zhenqing Zhang, and Robert J. Linhardt. "Glycosaminoglycans." In: *Handb. Glycomics*. Elsevier, 2010, pp. 59–80. ISBN: 9780123736000.

[297] Courtney Mycroft-West et al. "Glycosaminoglycans induce conformational change in the SARS-CoV-2 Spike S1 Receptor Binding Domain." In: *BioRxiv* (2020).

[298] E Fouissac, M Milas, M Rinaudo, and R Borsali. "Influence of the ionic strength on the dimensions of sodium hyaluronate." In: *Macromolecules* 25.21 (1992), pp. 5613–5617.

[299] Kanako Hayashi, Kikuko Tsutsumi, Fumio Nakajima, Takashi Norisuye, and Akio Teramoto. "Chain-stiffness and excluded-volume effects in solutions of sodium hyaluronate at high ionic strength." In: *Macromolecules* 28.11 (1995), pp. 3824–3830.

[300] Raniero Mendichi, Ladislav Šoltés, and Alberto Giacometti Schieroni. "Evaluation of radius of gyration and intrinsic viscosity molar mass dependence and stiffness of hyaluronan." In: *Biomacromolecules* 4.6 (2003), pp. 1805–1810.

[301] Lenka Musilová, Věra Kašpárková, Aleš Mráček, Antonín Minařík, and Martin Minařík. "The behaviour of hyaluronan solutions in the presence of Hofmeister ions: A light scattering, viscometry and surface tension study." In: *Carbohydr. Polym.* 212 (2019), pp. 395–402.

[302] Cristal I Gama, Sarah E Tully, Naoki Sotogaku, Peter M Clark, Manish Rawat, Nagarajan Vaidehi, William A Goddard, Akinori Nishi, and Linda C Hsieh-Wilson. "Sulfation patterns of glycosaminoglycans encode molecular recognition and activity." In: *Nat. Chem. Biol.* 2.9 (2006), pp. 467–473.

[303] Diana Soares da Costa, Rui L Reis, and Iva Pashkuleva. "Sulfation of glycosaminoglycans and its implications in human health and disorders." In: *Annu. Rev. Biomed. Eng.* 19 (2017), pp. 1–26.

[304] Yanjie Zhang and Paul S Cremer. "Interactions between macromolecules and ions: the Hofmeister series." In: *Curr. Opin. Chem. Biol.* 10.6 (2006), pp. 658–663.

[305] Raymond AA Smith, Kate Meade, Claire E Pickford, Rebecca J Holley, and Catherine LR Merry. "Glycosaminoglycans as regulators of stem cell differentiation." In: *Biochem. Soc. Trans* 39.1 (2011), pp. 383–387.

[306] Leyla Gasimli, Anne Marie Hickey, Bo Yang, Guoyun Li, Mitche dela Rosa, Alison V Nairn, Michael J Kulik, Jonathan S Dordick, Kelley W Moremen, Stephen Dalton, et al. "Changes in glycosaminoglycan structure on differentiation of human embryonic stem cells towards mesoderm and endoderm lineages." In: *Biochim. Biophys. Acta, Gen. Subj.* 1840.6 (2014), pp. 1993–2003.

[307] TR Rudd, MA Skidmore, M Guerrini, M Hricovini, AK Powell, G Siligardi, and EA Yates. "The conformation and structure of GAGs: recent progress and perspectives." In: *Curr. Opin. Struct. Biol.* 20.5 (2010), pp. 567–574.

[308] Eva Kutálková, Josef Hrnčiřík, Roman Witasek, and Marek Ingr. "Effect of solvent and ions on the structure and dynamics of a hyaluronan molecule." In: *Carbohydr. Polym.* 234 (2020), p. 115919.

[309] Benedict M Sattelle, Steen U Hansen, John Gardiner, and Andrew Almond. "Free energy landscapes of iduronic acid and related monosaccharides." In: *J. Am. Chem. Soc.* 132.38 (2010), pp. 13132–13134.

[310] Benedict M Sattelle, Javad Shakeri, and Andrew Almond. "Does microsecond sugar ring flexing encode 3D-shape and bioactivity in the heparanome?" In: *Biomacromolecules* 14.4 (2013), pp. 1149–1159.

[311] Sergey A Samsonov, Jan-Philip Gehrcke, and M Teresa Pisabarro. "Flexibility and explicit solvent in molecular-dynamics-based docking of protein–glycosaminoglycan systems." In: *J. Chem. Inf. Model.* 54.2 (2014), pp. 582–592.

[312] Balaji Nagarajan, Nehru Viji Sankaranarayanan, and Umesh R Desai. "Rigorous analysis of free solution glycosaminoglycan dynamics using simple, new tools." In: *Glycobiology* 30.8 (2020), pp. 516–527.

[313] Andrew Almond, Andy Brass, and John K Sheehan. "Dynamic exchange between stabilized conformations predicted for hyaluronan tetrasaccharides: comparison of molecular dynamics simulations with available NMR data." In: *Glycobiology* 8.10 (1998), pp. 973–980.

[314] Geraldine Cilpa, Marja T Hyvönen, Artturi Koivuniemi, and M-L Riekkola. "Atomistic insight into chondroitin-6-sulfate glycosaminoglycan chain through quantum mechanics calculations and molecular dynamics simulation." In: *J. Comput. Chem.* 31.8 (2010), pp. 1670–1680.

[315] Nicolas Sapay, Eric Cabannes, Maurice Petitou, and Anne Imberty. "Molecular modeling of the interaction between heparan sulfate and cellular growth factors: bringing pieces together." In: *Glycobiology* 21.9 (2011), pp. 1181–1193.

[316] Andrew Almond. "Multiscale modeling of glycosaminoglycan structure and dynamics: Current methods and challenges." In: *Curr. Opin. Struct. Biol.* 50 (2018), pp. 58–64.

[317] Nehru Viji Sankaranarayanan, Balaji Nagarajan, and Umesh R Desai. "So you think computational approaches to understanding glycosaminoglycan–protein interactions are too dry and too rigid? Think again!" In: *Curr. Opin. Struct. Biol.* (2018), pp. 91–100.

[318] Olgun Guvench, Elizabeth Hatcher, Richard M. Venable, Richard W. Pastor, and Alexander D. MacKerell Jr. "CHARMM Additive All-Atom Force Field for Glycosidic Linkages between Hexopyranoses." In: *J. Chem. Theory Comput.* 5.9 (2009), pp. 2353–2370.

[319] E Prabhu Raman, Olgun Guvench, and Alexander D MacKerell Jr. "CHARMM additive all-atom force field for glycosidic linkages in carbohydrates involving furanoses." In: *J. Phys. Chem. B* 114.40 (2010), pp. 12981–12994.

[320] Olgun Guvench, Sairam S Mallajosyula, E Prabhu Raman, Elizabeth Hatcher, Kenno Vanommeslaeghe, Theresa J Foster, Francis W Jamison, and Alexander D MacKerell Jr. "CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate–protein modeling." In: *J. Chem. Theory Comput.* 7.10 (2011), pp. 3162–3180.

[321] Sairam S Mallajosyula, Olgun Guvench, Elizabeth Hatcher, and Alexander D MacKerell Jr. "CHARMM additive all-atom force field for phosphate and sulfate linked to carbohydrates." In: *J. Chem. Theory Comput.* 8.2 (2012), pp. 759–776.

[322] Sunhwan Jo, Kevin C. Song, Heather Desaire, Alexander D. MacKerell, and Wonpil Im. "Glycan reader: Automated sugar identification and simulation preparation for carbohydrates and glycoproteins." In: *J. Comput. Chem.* 32.14 (2011), pp. 3135–3141.

[323] Sang Jun Park, Jumin Lee, Dhilon S. Patel, Hongjing Ma, Hui Sun Lee, Sunhwan Jo, and Wonpil Im. "Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank." In: *Bioinformatics* 33.19 (2017), pp. 3051–3057.

[324] Sang Jun Park, Jumin Lee, Yifei Qi, Nathan R. Kern, Hui Sun Lee, Sunhwan Jo, Insuk Joung, Keehyung Joo, Jooyoung Lee, and Wonpil Im. "CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates." In: *Glycobiology* 29.4 (2019), pp. 320–331.

[325] Sunhwan Jo, Taehoon Kim, Vidyashankara G. Iyer, and Wonpil Im. "CHARMM-GUI: A web-based graphical user interface for CHARMM." In: *J. Comput. Chem.* 29.11 (2008), pp. 1859–1865.

[326] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. "GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation." In: *J. Chem. Theory Comput.* 4.3 (2008), pp. 435–447.

[327] Justin Lemkul. "From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]." In: *Living J. Comput. Mol. Sci.* 1.1 (2019), p. 5068.

[328] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee Ping Wang, Thomas J. Lane, and Vijay S. Pande. "MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories." In: *Biophys. J.* 109.8 (2015), pp. 1528–1532.

[329] Suman Samantray and David L. Cheung. "Effect of the air–water interface on the conformation of amyloid beta." In: *Biointerphases* 15.6 (2020), p. 061011.

[330] James D Sterling, Wenjuan Jiang, Wesley M Botello-Smith, and Yun L Luo. "Ion pairing and dielectric decrement in glycosaminoglycan brushes." In: *J. Phys. Chem. B* 125.10 (2021), pp. 2771–2780.

[331] Jaibir Kherb, Sarah C. Flores, and Paul S. Cremer. "Role of Carboxylate Side Chains in the Cation Hofmeister Series." In: *J. Phys. Chem. B* 116.25 (2012), pp. 7389–7397.

[332] Emad F. Aziz, Niklas Ottosson, Stefan Eisebitt, Wolfgang Eberhardt, Barbara Jagoda-Cwiklik, Robert Vácha, Pavel Jungwirth, and Bernd Winter. "Cation-Specific Interactions with Carboxylate in Amino Acid and Acetate Aqueous Solutions: X-ray Absorption and ab initio Calculations." In: *J. Phys. Chem. B* 112.40 (2008), pp. 12567–12570.

[333] Alan D. Snow and Thomas N. Wight. "Proteoglycans in the pathogenesis of Alzheimer's disease and other amyloidoses." In: *Neurobiol. Aging* 10.5 (1989), pp. 481–497.

[334] John B. Ancsin. "Amyloidogenesis: Historical and modern observations point to heparan sulfate proteoglycans as a major culprit." In: *Amyloid* 10.2 (2003), pp. 67–79.

[335] Doris Lam, Heather A. Enright, Jose Cadena, Sandra K.G. Peters, Ana Paula Sales, Joanne J. Osburn, David A. Soscia, Kristen S. Kulp, Elizabeth K. Wheeler, and Nicholas O. Fischer. "Tissue-specific extracellular matrix accelerates the

formation of neural networks and communities in a neuron-glia co-culture on a multi-electrode array." In: *Sci. Rep.* 9.1 (2019).

[336] Michael C. Owen, David Gnutt, Mimi Gao, Sebastian K.T.S. Wärmländer, Jüri Jarvet, Astrid Gräslund, Roland Winter, Simon Ebbinghaus, and Birgit Strodel. "Effects of in vivo conditions on amyloid aggregation." In: *Chem. Soc. Rev.* 48.14 (2019), pp. 3946–3996.

[337] Papy-Garcia Dulce, Morin Christophe, Huynh Minh Bao, Sineriz Fernando, Sissoeff Ludmilla, Sepulveda Diaz Julia Elisa, and Raisman-Vozari Rita. "Glycosaminoglycans, Protein Aggregation and Neurodegeneration." In: *Curr. Protein Pept. Sci.* 999.999 (2011), pp. 1–11.

[338] Gerardo Castillo, Widia Lukito, Thomas Wight, and Alan Snow. "The Sulfate Moieties of Glycosaminoglycans Are Critical for the Enhancement of β-Amyloid Protein Fibril Formation." In: *J. Neurochem.* 72 (2001), pp. 1681–1687.

[339] Kazuchika Nishitsuji and Kenji Uchimura. "Sulfated glycosaminoglycans in protein aggregation diseases." In: *Glycoconj. J.* 34 (Aug. 2017), pp. 453–466.

[340] Lena Kjellén and Ulf Lindahl. "Proteoglycans: structures and interactions." In: *Annu. Rev. Biochem.* 60.1 (1991), pp. 443–475.

[341] Alexander-Maurice Illig and Birgit Strodel. "Performance of Markov State Models and Transition Networks on Characterizing Amyloid Aggregation Pathways from MD Data." In: *J. Chem. Theory Comput.* 16 (2020), pp. 7825–7839.

[342] Suman Samantray, Wibke Schumann, Alexander-Maurice Illig, Martin Carballo-Pacheco, Arghadwip Paul, Bogdan Barz, and Birgit Strodel. "Molecular dynamics simulations of protein aggregation: protocols for simulation setup and analysis with Markov state models and transition networks." In: *bioRxiv* (2020), 00i:10.1101/2020.04.25.060269.

[343] Man Hoang Viet, Son Tung Ngo, Nguyen Sy Lam, and Mai Suan Li. "Inhibition of Aggregation of Amyloid Peptides by Beta-Sheet Breaker Peptides and Their Binding Affinity." In: *J. Phys. Chem. B* 115.22 (2011), pp. 7433–7446.

[344] Dmitrij Frishman and Patrick Argos. "Knowledge-based protein secondary structure assignment." In: *Proteins Struct. Funct. Bioinforma.* 23.4 (1995), pp. 566–579.

[345] Jason Greenwald and Roland Riek. "Biology of amyloid: structure, function, and regulation." In: *Structure* 18.10 (2010), pp. 1244–1260.

[346] Gaetano Invernizzi, Elena Papaleo, Raimon Sabate, and Salvador Ventura. "Protein aggregation: mechanisms and functional consequences." In: *Int. J. Biochem. Cell Biol.* 44.9 (2012), pp. 1541–1554.

[347] Deling Shi, Anran Sheng, and Lianli Chi. "Glycosaminoglycan-protein interactions and their roles in human disease." In: *Front. Mol. Biosci.* 8 (2021), p. 64.