Molecular Recognition and Selectivity: Computational Investigations on the Dynamics of Non-bonded Interactions

Dissertation
Zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

von **Eric Schulze-Niemand, M. Sc.**

geb. am **01.08.1988** in **Wolmirstedt**

genehmigt durch die Fakultät für Verfahrens- und Systemtechnik
der Otto-von-Guericke-Universität Magdeburg

Promotionskommission:
Prof. Dr.-Ing. Kai Sundmacher (Vorsitz)
Hon.-Prof. Dr. rer. nat. Matthias Stein (Gutachter)
Prof. Dr. rer. nat. Michael Naumann (Gutachter)
Prof. Dr. Rebecca Wade (Mitglied)

Eingereicht am: **25.02.2022**

Promotionskolloquium am: **27.06.2022**

"When the power of love overcomes the love of power the world will know peace."

— Jimi Hendrix

ABSTRACT

Non-bonded interactions, such as hydrogen bonds, as well as hydrophobic and electrostatic interactions determine structure and dynamics of flexible molecules and multi-molecular assemblies. In single molecules, they selectively enable and stabilize rare, energetically unfavorable conformations which facilitate intramolecular chemical reactions or reaction with the solvent molecules. Such reactions often result in changes of surface charges with far-reaching effects on the molecular properties. Additionally, non-bonded interactions mediate the association of molecules to transient aggregates and stable complexes. The complementarity of interaction donors and acceptors on two molecular surfaces is the basis for their pairwise recognition. Selective recognition of distinct molecules or chemical groups within a single molecule is a fundamental aspect of cellular life as well as of artificial chemical systems.

Experimental methods often measure the macroscopic consequences of non-bonded interactions instead of the interaction themselves. More elaborate techniques are expensive and error-prone and still only yield limited insight. An experimental means to assess molecular interactions with high spatial and temporal resolution has not yet been proposed. In recent years, with the rise of graphics processing units and the increase in easily available computing power, the theoretical Molecular Dynamics (MD) method has emerged as standard tool to investigate the time-resolved behavior of molecular structures and interactions. A cornucopia of condensed phase molecular systems has been to the subject of MD simulations, yet with varying rigor in preparation, force-field selection, and quantitative analysis. Even though different questions require different analytics, an absence of comparable, generally applicable means to analyze and visualize non-bonded interactions and their effects from MD trajectory data can be stated.

In this work, dynamical aspects of non-bonded interactions as the basis for molecular selectivity and recognition are investigated by classical equilibrium and non-equilibrium MD simulation. With the aid of seven partially connected case studies on proteins and molecular layers, general conclusions on inter- and intramolecular non-bonded interactions are sought. For each system, customized MD-based workflows were developed and applied. The herein presented case studies encompass 1. the prediction of a small-molecule binding mode to a receptor protein, 2. the quantitative comparison of the protein-protein binding modes of two evolutionary divergent enzymes, 3. the site-resolved conformational analysis of N-glycans, 4. the site-selectivity of asparagine deamidation of two related proteins, 5. the aggregation of signaling lipids around an anchored peptide, 6. the phase transitions of the membrane anchoring components within mixed self-assembled monolayers (SAMs) and 7. the effects of such anchors on vesicles adsorbing to the mixed SAMs.

The problems were investigated with experimental support and theoretical insight from different research labs. Highlights of the computational methodology include the development of a feasible NMR-guided ensemble docking workflow for weak binders, the compilation of a fully automated, multi-scale modeling, simulation, and analysis workflow for mixed SAMs and the benchmarking and application of an embedded torsion-angle clustering approach. In general, it showed that, while the investigated issues were different, the necessary trajectory analysis means were related and or of general applicability. Initially, the most persistent intermolecular or non-neighboring intramolecular interactions were identified. Such analysis was accompanied by a high-resolution (bond-wise) analysis of conformational and in some instances also orientational preferences. A key

insight was that conformational analysis must be distribution-based to identify multimodality and avoid an artificial averaging. Instead, in MD trajectory analysis, quantile probabilities are the superior statistical means. Conformational clustering proved to be necessary to reveal the size of individual populations as well as unexpected statistical dependencies.

The individual case studies yielded valuable understanding and contributions to their respective fields and highlighted the diversity of types and their effects of non-bonded interactions. For example, the binding or bile acids to the receptor protein was mediated mostly by hydrophobic and electrostatic interactions. The binding was weak as reflected by significant dynamics and the accessibility of multiple possible binding modes. Upon acid binding, the C-terminus of the receptor transitions from a protein-bound to a more solvent-exposed conformation. Such a transition might facilitate the multimerization of the receptor proteins which are stabilized by C-terminal interactions. RavD and OTULIN are bacterial and human DUB proteases that bind to identical substrates. In the bacterial RavD equivalent, one of the binding sites substituted electrostatic for weaker hydrophobic interactions, with the result of a reduced binding interface area and stability compared to human. Transient protein-glycan interactions in human erythropoietin protein induce significant, site of glycosylation-specific changes to the conformational spaces of the glycosylation root but not on the glycan itself. Asparagine 373 of a viral coat protein undergoes exceptionally fast post-translational deamidation reaction. This residue is positioned in a specific loop region, which is characterized by a presence of a nearby threonine that forms strong hydrogen bonds with two successive backbone hydrogens. In this loop, the amino acid backbone adopts a rare conformation that enables a short attack distance as well as an increased backbone hydrogen acidity and thus promotes the chemical reaction. Mixed SAMs are used to tether lipid bilayers by inclusion of long acyl anchor-carrying alkanethiols to gold surfaces. Such molecules are engaged in strong hydrophobic intermolecular interactions, which lead to long-living self-aggregation and a highly ordered configuration with a collective surface normal-parallel orientation. This special configuration of the aggregated tethering molecules showed to be advantageous for tethered-bilayer preparations.

Overall, the results show that MD is the prime method of choice to study molecular interactions and their effects on the conformational space. However, recent advancements regarding the availability of more powerful computational resources and the resulting possibility to increase the time covered and the conformational space sampled affords the accessibility of more robust and elaborate trajectory analysis means. Such are suggested and recommended in this work.

ZUSAMMENFASSUNG

Intermolekulare Wechselwirkungen z.B. Wasserstoffbrücken, hydrophobe und elektrostatische Wechselwirkungen bestimmen die Struktur und Dynamik von flexiblen Molekülen und Molekülkomplexen. In isolierten Molekülen sorgen sie u.a. für die Stabilisierung von seltenen, energetisch ungünstigen Konformeren, die intramolekulare chemische Reaktionen oder auch Lösungsmittelreaktionen ermöglichen. Solche Reaktionen führen häufig zu einer Änderung der Oberflächenladung, was weitreichende Folgen für die molekularen Eigenschaften mit sich bringt. Dazu kommt, dass diese Interaktionen die Assoziation von Molekülen zu kurzlebigen Aggregaten und stabilen Komplexen veranlassen oder diese regulieren. Die Basis für die gegenseitige Erkennung von Molekülen liegt in der Komplementarität der oberflächlichen lokalisierten Wechselwirkungspartner Die selektive Erkennung von bestimmten Molekülen oder chemischen Gruppen innerhalb eines Moleküls durch einen Bindungspartner ist eine grundlegende Eigenschaft von zellulärem Leben und technisch-chemischen Systemen.

Experimente messen häufig lediglich die Resultate und zeitlichen Mittelwerte von nichtkovalenten Interaktionen anstatt der Interaktionen selbst. Aufwendigere Methoden sind teuer, eventuell fehleranfällig und meist limitiert auf wenige Atome oder Gruppen. Zurzeit gibt es kein Experiment, was in der Lage wäre, molekulare Wechselwirkungen mit hoher zeitlicher und räumlicher Auflösung darzustellen. Mit dem derzeitigen Aufstieg von Grafikprozessoren und dem damit verbundenem Wachstum von nutzbarer Computerrechenleistung, hat sich die theoretische Methode der Molekulardynamik (MD) zu einem Standardwerkzeug entwickelt. Die Technik wird regelmäßig benutzt, um die zeitliche Änderung von molekularen Strukturen und Wechselwirkungen zu untersuchen. Heute lässt sich aus einem Füllhorn verschiedenster Anwendungsbeispiele von MD schöpfen, wovon einige jedoch die nötige Fürsorge bei der Vorbereitung sowie der Auswahl von Kraftfeld-Parametern und Analysemethoden vermissen lassen. Es ist klar, dass verschiedene Fragestellungen auch verschiedene Techniken erfordern. Dennoch kann man feststellen, dass es zu wenige vergleichbare, allgemeinhin genutzte Ansätze für die Quantifizierung und Darstellung von intermolekularen Wechselwirkungen gibt.

In dieser Arbeit werden sowohl Gleichgewichts- als auch Nicht-Gleichgewichts-MD Simulationen durchgeführt, um eine dynamische Sichtweise von nicht-kovalenten Bindungen und ihren Beiträgen hinsichtlich molekularer Selektivität und Erkennung zu entwickeln. Anhand von sechs teilweise aufeinander aufbauenden Fallstudien zu Proteinen und selbstorganisierten synthetischen Mono- und Doppelschichten sollen allgemeine Erkenntnisse gewonnen werden. Es wurden für jede Studie maßgeschneiderte Abläufe der Präparation, Simulation und Analyse entwickelt und benutzt. Die einzelnen Fallstudien beinhalten 1. Vorhersage der Bindungsmodi eines kleinen Moleküls an seinen Rezeptor, 2. quantitativer Vergleich der Protein-Protein Bindungsstellen zweier analoger Proteine in einem bakteriellen und menschlichem Protein, 3. Analyse des Konformerenraums von protein-gebundenen N-Glykanen, 4. Dynamik-basierte Erklärung für die schnelle Deamidierung eines bestimmten Asparaginrests in zwei verwandten Proteinen, 5. Änderungen von Konformationen und Orientierungen von bestimmten langkettigen Komponenten einer gemischten selbstorganisierenden Monoschicht und 6. Auswirkungen der langkettigen Komponenten auf Vesikel, die an die Monoschicht adsorbieren.

Die Fragestellung und Herangehensweise wurden durch Experimente und theoretische Einblicke von verschiedenen anderen Laboren unterstützt. Besonders bemerkenswerte computergeschützte Methoden waren die Entwicklung eines Ensemble-Docking Protokolls für schwach bindende Moleküle, die Zusammenstellung eines automatischen Modellierungs-, Simulations- und Auswertungsprotokolls für gemischte selbstorganisierte Monoschichten, sowie die Entwicklung und Anwendung eines eingebetteten Gruppierungsalgorithmus für Torsionswinkel. Grundsätzlich

hat sich gezeigt, dass sich die nötigen Analysemethoden gleichen, auch wenn sich die untersuchten Probleme teils deutlich unterschieden. Dabei wurden zunächst langlebige inter- und intramolekulare Kontakte untersucht. Das wurde von einer genauen Analyse von Konformation und Orientierung verschiedener Bindungen begleitet.

Grundsätzlich hat sich gezeigt, dass sich ähnliche Analyse-Ansätze als nützlich erwiesen haben, unabhängig von der untersuchten Fragestellung. Zunächst wurden die wesentlichsten intermolekularen und intramolekularen Wechselwirkungen identifiziert. Diese Untersuchung wurde begleitet von einer hochaufgelösten Analyse der molekularen Konformationen und Orientierungen. Eine wichtige Erkenntnis war, dass geometrische Parameter immer eine Verteilungs-basierte Analyse erfordern, um künstliche Mittelwertbildung bei unerkannten Multimodalitäten zu vermeiden. Stattdessen ist es angebracht, Wahrscheinlichkeitsverteilungen zu benutzen. Außerdem hat sich die Clusteranalyse als nützlich erweisen, um Populationsgrößen zu bestimmen und unerwartete Abhängigkeiten zu identifizieren.

Aus den einzelnen Fallstudien konnten wertvolle Erkenntnisse und wissenschaftliche Beiträge abgeleitet werden. Außerdem wurde das Ausmaß der Unterschiede in Art und Wirkung von nicht-kovalenten Interaktionen deutlich. Zum Beispiel ist die von hydrophoben und elektrostatischen Wechselwirkungen dominerte Bindung von Gallsäuremolekülen an ein virales Rezeptorprotein von einer deutlichen Dynamik beider Moleküle begleitet. Insbesondere wird der C-Terminus des Rezeptorproteins von der bindenden Gallsäure verdrängt und wechselt in eine eher wasserzugängliche andere Konformation. Das könnte einen Einfluss auf die Multimerisierung des Rezeptors haben, welche durch C-terminale Interaktionen stabilisiert wird. Die Protease-Enzyme RavD und OTULIN binden dasselbe Substratprotein. Jedoch nutzt bakterielles RavD dafür eher unspezifische hydrophobe Wechselwirkung anstelle von gerichteten, komplementären elektrostatischen Wechselwirkungen in menschlichem OTULIN, was sich in einer verringerten Grenzfläche und Bindungsstabilität widerspiegelt. Bei dem menschlichen Wachstumsfaktor Erythropoietin wurden kurzlebige Wechselwirkungen zwischen den N-Glykanen und dem Protein identifiziert. Sie induzieren eine Veränderung des Konformerenraums an den Glykosierungswurzeln aber nicht so sehr in den N-Glykanen selbst. Im einem viralen Hüllenprotein gibt es eine spezielle Asparagin-Stelle, die spontan und ausnahmslos schnell die intramolekular chemische Reaktion der Deamidierung eingeht. Das konnte damit erklärt werden, dass sich dieses Asparagin in einem besonderen Schleifenmotiv befindet, welches durch starke Wasserstoffbrückenbindungen zwischen einem zentralen Threonin und zwei Rückgrat-Aminen hervorgerufen wird. Dieses Muster führt zu einer verzerrten Rückgrat-Konformation, die mit einer geeigneten Angriffsgeometrie sowie einer erhöhten Azidität des Amin-Wasserstoffatoms einhergeht. Mehrkomponentige selbstorganisierenden Monoschichten werden genutzt, um darauf Lipidmembranen zu fixieren. Dabei werden Alkanthiole beigesetzt, die mit weiteren langen Alkylketten funktionalisiert sind, um in die aufgebrachte Lipidmembran einzudringen. Derartige Moleküle zeigen aufgrund ihrer starken hydrophoben Interaktionen eine stabile Aggregation, was zu einer deutlichen Phasenänderung von einem ungeordneten zu einem geordneten Zustand führt. Diese Änderung hat sich als vorteilhaft für die Herstellung von fixierten Lipidmembranen erwiesen.

Zusammengefasst zeigen die Ergebnisse, dass Molekulardynamik Simulationen die Methode der Wahl zur Untersuchung der zeitlichen Entwicklung molekularer Wechselwirkungen bei konformationellen Änderungen ist. Jedoch haben jüngste Fortschritte in der Verfügbarkeit von zunehmender Hochleistungs-Rechenleistung dazu geführt, dass die zeitlichen Computersimulationstrajektorien deutlich an Simulationslänge gewonnen haben. Das wiederum erfordert robustere und geschicktere Analyse-Techniken, so wie sie in dieser Arbeit aufgezeigt und empfohlen werden.

CONTENTS

1 INTRODUCTION

Chemistry is often imaged as the science of chemical reactions in the sense of breakage and formation of covalent chemical bonds. In fact, per definition, chemistry studies the properties and the behavior of matter [1]. Generally speaking, any substance which has a mass and takes a volume can be considered as matter [2]. It can be of organic or inorganic origin, it can be present in different phases states such as solid, liquid, or gaseous, yet it will always be a composed of atoms which themselves are made of subatomic particles. The properties and behavior of matter is thus ultimately a consequence of the type of interactions – also called bonds - these atoms undergo with each other. The range of possible mutual interactions is restricted by the types of the involved atoms.

The general types of chemical bonds are I. covalent bonds, II. ionic bonds, III. metallic bonds and IV. hydrogen bonds. Covalent bonds form between atoms through sharing of electron pairs, which allows them to attain the equivalent of a full valence shell and thus a stable electronic configuration. Two or more atoms, which are connected via covalent bonds, are called molecules. Ionic bonds are based on electrostatic interactions and form between oppositely charged particles (atom or molecule ions) or between atoms with highly different electronegativity numbers. Assemblies of such particles are called salts. Metal bonds appear between atoms of the metal element groups and arise from electrostatic forces between ionized metal ions and conductible electrons. Finally, hydrogen bonds belong to the class of weak electrostatic interactions, which occur between a hydrogen covalently bound to a more electronegative atom and another nearby electronegative atom carrying a free pair of electrons (lone pair) [3].

In most of the substances surrounding us, multiple types of bonding come together to form most complex networks of interactions which explain their physical-chemical properties and their dependence on the physical environment (temperature, pressure). Water, for example, is basically only a 1:2 mixture of oxygen and hydrogen atoms. Nevertheless, it engages in various molecular interactions. Covalent binding of every oxygen to two hydrogen atoms defines the molecule's shape and explains its high molecular polarity which allows it to undergo intermolecular hydrogen bonds with each other. The strong polarity and hydrogen bonding propensity of water are the main reason for its exceptional properties and its ability to support life on planet earth [4].

1.1 Structure and dynamics

The undisputed importance of non-covalent interactions for biologic life [5, 6] but also technical advances [7] has led to a great wealth of structural data especially in the realm of proteins in which the global protein databank (RCSB PDB) might reach 200.000 entries by the end of the year 2022.

Even if the crystal- and the recently emerging cryo-EM - structures are invaluable for research and education, they do not account for naturally occurring molecular motions, also known as dynamics. Such motions are restraint within a crystal lattice or at low temperature yet are crucial to explain highly relevant functional phenomena such as catalytic activity [8, 9], membrane transport [10, 11] or allosteric signaling [12, 13], to only name a few.

Protein dynamics can be described as the harmonic or anharmonic deviation of atomic positions with respect to a reference state [14]. The reference state might be the energetic optimum or lowest-energy state, also referred to as *native state*, which is best represented by the crystal structure. Another reference can be a theoretical average state which, however, is not necessarily physically existent. However, the best description of the conformational landscape can only be achieved by a set of molecular conformations, being called *conformational ensemble*. Unimolecular motions take place on various time and length scales (**Figure 1.1**), ranging from ps to h and from pm to μm [15]. Motions involving multiple large macromolecular complexes appear on even larger scales. With reference to a protein, one can distinguish between local dynamics, regional dynamics, and global dynamics. Local dynamics encompass bond vibrations or sidechain rotations. Regional motions concern intra-domain or concerted and interdependent multi-residue dynamics driven by hydrogen bonds and ionic bonds. Global motions affect the whole protein and be of various extent, for example unfolding and refolding or large-scale "breathing" motions (normal modes).

1.2 Principles of molecular recognition

In a confined chemical environment, e.g. a test tube, a living cell, a solvent covered surface, a reaction vessel, etc., a finite number of different molecules is present. Based on their Brownian motion (diffusion), these molecules will inevitably collide with each other and mutually exert forces, which are either repulsive or attractive and vary in their magnitude [16]. It is easy to imagine, that stronger attractive forces will lead to the formation of more stable complexes, whereas weakly attractive complexes will quickly dissociate [17]. Based on their shape and exposed interaction areas as acceptors and donators, different pairs of molecules have a different interaction strength, which is reflected by their lifetime and often called *binding affinity*. Hence, a certain molecule will discriminate between the other surrounding molecules by means of their pairwise binding affinity. The binding affinity, and in consequence the lifetime of the molecular complex, can vary by many orders of magnitude. This concept is called molecular selectivity [18]. Here, the term *molecular recognition* will refer to the description of the many pairwise, non-covalent interactions which form the underlying physical foundation for molecular selectivity.

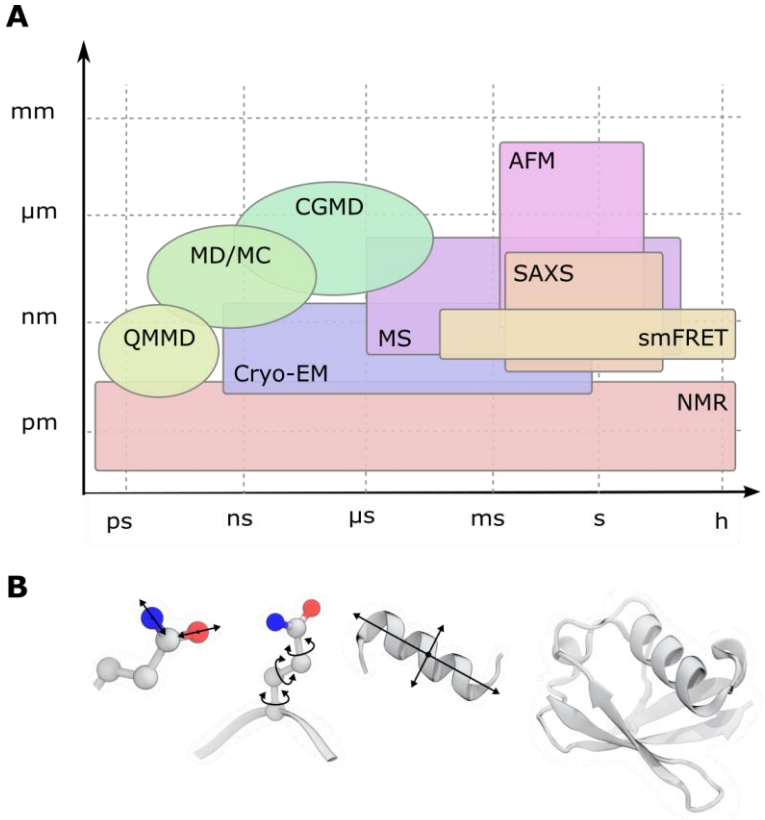

Figure 1.1: Molecular motions occur on different time and length scales. A) Overview of the accessible time and lengths scales of various computational (ellipses) and experimental (rectangles) methods. QMMD: Quantum mechanics molecules dynamics, MD: molecular dynamics, MC: Monte Carlo, CGMD: Coarse-grained molecular dynamics, NMR: nuclear magnetic resonance, EM: electron microscopy, MS: mass spectrometry, smFRET: single molecule Förster resonance energy transfer, SAXS: small-angle X-ray scattering, AFM: atomic force microscopy. B) Examples of molecular motions: bond vibration, sidechain rotation, domain motion and folding. The figure is adapted from [19].

Molecular recognition plays central roles in biological systems, in which it enables important cellular molecules such as enzymes and substrates, antigens and antibodies, sugars and lectins, or RNAs and ribosomes to robustly find each other in the crowded environment of the cell. [20] In the framework of the research presented in this thesis, we will extent the classical concept of bi-molecular recognition to a more generalized connotation which also allows interpretation of unimolecular (intramolecular) selectivity and multi-molecular selectivity in the context of molecular recognition.

1.2.1 Recognition of small molecules by protein receptors

Proteins are linear heteropolymers consisting of a genetically pre-determined sequence of amino acid that are linked by peptide bonds. In aqueous solvent and based on specific intramolecular interactions such as hydrogen bonds, ionic bonds, and hydrophobic interactions, that is, the desolvation of non-polar chemical groups through aggregation, the peptides locally adopt so called *secondary structure* elements such as alpha helices or beta sheets. Such elements further aggregate and form the tertiary structure of a protein domain. Aggregation of domains is called quaternary protein

structure. Each of the elements are connected by shorter or longer flexible segments called loops and random coils (**Figure 1.2 A-D)** [21].

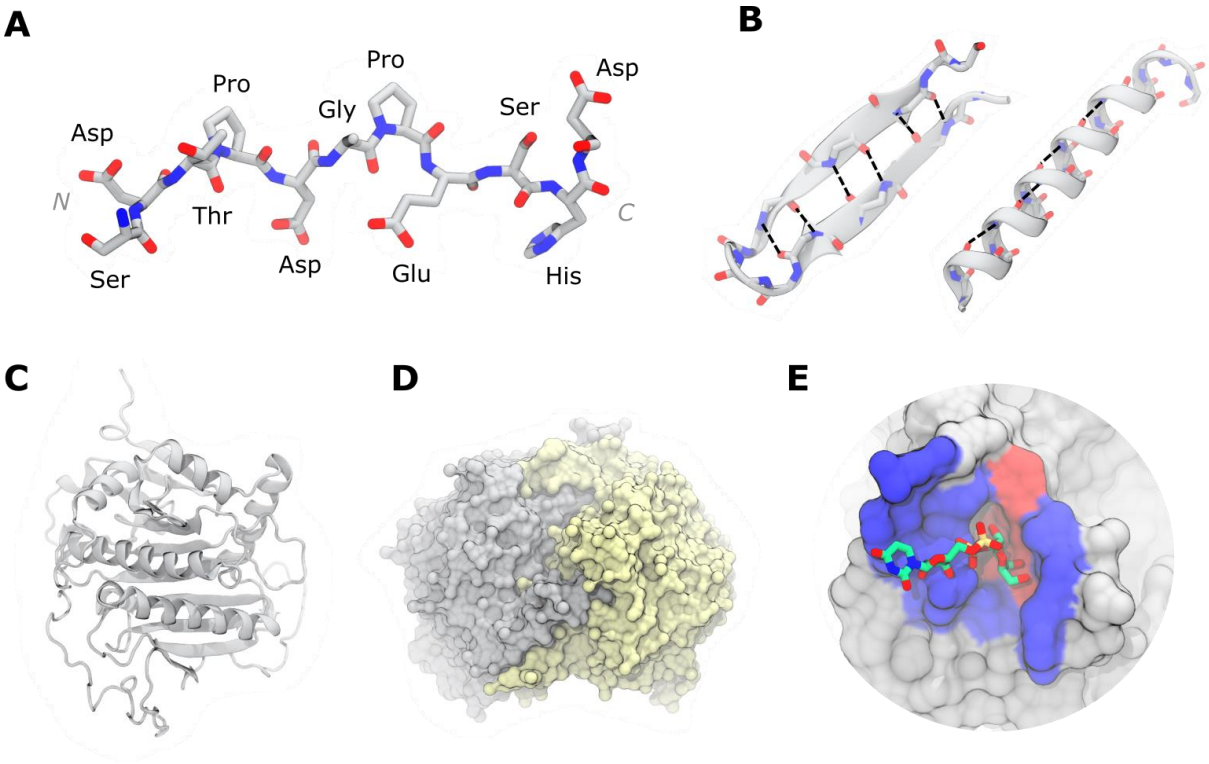

Figure 1.2: Proteins exhibit several levels of organization. A) Linear sequence of amino acids linked by peptide bonds (primary structure) B) Secondary structure elements of beta sheet and alpha helix stabilized by intramolecular hydrogen bonds (black dashed lines, only visible ones are drawn) C) Ordered and disordered regions fold into the tertiary structure. D) Multiple subunits consisting of individual peptide chains pack tightly to form homo- and hetero-oligomers (quaternary structure) E) A ligand binds to the recognition site (blue) so that the reactive chemical groups are in close contact with the catalytic residues (red). All images are generated from the PDB structure 1GUP.

Proteins have evolutionary developed to fulfil various delicate cellular tasks. One of the earliest recognized and best characterized protein function is to serve as a biocatalyst to reduce the energy barrier of a chemical reaction. Such "molecular machines" are called enzymes and they form the basis for cellular metabolism and signaling. Enzymes exhibit at least two distinct regions which beget their function: I. the substrate recognition site and II. the active site with the catalytic center (**Figure 1.2**). The substrate recognition site ensures high selectivity for a certain substrates and orients and remodels the substrate into a position and conformation that enables close proximity of the reactive groups of substrate and catalytically active residues or complexed metal ions. The active site ultimately undergoes a catalytic cycle once productive substrate binding is achieved [22].

As early as 1894, Emil Fischer observed enzyme selectivity and proposed a selectivity model based on complementarity of geometric shapes [23], which is often referred to as "lock-and-key" model. While the model is able to explain selectivity, it has a major caveat: enzymes achieve their function of lowering the activation energy barrier by e.g. stabilization of the transition state via favorable molecular interactions. This means that the binding is the strongest in the transition state and not

before or after the chemical reaction. In this way the product will have a lower affinity to the enzyme than the substrate and thus quickly dissociate from the enzyme. Such plasticity in the protein-ligand interaction strength cannot be explained by the static lock-and-key model, which is why it was largely replaced by the improved "induced fit model" by Daniel Koshland in 1958 [24]. This model allows conformational changes of the protein and the ligand during the different stages of binding, catalysis, and unbinding (dissociation). In fact, the model suggests that certain conformational rearrangements within the protein's active site appear only upon approaching of the ligand. However, as discussed in a recent article [25], advanced NMR and single-molecule spectroscopy experiments indicate that the conformational landscape of a protein is predetermined, and ligand-bound conformations can be adopted also without ligand. This observation suggests that a protein steadily transitions between its many possible conformations until the presence of a ligand stabilizes a certain conformation. In other words, the ligand enables the protein to stably adopt a conformation which was much more unlikely in absence of the ligand. This mechanism has been coined "conformational selection" [26].

A family of enzymes with tremendous current interest and exquisite selectivity are proteases, which selectivity recognize and cleave peptide or isopeptide bonds of distinct target peptides. Proteases play important roles in metabolism, cell signaling and protein homeostasis [27]. Additionally, they are frequently employed by pathogens to enable cell entry and immune escape [28]. One of such enzymes is the papain-like protease (PLPro) of the SARS-CoV-2 coronavirus, the cause for the coronavirus pandemic of 2019 and ongoing. Thus, it will be used here as an example to introduce protein-ligand interactions.

The crystal structure of PLPro in complex with a newly developed inhibitor (GRL0617) was recently solved (**Figure 1.3**) [29]. An inhibitor is a molecule which exhibits a high affinity towards the enzyme, thus it competes with substrate for binding sites and consequently reduces the activity of the enzyme. It shows that the inhibitor has three-dimensional shape which is highly complementary to the substrate binding grove of the enzyme. Additionally, it undergoes a range of distinct interactions. The bi-aromatic naphthyl-group is sandwiched between two proline residues (247 and 248) and tyrosine 268 via pi-pi interactions between the aromatic moieties. Hydrogen bonds are formed between aspartate 164 and the central amide and between tyrosine 268 and the aniline amino group. The two methyl groups undergo hydrophobic interactions with threonine 301, and leucine 162, respectively. Finally, the carbonyl oxygen on the inhibitor engages in a hydrogen bond with the backbone amide of glutamine 269. Comparison of the crystal structures of inhibitor-bound and free PLPro reveals the different conformations of the so called BL2 loop.

Upon binding, this loop, which comprises tyrosine 268 and glutamine 269, folds towards the protein core and covers a part of the ligand thus stabilizing the protein-ligand complex.

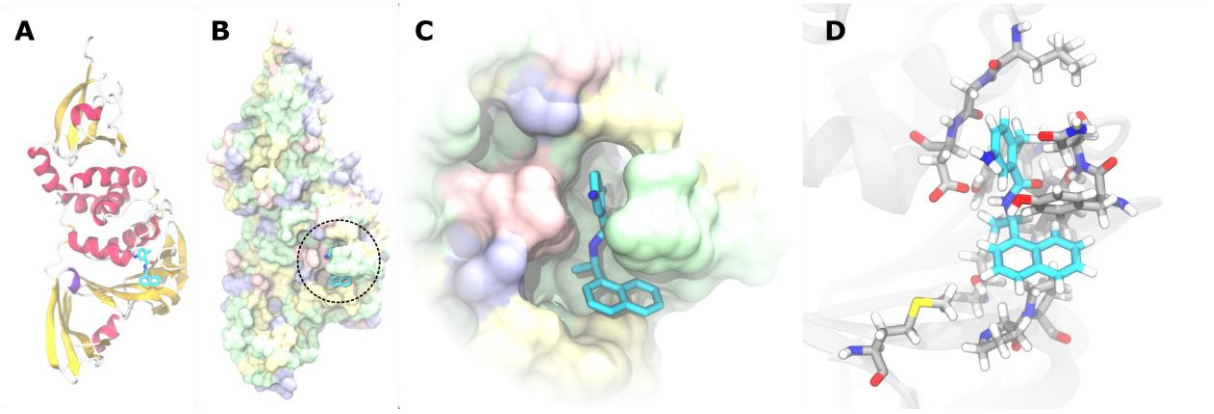

Figure 1.3: Different representations of the SARS-Cov-2 papain like protease bound to the GRL0617 inhibitor. A) Cartoon representation with alpha helices in red, 1-3 helices in purple, beta sheets in yellow and the inhibitor in blue. B) Surface representation with color coding according to residue type: green: polar, blue: basic, red: acidic, yellow: hydrophobic. C) Close-up view on the binding pocket. D) Stick model of the inhibitor in the binding pocket with interacting protein residues.

1.2.2 Specificity of protein-protein binding

In the cell, specific tasks such as signal transmission or metabolism, are not carried out by single, free-floating enzymes but rather by large, dynamic complexes of different, interacting molecular species [30]. One striking example for such a molecular machinery are the so called Cullin-RING E3 ligases (CRLs, **Figure 1.4**), which fulfil the essential task to specifically recognize and bind excessive target proteins and tag them with ubiquitin moieties [31]. Ubiquitin is versatile, small, and highly abundant globular protein in the eukaryotic cell. The post-translational modification of ubiquitin conjugation leads to the degradation of the target protein via the proteasome complex [32]. Cullin-RING ligases are heteromeric multi-protein complexes generally consisting of Cullin protein scaffold, which on one end tightly associates with a RING box protein, which in return binds to ubiquitin-carrying and transmitting enzymes called E2 ligases. On the other end, the Cullin binds to a certain pair of an adaptor and substrate receptor proteins. Depending on the emerging target that must be eliminated, CRLs rapidly adopt their substrate receptor via a complex regulatory cycle [33]. The function, regulation and possibilities for therapeutic intervention and utilization of CRLs has fascinated investigators for a quarter of a century [34]. However, only with the emergence of the recent biophysical technique of high-resolution cryo-EM combined with molecular-dynamics flexible fitting, the mechanism based on protein-protein interaction induced conformational transitions was finally elucidated [35]. Briefly, the conjugation of Nedd8 to the winged helix domain of Cullin leads to conformational rearrangement of the RBX-bound,

ubiquitin-carrying E2 ligase. Only this conformational transition allows the spatial vicinity of the ubiquitin and target protein, which is bound to the substrate receptor on the other site of the Cullin.

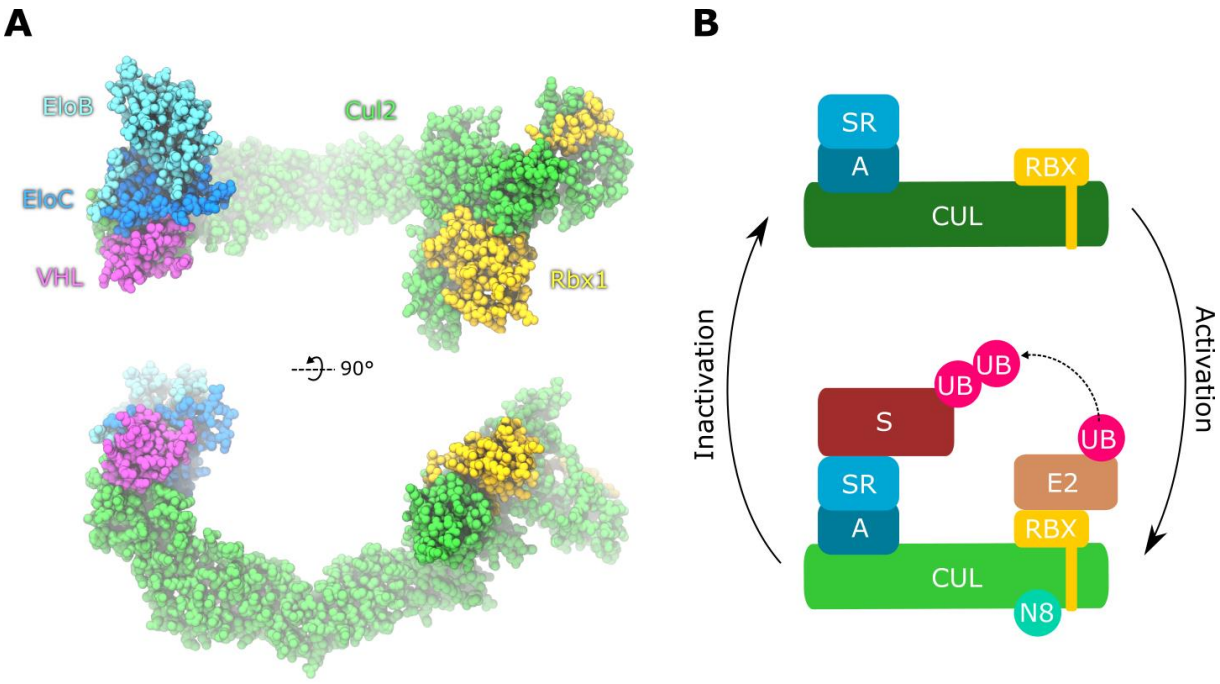

Figure 1.4: Multi-protein recognition within CRLs. A) Assembled CRL2^{VHL} in van-der-Waals representation (6TTU). B) Schematic representation of CRL activity regulation via the small modifier Nedd8. CUL: Cullin. SR: Substrate receptor. A: Adaptor. RBX: Really interesting new gene (RING) box: S: Substrate. Ub: Ubiquitin. E2: E2 ubiquitin ligase. N8: Nedd8.

The assembly and activation of CRLs is only one example on how protein-protein interactions control a majority of cellular tasks. It becomes apparent, that such interactions need to be highly specific and that a disruption thereof, for example by means of mutations, would dramatically affect cellular function. The interactions of small molecules with proteins and the interaction between two or more proteins generally follow the same physical principles and are driven by shape complementarity, as well as hydrogen bonds, electrostatic and hydrophobic interactions [36]. However, there are also striking differences. Whereas small molecules usually bind to distinct, spatially small yet deep pockets, protein-protein interfaces occupy large and shallow areas [37]. Small molecule binding is rather ensured by buried hydrogen bonds and hydrophobic interactions and to smaller extent by ionic bounds. Protein-protein interactions often rely on large hydrophobic patches surrounded or discontinued by complementarily charged areas. This is also the reason why the identification and development of small molecule drugs as protein-protein interaction inhibitors is challenging [38].

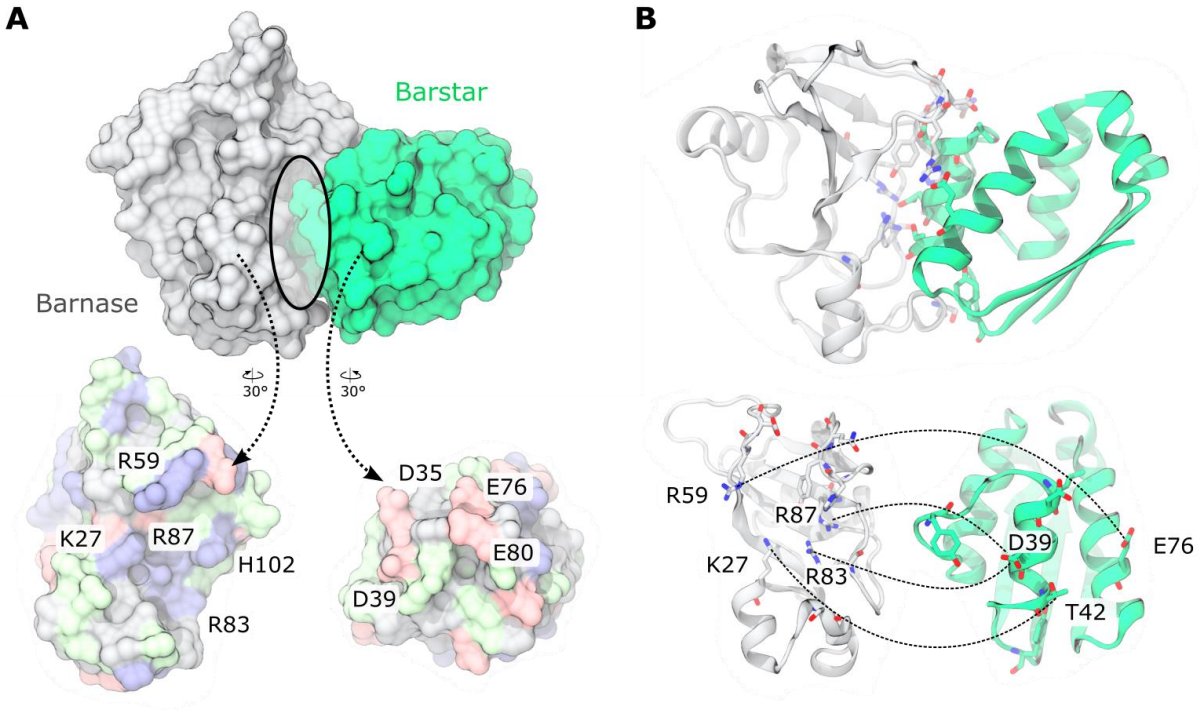

Figure 1.5: The Barnase-Barstar complex. A) Barnase (gray) and Barstar (green) in surface representation. The top image shows the complex, the bottom image the two interaction sites. B) Barnase-Barstar ribbon model and selected molecular interactions.

One example of an extraordinarily strong protein-protein interaction is the complex of the bacterial ribonuclease Barnase and its natural inhibitor Barstar (**Figure 1.5**) [39]. In the cell, Barnase would lethally dissect the bacterial plasmid, was it not always inhibited by its counterpart Barstar. The article by Guillet et al. elucidates the structural underpinnings for the highly stable protein-protein binding and can be considered as a prototype protocol for the quantitative description of such interactions. The authors identify 15 residues of Barnase being in close contact with one or multiple residues of Barstar, 14 distinct hydrogen bonds and a drastic decrease in solvent accessibility of 6 patches on Barnase and 5 patches on Barstar. Similar to the small molecule inhibitor GRL0617, Barstar packs a tremendous number of interactions into a comparably small volume, which renders it an exquisite binder.

1.2.3 Site-selectivity of intramolecular interactions

The conformational flexibility of a molecule is mostly pre-determined by the number of rotatable bonds. The likeliness of different rotational states (rotamers) of a given molecule is then affected by steric, non-bonded interactions. In the case of butane, for example, the interaction energy of the *syn* (*cis*) conformation, which corresponds to a C-C-C-C torsion angle of 0°, is the highest due to steric repulsions of the two terminal methyl groups. As biological systems thrive to low energy states, this conformation would be the most unlikely. The opposite conformation, called *anti* or *trans*, with a torsion angle of 180° is sterically less hindered and thus most likely.

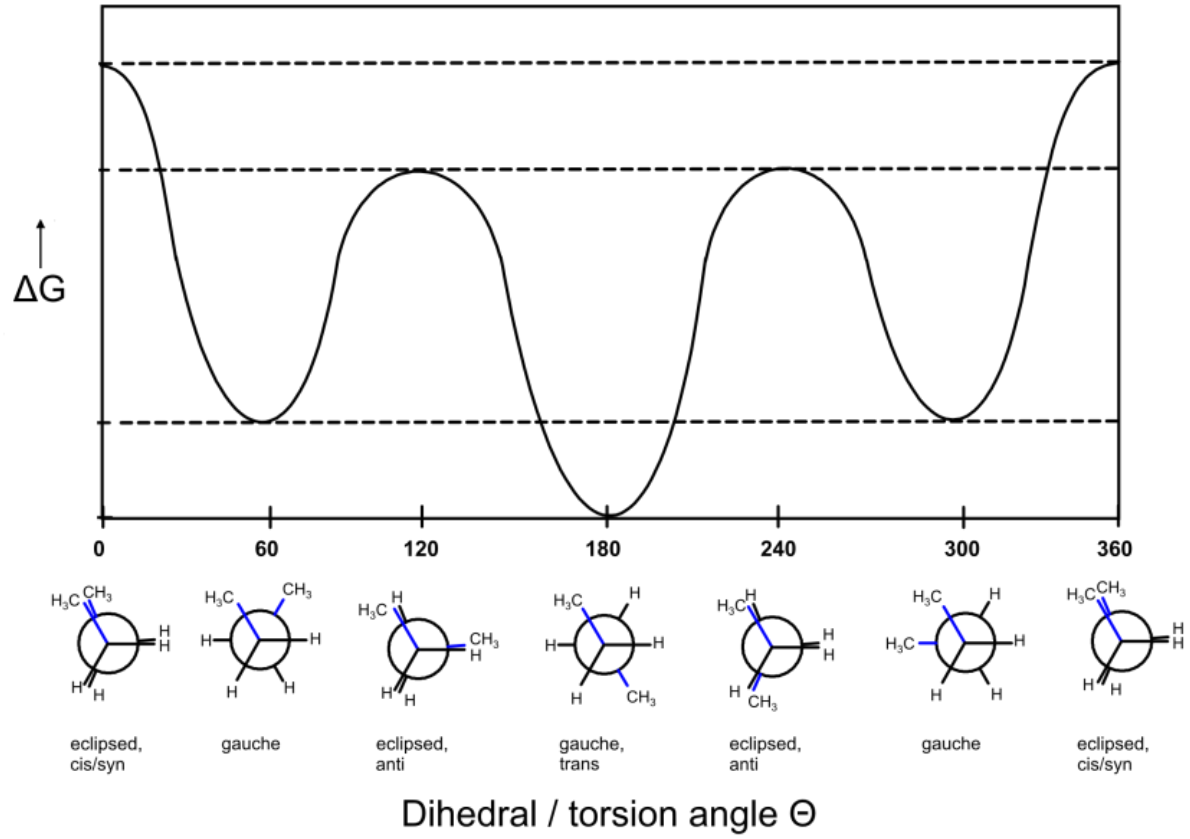

Figure 1.6: Energy diagram of the C-C-C-C dihedral rotation of butane. (Image by R. Mattern, 2020, CC BY 3.0, https://commons.wikimedia.org/wiki/File:Dihedral_angles_of_Butane.svg / cropped and removed axis scale)

Even though the terminal methyl groups contribute a substantial proportion of the steric repulsion energy term, the hydrogen atoms do also affect the conformational space. Eclipsed conformations, where two or more atoms are in a 180° torsion angle conformation are disfavored over gauche conformations where the atoms are rotated by 60° (**Figure 1.6**).

Similar consideration can be applied for the conformational space of e.g., protein sidechains or posttranslational modifications. Here, it frequently appears that the native conformational space is affected by intramolecular interactions of the sidechain atoms with the surrounding protein environment. A free, basic arginine sidechain for example might be fully exposed and highly solvent accessible and thus serve as an primer for protein-protein interactions [39]. Another arginine, with same chemical structure of course, could be in proximity of an acidic glutamate. The two amino acids would leave their native, extended conformation in favor of mutual electrostatic interactions and occupy an otherwise unlikely bond rotamers. Not only is the molecular conformation site-dependent, but also the protonation state of the amino acid histidine, for example. Depending on surrounding the chemical environment with hydrogen bonding acceptors and donators, the histidine residue can either be neutral or positively charged [40].

Posttranslational modifications (PTM) of proteins are often mediated by enzymes, which specifically recognize certain consensus sequence motifs such as H-B-B-B-X-S (H: hydrophobic, B: basic, X: arbitrary, S: serine phosphorylation site) by the serine phosphatase AMPK [41], Asn-X-Serine/Threonine for N-glycosylation [42] or the C-terminal sequences C-X-C and X-X-C-C on Rab proteins for the conjugation of a geranylgeranyl-group [43]. Additionally, there are PTMs, which do not require an enzymatic reaction but appear spontaneously if specific local conditions are fulfilled. In this case, reactions may be catalyzed by solvent molecules. One of such PTM is the deamidation reaction, which is thoroughly discussed in chapter 3. Another one is lysine carboxylation, which is estimated to concern about 1% of the larger proteins [44]. At basic pH and in presence of a CO_2-containing solvent [45], the lysine residue can undergo carboxylation, in which a carboxyl group is added to the sidechain amino group. In consequence, the charge changes from +1 to -1 with possibly dramatic effects on protein structure and function. Computational analysis of lysine carboxylation sites in the protein data bank identified that reactive lysine residues are rather buried and not solvent accessible. Additionally, acidic residues (Asp, Glu) as well as metal ions were frequently found within a 0.5 nm radius around the carboxylation site [44].

1.2.4 Preferred interactions in two-dimensional molecular assemblies

Amphiphilic molecules, that is, molecules with a polar and a non-polar part, are prone to interact with surfaces, interfaces or with each other and are thus considered surface-active. In a liquid phase, amphiphilic molecules such as phospholipids or detergents form micelles, vesicles and bilayers (**Figure 1.7 A**) [46]. At the interface between solvent and gas phase, they may assemble to give monolayers. This effect is amplified when e.g., alkanethiols interact with a gold surface. The interaction between the thiol groups and gold atoms is partly physical and partly chemical and remarkably stable (**Figure 1.7 B**) [40]. Thus, given enough time to assemble, alkanethiols form stable monolayers on top of gold surfaces, called self-assembled monolayers (SAMs) [47]. In the context of this thesis, lipid bilayer membranes and self-assembled monolayers are summarized under the term *molecular layers*.

Bilayers and monolayers have a variety of characteristics in common. Both consist of one or multiple species of amphiphilic molecules, which carry long aliphatic moieties and polar terminal groups. The acyl chains arrange in a parallel fashion to shield a large fraction of the layer from the solvent. Additionally, the terminal groups form electrostatic interactions and hydrogen bonds with the aqueous solvent and each other. In case of a monolayer, the hydrophobic portion is pointing towards the interface surface, whereas a bilayer consists of two mirrored monolayers on top of each other. In the context of a bulky phase within a finite observation volume, molecular layers have a large lateral but only a small normal direction extent and can thus be considered two-

dimensional. In contrast to SAMs, lipid bilayers are not confined to a surface but are liquid (or liquid-crystalline) and exhibit significant dynamics [48]. A bilayer can undergo undulations, wave-like motions on different time and length scales. Additionally, the lipid molecules exhibit lateral, rotational and inter-layer diffusion [49]. The matrix, i.e. the alkanethiol region of SAMs, is rather rigid and their dynamics are limited to transient defects in the densely ordered packing, which is dominated by all-trans conformations. Such a packing is also possible in bilayers and termed liquid-ordered phase which exists in contrast to the liquid-disordered phase [50].

In the case of mixed, multicomponent bilayers lateral diffusion can lead to formation of smaller or larger, transient or stable aggregates of certain lipid molecules. Additionally, certain molecules or aggregates might have preference for the liquid ordered or liquid disordered phase of the bilayer [51]. Furthermore, there are indications that such an ordering can be translated through the bilayer from one leaflet to the other [52]. Mixed SAMs are synthetic and designed to fulfil distinct functions, e.g. to support and tether a model lipid bilayer for research or analytical reasons. In this case, one component of the SAM consists of an alkanethiol portion, a tethering portion of hydrophilic polyethylene glycol (PEG) and an alkyl membrane anchoring portion to penetrate into the bilayer [53]. During the preparation of the mixed SAM, these moieties undergo interactions in the solvent phase or during the adsorption process, which may pre-determine the lateral distributions of the SAM components. Surface-exposed interactions of the PEG and alkyl portions of the assembled mixed monolayer further shape the surface properties of the monolayer [54, 55].

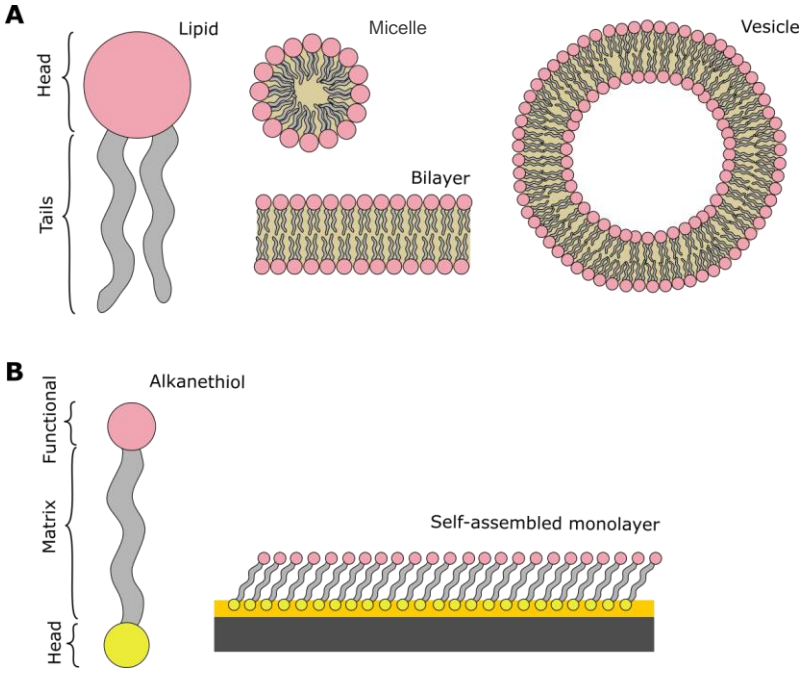

Figure 1.7: Architecture of layered assemblies. A) Schematic representations of a phospholipid molecule and assemblies. B) Schemes of an alkanethiol molecule and a self-assembled monolayer. Red beads: hydrophilic groups. Gray: hydrophobic acyl chains. Yellow beads: Thiol group. Gray and yellow rectangles: gold coated substrate.

1.3 Modeling molecular interactions

Long before the rise of computers in the 1960s, scientists had acquired knowledge of three-dimensional structures of molecules and physical molecular models were built form paper, wood, metal, glass and plastics (**Figure 1.8**). The first physical models of molecules date back to 1860 when August von Hofmann built a ball-and-stick model for methane, which was planar and the hydrogens were larger than the central carbon. Later, when the concepts of stereochemistry emerged, van't Hoff modeled the first three-dimensional, tetrahedral molecules. Arguably, one of the most iconic physical molecular model is the 1950s DNA double-helix model by Watson and Crick.

As soon as computers became more and more widely available, the physical model gave way for the virtual molecular model, although physical toy models still have their place in chemical education and in the heart of many chemists. Interestingly, when computers allowed visualization of complex molecules in the late 1960s, first attempts of a mathematical description of molecules had already been made. In fact, most of the still used mathematical models have their origin in two communications of Terrell Hill from 1946 and 1948, where he suggests to calculate the energy of small organic molecules in dependence of their conformation using a sum of a few simple terms [56, 57]. Nowadays, the terms structural modeling, molecular mechanics modeling and molecular dynamics are summed up under the umbrella of molecular modeling and are often used in combination. Here, for the sake of a detailed, discriminating methodological introduction, the three concepts are explained individually, even though they share many physical fundamentals with each other.

1.3.1 Structural modeling

The generation and visualization of the three-dimensional structure (stereochemistry) based on a two-dimensional chemical formula can be considered structural modeling in the broadest sense. For small molecules, such considerations can be done with only pen and paper and resulted in common projection methods such as the Haworth projection, the Natta projection, or the Newman projection, to name only a few. For the visualization of large, complex macromolecules, however, 2D projections are not sufficient and 3D models must be generated. When faced with multiple rotatable bonds, an *ab initio* structure prediction becomes highly ambiguous. That is why investigators have always relied on restraints as posed by either one or several experimental techniques such as X-ray diffraction, NMR spectroscopy, electron microcopy, IR spectroscopy, single molecule FRET, etc.

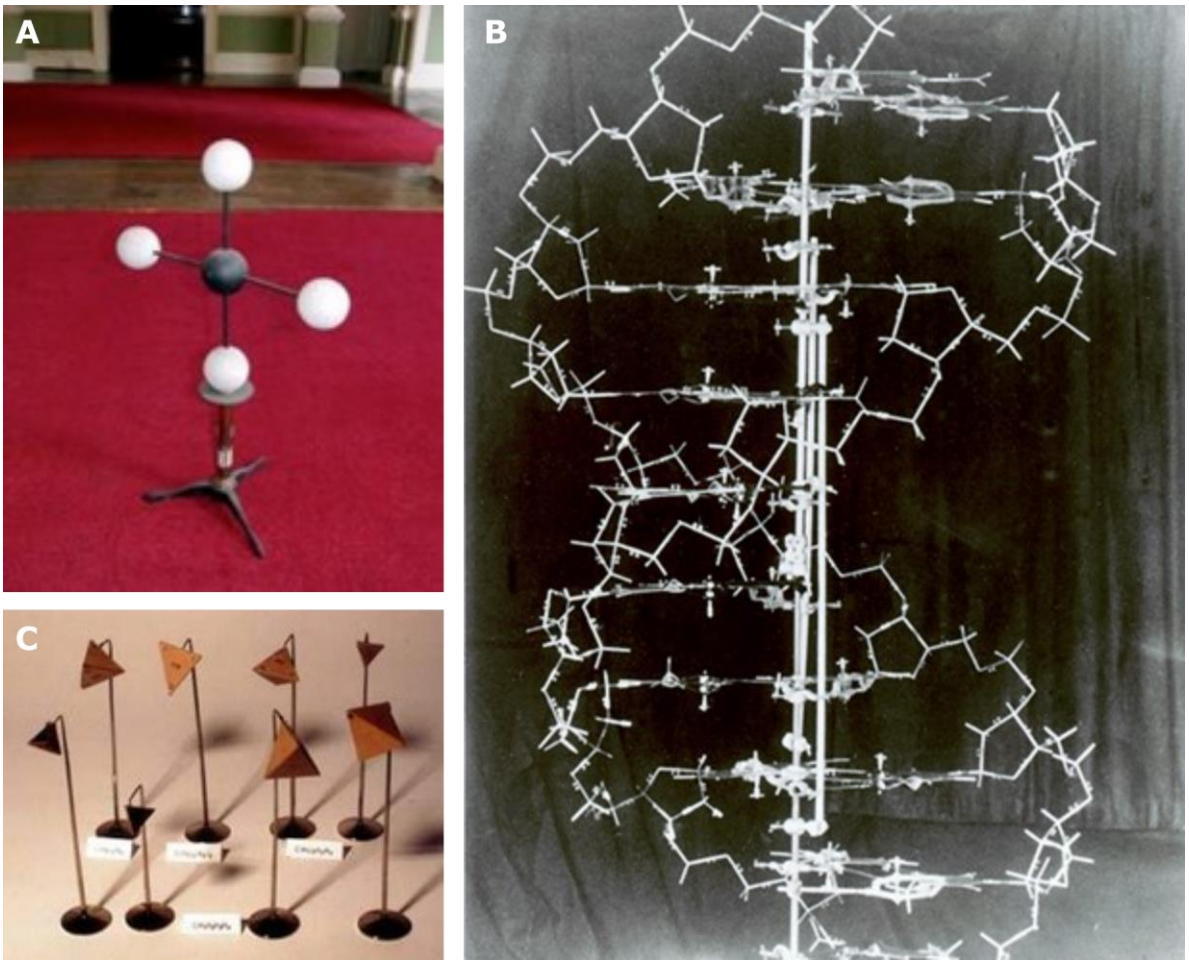

Figure 1.8: Photographs of selected physical molecular models. A) Molecular model of methane, created by August Wilhelm von Hofmann (ca. 1860). The square planar structure depicted is now known to be incorrect. Photograph by Henry Rzepa under GNU license (https://commons.wikimedia.org/wiki/File:Molecular_Model_of_Methane_Hofmann.jpg) B) The six feet tall metal DNA model made by Watson and Crick in 1953. Courtesy of Cold Spring Harbor Archives (https://dnalc.cshl.edu/view/16430-Gallery-19-DNA-model-1953.html) C) Van't Hoff disseminated his stereochemical ideas to leading chemists of the day by sending them 3-D paper models of tetrahedral molecules, like these now housed in the Leiden Museum, photograph By O. Bertrand Ramsay (https://www.sciencehistory.org/historical-profile/jacobus-henricus-vant-hoff)

Of these, the value of X-ray crystallography for the development of the fields of structural chemistry and biology cannot be understated. It was pioneered in 1912 by Max von Laue and awarded with the Nobel Prize in physics in 1914. In has ever since delivered fascinating insights into the molecular world and plenty of further Nobel Prizes were awarded to studies involving X-ray crystallography [58]. Very briefly, the technique is based on fact that the nuclei in molecular crystals are scattering X-ray radiation and the observation of diffraction patterns allow assertions about the molecular distances and angles in the crystal. Such information allows the construction of a 3D electron density map, into which the molecule of interest can be fitted. In the area of organic and biological molecules, the British chemist and Nobel Prize laureate Dorothy Hodgkin [59, 60] should not remain unnamed. She solved the structures of cholesterol, penicillin and vitamin B12 as well as that of insulin on which she worked for over 30 years. Theoretically, the size of molecules for X-ray diffraction is not limited. Instead, purification and crystallization of large

complexes of proteins or nucleic acids is a major bottleneck. Additionally, highly flexible molecules or regions within a protein can often not be resolved with X-ray crystallography.

Another, currently still emerging technique, is cryogenic electron microscopy (cryo-EM), which overcomes the size limitation of X-ray crystallography and was awarded with the 2017 Nobel Prize in chemistry to Jacques Dubochet, Joachim Frank and Richard Henderson. Here, biological samples are dehydrated, shock-cooled below -150°C and subjected to electron microcopy. Electron micrographs from many different angle of macromolecular complexes in various orientations allow the computational reconstruction of an electron density map similar to X-ray crystallography [61, 62].

The second major limitation of X-ray crystallography is conformational flexibility. Here, NMR spectroscopy has been shown to be a valuable tool investigate conformational ensembles of peptides and small proteins in solution. NMR is a spectroscopic technique, which observes the behavior of the local magnetic field around atom nuclei. Therefore, the magnetic nuclear spins are polarized via an external magnetic field and then perturbed by an oscillating magnetic field. The electromagnetic waves emitted by the perturbed atoms is detected. Of note, only certain isotopes such as ^{1}H, ^{13}C or ^{15}N with a nuclear spin are accessible for NMR and proteins must be labeled accordingly. The signal measured by NMR gives the chemical shift δ in units of ppm. It is affected by a range of factors such as electron density and electronegativity of neighboring groups, among others. The chemical shift is specific for certain chemical groups. Thus, NMR spectroscopy is routinely used for the identification of molecular species in a sample. NMR-based structure determination affords the recording of multiple multidimensional spectra, which can be used to calculate distance and angle restraints within a protein. These restraints aid the computational generation of a structural ensemble [63].

All of the above-mentioned, expensive and challenging experimental approaches yield valuable insight, which is however limited to the one investigated molecular system. Using computational modeling approaches, it is possible to extent the results of one experiment to a variety of homologous systems. For example, mutations can be introduced to the protein or chemical groups of a bound ligand can be altered. There is also the possibility to model the sequence of a structurally unknown protein into the structure of one with sequence similarity, which is called homology modeling. An application for homology modeling is the prediction of the binding mode of small molecules or ligands to a structurally unresolved receptor (only when in case the structure of a homologous protein complex is known).

The most frequently recurring problem in the field of structural modeling is the quantitative evaluation of the energy between different conformations i.e., the estimation of their macroscopic

probabilities. Such consideration ultimately leads to the calculation of energy differences between different conformational states of a molecule. This calculation can be undertaken on the electron level using quantum theory, i.e. the solution of the Schrödinger equation or one of its approximations such as the Hartree-Fock method or density functional theory [64]. Yet, given that the number of possible conformations of a linear molecule with N atoms and at least two rotamers per bond increases approximately by 2^N, the expense of using quantum level of theory would quickly exceed most university computing clusters. Instead, investigators have developed a simple approximation method, which treats atoms as spheres and bonds as springs: molecular mechanics.

1.3.2 Molecular mechanics

Molecular mechanics is a concept or mathematical framework, which applies equations from classical mechanics to molecular systems with the aim to quantitatively predict the energy differences between conformational states. With molecular mechanics, the single point energy of a molecular system can be calculated as a function of only the atomic positions. Therefore, the total potential energy of the system is represented as a sum of many terms which can be separated into bonded and non-bonded terms. The bonded terms include the deformation energies induced by bond stretching, angle bending and torsional rotation. The potential terms for bond and angle deformations are mathematically modelled as harmonic potentials with a certain equilibrium value and scaled by a force constant. The dihedral rotation terms employ a cosine potential instead. The non-bonded terms include van-der-Waals and electrostatic interactions, where the former is modeled using a 12-6 Lennard Jones potential and the latter by a Coulomb potential.

This mathematical modeling requires the definition of a large set of atom and chemical group dependent parameters for force constants, equilibrium geometries, partial charges, and van-der-Waals radii. Such a set of parameters, together with the sometimes slightly adapted potential equations is termed "Chemical Force Field". Various chemical force fields do exist, which vary in the way they were derived as well as their optimal areas of application. In the classical force fields, the parameters and topology, i.e., the type of covalent bonds, are predetermined and not conformation dependent. Yet, there are polarizable force fields, in which the partial charge of an atom can change depending on the surrounding chemical groups, as well as reactive force fields that allow formation and breaking of covalent bonds. Additionally, there are united-atom and coarse-grained force fields, in which multiple atoms are lumped together to larger interaction sites with the purpose of decreasing the computational expense at the cost of loss of accuracy. Some of the most commonly used force fields are summarized in **Table 1.1**.

Table 1.1. Non-exhaustive list of molecular mechanics and dynamics force fields

Name	Description	Ref.
UFF	Universal force field for small molecules in the gas phase	[65]
MMFF	Merck molecular force field is general purpose force field for small molecules	[66]
OPLS	Force field with optimized potentials for liquid simulations; is parameterized according to experimental bulk properties	[67]
CHARMM	Additive molecular dynamics force field with parameters for a wide range of molecules	[68]
AMBER	Molecular dynamics force field originally for proteins	[69]
BERGER	Parameter set optimized for lipid bilayer simulations; compatible with AMBER	[70]
GLYCAM	Force field for carbohydrates; compatible with AMBER	[71]
AMOEBA	Polarizable force field	[72]
GROMOS	United atom force fields for biomolecular simulations	[73]
MARTINI	Coarse-grained force field originally for lipid bilayer simulations	[74]

The choice of appropriate force field for the application is an important step in any molecular modeling effort. Therefore, many aspects must be considered. First and foremost, the force field should be accurate for the investigated molecular systems. Then, it must be supported by the molecular modeling or molecular dynamics software, which was chosen. When a new molecule needs to be parameterized, it is important that the parameterization method is consistent with the force field. Finally, the available computational resources and the size of the system further limit the choice of the force field. Furthermore, the selection of the solvent model plays a significant role for the accuracy and choice of the force field. In the framework of this research thesis plenty different molecular systems, ranging from small molecules, lipids and carbohydrates to proteins were investigated. For the sake of simplicity, comparability, and protocol transferability, it was decided to stick to the same force field for all different systems. Additionally, calculations with the coarse-grained MARTINI force field were performed, which was best supported by the GROMACS molecular modeling and simulation suite [75-81]. As force fields and the corresponding molecular modeling packages have always been developed in parallel and were mutually optimized, the transfer of one force field to a different modeling software is tedious and error prone. At the beginning of the thesis, both general molecular dynamics force fields AMBER and CHARMM were well-suited for the molecular systems of interest. However, based on the better transferability to GROMACS and the somewhat better accuracy for lipid bilayers [82], the CHARMM force field was chosen for all full-atomistic modeling efforts.

The CHARMM force field (and molecular mechanics engine), as initially developed by Nobel Prize laureate Martin Karplus and the 1980s, is an additive force field with parameter sets and equations for proteins [83], lipids [84], carbohydrates [85], nucleic acids [86], and for small, drug-like molecules [87]. The early full-atomistic protein force field, CHARM22 [68], employs the potential function of equation 1, which is still used in newer implementations. The terms for bonds, angles, dihedrals and non-bonded interactions are canonical. Additionally, CHARMM employs improper potentials, which applies to non-successively bound quartets of atoms. The angle omega corresponds to the out-of-plane angle. The Urey-Bradley terms are so called cross-terms, as they model the distance between the outer two of three bonded atoms (1-3 cross-terms).

$$\begin{aligned} V = & \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 \\ & + \sum_{dihedrals} k_\phi(1 + \cos(n\phi - \delta)) + \sum_{impropers} k_\omega(\omega - \omega_0)^2 \\ & + \sum_{Urey-Bradley} k_u(u - u_0)^2 \\ & + \sum_{non-bonded} \left(\epsilon \left(\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right) + \frac{q_i q_j}{\epsilon r_{ij}} \right) \end{aligned} \quad \textbf{(1)}$$

As stated before, the way of solvent treatment is equally important as the choice of the force field. Some force fields have been parameterized with implicit solvation and others with explicit solvation. In any case, the solvent model is mostly chosen to reproduce the qualities of liquid water. Several water molecule models exist, which differ widely in their accuracy and computational cost, yet are mostly transferable between the different force fields. Here, it must be noted that the solvation model has a tremendous effect on the accuracy of the model and that, considering the model volume and the ratio of atoms between solvent and solute, most of the computational effort goes into the calculation of the solvent. The complexity of water models can be classified by the number of interaction sites. There are 3-site (SPC, TIP3P), 4-site (TIP4P) and 5-site (TIP5P) water models. Water models with more than 3 sites, employ so called dummy atoms to model the electronic properties more precisely. The 3-site water models achieve high computational efficiency with reasonable accuracy and are thus heavily used in molecular dynamics simulations. In the CHARMM force field, a modified version of the TIP3P water model is used, in which the hydrogen atoms have Lennard-Jones potentials. The traditional SPC water model only has Lennard-Jones parameters for the oxygen [88].

The accurate quantitative description of flexible molecules in solution is clearly driven by the aim to identify the minimum energy conformation, i.e. the native state. However, when the forcefield

equation is subjected to mathematical optimization methods (i.e., minimization) with the atomic coordinates as variables, the potential function only converges to the closest local minimum. The global minimum though is much more challenging to identify. This is why efficient sampling methods had to be developed, which allow the generation of various conformations of which the energies can be compared. The two predominating methods are Monte-Carlo sampling (stochastic) and Molecular Dynamics (deterministic). It has been shown, that for gas-phase systems Monte-Carlo sampling is favorable whereas condensed phase systems, such as the ones investigated here, are better sampled using Molecular Dynamics (MD).

1.3.3 Molecular dynamics

The previous section was describing the potential energy of a molecular system, which allows the determination of locally optimal conformations (potential energy minima). This approach is however severely flawed because a complex molecule or molecular system can have a large number of local minima. A wide range of mathematical methods do exist, which allow global optimization of complex, multivariate systems, e.g., evolutionary algorithms [89], Bayesian optimization [90] or simulated annealing [91]. These methods were however never successfully applied to complex biomolecular systems. Possible reasons are the large number of variables (three times the number of atoms) and the high degree of inter-variable dependencies (bonds) and physical constraints. Instead, it is assumed that when a molecular mechanics model is allowed to also incorporate kinetic energy, i.e. a finite temperature, it would begin to traverse the potential energy landscape and eventually discover the global minimum (**Figure 1.9**). The concept of translating the molecular mechanics potentials into forces, which induce acceleration to the particles according to Newton's second law, is called Molecular Dynamics. Additionally, the method includes auxiliary algorithms to forward-integrate the atomic positions, to control temperature and pressure during the simulation, to deal with boundary conditions and to optimize computational efficiency.

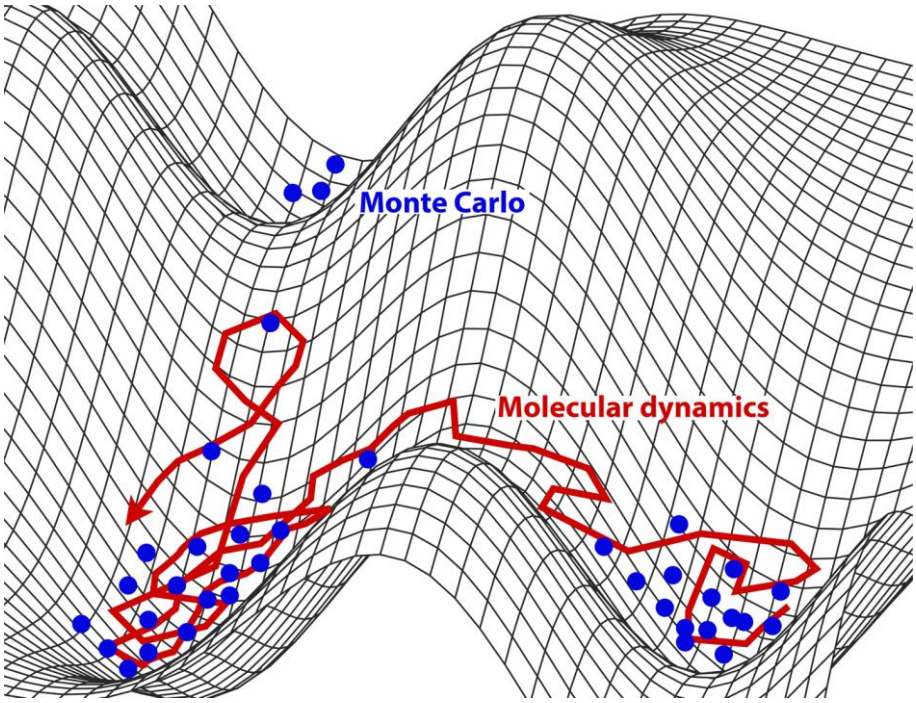

Figure 1.9: Schematic representation of a two-dimensional potential energy surface. The red line and blue points are visualizations of Molecular dynamics (red) and Monte Carlo (blue) sampling. (Image by Qx8134, 2020, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=94107414)

To numerically integrate Newton's equations of motion, i.e. approximate the atomic coordinates x and velocities v at a discrete future time step $t+\delta t$ based on the current time step t, the Verlet algorithm can be used. It can be derived from the Taylor series expansion:

$$x_i(t+\delta t) = x_i(t) + \delta t \frac{dx_i}{dt} + \frac{1}{2}\delta t^2 \frac{dx_i(t)^2}{dt^2} + \frac{1}{6}\delta t^3 \frac{dx_i(t)^3}{dt^3} + \cdots \quad (2)$$

The same can be done for the backward integration:

$$x_i(t-\delta t) = x_i(t) - \delta t \frac{dx_i}{dt} + \frac{1}{2}\delta t^2 \frac{dx_i(t)^2}{dt^2} - \frac{1}{6}\delta t^3 \frac{dx_i(t)^3}{dt^3} + \cdots \quad (3)$$

Addition of the equations (2) and (3), solving for $x(t+\delta t)$ and neglecting higher order terms yields:

$$x_i(t+\delta t) = 2x_i(t) - x_i(t-\delta t) + \delta t^2 \frac{dx_i(t)^2}{dt^2} \quad (4)$$

Equation (**4**) exhibits the accuracy of a third order Tayler approximation with terms of maximum second order. However, the Verlet algorithms has some striking downsides. The positions at the t-δt time step need to be known, which is usually not the case for $t=0$ in MD simulations. Also the velocities need to be calculated separately and require knowledge of the positions at $t+\delta t$ and t-δt, introducing a delay into kinetic energy (temperature) calculations, which introduces possible

instabilities into temperature coupling algorithms. Finally, equation (**4**) leads to numerical inaccuracies through the subtraction of two large terms.

Thus, the Verlet algorithm is utilized in MD integration in an optimized variant, called Velocity-Verlet algorithm. Therefore, in equation (**3**), *t* is substituted by *t+δt*:

$$x_i(t) = x_i(t + \delta t) - \delta t \frac{dx_i(t)}{dt} + \delta t^2 \frac{dx_i(t)^2}{dt^2} \tag{5}$$

Then, equation (2) plus equation (5) yields the equation for the velocity *v* at *t+δt:*

$$\begin{aligned} \frac{dx_i(t + \delta t)}{dt} &= \frac{dx_i(t)}{dt} + \frac{1}{2}\delta t \left(\frac{dx_i(t)^2}{dt^2} + \frac{dx_i(t + \delta t)^2}{dt^2} \right) \\ &= v_i(t + \delta t) = v_i(t) + \frac{\delta t}{2}\big(a_i(t) + a_i(t + \delta t)\big) \end{aligned} \tag{6}$$

The new coordinates are then given as:

$$x_i(t + \delta t) = x_i(t) + \delta t v_i(t) + \delta t^2 a_i(t) \tag{7}$$

Another way to eliminate the weaknesses of the original Verlet algorithm is the Leapfrog algorithm [92] which alternatingly calculates positions and velocities at different time points. Here it is noteworthy to point to out the importance of the selection of the time step *δt*. A too short time step will slow down the computation without increasing the accuracy. Oppositely, a too large time step might lead to uncontrolled atomic collisions, which result in extremely high velocities that technically displace atoms out of the simulation volume. The time step must be chosen as large as possible without risking such numerical instabilities. In classical molecular dynamics, time steps commonly reach from 1-5 fs.

As mentioned earlier, the velocities of the particles are corresponding to the set finite temperature and introduce kinetic energy to the system, enabling it to cross energy barriers on the potential energy landscape. Yet, how are the initial velocities generated and how is the temperature controlled during the simulation? Initial velocities must fulfil two criteria: I. they need to refer to the desired temperature and II. the total momentum of the system must be zero. Therefore, they are often stochastically generated by from a Maxwell-Boltzmann probability distribution:

$$f_v(v_x) = \sqrt{\frac{m}{2\pi kT}} \exp\left(-\frac{m v_x^2}{2kT} \right) \tag{8}$$

In equation (**8**), f is the probability of a velocity magnitude v in direction x, m is the mass of corresponding atom, k is the Boltzmann constant and T the temperature. Another method is to slowly heat the system from 0 K to the target temperature by gradually increasing the temperature of the coupled thermostat. One of the pioneering and still widely used approaches to maintain a target temperature was suggested by Berendsen et al in 1984 [93]. The Berendsen thermostat applies rescaling ($T_{scaled} = \lambda T_{current}$) to the particle velocities through a weak coupling with an external heat bath. The scaling factor λ in Berendsen coupling is defined as:

$$\lambda^2 = 1 + \frac{\delta t}{\tau}\left(\frac{T_0}{T(t)} - 1\right) \quad \textbf{(9)}$$

Here, δt is the time step, τ is the coupling constant, T_0 the bath temperature and T the systems temperature. The factor is based on a scaled temperature difference and allows a dampened (exponentially decaying) response. Hence, with Berendsen coupling, temperature fluctuations are explicitly possible. The coupling constant clearly plays a key role for the algorithm. If it is identical to the time step, the scaling appears immediately and not in a damped way. For the hypothetical case that τ becomes infinitely large, temperature coupling is disabled. Such a simulation would be called microcanonical or NVE, because the number of particles N, the phase volume V and the total energy E would remain constant. Finite Berendsen coupling constants yield an approximate but not a strict canonical (NVT) ensemble [94, 95]. For this reason, in recent years, Berendsen coupling was largely displaced by Parrinello-Rahman or Nose-Hoover temperature control. In a similar fashion to temperature control, the pressure can be controlled by scaling the simulation volume dimensions. In this case, the ensemble changes to NVP. A brief introduction to MD barostats is given in [96].

Whereas coupling algorithms decrease the performance of MD simulation, investigators have introduced many approximations to enhance it. In regard of short-range, non-bonded interactions, cutoffs are heavily employed in MD code and can even be considered part of the force field implementation. For example, CHARMM force field simulations frequently cut off Lennard-Jones and Coulomb interactions beyond distance of 1.2 nm because they elsewise converge zero asymptotically. To smoothen the transition from active Lennard-Jones potential to 0, a linearized switch function is implemented. Additionally, so called Verlet lists are employed, which encompass all neighboring atoms. These lists are only updated every couple of time steps. However, in biomolecular systems, long-range electrostatic interaction play important roles and cannot be neglected. This challenge is fortified by the fact, that biomolecular simulations heavily rely on periodic boundary conditions, in which the simulation volumes and the atoms therein are mirrored in all directions and distances are always measured between the closest mirror images. Thus, long-

range interaction must also take the mirror image convention into account. To solve this problem of computational expense, Ewald summation or more specifically the Particle Mesh Ewald method (PME) is applied [97-99]. It based on the idea, that the direct summation of interaction potentials is replaced by double summation of the short-ranged potentials in real space and the long range potentials in Fourier space. The long-range interaction can then quickly be calculated using the numerical algorithm of Fast-Fourier transform.

As the most important terms and concepts of MD simulation are introduced above, a general description of the MD workflow [100] will complete the chapter. Usually, MD begins with a structural model of the molecule of interest, which is taken from a structure database or, in the case of bulk liquids or two-dimensional assemblies, generated with a packing algorithm. In case of explicit solvation, solvent molecules are added to fill the desired volume. The volume must be chosen large enough to avoid interactions of the studied molecule with its replicates across the periodic boundaries. In biomolecular simulations, ions are added to neutralize the system and to avoid artifacts by an overly isolating medium. When all molecules (solvent, solute) are in place and the volume is defined, the system is subjected to minimization, mostly via the steepest descent algorithm. Is has the sole purpose to remove or at least highlight steric clashes and unphysical bond lengths which often occur when manual initialization is involved. When the energy has converged, initial velocities are assigned and the process of equilibration is initiated. During the first equilibration phase, the volume can remain fixed and only the temperature is equilibrated using a short time step of 0.5 – 1 fs. The temperature equilibration includes both the convergence of the average temperature toward the desired simulation temperature and a spatially isotropic temperature distribution. That is, no clusters of higher temperatures do exist. This phase usually takes no longer than a few ps. A second equilibration step is dedicated to the simulation box size, which is of course reciprocal to the density. Depending on the quality of the solvent packing, the pressure coupling algorithm and the simulated system (central solute vs bilayer), this phase can last from 0.5 to hundreds of ns. Temperature and volume equilibration can also be performed simultaneously. When temperature and volume have converged, final adjustments to coupling schemes and ensemble can be made and the production sampling can be started. In some cases it makes sense, to sample in NVT ensemble because it is slightly faster and the accuracy of results is not affected.

The final remarks of this section will be dedicated to advanced sampling strategies. In classical MD sampling, the system will be sampled for a certain time and conclusion will be drawn from the generated ensembles. In case of rare events, which require the overpassing of high energy barriers, classical sampling might not yield reliable results and is too much dependent on the starting

configuration. In such cases it is worthwhile to introduce some kind of bias. Such a bias can be either statistical/stochastic or physical. The simplest way to enhance sampling is to perform several replicates with slightly different initial conditions (velocities) [101]. This way, one can reduce the effect of artifacts occurring when a trajectory gets trapped in a deep local minimum. The procedure can be accentuated by a smart choice of restart configurations. For example, a region close to a certain transition state can be sampled by restarting the simulation from its proximity. Such sampling is called adaptive sampling [102, 103] and can be automatized to e.g. lead to transition path sampling [104]. On the other hand, it is possible to change the potential landscape via the addition of external forces or biasing potentials. Such simulations are called steered MD [105] or in a different context meta-dynamics [106]. A third way to enhance sampling is to alter the total energy of the system through a higher temperature [107]. This concept has culminated in replica exchange methods, in which several replica simulations are conducted at different temperatures, and transitions between the temperatures are commenced [108].

1.4 Quantification of molecular selectivity

A meaningful characterization and comparison with experiment of selective molecular recognition events can only succeed when a quantitative description is possible both from a macro- and a microscopic perspective. Furthermore, there must be a correlation between macroscopic and microscopic quantities. Here it is noteworthy, that physical, chemical, and biological experimentation rather yield macroscopic observables such as densities, energy differences or kinetic constants. The molecular simulations, on the other hand, yield trajectories (time series) of atomic positions and velocities. It is on the investigator to convert the microscopic state of the system into macroscopic quantities using principles of statistical mechanics and thermodynamics. In some cases, key quantities are directly accessible from both experimental measurements and ensembles of atomic coordinates. In others, the experimental observables are far beyond the time scales of MD simulation or require an ensemble of simulations with varying parameters.

1.4.1 Binding constants

In the area of biophysical chemistry, protein-protein or protein-ligand binding is usually quantified by either of the somewhat arbitrarily used terms affinity, binding energy and dissociation constant K_d [109]. But how are these terms derived? The aggregation of two molecules A and B into the complex AB can be described via the chemical reaction formula:

$$A + B \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} AB \qquad (8)$$

Here, the complex is in equilibrium with the free subunits. The association reaction appears with a rate constant k_{on}, whereas the reverse reaction, the dissociation of the complex, happens with a k_{off} rate constant. The association and dissociation rates would then be

$$r_{on} = k_{on}[A][B]$$
$$r_{off} = k_{off}[AB] \quad (9)$$

where square brackets denote the concentration. To fulfill the law of mass action, on- and off-rate must be equal:

$$k_{on}[A][B] = k_{off}[AB] \quad (10)$$

Hence, in thermodynamic equilibrium, the concentration ratio between dissociated and associated states is identical to the ratio of the rate constant, which is customarily lumped into the single dissociation constant K_D:

$$\frac{[A][B]}{[AB]} = \frac{k_{off}}{k_{on}} = K_D \quad (11)$$

As the dissociation constant is an equilibrium constant, the general expression

$$\Delta G_{bind} = RT \ln K_D \quad (12)$$

where R is the general gas constant and T the absolute temperature, can be applied. ΔG is then the change in free energy for the dissociation of the complex at constant temperature and pressure. Thus, when the equilibrium is on the site of the dissociated subunits, then $k_{off} > k_{on}$, $K_D > 1$ and $\Delta G > 0$. That is, energy is necessary to bring the subunits together. In the opposite case, $k_{on} > k_{off}$ leads to a $K_D < 1$ and thus $\Delta G < 0$. Energy will be released upon complexation and the reaction will take place spontaneously. In biochemistry, the term affinity is based on the reciprocal of the dissociation constant – the association constant. In a quantitative sense, nowadays only the dissociation constant is used.

The binding energy plays an essential role in various biochemical processes and is the driver for selectivity [110]. In fact, when a receptor A has an affinity for a ligand B, and an even higher affinity for a ligand C, ligand C is able to displace ligand B from the complex AB:

$$AB + C \rightleftharpoons AC + B \quad (13)$$

In this common scenario, where for example a drug molecule displaces the natural ligand from a receptor and blocks the binding site to inhibit signal transduction, the equilibrium constant can be calculated from the dissociation constants of the elementary reaction:

$$K_{B,C} = \frac{[AC][B]}{[AB][C]} = \frac{K_{AC}[A][B][C]}{K_{AB}[A][B][C]} = \frac{K_{AC}}{K_{AB}} \quad \textbf{(14)}$$

$K_{B,C}$ can hence be called selectivity coefficient.

The theoretical considerations raise the questions of what orders of magnitude typical dissociation rates and binding energies have and how they can be determined experimentally and computationally. Depending on the molecular system, i.e., protein-protein or protein-small molecule, a range of qualitative, semi-quantitative and quantitative experimental methods are available. Common qualitative or semi-quantitative experimental approaches for protein-protein interaction elucidation include ELISA, pull-down and co-immunoprecipitation assays (or band shift assay for nucleic acid binding proteins) [111]. They work by immobilization of one protein, which is allowed to interact with its labeled or tagged binding partner. Upon interaction, a signal is released. On the other hand, a large array of quantitative methods are available as reviewed in [112]. The review classified the methods into separative methods, where the ligand is separated from the receptor and the equilibrium concentration of either is directly measured and non-separative methods. The latter relies on the detection of a change in a physical or chemical property induced upon binding. Among the separative methods are equilibrium dialysis, ultrafiltration, liquid chromatography or capillary gel electrophoresis. Common non-separative methods are based on spectroscopy, calorimetry or surface plasmon resonance.

Dissociation constants span multiple orders of magnitude and ligands (protein or small molecules) are loosely classified as weak binding, moderate binding, strong binding, and very strong binding. Weak binders have dissociation constants in the millimolar (10^{0}-10^{-3} M) range. Moderate binding can be considered with micromolar (10^{-3}-10^{-6} M) dissociation constants. Strong, high-affinity binding is present when the dissociation constant is nanomolar (10^{-6}-10^{-9} M). Sub-nanomolar, i.e., picomolar binding is characterized by even smaller dissociation constant. It has to be noted, when determining K_D with one of the above methods, the values of k_{on} and k_{off} are not necessarily co-determined (except in SPR). Thus, a direct readout of the stability of the complex only based on K_d is not feasible. However, for most instances, the association of two species free roaming in solution is limited by diffusion. Such an upper limit for k_{on} is often quoted to 10^{9} M^{-1} s^{-1} [113, 114]. Depending on the environment, e.g. in a crowded cell or a lipid bilayer, it can be lower. In other cases, when strong, far-reaching, favorable electrostatic interaction between the binding partners

are present, the diffusion limit can also be surpassed [115]. Typical values are between 10^5 to 10^6 M^{-1} s^{-1} [116]. Considering a complex with a K_D of 21 pM, as for example recently reported for a bivalent ligand for a G-protein coupled receptor, and a k_{on} constant of 10^5 M^{-1} s^{-1}, k_{off} value would be 2.1 x 10^{-6} s^{-1} [117], which translates into a complex half-life of 90 h.

1.4.2 Computational estimation of binding constants

The time scales of binding and unbinding events are usually far beyond the limits of equilibrium molecular dynamics sampling. Additionally, such experimental values are averages and are difficult to interpret on the single molecule scale as investigated via molecular simulations. For instance, one would either need a gigantic simulation volume with such a large number of receptor and ligand molecules, that they would reach the same concentration as the in the test tube, or one would have to simulate one receptor-ligand system long enough to sample multiple binding and unbinding events so that the complex half-life would eventually converge. Neither is possible even with recent high-performance computing nodes. Thus, two branches of developments can be distinguished to partially resolve this issue. On the one hand, enhanced ensemble free energy methods allow a reasonably accurate, ab initio estimation of the binding free energy of a distinct ligand to its receptor using a set of molecular dynamics trajectories. Such methods include thermodynamic integration [118], exponential averaging [119] (free energy perturbation [120]) or umbrella sampling [121]. On the other hand, there are fast, simplified and rather approximate methods for screening purposes. These methods include docking, MM-PBSA [122] or PRODIGY [123]. For prediction of especially k_{on} there is e.g., the software SDA [124].

In the case of umbrella sampling, the conformational space of a molecule along a reaction coordinate is sampled. The individual trajectories are generated by application of an additional potential which is able to drag a ligand out of a binding site or lipid out of a bilayer. From the unbinding trajectory, equidistant snapshots are used as initial configurations for subsequent umbrella sampling. Here, in each trajectory, the initial point in terms of the reaction coordinate is restrained by the same potential as previously used. The trajectories are finally subjected to the weighted histogram analysis method (WHAM) [125, 126] with the purpose to calculate the potential of mean force, i.e. the free energy difference along some reaction coordinate. Umbrella sampling can yield a reasonable unbinding trajectory, however it is not suitable to efficiently compare different ligands to or predict a binding mode.

Thermodynamic integration allows the calculation of the free energy difference between two states. It is an alchemical method, which means that the system undergoes a non-physical transformation from one state to another. This is reasonable because the free energy difference is only dependent on the start and end states and not on the path itself. The alchemical path is characterized by its

path variable λ ($0 \leq \lambda \leq 1$), which is a scaling factor for the non-bonded interactions. Thus, λ allows to slowly couple and decouple a chemical group or a whole ligand from its surroundings. For example, to compare to the binding free energy difference of two ligands one might either decouple ligand A from the receptor and couple ligand B into the receptor, which would afford to simulate the (de)coupling of both ligands into (from) solvent, or to transform ligand A into ligand B **(Figure 1.10**). The free energy difference is ultimately calculated from integrating the derivate of the λ-dependent potential energy function with regard to λ over the whole λ-range from 0 to 1. The method is accurate yet expensive. The binding mode of the ligand must be known *a priori*.

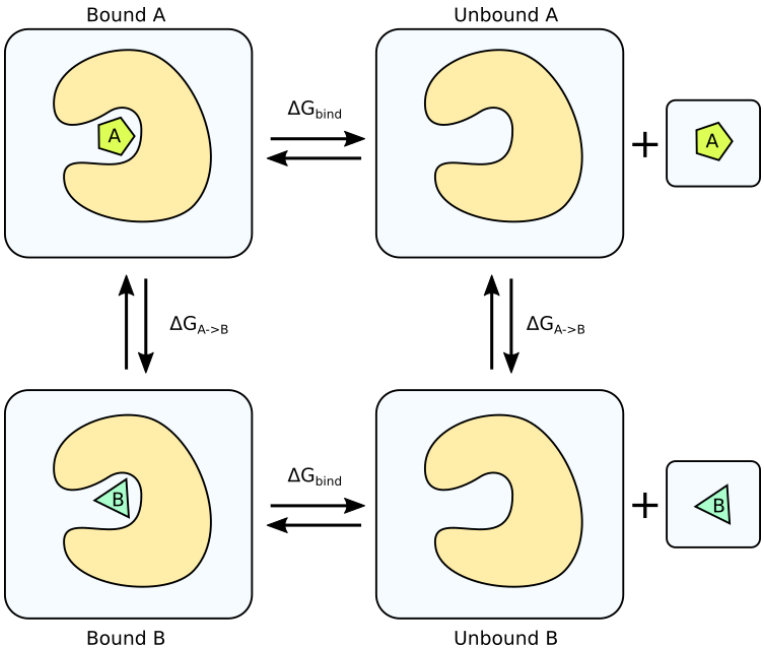

Figure 1.10: Thermodynamic path of two ligands binding and unbinding from the same receptor. The binding free energy difference between two ligands A and B can be modeled either by unbinding of A and binding of B or via morphing of A to B (FEP).

As opposed to the molecular dynamics-based free energy methods, molecular docking is a technique which allows rapid screening of ligand position, orientation and conformation relative to a receptor protein. Due to its quick energy evaluation, docking is suitable for small-molecules ligands as well as proteins. Docking is used to predict the binding mode of a ligand as the best scoring pose of a generated ensemble. The score is usually a semi-empirical binding energy function which was derived in close reference to molecular mechanics force fields. A canonical equation [127] can be summarized as the sum:

$$\Delta G_{bind} = \Delta G_{solv} + \Delta G_{conf} + \Delta G_{int} + \Delta G_{motion} \quad \textbf{(15)}$$

Here, the binding energy, or more precisely the free energy difference between unbound and bound state, is the sum of the solvation (hydration) free energy ΔG_{solv}, the free energy difference of conformational changes in receptor and ligand ΔG_{conf}, the interaction free energy ΔG_{int} and the free

energy contribution from the change in mobility of receptor and ligand ΔG_{motion}. The quality of a docking pose can be assessed by calculating the difference in binding free energy relative to the experiment as well as by the average atomic deviation of the ligand from the crystalized complex. Benchmark studies have shown that recent docking software packages (Autodock [128], Vina [129], Glide [130], RosettaDock [131], GOLD [132]) are capable of predicting the correct binding pose and to predict binding energies with an error of around 50% [133]. The quality of a docking simulation depends heavily on the dynamics of the protein and the number of rotatable bonds of the ligand. Especially, because protein flexibility is only sparsely taken into account, the selection of protein receptor input conformation is of utmost importance [134, 135]. In the best-case scenario, the free protein is in the same conformation when bound to the ligand of interest. This however is only possible in screening efforts were the structure of a ligand-bound complex is known. In cases when the binding mode of the ligand is to be predicted, multiple receptor conformations should be employed. Such an approach is called ensemble docking. The conformational ensemble can be generated via crystallography, NMR or MD simulation [136].

A complementary theoretical method to compute especially protein-protein binding free energy is PRODIGY [123], which can also be used to rescore results from protein-protein docking. PRODIGY takes a protein-protein complex as input and predicts the binding free energy by calculating inter-residue contacts. It is based on a regression model of the number of certain classes of inter-residues contacts versus experimental binding free energies of a benchmark set of 81 protein-protein complexes. The method is based on the observation that the correlation between binding free energy and certain inter-residue interactions (ICs) is higher than between the binding energy and differences in buried surface area. Thus, the final regression model only includes terms for inter-residue interactions and non-interacting surface area. The non-interacting surface area (NIS) is an important normalization term for taking the size of the interacting molecules into account. The resulting, optimized PRODIGY model is described as:

$$\begin{aligned} \Delta G_{calc} = {} & 0.095\, ICs_{Q/Q} + 0.100\, ICs_{Q/P} - 0.196\, ICs_{P/P} + 0.227\, ICs_{P/A} \\ & - 0.187\, NIS_A - 0.138\, NIS_Q + 15.94 \end{aligned} \tag{16}$$

The indices stand for Q: charged, P: polar and A: apolar. The model yields a correlation coefficient of -0.75 for rigid and -0.73 for flexible proteins. One recurring question in such calculations is the definition of an inter-residue contacts. In the development stages, it is customary to attempt various cutoffs e.g. heavy-atom minimum distances in the range of 0.5-0.8 nm. Other definitions are center-of-mass distances or alpha or beta carbon distances. The choice heavily depends on the application

and is of significant effect on the results. In the PRODIGY method, a distance threshold of 0.55 nm was employed.

Another method was suggested by Kolmann et al. in the late 1990s and combines molecular mechanics with the Poisson-Boltzmann equation and surface area solvation: MM-PBSA [137]. The Poisson-Boltzmann equation [138] and its commonly applied approximation, the linearized Poisson-Boltzmann equation allows the calculation of intermolecular interactions between molecules in an ionized solvent, such as NaCl dissolved in water. It is mostly used to compute the solvation free energy. The MM-PBSA method is based on the idea that the free energy of state G (not to be confused with free energy difference ΔG) can be estimated by:

$$G = \mathrm{E_{MM}} + G_{PBSA} - TS_{MM} \quad \textbf{(17)}$$

Here, E_{MM} is the molecular mechanics potential energy (see force field equation), G_{PBSA} is the solvation free energy from a Poisson-Boltzmann calculation and a surface area term [139], and TS_{MM} is the entropy, which can be estimated e.g. with normal mode analysis [140]. Normal modes describe oscillating movements of biomolecules such as bond vibrations of collective motions of domains (see protein dynamics) and can be elucidated with molecular dynamics and normal mode analysis [141]. Anyway, based on eq. **17**, one can calculate the free energy difference upon binding of a ligand to receptor via:

$$\Delta G_{bind} = G_{complex} - G_{receptor} - G_{ligand} \quad \textbf{(18)}$$

In practice, the terms are evaluated on an ensemble of snapshots generated by MD sampling and subjected to averaging. Theoretically, it is advised that complex, receptor, and ligand would be simulated separately, however it has been shown that the loss of accuracy by only simulating the complex is small. In their original articles, the authors investigate the binding free energies of a set of six protein-ligand complexes which all were analogs biotin binding to avidin. They reached a strikingly high correlation coefficient of 0.92 between experimental and calculated binding free energies. Since, MM-PBSA and the tightly related method MM-GBSA, (GB: Generalized Born) were heavily used to estimate binding free energies of small molecule ligands as reviewed in [122]. The review also reveals the major downside of the approach: poor precision. The precision can however be increased by using many replicates of sufficiently long trajectories, which inevitably increases computational cost and limits the area of application. The accuracy for protein-protein complexes was assessed by Chen et al. [142] using 46 complexes. Their studied revealed experiment-simulation correlation coefficients between -0.375 and -0.523, rendering the method inferior to PRODIGY for protein-protein complexes.

1.4.3 Indirect quantification of binding

From a molecular dynamics trajectory, or an ensemble thereof, many valuable information can be extracted. In the case of a complex of a protein and small molecule or protein ligand, the binding can most easily be characterized by the magnitude of the fluctuations of position and orientation of the ligand relative to the receptor. It can be imagined, that a ligand that binds tightly and stably to the receptor and retains its position and orientation throughout the trajectory, will most likely exhibit a high binding affinity. The one arguably most commonly used quantity for the quantitative comparison of molecular conformations is the root mean-square deviation (RMSD). Between two structures *i* and *j*, it is calculated as:

$$RMSD(i,j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left[\left(x_{k,i}-x_{k,j}\right)^2+\left(y_{k,i}-y_{k,j}\right)^2+\left(z_{k,i}-z_{k,j}\right)^2\right]} \quad \textbf{(19)}$$

Here, N is the total number of atoms of interest, and *x*, *y* and *z* correspond to the three spatial coordinates. To compare only the conformations by means of bonds, angles and torsions, global translational and rotational deviations must be removed. Therefore, the structure of interest *j* is structurally aligned to the reference *i* by least-square-fitting of the atomic coordinates under variation of the relative position and rotation. The Kabsch algorithm [143] can be used for the fast calculation of a starting structure for the superpositioning. For proteins, usually the alpha carbons atoms are used for fitting, whereas the backbone or whole protein can be used for the RMSD calculation.

From MD simulations, the RMSD is routinely calculated and plotted as a function of the simulation time frame and relative to the initial frame or the crystal structure. The time-evolution of the RMSD allows to make statements on the stability and convergence of the simulation, the conformational space of the protein (molecule) as well as the discrepancy between different replicas. For trajectories of molecular complexes, it can be expressive to align the whole complex by the coordinates of only the receptor atoms, and to then calculate the time evolution of the RMSD for the ligand atoms. With such an approach, rotational and translational motions of the ligand relative to the receptor are explicitly taken into account. Hence, the stability of the ligand binding mode as an indirect measure for binding affinity can easily be spotted from a set of equilibrium MD simulations. Another application for the RMSD is the calculation of the pairwise RMSDs between various conformations as a distance metric for clustering and embedding algorithms (see section "Advanced statistical analysis"). Of similar routine-use is the root mean-square fluctuation (RMSF), which allows the localization of stable and dynamic regions within a molecule. The difference to

the RMSD is that the positional deviation is averaged over the ensemble (time frames) instead of over the number of atoms.

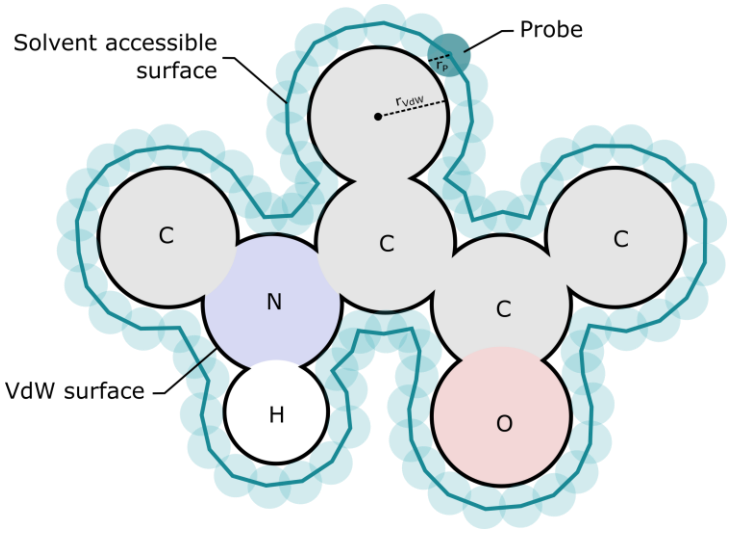

Figure 1.11: Construction of the solvent-accessible surface for an arbitrary molecule. A peptide backbone is drawn as an example. The probe is drawn unusually small for clarity. The number of generated probe points determined the resolution and smoothness of the calculated solvent accessible surface.

Of similar general interest and also as a quantitative means to evaluate molecular binding is the solvent accessible surface area (SASA) and its change upon complexation. The approach allows the calculation of a molecular hull or surface, which confines the molecule from the solvent and reveals the atoms at the boundary with frequent solvent interactions. It was first suggested by Lee and Richards in 1971 [145] and later mathematically optimized by Shrake and Rupley in 1973 [146]. It can be imaged as a surface generated by rolling a spherical probe over the Van-der-Waals surface of the molecule (**Figure 1.11**). The Shrake and Rupley algorithm employs the following steps: 1. Generation of a mesh of points at equidistant positions to the atoms. The distance is determined by the Van-der-Waals radii of the atoms and the probe radius. The probe radius is often set to 0.14 nm, corresponding to the size of water molecule. An elegant way to place points on a sphere are the Fibonacci Sphere also known as Golden Spiral method [147]. 2. Test, which points are not within the Van-der-Waals plus probe radii of the other atoms. These points are considered solvent accessible and their fraction is proportional to the solvent accessible surface area. Precisely, it is the fraction of accessible points multiplied by the sum of the spherical areas around the atoms defined by the Van-der-Waals radii and the probe radius. Besides the Shrake-Rupley method, few other approximations for the SASA have been proposed and are frequently used such as the LCPO method [148] or the power diagram method [149].

In a protein, it makes sense to accumulate the areal contributions of the protein residues to show which residues are more or less solvent exposed. This, however, affords a normalization strategy because larger amino acids would naturally occupy a larger area. Thus, the term relative solvent

accessibility (RSA, also called solvent exposure) was established, which expresses the SASA of an amino acid relative to its maximum possible SASA. The maximum SASAs have been investigated in a range of theoretical and empirical studies by [144],[145] or most recently by [146]. They are usually based on theoretical considerations of G-X-G tripeptides. Earlier studies employed an extended peptide conformation to estimate the maximum SASAs, whereas the recent values use bent conformation such as found in alpha helices. Intuitively, the curved conformation allows even higher maximum SASAs. The RSA takes values between 0 for fully buried residues and 1 for residues, which are most solvent exposed. Particularly for protein-protein interactions, the change in solvent accessibility upon complexation is of high interest and is an indirect measure of binding strengths. The surface area, which is accessible in the single proteins and inaccessible in the complex is called buried surface [147]. Additionally, the identification of certain residues that undergo large changes (become buried) are often key drivers for the association and recognition. They can thus be considered binding hot-spot residues and may be targeted by small-molecule ligands [148].

The analysis of the interface area can be accompanied by a thorough elucidation of the **intermolecular interactions** between receptor and ligand. The presence of a large number of stable, favorable interactions such as hydrogen and ionic bonds or hydrophobic interactions correlates with a high affinity [149, 150]. To quickly identify interactions within molecular dynamics trajectories, intermolecular contact occupation matrices or maps are frequently employed. [151] Therefore, initially the pairwise inter-residue distances are computed. As discussed before (see Chapter 1.3.4), the definition of inter-residue distance varies among the applications. For intramolecular residue-residue distances, alpha or beta carbon distances are used. For intermolecular distances, rather minimum-heavy-atom distances of the whole amino acid or only the sidechains are utilized. Center-of-mass or center-of-geometry distances are even further options. In a next step, for every residue-residue pair, the ratio of frames in the trajectory is counted, in which the pairwise distance is below a certain threshold. The threshold (also called cutoff) depends on the distance metric as well as the application. For a loose screen of possible interactions, a heavy atom minimum distance cutoff of 0.6 nm might be sufficient, for the identification of allosteric intramolecular networks an alpha carbon distance of cutoff 0.8 nm might be more expressive [152]. In the end, the contact matrix includes occupation values for each residue-residue pair based on the chosen threshold. It is obvious, that an extended cutoff will yield higher occupations and vice versa.

The contact matrix includes information on both the receptor and the ligand. It is usually rather sparse, because the majority of residues will not be involved in the interactions. Thus, it is

customary to filter the contact by either an occupancy cutoff e.g. 0.75 or 0.9, or by complementary information such as solvent accessibility or buried area. Additionally, the matrix can be projected to either the site of the protein or the ligand. Therefore, the pairwise occupancies can be cumulated row- or column-wise to yield residue-wise contact occupancies. Apparently, the residue-wise contact occupancy can exceed the value of 1, when e.g. one receptor residue is frequently in contact with multiple ligand residues. As this can be misleading, the residue-wise occupancy may be limited to 1, or the residue contribution to the total contact occupancy might be used. In the latter case, the occupancy matrix is normalized by its overall sum. The pairwise, or residue-wise contact contribution matrix (vector, respectively) indicates the relative importance of certain residues to the binding energy and can be considered another theoretical hotspot identification approach.

When the major interactions are identified, they can be further classified. For protein-protein interactions as investigated by MD simulations, it is feasible to roughly classify the interaction pairs based on the residue type. Amino acids are classically grouped into apolar (hydrophobic), polar (neutral), charged (basic and acidic) categories . The hydrophobic amino acids are characterized by aliphatic or aromatic sidechains with none or only weak hydrogen bonding capacities. This group includes alanine, valine, methionine, leucine, isoleucine, proline, tryptophan, and phenylalanine. Polar amino acids carry hydroxyl (thiol) or amino groups and can thus engage in hydrogen bonding. Tyrosine, threonine, glutamine, glycine, serine, cysteine and asparagine belong to this group. Lysine and arginine are protonated under physiological pH conditions and thus positively charged. Glutamate and aspartate carry carboxyl groups and are hence deprotonated and negatively charged at neutral pH. Histidine is an exception as its pKa is close to 7 but it is depending on the molecular environment and hydrogen bonds with neighboring residues [153]. Canonically, it is considered basic. It can, however, be positively charged in some instances. Considering protein-protein interactions which are mostly sidechain-mediated, the most favorable interactions occur between similar groups such as hydrophobic/hydrophobic, polar/polar or basic/acidic (acidic/basic, or more generally charged/charged). In particular, polar/charged interactions are also favorable, because charged groups can engage in so called short hydrogen bonds [154].

For a more detailed analysis, the presence and quality of hydrogen bonds must be explicitly investigated. In fact, the free energy contribution of a certain hydrogen bond towards the total binding free energy depends on multiple geometric and chemical properties and ranges between 2 and 7 kcal/mol in most biomolecular systems. Various methods have been proposed to estimate the hydrogen bonding energy [155]. In general, a hydrogen bond is characterized by three atoms, the donor heavy atom with the bound donor hydrogen, and acceptor heavy atom. The donor and acceptor atoms are more electronegative than the hydrogen. The distance between donor heavy

atom and acceptor heavy atom is around 0.3 nm and shorter. The angle between the three atoms is close to 180°. Briefly, the shorter (up to a certain threshold, i.e. covalent bond distances) and the closer to linear a hydrogen bond, the higher its dissociation energy [156]. Even though hydrogen bonds and their tremendous impact on biology and chemistry are known for many years, they are still subject of ongoing research and particular features remain a matter of debate [157].

1.5 Statistical analysis of conformational ensembles

Complex formation between two proteins or a protein and small molecule ligand can induce conformational changes in both the receptor (protein) and the ligand via an induced fit or conformation selection mechanism. In enzymes, the ligand-induced conformational changes might directly affect the catalytic center and thus be substantial part of a selectivity mechanism [158]. Additionally, the catalytic cycle of an enzyme may encompass steps, in which, after catalytic reaction, the protein adopts a conformation that literately ejects the product from the binding site [159]. The conformational flexibility in the realm of biomolecular reactions is often based on bond rotations rather than bond stretching or angle bending. Thus, conformational analysis often includes dihedral angles.

1.5.1 Dihedral angle analysis

One of the earliest and most commonly conformational analysis of proteins is the Ramachandran plot [160]. It is a scatter or contour plot of the peptide backbone torsions angles ψ over φ (**Figure 1.12**). ,The angle ω corresponds to the torsion angle over the planar peptide bond. In most structural biology applications, it remains 180° and is thus rarely explicitly considered. The position of a point in the 2D torsion angle map, which corresponds to the backbone conformation of a distinct amino acid, allows statements on the secondary structure of the protein at the position. Thus, the Ramachandran plot of a single protein yields insight on its global structure and degree if disorder. Ramachandran et al. realized that the torsion angular space of the protein backbone is constrained to distinct regions (fully allowed and outer limit regions) due to steric hindrance.

In MD simulations, it can be revealing to monitor certain torsion angles of a single residue over the whole trajectory and to visualize the data in the same way as Ramachandran et al. Residues in a rigid region of a molecule will yield only points in a locally confined area of the plot, whereas flexible regions might occupy multiple, distinct areas. When sufficiently large ensembles of tens of thousands of conformations are available, it is customary to estimate the density of points at distinct positions by either binning or Kernel density estimation [161]. Naturally, the density of the points is then equivalent to the probability density of the conformation at the particular position. A

visualization of the point density or the probability density function over the conformational space (φ, ψ) is called probability map.

Analog to equation **12**, we can use the probabilities to construct a free energy map or free energy landscape

$$\Delta G_{(1\rightarrow 2)} = -k_b T \ln \frac{P_1}{P_2} \quad \textbf{(20)}$$

where $\Delta G_{(1\rightarrow 2)}$ corresponds to the free energy difference for the transition from conformation 1 with the probability density P_1 to the conformation 2 with the probability density P_2. The factor k_B is the Boltzmann constant with a value of 1.38×10^{-23} J/K and T is the absolute temperature. In practice, state 2 is frequently chosen as minimum energy state, i.e. the point of the highest density, and the free energy map is constructed relative to it [162].

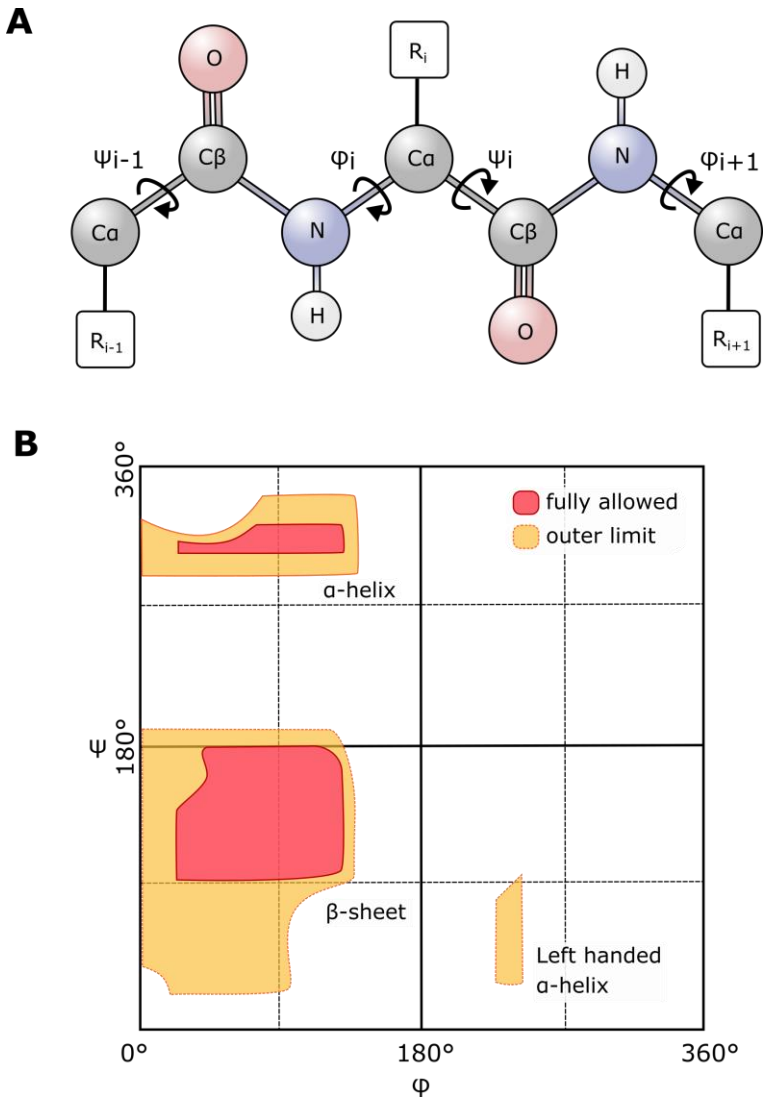

Figure 1.12: Backbone torsion angles define predetermine secondary structure. A) Schematic representation of a protein backbone with annotated torsion angles φ and ψ. The Cα hydrogen is omitted for clarity. B) The original 1963 Ramachandran diagram redrawn in modern aesthetics.

1.5.2 Dimensionality reduction

For one or two dimensions, the above considerations are easy to understand and mathematically implemented. However, in practice, often more than two torsion angles are necessary to describe the dynamics of a protein. The fact, that molecular torsion angles are often not statistically independent from each other, adds to complexity. Thus, the estimation of the joint probability must be performed in the high dimensional space. The number of bins increases exponentially with the number of dimensions and the density tensor becomes increasingly sparse. Kernel density estimation could theoretically be a reasonable alternative but suffers from numerical problems especially with infinite Kernels (such as Gaussian) or when fast Fourier transformation as applied to speed up the expensive calculation. Therefore, two approaches are increasingly employed individually or in combination: dimensionality reduction and clustering. Here, it is important to

mention that torsion angles are periodic (circular) quantities. That means the two-dimensional flat reorientation of the Ramachandran plot is actually a distortion of the true topological space, which would be better reflected by a torus. This must be taken into account, when the above methods are applied. For low dimensions, it can be sufficient to transform angles φ into two-dimensional numbers z = (cos φ, sin φ) [163, 164].

Dimensionality reduction methods utilize the intrinsic structure of the high-dimensional data to generate a lower dimensional projection or embedding. The aim of dimensionality reduction techniques is to visualize the high-dimensional data under minimal loss of variance and maximum conservation of the pairwise differences. This would allow that points, that are pairwise similar (by whatever similarity metric) in the high-dimensions, are also mapped together in the low dimensional representation. Dimensionality reduction is nowadays considered a subfield of machine-learning, whereas it was traditionally considered multivariate statistics. One if not the most commonly employed technique is the *principal component analysis* with its many variations and generalizations [165]. Mathematically it is equivalent to singular value decomposition and based on the projection of the data along the vector of the highest variance. Considering the data is defined in n-dimensional Cartesian space and is standardized i.e. centered and normalized, the PCA algorithm is simple and robust. Initially, the covariance matrix is computed:

$$\mathrm{C} = \begin{bmatrix} Cov(x_i, x_j) & \cdots & Cov(x_i, x_n) \\ \vdots & \ddots & \vdots \\ Cov(x_n, x_j) & \cdots & Cov(x_n, x_n) \end{bmatrix} \tag{21}$$

The vectors x_i and x_j ($i,j \leq n$) correspond to columns or features of the data matrix X of shape (m,n) where m is the number of samples and n the number of features (dimensions). The covariance matrix is symmetrical and the main diagonal holds only the variances of the data because $\mathrm{cov}(x_i, x_j) = \mathrm{var}(x_i)$.

Solving the eigenvalue problem

$$Cv = \lambda v \tag{22}$$

yields n eigenvalues λ_i and eigenvectors v_i of the covariance matrix. The eigenvector corresponding to the largest eigenvalue yield the projection with smallest projection error and largest variance. The second eigenvector is orthogonal to the first one, and has the second largest variance, and so on. To project the data into e.g. two dimensions, the first two eigenvectors are multiplied with the original data X to attain the transformed data X*.

$$X^* = [v_1, v_2]^T * X^T \tag{23}$$

PCA has for example been be employed to project torsion angle data of peptides and to calculate the free energy map in the projected space [162]. In this way, the conformational space of a molecule can be visualized in a two-dimensional map and free energy differences between distinct conformations can easily be estimated. The downside is, that especially when features are periodic, PCA is not able to separate point clouds in the projection and energy wells appear deeper than they actually are.

Depending on the application, variants of PCA may be employed. Especially in the field of Markov state modeling, the time-lagged independent component is used. Markov state modeling is process of generating a Markov or hidden Markov state model by discretization of the state space and subsequent estimation inter-state transition rates. Markov state models are versatile tools to analyze, understand and visualize the dynamics of complex systems and have enjoyed a reasonable hype in the late 2010 years. However, their generation is tedious and accurate Markov state models afford extensive sampling. Additionally, the number of states and transitions should be humanly comprehensibly. Thus, traditional Markov state modeling is slowly replaced by elegant, self-enforcing machine learning methods. The field of canonical Markov state models is nicely reviewed by Husic and Pande [166]. Anyway, time-lagged independent component analysis (tICA) still is of value to reveal slow transitions in molecular dynamics trajectories and is therefore utilized to identify collective variables for enhanced sampling methods [167, 168]. The advantage of tICA over PCA is that it includes time information via a lag parameter τ in a way that the eigenvalues and eigenvectors are not computed from the covariance matrix, but a time lagged covariance matrix [169].

Not a variant but rather generalizations of PCA are the manifold learning embedding methods t-distributed neighbor embedding (tSNE) and its successor UMAP: uniform manifold approximation and projection for dimension reduction. The former was proposed in 2008 by Van der Maaten and coworkers [170] and has since seen extensive use in all fields of data science as well as in cell biology and molecular simulation. The method allows embedding and exploration of large high dimensional datasets with unprecedented conservation of local and global structure. The key idea of t-SNE is that it carries over probabilities distribution from the high dimensional data to the low dimensional embedding. This is in contrast to the most basic embedding approach (multidimensional scaling), which aims to reproduce the distances. The probability distributions are two dimensional and encode the pairwise similarities. The algorithm then minimizes the Kulback-Leibler divergence [171] between the high-dimensional and the embedded probability distribution. Briefly, for discrete distributions as in the above case, the Kulback-Leibler divergence is the expectation value of the logarithmic differences between two probability distributions. Ten

year later, the UMAP algorithm has been proposed [172] and quickly reached tremendous interest in especially the single cell community [173]. It is largely considered superior to t-SNE based on its better performance, lower memory consumption and better preservation of the global data structure. UMAP is centered on complex topological data analysis and a full derivation and description of the algorithm is beyond the scope of the thesis. Basically, the initial step is the generation of k-simplices by the high-dimensional data points. A k-simplex is the convex hull spanned by k+1 points, i.e. a line connecting two points or a triangle spanned by three points, etc. In the next step, the simplices are connected to a simplicial complex if they share a face. Comparing the simplicial complex with a nearest neighbor approach shows that most information is encoded in the 0- and 1- simplices. Thus, the topological problem of the simplices can be translated into a graph and the embedding converges to a graph layout problem which can traditionally be solved via spectral methods such as Laplacian eigenmaps or diffusion maps. The full algorithm is of course much more complicated to be able to deal with the complexity and diversity of real world data. For instance, mathematical finesses, such as locally varying metrics and fuzzy open sets are utilized.

The three methods (PCA, t-SNE and UMAP) are exemplarily shown for a subset of the MNIST digit dataset. It contains 70.000 images of handwritten digits in 28x28 pixel resolution. Each image can be converted to a feature vector with a length 784 and entries between 0 and 1. (**Figure 1.13 A**) In the two-dimensional projection by PCA, a weak separation of the different clusters can already be seen. This is significantly improved in the t-SNE and UMAP embedding. It shows that both the embedding methods still have difficulties to distinguish between 7 and 9 as well as between 9 and 3. The clusters of 0 and 1 are well separated. (**Figure 1.13 B**) Anyway, the visualization highlights the huge advantage of embedding methods to visualize high-dimensional data in a low dimensional space.

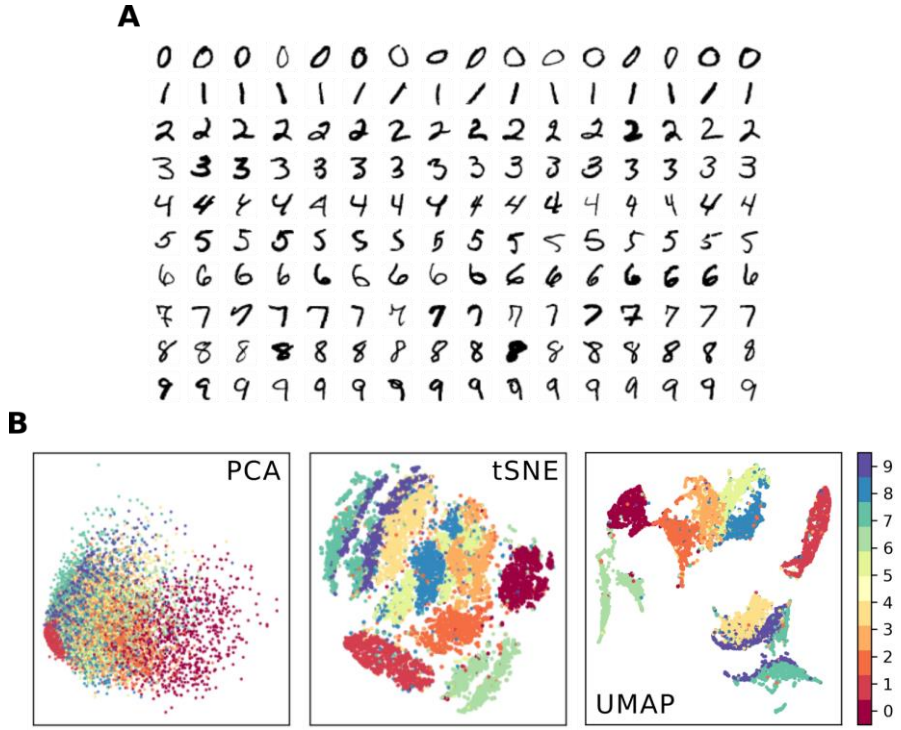

Figure 1.13: Comparing different data dimensionality reduction approaches. A) A subset of the handwritten digits in the NIST digits dataset and B) corresponding two-dimensional representations by PCA, T-SNE and UMAP embedding (own calculations).

1.5.3 Clustering

Whereas dimensionality reduction methods largely fulfil data visualization purposes, clustering algorithms are highly useful to discretize the conformation space based on the structure of the data. From a machine learning perspective, clustering belongs to unsupervised learning methods and stands in opposition to classification, which is mostly supervised. In the case of classification, one aims to find a mathematical function, which allows to separate *labeled* data points based on a set of features. This mathematical function is called a classifier and, if well trained, is able to sort a new, unlabeled points into the previously learned classed. Classification is one of the most central problems of machine learning and plenty approaches such a logistic regression, k-nearest-neighbor, decision tree, random forest and deep learning, have emerged and are heavily used [174, 175]. In traditional trajectory analysis of molecular conformations, the data is rather unlabeled. Thus, classifiers cannot be trained via supervised learning methods and unsupervised clustering techniques are preferred. It has to be noted that, in some exceptional cases, large databases of orthogonal information, e.g.experimental binding constants or solubilities are available and supervised learning is possible and recommended.

Anyway, analysis of the conformational space using clustering methods, also called conformational clustering, is a common and intuitive way to not only understand but also process molecular dynamics data in automatized workflows. It generally aims to identify groups of pairwise similar data points. Clustering can be combined with dimensionality reduction methods to help evaluation

of the clustering via visual inspection of the outcome on lower dimensional representation. To what extent clustering can be used in the embedded space is still a matter of debate. It appears intuitive to exploit the quality of the embedding to reduce the mathematical and computational requirements for the clustering algorithm. However, it is important to harmonize the applied metrics. That is, if the embedding preserves probability distribution and not distances, a distance-based clustering in the embedded space might be misleading. The metric lies at the heart of every clustering algorithm and drastically affects quality and performance. It is the mathematical function used to quantify how similar or how different two points in the feature space are. Differences are also called distances and are defined in the range of 0 (identical) to infinity (very different). A matrix containing the pairwise distances between all points in the set is called the distance matrix. A metric is also called a norm because the distance between two points is the absolute of the connecting vector, and the vector can be normalized to unit length by dividing it by its length. In a low dimensional Cartesian space, e.g. spanned by the two H-O bond distances of water molecule, the Euclidean distance metric, also called standard norm or 2-norm is an intuitive and often expedient choice. The n-dimensional 2-norm is defined as

$$d(i,j) = |x_i - x_j| = |\overrightarrow{v_{i,j}}| = \left(\sum_{k=1}^{n} (v_k)^2\right)^{\frac{1}{2}} \quad (24)$$

where d(i,j) is the distance between the two points x_i and x_j, $v_{i,j}$ is the connecting vector, and k indexes the different dimensions (features). Such a norm is a representative of the so-called p-norms, with p=2. It can be easily generalized by changing the value of the exponent p ($p \geq 1$). How the choice of p affects the outcome in higher dimensions was systematically investigated by Aggarwal et al. [176]. In their mathematical experiments, they draw points from an n-dimensional uniform distribution and calculate the distances to the closest and the farthest points from the origin. The normalized difference between such distances is measure for the contrast of the applied metric. Interestingly, with the Euclidean metric (2-norm), the contrast increases only initially with the number of dimensions. When the dimensions exceed of number of 10, the contrast only increases marginally. Thus, when points in e.g. a 20 dimensional space significantly differ in only a few of them, the Euclidean norm will not be able to distinguish between them. This observation has been coined "Curse of Dimensionality". Fortunately, the magnitude of this effect can be reduced by employing a different norm, e.g. the extreme cases of the p-norm: the 1-norm also known as Manhatten or Cityblock metric or the infinity-norm, which is also called maximum norm. Besides the p-norms, a broad range of different metrics exist which are optimized for their specific fields of application. For example, the Haversine metric, also called great circle distance allows the calculation of the shortest path between two points on the surface of sphere given their longitude

and latitude values. It was not yet accomplished to generalize the Haversine formula into higher dimensions. Other examples are distance metrics for Boolean vectors such as the Jaccard distance also known as Tanimoto score.

With the choice of an appropriate distance metric, the clustering can finally be performed. One of the most used and most intuitive clustering algorithm is *k-means* [177]. In contrast to other clustering algorithms, it operates with the actual data points and not only their pairwise distance matrix (**Figure 1.14 A**). The aim is to find a partitioning of the data points, so that the sum of the distances between points within a cluster and their mean is minimized. The number of clusters *k* is an input parameter for the clustering. The algorithm operates in three steps, of which 2 and 3 are repeated until convergence is reached:

1. Initialization: The first k "means" (seeds) are chosen
2. Assignment: All data points are assigned to their nearest mean
3. Update: The cluster means are re-calculated

Convergence to the global minimum is not guaranteed. Hence, in practice multiple attempts with different initializations are performed and the sums of intracluster-distances to the centroids are compared. Apparently, the initialization method is key to find the global minimum. In an assessment of Celebi et al. [178], they figured that the k-means++ variant performs generally well. This approach was suggested by David and Vassilvitskii [179] and exploits the data structure to place the seeds in well separated areas.

Another class of clustering algorithms is density-based clustering with one massively used representative being DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [180]. The number of clusters as well as the ratio of noise (unclustered data points) is determined by the algorithm based on the tuning parameters ε and min_samples (**Figure 1.14 B**). DBSCAN and derivatives are attractive for molecular dynamics trajectory data because the density is equivalent to the energy as described earlier. Thus, clustered conformations are likely to belong to the same energy minimum and not density-connected points are disconnected by high energy-barriers. Popular extensions to DBSCAN are OPTICS [181] and HDBSCAN [182]. In the original approach, the density at each point is estimated by the number of points within its ε-environment. This information is readily available from the pairwise distance matrix, independent of the utilized metric. When a point has at least min_samples points in its ε-neighborhood, it is considered a core point. Every point that has less than min_samples points within ε but is within ε of another point is a bordering point. Every other point is considered noise. Finally, a reachability graph is constructed through connecting core points within ε-environments by edges. Pairwise reachable points are clustered together. Boundary points are assigned to the cluster of their connected core

point. DBSCAN reliable recognizes clusters with different shapes and sizes and filters noisy points. It has however difficulties to distinguish clusters with different densities. The latter topic is partially addressed by hierarchical DBSCAN (HDBSCAN).

In general, hierarchical clustering is a family of distance-based cluster analysis, which can be further divided into agglomerative and divisive approaches. Both ways aim to build a hierarchy of clusters, traditionally visualized via dendrograms (**Figure 1.14 C**). The standard method is hierarchical agglomerative clustering. Here, each observation (point) starts in its own cluster. In each iteration, the most similar clusters are merged. The algorithms are often implemented recursively. In complete-linkage clustering, the distance between two clusters is the maximum of the distances between the points within the respective clusters. In single-linkage clustering, it is the minimum distance. Another popular linkage criterion is Ward's minimum variance method [183]. In Ward's method, the cluster distance is the sum of squared pairwise distances between the points of different clusters.

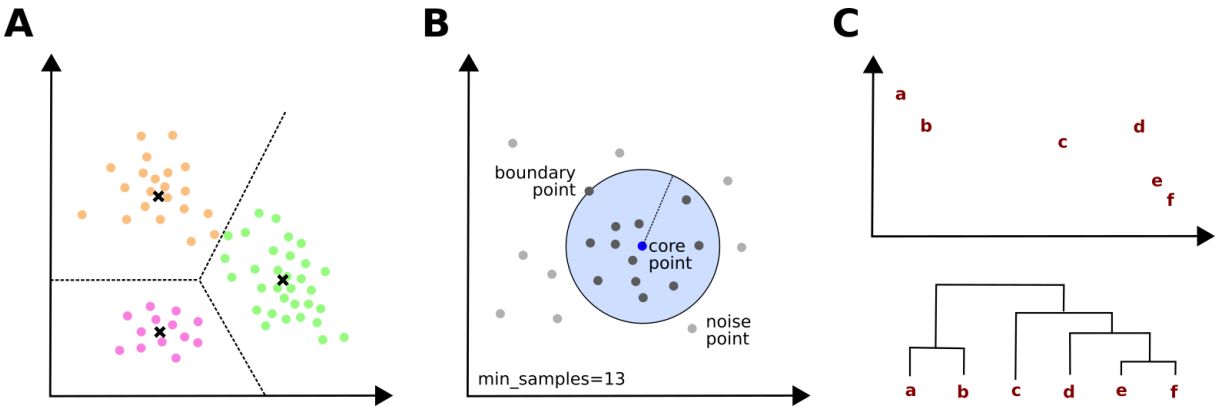

Figure 1.14: Schematic representation of different clustering algorithms. A) A point cloud is clustered via k-means (k=3) clustering. The assignment is chosen so that the distances of the points to their means (black x) is minimized. B) Visualization of a DBSCAN calculation for certain core point and a min_samples value of 13. C) Distribution of six points (a-f) and the corresponding dendrogram.

1.6 Physical characterization of thin molecular layers

Molecular layers, such as self-assembled monolayers (SAM) and or lipid bilayers must be considered separately from globular molecules or a bulk phase, due to their confined, almost planar geometry. These almost two-dimensional assemblies form a distinct, clearly bounded phase. Physical characterization included the description of the phase as a whole as well as the conformations and motions of the molecules within.

1.6.1 Global descriptors

The distance between the two outer boundaries (phase boundaries) is called the thickness or in case of SAMs also the height of the layer. Experimentally, the thickness of lipid bilayers is often measured via NMR spectroscopy [184] whereas the height of SAMs is rather investigated by ellipsometry [185] or neutron scattering [186]. Depending on the experimental method, the results may differ. In NMR and neutron scattering, the relative position of the heavy atom nuclei is analyzed, whereas ellipsometry rather recognizes the interfacial boundaries. Computationally, both observables can be read off from normal direction density profiles. The density profile is generated from a molecular dynamics trajectory by dividing the simulation box into a number of horizontal slices and determining the mass-weighted atom number distribution. Therefore, it is customary to center the layer at a certain z-coordinate to avoid smearing of the density peaks due to normal directed motions of the whole layer. The density profile of a lipid bilayer has two significant peaks, which correspond to the positions of the phosphate groups and mark the position of the head groups. The distance between these peaks is the thickness (PO4-PO4 thickness) of the lipid bilayer. Interestingly, the closer towards the center of the bilayer, the density decreases which is a results of decreasing order and increasing flexibility of lipid tails.

While the thickness describes the layer in normal direction, the lateral directions are characterized by the area per molecule, also known as molecular area or in the case of lipid bilayers lipid area. The lipid area is a measure of the areal density of the layer. In a molecular simulation, the molecular area can be obtained from the number of molecules in a layer divided by its lateral dimensions. This affords that the layer had enough time to sufficiently equilibrate in an NPT ensemble (flexible volume).

A third experimentally and computationally observable figure is the bond order parameter [187, 188]. It is usually investigated for the carbon-hydrogen bonds of the fatty acyl chains and indicates the degree of packing, or in other words, it quantifies if the molecules rather adopt liquid-crystalline (gel-like) conformations (order parameter is high) or if they are mostly disordered (order parameter low). The order parameters can accurately be measured via NMR and are geometrically defined as

$$S_{CH} = \frac{\langle 3\cos^2\varphi - 1\rangle}{2} \quad (25)$$

where φ denotes the angle between a bond vector and the layer normal. Triangular brackets represent time and ensemble average. In the case of a multi-molecular assembly, an ensemble is already present from the number of different molecules. Thus, in contrast to single-molecule simulations where the ensemble is generated via temporal sampling, here, an ensemble is generated by different instances of the same molecule plus temporal sampling. In some cases, it makes sense to distinguish between time and ensemble (molecule) averaging. Inspection of equation **25** is revealing. The angle φ range is between 0 and 180° (π). Hence, cos φ yields an almost linear projection of the angle to the range of [-1, 1]. The square raising, however, folds the negative values for φ>90° into the first quadrant and the distinction between values of below and above 90° is lost. Instead, the $\cos^2\varphi$ shows how far the angle is away from 90 degree on a scale of 0 to1. Finally, the parameter is mapped to the range of -0.5 to 1, where 1 indicates a bond-normal angle of 90°, 0.5 of 60°, 0 of 30° and -0.5 of 0°. The introduction of the order parameter and the consequent avoidance of directed angles only allows meaningful averaging. Thus, an average CH bond order parameter 0.3 and higher can only be achieved by a collectively similar conformation and orientation of many acyl chains, which corresponds to a gel like phase.

It has to be noted that, the three descriptors thickness, area and order parameters are physically interdependent. A layer of molecules with for example 16 carbon acyl chains can adopt different phases: a gel- or wax-like, so-called lo liquid ordered l_O phase or an amorphic, liquid disordered l_D phase depending on the degree of saturation. In the liquid ordered phase, all tails are in a similar, extended conformation. The order is high, the lipid area is small and the thickness is also high. In the liquid disorder phase, the acyl tails are literally melted and cover a larger area. The thickness and order parameters decrease. Excitingly, both phases can coexist in a multi component layer. A liquid ordered phase within a liquid disordered bilayer for example is called lipid raft and can in cells serve as a docking, recognition and signaling platform [189].

1.6.2 Spatiotemporally resolved description

In lipid bilayers, molecules underlie a slow lateral diffusion [190]. In contrast to a protein in water, the molecules of interest are solvated in their own kind and additionally constrained to the molecular layer. Interactions are mediated by head groups and the acyl chains. The acyl chain interactions are predominantly formed between saturated (more ordered) or between unsaturated (more disordered) chains and usually short-lived. Hence, a temporal average-based description must be employed for the quantification. One way is the spatiotemporal analysis of the bilayer properties. For example, it is possible to analysis the local density using a discretization-based

approach and elucidate the temporal averages over various lag times. Such an approach reveals transient, locally constrained, cholesterol-rich phases in model bilayers. When a distinct solute, such as a membrane protein peptide or protein, is of interest, radial or cylindrical distribution functions (cdf) might be revealing. They are constructed by counting representative atoms within spherical or cylindrical shells of increasing radii around the solute. For large radii, the cdf converges to the bulk (global layer) density. For small radii, i.e., first and second solvation shell, differences among the various molecules in the bilayer show preferential interaction partners.

In SAMs, lateral diffusion is severely limited. Hence, the spatial arrangement of mixed SAMs is pre-determined by interactions in the solvent phase before the adsorption stage, or through interactions during the adsorptions. Unfortunately, the SAM are mostly probed after adsorption and earlier effects are neglected. In mixed SAMs, in which one component is functionalized with large moieties such as acyl chain anchors connected via a polyethylene glycol (PEG), interesting intermolecular interactions can be indirectly observed with spectroscopic methods. For example, the degree of order is proportional to the concentration of the anchor-carrying components because of pairwise stabilizing interactions.

1.7 Structure and aims of the thesis

In this research thesis, the computational technique of molecular dynamics simulation is critically assessed for its capabilities to quantitatively disclose the molecular fundamentals of macroscopic observations. However, it was decided that the work should not be conducted in a way of a benchmark study but as a set of applied molecular dynamics studies on recent, practical problems in the fields of biology and chemistry. Therefore, we collaborated with different labs who were interested in specific questions and in return shared their expertise and experimental insight. This way, we could study the MD method and its limitations, develop novel, integrated MD workflows, and thereby also perform fundamental biology and chemistry research. This work is truly interdisciplinary, as it combines theories from the fields of numerical mathematics, computational physics, machine-learning (multivariate statistics), programming, physical chemistry, and biochemistry. The interpretation of the results additionally affords knowledge in the fields of spectroscopy, preparative chemistry and systems biology. It is apparent, that in such an interdisciplinary framework, not all theories can be studied and described in extensive detail and some methods herein are applied and interpreted without a complete description and discussion of their physical or mathematical foundations.

We have applied MD simulation to three classes of molecular systems, namely bimolecular, unimolecular, and multimolecular systems. The uni- and biomolecular systems can be considered

of biological interest, whereas the multimolecular systems are rather chemistry-oriented. The three classes make three blocks within the thesis, in which similar modeling and analysis methods are employed. However, each individual problem has its own chapter because it corresponds to a manuscript which has either been already published or is still in the final stages of preparation.

The two bimolecular system chapters deal with the intermolecular interactions between a protein receptor with small molecule ligand and protein receptor with a substrate protein. The former receptor is the norovirus protruding domain which interacts weakly with putative allosteric modulators that a present in the human gastrointestinal system: bile acids. The binding mode of the bile acids was not a priory known and was predicted using an ensemble docking workflow. Based on the prediction, the protein-ligand interactions were studied using MD simulation and compared to NMR chemical shift perturbation experiments. The results shed light on the bile acid recognition of human norovirus and deepen the understanding of the importance of conformational selection for docking approaches. In the latter system, selective binding of a diubiquitin protein substrate to a either a human or a bacterial isopeptidase enzyme, crystal structures of the complexes were available. MD simulations of the enzymes with and without the bound substrate proteins were performed and analyzed in unprecedented detail. The effect of substrate binding on the competence of the catalytic triad was assessed as well as the composition of protein-protein interactions at the two major binding surfaces. The results were contrasting some experimental conclusions and yielded the insight that an intentionally induced mutation in the substrate on the bacterial protein complex has led to a misleading interpretation of the crystal structure.

The unimolecular systems are comprised of a single protein molecule within a solvent phase. One of these proteins is heavily-glycosylated human growth factor erythropoietin (EPO). Here, 16 models of EPO were generated, each with a different glycosylation pattern. The mutual interactions between the glycosylation and protein were studied with the aim to understand how glycosylation alters the proteins biophysical properties and how the site of the glycosylation selectively affects the local protein and glycan conformational spaces. The other protein is the protruding domain (P-domain) of the norovirus VP1 capsid protein. It was used to study the posttranslational modification of deamidation (auto-reaction of asparagine to aspartate or glutamine to glutamate), which is a key driver of protein degradation and reduces shelf lives of therapeutic proteins. Here, the P-domain is an exquisite model because only one residue selectively undergoes deamidation with especially high rate. We use extensive MD simulations and advanced statistical and geometrical analysis to study the conformational spaces of the various deamination candidates to allow a first dynamics-based deamidation prediction model.

In the third block, multimolecular, layered systems are studied. The initial, short chapter serves as a transition from biological to chemical systems and introduces the concept of coarse-graining and the MARTINI force field. It deals with the preferred interactions of membrane-bound peptide with the lipid molecules in the surrounding bilayer. The second part shows how the methods of coarse-grained bilayer and membrane anchor modeling were transferred to a new class of systems: self-assembled monolayers (SAMs). We have developed a novel and complete modeling, simulation and analysis workflow for SAMs based on coarse-grained representation, a hexagonal packing and a flat surface model. This automatized protocol allowed rapid computational screening of the variety of mixed SAMs and predicted physical properties with high accuracy. After resolution transformation to full-atomistic detail, analysis of intramolecular interactions and resulting changes in conformation and orientation rationalized molecule-specific differences in infrared reflection adsorption spectra. The simulations together with experimental spectra led to the conclusion that certain SAM components aggregate either in the solvent phase or during the adsorption stage, which leads to highly non-isotropic SAMs. In the final chapter, the advantages of the novel SAM model are utilized to investigate interactions between such mixed SAMs and adsorbing lipid vesicles. The interactions occur during the preparation of tethered bilayer membranes, when lipid vesicles are titrated onto SAMs. At a certain vesicle concentration and with a correct SAM composition, the vesicles rupture and the released lipids aggregate into a bilayer. How the SAM composition affects the critical vesicle concentration is unclear. Steered, coarse-grained MD simulations were carried to initiate interactions of vesicle and SAM. In subsequent equilibrium simulations, the SAM-vesicle contacts were classified and quantified. Together with quartz crystal microbalance studies, the results indicate that SAM components with a proficiency to form

The thesis is structured in a way that the individual research chapters are framed by statements on how they fit into the main topic of molecular recognition and comments on the accuracy and feasibility of the molecular dynamics method for the particular problems. The global aim of the thesis is to develop a generalized, dynamics-based perspective on selective molecular interactions and their effect on the macroscopically observable behavior.

2 NOROVIRUS RECOGNITION OF BILE ACIDS

In this chapter, the binding of bile acid molecules to a norovirus capsid protein is investigated using molecular dynamics, docking, and NMR spectroscopy. The NMR experiments were performed by collaborators at the University of Lübeck. The developed workflow may be of general interest for the estimation of binding modes of low affinity binders. The chapter is published in [191].

2.1 Introduction

Norovirus infections are among the leading causes of infectious gastroenteritis [192, 193] and pose a major health threat for immunocompromised patients [194] and people in developing countries [195]. To enable infection, it is essential that the protruding domain (P-domain) of the norovirus capsid protein VP1 attach to presented histo-blood group antigens (HBGAs) of the host cells [196-198]. Hence, the HBGA binding site has been employed as a target for the development of potential virus entry inhibitors [199-201], however with no clinical significance. Interestingly, recent successes in the development of norovirus cell culture systems [202, 203] yielded the insight, that besides HBGAs, bile acids function as important infection-promoting cofactors [204, 205].

For particular norovirus strains, bile acid bindings sites with micromolar affinity have been identified crystallographically [204, 206] and localized in proximity of the HBGA binding sites. Here it is to mention, that the VP1 protein appears as a dimer. Thus, the dimer carries two symmetrical pockets for HBGAs and bile acids. To our surprise, the dominant human disease-causing norovirus strains GII.4 and GII.17 do not display bile acid binding to these pockets. Yet still, bile acids are essential (GII.17) or at least promotive (GII.4) for norovirus infection. In this chapter, microsecond scale molecular dynamics (MD) are employed in combination with NMR spectroscopy by collaboration partners to identify a novel bile acid binding site of GII.4 and GII.17 P-domains and to predict the ligand binding mode. Such MD simulation can reveal transient ligand binding pockets [207], which are also referred to as sub-pockets, adjacent pockets, channel/tunnel, or allosteric pockets [208]. Small molecule ligands (such as bile acids) may selectively bind to one or to an ensemble of such pre-existing conformations [135]. The theoretical methods were guided and restrained by NMR experiments, which are described in the following section.

The selective, yet low affinity binding bile acids by the norovirus P-domain dimer is an archetypical problem in the area of early-drug discovery, in which a drug candidate was identified through a screening experiment. This so-called hit would then be optimized using medicinal-chemistry or structure based approaches or a combination thereof. The aim of hit-to-lead optimization is to establish and to strengthen specific protein-ligand interactions by the introduction or alteration of chemical groups [209]. The chemical effort can be lowered by structure-guided restraints [210].

However, co-crystal structures of the protein-ligand complex at this low-affinity binding stage are often impossible to generate[211]. Here, we show an NMR-driven ensemble-docking workflow as an alternative approach to estimate both binding site and mode of a low-affinity ligand.

2.1.1 NMR spectroscopy

Three classes of NMR experiments were conducted by R. Creutznacher to investigate bile acid binding to Norovirus P-dimers. 1H15N TROSY HSQC and methyl TROSY chemical shift perturbation (CSP) experiments were performed to identify the bile acid binding region (**Figure 2.1**). Saturation-Transfer Difference (STD) NMR allowed the estimation of a binding epitope of cholic acid and NMR titration provides binding affinities for the various cholic acid species (**Figure 2.2**).

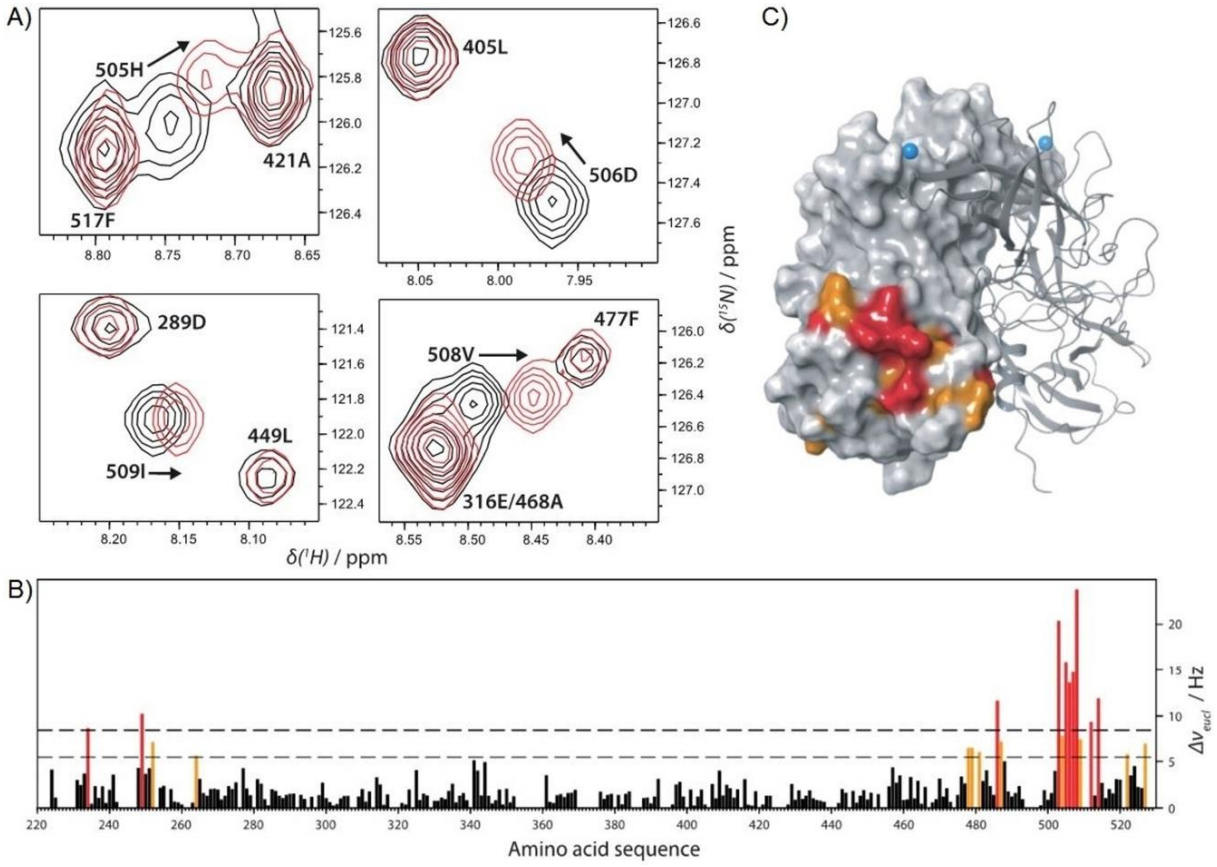

Figure 2.1: Chemical shift perturbation upon cholic acid binding to Norovirus P-dimer. A) 2D NMR spectra of selected residues in bound (red) and unbound (gray) state. B) Euclidean shifts of each residue. Orange bars are larger than the mean plus the standard deviation. Red bars are larger than the mean plus two times the standard deviation. C) Surface model of Saga P-Dimer with residues colored according to the significance of chemical shifts. Colors identical to panel B.

CSP NMR experiments allow the observation of the change in the NMR spectra of especially proteins upon complexation e.g. with a small molecule ligand. As the herein recorded spectra are two-dimensional (2H-15N and 2H-13C), the chemical shift perturbation is calculated as the Euclidean distance between the peaks. Usually, many of the NMR spectral peaks undergo a slight change in such experiments. Therefore, the significant CSPs are selected as the ones which are

larger than the average plus the standard deviation of all measured CSPs [212]. The CSP measurements require near-complete NMR peak-to-residue assignments of backbone and methyl groups. For the herein considered P-domains of the GII.4 Saga strain, this was previously achieved by Mallagaray et al [213]. In presence of 8 mM CA, the chemical shifts of the P-dimer are perturbed in a distinct region of the protein, which is on the opposite of the regular substrate binding pocket. The highest significant backbone TROSY CSPs are reported for Val508, Gly503, His505, Leu507, Asp506, Tyr514, Leu486, Leu249, Asn512, and Val234 (in order of decreasing CSP). Their CSPs are larger than the mean plus two standard deviations. For, Gln504, Ile509, Phe487, Gly252, Leu527, Asn479, Val478, Asp481, Asn522 and Gly264, the CSP were also significantly high, yet only higher than the mean plus one standard deviation. Most of the high-CSP residues form a large, continues patch on the protein surface and only Val234, Leu249, Gyl252 and Gly264 are in remote positions. In addition to the high backbone CSPs, Val478, Leu486 and Leu507 show also significantly high methyl TROSY CSPs. The c-terminal residues Leu527, Ala528, and Met530 have high CSPs exclusively in the methyl TROSY spectra.

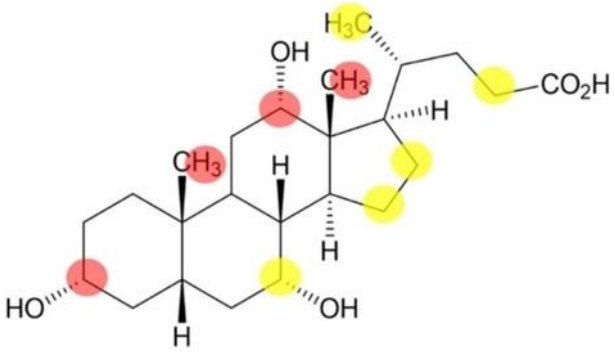

Figure 2.2: Epitope mapping of cholic acid. The red atoms receive high saturation (80-100%) and become most buried. Yellow atoms correspond to 60-80% received saturation.

On the bile acid site, the binding can be explored by STD NMR. The highest saturation is achieved at the hydrogens of C3, C12, C18 and C19 which belong to the more hydrophobic site. However, also hydrogens at C7, C16, C16, C21, and C23 on the opposite site receive saturation from the protein. Additionally, dissociation constants K_D of CA and GCDCA were determined using both CSP NMR and STD NMR titration. The K_D of CA against GII.4 Saga P-dimers is in the order of 10 mM, and around 4 mM against GII.4 virus like particles (VLPs). The affinity against VLPs and P dimers of other Norovirus strains is similar or lower with K_D values ranging from 6 to 32 mM. GCDCA has slightly higher affinity to GII.4 Saga P dimers (1.5 mM).

2.2 Method

2.2.1 Model generation

Cholic acid was modeled using CHARMM-GUI ligand reader [78] with a coordinates from RCSB ligand expo [214]. The parameters were generated using CGenFF [87] as implemented in CHARMM-GUI. Cholic acid is structurally related to cholesterol, which is heavily studied in bilayer simulations as frequently performed using the CHARMM lipid force field [84, 215]. Thus, parameters for cholic acid were considered sound and reliable.

The protein models were generated using the CHARMM-GUI PDB reader [216]. For the GII.4 Saga strain, the crystal structure from PDB-ID 4OOX [217] was used. Histidine residues 292, 347, 417, 460 and 501 were protonated at Ne position. Histidine residues 378, 396, 414, 490 employed the standard scheme: protonation at the delta nitrogen. Histidine 505 was double protonated and thus charged. Both subunits, i.e., the whole dimer was modeled. Both the termini were charged (R-NH3+, R-COO-). Using the CHARMM-GUI quick MD simulator [218], the cubic simulation box was generated with at least 2 nm space in every direction from the protein. The box was filled with TIP3P water [219, 220] and ionized to 0.15 M NaCl using random ion placement. Protein topology files including CHARMM36 parameters were generated by CHARMM-GUI. To investigate possible further binding hotspots for cholic acid, five instances of the molecule were added to the simulation box. They did not undergo stable interactions with the protein or themselves and were thus neglected.

The docking scores and poses were computed with AutoDock Vina [129] (ver. 1.12). Receptor (GII.4 P-dimer) and ligands (bile acids CA, DCA, CDCA, and GCDCA, **Figure 2.3**) were prepared for docking with AutoDock tools [128]. The protein was kept rigid, whereas all rotatable bonds of the bile acids were flexible. Gasteiger partial charges [221] were assigned to receptor and ligand. The cubic search space was set up to encompass all perturbed amino acids in the binding region (**Figure 2.4 D**). The search was performed with an exhaustiveness of 64. Based on the generated topologies for the P-dimer and cholic acid by CHARMM-GUI, the complex systems were solvated with TIP3P water, and ionized to 0.15 M NaCl using GROMACS tools ver. 5.1.5. While the whole complex was simulated, the bile acid was only bound to one monomeric subunit. The box size was identical to the initial protein conformational sampling.

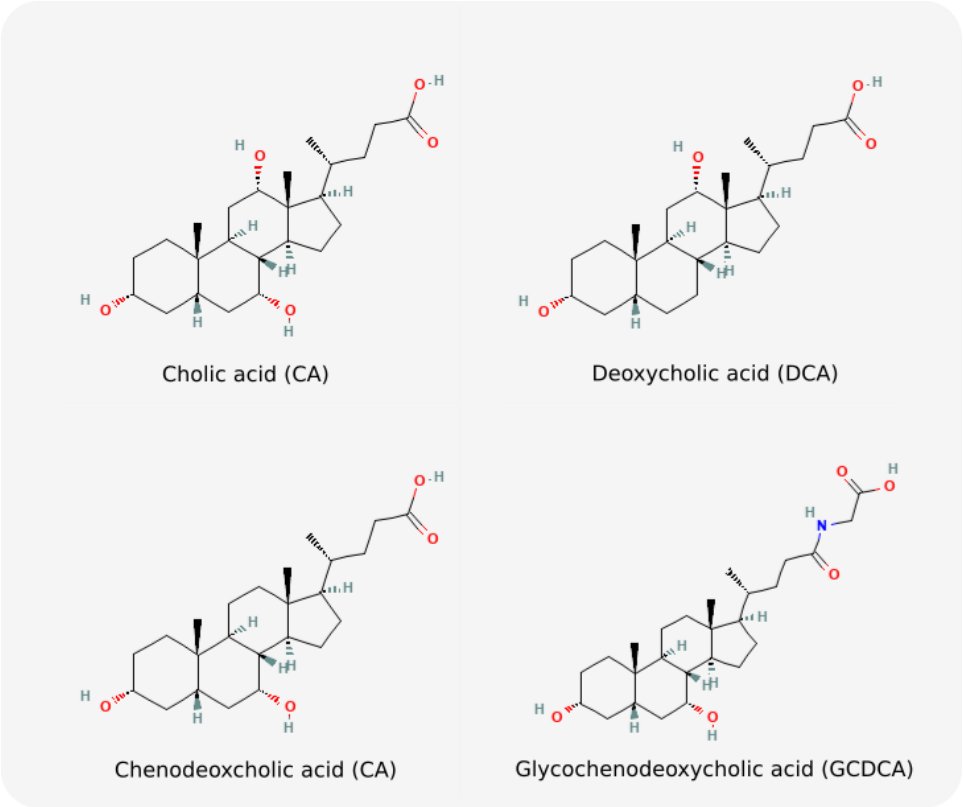

Figure 2.3: Chemical structures of the herein studied bile acid species. The structures differ in the hydroxylation pattern and the tail group.

2.2.2 Simulation protocol

The simulations in this chapter employed the CHARMM36 force field [83]. Verlet cutoff scheme [222] was employed with neighbor list updates every 20 steps. The short-range neighbor list cutoff was 1.2 nm. Coulomb interactions were treated with the Particle Mesh Ewald method [98, 99, 223, 224], where short range interactions were cutoff after 1.2 nm. Van-der-Waals interaction were cutoff via a force-switch modifier between 1.0 and 1.2 nm. The conformational sampling of the protein was achieved by an initial minimization for 5.000 steps using the steepest descent algorithm, followed by 100 000 steps NVT and 100 000 steps in NPT equilibration (time steps 0.01 and 0.02 ps). Production sampling was generated for 1 μs using a 0.02 ps time step. Temperature coupling was achieved with the Nosé–Hoover method [225] (target temperature of 303.15 K, coupling constant of 0.4 ps during equilibration and 2.0 ps during production). Protein (plus the bile acid ligand, if present) and solvent (including water and ions) were coupled individually. The initial temperature distributions were generated according to a Maxwell–Boltzmann distribution at 293.15 K. Pressure coupling was achieved by the Berendsen barostat [93] during equilibration and by Parrinello–Rahman [226] during production (both using a coupling constant of 2 ps and a reference pressure of 1 bar). Protein and ligand heavy atoms were restrained during equilibration. Hydrogen bonds were constrained, with constraints solved by LINCS [227, 228].

2.2.3 Trajectory analysis

Backbone RMSDs and RMSFs were calculated using GROMACS tools *gmx rms* and *gmx rmsf*. Translational and rotational displacements of the entire complex were removed by fitting the trajectory to the crystal structure. The differences in pocket volume was calculated with POVME 3.0 [229]. The search space for cavities was set up manually as six spheres with varying radii to fully encompass the expected binding cavity. For the pocket volume and docking calculation, the trajectory was aligned using only the significantly perturbed residues (**Figure 2.4 C-D**).

Based on the MD simulation of the cholic acid – P-domain complex, contact analysis was performed with MDTraj [230]. For each frame, for each amino acid, a contact was counted if at least one heavy atom of CA was in proximity of 0.6 nm or less to its backbone nitrogen. The contact occupancy of an amino acid is the number of counted contacts dived by the number of frames. The stability of the complex was assessed using the relative RMSD of cholic acid to the P-dimer. Therefore, the complex was aligned by the P-dimer backbone coordinates and the RMSD of the cholic acid was computed. This RMSD included translational and rotational motions. The average of the last 10 ns were considered to make a decision on stability.

2.3 Results

2.3.1 Molecular dynamics sampling of the P-dimers

The crystal structures of the GII.4 Saga P-Dimers [217] do not exhibit accessible binding pockets in the suggested bile-acid binding region. Thus, 1 μs molecular dynamics sampling of the P-dimers was conducted. The protein and binding site dynamics were monitored by means of backbone root-mean-square deviation (RMSD) and fluctuation (RMSF) relative to the crystal structure Additionally, the relative change in binding cavity volume was calculated for every time step. Furthermore, molecular docking of cholic acid toward conformational snapshots of the P-dimers throughout the dynamics simulation was performed, and the best docking scores were monitored (**Figure 2.4 A-B**).

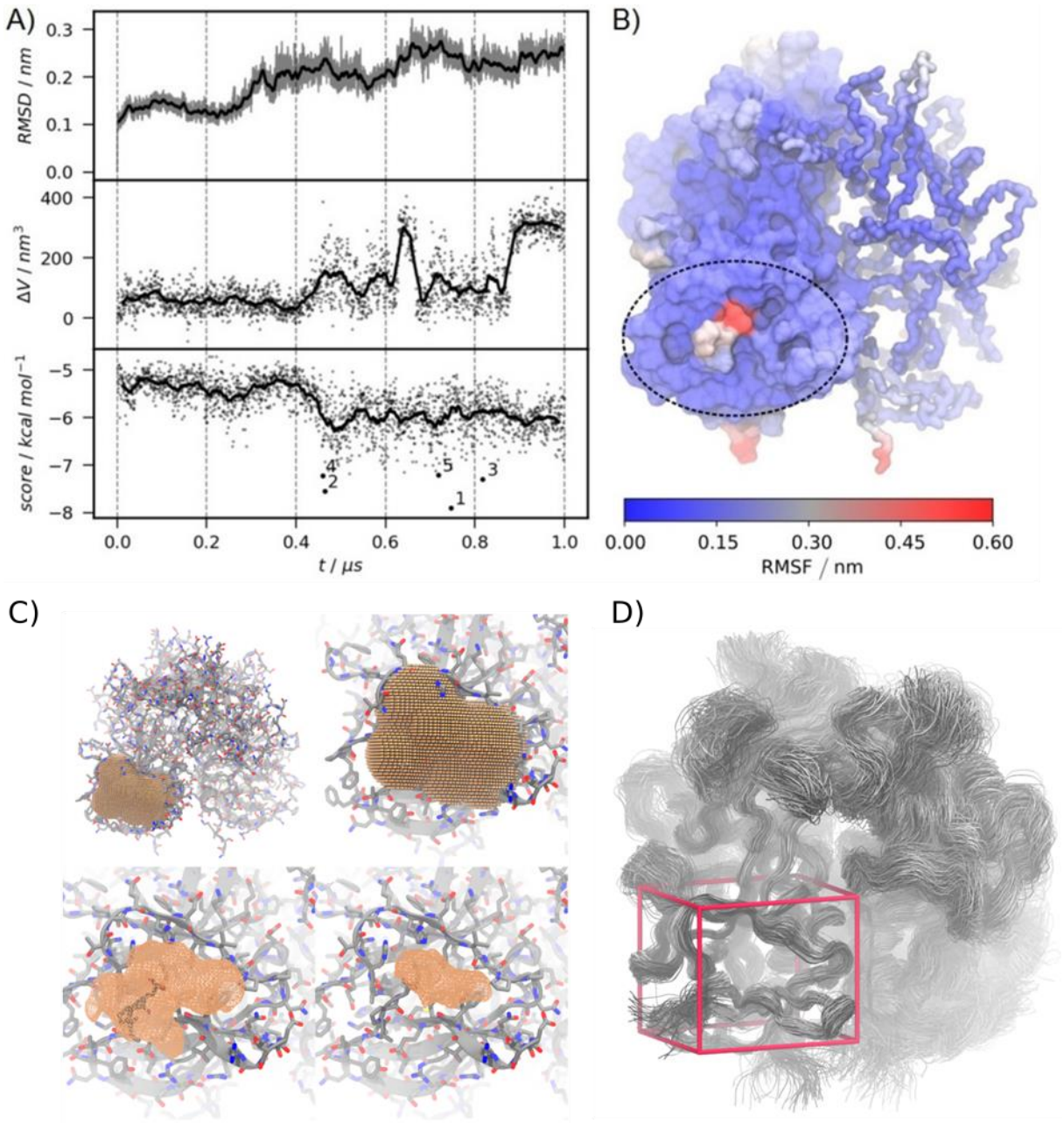

Figure 2.4: Molecular dynamics sampling and pocket measuring of the Saga P-Dimer. A) RMSD, relative pocket volume and docking scores during the trajectory. Thin lines and scatter points represent the determined values at 0.1 ns (RMSD) and 1 ns (ΔV, score) sampling rate. The thick lines are the moving averages (N=50). B) Residue-wise RMSF mapped onto the surface (left) and backbone (right) of the protein. C) Pocket search space (top) and occupancy isosurfaces for 0.5 and 0.9 pocket occupancies (bottom). D) Protein alignment to the significantly high CSP residues and docking search space.

The RMSD in the first 200 ns is as low as 0.11 nm and then slowly increases to 0.2 nm during 300 and 400 ns. The value of 0.2 nm is stable until 600 ns, after which the RMSD further increases to 0.25 nm. The changes in the RMSD are low and show the absence of large structural rearrangements concerning the whole protein. However, the small increases in the RMSD are rather stepwise. Thus, the RMSD reflects small and localized conformational transitions. The RMSF is used to localize conformational flexibility. Most of the protein is rigid with RMSF values of 0.1 nm smaller. Typically, higher flexibility is localized in the loop regions and the termini.

Interestingly, though, the C-terminus is a major part of the suggested bile acid binding region. Upon alignment of the trajectory along the high CSP residues, the conformational space of the binding site becomes even more apparent. The protein backbone forms a large elliptic grasp with an open center. However, the center is partially blocked by the C-terminus. The outer scaffold is rather rigid and underlies only marginal fluctuations. The C-terminus on the other hand is highly flexible can adopt multiple conformations – both bound to the protein and mostly solvated.

These conformational dynamics are well reflected by the relative change in cavity volume. It must be noted that the relative cavity volume is considered, because an absolute volume for such a shallow region is difficult to define [231]. Thus, changes in the volume of an arbitrary search space, defined as a set of spheres, are used instead [229] (see Method). The methodology allows not only the quantification of the pocket size but only detection of transient cavities, which are only rarely accessible. In the first 400 ns of the trajectory, it increased by 100 nm^3 relative to the crystal structure. After the first conformational transition, it increases even further and deviates 200 nm^2 from the initial value. During and after the second conformational change, the relative cavity volume increase can even reach values of up to 400 nm^2.

2.3.2 Ensemble docking of bile acids

To assess the quality of the sampled pocket conformations to accommodate bile acid molecules, they were utilized in small-molecule docking. Each CA, DCA, GDCA and GCDCA were docked against each conformation of the sampled PP-dimer conformational ensemble. The bile acids were fully flexible, whereas the protein was rigid. Such an approach can be advantageous over flexible protein docking, when enough receptor conformations can be sampled [232]. Each of the best docking scores throughout the trajectory were averaged among the four bile acid molecules and monitored. Initially, when the pocket volume is small and the protein conformation is close the crystal structure, the scores are in the range of -5.5 kcal/mol. After 400 ns, however, they drop to -6.5 kcal/mol in average and can even achieve scores of -7 kcal/mol and lower. Interestingly, the best scores are not achieved when the pocket volume is maximum, but when it is around 200 nm^2 larger than in the crystal structure. Such a volume is realized by conformations in which the C-terminus has moved away from the center of the pocket and towards the bottom. The five best average docking scores are subjected to further analysis by visual inspection of the binding poses.

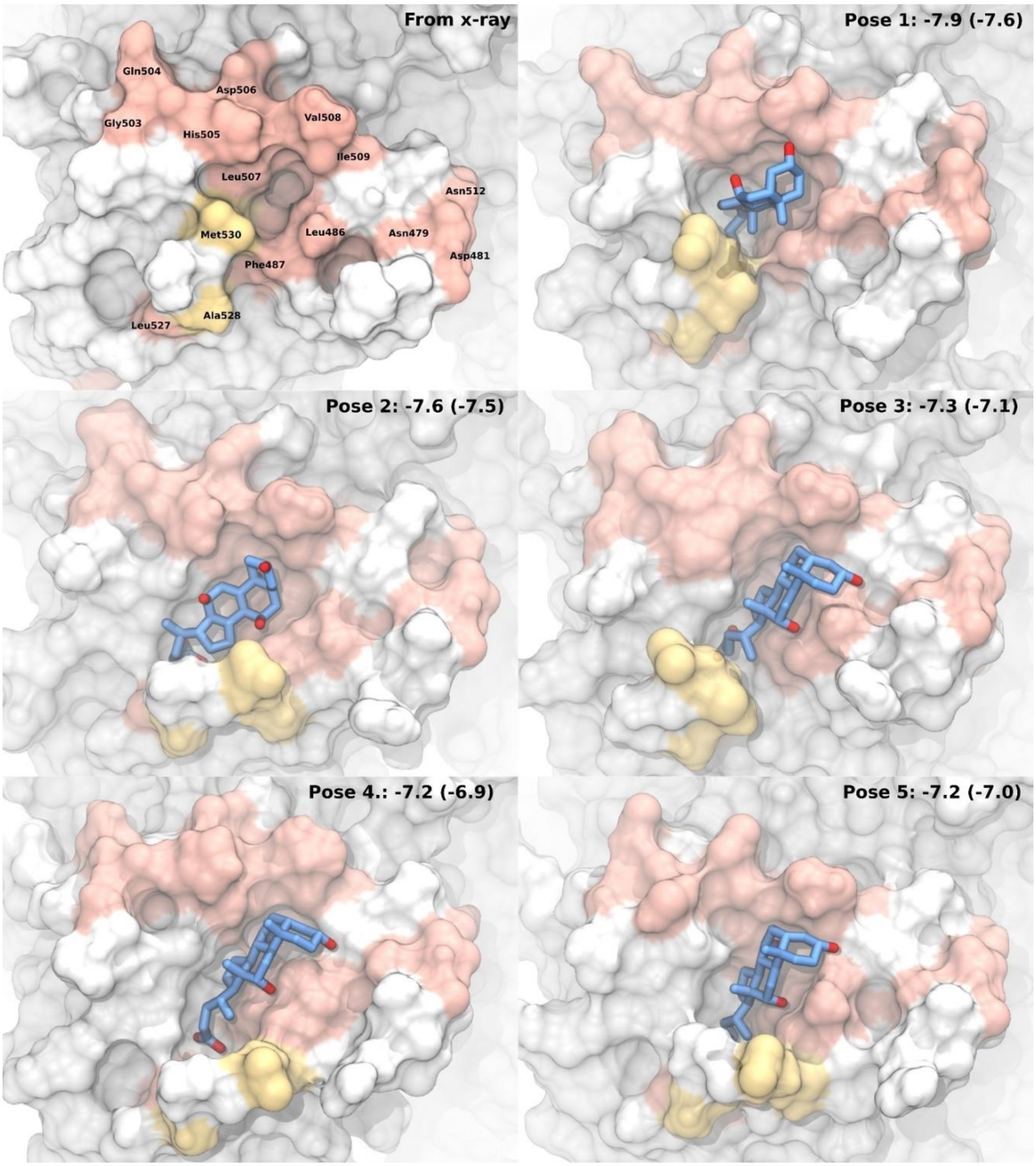

Figure 2.5: P-domain bile acid binding site. X-ray structure (in absence of small molecules) and the five top-scoring poses resulting from dynamic docking of CA to MD snapshots of P dimers. The protein surface is color coded according to experimental CSPs. Backbone CSPs larger than $\mu+2\sigma$ are shown in pale red. CSPs larger than $\mu+\sigma$ from methyl TROSY experiments are yellow. CA is shown in blue with oxygen atoms highlighted in red and hydrogen atoms omitted for clarity. The numbers represent the average of all bile acid docking scores with the CA docking score in brackets (kcal mol^{-1}).

The docking poses of four molecules are broadly identical with only few exceptions (**Figure 2.5**). Thus, only the predicted binding modes of CA are discussed in detail. All poses have in common that the hydrophobic cavity around Leu507 is accessible due to the C-terminal residues being moved to the side or the bottom. This allows deep burial of the carboxylate group of the bile acid to form interaction with Leu527 and neighboring residues. Otherwise, the poses mostly differ by

their rotation around the sterol backbone. Here it is recapitulated that the sterol backbone is of rather flat geometry and structurally rigid. It has perpendicularly oriented methyl groups on the one face, and a number of hydroxyl groups on the opposite face. The number of hydroxyl groups determines the type of bile acid and its solubility. The hydroxyl group at C3 is common among all bile acids. Cholic acid, for instance, has two additional hydroxyl groups: one at C7 and one at C12. In pose 1, these hydroxyl groups are oriented towards top residues His505, Leu506 and Val508. In pose 2, these groups are facing in direction of the solvent. Pose 3 to 5 have the hydroxyl groups oriented towards Leu486 and Phe487. The docking score only included non-bonded interactions and not conformational distortion of the receptor protein. Thus, a conformation which allows a perfect fit for the ligand molecule, might be energetically unfavorable. The balance between ligand binding energy and energy of the conformational change determines the true quality of binding. Unfortunately, the energy of the conformational change is computationally difficult to assess. Instead, the docking will be further evaluated by molecular dynamics simulations of the protein – bile acid complexes.

2.3.3 Resampling of protein-ligand complex dynamics

The complexes of the top five poses were each utilized as initial configurations for ten replicates of 20 ns explicit solvent molecular dynamics simulations. The stability of the docking pose was assessed by means of ligand RMSDs (**Figure 2.6**). The ligand RMSD is computed as the RMSD of the ligand upon least-square fitting of the receptor-ligand complex by the CA atom of the receptor protein. It thus explicitly includes translational and rotational motions of the ligand relative to the protein. The average ligand RMSD of the last 10 ns is used to filter out stabilized trajectories. Depending on the pose as well as on the initial, randomly generated velocity distribution, the bile acid ligand may remain stable at the initial position (RMSD<0.3nm), it may rearrange and then stabilize (RMSD < 1nm) or it may dissociate (RMSD > 1 nm). For pose 1, ligand RMSDs of 0.9 to 7.2 nm are reported and only one trajectory replicate (Rep. 9) stabilized at a RMSD of 1 nm or lower. Starting from pose 2, replicate 1 and 5 show RMSD of lower than 1 nm. The other trajectories have RMSD of either around 2 nm or 5 nm. For pose 3, the RMSDs arrange between 0.8 and 5 nm. Rep. 5 is the only one replicate to converge with a RMSD below 1 nm. The RMSDs of the trajectory with pose 4 as initial configuration range between 0.5 and 4 nm. Three replicates remain RMSDs lower than 1 nm (Rep. 3, Rep. 4., and Rep. 8). Pose 5 yields RMSD between 0.2 and 6 nm. The lowest RMSD replicate (Rep. 2) has the overall lowest value of 0.2 nm.

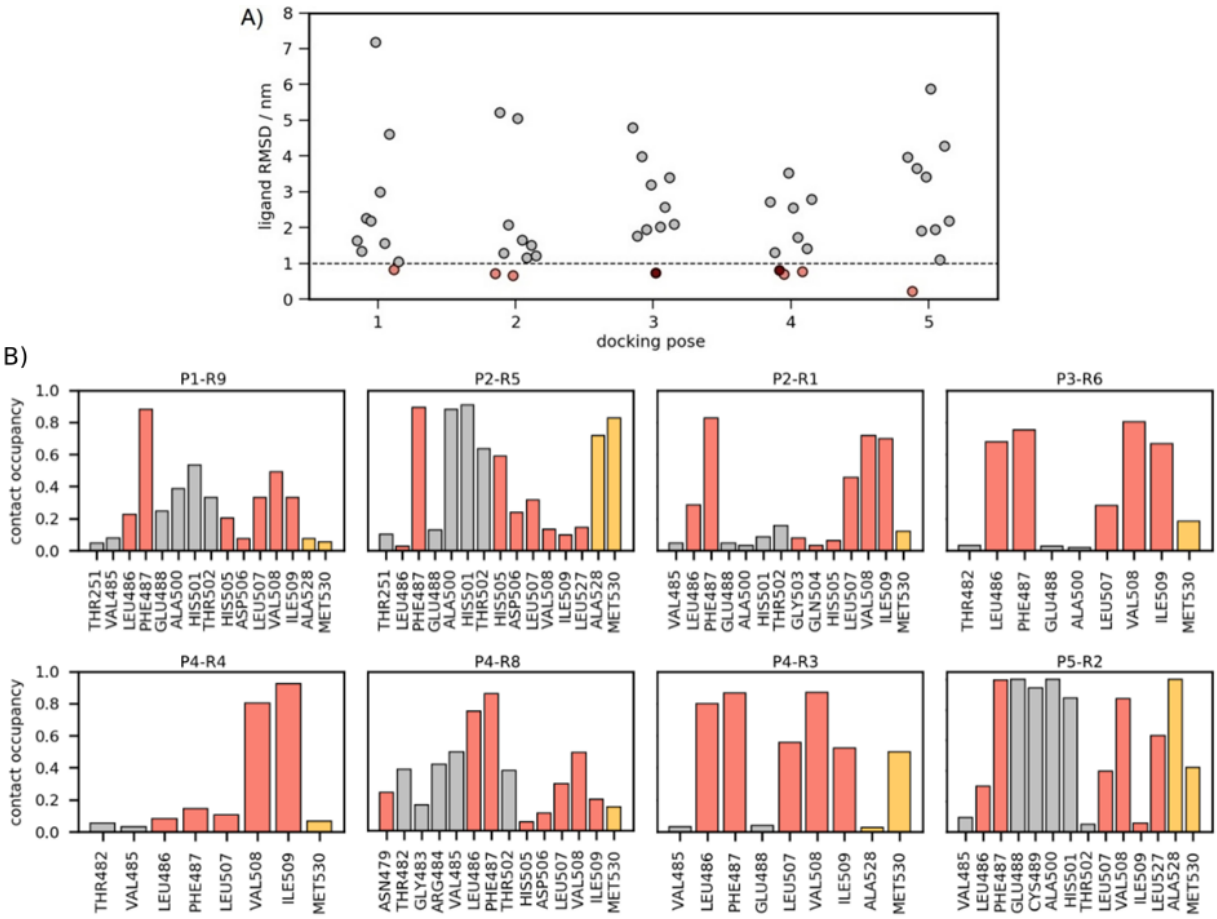

Figure 2.6: Outcome of refinement simulations. A) MD refinement of docked ligand poses of CA. Each point represents the mean RMSD of the final 10 ns of a 20 ns MD trajectory using the corresponding docking pose as initial coordinates. For each pose 1–5 ten independent simulations with varying initial velocity distributions were performed. Trajectories with RMSD<1 nm are highlighted as orange or dark red filled circles and were further analyzed for CA–protein contacts. B) CA–protein contacts for the red filled circle trajectories are shown in the bottom panels in detail. Contact criterion is a distance ≤0.6 nm between the backbone N and at least one heavy atom of CA. Contact amino acids that exhibit significant CSPs are highlighted in red (backbone HSQC) and gold (methyl TROSY), respectively. Proline residues are not considered as they show no NMR signals. Only amino acids with an occupancy >0.02 are shown.

The stabilized trajectories are subjected to a detailed contact analysis. Therefore, the inter-residue distances between the bile acid and the protein are calculated throughout the trajectory by means of minimum pairwise heavy-atom distances. The ratio of frames in which the inter-residue distances is below a threshold of 0.6 nm is considered the contact occupancy of this residue pair. In Pose 1 Rep. 9, cholic acid maintains only one persistent interaction with Phe487. It additionally forms weaker, and thus more short-lived contacts with a patch among Ala500, His501 and Thr502, and the residues Leu507, Val508, and Ile509. The cumulated number of interactions is low. The interactions in Pose 2 Rep. 5 are more numerous and stronger. Here, here cholic acid engages in highly persistent interaction with Phe487, Ala500, His501 and the c-terminal residues Ala528 and Met530. Considerable interactions are also observed between Thr502 and His505. Pose 2 Rep 1 shows a similar interaction pattern as Pose 1 Rep 9, however the interactions around His501 are weaker and the around Val508 stronger. In Pose 3 Rep 6, the interaction pattern is similar yet more

distinct with few but persistent interaction with Leu486/Phe487 and Val508/Ile509. The interactions around His501 are negligible. In Pose 4 Rep, cholic acid only interacts persistently with Val508 and Ile509. Pose 4 Rep8 is the first trajectory in which considerable interactions with Thr482, Arg484, and Val485 are reported. Yet, the dominating binding interactions are mediated by Leu 486/Phe487 and they are supported by many weaker interactions around Val508. Pose 4 Rep 3 yields distinct and stable interactions with Leu486/Phe487 and the trio of Leu507/Val508/Ile509. Additionally, Met530 exhibits a contact occupancy of 0.5. Pose 5 Rep 2, which was the most stable trajectory, has a unique interaction pattern. It bonds tightly with Phe487, Glu488, Cys489, Ala500, and His501. Additionally, it is involved in strong interactions with Val508, Leu527 and Ala528.

2.3.4 Binding mode decision by MD and NMR

Bile acids bind the norovirus P-domain only weakly with low affinities in the millimolar range. Such a binding must be considered transient and ambiguous. This is additionally reflected by the large spatial distribution of residues with significantly high CSPs as well as the general instability of the complexes in the MD. Thus, it must be assumed that bile acid binding is not limited to a single binding mode but an unspecific, flexible binding. Furthermore, it is even possible that bile acids bind not only as single molecules but as disordered aggregates or micelles. Based on the MD, a picture arises in which the bile acids compete with the c-terminus for the hydrophobic binding site center which consists mostly of the residue pairs Leu507/Val508 and Leu486/Phe487. When the C-terminus makes room to accommodate the bulky bile acid molecule, the binding is mostly driven by hydrophobic interactions with either or both these residue pairs. This is consistent in all the stabilized MD trajectories. Additional interactions might occur with the C-terminal residues Ala528 and Met530. Such interactions are interesting because they suggest that the C-terminus may cover the tail-like portion of the bile acid molecule to bury some polar interactions and shield them from the solvent. The dominating interactions from the complex MD simulation are in excellent agreement with the measured CSPs. Especially the conformations sampled by Pose 3 Rep 6 and Pose 4 Rep 3 yield contacts exclusively with residues whose chemical shifts are significantly perturbed upon binding (**Figure 2.7**). Thus, they are considered the predominating binding modes. They are distinguished by a tight binding of the methyl groups into the pocket and the hydroxyl groups facing outwards. The carboxylate tail faces outwards to either His505 or Arg484, however these interactions are weak due to their high solvent accessibility.

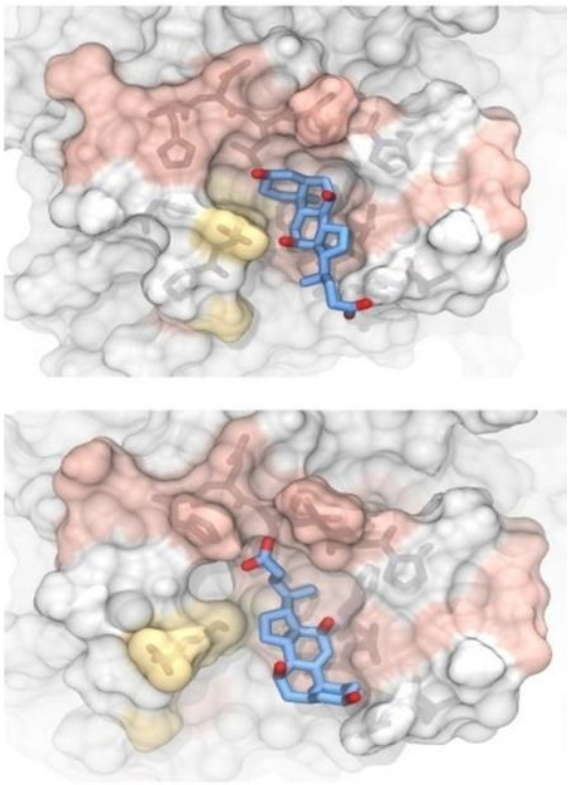

Figure 2.7: Dynamic binding mode of cholic acid. Two representative snapshots of stable protein–CA complexes for Pose 3 Rep 5 (top) and Pose 4 Rep 3 (bottom).

Not all the CSPs can be explained by contact occupancies as sampled via MD. CSPs reflect changes in the magnetic environment in proximity to the nuclei which depends on many variables such as nearby aromatic sidechains, charged groups, hydrogen bonding, etc. Hence, there is no unequivocal correlation between CSPs and inter-residue contacts. Especially, the absence of contacts with the polar residues surrounding the hydrophobic binding core (Gly503, Gln504, Asp506 and His505 on the one site and Asn470, Asp481 and Asn512 on the other) is eye striking. They are rarely touched by cholic acid; neither do they undergo large conformational changes. Possibly, a distinct electric field defines the chemical shifts in the unbound state, which would be clearly altered upon insertion of an amphoteric bile acid molecule. The methyl CSPs of Ala528 and Met530 are most likely not explained by contacts but by their conformational dynamics. The bile acids clearly displace these C-terminal residues from their native state. Finally, the predicted binding modes also agree well with the STD-NMR based epitope mapping. The hydrogens which receive the most saturation from the protein, are exactly the ones which are most buried in the complex.

2.4 Discussion

The prediction of the binding mode of small molecules to receptor proteins is an elemental step in every target-based drug discovery campaign[233]. It is the basis for the calculation of the binding affinity of the ligand which is important to recognize high-affinity binders in a large molecular

library. For novel targets, crystal structures might exist, however they show the protein only in its unbound native state. A targetable site might not be known yet or it might only become accessible through a conformational change which is only stabilized upon ligand binding (transient, or cryptic pocket) [234]. In such a scenario, a conformational ensemble of the target protein is indispensable [235]. Depending on the type of the protein, it can be acquired by high-throughput crystallography or NMR spectroscopy. Both experimental methods are however limited and costly. Molecular dynamics has proven its feasibility to sample protein conformations for decades [236]. Due to its high computational costs and the requirements of large computing clustering in the past, it took until the late 2010s years for it to be routinely employed in drug research labs[237].

In this work, only four different molecules were studied. In a drug discovery effort, the number of molecules is frequently exceed the order of millions [238]. Thus, here the limiting factor was the sampling and not the docking. A large ensemble of receptor conformations was assessed, and the docking results were compared to conformation and binding site volume. Usually, the situation is the opposite, and the virtual screening, i.e., docking of a large library, is the most time-consuming step. Then, few receptor conformations must be pre-selected to cover a broad range of the conformational space. Such a selection is non-trivial and must be undertaken with care [239, 240]. The herein presented results highlight that the docking quality does not necessarily depend on pocket shape and volume or certain representative conformation of usually low energy. Thus, it is suggested that the conformation ensemble of the receptor should be as large as possible and importantly also cover higher energy states. Additionally, in the docking, sidechain flexibility should be taken into account [241].

To deal with the energy contribution of the deformed protein, the docking poses and energies alone are not sufficient for a ranking and further calculations must be performed. Especially when it comes to fragment docking to shallow binding sites as frequently the case in protein-protein interactions, MD simulations lead to dissociation of the complex. Thus, it is preferable to run plenty of replicates with different initial velocity distributions [101]. In the end, a small set of possible binding modes still remains, and a final decision must be made. Based alone on the stability of the MD and the number of contacts, probably a different binding mode would have been chosen. Only with the support of the NMR restraints, the binding mode could be narrowed realistically.

2.5 Conclusion

Thus, the key message is that prediction of a realistic binding mode of a weak binder is still challenging. This is probably due to the fact, that a weak binding cannot be abstracted by a single binding mode and dynamics must be taken into account [242]. The true binding mode might be an

equilibrium between various binding orientations. For a stronger binding with more distinct interactions and a deeper, less flexible pocket, the employed workflow might have yielded less ambiguous results. Nevertheless, with the aid of the NMR experiments, a detailed picture of bile acid binding to norovirus P-dimers has been provided.

3 SELECTIVE CLEAVAGE OF LINEAR POLY-UBIQUITIN

This chapter focuses on the structural and dynamical basis for the recognition of linear di-ubiquitin by two evolutionary divergent proteases. The herein performed molecular dynamics simulations and their interpretation built upon pre-existent, published crystal structures and biochemical experiments. The resulting data lead to a different conclusion than one of the original articles and highlight the drastic structural effects of seemingly insignificant mutations when introduced in critical protein regions. The chapter is published [243] and [244].

3.1 Introduction

Ubiquitinylation is among the most abundant prost-translational modifications (PTM) and has key regulatory functions in most cellular processes [32]. The reversible PTMs are widely recognized to target proteins for proteasomal degradation or regulation [245] and have emerged as critical signals in innate immune response [246, 247], in which they initiate inflammation, impede pathogen growth, and trigger cell death [248].

Ubiquitinylation is achieved through the concerted action of specific E1 activating enzymes, E2 conjugating enzymes and E3 ligases [249]. Multiple ubiquitin (Ub) units can be conjugated via (iso)-peptide linkage of one of the exposed lysine residues (K6, K11, K27, K29, K33, K48, K63) or the N-terminal methionine (M1) of the proximal Ub with the C-terminus of the distal Ub [250]. This chemical bonding determines the fate of the ubiquitinylated protein, e.g., degradation, trafficking or signaling [251].

The reverse process, i.e. the cleavage of ubiquitin chains, is performed by deubiquitinase enzymes (DUBs). There are ~100 human DUBs, which are categorized into seven different families based on structural and functional characteristics [252]. While most DUBs are cysteine proteases, the JAMM subfamily employs a Zn^{2+} ion in its catalytic center [253]. DUBs possess multiple factors of selectivity control regarding protein recognition and Ub-linkage type[252, 254]. Some members of the family of OTU DUBs are known for their high selectivity towards various types of ubiquitin linkages [255]. OTUD7B is highly selective towards K11 linkage [256], OTUB1 [257] and OTUD1 [255] exclusively cleave K48- and K63-linkages, respectively, and OTULIN [258, 259] exhibits a unique activity against linear (M1-linked) polyubiquitin chains. Met1-linked ubiquitin chains are critical regulators of inflammation and immunity to pathogens [260]. Such a unique M1-selectivity requires a sophisticated multi-factorial mechanism involving multi-site recognition and substrate-assisted catalysis [259]. This high-level control of selectivity can only be addressed by structural investigations [252, 255].

Viral and bacterial pathogens have independently evolved numerous deubiquitinylating effector proteins to mimic host DUBs as a strategy to counteract innate immune response [28, 261-263]. Recently, the *L. pneumophila* effector protein RavD [264] was identified to cleave linear polyubiquitin chains and thus inhibit downstream M1-ubiquitinylation-dependent NF-κB signaling.

To extent the insight gained from available crystals structures and to put them mechanistically into perspective, molecular dynamics simulation of RavD, RavD in complex with mutant diubiquitin, RavD in complex with wildtype diubiquitin, and OTULIN in complex with wildtype diubiquitin, were performed. The MD trajectories were analyzed in regard of binding stability and energy, interaction residue number and composition as well as of conformational dynamics of the catalytic triad. The results are discussed under the aspect of how RavDs exquisite specificity for linear diubiquitin is structurally rationalized.

3.1.1 Quantitative comparison of the crystal structures

Protein crystal structures of RavD and OTULIN in absence and when in complex with substrate M1-di-ubiquitin (DiUb) are available but not conclusive in terms of structural control of selectivity and its link to the activation mechanism [259, 264]. For OTULIN, a substrate-assisted catalysis activation mechanism to bring the active site triad into a catalytically competent state was confirmed [259]. RavD, however, did not show the change in inter-residue distances as the substrate approaches the active state. In addition to the active site binding, the DUB interacts with the ubiquitins by forming protein-protein contact interfaces with the proximal (S1 binding site) and distal (S1') ubiquitin molecules [259].

RavD is a papain-like deubiquitinylase with a Cys–His–Ser catalytic triad, exhibits an overall structure that is dissimilar to OTULIN and may be considered the founding member of a novel class of DUBs [265]. The binding mode of linear DiUb to RavD is, nonetheless, almost identical to the one in OTULIN (**Figure 3.1**). Different activation mechanism for RavD and OTULIN, however, were reported.

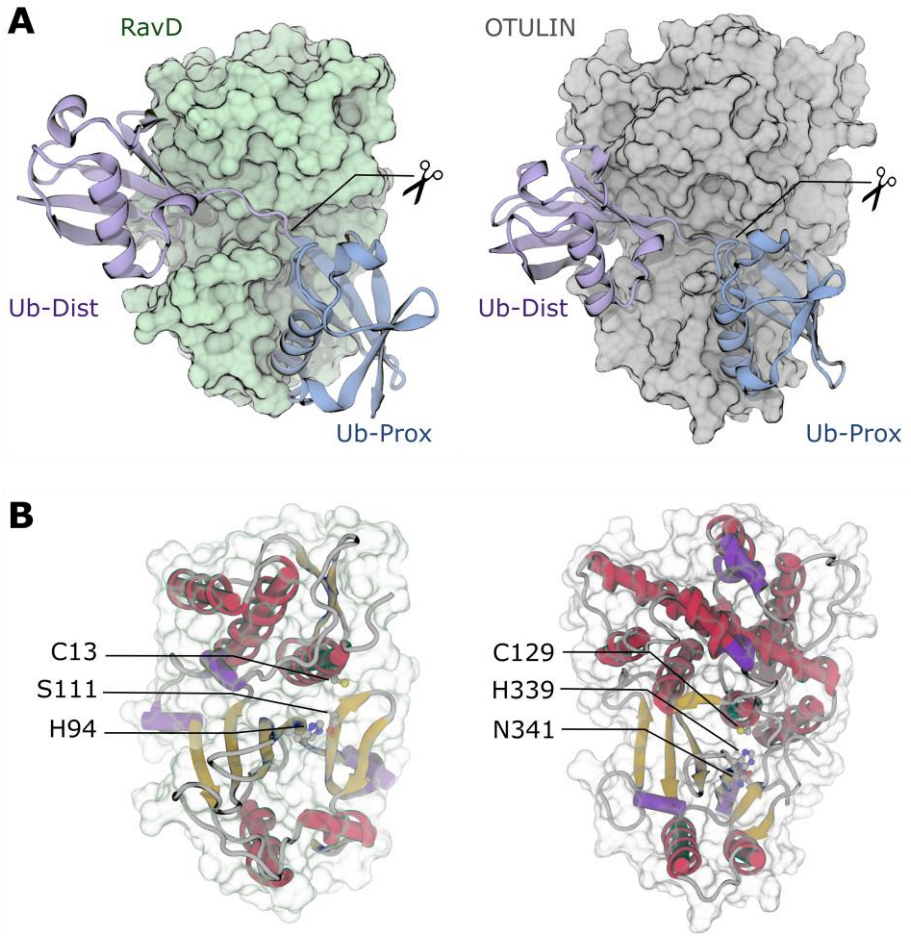

Figure 3.1: Comparison of M1-selective deubiquitinylases from *L. pneumophila* RavD and human OTULIN. (A) Surface representation of the DUBs plus ribbon diagram of the diubiquitin. The scissile bond of diubiquitin is labelled. (B) Annotation of the conserved catalytic triad residues.

The prime criterion for a clear identification of a substrate-assisted enzymatic reaction mechanism is the analysis of the structural arrangement within the catalytic center of the unbound enzyme in comparison with the enzyme-substrate complex [164]. For example, human OTULINs exquisite M1-linkage specificity originates from two Ub-recognition sites S1 and S1' and, on top, from a substrate-assisted catalysis, in which the catalytic triad is only activated upon tight substrate binding.[273] In particular, when the substrate is in a reactive conformation, the scissile bond comes as close as 4.0 Å to Cys, and the inter-residue distances for the catalytic triad residues decrease from 6.5 Å to 3.4 Å for Cys-His and from 8.3 Å to 3.1 Å for His-Asn residues **(Figure 3.2**). Only this tight complex allows deprotonation of Cys by His and thereupon the activation of the catalytic triad to form the zwitterionic state.

However, compared to OTULIN-DiUb, in the RavD-DiUb crystal structure, catalytic inter-residue distances are as large as 8.8 Å from the scissile bond to cysteine, and 7.1 Å (for Cys-His) and 2.7 Å (His-Ser). This is beyond a reactive distance for a cysteine protease [275]. The co-crystal of RavD-DiUb was obtained when complexed with a non-hydrolysable DiUb substrate analogue, in which the two terminal glycine residues of the distal Ub are mutated to serine residues (referred to as

RavD-DiUbGGSS in the following). These changes may be responsible for a hindered substrate insertion into the catalytic groove, and hence the conformational change of the catalytic triad.

Molecular dynamics simulations are able to give insight into the dynamics and accessible conformational ensembles of proteins and protein-protein complexes in aqueous solution and at finite temperature[148]. This information is complementary to that of static protein crystal structures in the solid form. We here also recover the physiological RavD-DiUbGG complex, which we refer to as 'RavD-DiUb' in the following.

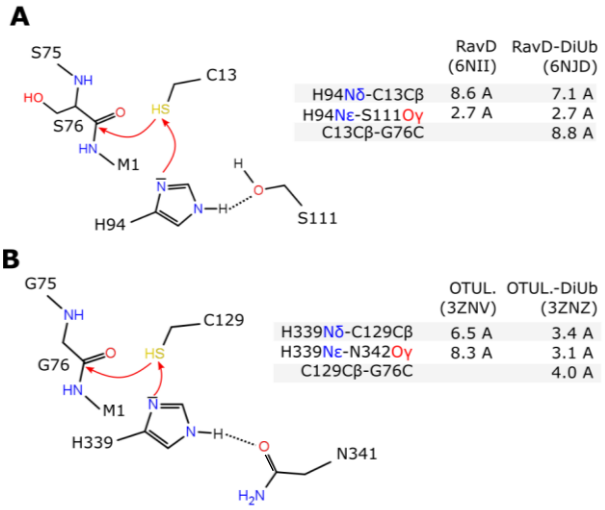

Figure 3.2: Comparison of structural parameters of catalytic triad residues in M1-specific DUBs (A) RavD (top) and (B) OTULIN (bottom). All distances to cysteine were measured from the Cβ atom for reasons of comparability with the OTULIN-DiUb crystal structure in which cysteine was mutated to alanine.

3.2 Method

3.2.1 Model generation

Both RavD and OTULIN were modeled in their free, substrate-unbound state and in complex with diubiquitin. The models are based on the crystal structures of free RavD (6NII), diubiquitin-bound RavD (6NJD), free OTULIN (3ZNV) and diubiquitin bound OTULIN (3ZNZ). The crystal structure of substrate bound RavD exhibits a double mutation in the substrate: G75S/G76S. The crystal structure of substrate bound OTULIN shows a mutation in the DUB: C13A. The OTULIN mutation was reversed and the OTULIN model resembled the wildtype OTULIN. For substrate bound RavD, two models were generated: I. the complex with the mutated substrate, and II. The complex with the wildtype substrate. Introduction of the re-mutations, as well as the addition of missing atoms (mostly hydrogen) and the protonation of the termini (R-NH3+, and R-COO-) was achieved using the software PSFGEN within VMD ver. 1.9.3 [266]. Histidine residues

were protonated at the delta position. The simulation boxes were set to a size of 10x10x10 nm^3. They were filled with TIP3P water and 0.15 M NaCl ions. CHARMM36 force filed parameters[83] were assigned.

3.2.2 Simulation protocol

Molecular mechanics and molecular dynamics simulations were carried out using openMM 7.4.1 [267]. We used PME [97] for electrostatic interactions, a non-bonded cutoff of 1.2 nm and a switch distance of 1.0 nm. Covalent bonds to hydrogen atoms were constrained. The molecular system was initially minimized for 5000 steps. Equilibration was carried using Langevin integration [268] with a 300 K thermostat, a 1.0 ps^{-1} friction coefficient and 1 fs time step. During equilibration, all the CA atoms were restrained with a force constant of 500 kcal / mol nm^2. Initial velocities were set to suite a Maxwell distribution corresponding to 300 K. Production data were generated using the same integrator but with a time step of 2 fs. Additionally, a Monte-Carlo barostat [269] was engaged to maintain a constant pressure of 1 ATM. The barostat was coupled every 25 integration steps. Trajectory snapshots were saved every 0.2 ns. Four replicate simulations were performed with individual initial velocity distributions. Every replicate had a run time of 250 ns (500 ns for RavD-DiUb), cumulating to 1 μs (2 μs for RavD-DiUb) total sampling time for the various systems. In the case of RavD-DiUb, the protocol was adapted to consider large initial fluctuations of the proximal Ub. Here, we performed an initial 500 ns equilibration run from which the four replicates of 500 ns each were initiated (with new and varying velocity distributions). The integration was performed on Nvidia GTX1080 GPUs in single precision mode.

3.2.3 Trajectory analysis

For trajectory analysis, MDTraj 1.9.5 [230], MDAnalysis 0.20.1 [270, 271], NumPy 0.51.2 [272] and SciPy 1.2.1 [273] were used. Visualization was achieved with VMD 1.9.3 [266] and MatplotLib 3.0.2 [274]. Ubiquitin RMSDs were calculated using the CA coordinates. For reference coordinates we used the initial model from the crystal structure. Before the RMSD calculation, the whole protein complex was aligned by the CA atoms of the DUB (RavD, or OTULIN respectively). For the computation of the RavD proximal Ub pairwise RMSD matrix, 10.000 frames where chosen (every 200 ps). The binding energy was estimated using the PRODIGY[123] of multiple conformations extracted from the MD simulations. For RavD, 1.000 conformations and for OTULIN 1.000 conformations (every 1 ns) were chosen. The interface areas were calculated from the differences in the solvent accessible surface areas (SASA), i.e. the sum of the SASAs of two proteins alone minus the SASA of the complex. For the SASAs of the proteins alone, also the complex trajectory was used. The per-residue SASAs were computed using the algorithm of Shrake and Rupley [275] with a probe radius of 0.14 nm and 512 sphere points. The residue-wise buried areas were

computed as the relative difference between SASA in the unbound and the bound forms. For intermolecular interactions, heavy-atom contacts were calculated with MDTraj, employing a distance cutoff criterion of 0.55 nm. The residue type contributions were calculated as the ratio of contacts of the single residues to the total contacts. The classification followed the standard convention. Histidine, cysteine, glycine, and proline were always considered polar. The free energy maps were constructed from the 2D joint probability distribution functions (PDF) of the interatomic distances of Cys-S:His-Nδ and HIS-Nε:Ser-Oγ (HIS-Nε:ASN-Oγ for OTULIN). The PDFs were estimated using binning to 50x50 bins. The free energy difference was calculated as the negative natural logarithm of the PDF.

3.3 Results and Discussion

3.3.1 Stability of diubiquitin binding

The stability of a protein-protein complex in solution can give qualitative insight on its binding energy. Intuitively, a stable binding without large fluctuations would correspond to a high absolute binding energy. From four replicates of each 250 ns of the complexes RavD-Diubiquitin and OTULIN-diubiquitin, the relative ubiquitin RMSDs were calculated. In contrast to the standard RMSD of atomic coordinates, which is calculated after removing translational and rotational motions by least-square fitting to a reference conformation, the herein considered RMSDs include translation and rotation of the individual ubiquitin units (proximal and distal), relative to their bound DUB (RavD or OTULIN). This is achieved by fitting the whole complex by only the Cα atoms of the DUB. The high-frequency and low-amplitude fluctuations based on conformational dynamics can be removed from the RMSD by employing a low-pass filter. This way, the RMSD only reflects orientational contributions by means of relative translation and rotation away from the crystal structure. The RMSDs of the distal RavD-bound ubiquitin vary between 0.3 and 0.5 nm relative to crystal structure depending on the replicate (**Figure 3.3 A**). Every replicate RMSD has stabilized but still shows fluctuations in the order of 0.1 - 0.2 nm. For reference, the distal ubiquitin of the OTULIN-DiUb complex stabilized at a value of 0.2 - 0.3 nm relative to the crystal structure and fluctuations are only the order 0.05 to 0.1 nm. Thus, the initial distal ubiquitin binding mode as seen in the crystal structures is more dynamic in the RavD-DiUb complex than in the OTULIN-DiUb complex. Similar can be observed at the proximal binding sites. On RavD, the ubiquitin RMSDs exhibit baseline values of 0.2-0.3 nm but show large fluctuations of up to 0.7 nm. The RMSDs of the OTULIN bound proximal ubiquitin stabilized to values of 0.2 nm and exhibit only minor fluctuations. Based on the RMSDs, the proximal binding of Ubiquitin to RavD is less stable than to OTULIN.

In addition to the RMSD analysis, binding affinities of large conformational ensembles (1.000 snapshots) were estimated using an interaction-based method called PRODIGY [123, 276]. It approximates the binding energy using a regression model based on a benchmark set of 81protein-protein complex with energies between –6 and –16 kcal/mol. The regression model mostly makes use of the numbers of the various possible types of inter-residue interactions (e.g. hydrophobic-hydrophobic or charged-charged) and is among the most accurate methods for binding energy estimation. For the distal ubiquitin binding to RavD, dissociation constants K_D were mostly in the order of 3 to 10 nM, and the whole ensemble covered values between 0.8 and 900 nM (**Figure 3.3 B**). The distal ubiquitin binding to OTULIN is surprisingly slightly weaker. Sampled conformations yielded dissociation constants between 2 nM 800 nM with the highest probability density around 10 and 20 nM. On the proximal site, the differences between RavD and OTULIN are much clearer. RavD binding of the proximal ubiquitin showed dissociation constants in the range of 70 nM to 5 µM, with a peak between 300 and 600 nM. The dissociation constants computed for proximal ubiquitin binding to OTULIN are significantly lower. Most of the sampled conformations yield values of 7 nM and the whole sample exhibits a K_D bandwidth of 2 to 20 nM. In summary, based on the K_D prediction, the binding of the distal ubiquitin to RavD is as strong as to OTULIN. The distal and proximal binding sites of OTULIN are of similar strength. RavDs proximal site significantly weaker than its distal site and thus also significantly weaker than OTULINs proximal site.

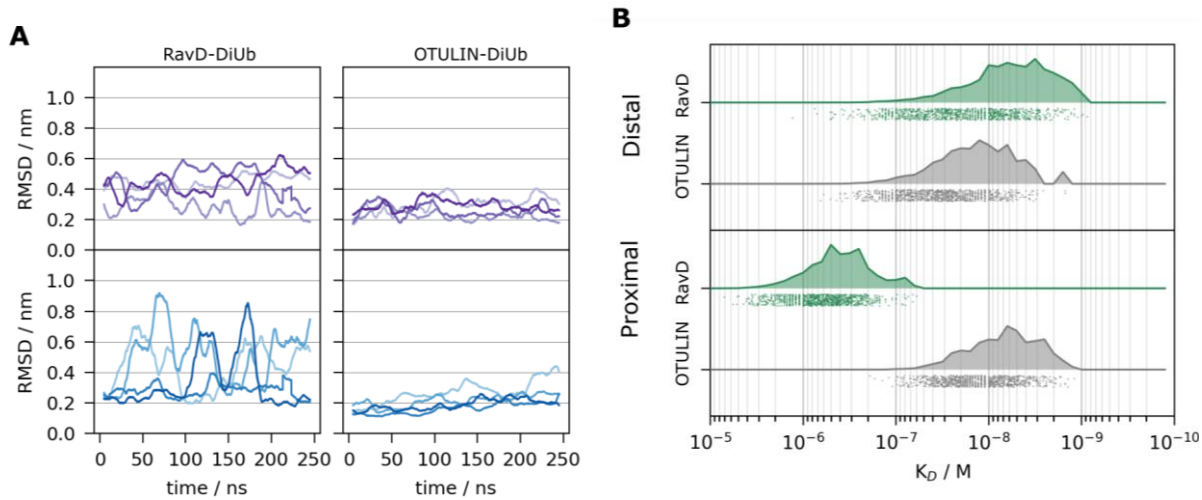

Figure 3.3 Comparing diubiquitin binding strength between RavD and OTULIN (A) Cα root mean squared deviation (RMSD) for the individual ubiquitin units relative to the DUB (see text for details). A moving average is drawn for clarity. Ubiquitin binding to the S1 site is given in purple in the top panels, proximal ubiquitin binding in blue in the bottom panels. (B) Calculated binding free energies for distal and proximal ubiquitin to RavD and OTULIN. The points represent individual results for 1.000 snapshots for each RavD and OTULIN. The filled curves represent the probability density functions estimated from 50 equidistant (log space) bins.

Putting the estimated binding affinities into perspective is difficult because such data is not available for deubiquitinase-ubiquitin complexes. However, the dissociation constants between ubiquitin and various ubiquitin-binding domains are frequently in the micromolar range [286, 287]. Micromolar binding can be considered transient whereas high-nanomolar binding, as predicted for the proximal ubiquitin to RavD, is more stable yet still not permanent but dynamic [288].

Another means to assess the quality of protein-protein binding is the identification residues which bury deeply into the interface upon binding. Usually, this can be achieved by comparing the solvent exposed area of residues in the free and complexed state. However, such a computation is not straightforward because the solvent accessible area of a residue depends on its size and flexibility as well as its position within the protein. Here, the relative reduction of the solvent accessible area upon binding is considered sufficient (**Figure 3.4 A-B**). This way, the distal ubiquitin residues Gly47, Ile44, Val70, Leu8 and Thr7 were identified to bury more than 80% of their surface area into the protein-protein interface when complex with RavD. OTULIN on the other hand enables deep burial of Ile44, Val70, Leu8 as well as Leu73 and Gly75. Whereas the RavD-bound ubiquitin residues form one large patch, they are spatially more distributed in distal the OTULIN binding. This is also reflected by somewhat smaller interface area of the OTULIN-DiUb complex of 22 nm^2 relative to the distal RavD-DiUb interface of 23 nm^2. It is striking, however, that OTULIN deeply buried the c-terminal ubiquitin residues whereas RavD did not, or at least not to a significantly lower extent. On the proximal site, OTULIN buries a large, continues and distinct patch stretching over residues Gln78, Met77, Ala93, Glu92, Lys105, Asp108, and Lys109. This patch forms most of the interface area with a size of 17 nm^2. RavD accommodates broadly identical proximal ubiquitin residues but fewer residues are deeply buried (only Ala93 and Asp108). The interface area is smaller and has a median size of 12 nm^2 (**Figure 3.4 C**).

The buried residue analysis yields the picture that both RavD and OTULIN have developed binding surfaces to recognize identical patches on their specific ubiquitin units (distal or proximal). The quality of the distal binding site is similar but RavD seems to put more focus on the hydrophobic ILE44 patch whereas OTULIN also includes the C-terminus. The interface area on the distal site is similar yet slightly larger on RavD. Its size fluctuates considerably in the simulation on both RavD and OTULIN and despite its large size, only few ubiquitin residues are fully buried. Compared to OTULIN, the proximal binding site of RavD is substantially smaller, more dynamic and buries fewer residues. The binding is weak and mediated by only few strong interaction anchors. This leads to orientational tumbling and temporal solvent accessibility of the other residues.

The total interface size of the distal binding sites in the order 22-23 nm^2 is large as compared to other single-patch protein-protein complex interfaces which are commonly in the order 15 nm^2 [277]. Others have described that the standard interface size is between 12 to 20 nm^2 [278]. Smaller patches in the order 11.5 – 12 nm^2 indicate short-lived, low-stability complexes [279]. This fits well to the low RMSD stability, and the high nanomolar affinity of the proximal ubiquitin to RavD, which has exhibits an interface area of 12.5 nm^2. The largest interface areas are reported for proteases and their highly specific inhibitors. Such areas can attain values between 20 and 46.6 nm^2. The total binding interface area of OTUIN-diubiquitin is 40 nm^2 [280], and is thus among the largest interfaces. It must be noted, that OTULIN of course belongs to the class of proteases. RavDs total interface area with diubiquitin is 35 nm^2, which can still be considered very large. In conclusion, both protein RavD and OTULIN will engage in tight and stable complexes with their linear linked diubiquitin substrate. After bond cleavage though, the OTULIN would still stably bind both ubiquitin moieties, whereas for RavD it is likely that the proximal Ubiquitin would dissociate.

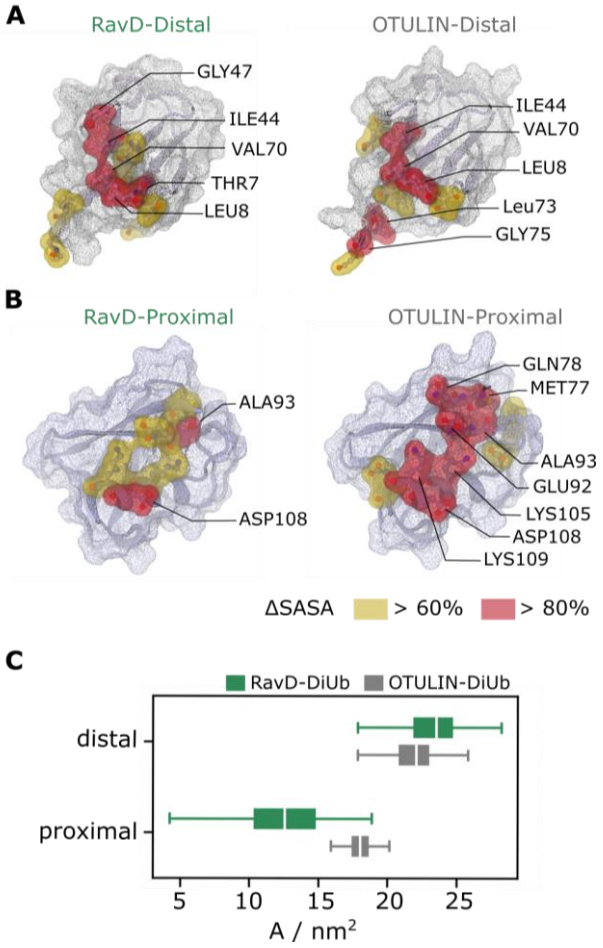

Figure 3.4: Buried residues and interface areas. Changes in protein solvent-accessible surface area (ΔSASA) upon DiUb binding to RavD and OTULIN are shown in perecent. DUB binding to (A) distal and (B) proximal ubiquitin. The boxplot (C) shows the total protein-protein interface areas. Whiskers show lowest and highest sampled areas of the entire ensemble. The box length represents the interquartile range (50% of data). The central line is the median.

3.3.2 Inter-residue interactions

For the identification of key interaction residue pairs and the quantitative comparison of binding interfaces, it is useful to count all inter-residue contacts throughout the MD trajectory. On the molecular scale, a contact is not unambiguously defined, and different contact criteria lead to different results. In here, a contact between two residues is count, when at least one of the pairwise heavy-atom distances was shorter than 0.55 nm. All heavy atoms, sidechain and mainchain, were considered. It frequently appears that a residue of the one protein is in contact with more than one of the other. In this case, multiple contacts were counted. The number of contacts can be expressed as the contact contribution by normalizing the residual number of contacts by the total number of contacts. This representation allows a direct discrimination of important drivers of the protein-protein interaction. Based on the contributions of all monitored inter-residue interactions and the residue types (charged, polar, apolar) the contribution of the interaction types can be calculated and used to assess how favorable the protein-protein interaction is. The results are presented in **Figure 3.5**.

On the distal interfaces, in total average, RavD engages in 78 inter-residue contacts and OTULIN in 81. The S1 site of RavD recognizes the distal ubiquitin by four sites. The first resolves around Leu8 and Thr9, which together contribute to 15% of the total contacts. A second patch belongs to the c-terminal residues of Leu73, Arg74 and Gly75, together contributing 20% of the total contacts. The Ile44 hydrophobic patch contributes less to the total number of contacts. A fourth binding patch is recognized at Glu34. Interestingly, OTULIN binds essentially identical residues with only marginal differences. The patch at Leu8 is locally more constrained, the Glu34 patch is slightly weaker, and the C-terminal patch is stronger and more extended towards the last residues Gly75 and Gly76. Analysis of the types of the involved residues yields a similar picture for RavD and OTULIN. About 19%, respectively 17%, of the interactions belong to the favorable group of apolar-apolar, i.e., hydrophobic interactions. Each 12% are charged-charged interaction, and 12% and 16%, respectively, belong to polar-polar interactions. Both RavD and OTULIN exhibit 19% charged-polar interactions at the distal binding interface. The total amount of less favorable apolar-polar and apolar-charged interactions is 37% for RavD and OTULIN. Thus, the composition of interaction types is similar between RavD and OTULIN and the majority of interactions is favorable.

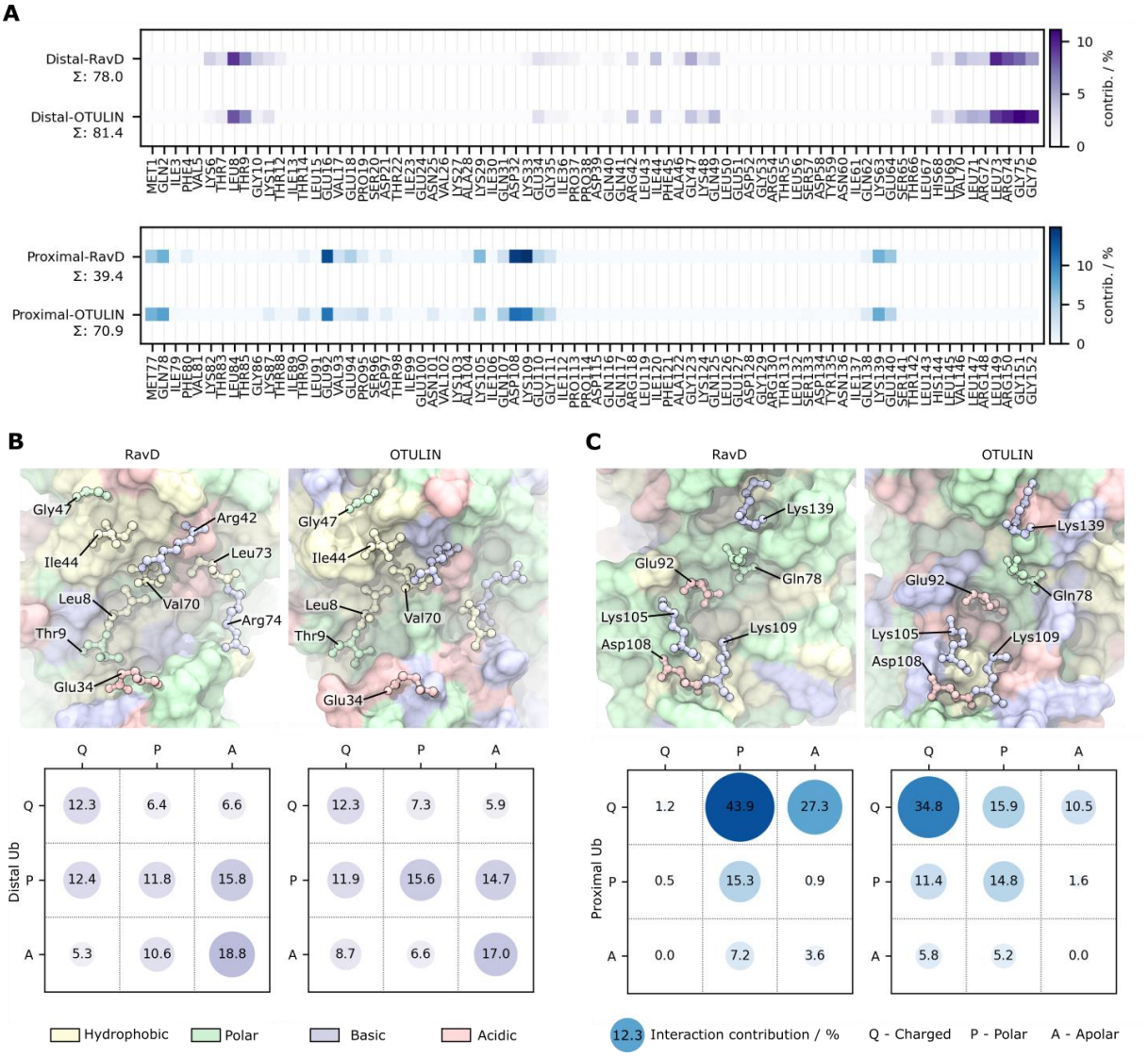

Figure 3.5: Analysis of persistent DUB-DiUb interactions. (A) RavD and OUTLIN residues interaction with the distal and proximal ubiquitin molecules. Key protein-protein contact residues of DiUb are given by their residual contribution to the total number of inter-residue interactions in percent (see Method for details). (B) Snapshots of distal Ub binding to the S1 sites of RavD and OTULIN and inter-residue interactions between charged (Q), polar (P) and apolar (A) DUB S1 residues with their distal Ub counterparts. (C) Snapshots of proximal Ub binding to the S1' sites of RavD and OTULIN and inter-residue interactions between charged (Q), polar (P) and apolar (A) S1' DUB residues with their proximal Ub counterparts.

On the proximal site, the total number of interactions is much lower on RavD (39) relative to OTULIN (71). However, the residues of ubiquitin which come into contact with the DUBs are, again, identical, yet with slightly different contributions. The biggest drivers of the interactions are Glu92 with about 10% contribution and Asp108/Lys109 with 25% and 20% contribution for RavD and OTULIN, respectively. A third recognition site for Lys139 is identified on both RavD and OTULIN. Surprisingly and in stark contrast to the distal binding site, the residue and interaction composition of RavD and OTULIN are widely different. OTULIN exposes more than 50% charged residues, of which 70% are engaging in favorable charged-charged interactions. The amount of apolar residues is as low as 12 % and not a single hydrophobic interaction was monitored. RavD employs a mere 2% charged residues for the protein-protein interaction. Instead,

the charged residues of ubiquitin are overwhelmingly often (44%) bound to polar residues and to 27% to apolar residues. Such an interactions pattern is much less favorable than the one at the OTULIN binding site and explains why RavD does not manage a stable proximal ubiquitin binding and why the binding affinity is so low. On the other hand, it is surprising, that RavD accommodates Glu92 and Asp108/Lys109 so well. The role of Glu92 in OTULIN is special: it binds close to the active site and stabilizes the competent conformation of the catalytic triad [259]. It is thus a crucial part of the substrate catalytic mechanism of OTULIN.

3.3.3 Substrate binding and catalytic triad

In contrast to OTULIN, a substrate assisted activation of the catalytic triad was disregarded for RavD. This assumption was based on a low RMSD between diubiquitin-bound and unbound RavD. Such a criterion is not a sufficient, and distances between catalytic residues must be considered in the different states. Molecular dynamics allows sampling of various conformations and the quantification of the free energy differences between them (**Figure 3.6**). For diubiquitin bound OTULIN, most of the sampled conformations had cysteine-histidine distances of 0.4 nm and cysteine-asparagine distances of 0.25 nm. A second population with a histidine-asparagine distance of 0.8 nm is two order is two orders of magnitude less likely. Such close distances allow proton transfer from cysteine to histidine to activate the catalytic triad. In unbound RavD, the catalytic distances show large flexibility, and many conformations are accessible without high energy barriers between them. There are two shallow energy wells: one at cysteine-histidine of 0.9 nm and histidine-serine of 0.3 nm, and one at cysteine-histidine of 0.9 nm and histidine-serine of 0.9 nm. Both energy wells would not lead to a competent catalytic triad. Upon binding of diubiquitin however, the energy landscape changes, and the energy minima shift towards shorter cysteine-histidine distances. Only when diubiquitin is bound, a substantial number of sampled conformations would allow deprotonation of cysteine. This observation is consistent for the binding of mutated and wildtype diubiquitin. However, the effect is more pronounced for wildtype diubiquitin. Thus, binding of diubiquitin to RavD clearly leads to a stabilization of the catalytic triad as well as a facilitation of short cysteine-histidine distances. To be even more precise, the catalytic triad of RavD is inactive in the unbound, and at least partially active in the substrate bound state. However, the probability for an active catalytic triad arrangement is still one order of magnitude lower in substrate bound RavD compared to substrate bound OTULIN.

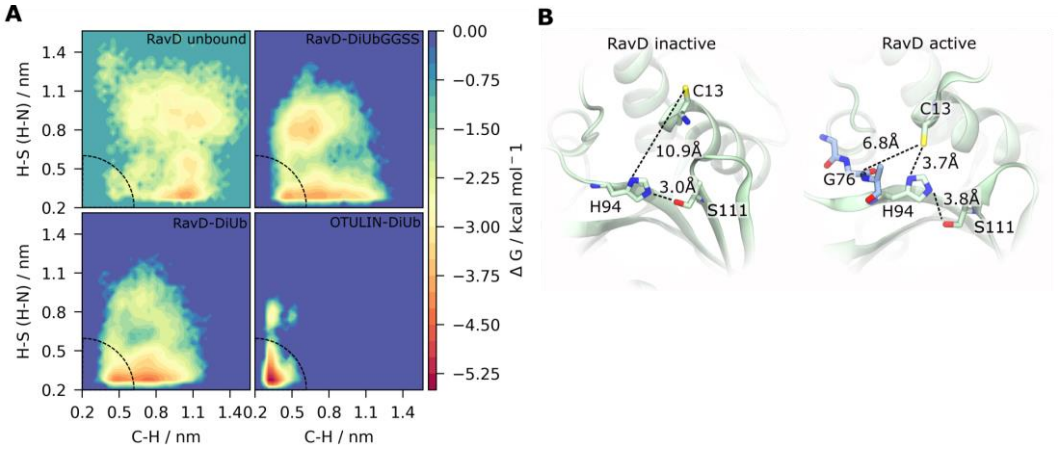

Figure 3.6: Diubiquitin binding induces partial catalytic triad activation. Free energy maps (A) of relevant inter-residue distances of catalytic residues for unbound RavD, and substrate-bound RavD and OTULIN (see text for details). The area in which the catalytic triad is in a catalytically competent state is marked. The free energy landscapes were clipped at the highest sampled energy. (B) Representative MD snapshots of catalytically active and inactive states based on the energy wells of unbound RavD and RavD-DiUb.

3.3.4 Conclusion

The bacterial effector protein RavD hydrolyses linear polyubiquitin chains with high selectivity. Therefore, it has developed a highly specific distal binding site which is in every aspect comparable to that of OTULIN. Such a S1 binding sites selectively binds ubiquitin in an orientation in which the ubiquitin C-terminus is positioned into the catalytic grove. This however does not yield linkage selectivity because the distal binding mode is identical independent of the polyubiquitin linkage and the C-terminus is always the "substrate" which holds the bond to be cleaved. Thus, OTULIN for example employs a second, highly specific binding site for the proximal ubiquitin to be oriented so that Met1 (M77 in case of diubiquitin) comes close to the active site. In OTULIN this is achieved by three distinct recognition sites to selectively bind ubiquitin residues Glu92, Asp108 and Lys139. RavD has also developed a second ubiquitin binding site to accommodate exactly the same residues. However, the proximal binding site on RavD is substantially weaker. It makes fewer contacts, it buries a smaller area, and it engages in less favorable interaction types. Thus, the selectivity and activity of RavD towards linear linkages would be significantly lower than of OTULIN. As this is not reported and RavD performs similar to OTULIN, another factor must be considered: substrate-assisted catalysis. It was originally disregarded for RavD, but it is inevitable to explain its selectivity. Additionally, it is clearly suggested by the molecular dynamics simulation. The question arises why the substrate-assisted catalysis is not clearly seen in the crystal structures.

The reasons for this misconception are to be sought in the mutation of the substrate. The mutation, a double exchange of glycine by serine at both the c-terminal residues of the distal ubiquitin, was necessary to avoid hydrolysis and hence allow crystallization of the wildtype protein with a full diubiquitin. However, as only the glycine-glycine motif is small enough to bury into the active site, the serine-serine motif remained outside of the binding grove and engaged in s few superficial

interactions. The C-terminus and initial residues of the N-terminus are of sufficient flexibility, so that the absence of a tight central binding, did not interfere with the remote binding of distal and proximal ubiquitin. Anyway, the absence of a close approaching of the substrate towards the catalytic center inhibited the conformational transition within the catalytic triad to the activated state. This lead the original authors to the assumption that substrate assisted catalysis would not take place. In the MD simulation, the effect of the proximity of the substrate on the dynamics of the catalytic triad can clearly be seen. Only the final transition, in which the substrate buries into the catalytic grove and fully stabilizes the active catalytic state is not achieved within sampling time. To be clear, such a transition is impossible for the mutant substrate because it is too large to fit into the tight grove. It is possible, though, for the wild type substrate, but could not be sampled within the cumulated sampling time of 1 μs.

To summarize, the mutation of the substrate lead to an artificial binding mode, in which both the protein-protein recognitions sites for distal and proximal ubiquitin are correct, but not the positioning of C-terminus and thus the arrangement of the catalytic triad. Based on this artificial crystal structure, the author came to the wrong conclusion about the activation mechanism of RavD which lead to large ambiguities in the structure-selectivity relation. Here, molecular simulation was employed to sort out this discrepancy by showing that RavD activation must be substrate-assisted.

4 GLYCAN CONFORMATION AND GLYCOSYLATION SITES

Complex N-glycans are commonly considered highly flexible and disordered. In this chapter, the site-specific, intramolecular protein-glycan interactions are investigated and their effect on the degree of disorder is quantified. Therefore, an embedded clustering workflow based on the glycosidic torsion angles was developed, tested on artificial data and employed on large molecular dynamics ensembles. The workflow allowed the identification of more than 100 conformational clusters within the glycan conformational space. Whereas the adopted conformations were broadly identical, the glycosylation site had a significant effect on the abundance of certain conformations. The results highlight that protein N-glycosylation are not fully disordered, but transition quickly between many distinct conformational states, some of which are stabilized by specific protein-glycan interactions.

4.1 Introduction

Protein glycosylation presents a posttranslational modification that occurs in all domains of life [281] and is characterized by its tremendous diversity and abundance [282]. Glycosylation is mostly defined as the concerted enzymatic conjugation of one or more glycan moieties (carbohydrates) to either asparagine (N-glycosylation) or serine/threonine (O-glycosylation). Common glycosylation types are shown in **Figure 4.1**. The posttranslational modification takes place in the endoplasmic reticulum and Golgi system and affects over 85% of secretory proteins by N-glycosylation [283] and a majority of nuclear and cytoplasmic proteins by additional O-glycosylation [284].

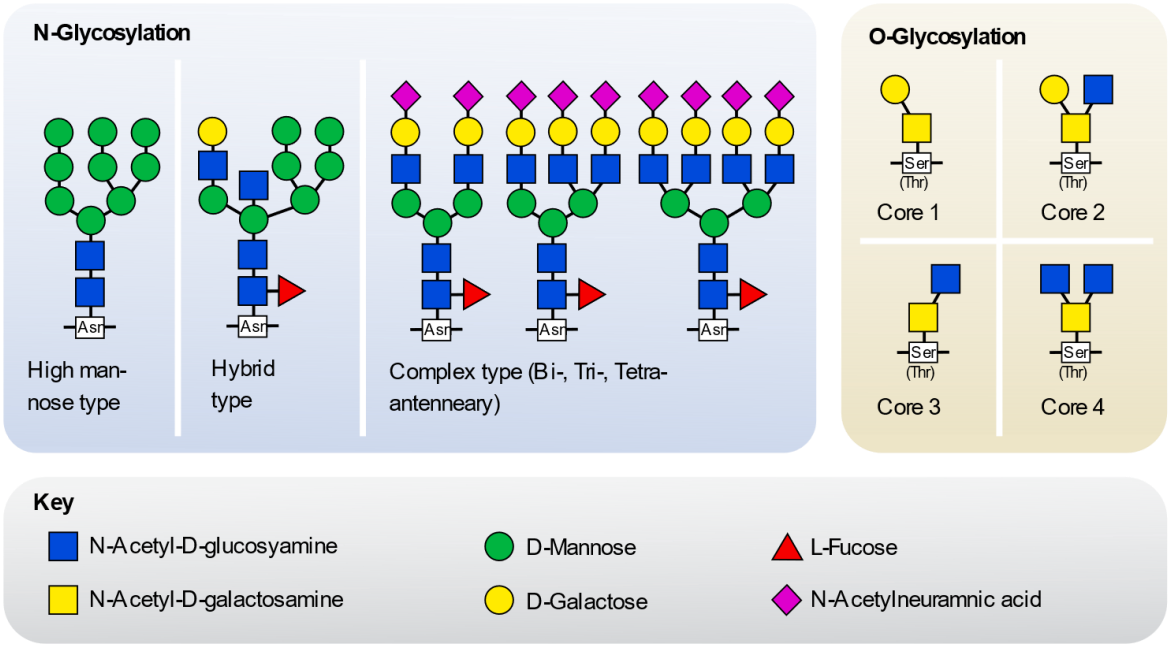

Figure 4.1: Overview of common N- and O- glycans. N-glycans are conjugated to asparagine, O-glycans to serine residues. The shown moieties represent only a small subset of the most commonly occurring sugars. Symbols and nomenclature according to https://www.ncbi.nlm.nih.gov/glycans/snfg.html.

On the molecular scale, glycosylation of proteins is important for folding, quality control, stability, function and transport [285]. On the cellular level, the roles of glycans comprise, among others, cell-adhesion, ligand binding, effector function and receptor dimerization [282]. Frequently, pathogens specifically recognize glycosylation motifs on host cells (e.g. norovirus, see Chapter 2) or produce host-mimetic glycoconjugates to evade immune response [286]. Proteome-wide glycosylation patterns are often disturbed in disease and can be used as diagnostic and prognostic markers [287]. Recent reviews summarize the mutual effects of glycosylation and cancer [288] and auto-immune diseases [289].

Structurally, glycosylation and especially N-glycosylation underlies a certain microheterogeneity, which is based on the somewhat unregulated enzymatic cascaded in the ER and the Golgi [290]. Thus, in vivo glycoproteins exist as a population of different variants, so called glycoforms, which exhibit glycosylation site-dependent distributions of different glycan types. Development of methods for the site-resolved, quantitative analysis of different glycosylation motifs is subject of ongoing research [291, 292] [293]. Such heterogeneity and the obstacles in both analysis and preparation of pure glycoproteins pose a major challenge for experimental structural studies. Thus, in the protein data bank (PDB), structures containing glycans are heavily underrepresented [294].

Theoretical methods, such as molecular modeling and simulation, historically played and still play an important role in the field of structural glycobiology [295-297]. Especially in the recent years during the Corona virus pandemic (2019 and ongoing), computational, structural glycobiology

regained tremendous interest because of the heavy glycosylation of the SARS-COV-2 spike protein [298-300]. However, a thorough understanding of the conformational landscape of complex glycans remains largely elusive. Recent experimental advancements enable insight to structurally well-defined glycans using an imaging method [301] or in the gas phase via IR spectroscopy[302]. Hence, molecular simulation is still the method of choice to investigate conformational ensembles of complex carbohydrates. It is, however, limited by three key factors: sampling, force-field accuracy, and data analysis. Especially ring puckering is a challenge for MD simulation, as with current force fields it happens on very slow time scales in the order 0.07 per microsecond [303]. With current computing resources, such sampling can only be achieved for small oligosaccharides and not for large glycoproteins. Additionally, in the case of complex carbohydrates with many interdependent degrees of freedom, data analysis of the trajectory is tedious and a gold standard does not exist. Thus, it is unclear if and how intramolecular interactions would mutually affect protein and glycan conformation.

In this research, multi-microsecond scale sampling of complex a bi-antennary, sialylated complex N-glycan attached to human erythropoietin (EPO) as a model glycoprotein is achieved (**Figure 4.2 A**). The mutual effects of the protein core of EPO and its glycan shield are studied. Here, the effects of the glycosylation on the molecular properties are probed. Additionally, the effect of the local protein environment on the structure and dynamics of complex glycans at the three glycosylation sites is examined. The second task revolves around an unprecedentedly detailed analysis of the conformational space of the glycan. Therefore, a dimensionality reduction and clustering workflow was developed, tested and employed.

4.2 Method

4.2.1 Glycoprotein modeling

To understand site-specific effects and allow site-specific conformational analysis, 16 different EPO glycol-isoforms (**Figure 4.2 C**) were structurally modeled and each subjected to 3 x 150 ns molecular dynamics sampling. The simplest model contains no glycosylation at all. At the N-glycosylation sites, a complex-type, bi-antennary, double sialylated, and fucosylated glycan was added (**Figure 4.2 B**). At the O-glycosylation, a standard core 2 type O-glycosylation was conjugated. Every permutation of the 4 glycosylation sites: Asn24, Asn38, Asn83 and Ser126 was modeled. The corresponding N-glycan was selected from the database of the glycosciences.de [304] webserver and *in silico* glycosylated to the structural model of human EPO. The initial glycosidic torsion angles were chosen to avoid atomic overlap. The EPO model was based on the crystal structure 1EER [305]. It was subjected to 10 ns molecular dynamics simulation in the non-glycosylated form. To enable *in silico* glycosylation, the relative solvent accessibilities of Asn24,

Asn38 and Asn83 were monitored and a snapshot with high accessibility of all three residues was selected. The resulting PDB files were parsed for the CHARMM-GUI glycan reader [306] by removing empty lines and adjusted TER statements and chain IDs. This way, CHARMM-GUI [307] could be used for solvation and ionization of the glycoprotein models as well as parameter assignment with the CHARMM36 [83 and CHARMM carbohydrate force field, 85, 308]. Additionally, CHARMM-GUI glycan modeler [306] was utilized to manually add O-glycosylation to Ser126.

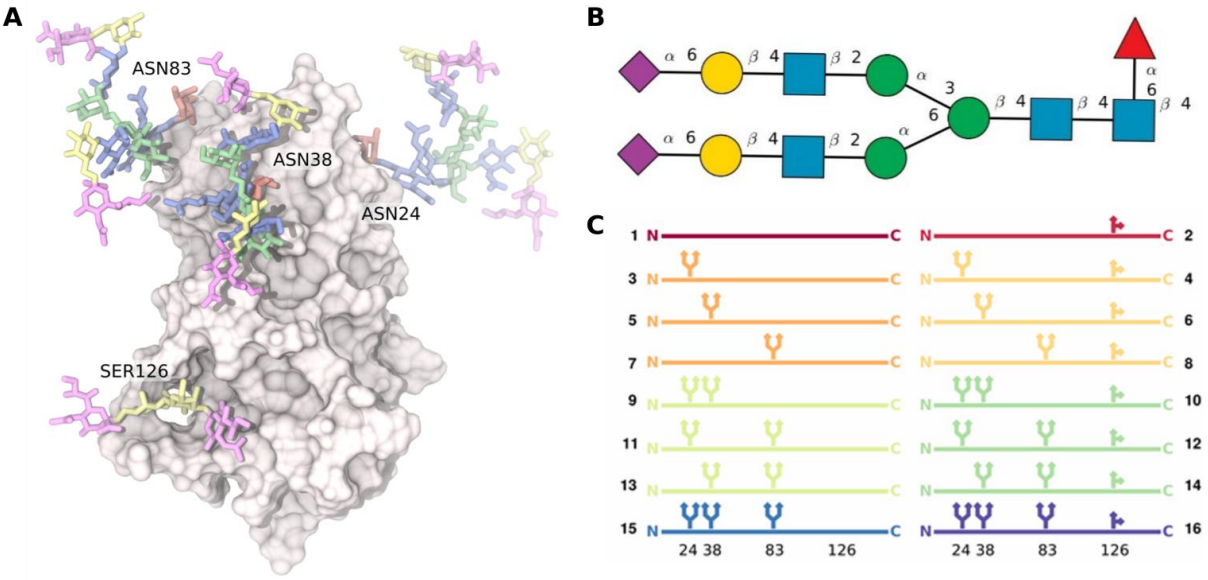

Figure 4.2: Structural models of fully glycosylated recombinant human EPO. A) Surface model of EPO with annotated N- and O-Glycosylation sites and the herein investigated glycans in licorice representation (colored by residue type) B) Schematic representation of the complex N-glycan with glycosidic bond annotation. Residue type is coded by shape and color according to conventional glycan nomenclature (https://www.ncbi.nlm.nih.gov/glycans/snfg.html). C) Schematic representation of the 16 modeled EPO glycol-isoforms with the color coding corresponding to the number of glycan modifications.

4.2.2 Simulation protocol

MD simulations followed a standard protocol of 5000 steps steepest descent minimization, 25000 steps of NVT equilibration with a short 0.001 ps time steps followed by and 150 ns production sampling with a 0.002 ps time step and NPT ensemble. Three replica of production runs were performed for each system. All simulation were carried using GROMACS ver. 5.1.5 [75-77, 80, 81, 309]. In all instances, a Verlet cutoff scheme with a neighbor list update step of 20 was employed. Short range interactions were cut off after 1.2 nm. Van-der-Waals interactions used a force-switch modifier between 1.0 nm and 1.2 nm. Long range electrostatic interactions were treated with the particle mesh Ewald method [97-99]. Temperature coupling was achieved with a Nose-Hoover thermostat [225, 310], in which the glycoprotein and the solvent were coupled separately. The coupling constant was 1 ps and the reference temperature was set to 298.15 K. Initial velocities were assigned to fit a Maxwell-Boltzman distribution at a temperature of 298.15 K. Bonds to

hydrogen atoms were constrained via LINCS [311]. Center of mass motions were removed. During minimization and equilibration, the backbone and sidechain atoms were restrained using force constants of 500 kJ/mol nm and 50 kJ/mol nm respectively. In the production phase, isotropic Parrinello-Rahman pressure coupling [312, 313] with reference pressure of 1 ATM, a coupling constant of 5 ps and a compressibility of 4.5e-5 was used.

4.2.3 Trajectory analysis

The global glycoprotein properties were elucidated using GROMACs *tools*. The backbone RMSD was computed with *gmx rms* after least square fitting of only the protein backbone coordinates. The radius of gyration was determined with *gmx rgyr* and the considered the whole glycoprotein molecules. The number of all intramolecular H-bonds were computed with *gmx hbonds* and a minimum heavy atom distance of 0.3 nm and a minimum angle of 150°. The total solvent accessible area was monitored with *gmx sasa* and a probe radius of 0.14 nm. Translational diffusion coefficients were estimated from linear fitting of the mean squared displacement functions using *gmx msd*. The dipole moments of whole molecules were computed with *gmx dipole*. More localized effects were assessed using the residue-wise descriptors of relative solvent accessibility and backbone RMSF. The values were calculated via *gmx sasa* (probe radius of 0.14 nm) and *gmx rmsf* (alignment via protein backbone atoms). The data were pooled from all trajectory replicas and models and averaged. The differential values were calculated for all glycosylated proteins with respect to the non-glycosylated protein. Furthermore, the residue-wise protein contact occupancy with the glycan with a heavy-atom minimum distance criterion of 0.6 nm were calculated using MDtraj Ver 1.9.3.[230] The interactions of the N-glycans with their protein were analyzed similarly. Here, the contacts were resolved individually for the different glycosylation sites and categorized by protein residue type. The software MDTraj was also used to compute the glycosidic torsion angle distributions of the N-glycans. Therefore, list files containing the atom indices according to figure 31 A were generated and utilized. In total, 26 glycosidic and 3 root asparagine sidechain torsion angles were analyzed. For clustering and embedding, they were transformed into the cosine, sine space via the transformation $z(\varphi) = [\cos(\varphi), \sin(\varphi)]$. Pairwise circular correlation coefficients were calculated using instructions by Johnson et al. [314] Embedding was performed with UMAP[172, 173]. Clustering was achieved with HDBSCAN [182] within the embedded space. Different combinations of parameters were investigated and can be found in the results section. Generation of figures and structural render images as utilized with Matplotlib [274] and VMD ver 1.9.3 [266] respectively.

4.3 Results and discussion

4.3.1 Global effects of N-glycosylation

Molecular dynamics simulation of the above given 16 different EPO glycoforms were carried out in order to systematically investigate the effect of glycosylation on global protein properties such as backbone RMSD, radius of gyration, number of intramolecular hydrogen bonds, solvent accessible surface area, translational diffusion coefficient and electric dipole moment (**Figure 4.3**).

The overall backbone RMSD is between is 0.24 ± 0.02 nm for the non-glycosylated EPO protein. Upon glycosylation, the backbone RMSD is always between 0.20 and 0.27 nm for all glycosylated models of EPO independent of site, number and type of glycosylation. The differences between the glycosylated models is only marginally larger than the differences between the replicates of each system. This shows that glycosylation does not affect the overall protein backbone RMSD and there are no larger conformational changes upon glycosylation. A RMSD of 0.3 nm or below is common for small globular proteins.

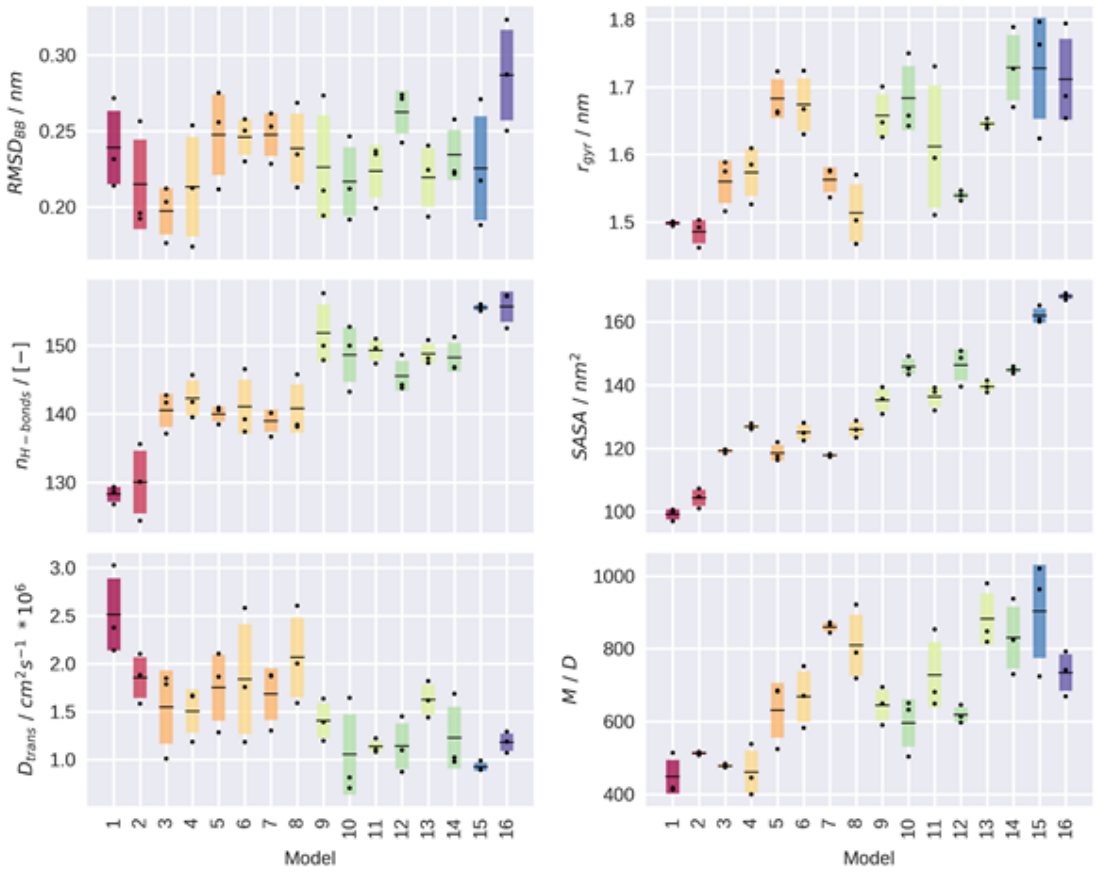

Figure 4.3: MD simulations yield structural descriptors for 16 different EPO glycoforms. The black circles represent the individual values of three replicas, the vertical lines the averages and the colored rectangles the standard deviations. The panels from left to right and top to bottom present the average equilibrated backbone root-mean-squared-deviation $RMSD_{BB}$, the radius of gyration r_{gyr}, the number of hydrogen bonds between protein and glycan $n_{H\text{-}bonds}$, the total solvent accessible surface area of the glycoprotein SASA, the lateral diffusion coefficient of the glycoprotein D_{trans} and the dipole moment of the glycoprotein M.

The radius of gyration, which is a measure of the effective molecular size in solvation, increases from 1.5 nm for the non-glycosylated EPO to 1.7 nm for the fully glycosylated model. O-glycosylation does not significantly increase the radius of the protein in solution. This is indicative of a glycan orientation remarkably close to the protein surface. Interestingly, N-glycosylation at the ASN38 position appears to have the strongest effect as it leads to increased radii as compared to its isoform group, i.e. models with same number of glycosylations but at different sites. This indicates a more extended conformation of the glycan at ASN38 position as compared to the ASN24 and ASN83 positions.

From **Figure 4.3**, it becomes clear that the number of intramolecular hydrogen bonds is determined by the number of glycans rather than their positions. Every N-glycosylation adds between 5 and 10 more hydrogen bonds, which are found both between the glycans and between the glycans and the protein. The number of hydrogen bonds upon ASN38 glycosylation is not significantly lower than for its isoforms, leading to the conclusion that the conformational differences are not governed by the formation/absence of hydrogen bonding.

The SASAs are reproducible in all the three replica and critically dependent on the number of glycans added but less so on the site of glycosylation. A first, N-glycosylation is most prominent and increases the SASA by around 20 nm^2 to the proteins SASA starting from 100 nm^2 in the non-glycosylated form. O-glycosylation increases the molecular SASA by an additional 5 to 10 nm^2. The second N-glycosylation only adds around 15 nm^2, indicating only weak interactions between neighboring glycans. The fully glycosylated protein has a SASA of 165 nm^2.

Upon glycosylation, the translational diffusion coefficient decreases with increasing molecular weight due to glycosylation. This decrease depends more on the number of glycosylations than on a particular site. The computed diffusion coefficient of non-glycosylated EPO is around 2.5 x 10^6 cm^2/s, which is in good agreement with other proteins of similar size. It decreases to 1.7 x 10^6 cm^2/s with one N-glycosylation, and to 1.2 and 1 for two, respectively three N-glycans. O-glycosylation does not seem to influence the protein diffusion coefficient.

The electrical dipole moment on the other hand is clearly dependent on the position of glycosylation. Whereas an N-glycan at the ASN24 position does not alter the moment, an N-ylycan at the ASN83 position almost increases the dipole moment by a factor of two (by 200 D). In most models, the effect of the O-glycan is negligible. An exception is the fully glycosylated form, where it reduces the moment by 70 D.

To summarize the above observation, protein RMSD, number of hydrogen bonds, SASA, and translational diffusion are rather determined by the number of glycosylations and not the position.

Only the radius of gyration and the electrical dipole moment are significantly glycosylation site-dependent. Apart from the diffusion coefficient, the standard deviation between the three replicates indicates the how much that property is based on the conformational flexibility of the glycans. This demonstrates the complexity of glycan torsional flexibility and the challenge of a full conformational space sampling. We note in passing, that all of the aforementioned properties influence the bio- and physicochemical behavior of the glycosylated protein such as aggregation, solubility, receptor-affinity and kinetics of binding as well as complex formation.

4.3.2 Intramolecular protein-glycan interactions and effect on protein structure

We have shown how N- and O-glycosylation alters the global protein properties. In this section, the localized effects of the glycan on certain amino acids are discussed. The results of all 16 simulations are summarized in the single figure **Figure 4.4**, which visualizes the effect of glycan contacts on residual RMSF and relative SASA. The backbone RMSF is altered in a range from -0.1 to 0.3 nm. Interestingly, residues that show a high increase in RMSF fo more than 0.5 nm never show any contact with the carbohydrate. In fact, further investigation revealed that these residues are mostly N- and C-terminal and thus naturally more flexible. Looking at the residues with a decreased RMSF, we note an accumulation of frequently contacted residues. However, most residues can be found close to an RMSF change of 0, independent of glycan contact frequency. As compared to the non-glycosylated form, the difference in relative SASA ranges between -0.4 and 0.3, where most of the residues are normally distributed around 0. Only one isle of residues is eye-strikingly off-centered. Namely the residues with a SASA reduction of more than -0.3, and a high glycan contact frequency. Unsurprisingly, these residues correspond to the glycan attachment sites ASN24, ASN38, ASN83 and SER126 and neighboring residues. This leads to the conclusion, that at least in the case of EPO and most likely also for other globular proteins, glycosylation does not alter solvent accessibility and fluctuation of residues unless they are close to the glycosylation root. In this case, the solvent accessibility and RMSF are decreased.

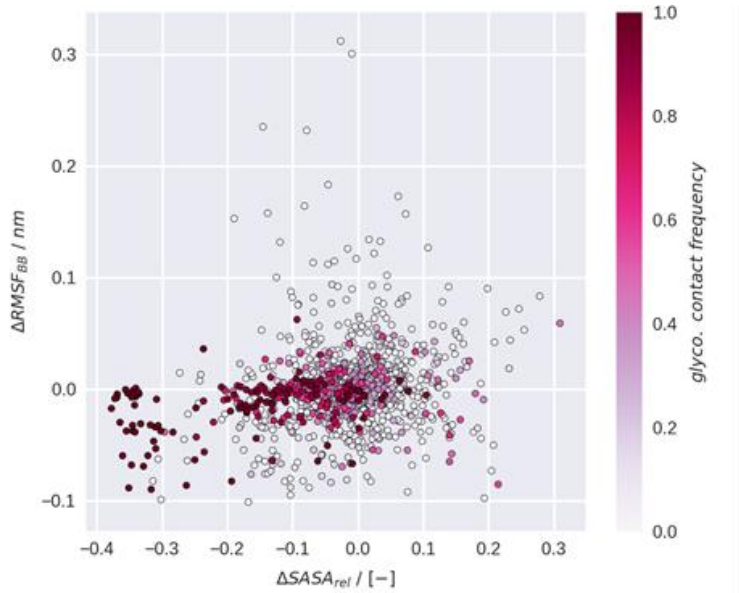

Figure 4.4: Effect of glycan contacts on residual RMSF and relative SASA relative to non-glycosylated EPO. The scattered points represent the amino acids pooled from all the 15 simulations including glycans (averages from the three replica). The abscissa is the difference in relative SASA as compared to the non-glycosylated EPO. The ordinate is the difference in RMSF as compared to the non-glycosylated EPO. Every point is colored according to the contact frequency of the amino acid with any glycan residue.

Even though protein-glycan interactions did not introduce drastic changes to structure and dynamics of the amino acids, the interaction partners on the carbohydrate site were further itemized. The number of contacts of each glycan residue with the protein were monitored and classified by amino acid residue types (**Figure 4.5**). The total number of contacts and the residue types are largely consistent among the glycosylation sites with only minor differences. The most contacts are mediated by the glycan core residues N-acetylglucosamine (GlcNac) and fucose (Fuc). The interaction residue types resemble the local protein surface composition. The number of glycan-protein interactions is highly decreased for the second GlcNac, the b-mannose (bMan) as well as both the a-mannoses (aMan). Only the sialic acid (Neu5Ac) of the 3-arm and the galactose (Gal) and Neu5Ac of the 6-arm show frequent contacts with the proteins. Thus, in most instances, the 6-arm folds towards the protein, whereas the 3-arm remains in solvation. The 6-arm contributes about 2000 contacts throughout the trajectories - a quarter of the number of contacts of the core GlcNac, which is in contact with protein all the time. Thus, the protein-glycan interaction between the two terminal residues of the 6-arm is significant. Surprisingly, the interactions are not only of the charged type between acidic Neu5Ac and basic protein residues but also between polar and hydrophobic residues. The composition of the interactions is quite similar also for the terminal glycan-protein interactions. Systematically, the number of contacts mediated by the glycan at the 38 position is slightly larger as compared to the other two sites. Additionally, only the glycan at the 24 position is enabled to form persistent interactions via the 3-arm. However, in total the differences between the various glycosylation sites are inconspicuous.

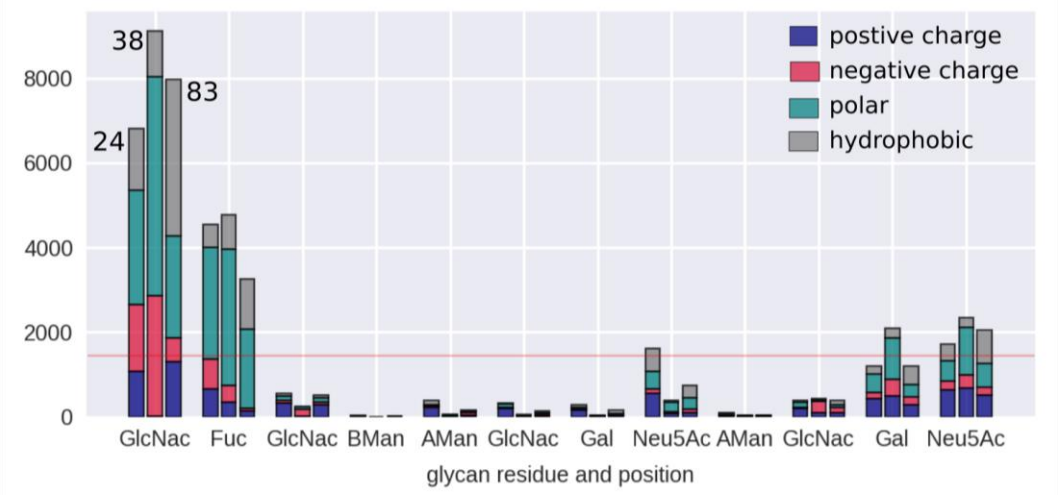

Figure 4.5: Amino acid type composition of residues in contact with the glycan. The stacked bars represent ASN24, ASN38 and ASN83, distinguished either by core (GlcNac, Fuc, GlcNac, Bman) and arm (3-arm: AMan, GlcNac, Gal, Neu5Ac and 6-arm: Aman, GlcNac, Gal, Neu5Ac) domain of sugar moieties. Results are averaged over all simulations.

4.3.3 Conformational analysis

Even though the site-specific differences in the intramolecular protein-glycan interactions are small, significant contacts could be identified, especially by the terminal Gal and Neu5Ac residues of the 6-arm. It thus is of interest, if and how such interactions interfere with the conformational space of the glycans. This is a tedious task though, because glycans are considered highly flexible and the mathematical descriptions affords many variables. In the case of the herein examined complex N-glycans, 29 torsion angles are necessary to fully describe their conformational state (**Figure 4.6 A**). Three of them belong to the root connection of GlcNac to asparagine, 26 are glycosidic bond torsion. The three root bonds are conformationally flexible and each exhibit two states. Thus, a major contributor to the flexibility of a glycosylation arrives from the glycan-protein linkage alone. The glycosidic linkages are in fact broadly rigid and many of them only populate a single conformational state with a low conformational variance. The highest contributors to the conformational variance are the linkages to the carbon 6 as located at the core fucose, the mannose of the second antenna (6-arm) and the sialic acid residues. Such bonds were found to exhibit up to three conformational populations (**Figure 4.6 B**). While these bonds contribute to large amounts of conformational variance, they are easy to identify and to distinguish. In other bonds, the torsion angles are only slightly distorted and dragged in one or the other direction. In the histograms, this shows as two closely overlapping or one broad peak.

All of the above is reflected by the circular variance (**Figure 4.6 C**) which can take values between 0 and 1. For the torsion angle distributions, three classes of values are identified: I. large circular variance with a value above 0.2, which corresponds to two or three clearly distinct populations; II. Slightly increased values between 0.05 and 0.2 that indicate broadened or heavily overlapping distributions, and III. Low values below 0.05 to show single populations. Thus, when it comes to

the identification and classification of conformations, the bond torsion angles with large circular variance will be more important than the lower ones. Or, in other words, some bond torsion angles are negligible in the conformational analysis because they are constant. Additionally, the analysis of the pairwise circular correlation matrix (**Figure 4.6 D-E**) reveals a low degree of statistical interdependence. The correlation coefficients are largely close to zero, and correlation is only reported for consecutive torsion angles, that is angles of same glycosidic bond.

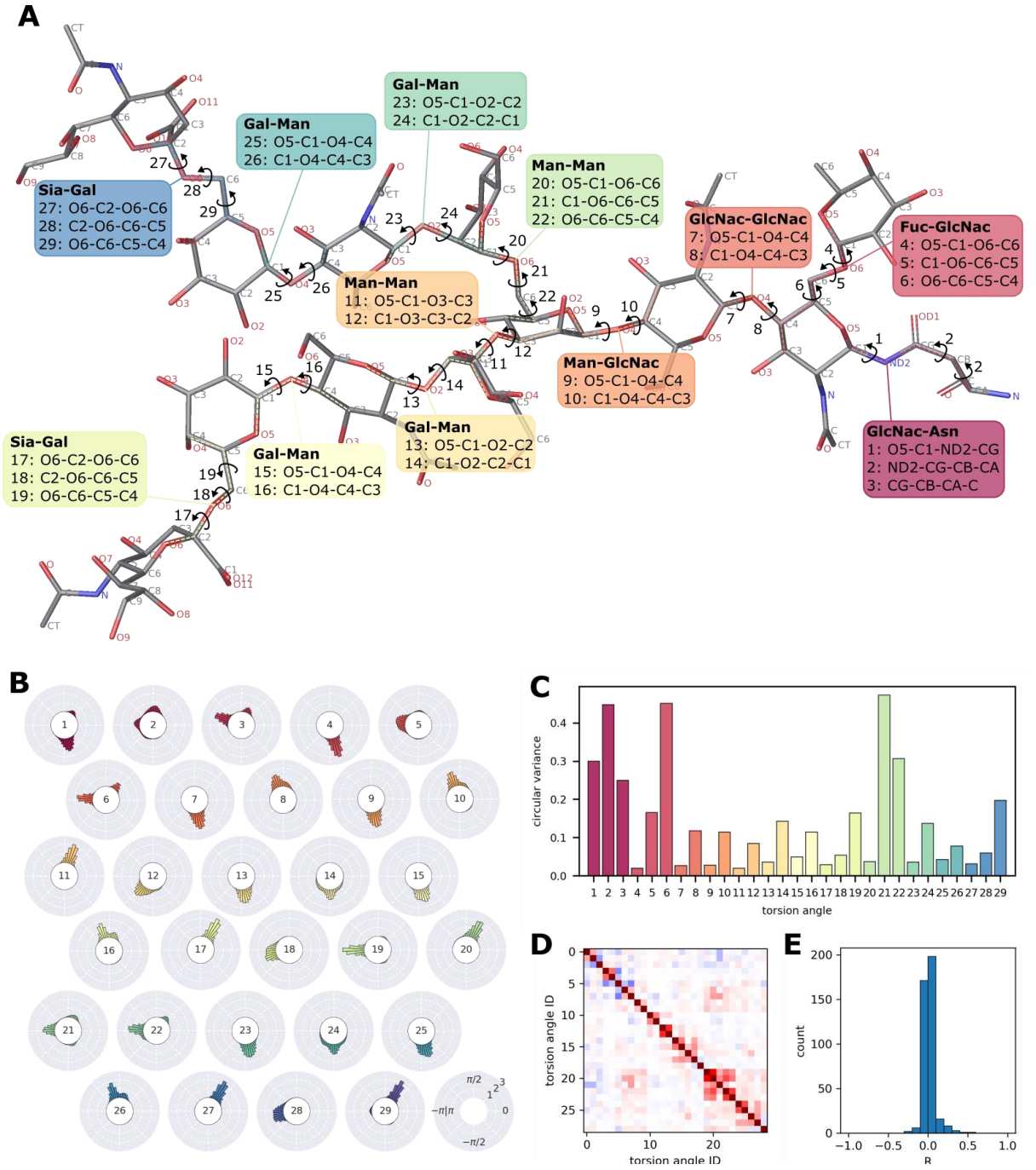

Figure 4.6: Torsion angle sampling of the N-glycans. A) Atom numbering and annotated investigated torsion angles. B). Histograms of all sampled torsion angles for each bond. Histograms are wrapped around a circle to emphasize on their periodicity. Axes tick labels are only drawn on the last panel for clarity and are identical for all histograms. C and D) Circular correlation coefficients between all torsion angles and their distribution. E) Circular variance of the different glycan torsion angles.

4.3.4 Conformational embedded clustering

To discriminate different conformational states, the torsion angles were utilized as features for a density-based clustering method. As the numbers of features is high, and the data is constrained to a circular manifold, the clustering method was benchmarked using an artificial dataset. This dataset is supposed to resemble the real data. Thus, an assumption for the origin of the real data was thought out. The basic idea was, that a molecular conformation corresponds to an energy minimum in a high dimensional energy landscape. If the conformation changes in any of the dimension, its energy increases. Thus, the closer the conformation is to the minimum energy conformation, the more likely it is. This way, it can be imagined that conformations of a molecule are drawn from high-dimensional normal probability distribution functions, in which the marginal distributions correspond to single features, such as torsion angles. The dimensions (features) might be pairwise correlated for example due to favorable non-bonded interactions such as hydrogen bonds or steric hindrance. Additionally, it is possible that certain features have multiple local energy minima so that one multi-dimensional Gaussian is not sufficient to describe the whole conformational space. Then, a mixture of such Gaussians must be employed to model all the different conformations. Here, mixtures of increasing numbers of Gaussians and dimensions were generated to sample large artificial dataset to mimic molecular conformations. These datasets were subjected to a combination of density-based clustering and dimensionality reduction by means of nonlinear embedding. For the clustering, different distances were probed.

Briefly, it could be shown that a combination of an initial UMAP embedding, combined with hierarchical density-based clustering is well suitable to re-separate up to 265 Gaussians as long as they differ in at least one of 32 dimensions (**Figure 4.7**). In terms of accuracy, it did not make a great difference, if cos-sin transformation or a periodic city-block metric was employed. However, cos-sin transformation was computationally much more efficient. While this is well in the limits of the real data, we still decided to split torsion angle space into two categories: One for the root connection between Asn and the glycan, and one for the glycan itself. This is feasible because the torsional angles do not significantly correlate with each other (**Figure 4.6 D**). Additionally, the distinction is of interest to understand if the glycan conformation or only orientation as defined by the root torsion angles is affected by the protein. Furthermore, it is supportive for the algorithm by a drastic reduction of the number of clusters. Due to the statistical independence of the different torsion angles, the probability for the joint clusters could later be calculated via simple multiplication.

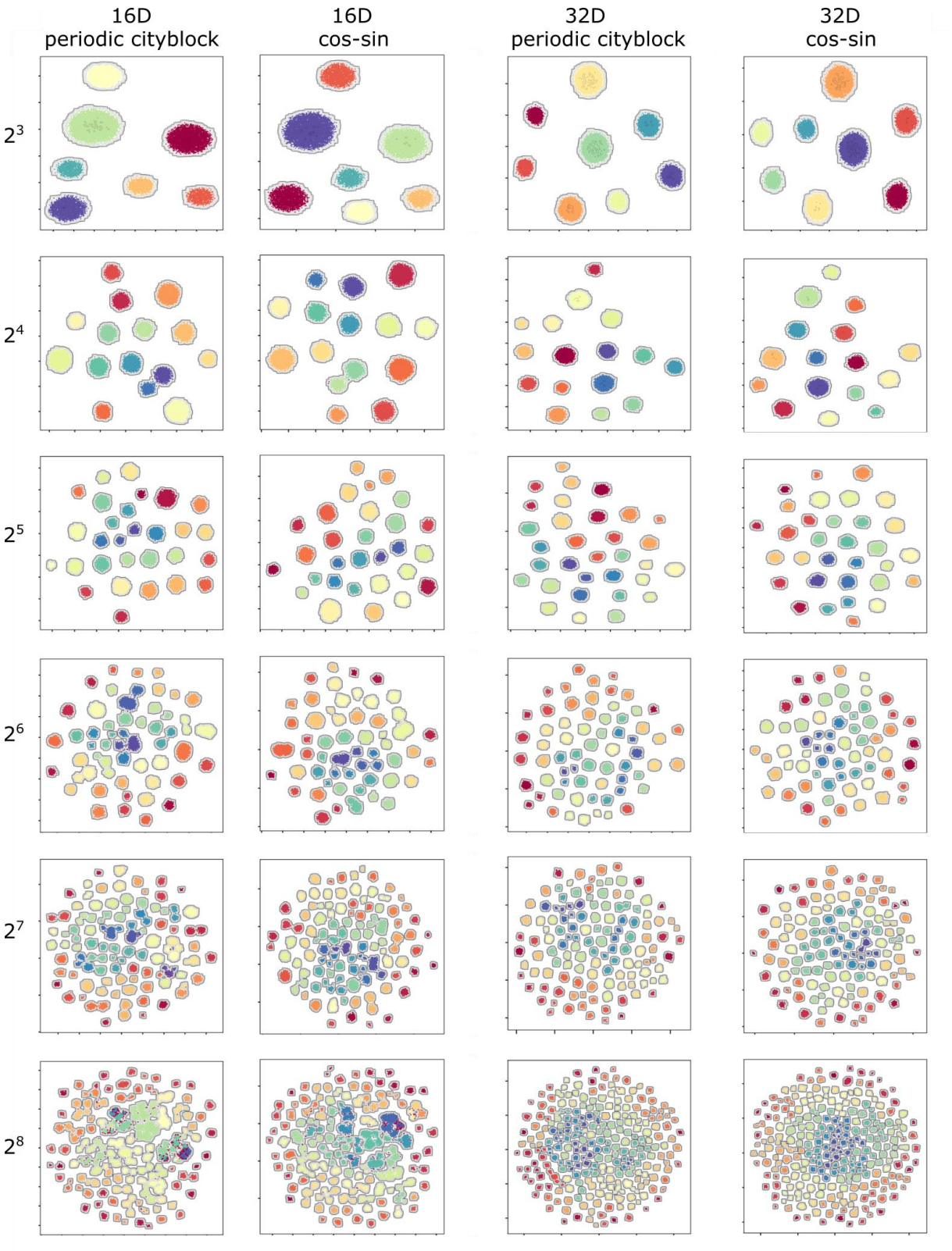

Figure 4.7: Benchmark of the lower dimensional embedding and clustering approach. The scatter plots show the 2D embedding of 20.000 16- (two left columns) or 32- (two right columns) dimensional artificial data points. The points are drawn from a mixture of 8, 16, 32, 64, 128, and 256 Gaussian probability distributions. The embedded points are clustered using HDBSAN and colored accordingly. The estimated density of the 2D embedding is drawn as black contours. The 1st and 3rd column employ a periodic Cityblock metric, whereas the 2nd and 4th column use a cosine-sine extension and Euclidean distances.

Concerning the conformational space of the asparagine sidechains which conjugate the N-glycans, nine clusters of different sizes were identified (**Figure 4.8 A**). The shape of clusters in the embedding differs widely from the artificial data. This indicates that the model to generate the artificial data was too simple and the real data are based on more complex probability density functions. However, with the proposed approach it is possible to algorithmically determine distinct conformational states. The largest clusters, i.e. the most likely conformational states correspond to cluster 4 and 3 with each about 30% contribution. Clusters 1 and 0 each contribute 10% to the total conformational space. The other clusters are significantly smaller.

The initial question was how the conformational clusters distribute over the three glycosylation sites (**Figure 4.8 B**). The most conformational variability is reported at Asn24, which adopts conformations from five of the clusters (0, 1, 2, 3, 5). Asn38 adopts only the one exclusive conformation of cluster 4. Asn83 adopts similar conformations as Asn24 yet with a different composition. Approximately, six conformational states are visited with some exclusive to certain glycosylation sites. The conformation of the glycosylation root has tremendous effects on the overall shape of the glycan because it determines its orientation. Thus, the apparent flexibility of the glycan is heavily influenced by the asparagine-GlcNac linkage. The root connection at Asn38 is the most rigid, which is reflected by the high propensity of the glycan to interact with the protein.

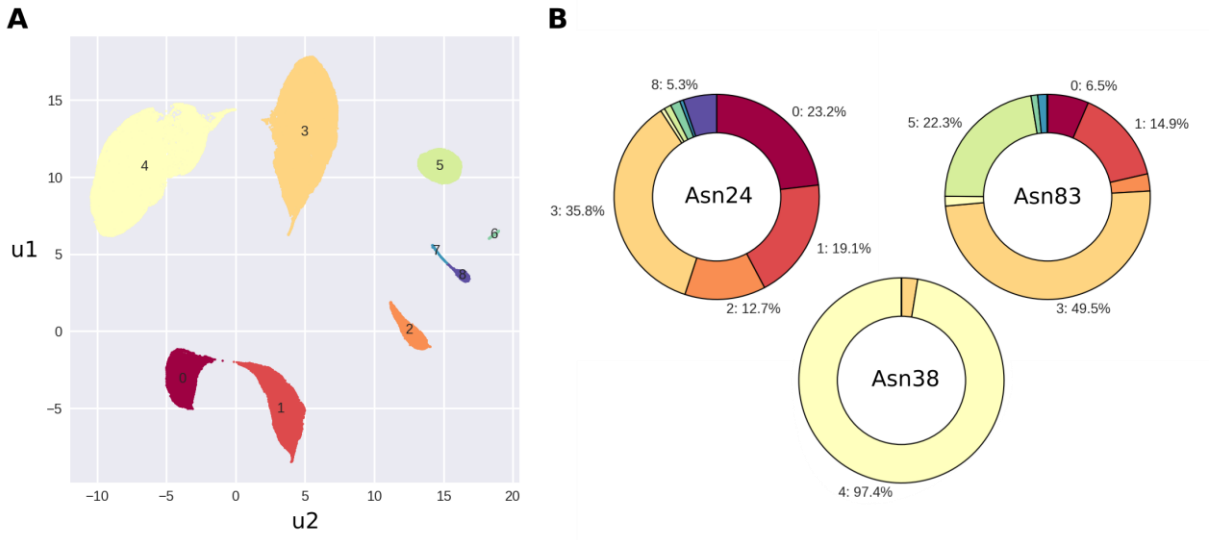

Figure 4.8: Clustering of the Asn-Glycan torsion angles within the 2D embedded space. A) UMAP 2D embedding of the three bond torsion angles between Asn and glycan. The points are colored according to HDBSCAN clustering. The cluster ID is plotted at the average cluster coordinates. B) Cluster contribution at the different glycosylation sites.

The molecular representations of conformational states as distinguished by the clustering prove the quality of the method to identify and classify molecular conformations (**Figure 4.9**). Additionally, they highlight the flexibility induced by the Asn sidechain which incorporates two rotatable bonds. Another rotatable bond appears through the linkage with GlcNac. One key feature that is common among most of the clusters is that the glycan root bents to engage in protein-glycan interaction via the fucose residue on the one site or the acetylamino-group on the other. The direction of bending depends on the local protein environment. A fully upright and well solvated conformations is also possible (**Figure 4.9**, cluster 3), however fewer variability of such conformations exists.

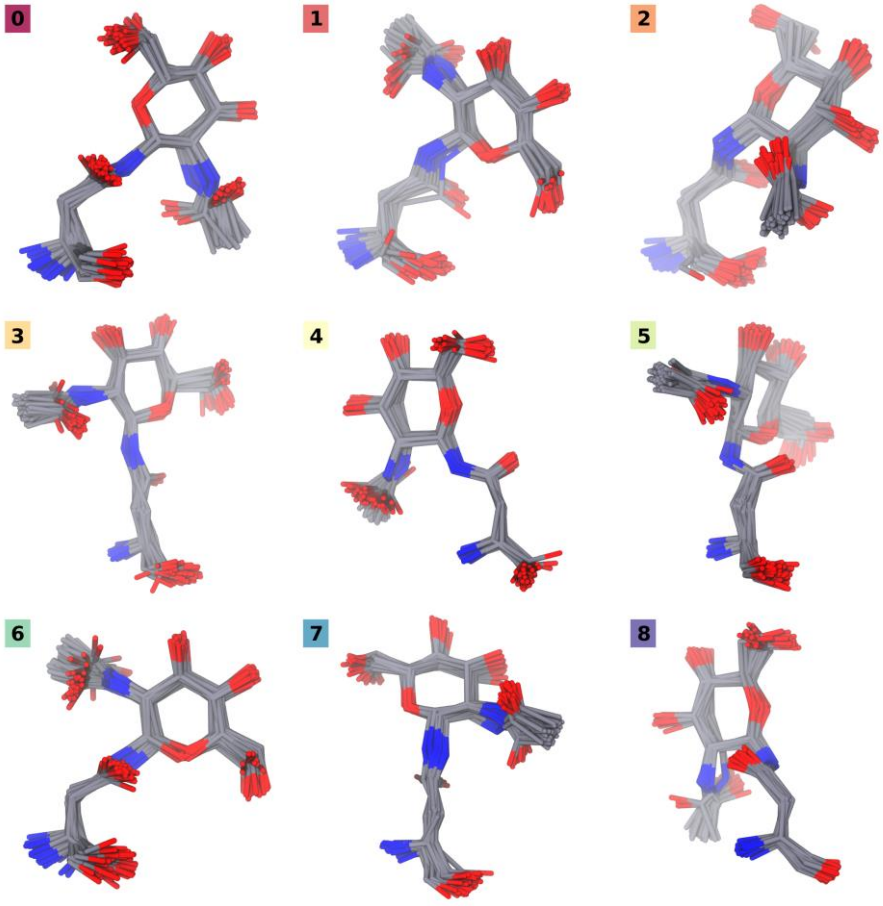

Figure 4.9: Representative conformations of the glycosylation root. The 50 best cluster representatives are overlaid in licorice representation. The Asn backbone is oriented similarly for all clusters.

As the glycosylation roots showed significant differences in their conformational space, it is of interest if the glycan itself is similarly affected by the protein conjugation site. In the UMAP embedding 77 clusters were identified via HDBSCAN clustering (**Figure 4.10 A**). The largest conformational clusters correspond to the numbers 3, 13, 20, 54, 67, 74 and 75. All the larger clusters are present at all glycosylation sites (**Figure 4.10 B**). The contribution of each cluster to the whole conformational space is similar for the different sites. Only cluster 3 and cluster 75 at the Asn38 site are larger as compared to Asn24, and Asn83. On the other hand, the number of sampled conformations within cluster 54 is decreased at the 38 position. However, in general, the differences are small.

The cluster representative conformations are identified as the conformation which is closest to the cluster arithmetic mean (medoid). Comparing the representatives of the 7 largest clusters reveals that the major source for variability are the two arms (**Figure 4.11**). Here, the arms themselves can fold back toward the glycosylation root to form a parallel arrangement with the core GlcNac residues. Only one arm can fold down at the same time and this is preferentially the 6-arm (Cluster 54). Additionally, the arm itself can bend around an "elbow" composed of GlcNac, Gal, and Neu5Ac. Conformations with a bent elbow are most likely, whereas an extended elbow is the exception. In cluster 3, the 6-arm is bent down towards the protein and the 3-arm is bent upwards towards the solvent. Both elbows are in folded conformation. Cluster 13 has both arms in upward conformation and both elbows bent. Cluster 20 is similar to cluster 3 yet with different rotations around the elbows. In cluster 67, the 3-arm is facing upwards and the 6-arm takes a mediocre orientation between down- and upfolded. Both elbows are bent. Cluster 75 is the exception were the 3-arm is downfolded and the 6-arm is oriented towards the solvent. Cluster 75 is similar to cluster 13 but with a clearly more extended 6-arm.

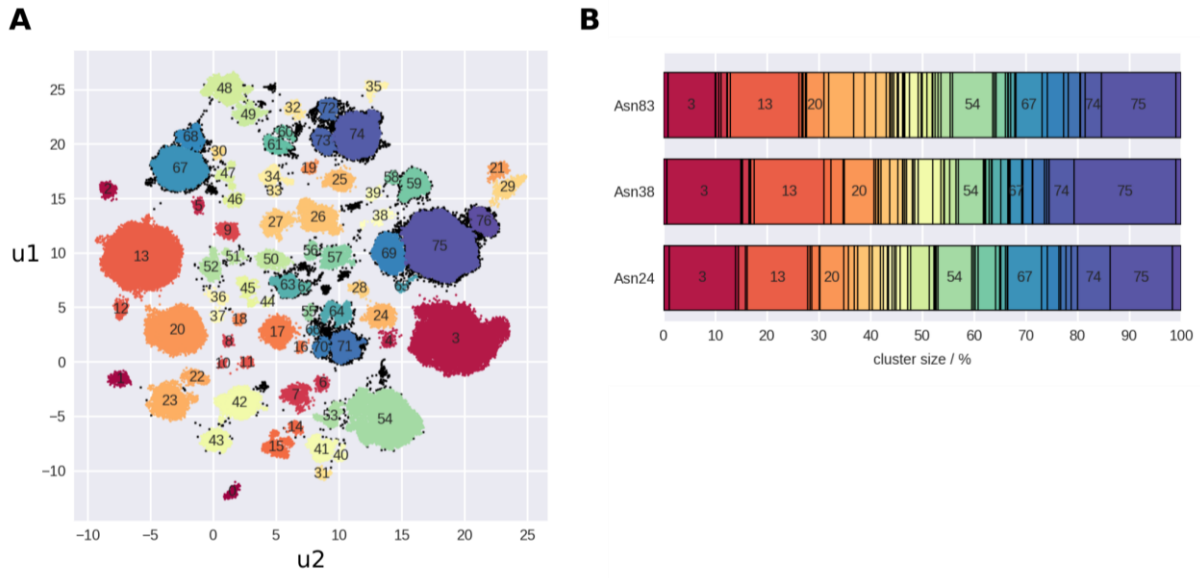

Figure 4.10: Clustering of glycan conformations. A) 2D UMAP embedding of the glycosidic bond torsion anlges of the N-glycans pooled from all replicates, glycoforms and sites. The points are colored according to HDBSCAN clustering in the embedded space. Black points mark noise. B) Cluster contribution at the three different glycosylation sites. Clusters with a contribution of more than 5% are annotated.

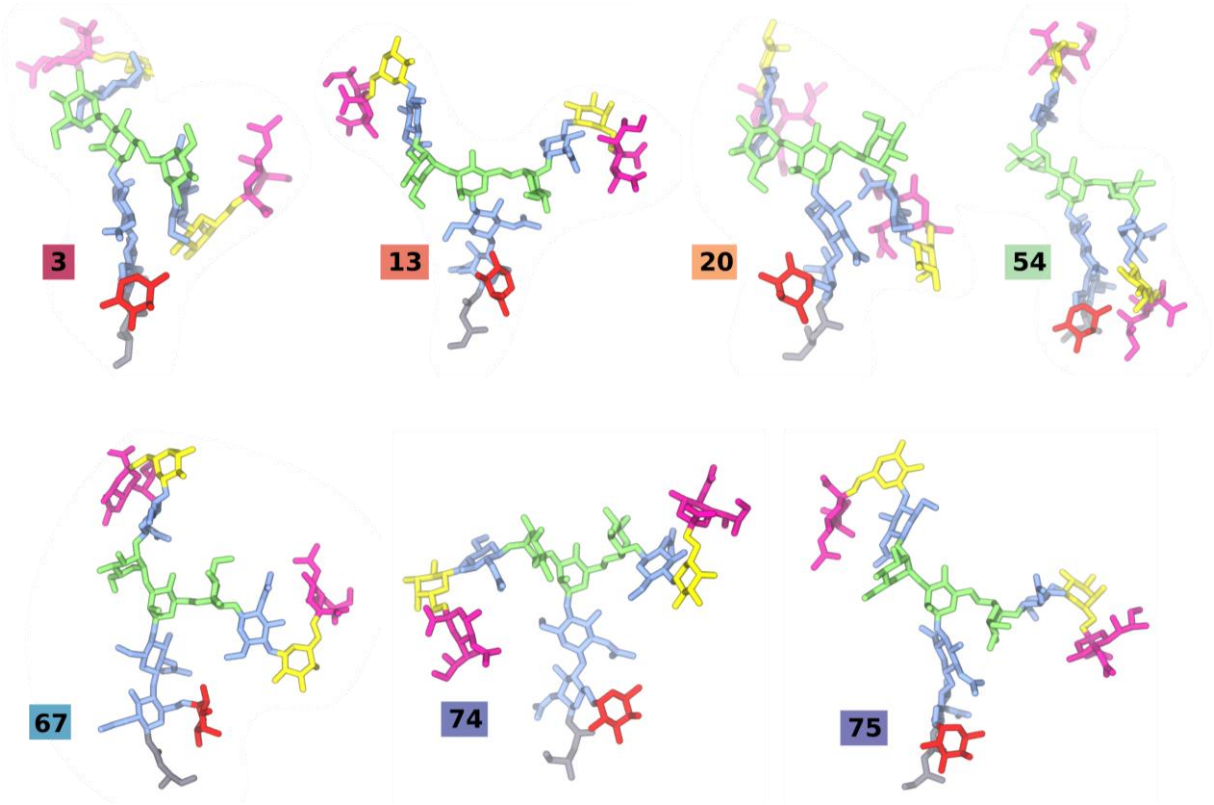

Figure 4.11: Representative glycan conformations of the 7 largest clusters. The glycan is shown in licorice representation with the residue type color coding corresponding to the SNFG nomenclature (Fuc: red, GlcNac: blue, Man: green, Gal: yellow, Neu5Ac: magenta) Asn is always at the bottom, the central branching mannose is in the center of the view and the 6-arm is oriented to the right.

It is interesting to compare the molecular conformations with the relative positions of their clusters in the 2D embedding. Visually similar conformations are mapped to much different areas in the embedding because they do not share neighbors. However, a large distance in the embedding does not necessarily mean a large conformational difference, but rather a large energy barrier between conformations because transitions were not sampled. Additionally, it must be noted that the high dimensional local density of the points is not preserved. Instead, points with high likeliness to be neighbors are embedded at a predefined minimum distance. Thus, a cloud of nearly identical points in the high dimensional space would not stack on top of each other in the embedding but rather form a large, extended cluster. To further improve the information gained from the clustering, the next logical step would be to model a Markov state model based on the transition frequencies between the clusters. This however requires even more sampling and does not reveal much more information regarding the conformational space.

4.4 Conclusion

The site-specific, mutual effects of protein-glycan interactions were analyzed in unprecedented detail using *in silico* methods. Global glycoprotein properties were generally more affected by the number of glycosylations than the glycosylation sites. Only in the case of gyration radius and dipole moment, the glycosylation site was of similar significance. The overall effect of the glycosylation was most notable on the SASA, lateral diffusion and dipole moment. Full glycosylation increased the SASA by 60%. The diffusion was reduced by a factor of 3 and the dipole moment was doubled. Interestingly, the O-glycosylation is able to significantly reverse the effect of N-glycosylation on the dipole moment by up to 20%. The RMSD of the protein is not significantly or systematically touched by the introduction of glycans. The same was observed for differences of RMSF and SASA of protein residues which interact with glycan. Only the few residues, to which the glycosylation is conjugated to, are substantially effect. For other residues, the protein-glycan interactions are short-lived. This is underlined by the observation that the major contributors to the protein-glycan contacts are the root GlcNac and Fuc. Further interactions are only mediated by the terminal Gal and Neu5Ac of the 6-arm and to lesser extent also by the 3-arm. The ratio between root and arm interactions is about 4:1. The glycosylation site where the most protein interactions were monitored is Asn38. The small amount of protein-glycan interactions as well as the undisturbed protein lead to the question if the glycan is still conformational affected by the local protein environment.

Thus, a novel conformational clustering workflow based on a lower dimensional embedding of the glycoside torsion angles and hierarchical density-based clustering was composed and proved its feasibility on an artificial dataset. The workflow revealed an expectedly large conformational space with 77 distinct clusters. The contributions of each cluster to the local, site-specific conformational

space were broadly similar. We conclude that the conformational space of the complex N-glycan is only marginally influenced by protein-glycan interactions. Oppositely, and identified with the same method, the conformations of the connection between Asn and the glycan differs heavily from site to site. Nine conformational states were identified, two of them exclusive to certain sites. This insight will simplify further conformational studies on glycosylation because such differences in the glycosylation site-dependent behavior might be general. Hence, it is recommended to start conformational analysis of the glycan with the sidechain torsions of the Asn it is conjugated to. The workflow produced exceptionally well separated clusters in a short computation time and should generally be employed for clustering of high-dimensional torsion angle data.

5 SITE SPECIFICITY OF ASPARAGINE DEAMIDATION

Besides the already covered post-translational modification of ubiquitination and glycosylation, which afford enzymes for their conjugation, deamidation of asparagine and glutamine residues appears spontaneously. It is considered the main driver for *in vitro* protein degradation and short shelf lives of protein drugs. Interestingly, only few residues in highly specific protein microenvironments show fast deamidation. Understanding and prediction of such deamidation sites is of tremendous interest for highly value pharmaceutical compounds such as antibodies. Interestingly, some pathogenic proteins undergo deamidation to escape immune response. For the GII.4 strain Norovirus protruding domain of the VP1 protein, fast deamidation was reported by Mallagaray and coworkers using NMR spectroscopic methods. A rationalization for these observations in regard of sequence- and structure-based prediction methods was not possible, though. In this chapter, extensive molecular dynamics sampling and thorough analysis of inter- and intramolecular interactions in proximity to potential deamidation sites were conducted. The data allow the conclusion that favorable attack geometries are common among solvent exposed asparagine residues and are insufficient predictors. Instead, backbone hydrogen acidity and peptide bond deformation appear to be the discriminating quantities. The NMR and chromatography experiments of the full size proteins were conducted by R. Creutznacher from Lubeck University. Additioanlly, M. Schubert from the University of Salzburg synthesized peptides of identical primary sequence as the Norovirus deamidation site and observed them over weeks with NMR spectroscopy.

5.1 Introduction

Deamidation is a spontaneous posttranslational modification of the protein backbone, in which the asparagine sidechain carbonyl of asparagine (or glutamine) undergoes an intramolecular reaction with the backbone nitrogen of the succeeding residue [315]. Thereby, asparagine (glutamine) is converted to aspartate or iso-aspartate (iso-glutamate) under formation of a succinimide intermediate and dissociation of ammonia (**Figure 5.1**) [316]. Such a reaction is considered the most abundant driver of protein degradation, especially of monoclonal antibodies [317-319]. Opposite examples, where deamidation induces protein function, are the activation of a fibronectin-integrin binding site [320] or the cellular stability of the bacterial pyruvate transferase MurA [321].

Despite plenty observed instances of deamidation on the protein and peptide level, complete understanding of the factors facilitating deamidation has not been achieved. Clearly, the physicochemical microenvironment in terms of neighboring amino acids and solvent accessibility

has principal effects on site and rate of deamidation. Most importantly, the Asn side chain must be able to adopt a suitable, reactive conformation which enables a short attack distance between the nucleophile backbone nitrogen of the residue N+1 and the electrophilic sidechain carbonyl carbon [322]. Thus, that the reaction rate must be dependent on the secondary and tertiary structure of the protein. These conformational restraints are often neglected or only indirectly attributed by sequence-based prediction methods, which are mostly regression models based on experimental deamidation rates of peptides [323, 324]. Other prediction attempts explicitly incorporate protein tertiary structure using physics-based [325], or data-based [326, 327] methods, some of them even with quantum level detail [328]. Ugur et al. also employed structural dynamics to address the deamidation rates of two asparagine residues of triosephosphate isomerase [329]. However, even within two decades of deamidation prediction efforts, the array of deamidation descriptors remained largely identical and includes the type of the succeeding amino acid, solvent accessibility, nucleophilic attack distance, as well as backbone and sidechain torsion angles [322] [325, 330].

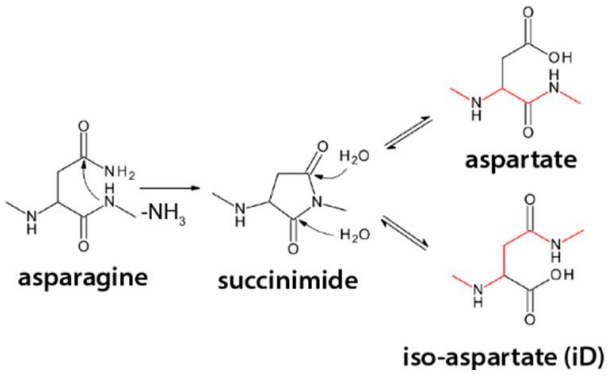

Figure 5.1: Reaction scheme for asparagine deamidation. The reaction is initiated by nucleophilic attack of N+1 backbone nitrogen at the N sidechain carbonyl carbon. Under release of ammonia, a succinimide intermediate is formed, which is reversibly converted to aspartate or iso-aspartate.

Interestingly, especially the recent work of Delmar and colleagues [327] reveals the flaw of machine-learning and structure-based prediction methods. There appear to be two distinct population of deamidation residues: one for which the regression models are predictive over a wide range of half-lives, and one that is characterized by very fast deamidation and is not well captured by the models. This leads to the assumptions, that the models are trained too much on the slower deamidation events, which might be easier to predict (large attack distance, buried residues). Nevertheless, Delmar et al reach a fantastic quantitative prediction accuracy of over 93% for their benchmark set of asparagine half-lives. However, they leave out molecular rationalizations on why especially the weights for the N+1 residue type and torsion angles are so high.

Recently, Mallagaray and colleagues discovered that the P-dimers of the GII.4 Saga Norovirus strain undergo fast, and site specific deamidation at the asparagine 373 (N373) position [213]. A molecular explanation was however not yet sought out. This residue N373 is highly conserved among GII.4 NoV strains and deamidation is also observed for P-domains of other GII.4 strains [331]. N373 is only one of an abundance of asparagine residues within the P-domains (**Figure 5.2 A**), some of which are predicted to be prone for fast deamidation (glycine as N+1 residue). However, the fact that deamidation is only observed for N373, raises the question of what makes N373 so special. Analysis of the crystal structure did not yield any exclusive topological properties of N373. Thus, we set out to perform dynamics-based analysis via extensive microsecond scale molecular dynamics (MD) simulations. Here, we describe how the combination of experimental data from protein NMR spectroscopy, ion exchange chromatography and quantitative analysis of MD trajectories lead to a rational and physics-based explanation for site-selective deamidation in GII.4 Saga and V387 NoV capsid proteins and selected point mutants.

5.1.1 NMR spectroscopy and kinetic modeling

N373 of the P-domain of GII.4 Saga undergoes fast spontaneous deamidation. N373 is located at the outward-facing part of the NoV capsid, at the tip of the P-domain homodimers (**Figure 5.2 A**). Deamidation of N373 does not result in the expected mixture of aspartate and iso-aspartate. Instead, only formation of the iso-aspartate product has been observed. In addition, the fact that the ^{15}N TROSY-HSQC NMR spectra of N373iD do not change over time suggests that it cannot interconvert into the aspartate form N373iD (**Figure 5.2 B**). Based on cation exchange chromatograms, R. Creutznacher developed a kinetic model for the deamidation of the P-domains. Of note, introduction of a dimer association and dissociation terms increased the quality of the kinetic model. Global least-square fitting of the model to the experimental data yielded a deamidation rate constant for the SAGA P-domain of 4.5 x 10^{-7} s^{-1}. Interestingly, it was also recognized that acidic buffer conditions drastically increase the half-life of N373. In order to dissect a possible influence of the primary amino acid sequence on the deamidation rate of N373 from structural, through-space effects, M. Schubert synthesized a 13-mer model peptide for GII.4 Saga P-domains comprising the entire sequence of the deamidation site loop. Notably, this peptide contains a second Asn residue allowing us to probe the selectivity for deamidation of N373 as well as the corresponding reaction kinetics. To this end, we monitored 2D NMR spectra of the peptides under the same experimental conditions as applied to the P-dimers, except for the temperature. In contrast to the P-dimers, all Asn residues in the peptides deamidate over time and both, isoAsp and Asp reaction products are detected with a ratio of ca. 4:1. We find that even at higher temperature, Asn deamidation at the position corresponding to N373 is significantly slower than in the P-dimer, i.e., 61 d at 37°C for the Saga peptide compared to 1.6 d for the protein [191].

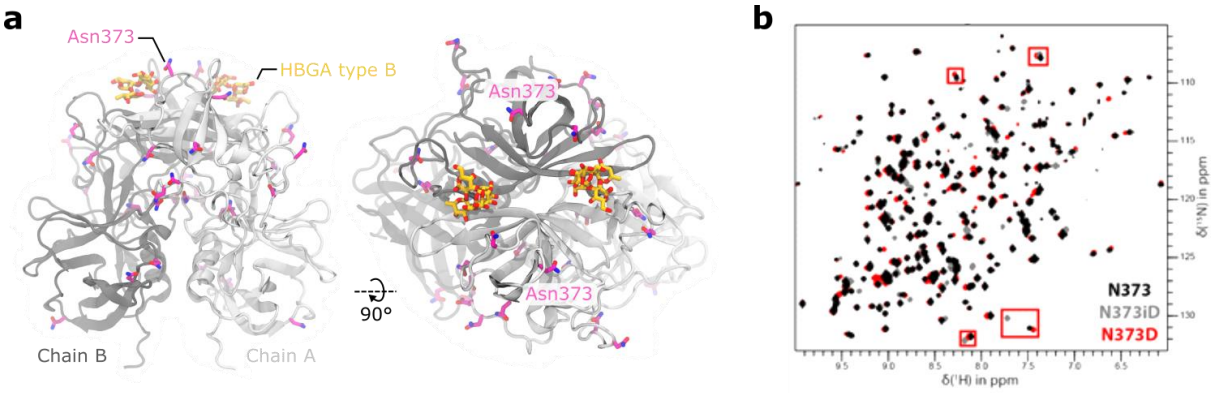

Figure 5.2: Deamidation of the SAGA P-dimer. a) Crystal structure of the SAGA P-Dimer (PDB 4X06) in ribbon representation. All asparagine residues are explicitly shown as well as the natural HBGA type b ligand. Asn373 is annotated. b) 2D 1H15N NMR spectrum with annotation of the peak shifts upon deamidation of N373 to iso-aspartate (N373iD) and aspartate (N373D).

The Asn residue at the position corresponding to N380 deamidates even slower with a half-live of 100 d. As deamidation in the model peptide is neither fast nor exclusive for N373, we conclude that fast deamidation of N373 of Saga P-dimers is primarily caused by conformational effects.

5.1.2 Deamidation of the related strain VA387

Selection pressure of the host immune system causes considerable sequence variation within the outward facing parts of the capsid NoV [332], including the loop containing N373. High conservation (64%) of N373 among GII.4 strains suggests a functional advantage of asparagine in this position. Therefore, we investigated the impact of sequence variation in neighboring positions on the deamidation behavior of a natural GII.4 NoV variant, the VA387 strain. P-domains from the Saga and VA387 strain are remarkably similar in terms of primary sequence (94% identity) and 3D structure (0.4 Å RMSD). However, some amino acid substitutions can be identified close to the critical position 373 (R297H and E372N), allowing us to study deamidation in the context of two naturally occurring protein homologs (**Figure 5.3 A**). The same changes in the NMR spectra characteristic for N373 deamidation in Saga P-domains were observed, demonstrating that site-specific deamidation is conserved among the two GII.4 NoV strains. Likewise, for both P-dimers only iso-aspartate (iD) was detected as reaction product. A model peptide of the VA387 loop containing N373 did neither reflect specific nor fast deamidation. A half-life of 27 days at 20°C was measured for the N373 VA387 P-dimers compared to 9 days at 20°C for GII.4 Saga P-dimers (**Figure 5.3 B**). To probe differences in local conformations of the loop that could account for this divergence, we determined the dissociation constant K_D for binding of methyl α-L-fucopyranoside to VA387 P-dimers. It is known that D374 is critical for binding to L-fucose containing glycans, and, therefore, such conformational changes may reflect on binding affinity. However, NMR spectroscopy yielded a dissociation constant K_D of 21 mM which is almost identical to the value previously determined for GII.4 Saga P-dimers [191].

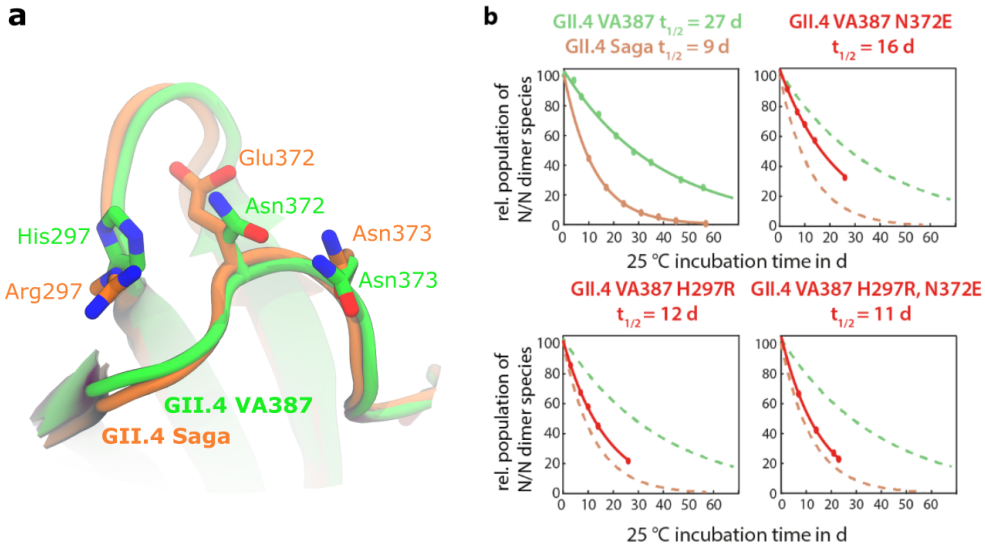

Figure 5.3: Deamidation of GII.4 VA387 P-domains. a) Structural alignment of GII.4 VA387 P-domain dimers (green) with GII.4 Saga P-domains (PDB 2OBT, 4X06). The deamidation site and nearby amino acid substitutions are highlighted. b) P-domain incubation at 25 °C and subsequent IEX chromatography yield N373 half-life times $t_{1/2}$. Deamidation of Saga P-domains is three times faster than that of VA387 P-domains. Mutating VA387 amino acids close to N373 into their Saga counterparts reveals a strong influence of R297 on the deamidation rate of N373.

This suggests that the local fucose recognition site is not affected by the deamidation. Additionally, point mutants were created to further scrutinize the cause for the observed difference in deamidation rates, which can be considered intermediates between SAGA and VA387. The two-point mutants are E372N and H297R relative to SAGA. Both mutations substantially increased the deamidation rates. Surprisingly, the H297R mutant alone almost fully restored the fast deamidation kinetics of the Saga strain, clearly indicating that deamidation of N373 is partially controlled by an interaction with a neighboring surface loop. As expected, the behavior of the VA387 H297R N372 double mutant closely resembles that of the Saga wild type protein.

5.2 Method

To understand how the experimental observations are structurally rationalized, extensive molecular dynamics of the two P-dimers was conducted and assessed via elaborate geometrical and statistical analysis.

5.2.1 Molecular dynamics sampling

Theoretical conformational sampling was achieved using explicit-solvent, full-atomistic equilibrium molecular dynamics. Two molecular systems were subjected to molecular dynamics integration: 1. the P-protein dimer from strain GII.4 SAGA, and 2. the P-protein dimer from strain VA387. For both systems, data were collected from five trajectory replica of 1 µs length each, which were individually equilibrated using different initial velocity distributions. One Saga MD 1 µs trajectory was used previously to sample protein conformations for ensemble docking to the bile acid binding grove [191].

For the SAGA P-dimer, the molecular simulation tasks were performed with GROMACS 5.1.5 [75-77, 80, 81, 309] using CHARMM36 force field parameters [83]. Modeling of the initial system was attained with CHARMM-GUI solvation builder [307] using the X-ray structure PDB 4OOX [217]. CHARMM-GUI was also used for solvation with TIP3P water and ionization to 0.15 M NaCl. The periodic simulation box was set to a cubic shape of 9.3x9.3x9.3 nm^3 volume, corresponding to 2 nm water layers in each direction around the central protein dimer. Prior to dynamics integration, the system was minimized for 5000 steps using a steepest descent algorithm. Dynamics were initiated by assigning velocities according to a Maxwell-Boltzmann distribution at 303.15 K, followed by NVT equilibration for 100 ps (time step 0.002 ps) using Nose-Hoover [95, 225, 310] temperature coupling (coupling constant 0.4 ps-1, reference temperature 303.15 k). Protein and solvent are coupled to individual baths. To relax the box volume, 100 ps of NPT (time step 0.002 ps) sampling using an isotropic Berendsen coupling [93, 94] with a reference pressure of 1 ATM and a compressibility of 4.5e-5 ATM-1 were attached. The box volume was adjusted every 0.5 ps. During minimization and equilibration, the backbone atoms were restrained by 400 kJ/mol/nm and the sidechain atoms by 40 kJ/mol/nm. For the 1 μs of unrestrained production (time step 0.002 ps), Parinello-Rahman pressure coupling [312, 313] was applied instead and the temperature coupling constant was increased to 2 ps. Snapshots were stored every 20 ps. During all the steps, covalent bonds to hydrogen atoms were constraint using LINCS [227, 228] as a solver. Coulombic interactions were computed using the PME method [97-99] and a cutoff of 1.2 nm. Van-der-Waals interactions had a cutoff 1.2 nm with a force-switch modifier starting at 1.0 nm. Center of mass movement of the whole system was removed every 100 steps.

The simulation protocol was only marginally updated for the VA387 P-dimer MD calculations. In particular, we here used GROMACS 2018.3 and discarded the NPT equilibration step because the long simulation time renders the initial few nanoseconds of box size equilibration negligible. However, we increased the initial NVT equilibration by 25.000 steps. Additionally, we applied restraints to the protein dihedrals during the equilibration (4.0 kJ/mol/deg). Also, the solvation box size was slightly smaller (9.2x9.2x9.2 nm^3). Additionally, we note that the SAGA calculations were performed on multiple CPU nodes, whereas the VA387 simulations were achieved using single CUDA GPU nodes. However, we do not expect the adjustments to affect the overall outcome of our calculations. Regarding the length of the trajectories and the substantial computational effort, we justify the adjustments by a better utilization of compute resources.

5.2.2 Trajectory analysis

Data analysis and visualization were carried out with GROMACS *tools* and the Python libraries NumPy [272, 333], MDTraj [230]and MatPlotLib [274]. The root mean squared fluctuation (RMSF)

was computed using *gmx rmsf*. Only the backbone atoms C, N, CA and O were considered. Rotational und translational motions were removed using least-square super positioning of the backbone atoms. The per-residue solvent accessible surface area (SASA) was computed with *gmx sasa* and probe radius of 0.14 nm and 24 dots. To calculate the relative surface accessibility of the Asn residues, we divided their absolute SASA by 1.95 nm², corresponding to the theoretical maximum SASA for Asn [146]. The sidechain torsion angles of the Asn residues, as well as the distances from the Cγ atoms of Asn to the backbone nitrogen atoms of the subsequent amino acids were computed with MDTraj. The Asn torsion angles are defined as: φ: C_{i-1}-N_i-Ca_i-Ci, *ψ*: N_i-Ca_i-C_i-N_{i+1}, χ_1: N-Ca-Cb-Cc, and χ_2: Ca-Cb-Cc-Od (**Figure 5.4 A**). The free energy maps were constructed from the 2D probability densities as estimated by binning the data to 100 x 100 bins of $2\pi/100$ widths. The relative free energies in units of k_BT are computed as the negative natural logarithm of the probability density. Clustering was performed in the cosine-sine feature space spanned by the transformation of the four torsion angles z(φ)=[cos (φ), sin(φ)]. Hierarchical density-based clustering was calculated using the HDBSAN [182] algorithm with cluster selection alpha value of 0.5, a minimum sample size of 100 and a minimum cluster size of 100 (other parameters were default).

The Burgi-Dunitz (BD) [334] and Flipping-Lodge (FL) [335] angles were calculated to better describe the nucleophilic attack geometry (**Figure 5.4 B**). They are based on a coordinate transformation that centers the carbonyl carbon to the origin and the carbonyl plane onto the XY-plane. Then, the BD angle is defined as the angle between the vectors connecting the carbonyl carbon with carbonyl oxygen and the carbonyl carbon with nucleophile. It can be described as the as altitude angle of the nucleophile when the electrophile (carbonyl carbon) is the reference. It is between 0 and 180°. The FL angle can be imaged as the inclination angle of the nucleophile relative to the normal of the carbonyl plane. Here, it is calculated as the pseudo-torsion angle between the carbonyl plane normal vector, the carbonyl carbon to carbonyl vector and the carbonyl carbon to nucleophile angle. It defined in a way that it is positive for a rotation towards the Cβ and negative towards Nδ2. Its range is between -180° and 180°.

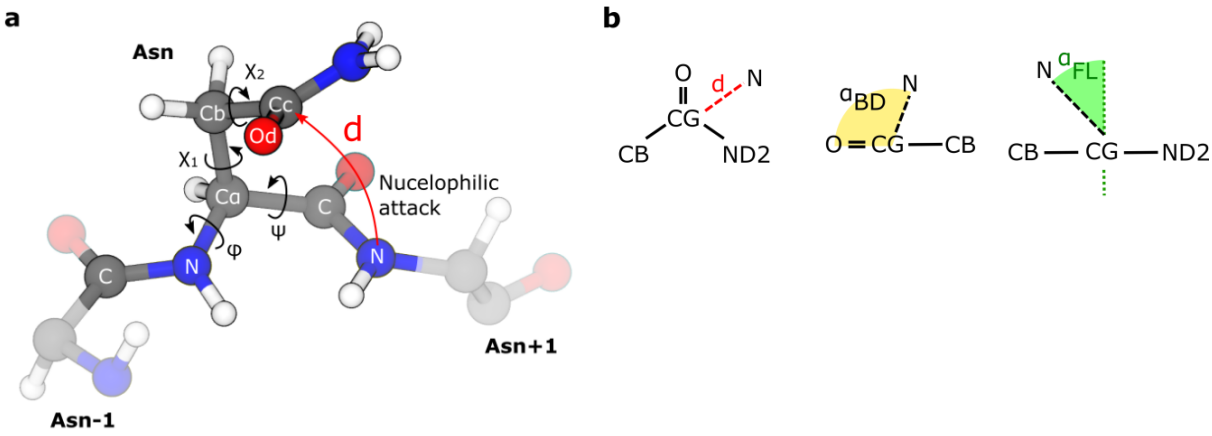

Figure 5.4 Attack geometry and key conformational descriptors. a) φ: C_{Asn-1}-N_{Asn}-$C\alpha_{Asn}$-C_{Asn}, ψ: N_{Asn}-$C\alpha_{Asn}$-C_{Asn}-N_{Asn+1}, χ_1: N_{Asn}-$C\alpha_{Asn}$-$C\beta_{Asn}$-$C\gamma_{ASN}$, χ_2: $C\alpha_{Asn}$-$C\beta_{Asn}$-$C\gamma_{ASN}$-$O\delta_{ASN}$ b) Representation of d and BD and FL angles in the CG-centered coordinate system.

5.3 Results and discussion

5.3.1 Solvent accessibility and flexibility

To extend structural information of the crystal structures to a conformational ensemble of the protein in solution, extensive molecular dynamics simulations (MD) were conducted. From the MD ensembles, initially, relative solvent accessible areas (SASA) and backbone flexibilities by means of alpha carbon root-mean squared fluctuations (RMSF) were calculated (**Table 5.1**). The RMSF values of the various Asn residues values range from 0.05 nm to 0.2 nm. The highest flexibility with an RMSF of 0.2 nm is located at Asn307, Asn309 and Asn310. Asn373 is slightly less dynamics and shares a RSMF of 0.15 with Asn412. The highest relative solvent accessibility (ranging from 0: fully buried to 1: fully exposed) is observed for Asn412 (rel. SASA = 0.8). Asn309, Asn373, Asn398 exhibit a similarly high solvent exposure of 0.6. Asn282, Asn309, Asn415 and Asn512 are slightly less, yet still well solvent accessible (rel. SASA = 0.5).

5.3.2 Conformational space and attack geometry

MD allows sampling of many conformational states. In the case of an intramolecular reaction such as the cyclization step of the deamidation, the conformation dictates the geometry of the attack trajectory. Whereas the residue conformation is sufficiently described by the backbone torsion angles φ and ψ and the side chain torsion angles χ_1 and χ_2, the attack geometry is more precisely defined by the attack distance, and the two nucleophilic attack angles α_{BD} (Bürgi-Dunitz) and α_{FL} (Flippin-Lodge). It is important to note, that backbone conformation and sidechain conformation can be mechanistically dependent and may not be interpreted separately. The attack distance and angles directly follow from the backbone and sidechain torsion angle and may be explicitly expressed using geometric transformations. Here, however, we calculated the attack trajectory directly from the Cartesian coordinates of the atoms.

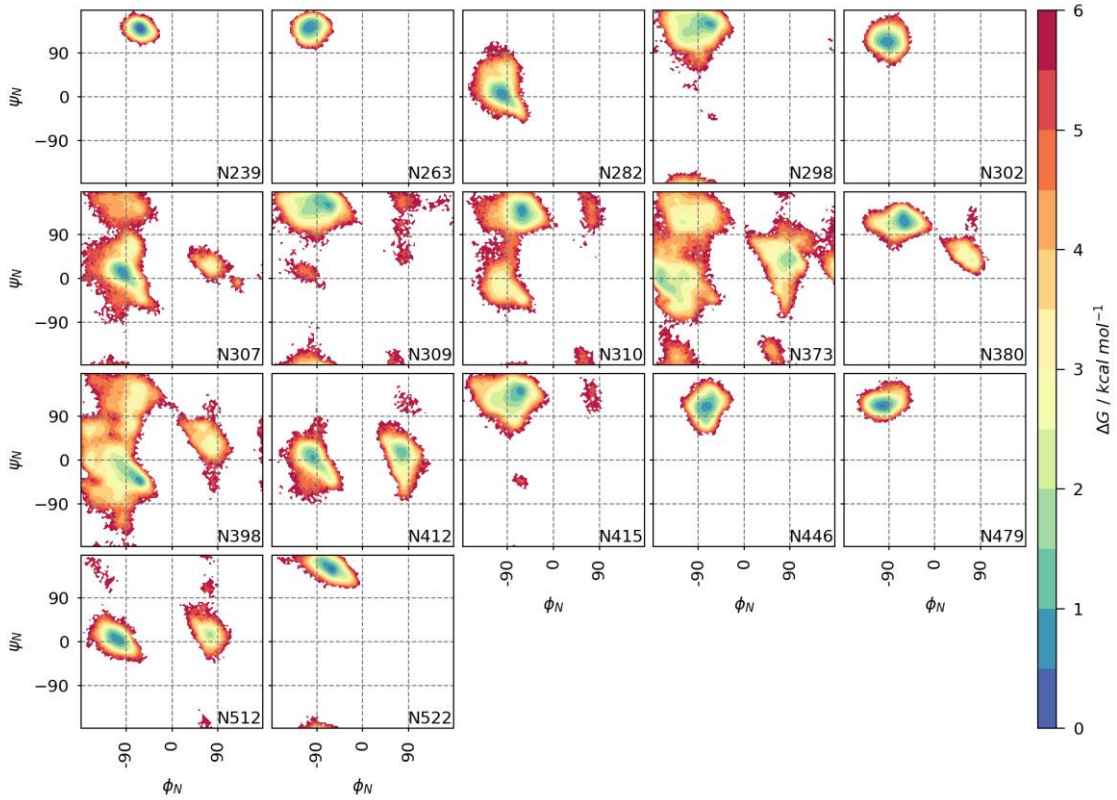

Figure 5.5: Backbone torsion φ, ψ free energy landscapes of all Asn residues of the Saga P-domains. Data was pooled from the 5x1 μs simulations and both P-domain monomers. The color scale applies to all panels.

Intuitively, Asn residues with large backbone RMSF also populate more conformations in the φ,ψ free energy landscape (**Figure 5.5**). Most of the sampled conformations can be attributed to the common Ramachadran regions, with high propensities for beta sheet like (φ, ψ = -90°, 90°) and alpha helical conformations (φ, ψ = -90°, 0°). Some residues, such as Asn373, Asn398 and Asn412, also exhibit energy minima in proximity to the left-handed helix region (φ, ψ = 45°, 45°). Interestingly, the backbone torsion free energy map of Asn373 is outstanding, and only weakly matched by the one of Asn398. Particularly, Asn373 exhibits the shallowest free energy landscape with three distinguishable minima, each belonging to the before mentioned Ramachadran regions. However, the population of the alpha-helical region is significantly shifted from φ ≈ -90° towards φ ≈ -180° with ψ remaining 0°. Hereafter, this unique backbone conformation, which is almost exclusively accessible to Asn373 will be referred to as 'anti/syn' conformation.

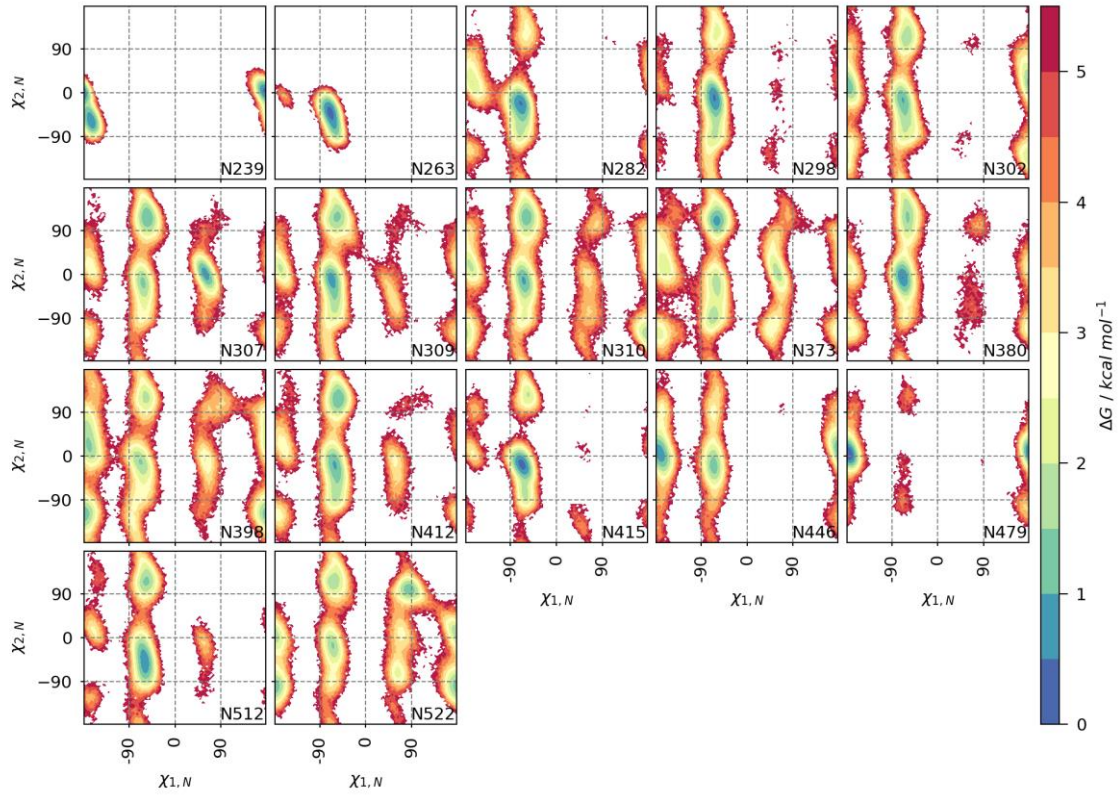

Figure 5.6: Sidechain torsion χ_1, χ_2 free energy landscapes of all Asn residues of the SAGA P-domains. Data was pooled from the 5x1 µs simulations and both P-domain monomers. The color scale applies to all panels.

Among the Saga asparagine residues, the sidechain conformational space of Asn373 is not unusual (**Figure 5.6**). All well solvent-exposed Asn sidechains exhibit significant dynamics and transition between various conformational states, which are similar among all Asn residues. The χ_1 angle has energy minima at -180°, -60° and 60°. At χ_1= -180°, χ_2 is mostly around 30° or -120°, whereas at χ_1= -60°, it is rather close to 120° or -30°. When χ_1 becomes 60°, χ_2 is mostly 0° or 90°. The energy barriers between the different rotational states of χ_2 are low ($\Delta G < 3$ kcal/mol), whereas transition states between different χ_1 angles were rarely sampled ($\Delta G > 8$ kcal/mol). The sidechain of Asn373 populates all the six minima. However, the contribution of conformations with χ_1 around 60° is comparably high and only matched by Asn307, Asn310, Asn398 and Asn522.

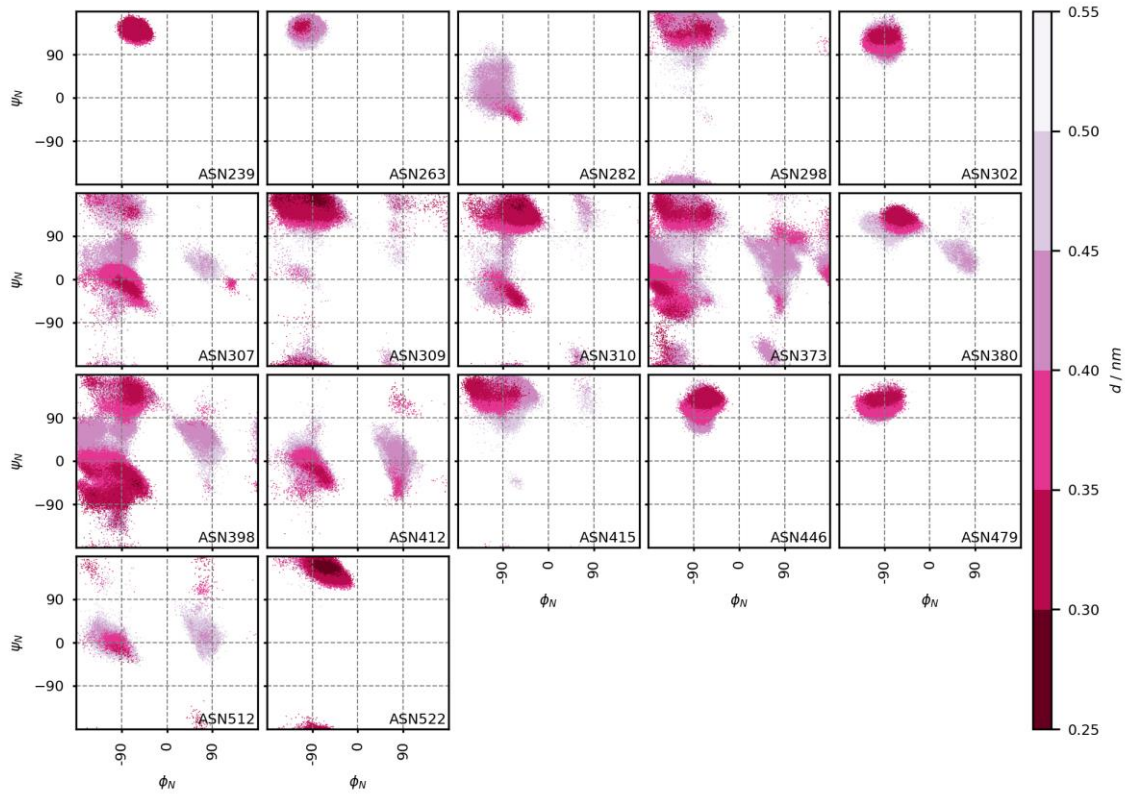

Figure 5.7: Attack distance d in dependence of backbone conformation φ, ψ. The points are ordered in a way, that the lowest distances are plotted to the top. Data was pooled from the 5x1 µs simulations and both P-domain monomers. The color scale applies to all panels.

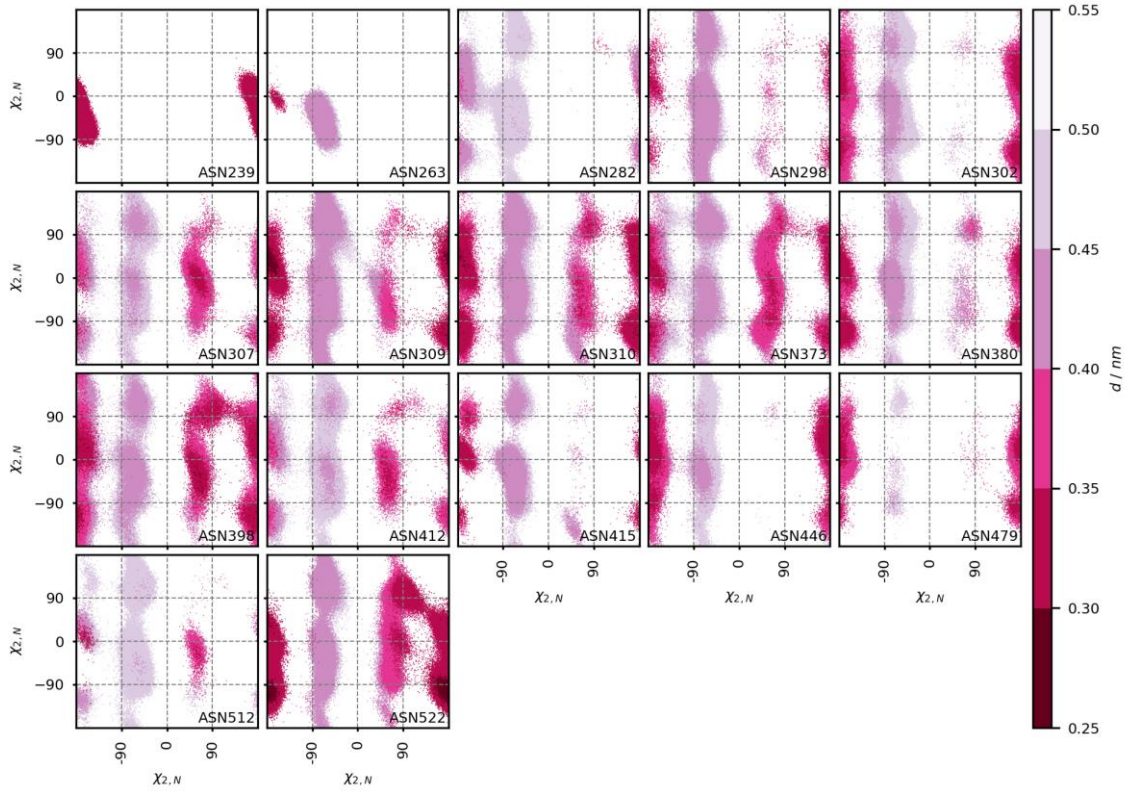

Figure 5.8: Attack distance d in dependence of sidechain conformation χ_1, χ_2. The points are ordered in a way, that the lowest distances are plotted to the top. Data was pooled from the 5x1 µs simulations and both P-domain monomers. The color scale applies to all panels.

Regarding attack distances, it is apparent that the χ_1 torsion angle is of key importance because it describes the rotation of the Ca-Cb bond and largely determines if the sidechain amide is oriented towards or away from the backbone nitrogen (**Figure 5.8**). A stark effect of the carbonyl group rotation χ_2 on the attack distance is not reported (**Figure 5.8**). Interestingly, also the backbone angles appear to have a significant effect on the attack distance (**Figure 5.7**). Within the sampled conformational space of all Saga P-dimer Asn residues, short attack distances can be achieved by two distinct sets of conformations. In the first, χ_1 is close to 180° (+/-30°) and the backbone is in a φ/ψ = -90°/120° (beta-sheet) conformation. In such an arrangement, χ_2 is often free to rotate, without having a dominant impact on the attack distance. The second set is characterized by narrow χ_1 angles between 45° and 90°, and a right-handed alpha helical conformation of the backbone (φ/ψ = -90°/0°). Again, the impact on χ_2 on the attack distance is negligible. However, χ_2 is of high importance for the reaction because it determines the attack geometry by means of the FL angle.

The distribution of attack angles is similar throughout most Asn residues (**Figure 5.9** and **Figure D.1**). Two distinct populations are present: I. a higher distance, non-reactive population (d>0.4 nm), in which the Asn carbonyl is folded away from the backbone and α_{FL} is centered around 90° and α_{BD} around 120°; and II. a close distance (d<0.4 nm), crescent-shaped population, in which the FL angle is dependent on the BD angle. When α_{BD} becomes smaller toward 45°, α_{FL} centers around 90°. This corresponds to conformations, in which the carbonyl folds over the backbone in way that the backbone N+1 nitrogen is close to the carbonyl oxygen and thus far away from an ideal attack trajectory. However, when α_{BD} increases to 90° and above, the FL angles diverges to either 0° or 180° degree. Such conformations appear highly favorable for nucleophilic attack and are accessible to a variety of Asn residues and are not restricted to ASN373.

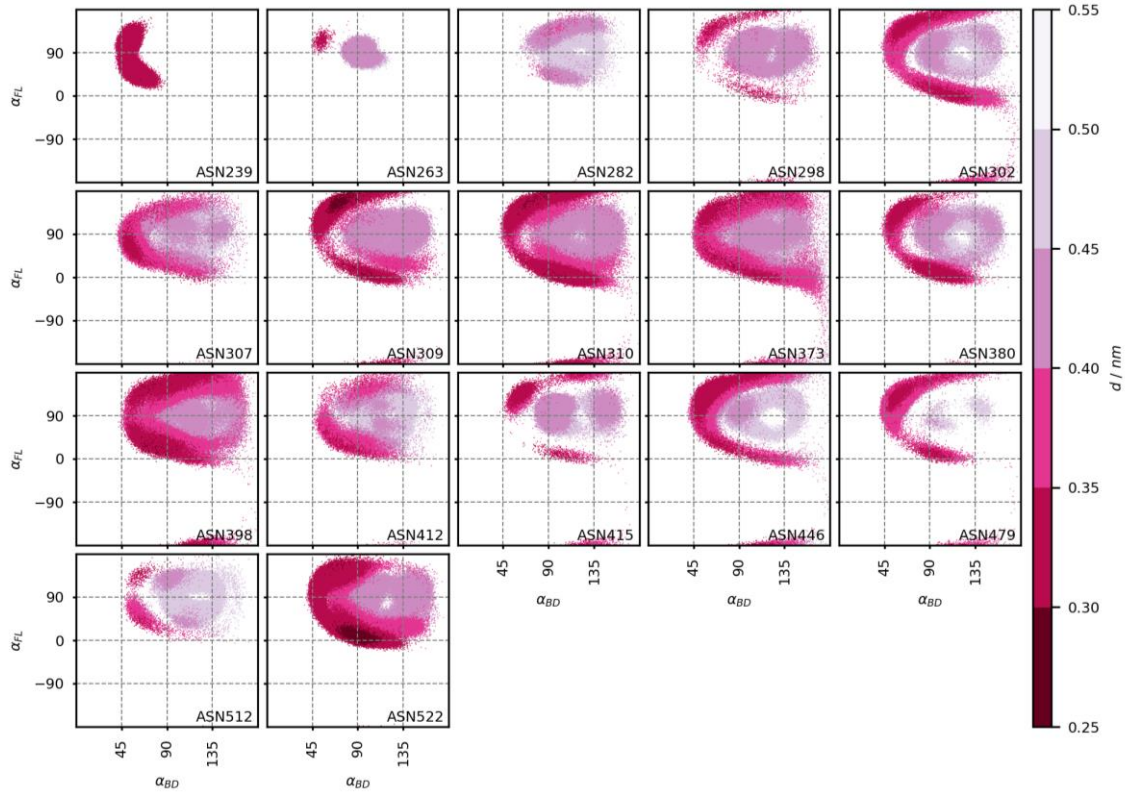

Figure 5.9: Attack distance d in dependence of the attack angles α_{BD}, α_{FL}. The points are ordered in a way, that the lowest distances are plotted to the top. Data was pooled from the 5x1 µs simulations and both P-domain monomers. The color scale applies to all panels. The corresponding free energy maps can be found in Figure D.1 (appendix).

To further accentuate how the conformational states (φ, ψ, χ_1, χ_2) translate into attack trajectory, conformational clustering was employed (**Figure 5.10**). It shows once more that χ_1 torsion angles between -90° and -45° always yield to the higher-distance populations in the BD-FL space at around 135°/90°. Such a sidechain conformation is strongly favored by a beta-sheet like backbone with φ/ψ =-90°/90°. Some Asn residues, however, also populate the same sidechain conformation with backbone of φ/ψ = -90°/0° or even 90°/0°. The crescent-shape population in the BD-FL space frequently allows short distances and favorable, possibly reactive geometries. Conformational clustering reveals, that such attack geometries result from Asn sidechains with χ_1 torsion angles of 180°. In this case, the top arm with FL angles of 45° and higher correspond to χ_2 angles ≥ -45°. The opposite, bottom arm thus belongs to χ_2 angles ≤ -45°. The χ_2 = 180° conformations are highly common among the different Asn residues and often realized by backbone conformation of φ/ψ = -90°/90° or -90/0°. Additionally, the crescent-shape population is based on sidechain χ_1 angles of 60°. Then, the boundary χ_2 angle for the lower and upper arm in the BD-FL space is 0°. This conformation is more unlikely and not accessible to all Asn residues. Interestingly, for some residues, the conformation yields a slight increase in the FL angle. At Asn309, Asn310 and Asn522, the sidechain conformation is based on a backbone with φ/ψ angles of -90°/90°. At Asn307, Asn373 and Asn398, the corresponding backbone adopts a rare state with φ angles of -90° and below with a psi of 0°.

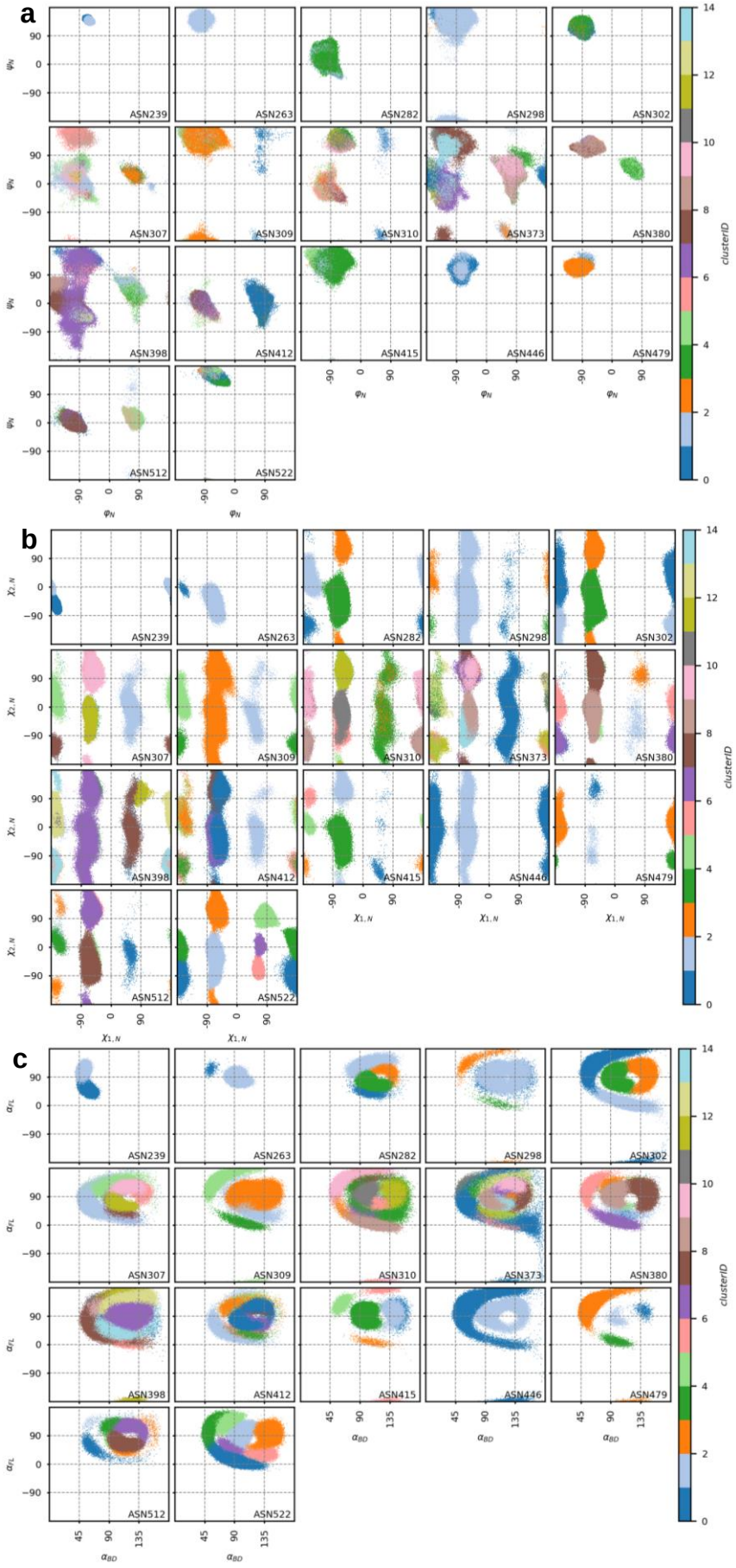

Figure 5.10: Clustering of the Asn conformations. The HDBSCAN clustering was performed in the four-dimensional torsion angle space (φ, ψ, χ1, χ2). The results of the clustering are separately shown in the a) backbone torsion angle space, b) sidechain torsion angle space, and c) attack angle space. The color map represents the identified conformational clusters and is consistent throughout a-c and all panels therein.

5.3.3 Backbone acidity

The acidity of the backbone amide hydrogen is affected by bonded and non-bonded interactions. It has been calculated in two recent, independent studies using quantum mechanical methods that the hydrogen affinity decreases when the backbone adopts a *syn* conformation, i.e. ψ angles close to 0 and φ angles close to 180°. The order of magnitude of this decrease is reported to 24 kcal/mol by [329] and 18 kcal/mol by [336]. Such backbone conformations can only be reported for a few Asn+1 residues of the Saga P-Dimer **(Figure 5.11** and **Table 5.1**). Additionally, the propensity to release the hydrogen is increased by hydrogen bonding acceptors in proximity. The probability of the hydrogen to be involved in hydrogen bonds is around 0.5 or higher for a broad array of Asn residues (**Table 5.1)**. Furthermore, we have resolved the attack distance distribution in dependence of the presence of the backbone hydrogen bond. However, significant changes dependent on the hydrogen bonding state are only observed for Asn307 and Asn398 (**Figure D.2**). The joint probability to adopt a *anti*/*syn*-backbone conformation (φ=180°, ψ=0°) and to simultaneously undergo hydrogen bonding is only above zero for Asn373 and Asn309. It is noteworthy that the relative contributions of backbone conformation and hydrogen bonding to the acidity of the backbone amide hydrogen cannot be accurately quantified from only one "training" system.

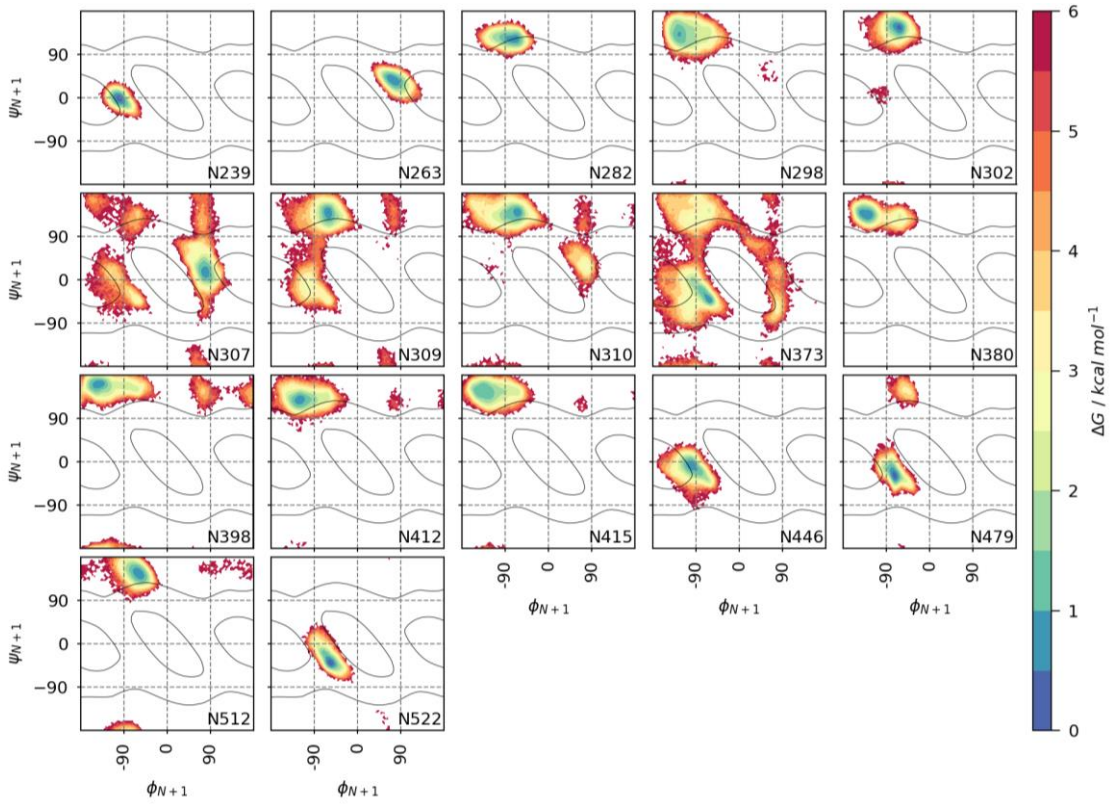

Figure 5.11: Backbone torsion free energy landscape of N+1 residues. In the panels, the Asn residues are annotated, however the torsion angles of the succeeding residues are shown. The contour lines indicate areas with increased backbone acidity. The central ellipses have the highest acidity. Data was pooled from the 5x1 μs simulations and both P-domain monomers. The color scale applies to all panels.

The results for all asparagine residues within the Saga P-domains are summarized in **Table 5.1** in term of RMSF, relative solvent accessibility, and probabilities to adopt certain conformations. The conformational probabilities are categorized into attack geometry and N+1 hydrogen acidity. Based on the previous structural-geometrical considerations, we decide that a favorable attack geometry is present, when the attack distance is shorter than 0.4 nm, the BD angle is between 45° and 135°, and the FL angle is close to 0° or 180° (each ± 45°). We have additionally calculated the joint probabilities. The N+1 hydrogen acidity is quantified by its capacity to undergo hydrogen bonds with neighboring residues as well as is backbone conformation. The hydrogen bonding criteria were chosen to a maximum 0.33 nm heavy-atom distance between donor and acceptor and an angle of 120° or larger. Acidic backbone conformation were chosen to resemble the quantum chemistry results of [329]. They reported high hydrogen acidity when ψ is between -45° and 45° and φ is between 120° and 240°. Here, again a joint probability was computed.

Table 5.1: RMSF, SASA and probabilities to adopt conformations favorable for nucleophilic reaction of Saga. The employed intervals for the angles α_{BD}, α_{FL}, ψ and φ are 90° ± 45°, 0° ± 45° (or 180° ± 45°), 0° ± 45° and 180° ± 60°, respectively. The criteria for hydrogen bonds are a maximum distance between N and O of 0.33 nm and an angle N-H-O of at least 120°.

			Attack geometry				N+1 hydrogen acidity			
Residue	**RMSF**	**rel. SASA**	**d<0.4**	**α_{BD}≈90°**	**α_{FL}≈0°|180°**	**joint**	**HB**	**ψ≈0°**	**ϕ≈180°**	**joint**
N239	0.048	0.001	1.000	0.985	0.101	0.101	0.018	1.000	0.007	0.001
N263	0.048	0.001	0.001	1.000	0.000	0.000	0.266	0.840	0.000	0.000
N282	0.058	0.084	0.001	0.993	0.006	0.000	0.280	0.000	0.014	0.000
N298	0.102	0.496	0.007	0.954	0.003	0.001	0.837	0.000	0.470	0.000
N302	0.064	0.187	0.313	0.873	0.062	0.050	0.956	0.000	0.001	0.000
N307	0.183	0.346	0.314	0.974	0.036	0.018	0.905	0.920	0.007	0.000
N309	0.183	0.481	0.094	0.822	0.023	0.014	0.101	0.066	0.007	0.003
N310	0.188	0.583	0.221	0.745	0.120	0.105	0.118	0.025	0.029	0.000
N373	**0.143**	**0.584**	**0.119**	**0.879**	**0.085**	**0.026**	**0.551**	**0.719**	**0.037**	**0.004**
N380	0.080	0.226	0.070	0.827	0.037	0.032	0.895	0.000	0.836	0.000
N398	0.109	0.561	0.332	0.938	0.195	0.125	0.487	0.000	0.841	0.000
N412	0.145	0.825	0.023	0.972	0.029	0.012	0.461	0.000	0.432	0.000
N415	0.103	0.509	0.015	0.942	0.002	0.002	0.018	0.000	0.670	0.000
N446	0.058	0.106	0.552	0.981	0.008	0.006	0.047	0.992	0.045	0.000
N479	0.123	0.089	0.921	0.995	0.006	0.006	0.000	0.969	0.000	0.000
N512	0.126	0.478	0.005	0.984	0.011	0.002	0.690	0.000	0.001	0.000
N522	0.121	0.302	0.647	0.858	0.301	0.279	0.261	0.837	0.000	0.000

We note that the selection of the criteria is of striking importance on the calculated probabilities. Such criteria are challenging to define, especially for the attack geometries as well as hydrogen bonds and conformation-based hydrogen acidity. Here, it would be necessary to model explicit free energy functions based on the geometric variables. However, the computational effort to calculate such energies is beyond the scope of this research item. Such endeavor could indeed be accelerated by machine learning techniques to avoid extensive mathematical modelling.

Nevertheless, the calculated probabilities with the somewhat arbitrary defined boundaries are still reasonable indicators for the deamidation of Asn373. They do not clearly reveal why it so exclusive to Asn373, though. It shows that Asn373 is among a small set of residues that fulfill all the necessities: solvent-accessibility, a chance to adopt favorable attack geometry *and* additionally an acidic N+1 backbone. Here, we must declare, that we also calculated the joint probability of favorable attack geometry and backbone N acidity. It was zero for all Asn residues. This is however quite reasonable. Assuming the two events favorable attack geometry and hydrogen acidity are statistical independent, the joint probability can be calculated as the product probability. The theoretical, independent probability for Asn373 would then be 0.0001. This is too small to be sampled within the 5 μs of MD simulation. We have to bear in mind that the half-life of Asn373 on Saga P-domains is in the order of days. Thus, it would naturally be very unlikely to observe a large population of fully reactive configurations. Additionally, the assumption is flawed. We monitored a significant effect of the backbone torsions at Asn373 position on the attack distance and in proteins it is known that at least the neighboring residues mutually affect their backbone conformations. Thus, the backbone conformation of the N+1 residue and the N sidechain conformation (attack geometry) must be at least to some extent allosterically connected.

5.3.4 The VA387 protruding domain dimer

Extensive MD simulations of VA387 P-dimers show that both solvent accessibility and flexibility of N373 alone are insufficient descriptors to explain its fast deamidation (**Table 5.2**), as described above for Saga P-dimers. We calculated the same structural observables for VA387 (F**igures D.3-D.6**) The probability to adopt a favorable attack geometry is slightly lowered as compared to Saga Asn373 but the hydrogen acidity is increased. Compared to other Asn residues within the VA387 P-domains, Asn373 is one of two residues that show significant probabilities for both attack geometry and hydrogen acidity. Interestingly, the second residue is Asn372. It has significantly higher probabilities for both events. Yet, it does not show fast deamidation in the NMR experiments.

These results confirm previous considerations. The accomplished microsecond scale sampling might still not suffice to fully equilibrate all probability distributions and sampling of the rarer

conformations might not have been fully converged. Additionally, we were not able to learn enough about the geometric criteria for the two different aspects attack geometry and hydrogen acidity. Finally, we do not know to what extent the different descriptors contribute to the deamidation. For example, one of the attack angles could be much less significant than the other. Hydrogen bonding could be more important than backbone conformation for the acidity or the opposite could be true. Such points must be addressed in further research by either physics driven assessment of the energetic contributions or data-structure based screening of more deamidating proteins with the herein presented methods.

Table 5.2: RMSF, SASA and probabilities to adopt conformations favorable for nucleophilic reaction of VA387. The employed intervals for the angles αBD, αFL, ψ and φ are 90° ± 45°, 0° ± 45° (or 180° ± 45°), 0° ± 45° and 180° ± 60°, respectively. The criteria for hydrogen bonds are a maximum distance between N and O of 0.33 nm and an angle N-H-O of at least 120°.

			Attack geometry				N+1 hydrogen acidity			
Residue	**RMSF**	**rel. SASA**	**d<0.4**	**α_{BD}~90**	**α_{FL}~0/180**	**joint**	**HB**	**ψ~0**	**ψ~180**	**joint**
N239	0.056	0.000	1.000	0.994	0.181	0.181	0.026	1.000	0.012	0.002
N263	0.052	0.001	0.003	0.999	0.000	0.000	0.311	0.824	0.001	0.001
N282	0.067	0.037	0.001	0.984	0.013	0.000	0.349	0.000	0.021	0.000
N302	0.069	0.221	0.218	0.834	0.037	0.030	0.967	0.000	0.000	0.000
N307	0.223	0.428	0.236	0.975	0.040	0.014	0.833	0.880	0.052	0.002
N309	0.216	0.626	0.076	0.834	0.019	0.012	0.072	0.019	0.006	0.000
N310	0.212	0.482	0.234	0.778	0.121	0.107	0.077	0.000	0.021	0.000
N372	0.202	0.556	0.084	0.849	0.034	0.018	0.422	0.693	0.130	0.026
N373	**0.181**	**0.401**	**0.140**	**0.890**	**0.034**	**0.010**	**0.382**	**0.815**	**0.069**	**0.009**
N380	0.079	0.138	0.087	0.868	0.056	0.051	0.955	0.000	0.886	0.000
N393	0.244	0.605	0.084	0.952	0.044	0.021	0.133	0.070	0.085	0.000
N394	0.215	0.566	0.143	0.824	0.040	0.028	0.081	0.115	0.199	0.003
N397	0.121	0.360	0.193	0.941	0.131	0.087	0.498	0.000	0.776	0.000
N406	0.065	0.184	0.164	0.852	0.005	0.002	0.726	0.971	0.002	0.001
N414	0.129	0.309	0.047	0.911	0.009	0.006	0.036	0.000	0.682	0.000
N445	0.073	0.193	0.677	0.988	0.028	0.026	0.114	0.884	0.023	0.000
N447	0.073	0.146	0.041	0.848	0.028	0.022	0.929	0.000	0.576	0.000
N478	0.124	0.211	0.929	0.998	0.006	0.006	0.000	0.983	0.000	0.000
N511	0.139	0.214	0.014	0.981	0.015	0.006	0.679	0.000	0.001	0.000
N521	0.129	0.435	0.659	0.876	0.294	0.272	0.204	0.778	0.000	0.000

5.4 Conclusion

The probability for Asn373 to adopt favorable attack geometries is 0.026, which ranks it at the 6th position after Asn522 (0.279), Asn398 (0.125), Asn310 (0.105), Asn302 (0.05), and Asn380 (0.032). Thus, clearly, attack geometry alone is not a sufficient to explain deamidation. The likeliness for the succeeding residue of Asn373 to undergo backbone amide hydrogen bonding is 0.55, for Asn522 0.26, for Asn398 0.487, for Asn310 0.118, for Asn302 0.956 and for Asn380 it is 0.895. Only when accounting for N+1 backbone conformation, Asn373 becomes among the top candidates for deamidation.

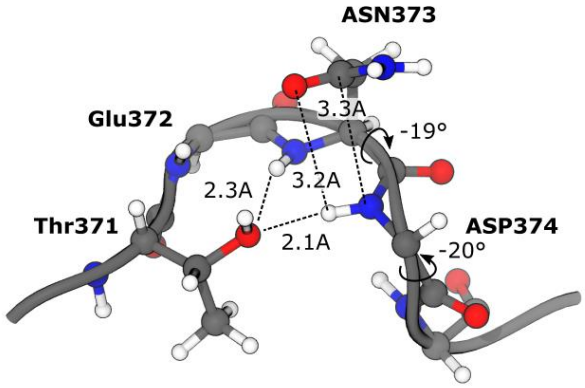

Figure 5.12: Snapshot of a productive conformation of the Asn373 loop. The two torsion angles ψ_N and ψ_{N+1} are annotated as well as the attack distance, the two stabilizing hydrogen bonds and the carbonyl oxygen to backbone hydrogen distance.

Furthermore, the likeliness for the N+1 backbone hydrogen bond is slightly reduced. N373 of both Saga and of VA387 P-dimers is part of a type II' ST turn 35. These loops are characterized by a hydrogen bond between the side chain OH-group of Ser or a Thr residue i-3, and the backbone NH of residue i+2. The ST turn motif is associated with the backbone dihedral angle of residue N+1 of 0°, i.e., a syn orientation. In crystal structures of the P-dimers, residues T371, E372, and N373 in the case of Saga29, and T371, N372, and N373 in the case of VA387 28 form a ST turn with a hydrogen bond between the side chain OH of T371 and the backbone NH of N373. Interestingly, T371 is not only engaged in hydrogen bonding as part of the ST turn but at the same time forms a hydrogen to the backbone NH of D374, which is only possible with N373 residing in a *syn* conformation. This leads to two consecutive *syn* backbone orientations of residues E/N372 and N373 (**Figure 5.12**).

Thus, Asn373 is the only solvent accessible residue that can attain the generally highly unfavorable conformation of a double anti/syn backbone orientation at the N and N+1 position with simultaneous productive sidechain placing and stable hydrogen bonding of the N+1 backbone

hydrogen. However, it is not clear yet if such a cause for fast deamidation is a general feature or if it is exclusive to norovirus P-dimers.

6 MEMBRANE RECRUITMENT OF PI3P

Protein-protein complexes often show nanomolar affinities based on a range of electrostatic and hydrophobic interactions as well as hydrogen bonds. The interaction partners diffuse in a 3-dimensional environment and are surrounded by highly polar water molecules. The relative unlikeliness of prolific collisions is compensated by high-affinity binding, long-lived complexes and fast diffusion driven by the aqueous environment. The environment of a lipid bilayer membrane is fundamentally different. The diffusion is limited to two dimensions, which increases the probability of collisions. The medium in the center of the bilayer is apolar and hydrogen bonding acceptors and donators are largely absent. Hence, interactions between two discrete molecules are weak and short-lived, and interaction studies afford the presence of multiple molecules of a certain species as well as sufficient sampling and adequate analysis. Most frequently, interactions within a bilayer are not limited to the central fatty acid region, but also happen between the solvent exposed chemical groups. In this chapter, the mutual effects of a membrane anchored peptide with its surrounding lipid molecules are investigated using coarse-grained molecular dynamics. The coarse-grained MD simulation were performed in parallel to full-atomistic simulations by E. Münzberg and are published in [337] and are partially discussed in her dissertation thesis [338].

6.1 Introduction

The Rab5 protein is a member of the Rab-GTPase family and has important roles in the regulation of endosomal trafficking [339]. As a peripheral membrane protein, it is anchored to the endosomal membrane via two subsequent geranylgeranyl posttranslational modification at the Cys212 and Cys213 positions. The C-terminal geranylgeranylation sites are part of a hypervariable region, which flexibly links the Rab5 GTPase domains with the bilayer [340] The early endosome membrane as well as the plasma membrane are a mixture of plenty different molecules including lipids and proteins [52, 341]. In general, the lipid composition is highly dynamics, asymmetric between inner and outer layer and largely depends on the cellular localization of the membrane [342]. On early endosomes, the anchoring point for Rab5, the membrane composition is accentuated by the presence of the signaling lipid PI3P [343, 344].

It is of interest if the intersection of Rab5 and PI3P signaling can be structurally rationalized by the formation of nanoscale membrane domains. To test this hypothesis, coarse-grained MD simulations were carried out using the MARTINI force field, which proved valuable for long- and large-scale simulations of lipid bilayers [52, 74, 345, 346]. Therefore, it was pragmatic to truncate the C-terminal Rab5 peptide with the membrane anchors instead of simulating the whole protein. At the time when this research was carried out, no coarse-grained parameters for the

geranylgeranyl-cysteine were available. They were thus newly developed by comparison to the full atomistic simulations of E. Münzberg. With the new parameters, the interactions between the lipidated peptide and surrounding membrane compounds were sampled and quantified using radial density distributions. Five different model membranes with increasing complexity were assessed (see Method) by five μs MD simulations each to ensure thorough mixing of the membrane.

6.2 Method

6.2.1 Coarse-grained system setup

The coarse-grained model of the double geranylgeranylated peptide was generated using the MARTINI software *martinize* [347, 348] on the full-atomistic model generated by E. Münzberg [337]. The c-terminal peptide of Rab5 has the following sequence 205-QPTRNQCCSN-215. The geranylgeranyl moieties were covalently linked to the cysteine residues 212 and 213. Thus, for the MARTINI modeling, geranylgeranyl-cysteine was added as another amino acid. No secondary structure by means of an elastic dynamic network was assigned to the peptide, i.e., it was fully flexible. The termini were charged.

Three different, symmetric lipid bilayers were modeled: 1. A ternary mixture of 40 mol% palmitoyl-oleoyl-phosphocholine (POPC), 20 mol% palmitoyl-sphingomyelin (PSM) and 40 mol% cholesterol; 2. A mixture of POPC, cholesterol, PSM, palmitoyl-oleoyl-phosphoethanolamid (POPE), palmitoyl-oleoyl-phosphoserine (POPS); 3. A mixture of POPC, Cholesterol, PSM, POPE, POPS, and phosphatidylinositol-3-phospaht (PI3P). The bilayers were modeled to a size of 10x10 nm using the *insane* software [349]. The software was extended to include PI3P. The geranylgeranylated peptide was introduced to the center of the bilayer patch in a way that the membrane anchors were at the same height as the alkyl group particles of the lipid molecules. The spatial distribution of the lipids was randomized with every lipid having a uniform probability distribution. Also with the insane software, the bilayers were solvated with standard MARTINI water (a 9:1 mixture of MARTINI water particles and MARINI anti-freeze particles) and ionized to 0.15 M NaCl. The topology files with the force-field parameters were downloaded from the MARTINI website cgmartini.nl. The non-polarizable MARTINI 2.0 parameters were used [74, 345]. The topology for PI3P was generated as a combination of the available topologies "PAPI" and "POP1". For PSM, the topology "DPSM" was used.

6.2.2 Coarse-grained MD simulations

The coarse-grained simulations all were performed with the GROMACS [75-77, 80, 81, 309] ver. 5.0.7 and recent MARTINI input parameters. In all simulations, a Verlet cutoff scheme [222] was employed with a neighbor list update at every 20 steps. Periodic boundaries were employed in all

directions. Van-der-Waals interactions were cutoff at 1.1 nm using a potential shift Verlet modifier. Coulomb interactions were treated using a reaction field with a cutoff of 1.1 nm. The dielectric constant was set to 15. Temperature coupling was achieved with the velocity rescaling algorithm [350] and a coupling constant of 1 ps. If not described otherwise, temperature coupling was performed individually for solvent and bilayer/monolayer. During equilibration, Berendsen pressure coupling was employed [93] to realize fast box volume convergence. For production, Berendsen coupling was replaced by Parrinello-Rahman [312] coupling. Reference pressure was 1 ATM. Berendsen coupling had a coupling constant of 5 ps and Parrinello Rahman of 12 ps. In all cases, the pressure coupling was semi-isotropic, i.e., different in lateral and normal directions. For the bilayers the compressibility was 4.5e-4 in both directions. In all instances, the modeled systems were initially minimized using a steepest descent algorithm [351] of 1.000 steps. Velocities are assigned to suite a Maxwell-Boltzmann distribution of the target temperature. The bilayer simulations were conducted at a temperature of 310 K. NVT equilibration had length of 0.5 ns with a 10 fs time step. The box was relaxed for 30 ns with a 20 fs time step. Production sampling was generated over 5 μs with a 20 fs time step.

6.2.3 Trajectory analysis

To assess the local lipid bilayer composition around the anchored peptide, the GROMACS tool *gmx select* was employed. The local concentration of a lipid species within a certain cutoff around the peptide was calculated as the number of molecules of a certain species divided by the total number of lipids. A lipid was considered as within the cutoff, if any pairwise atom-atom distance between lipid and anchored peptide was closer than the cutoff. Only lipids of the upper leaflet, in which the peptide was anchored, where considered. The calculation took periodic boundary conditions into account. For the lateral density profile, the trajectories were preprocessed to center the peptide in the box. With the transformed coordinates the 2D lateral density of PI3P was calculated and averaged over the whole 5 μs trajectory using the GROMACS tool *gmx densemap*.

6.3 Results and discussion

To investigate the pairwise effects between the Rab5 lipid anchor and the local membrane environment, three multi-microsecond coarse-grained MD simulations of the anchor within three model bilayers were performed. The double-geranylgeranylation was attached to the 10-residue C-terminal peptide of Rab5. The model bilayers were pure POPC, a ternary mixture (4:4:2) of POPC, PSM and cholesterol, and an early endosome model membrane consisting of POPC, POPE, POPS, PSM, Cholesterol and PI3P (see Method section for details). Coarse-grained parameters of the lipid molecules were available or easily accessible through simple recombination of existing parameter sets.

6.3.1 Coarse-grained force field development

Coarse-grained parameters and mapping of the geranylgeranyl anchor were newly developed and optimized against full-atomistic simulations by E. Münzberg [337]. The bead mapping and typing defines the non-bonded parameters. It is recommended to map distinct chemical groups of 4-5 atoms together [345]. In the case of geranylgeranyl-cysteine, this was achieved by mapping the amino acid backbone atoms, the sidechain atoms, and each isoprenoid group into single beads (**Figure 6.1).** The bead types of the backbone (P3) and sidechain (Na) were predetermined due to restrictions of the force field [347]. The choice of C3 as the bead type for the prenyl-group beads was finally rationalized by free energy estimation using umbrella sampling [352] and weighted histogram analysis[126]. Optimum bonded parameters in the MARTINI force field are always a compromise between accuracy and long time-step stability. Thus, for long alkyl chains it unnecessary and even disadvantageous to use dihedral angle potentials. Thus, the geranylgeranyl anchor was parameterized by only bond length and angle parameters, which are displayed in (**Table 6.1**).

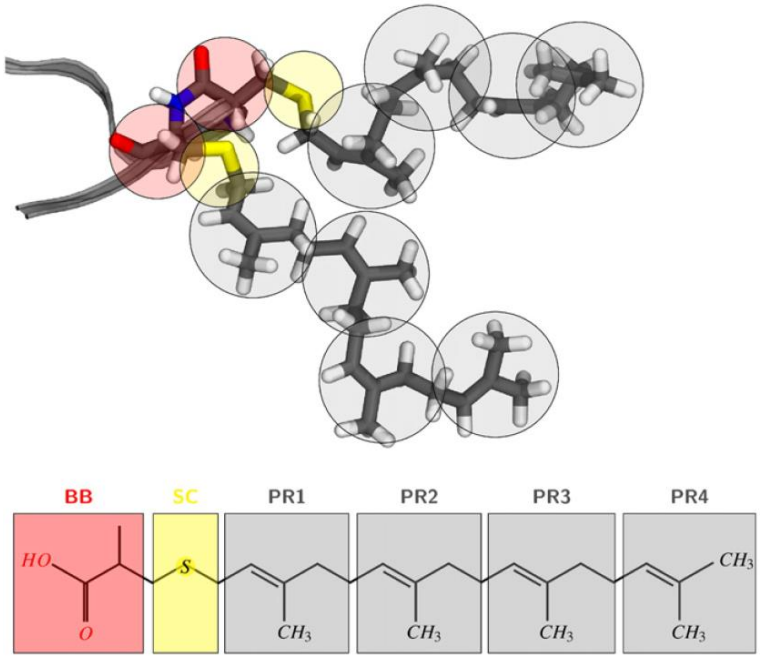

Figure 6.1. Chemical structure and coarse-grained mapping of the double geranylgeranylated peptide. The top image represents the three dimensional structure of the geranylgeranyl-cysteine residues in full-atomistic detail in sticks representation. The peptide is represented as a ribbon. The resulting coarse-grained beads are shown as circles around their respective chemical groups. The bottom scheme shows the coarse-grained mapping and assigned bead names in more detail.

Table 6.1. Developed coarse-grained parameters for geranylgeranyl cysteine. The stared variables are the equilibrium values and F the force constants.

Bond	r* / nm	F_r	Angle	φ* / °	F_φ
BB-SC	0.31	20.000	BB-SC-PR1	135	20
SC-PR1	0.37	10.000	SC-PR1-PR2	115	25
PR1-PR2	0.46	7.500	PR1-PR2-PR3	105	20
PR2-PR3	0.46	7.500	PR2-PR3-PR4	100	20
PR3-PR4	0.46	7.500			

With the developed coarse-grained parameters, the bond length distributions in solution of the full-atomistic model could be reproduced very well (**Figure 6.2**). The angles distributions were also well reproduced. However, minor differences between the two geranylgeranylations were not captured and both moieties behaved identically. It is unclear if the bond angle differences are a cause of insufficient sampling or unexpected intermolecular interactions. As the inaccuracies were so small, the issue was not further investigated. The torsion angle distributions were not accurately reproduced as they were not explicitly parametrized. However, there is a clear tendency in both coarse-grained and full-atomistic simulation, that the BB-SC-PR1-PR2 and SC-PR1-PR2-PR3 dihedrals favor an angle of 180°. The full-atomist torsion between the four prenyl-units also has significant population at 0°, which however was not covered by the coarse-grained simulations. Overall, the coarse-grained parameters were considered sufficient.

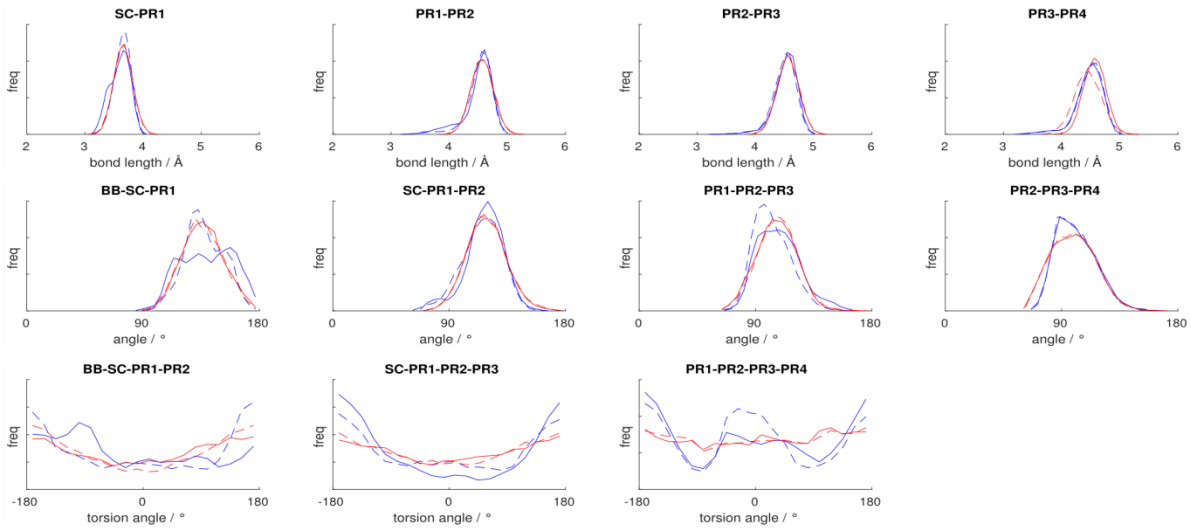

Figure 6.2: Validation of the bonded parameters. The top row represents the bond length, the central row the bond angle and the third row the bond dihedral distributions. The sampled values from the coarse-grained simulations are colored red, and the all atom values blue. The solid lines correspond to the first geranylgeranyl anchor at the 212 position, and the dashed lines to the second one at the 213 position. The full atomistic distributions were computed from identical groups using the center-of-mass of the atoms which correspond to a coarse-grained particle.

6.3.2 Lipid enrichment and domain formation

The twofold geranylgeranyalted C-terminal Rab5 peptide was placed in the center of each coarse-grained membrane model and subjected to dynamics simulation. The lateral radial distribution function (see method) of each lipid species and cholesterol were calculated over the whole trajectory (**Figure 6.3**). In the ternary bilayer, cholesterol was significantly enriched in proximity to the peptide (55% in a 1 nm radius relative to 40% in the bilayer composition). POPC and PSM concentrations close to the anchor were decreased. In the second shell (2-3 nm radius), the cholesterol concentration is slightly reduced whereas POPC is enriched. In the early endosome membrane model without PI3P, again cholesterol is enriched in the first shell and depleted in the second. Here, PSM is the dominating lipid and is slightly enriched in the second shell. In the early endosome model membrane with PI3P, the signaling lipid is substantially enriched in proximity to the anchored peptide. It is, in fact, four times more likely to find PI3P close to the peptide than anywhere else in the bilayer.

The enrichment of certain molecular species around the anchor can be attributed to both interactions within core region of the bilayer and head group region in solvent phase or a combination thereof. For the enrichment of cholesterol in close proximity to the anchored peptide, it is most likely that the missing head group of cholesterol was decisive. The peptide was sitting in the head group region and displaced neighboring phospholipids. Thus, the space under the peptide could only be occupied by cholesterol. The PI3P enrichment can be explained by two mechanisms. I. Carbohydrate head groups show a high tendency to self-aggregate as observed *in vivo* and *in vitro* especially for gangliosides [353]. II. Electrostatic interactions between acidic lipid head group (phosphate) and basic arginine residues on the peptide allow the formation of a stable microdomain as also experimentally observed for PI2P [354]. In the case of PI3P, such an accumulation aids the recruitment of effector proteins such as Vps34 [355]. The coarse-grained simulation results yield a molecular rationalization of biological observations which were not accessible using full-atomistic simulations. The enrichment of PI3P may be the initial step for the formation of an effector protein recognition platform.

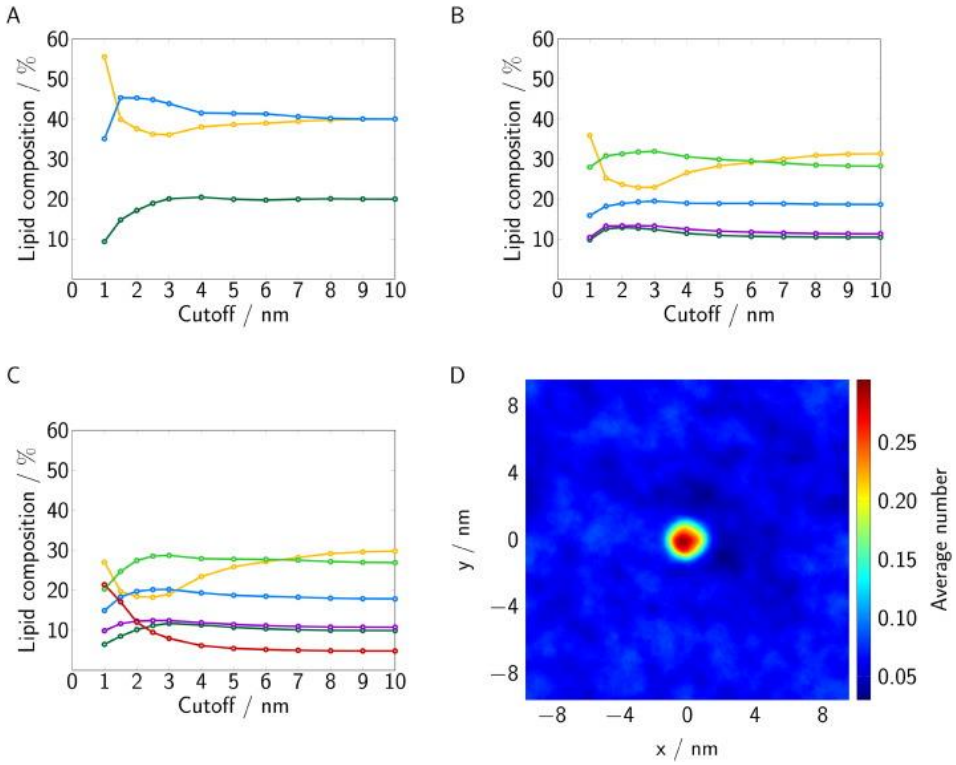

Figure 6.3: Lipid enrichment around the anchored peptide. A-C) Radial distribution functions of lipids around the anchor in the ternary bilayer (A), the early endosome model bilayer without PI3P (B) and the early endosome model bilayer (C). The distribution function is expressed as the relative lipid composition. Only the leaflet is considered in which the peptide was anchored. Lipid types are colored as follows: POPC (blue), cholesterol (yellow), PSM (dark green), POPE (light green), POPS (violet), and PI3P (red). D) Lateral distribution function of PI3P sampled over 5 μs. The anchored peptide was centered before the calculation.

6.4 Conclusion

In this research, a coarse-grained model for geranylgeranyl-cysteine was developed and its effects for the membrane segregation of the Rab5 C-terminal peptide were probed. The coarse-grained model allowed multi microsecond sampling of the peptide embedded in four different model bilayers. Such long simulation times were necessary to sufficiently sample mixing and allow convergence of radial distribution functions. The coarse-grained modeling approach using insane.py packing software proved to be robust, simple, and fully automatable. All atom conformation distributions were reproduced with high accuracy. The coarse-grained simulation revealed a significant aggregation of the PI3P signaling lipid in proximity of the Rab5 peptide. However, the effects could be attributed to superficial electrostatic interaction between the negatively charged head group of PI3P and the positively charged arginine residues on the peptide. A phase segregation by means of separation of liquid ordered and liquid disordered phases was not observed. The interactions between the fatty acyl tails were thus insignificant. Nevertheless, the coarse-grained modeling approach was considered highly feasible for the modeling and equilibration of molecular layers, so that further research projects were inspired.

7 MONOLAYER PHASE SEGREGATION AND TRANSITION

Coarse-grained simulations are frequently employed to study molecular interactions within a lipid bilayer. As self-assembled alkanethiol monolayers (SAMs) share a substantial amount of properties with lipid bilayers, it can be assumed that coarse-grained modeling and simulation strategies for bilayers can also be utilized for SAMs. In this chapter, the development and application of such an approach is described for two specific binary SAMs. The studied SAMs consist of a matrix of C16 alkanethiols adsorbed on a Gold(111) surface. The alkanethiols are functionalized with hydroxyethyl groups conjugated via amide linkages to the alkane chains. Additionally, some of the matrix compounds are replaced by anchoring compounds, which tether either C8 or C16 alkyl groups to the alkanethiols via oligoethylenglycol (OEG) polymer chains. The application for such mixed SAMs lies in the tethering of protein-carrying lipid bilayers for pharmaceutical research. However, spectroscopic analysis of the mixed SAMs reveal unexpected interactions between only the C16 alkyl-anchor compounds. As the structural basis of these interactions could not be fully resolved by experiments, coarse-grained and full-atomistic simulation were conducted. Experiments were performed by Martynas Gavutis and colleagues at the Center for Physical Sciences and Technology in Vilnius. The experimental and theoretical results are published in the two separate articles [356] and [55]. The following sections and figures are reprinted with permission from Schulze, E. and M. Stein, *J Phys Chem B*, **2018**. 122(31): p. 7699-7710. Copyright 2018 American Chemical Society.

7.1 Introduction

The phenomenon of an adsorbing (self-assembling) monolayer (SAM) onto a metal surface was first observed more than 80 years ago by Bigelow and co-workers, [357] and it later regained interest as a model system to investigate the fundamentals of intermolecular interaction and adsorption. [358] The ability to modify both the head and tail groups of the layering molecules makes SAMs excellent systems to probe the numerous competing effects such as hydrophilicity versus hydrophobicity and order versus disorder. Apart from silanes or fatty acid derivatives on hydroxylated surfaces, a large number of publications are devoted to the studies of organothiol/disulfide assemblies on semiconductor or metal surfaces, in particular gold (for relevant reviews, see [358] or [116]).

The high degree of control and reproducibility of SAM formation make them an ideal model system to design a controllable microenvironment in nanotechnology and biology. For example, SAMs with properly designed terminal groups (herein referred to as "anchor"' groups) serve as an excellent starting point for the development of tethered bilayer lipid membranes (tBLMs) [359].

Such tethered bilayers prepared by the fusion of liposomes [360] can be regarded as a first-order "synthetic" model of a biological cell membrane [361] as they can provide a close-to-native and protective environment for membrane-associated proteins [362]. For a review article about the structure and function of supported BLMs, see [363]. Highly reproducible and standardizable protocols are available, which turn SAM–tBLMs [364] into an attractive platform for the structural and functional characterization of membrane proteins and also for the screening of drug candidates targeting these proteins [365, 366]. For example, tBLMs have proven to be useful for monitoring the dynamics of a ternary cytokine receptor protein complex [367]. In addition, SAMs and tBLMs on solid support are also well-suited for an extensive characterization using established surface analytical techniques such as surface plasmon resonance spectroscopy, X-ray and neutron reflectometry, contact angle goniometry, infrared (IR) spectroscopy, fluorescence microscopy, scanning probe microscopy, atomic force microscopy, and electrochemistry, to mention only a few [368, 369].

Organothiol SAMs are typically prepared by immersing a gold substrate overnight in a dilute ethanolic solution of the alkanethiol compound. The alkanethiol can be functionalized with soluble polymer/oligomer moieties, for example, oligo ethylene glycols (OEGs), and such SAMs have been extensively used for fundamental protein adsorption studies [370-372]. OEG–alkanethiols (as a matrix compound) also have been proposed to function as a soft cushion for tBLMs by Lee et al. [373]. In this case, the matrix compound is mixed with a longer intercalating molecule that serves as an anchor for the lipid bilayer (**Figure 7.1**).

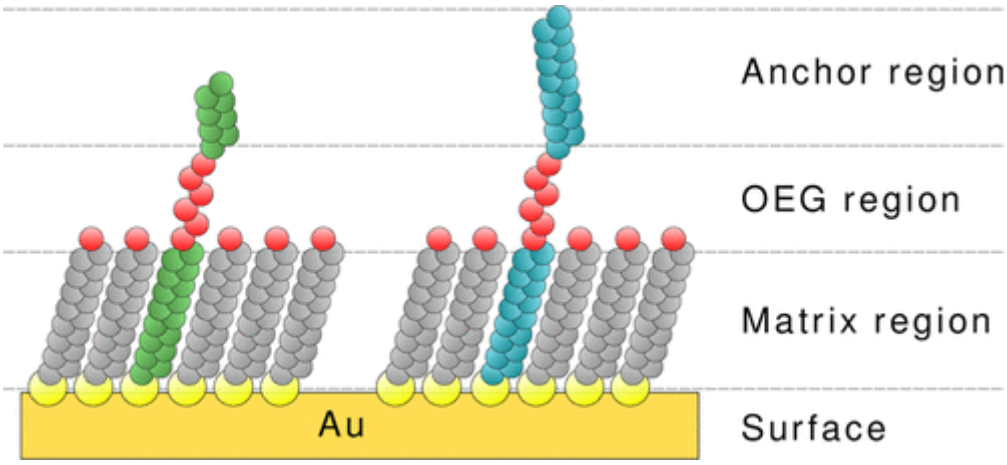

Figure 7.1: Schemes and nomenclature of the mixed self-assembled monolayers. The orange box represents the gold-coated surface. The molecules represent the different compounds. In all cases, the yellow spheres correspond to the adsorbing thiol groups and the red spheres to the terminal or OEG oxy-ethylene groups. The gray, green and blue spheres represent alkyl chains of the matrix, C8-anchoring and C16-anchoring compounds, respectively. The mixed SAMs are composed of matrix with either C8- or C16-anchoring compounds.

Molecular dynamics (MD) simulations can provide deep insight into the structure and dynamics of SAMs in atomic detail. In the past, it led to a clearer understanding of the architecture of the monolayers of alkanethiol chains with different terminal groups [374, 375]. Also, large-scale MD simulations of the alkanethiol self-assembled monolayers have been performed to study the effects of temperature and packing density on the structural parameters [376]. Sufficient sampling and convergence to equilibrium configurations are difficult to achieve because of slow molecular reorientations, and for the reproduction of lateral diffusion, advanced conformational sampling algorithms are required [377, 378]. The accuracy of the different force fields varies to a great extent, especially in the reproduction of chain length-dependent tilt and twist angles [379]. Nevertheless, MD is the appropriate tool to study the SAM structure and dynamics at high spatial and temporal resolution. To overcome the complexity and the necessity of long equilibration periods, especially in systems involving membranes or vesicles, the MARTINI coarse-grained (CG) force field has recently been of extensive use [380]. Its application on phospholipid bilayers [52, 349] and polymers, [381], PEGylated lipids [373] and alkanethiol-covered nanoparticles [382] demonstrate the feasibility of this CG force field for molecular systems comparable to ours.

Although many of the previous MD studies focused on single-component matrix molecule monolayers, we here present a novel approach for the systematic simulation of a multitude of mixed monolayers at different mole fractions of different anchoring molecules using various representations and a concise transformation between them. Our simulations are most helpful to complement the experiments because they can probe mixed concentrations and mole fractions that are experimentally hard or impossible to access. In this article, we address for the first time the structure and dynamics of mole fraction-dependent orientational and conformational transitions of the alkyl anchor compounds in a matrix environment attached to a gold (111)-like surface. We here show that structural parameters such as lattice constant and tilt angle are almost independent of the type of anchoring molecule. The anchor alkyl-chain orientation, however, is critically controlled by the composition and its chemical nature. Although the short-chain C8 anchor molecules adopt a random, disordered conformation at low anchor densities and transition into a more ordered conformation with increasing anchor density, the C16 anchors already adopt highly ordered conformations at low anchor densities, which reoccur in the monolayers of higher anchor density. The results of our large-scale, two-step, multiscale procedure are in excellent agreement with the recent experiments by Lee et al. and explain the observed spectral features. Finally, our newly composed modeling procedure forms the basis for future computational investigations on SAM–tBLMs and SAM-tethered lipid vesicles of various anchoring molecules and monolayer compositions.

7.1.1 Experimental characterization of mixed SAMs

Binary mixed SAMS with the purpose to tether lipid bilayer membranes, consist of matrix compounds and anchor compounds (**Figure 7.2**). Both compounds share a common alkanethiol backbone. The anchor compounds additional exhibit terminal aliphatic moieties. The alkyl anchors are not directly conjugated with the alkanethiol backbone but rather flexibly tethered via oligoethylenglycol (OEG) chains. The central OEG portion is bound to the aliphatic portions via amide bonds. The amide bonds allow the formation of hydrogen bonding network to further stabilize the SAM. To investigate the systematic structural effects of the mixed SAM composition, Lee et al. [54] have synthesized SAMs with increasing anchor compound concentration. Furthermore, two different anchor compounds with varying alkyl chain lengths of 8 and 16 carbon atoms, with the chemical formulas $HS(CH_2)_{15}CONH((CH_2)2O)_6CH_2CONH(CH_2)_7CH_3$ and $(HS(CH_2)_{15}CONH((CH_2)_2O)_6CH_2CONH(CH_2)_{15}CH_3)$, respectively, were investigated. The authors determined the surface hydrophobicity by means of contact angle goniometry, as well as the layer thickness via ellipsometry. To subject the monolayers to IRRAS, the anchoring alkyl portions were deuterated to distinguish them from the matrix.

Contact angle measurements allow assessments of the hydrophobicity of the surface, i.e., the amount of favorable surface-solvent interactions. Two angles are determined, the advancing and the receding contact angle. Generally, larger angles correspond to a more hydrophobic surface. Low angles are achieved at hydrophilic, well wetting surfaces. Angles larger than 90° are considered hydrophobic, and smaller than 20° rather hydrophilic. For example, polystyrene yields a contact angle of 85-92° [383, 384], whereas a clean gold surface yields values below 10° [385]. In the case of pure a SAM of the before mentioned matrix compound, contact angles of 30°/40° (advancing/receding) were measured. The pure anchor monolayer had contact angles of 105°/115° independent of anchor length (C8 or C16).

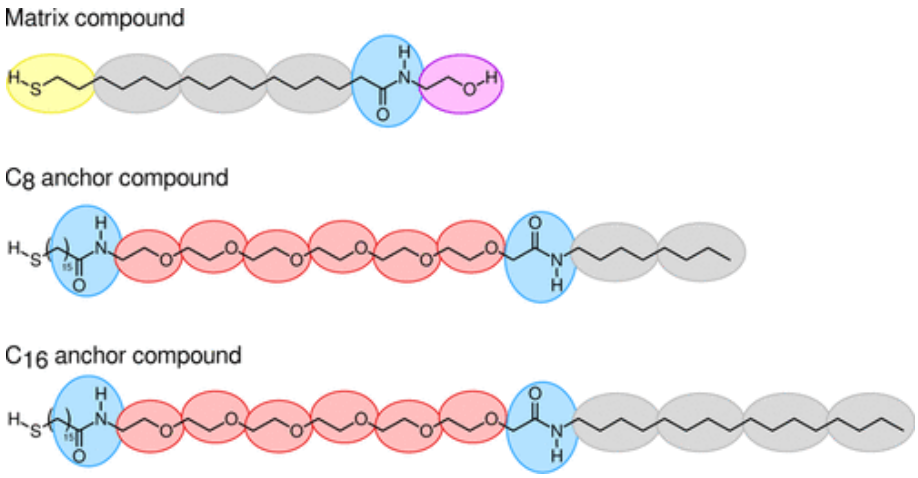

Figure 7.2: Chemical structures of the three monolayer compounds. The coarse-grained mapping is shown as colored circles. The matrix part is identical in both anchor compounds. The colors correspond to the utilized MARTINI particle types. Yellow: C5, Gray: C1, Blue: P5, Red: SN0, Purple: SP2.

In case of the mixed monolayers, it must be mentioned that only the concentration of the compounds in the supernatant are truly known. The composition of the adsorbed SAM deviates from the supernatant due to different adsorption isotherms [386] which result mostly from different solubilities. Thus, a supernatant C16 anchor mol fraction of 0.5 for example, results in an assembled monolayer with an anchor mol fraction of 0.1-0.2. This behavior is weaker for the C8 anchor (mol fraction of 0.5 in solution leads to 0.4 in SAM). It shows however, that high anchor mol fractions are experimentally not accessible and thus demonstrating on the importance of molecular simulations. Anyway, the contact angles increase with increasing anchor mol fractions. Interestingly, the contact angle hysteresis is much higher for the C16 anchor monolayers, which suggests a higher roughness on the molecular scale [387]. The magnitude of the contact angles appears higher for the C8 anchor SAMs (60°/70° at 0.5 supernatant mol fraction vs. 30°/50° for C16). This however is most likely the consequence of the worse adsorption of the long anchor compounds. Instead, the contact angles should increase linearly and with about the same incline for C8 and C16 anchoring chain length. The ellipsometric monolayer thickness is also more meaningful for the pure SAMs than for the mixtures. The matrix-only SAM has a thickness of 25 Å, the pure C8 anchor SAM of 49 Å, and the pure C16 anchor SAM of 58 Å.

Interestingly, the IRRAS spectra (**Figure 7.3**) of the deuterated anchor portions reveal significant differences between the two anchor compounds of different lengths. Due to the polarization of the radiation source, IRRAS spectra show the absorption of certain chemical bond vibration in dependence of their orientation. Thus, the band intensity is not only determined by the abundance of the bond but also additionally their tilt against the surface normal so that IRRAS spectra allow statements about molecular conformations. The IRRAS spectra of the matrix region show peaks at 2918 cm^{-1} and 2850 cm^{-1} which correspond to asymmetric and symmetric C-H stretching of alkyl

chains in highly ordered, densely packed, all-trans arrangement. In the pure anchor SAMS, the OEG linker portion can be seen in the IRRAS spectra as an additional band at 2890 cm^{-1}, it corresponds to asymmetric C-H stretching and only appears when the ethylene glycol elements form a helix which is orientated along the surface normal. The band is not visible in the mixed SAMS. The spectra of the anchor portions are similar to the matrix spectra but naturally at lower wavenumbers. The C16 anchor monolayer has a high peak at 2193 cm^{-1} and a smaller on close by at 2216 cm^{-1}. The second major peak is at 2090 cm^{-1} with a neighboring, smaller peak at 2074 cm^{-1}. The major peaks correspond to CD_2 asymmetric and symmetric stretching, whereas the smaller peak show the CD_3 vibrational modes. For the C16 anchor, the CD_2 bands are higher than CD_3 bands also for lower anchor concentrations. For the C8 anchor, though, The CD_3 peaks are higher than the CD_2 peaks. The spectra of the C16 anchors indicate highly ordered, perpendicularly oriented d-alkyl chains [395], even for anchor molar fractions of < 0.5. The C8 SAM spectra are rather similar to bulk d-hexane, and thus indicate partial disorder especially for lower anchor fractions.

All in all, the experimental examination of the mixed SAMs yields the hypothesis that the long alkyl anchor compounds form domains of unknown size, in which the anchor alkyl chains pack tightly and orient perpendicular to the surface. Thus, the spectra of low anchor mol fractions are similar to the pure anchor monolayer spectra. Additionally, the larger contact angle hysteresis of the C16 anchor monolayers would be explained by such a rough surface of spatially clustered anchor domains [396] versus a flatter surface of spatially equally distributed anchors. Molecular modeling and simulation can help the visualization of such effects on the nanoscale and favor or disfavor the hypothesis.

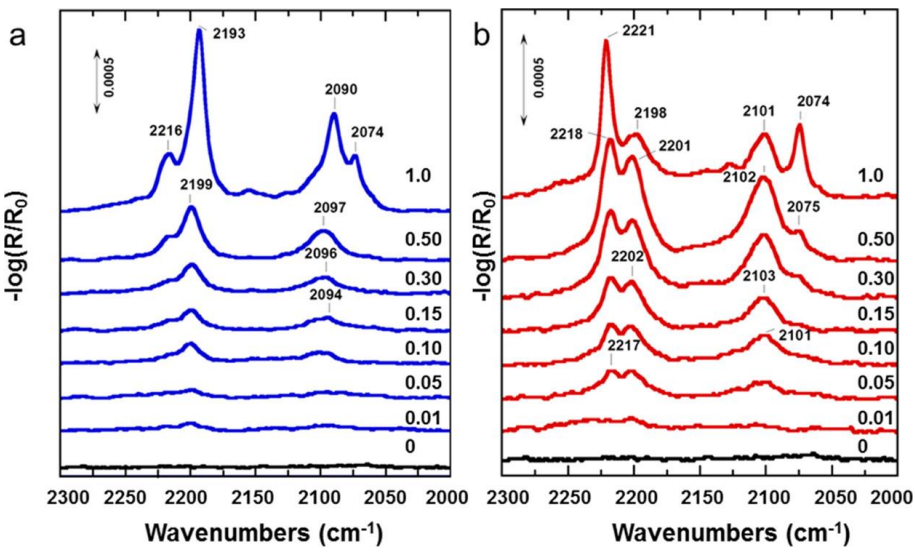

Figure 7.3. IRRAS spectra of the deuterated alkyl anchors. a) C16 anchor mixed SAMS. b) C8 anchor mixed SAMS. Various supernatant anchor molar fractions are shown. Reprinted with permission from Lee et al. *J Phys Chem B*, 2018. 122(34): p. 8201-8210. Copyright 2018 American Chemical Society.

7.2 Method

7.2.1 Initial configuration

To model SAMs, the insane software [349] was repurposed. The asymmetry option was exploited to only build a single layer. The packing procedure of the molecules was rewritten so that the lipids are placed into a hexagonal instead of a rectangular grid. Therefore, the initially generated rectangular grid point coordinates were transformed using the following equations.

$$y_{hex} = \frac{\sqrt{3}}{2} y_{rect} \tag{26}$$

$$x_{hex} = \begin{cases} x_{rect}, & x \bmod 2 = 0 \\ \frac{1}{2} x_{rect}, & x \bmod 2 = 1 \end{cases} \tag{27}$$

All the MARTINI particles were initially put to identical x and y coordinates. The initial z-directional distance between two standard sized particles was set to 1.5 nm, for small particles to 1.0 nm. To model molecular aggregates, a stochastic algorithm was developed. The algorithm takes an additional input argument: the target domain size. Initially, it estimated the number of cluster seeds as the number of clustered molecules divided by the target domain size. In a second step, it randomly samples 10.000 seed coordinates and calculates the pairwise distances, taking periodic boundaries into account. The configuration in which the pairwise distances are maximized is employed for molecular placement. A molecule, which is to be positioned into clusters, occupies each seed positions. Then additional molecules of the same kind are added one after another until the target concentration is reached. The molecules are placed in the nearest neighborhood positions of a random occupied spot in each cluster. What exact positions are occupied is decided randomly using a uniform probability distribution. The algorithm is supposed to loosely resemble the selective adsorption of molecules only in the neighborhood of their own kind, or the collective adsorption of aggregates that pre-formed in solution. Example configurations are presented in **Figure 7.4**.

With the adapted insane software [349], a broad range of monolayers was modeled. For the coarse-grained only simulation 20x20 nm^2 binary SAMs with anchor and molar fraction of 0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 were modeled. The SAMs consisted of the matrix compound mixed with either of the C8 or C16 anchor compound. For each anchor type, all molar fractions were probed. To enhance sampling, each SAM was built in three replicates with different randomizer initiations. The 20x20 nm^2 SAMs were all in a randomized spatial configuration (no clusters) using a uniform probability density function. SAMs with the purpose of resolution transformation to full atomistic detail, were built to a lateral dimension of 10x10nm^2. They had anchor molar fractions of 0, 0.05, 0.1, 0.2, 0.5, and 1.0. The mixed SAMs were modeled each in

randomized and in aggregated configuration. In the aggregated case, only a single central domain was modeled. For each concentration, and each spatial arrangement, SAMs with C8 and C16 anchors were generated. The boxes were filled with polarizable MARTINI water [388].

7.2.2 Coarse-grained simulations

The coarse-grained simulation protocols largely followed recent suggestions by de Jong et al. [389] and were conducted with GROMACS [75-77, 80, 81, 309] ver. 5.1.4. We used a Verlet neighbor list scheme [390] (updated every 20 steps) with short cutoffs of 1.1 nm for Van-der-Waals and Coulomb interactions. For long range electrostatic interaction, a reaction field potential with an infinite permittivity constant and a dielectric constant of 2.5 was employed [391]. Periodic boundaries were applied in all direction. Minimization was performed over 5.000 steps. A temperature of 298.15 K was initially set according to Maxwell-Boltzmann distribution and controlled by velocity rescaling [350]. Solvent and monolayer were independently coupled with coupling constants of 1.0 ps. NVT equilibration was achieved within 10 ps with 10 fs time step. NPT equilibration was performed for 10 ns with 20 fs time step. Production simulations were run for 60 ns with a 30 fs time step. Semi-isotropic Berendsen [93, 94] pressure coupling was only applied in z-direction by setting the lateral compressibility to 0 and the normal one to $4.5e10^{-5}$ bar^{-1}. Additionally, the thiol-group particles were restraint to their initial positions by harmonic potentials with force constant of 5.000 kJ mol^{-1} nm^{-1}. The restraint was only applied in z-direction and lateral movement was allowed. The reference coordinates were scaled according to box size fluctuations. For production, we switched to Parrinello-Rahman [226] coupling with 12 ps coupling constant. The SAM simulations were conducted in triplicates where each system was individually modeled, minimized, and equilibrated.

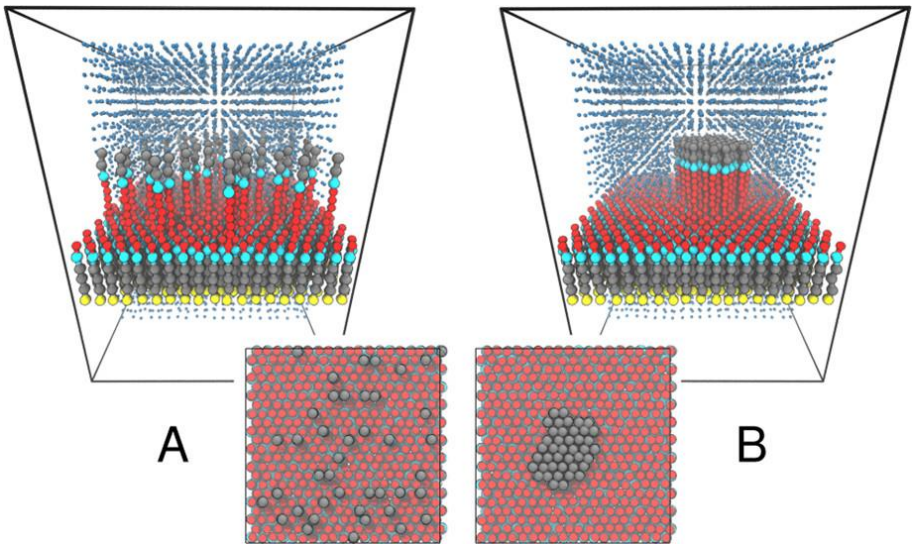

Figure 7.4: Initial packing of mixed SAMs. A) Random arrangement of C8-anchors. B) Clustered arrangement of C16 anchors. Water particles are left out for clarity. Color coding according to MARTINI bead type as in Figure 7.2.

7.2.3 Full-atomistic simulations

Before dependable full-atomistic conformations can be sampled, the initial back-mapped configuration must be refined. Therefore, the protocol as proposed by Wassenaar et al [392] was satisfied: I. 5.000 steps steepest descent minimization of only the bonded interactions, II. 5.000 steps steepest descent minimization with all interactions, III. 1.000 molecular dynamics steps with 0.1 fs time step including Nose-Hoover temperature coupling [225] to a 298.15 K bath with a 1 ps coupling constant (SAM and solvent coupled separately), and IV. 1.000.000 MD steps with a 1 fs time step with Berendsen [93] pressure coupling (coupling constant 5 ps, reference pressure 1 bar, normal direction compressibility 4.5e-5 bar^{-1}). Finally, production sampling was performed for 10.000.000 MD steps with a 2 fs time step. Parrinello-Rahman pressure coupling [226] was employed in normal direction with a coupling constant of 12 ps. All H-bonds were constrained. LINCS [227, 228] constraint solver was utilized. The full atomistic SAMs were simulated with GROMACS ver. 5.1.4. The full-atomistic topologies were generated via CHARMM-GUI [79, 307] and its CGENFF [87] implementation. The CGENFF-based parameters were identical to expected parameters from the CHARMM lipids force field and yielded penalty scores of 32 (charges) and 42 (bonded).

7.2.4 Trajectory analysis

Of each SAM molecule, the distances between its thiol group bead (sulfur atom, in case of all atom simulation) to all the other thiol group beads were computed taking mirror images into account. The average of the six shortest distances was monitored as molecular lattice parameter. The lattice constant was determined as the time and ensemble average of the molecular lattice parameters. The distances were calculated with MDTraj ver. 1.9.3.[230] The collective tilt angle of the monolayer was calculated as the time and ensemble average of the molecular tilt angle. The molecular tilt angle was measured as the angle between the first principal axis of the matrix alkyl chains with the surface normal. The principal axis was determined as the first principal component eigenvector as calculated by scikit-learn ver. 0.22.2 [393]. The thickness is determined from time average normal-direction density profiles of the SAMs and the surrounding solvent. The density profiles were computed with GROMACS tool *gmx density*, in which the box was divided in 50 slabs for the coarse-grained simulations and 100 slabs for the all-atom simulations. From the density profiles, the intercepts between SAM profile and solvent profile were numerically determined. The distance between these intercepts was considered the SAM thickness. The surface hydrophobicity was estimated from the solvent accessible area of the alkyl group beads (alkyl carbon and hydrogen atoms) of the anchors relative to the total solvent accessible area of the SAM. The solvent accessible areas were computed with GROMACS *gmx sasa* using a 0.14 nm probe size. The MARTINI particle Van-der-Waals radii were 0.26 nm and 0.23 nm for standard and small particles, respectively.

Carbon-carbon bond order parameters [188] were calculated using equation **24**. In case of the molecular portion analysis, bond order distributions are shown to avoid artifacts from average over multimodal distributions. Elsewhere, they are shown as ensemble averages. Conformations of distinct parts of the matrix and anchor molecules were determined by certain torsion angles. For the alkyl chains the C-C-C-C torsion angles were determined and the fraction of angles with absolute values of 150° or more is considered the trans fraction [394]. In the OEG portion, the helicity is determined as the number of consecutive, same-directed O-C-C-O bond in gauche orientation normalized by the total number of O-C-C-O dihedrals (here 5). Gauche orientation corresponds to absolute torsion angle between 50° and 150° [394]. Graphics were generated with MatPlotLib [274] and molecular images were rendered with VMD ver. 1.9.3. [266].

7.3 Results

7.3.1 Coarse-grained force field parameters

To minimize the systems and to perform dynamics calculations, force field parameters need to be assigned. In the case of the SAM molecules (matrix and anchor compounds), most parameters could be taken over from already published MARTINI lipid [74, 345], protein [347] and polymer [373] force fields. Thus, the mapping was also mostly predetermined (**Figure 7.2**). C5 beads were used for the thioethyl head groups and C1 beads for the n-butyl groups. The acetamide groups were modeled by P5 beads and the oxyethylene groups of the OEG chain were mapped into SN0 beads (SN2 in case of terminal ethanol groups at the matrix compounds). If note, SN0-SN0 and SN0-P4 non-bonded interactions were made more attractive (to the level of Nda particles). As many of the bonded interactions could be re-applied from the literature, only the amide linkage between alkyl and OEG chain were newly parametrized. Here, the coarse-grained simulations were optimized against united-atom simulation with the GROMOS 53A force field [374, 395]. The final parameters are summarized in **Table E.1** in the appendix. To ensure stability at long integration time steps, dihedral parameters were not employed. To overcome the general unavailability of physically correct gold parameters, the surface was modeled implicitly using a flat surface model [374]. Therefore, the thiol head groups are restrained to a flat plane using a harmonic potential. The utilized force constant of 5000 kJ mol^{-1} nm^{-2} proved to be the best compromise between stability and correct packing.

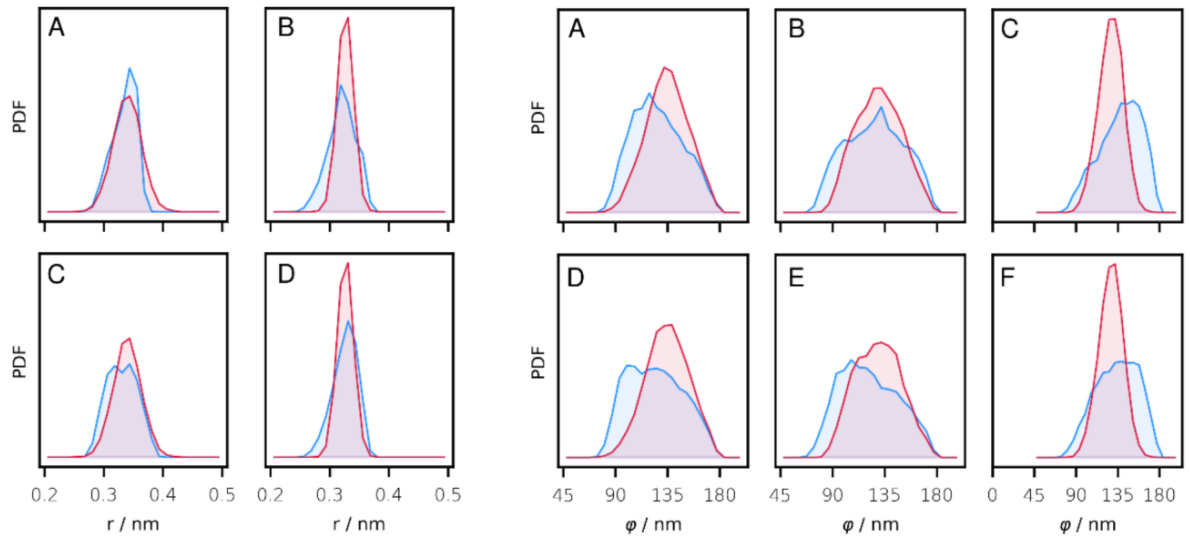

Figure 7.5: Coarse-grained parameter optimization of the anchor amide linkage. The top row corresponds to the parameters at the matrix portion, the bottom row to the anchor portion. The four panels to the left show the bond distance distributions. A and C: amide-alkyl, B and D: amide-hydroxy ethylene. The six panels to the right show the angle distributions. A and D: alkyl-alkyl-amide, B and E: alkyl-amide-hydroxy/oxyethylene, C and F: amide-oxyethylene-oxyethylene.

The established coarse-grained parameters yielded good agreement of the mapped distances and angles as compared to the united atom calculations (**Figure 7.5**). The bond distances were well reproduced. Of the angles, only the one between the amide, and the two succeeding oxyethylene groups produces slightly too small values even when the force constant is chosen high. Here, consistency with the other parameters was considered more important than accuracy, which is in line with MARTINI design principles.

7.3.2 Coarse-grained sampling

To assess the accuracy of new coarse-grained model, 20 x 20 nm² monolayers of increasing anchor mol fractions and both types (C8 and C16) were sampled in triplicates for 60 ns, of which first 10 ns were discarded as equilibration. Structural characteristics, such as lattice constant, tilt angle, thickness and surface hydrophobicity were calculated (**Figure 7.6**). The lattice constant is in fact directly modeled in terms of packing and then effectively restrained by disallowing lateral box size fluctuations. However, lattice constant and tilt angle are correlated in way that a larger pinning distance result in a larger tilt angle [396, 397]. The coarse-grained force-field with its large spherical beads for butyl groups is not capable to accurately reproduce both a lattice constant of ~4.9 nm [396] and a tilt angle of ~30°. Thus, it was decided that the reproduction of the tilt angle antecedes a correct lattice constant, and the molecules were placed with an initial distance of 5.4. This way a mean tilt angle of around 24° for all compositions was realized, which is in good agreement with experimental findings of 20°-35° [398, 399].

For the pure matrix compound monolayers, the CG simulations yield a thickness of 23 Å (**Figure 7.6 C**). The pure C8 anchor SAM has thickness of 46 Å and the pure C16 anchor SAM a thickness of 56 Å. The thickness values are in good agreement to the experiments (25 Å, 49 Å, and 59 Å [54]) and systematically about 3 Å too small. The small discrepancy might be the results of the missing gold substrate. As expected, the thickness increases with increasing anchor mole fraction and is mostly higher for the C16 anchors. Interestingly, for small anchor coverages (x<0.2), the thickness increases slowly, linearly, and similarly for both anchor types. After x=0.2, the thickness soars until it levels out around x=0.8. The slow initial increase followed by a rapid rise can be explained by hydrophobic interactions of the anchor moieties. At low coverages, they lie flat on the SAM surface to minimize water contacts and to be able to interact over long distances. Once a critical concentration is reached, the alkyl chains are allowed to pack side-by-side in a surface-perpendicular orientation. This behavior is resembling that of an assembling monolayer [400]. Interestingly, the aggregating and re-orienting anchor alkyl chains at mol fractions of 0.3 and higher form small islets and domains of larger thickness (**Figure 7.7**). However, the differences between C8 and C16 anchors in their ability to form domains are too small to explain the larger contact angle hysteresis of the C16 anchor and the differences in the IRRAS spectra. Therefore, the partitioning would need to be much more prominent.

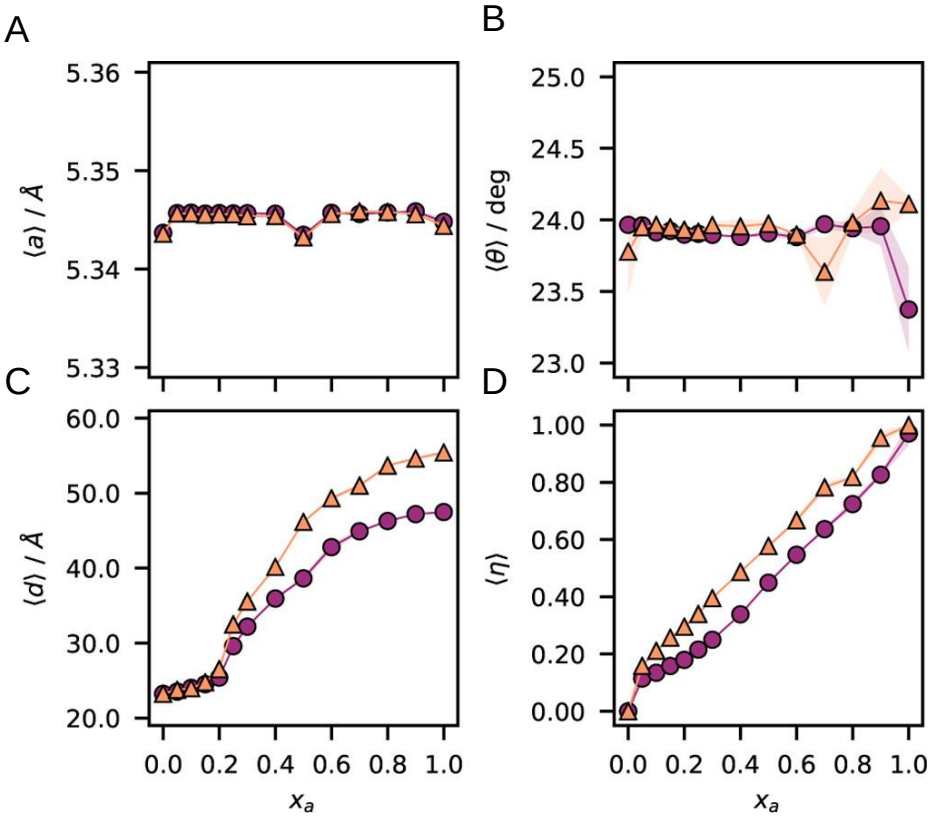

Figure 7.6: Structural characteristics of the mixed SAMs as sampled by coarse-grained MD. A) Lattice constant ⟨*a*⟩, B) tilt angle ⟨θ⟩, C) thickness ⟨*d*⟩, and D) surface hydrophobicity ⟨η⟩ of SAMS with C_8 and C_{16} anchoring compounds as functions of molar fraction. Orange filled circles: C_8 anchors; purple filled triangles: C_{16} anchors. The shaded area is the standard deviation from the mean of three replicate trajectories.

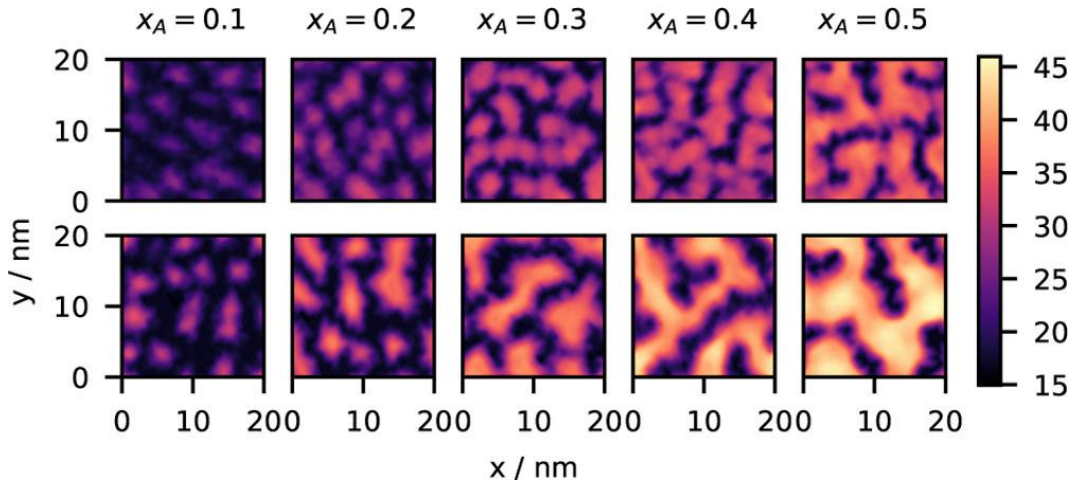

Figure 7.7: Spatial distribution of the monolayer thickness. Results are shown inn Å for increasing anchor mole fractions (from left to right) and the two anchor compounds. Top: C_8, bottom: C_{16}.

To compare the wetting or surface hydrophobicity to the contact angles measurements, the solvent exposure of different chemical groups was calculated **(Figure 7.6 D**). The ratio of solvent accessible area, which belongs to the alkyl group beads relative to the total solvent accessible area is taken as measure for hydrophobicity. The hydrophobicity of the C16 anchor monolayers is systematically higher than that of the C8 anchors. It is however identical for the pure anchor monolayers. The hydrophobicity increases linearly with the anchor mol fraction. The CG model here is in line with experimental observation and intuition. It however cannot capture the contact angle hysteresis. There are methods to simulate contact angles using a water droplet on top of a surface [410-412]. Unfortunately, the polarizable MARTINI force field water model is not suitable for such an approach because the particles do not form stable droplets. To sum up, the coarse-grained results are broadly in good agreements with experimental findings. Tilt angles, layer thickness and hydrophobicity are of the pure monolayers are well reproduced. For the mixed SAMs experimental data is sparse and ambiguous because the true monolayer composition is difficult to determine. The accuracy and reliability of the molecular model can be increased by transforming it to full atomistic representation.

7.3.3 Full-atomistic refinement

For resolution transformation from the coarse-grained representation to full atomistic resolution (**Figure 7.8**), the atoms must be placed at the position of the corresponding coarse-grained particles. This is initially done in a crude way, in which bond distances, angles and dihedrals are just approximated based on the coarse-grained coordinates. The coordinate transformation is followed by a multistep minimization protocol, in which initially only bonded potentials are considered and non-bonded potentials are nullified. This way, bond lengths, angles and dihedrals can relax before

non-bonded atoms will start interacting which each other. Finally, the non-bonded interactions are slowly ramped up [392].

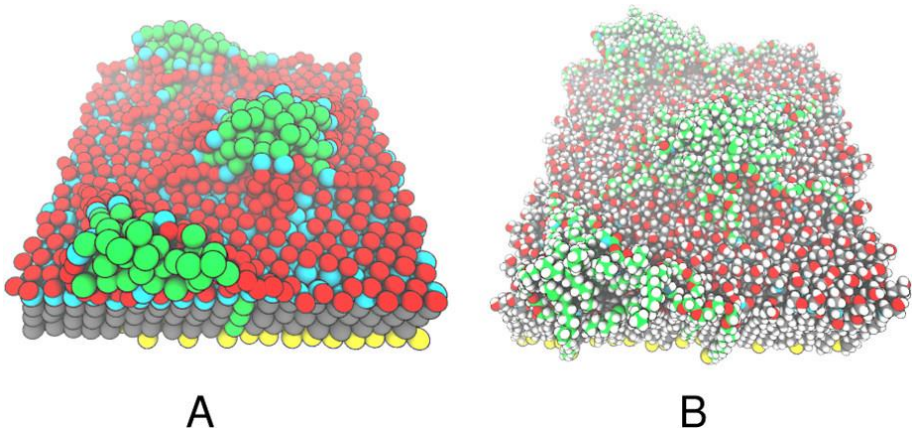

Figure 7.8. Resolution transformation of the SAM model. Shown are snapshots of an x_{C16} = 0.1 monolayer before (A) and after (B) resolution transformation in Van-der-Waals representation. The anchor compound carbon atoms (beads) are colored green.

Again, the structural features of the modeled pure and mixed monolayers were calculated (**Figure 7.9**). The lattice constant shows only minor deviations from the experimentally reported value of 4.97 Å. The tilt angle is decreasing slightly from 27° to 26° with increasing anchor mol fraction based on stark interactions of the anchor alkyl chains. The thickness values of the full-atomistic SAMs resemble the ones of the coarse-grained systems. This highlights the accuracy of the coarse-grained model. The pure matrix compound SAMs have a thickness of 25 Å, the pure anchor monolayers of 51 Å and 61 Å, respectively, for C8 and C16. Oppositely to the coarse-grained model, the full atomistic model slightly overestimates the thickness relative to the experiment, which can be attributed to the decreased helicity of the OEG portion, as described in the next section. Interestingly, the difference in the slope of the thickness over the anchor concentration between C8 and C16 anchor is even more pronounced in the full atomistic model. The thickness of both anchor types is similarly small up to an anchor concentration of 0.1. At higher concentrations, the C16 anchor forms monolayers with a significantly larger thickness compared to the C8 anchor. The interactions between the C16 anchors are stronger so that they can form upright standing aggregates at lower concentrations. Additionally, due to the longer chain length, they are more likely to encounter each other at low concentrations. Similar deviations from a linear increase can be seen for the surface hydrophobicity. Initially, at low anchor concentrations, the values are identical for C8 and C16 anchor compounds. At a mol fraction of 0.2 however, the effect of the C16 anchor becomes stronger. The pure anchor monolayers are both fully covered by aliphatic moieties.

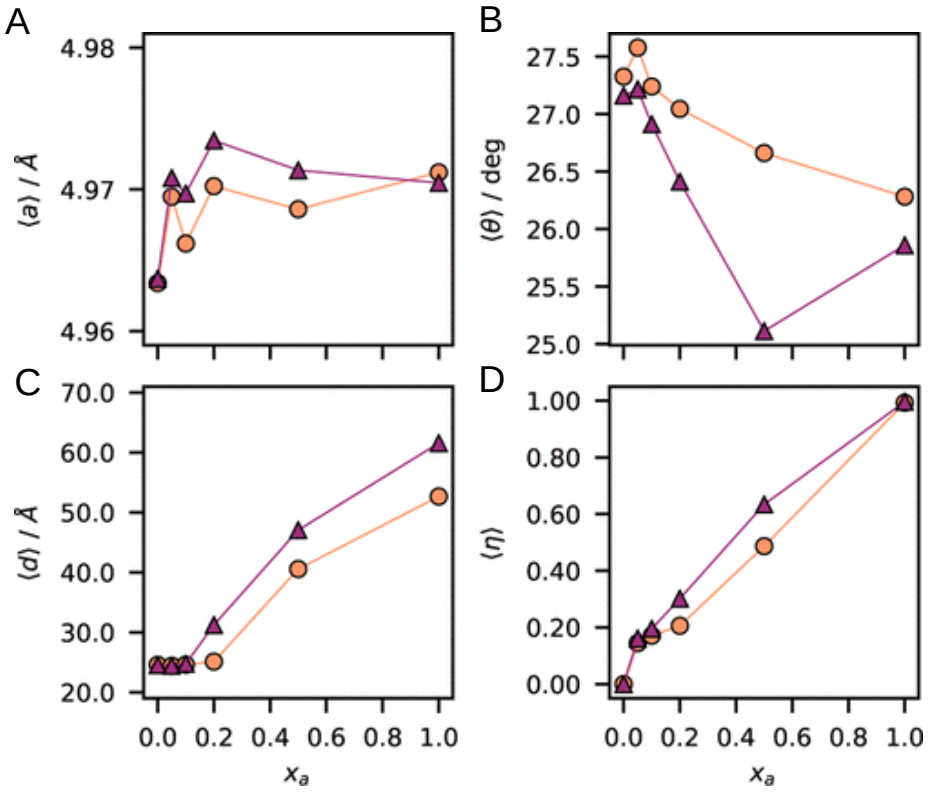

Figure 7.9: Structural characteristics of the mixed SAMs as sampled by all-atom MD. A) Lattice constant ⟨*a*⟩, B) tilt angle ⟨θ⟩, C) thickness ⟨*d*⟩, and D) surface hydrophobicity ⟨η⟩ of SAMS with C_8 and C_{16} anchoring compounds as functions of molar fraction. Orange circles represent the C_8 anchors and purple triangles the C_{16} anchors.

Compared to the coarse-grained model, the full atomistic simulations further increase the accuracy of the simulation procedure and reveals small additional differences and details. In particular, because of the full atomistic refinement, the lattice parameter decreased from 5.35 to 4.97 Å, the tilt angles increased from ~24° to ~26°, and the thickness systematically increased by 2–5 Å depending on the anchor mole fraction, whereas the hydrophobicity of the monolayer surface is not affected.

7.3.4 Conformation and orientation

The outcome of the full atomistic sampling is in such a good agreement to the literature and recent experiments of Lee et al. [54], that the bond orientations of the matrix, OEG and anchor portion should be comparable with the IRRAS spectra. Thus, bond order parameters for the whole molecules were calculated. The bond order parameter is a measure of the average angle between a certain chemical bond and the surface normal. Due to its mathematical description, it maps the angle to a range between 1.0 and –0.5, where 1.0 corresponds to parallel to surface normal (0°) and –0.5 to perpendicular (90°). The bond order parameters are frequently used to compare atomistic models of proteins and lipid bilayers to NMR experiments. Here, it is used to visualize order and orientation of the different molecular portions.

Bond order parameters are exemplarily discussed for a 5% C16 anchor SAM (**Figure 7.10**). Here, in the matrix portion, the C-C bonds have order parameters alternating between 0.1 and 0.6. This means, that the average angle of the even C-C bonds relative to the surface normal is 25°, and the average angle of the uneven bonds is 50°. This corresponds to a defect-free packing of all-trans alkyl chains with a molecular tilt angle of around 30°. In the OEG region, the order parameter decreases but the alternating behavior is maintained for 8 bonds. Afterwards it depends if the anchors are in random or clustered spatial arrangement. In the random arrangement, the order parameter stays as low as –0.1 until the end of the molecule. Such a value might correspond to an average angle of 60° relative to the surface normal but is also an indicator for disorder and thus a broad distribution of bond orientations. The clustered arrangement, however, leads to an increase of the bond order after the 32nd bond and the re-appearing of the alternating behavior. The anchor alkyl chains show an order parameter pattern similar to the matrix alkyl chains, yet less distinct. The example stresses, that while the order parameter is good indicator for the orientation as well as order and disorder, it still suffers the flaws of an arithmetic mean of possibly multimodal or skewed distributions. Therefore, it is advisable to explicitly look at the probability density functions of the bond-normal angles or bond order parameters.

The distributions of CC bond order parameters for the different molecular portions (matrix, OEG and anchor) were calculated from the full atomistic MD sampling and contrasted by anchor length (C8, C16) and spatial distribution (uniform random, clustered). The results are presented in **Figure 7.120**. In the matrix region, the order parameter has two distinct population with means of 0.8 and –0.1 corresponding to the odd- and even-numbered bonds. The populations are well separated, especially for the pure monolayers. In the mixed SAMs, there is a small amount of disorder, which is reflected by smaller broader peaks, which may slightly overlap. For random spatial distribution, the amount of disorder is highest at low anchor coverage, whereas it is reversed in clustered arrangement. Possibly, the interaction of OEG and even more of anchor alkyl portions induce stress ("pull") at the matrix alkanethiols and thus locally change tilt angle and dihedrals. The effect is however small and can be neglected. Generally, the matrix region order parameters show a highly ordered monolayer with a collective tilt angle of 25° and thus alternating bonds of 10° and 70° relative to the surface normal.

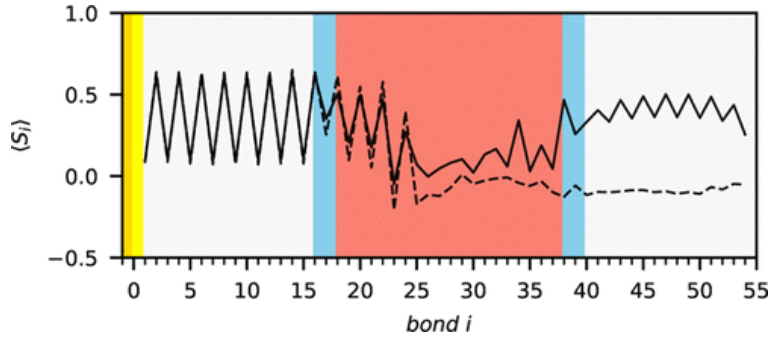

Figure 7.10: Bond order parameters of the C16 anchor at 5% coverage. Solid: clustered configuration. Dashed: random configuration. Background colors represent the molecular areas. Gold and yellow: Surface and thiol group. Gray: alkyl chains. Blue: Amide linkage. Red: OEG.

In the OEG region, disorder is dominant. The distribution are much broader and clear peaks can only be seen in the pure anchor SAMs or the SAMs, in which the anchor coverage is over 50% and the anchors are arranged in a local cluster. For low anchor concentrations and random spatial distributions, the order parameters are uniformly distributed with a slight increase in probability density for order parameters close to –0.5. This is consistent for both the C8 and the C16 anchor. The formation of the peak at 0.5 with increasing anchor mol fraction is indicative of a phase transition from disordered to ordered. The phase transition begins as with 20% anchor coverage when the anchors are locally aggregated. For the random distributed anchors, the initiation occurs only at coverages of over 50% (not probed with MD).

The order parameter distributions of the anchor alkyl chains are the pure anchor monolayers are similar to the ones of the matrix compounds. They exhibit to peaks, one at 0.8 and a second at –0.1. The single peaks of the C16 anchors are even more separated than the matrix compound peaks. The C8 anchor order parameter distribution of the pure anchor SAM shows more disorder than the C16 anchors. In the mixed SAMs with random distribution, the order parameters are centered around a value of –0.5 and follow an approximately uniform distribution elsewhere. The likelihood for higher order parameters close to 1 is lower. It however increases with increasing anchor molar fraction. This effect is identical between the two anchors. Some variety between the order parameters of the monolayers with clustered anchor compounds is noticeable. The 5% clustered C8 anchor SAM has broad distribution with a flat peak at 0.6, which splits into two peaks at higher anchor concentrations. The peaks are at 0.8 and –0.1, similar to matrix compounds. The clustered C16 anchor monolayers exhibit two overlapping populations at 5% coverage, which merge to a single peak at an order parameter of 0.5 for 10% coverage and then separate again at 50% coverage to two peaks at 0.9 and –0.25.

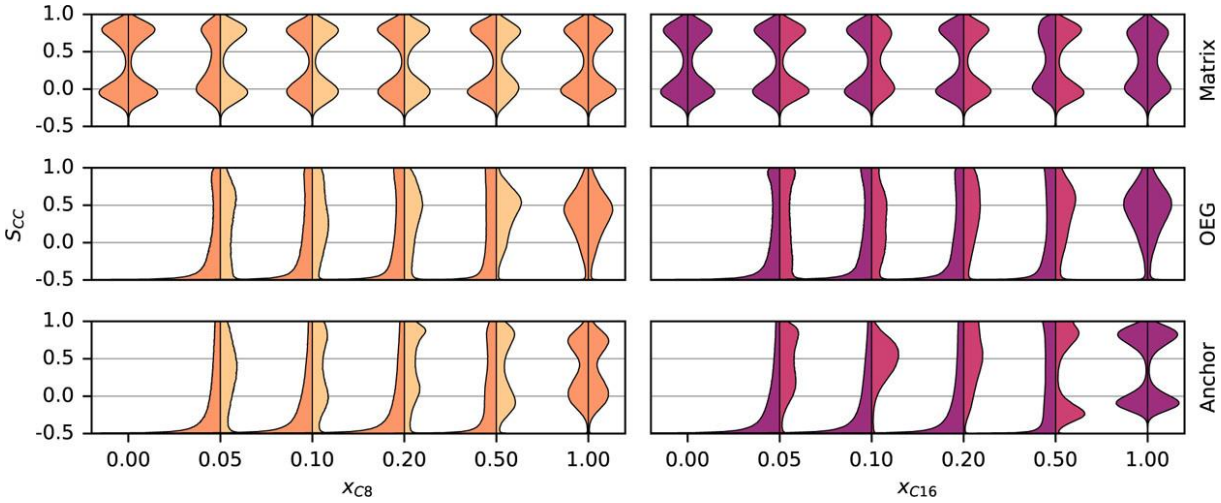

Figure 7.11: Violin plot of CC bond order parameter distributions. The distributions show the different modules of the SAMs depending on the anchor molecule chain length (C_8: orange, C_{16}: purple), spatial distribution (random: darker colored left-hand side; clustered: bright colored right-hand side) at various molar fractions.

While the order parameters reveal much about orientation of the bonds, they still leave ambiguities regarding the conformations of the molecules. Thus, torsion angle distributions were monitored for the three molecular portions and the ratio of *trans* and *helix* conformations were computed (**Figure 7.132**). For both spatial arrangements, the matrix alkyl chains are virtually in all-trans conformations, with a *trans* ratio of 98% independent of anchor molar fraction. The helicity of the OEG portion decreases with increasing anchor coverage. For both random and clustered spatial distributions as well as both anchors, it is around 0.5 at low coverages < 20% and decreases to 0.3 for the pure anchor SAMs. The decrease is linear for random distribution and asymptotic for the clustered distribution. Thus, the minimum helix ratio of 0.3 is already reached at an anchor mol fraction of 0.5 for the clustered anchor SAMs. Helical conformations of PEG chains stabilized by intramolecular hydrogen bonds have been confirmed in earlier experimental [401, 402] and quantum chemical [403] studies. Helical elements in the OEG chain of the anchoring compounds are strongly pronounced in the IRRAS spectral by Lee et al. Similar is observed for the *trans* ratio of the anchor aliphatic moieties. It raises linearly in the random distribution SAMs with a slightly higher slope for C16 anchors. In the clustered C16 monolayers, the anchor *trans* ratio is over 0.9 already at 5% coverage. For the clustered C8 anchors it is systematically lower.

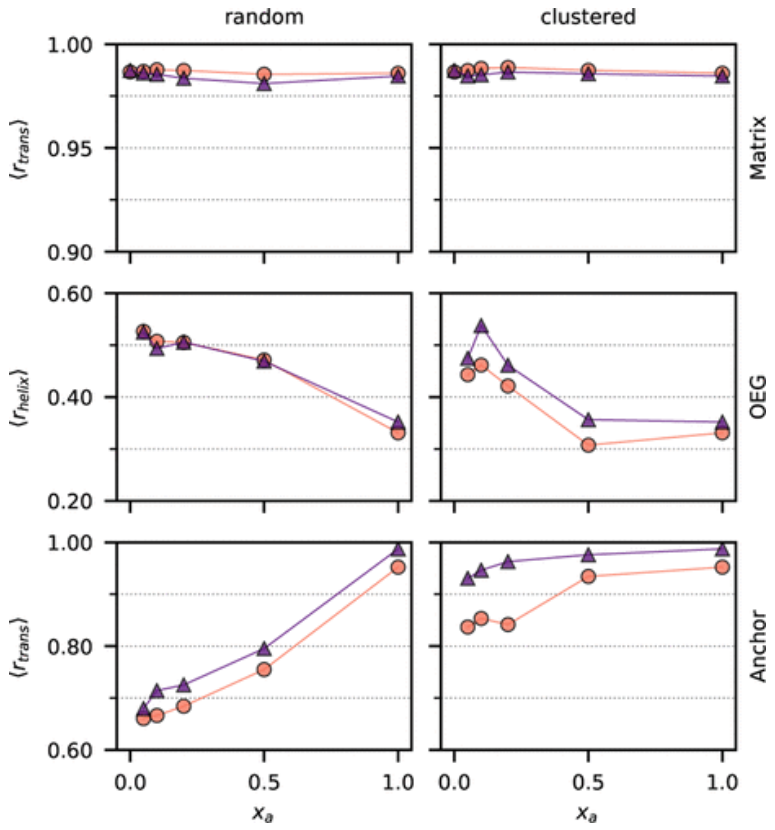

Figure 7.12: Molecular conformations of the different molecular regions. Top: matrix region, middle: OEG, bottom: anchor molecules. C_8 orange. C_{16} purple.

7.4 Discussion

A central part of any computational theoretical approach is the repetitive comparison with experiments. The known variables must be reproduced so that the prediction of the unknown is rationalized. In case of the SAMs, which were initially modeled in coarse-grained representation, subjected to resolution transformation to full atomistic detail and finally sampled for a large ensemble of conformations, such variables are intramolecular conformation and orientation relative to the surface normal. They are experimentally well accessible for the pure SAMs (matrix and anchors) but not for the various mixtures, especially of higher anchor concentration.

The molecular architecture of the matrix region is characterized by highly ordered, densely packed, collectively tilted alkyl chains which attain an all-trans conformation, as proven by the two dominating IRRAS bands at 2918 and 2859 cm^{-1}. In the MD ensemble, only 2% of the dihedrals deviated from trans conformation and the bonds were either oriented in 10° or 70° relative to surface normal. As this is only possible for collectively tilted alkanethiols, the modelling of the matrix portion can be considered to excellently agree with experiment. Orientation and conformation of the matrix compounds are only weakly disturbed by the superficial interactions of OEG and anchors, further highlighting their tight packing and stability.

For the OEG portion, a phase transition can be seen upon increasing anchor molar fraction in both simulation and experiment. In the IRRAS spectra, the OEG portion C-H stretching band at 2890 cm^{-1} s only evident in the pure anchor SAMs, but not in the mixed SAMs. Thus, a conformational and orientational transition must happen which only allows this bond to absorb the polarized light. In the MD ensemble, the transition is twofold. While the helicity decreases from 0.5 to 0.3, the orientation changes from recumbent to upright. Only in the upright orientation, the helical packing can be seen as a number of bands in the IRRAS at 2890, 1464, 1346, 1244, 963, and 115 cm^{-1}, even if the number of helical elements is reduced. Additionally, the helical elements might be more pronounced in experiments because they are measured in the gas phase. Here, the helical conformation of OEG is favored because surrounding water molecules cannot disturb the intramolecular hydrogen bonding network of such an arrangement. This is also reflected in the slightly underestimated monolayer thickness at higher anchor concentrations. The reason for the decrease in helical elements at high anchor concentration are steric clashes. The anchor moieties pack so tightly, that the OEG elements must adopt a more extended conformation. Favorable intramolecular interactions that were the driving force for the helical arrangement are displaced by intermolecular interactions.

The IRRAS spectra of the deuterated alkyl anchors yielded substantial differences between the long C16 and the short C8 anchor. In the pure C8 anchor SAMs, the CD2 asymmetric and symmetric stretching bands are weaker than the CD3 asymmetric and symmetric stretching bands. The order parameter distributions show two peaks for both C8 and C16 anchor monolayers. However, the peaks are significantly more separated for the C16 anchors and more overlapping for the C8 anchors. Thus, both IRRAS and MD reveal a substantial amount of disorder only in the pure C8 SAM and not in the C16 SAM. For anchor molar fractions of 0.5 and below the anchoring chains are disordered independent of the length. This can be seen in the broad distribution of order parameters and the low trans ratio. In the IRRAS spectra the effect is reflected by the relative band intensities of the CD2 and CD3 bands, which become more alike with lower anchor concentration. Most striking are the differences between C8 and C16 for the clustered, mixed SAMs. While the clustered C8 anchors are still substantially disordered, and phase transition begins only at molar fraction of 0.5, the C16 anchors already begin to reorient at a molar fraction of 0.5. At a molar fraction of 0.1, the order parameter distribution becomes unimodal, which is only possible if the anchors are oriented parallel to the to the surface normal, and both odd- and even-numbered bonds exhibit an identical absolute angle (relative to the surface normal). With higher anchor concentration, the distribution broadens and separates, because the anchors begin to collectively tilt. This early-onset phase transition, only of the clustered C16 anchors and not the C8 anchors, strongly favors the hypothesis that domain formation is exclusive to the C16 anchors (**Figure 7.13**).

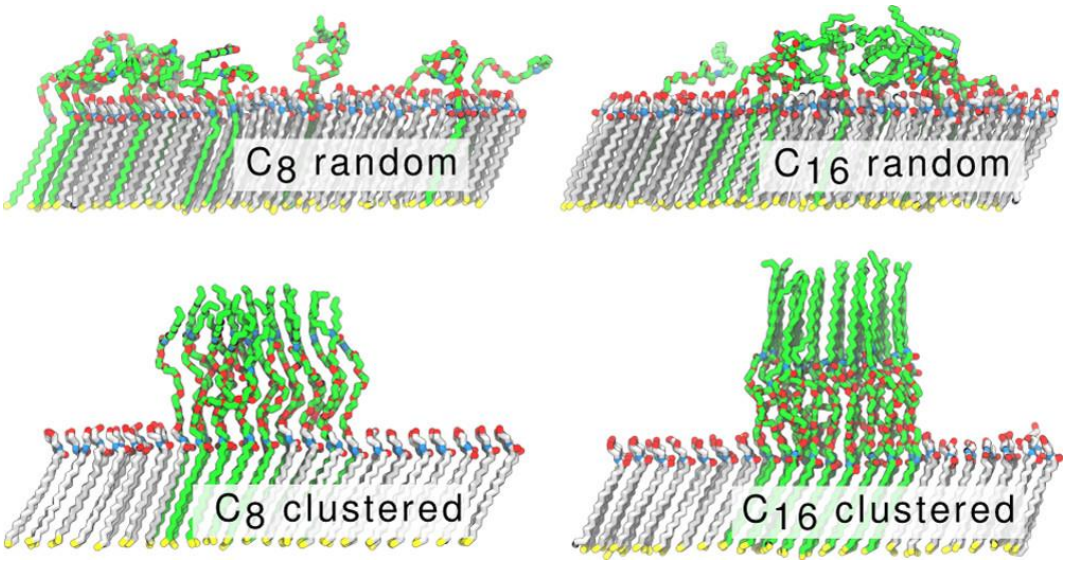

Figure 7.13: Representative snapshots of C_8 and C_{16} monolayers at x_A = 0.1 after 20 ns full atomistic simulations. Only a 1 nm slab is shown in this representation. Molecules are displayed as licorice. Hydrogen atoms are omitted for clarity. The matrix compound is displayed with gray and the anchor compound with green carbon atoms.

7.5 Conclusion

SAMs provide an ideally suitable and well-controllable microenvironment for the incorporation and immobilization of large biomolecules. Matrix molecules and long-chain anchoring components consisting of OEG (here 6 units) and C8 and C16 long alkyl chains provide the possibility for tethering phospholipid bilayers and vesicles. The design and setup of a consistent simulation protocol for full atomistic simulations of mixed self-assembled monolayers on flat model surfaces, based on previous CG initialization, modeling, and equilibration, were developed. The results of both the CG and the full atomistic models agree well with the experimentally reported structural parameters of the self-assembled monolayers. We demonstrate the feasibility of a CG model to bring any mixed SAM system including the membrane-anchoring compounds into thermodynamic equilibrium to start subsequent full atomistic simulations. By simulating diverse types of C8 and C16 anchor monolayers over the complete range of molar fractions, significant differences in packing and interactions can be observed. The structural parameters such as lattice constant and tilt angle are almost independent of the type of the anchoring molecule. The anchor alkyl-chain orientation, however, is critically controlled by the composition and its chemical nature. Although the short chain C8 anchor molecules adopt a random, disordered conformation at low anchor densities and undergo a transition into a more ordered conformation with increasing anchor density, the C16 anchors already adopt highly ordered conformations at low anchor densities, which reoccur in the monolayers of higher anchor density. The results from our MD simulations are in excellent agreement with current experiments by Lee et al. and explain the observed spectral features. In addition, the simulations are able to probe mixed concentrations and mole fractions

that are experimentally hard to access. Complementing the experimental IRRAS measurements, our simulations results provide additional insights into the arrangement and degree of ordering of the anchoring compounds in the monolayer. Depending on the initial spatial anchor compound distribution, we find distinct orientational and conformational phase transitions. Our simulations support the hypothesis that the C16 anchor compound exclusively forms self-aggregates, whereas the C8 anchor compound scatters randomly in the matrix.

8 MONOLAYER-VESICLE AFFINITY

One method to prepare tethered bilayer membranes is to adsorb lipid vesicles on mixed self-assembled monolayers (SAMs), which will rupture upon reaching a critical concentration and form a lipid bilayer. In this approach, the composition of the SAM is critical. Without any anchoring compounds the lipid molecules are not able to form stable interactions with the SAM and are washed away. With an overly high number of anchoring compounds, on the other hand, hydrophobic interactions are so dominant that lipid molecules will form another monolayer, in which the fatty acid tails in are in contact with the exposed anchor alkyl chains. Only with the adequate anchor compound density, a bilayer will assemble where the anchor alkyl chains intercalate with the phospholipid tails in bottom leaflet. The adequate density is also affected by the anchor chain length and the spatial distribution of anchor molecules. In this chapter we employ the previously developed coarse-grained model for mixed SAMs to investigate differences in the vesicle adsorption as consequence of the SAM composition. The chapter is the result of a joint work of the Department of Nanoengineering at the Center for Physical Sciences and Technology in Vilnius, Lithuania with main experimental contributions by Martynas Gavutis and the Group for Molecular Simulation and Design at the Max-Planck-Institute, Magdeburg. Our contributions are the design, execution, analysis, and discussion of the MD simulations. The chapter is prepared for publication to Journal of Physical Chemistry C with M. Gavutis and E. Schulze-Niemand as shared first authors.

8.1 Introduction

Lipid bilayer membranes and membrane proteins are experimentally challenging to study because they are difficult to isolate and stabilize [404]. Existing biophysical techniques include lipid nano-discs [405-407], unilammellar vesicles [408, 409], black lipid membranes [410, 411] and supported lipid bilayers [363, 412, 413], all of which are accessible to different analytical methods. Lipid nano-discs are mostly used to study membrane proteins by means of Cryo-EM [414], NMR [415] and SAXS [416] but do not allow extensive insight on the bilayer itself. Unilammellar vesicles vary in size between tens of nanometers and several micrometers and can be prepared in large quantities [417]. They are frequently studied via NMR [418] and flow cytometry [419]. Black lipid membranes are useful for electrochemical surveys to study membrane proteins such as ion channels [420, 421]. However, they are usually short lived and not suitable for necessary long-time experiments. Supported bilayers are flat layers that sit on top of a solid surface and are thus highly stable and resistant to high flow rates and vibrations (unlike black lipid membrane) [422, 423]. In fact, they are feasible to use in experiments that last even weeks to months. Furthermore, pore formation does not destroy the bilayer [363]. Supported bilayers can be well studied using surface chemistry-

and physics-based methods such as AFM [424], IRRAS [425], and QCM [426], as well as surface plasmon resonance and TRIF [368, 369]. Therefore, the bilayer membrane is attached to artificial thiol-lipids which adsorb onto a gold surface [361]. In some instances, the gold surface is fully covered by a self-assembled monolayer of alkanethiols, which are partially functionalized with membrane anchors [373]. The bilayer will then self-aggregate on top of the monolayer with the membrane anchoring portions intercalating into its fatty acid regions (**Figure 8.1**). In practice, this is achieved by flooding the SAM with vesicles that will rupture upon a critical concentration is reached. This process is also called vesicle fusion [427, 428]. Even though this method is central for tBLM formation, literature is scarce. Naturally, kind of SAM and the degree of functionalization, i.e., the concentration of anchoring molecules, will have a tremendous effect on the bilayer in terms of formation but also stability and equilibrium properties such as curvature, thickness, lipid area and lateral diffusion coefficients.

While global characteristic of the SAM and bilayer, as well as the architecture of the tethered bilayer membrane can be well studied experimentally, a molecular-scale description at the interfaces: solvent-anchors, anchor-anchor and membrane-anchor remain elusive. With current developments such as surface functionalization, the SAM molecules become more complex and dynamic. SAM-SAM as well as SAM-solvent interactions affect the structure and often lead to unexpected experimental observations [54]. Thus, the number of articles featuring MD simulations of SAMs steadily increased in the past decade [374, 376, 429-432]. MD aids the understanding especially of structural dynamics of solvent exposed, functional groups. The accuracy of developed coarse-grained modeling and simulation protocol for mixed SAMs [55] proves its feasibility for further extensions such as a tethered lipid bilayer membrane. Even more so, the initial adsorption of a vesicle to a mixed SAM can be well modeled using the coarse-grained protocol.

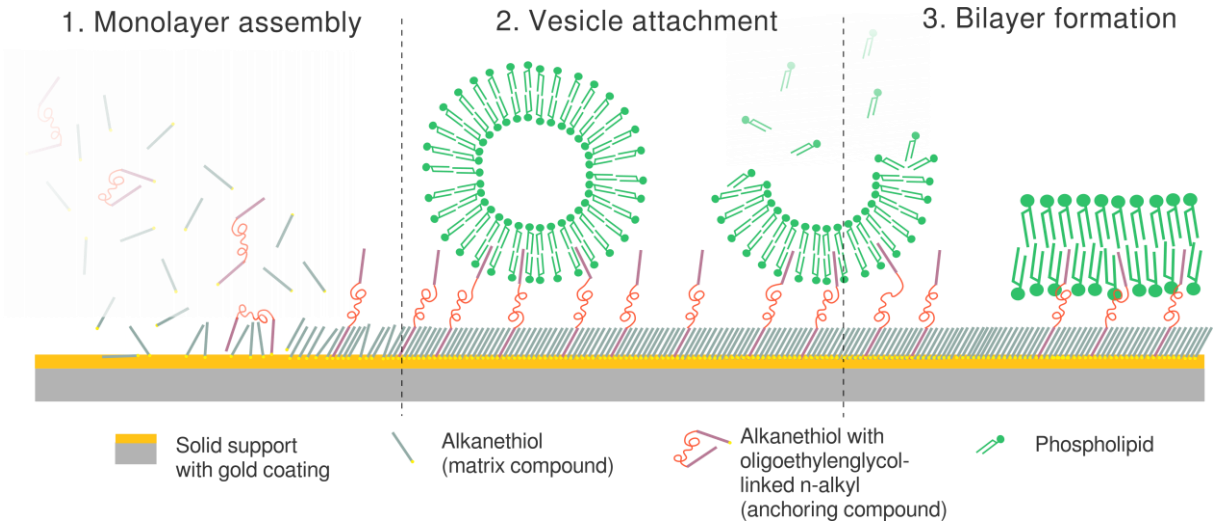

Figure 8.1: Schematic representation of the formation of tethered bilayer by vesicle adsorption and fusion.

Here, a comparison of a lipid vesicle attachment and equilibrium anchor dynamics yields valuable insight into the formation of tethered bilayers as well as the response of a lipid vesicle to physical stress due to deformation. Different SAM composition may be probed to theoretically optimize the tethered bilayer formation protocol in terms of duration and bilayer stability and comparability to a non-tethered, supported lipid membrane. The aim of the present study is to elucidate the interactions between model lipid vesicles and the linear tethers in the process of tBLM formation. For this purpose, we chose a tandem of two complementary techniques: quartz crystal microbalance with dissipation monitoring (QCM-D) and molecular dynamics (MD). The first technique reveals the events related to the mass of the adsorbed material and probes the lipid states on the substrate. While QCM-D allows monitoring real time events at a scale of seconds and minutes, the molecular level information about the immediate tether-vesicle interactions can be efficiently derived from MD simulations. In combination, the experimental model surface system, the QCM-D analysis, and the computational means provide new insights into small unilammellar vesicle (SUV) interactions with membrane-supporting molecular assemblies.

8.1.1 QCM-D Studies

Experimentally, the vesicle adsorption and fusion process can be monitored using QCM-D [433]. With the method, lipid vesicles and lipid bilayers can be well distinguished. Here, the molecular mass affects the frequency of the crystal, and the viscosity is reflected in the dissipation. Generally, during the formation of a bilayer membrane from vesicles, the frequency will initially rise from zero to a maximum, then decrease until a plateau is reached. The dissipation follows a similar pattern. However, the onset is earlier, and the curve is flatter. The initial increase corresponds to the loading of vesicles until rupture. Then, the free lipid molecules are washed away, and the tethered bilayer remains. It is of interest how the SAM composition in term of anchor concentration and length (and thus spatial distribution), will affect peak height, which corresponds to the critical vesicle concentration and the time it takes to reach it. Such studies were recently performed by M. Gavutis and colleagues using mixed alkanethiol SAMs which include different concentrations of membrane tethering/anchoring compounds with two chain length herein called EG_6AC_8D and $EG_6AC_{16}D$ (**Figure 8.3 A**).

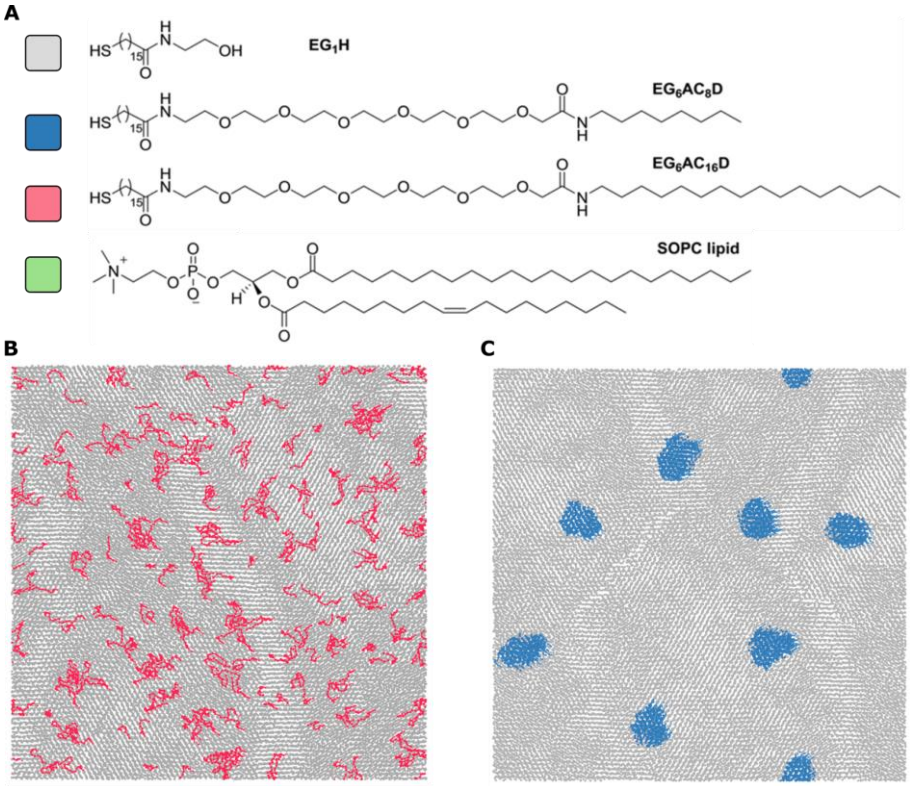

Figure 8.2: System components and SAM packing. A) Molecular structures of the included molecules and subsequently used color coding. B and C) Top view on the employed SAMs with the EG6AC8D and EG6AC16D tethered, respectively.

The titrated vesicles are composed of 1-stearoyl-2-olyeolyphosphatidylcholin (SOPC). Their QCM-D sensograms are shown in **Figure 8.3**. Interestingly, at the anchorless SAM, the vesicle fusion peak is never reached. Thus, without anchors the vesicles do not rupture. For anchor compound molar fractions of 0.01, 0.05 and 0.1, the maximum frequencies are 65, 50, and 35 Hz and it takes 800, 500 and 450 s to reach them. After rupture, the frequency stabilizes to 30 Hz, indicating the formation of similar bilayers. For 0.15 anchor molar fraction, the peak is absent, and the vesicles do not accumulate excessively and only few lipid molecules are washed away. At even higher anchor coverages, the final frequency of 30 Hz is not reached, and it is likely that bilayer formation is aggravated. Instead, the lipid molecules form a monolayer on top of the hydrophobic anchors, which cover a substantial amount of the SAM. The same appears for the pure anchor monolayer, however, due to the flow, the supported monolayer is unstable, and a high proportion of the lipids are washed away. While a dynamic structural modeling of vesicle attachment, rupture and bilayer formation is not accessible- even in the coarse-grained resolution, valuable information may be gathered from equilibrium simulations of the vesicle anchored to the mixed SAM.

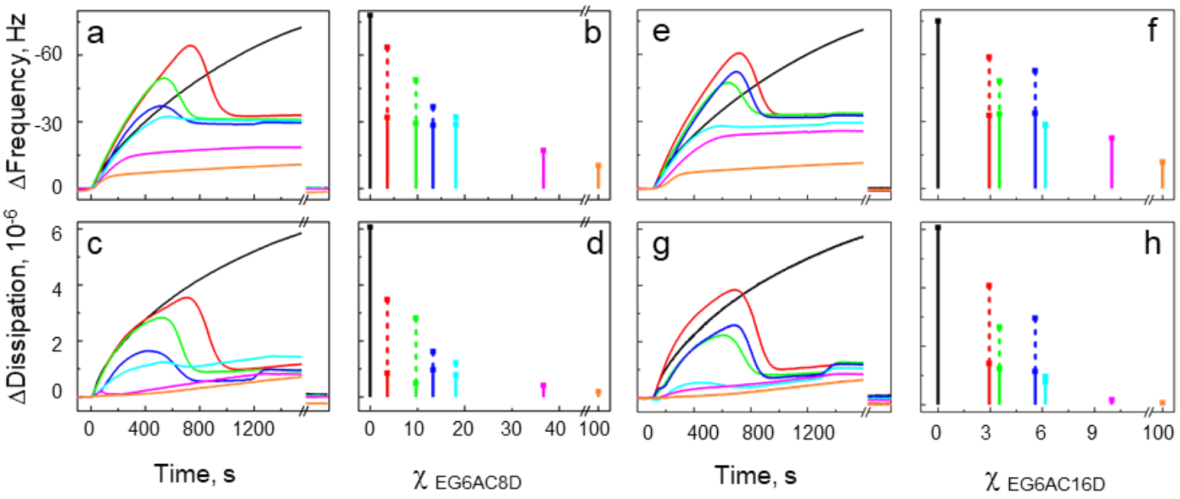

Figure 8.3: QCM-D sensograms of vesicle fusion to the mixed SAMs. The data for the EG6AC8D tethers are shown in panels a, b, c, d. Panels e, f, g, h show the results for the EG6AC16D tethers. a, c, e, g present sensograms, b, d, f, h show the peak and terminal values of the corresponding sensograms.

8.2 Method

Details regarding the experimental studies will be published in our joint article. The computational methods are given below.

8.2.1 System setup and configuration

The lipid vesicles interacting with the mixed SAMs were computationally modelled and subject to long time-scale coarse-grained (CG) molecular dynamics (MD) simulations, each consisting of a 30 nm diameter SOPC vesicle and three different compositions of the SAMs. An overview of all the modeled and simulated systems with molecule numbers and simulation times is given in **Table 8.1.** The vesicle model was generated using the CHARMM-GUI MARTINI maker [434]. Of note, the vesicle employed the palmityol-oleolyl-phosphatidylcholin (POPC) coarse-grained lipid definitions which would be identical to the ones of SOPC and only the one for POPC does exist. The mixed SAMs were modeled using our protocol as introduced in the mixed SAM study [54] (see chapter 7), where the SAM-forming molecules are positioned on a hexagonal grid and the tethering compounds can initially be placed either in a random spatial distribution or into domain-like aggregates (**Figure 8.2 B**).

Table 8.1: Overview of simulated systems. F is the force constant with which the vesicle was dragged towards the SAM.

ID	Name	F / kJ mol^{-2} nm^{-2}	Matrix molecules	EG_6AC_8D Tethers	$EG_6AC_{16}D$ Tethers	SOPC Lipids	Simulation time / ns
1	Tetherless	100	7020	0	0	7532	500
2	Tetherless	1000	7020	0	0	7532	500
3	5% C8	100	6669	351	0	7532	500
4	5% C8	1000	6669	351	0	7532	500
5	5 % C16	100	6669	0	351	7532	500
6	5% C16	1000	6669	0	351	7532	500

The first of the modeled systems consists of an anchorless SAM (formed by EG_1H), the second one is enriched with 5 mol % of the EG_6AC_8D tether compounds in a random spatial distribution, and the third one is enriched with 5 mol % of $EG_6AC_{16}D$ tether compounds in a clustered configuration. These correspond to the experimentally observed configurations of the matrix and the EG_6AC_8D and $EG_6AC_{16}D$ tethering molecules. The self-assembling monolayers were modeled to a cover a lateral area of 42 x 42 nm². Both constituents (SUV, SAMs) were independently solvated with a MARTINI polarizable water model [405] and concatenated in the z-dimension. The initial distance between the vesicular center and the monolayer surface was set to 24 nm in the z-direction. We note that the lipid vesicle, especially the number of lipids per leaflet, was fully relaxed before simulation of the entire system. The vesicle relaxation protocol was adopted from Hsu and coworkers [455]. For subsequent simulations, we use the standard polarizable MARTINI mappings and parameters [75, 359, 363]. For the SAM molecules, the previously optimized parameters were used [55].

8.2.2 Molecular Dynamics Simulations

After merging the lipid vesicle (SUV) and the SAM of the chosen composition, the entire system was carefully minimized and equilibrated according to the following protocol. First, 10 steps of the steepest descent minimization were conducted without domain decomposition for improved stability. Second, 500 steps minimization using the steepest descent algorithm with regular domain decomposition were performed. Third, a short equilibration for 100.000 steps in an NVT ensemble with a time step of 0.01 ps was conducted. Fourth, an NPT ensemble equilibration (Berendsen coupling [93], coupling constant: 5 ps, compressibility: 4.5e-5 bar^{-1}) for 200.000 steps with a 0.02 ps time step was connected. In the NPT simulations, the pressure coupling was only allowed in z-dimension to follow closely the SAM simulation protocol [54], which is based on a fixed lateral area. The initial velocities were generated according to a Maxwell-Boltzmann distribution at 303.15 K. The temperature was controlled with velocity rescaling [350] using a coupling constant of 1 ps.

Vesicle, SAM and solvent were each coupled independently. Other parameters were set according to recent suggestions [389]. If not stated otherwise, we applied reaction field electrostatics [435].

After equilibration, the vesicle was gently pulled toward the SAM by applying a constant, periodic (wrapped around box limits) external force between the PO4 beads of the phospholipids and the THIO beads along the monolayer normal direction. The force constant was set to 100 kJ mol^{-1} nm^{-2} or 1000 kJ mol^{-1} nm^{-2}, respectively. For the steering part, 4.000.000 steps (time step: 0.025 ps) were simulated. Pressure coupling was realized with the Parrinello-Rahman [226, 313] method using a 12 ps coupling constant. Finally, the external force was released, and the vesicle was allowed to relax for 500 ns (20.000.000 steps with a 0.025 ps time step). All simulations were conducted with GROMACS version 5.1.5 [75-77, 80, 81, 309].

8.2.3 Trajectory analysis

During the simulation, we carefully monitored the vesicle particle sphericity and two sets of SAM-vesicle contacts. The vesicular sphericity is calculated for the inner and outer leaflets and then averaged. We here define the leaflet sphericity as:

$$s_{leaflet} = 1 - \frac{\sigma(r_{PO_4})}{\mu(r_{PO_4})} \quad (28)$$

where σ is the standard deviation, μ the mean, and r are the radial distances of the leaflet PO_4 beads, i.e., the distance from the center of geometry of the vesicle. This way, the value would only become one if all PO4 beads were perfectly spherically distributed. As distance and interaction criteria for the vesicle approaching the SAM, two classes of contact intermolecular constants were monitored. The first set corresponds to the hydrophilic vesicle-SAM contacts and is defined as the number of interactions between the SOPC head group beads (PO_4 and NC_3) and the terminal beads of the matrix compound molecules. The second set is a measure for hydrophobic vesicle-SAM contacts and is calculated as the number of contacts between the SOPC phospholipid tail beads and the tether alkyl chain beads. The number of contacts were computed using the GROMACS tool *gmx mindist* with a distance cutoff of 0.6 nm. Visualization and rendering was performed with VMD ver. 1.9.3. [266].

8.3 Results

8.3.1 Vesicle adsorption

We have employed molecular dynamics simulation to rationalize the findings regarding the effect of the tether length and coverage (surface density) on the vesicle rupturing kinetics. Computational resource restraints disallow modeling of the entire process of the multiple-vesicle adsorption until their rupture upon critical mass attainment. Thus, a steered MD workflow was developed in which

the steady state of a tethered vesicle was approximated from two directions. The steady state here corresponds to the amount of tether alkyl chains that are inserted into the vesicle. In one set of the simulations, the vesicle is only weakly directed toward the surface of the SAM (steered MD force constant k=100 kJ/mol), so that the insertion of the tether must happen spontaneously in the subsequent equilibrium MD (k=0 kJ/mol). In the second set of the simulations, the vesicle is force-squeezed (k=1000 kJ/mol) onto the SAM, so that more tethers are inserted than in the steady state (**Figure 8.4**). Thus, here, the tethers must extract from the vesicle spontaneously in the equilibrium MD. After infinite sampling both approaches should yield the same equilibrium. However, as the transitions are extremely slow, it is expected that such a perfect sampling cannot be reached.

During steered attachment and relaxation phase, the vesicle shape in terms of sphericity, as well as the number of hydrophilic and hydrophobic interactions were monitored. Hydrophilic interactions are superficial interactions between the phosphocholine head group and the oligoethylenglycol and terminal ethanol group particles of the SAM. Hydrophobic interactions are only achieved by penetration of the anchor alkyl chain into the vesicle bilayer fatty acyl core. Expectedly, depending on the initial attachment force, the outcome after the relaxation is different. Thus, both force constants can only approximate the true vesicle binding mode.

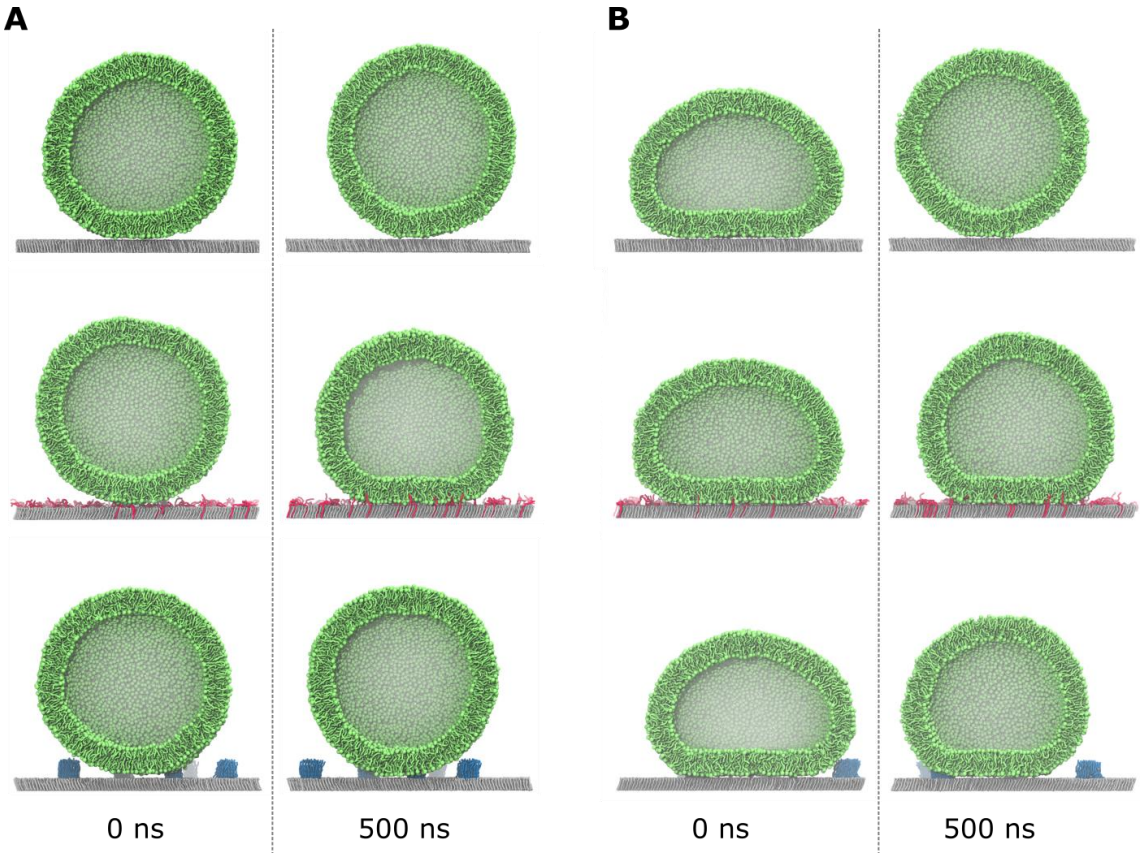

Figure 8.4: Vesicle adsorption and relaxation protocol. A) Low pulling force constant (100 kJ mol^{-1} nm^{-1}). B) High pulling force (100 kJ mol^{-1} nm^{-1}) Left to the dotted line: After pulling (begin of relaxation). Right to the dotted line: After 500 ns relaxation. Top: anchorless SAM. Center: 5% C8 tether SAM. Bottom: 5% C16 tether SAM.

In the case of the tetherless SAM, the SOPC vesicle stays in proximity to the SAM surface after adsorption independent of the attachment force constant. After the release of the force, it quickly relaxes to nearly perfect sphericity (**Figure 8.5 A**). In equilibrium, the SOPC vesicle adsorbs only weakly on the tetherless SAM surface, as shown by few transient interactions mediated by the headgroups of the lipids and SAM surface (**Figure 8.5 B-C**).

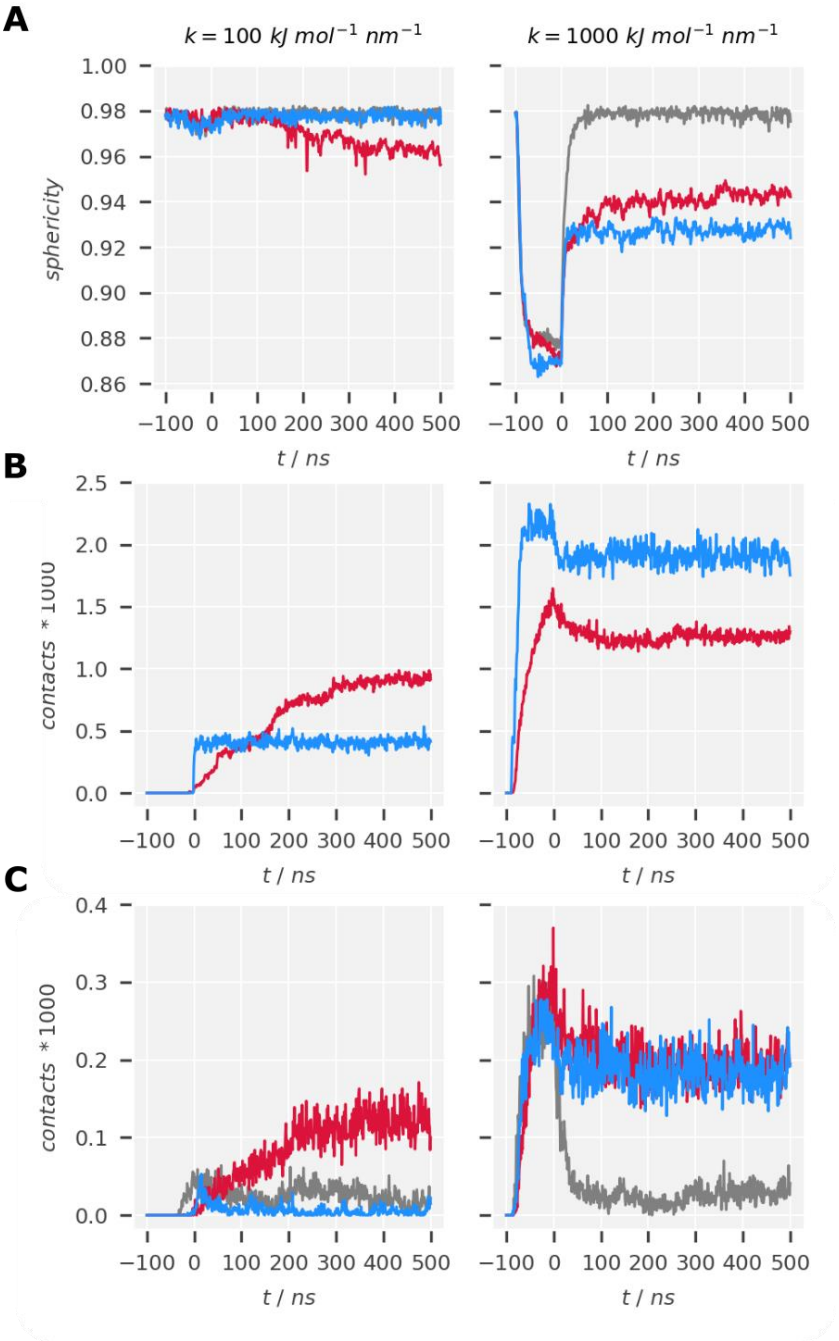

Figure 8.5: Vesicle shape and interactions during adsorption and relaxation. Two different force constants are shown: left column: 100 kJ/mol nm, right column: 1000 kJ/mol nm. A) shows the sphericity. B) shows the polar interactions of lipid head groups and SAM OEG and ethanol terminal groups and C) the apolar (hydrophobic) interactions of anchors and lipid tails. The color corresponds to the nature of the SAM. Gray: anchorless. Blue: 5% C16 anchors. Red: 5% C8 anchors.

Upon adsorption of the vesicle on the randomly distributed EG_6AC_8D tethers ($\chi_{EG6AC8D}$= 5 mol%,), the initial deformation is identical to the tetherless and the $EG_6AC_{16}D$ SAM. Using the weak force constant for attachment, no tethers are integrated into the vesicle fatty acid region (**Figure 8.5 B**). With the high force constant, however, already 1500 pairwise interactions are monitored until the end of the attachment step. Upon release of the force, the weakly attached vesicle begins to deform by spontaneous insertion of increasing amounts of EG_6AC_8D tethers. The sphericity converges to a decreased value of 0.96 whereas the number of hydrophobic contacts increases to 900. Similarly, the hydrophilic contacts between the surface beads of the SAM and PO4 beads of the vesicle increase to a value of 140. Thus, the insertion of more tethers drives a deformation of the vesicle, however, only until a certain equilibrium is reached. At this point, the favorable energy contribution of desolvation of the hydrophobic tethers is identical to the unfavorable energy induced by vesicle shape distortion (surface tension).

Starting the equilibrium MD from the forcefully attached vesicle (k=1000 kJ/mol) yields an even more squeezed vesicle with a sphericity of 0.94, which remains stable during the simulation. The number of polar interactions immediately decreases after release of the steering force and stabilizes to a value of 200. The number of hydrophobic interactions, which is a measure of inserted tether moieties, slowly decreases after the force release to a value of 1300. Thus, the achieved equilibria are slightly different (sphericity 0.96 vs. 0.94, polar contacts 140 vs. 200, hydrophobic contacts 900 vs. 1300). However, the tendency suggests they would converge into each other with further sampling. Thus, we conclude that insertion and extraction of randomly distributed, single EG_6AC_8D tethering molecules would happen spontaneously and is only limited by the surface tension of the spherical vesicle.

As mentioned above, our previous data suggests a clustered spatial distribution of the $EG_6AC_{16}D$ tethers in the mixed SAMs. Interestingly, on the SAM presenting the $\chi_{EG6AC16D}$= 5 mol% surface density of the $EG_6AC_{16}D$ tether, the vesicle behaves significantly different as compared to the EG_6AC_8D SAM. After a weak initial attachment, it quickly relaxes to its spherical shape. The number of polar interactions stays low – even lower as in the case of the tetherless SAM. At the end of this attachment phase, one of the $EG_6AC_{16}D$ tether domains is inserted, leading to 400 contacts between the alkyl chains of the tether and SOPC, respectively. This number of contacts is stable until the end of the trajectory. Thus, no further tether domains get spontaneously inserted. The discrepancy between the hydrophobic and polar contacts leads to the conclusion that the vesicle is elevated by the $EG_6AC_{16}D$ tethers. This way, it is stably tethered in a distance from the SAM and superficial interactions between the SAM and vesicle are absent. When the vesicle is forcefully squeezed onto the EG6AC16D SAM, it shows features similar to those on the EG_6AC_8D

SAM. The number of the inserted $EG_6AC_{16}D$ tethers is initially high, and then decreases after force release. One peripheral $EG_6AC_{16}D$ domain was spontaneously extracted from the vesicle. The differences between the equilibria at the $EG_6AC_{16}D$ SAM are even more pronounced than at the EG_6AC_8D SAM. Apparently, the energy barrier of penetration of the large $EG_6AC_{16}D$ tether domains through the hydrophilic headgroup region of the vesicle is much higher. Additionally, the ratio between solvent accessible surface and volume of a tether domain is lower than for individual tethers. Thus, the energy gain through desolvation is lower as compared to the EG6AC8D tethers.

8.3.2 Vertical architecture

Further insight on the equilibrium vertical architecture of the SAM-tethered lipid vesicles can be gained from the normal direction density profiles (**Figure 8.6**). Here, we use the simulations with the high force constant because they established a larger interface area between SAM and vesicle. The density profiles reveal that the architecture of the matrix region is identical for all three SAMs and the thickness is 2 nm. The first solvent peak, corresponding to the first layer of interface water is most pronounced for the tetherless SAM and most ambiguous for the EG_6AC_8D SAM. The second solvent layer is noticeable for all three SAMs, yet it is significantly less clear as compared to the first solvent layer. The third solvation layer is only identified at the tetherless SAM. This can be explained by the small interface area, which leads to a largely undisturbed SAM surface. In any instance, the total interface water layer thickness is around 1.5 – 2 nm. The density profiles also yield insight about the anchor penetration depth. The C8 anchor enters the bilayer to 1 nm, whereas the C16 anchor penetrates as deep as 2 nm. Hence, the long anchor affects the bilayer structure on both leaflets. The density profile of the SOPC vesicle is the most difficult to interpret because the vesicle is partially spherical, and the SAM-tethered lipid layer cannot be distinguished from the rest of the vesicle. The density profile of the spherical vesicle at the tetherless SAM indicates that the normal direction density linearly increases with z until ring like cuts through the vesicle are present, for which the density remains constant. This observation is different at the mixed SAMs. Here, the density profile resembles that of bilayer, with a SOPC density peak at the inner leaflet boundary. At the outer leaflet, interestingly, no such peak is reported. Anyway, the differences in the solvent and SOPC density profiles between the EG_6AC_8D and $EG_6AC_{16}D$ tethers are marginal.

The results from the normal-direction density profile are well reflected by close-up snapshots of the interface after 500 ns relaxation. The vesicle at the tetherless SAM remains spherical and the superficial water layer is undisturbed. For both the tethered bilayers, the vesicles were forced onto the bilayer, which results in a planar geometry in proximity in top of the SAMs.

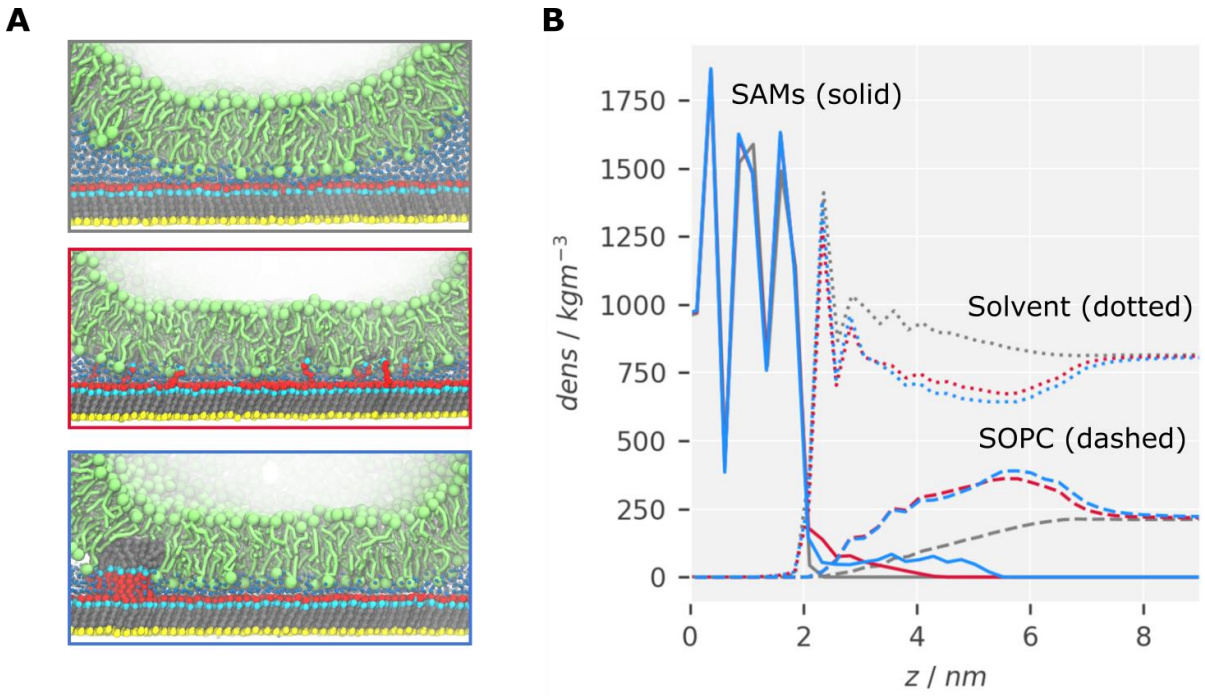

Figure 8.6: SAM-vesicle interface characterization. A) Surface normal direction density profiles. Solid lines: SAM. Dotted lines: water. Dashed lines: vesicle. Gray: anchorless. Red: 5% C8 anchors. Blue: 5% C16 anchors. B) interface snapshots.

8.4 Discussion

The QCM-D data show that that tethered lipid bilayer formation through vesicle fusion is robust to small deviation in the SAM composition. Bilayers do form at SAMs with EG_6AC_8D tether coverages of 5-20 mol% and $EG_6AC_{16}D$ concentration of 3-10 mol%. For smaller concentrations, vesicle rupture is not observed. For larger concentrations, the final mass on the microbalance is smaller and a reverse monolayer is formed. Our data indicate that with $EG_6AC_{16}D$ tethers, vesicle fusion appears at lower vesicle concentration and with lower tether coverage. This can be rationalized by various observations from the molecular simulations.

On the tetherless SAM, hydrogen bonds are the predominant interaction type. They are mostly formed between the terminal hydroxyl groups of the SAM and the phosphate group of the lipids. The transient interactions lead to localized deformation of the vesicle (flattening) only at the interaction site. The vesicle is separated from the SAM by a continuous water layer, which is locally disconnected by a small number of direct SAM-vesicle contacts. The incorporation of EG_6AC_8D tether molecules leads to an increased SAM-vesicle interface area and more pronounced planar distortion of the vesicle. The tether alkyl chains penetrate the outer leaflet of the local phospholipid bilayer zone but do not advance into the inner layer. The tether amide groups largely intercalate with the phosphate groups of the lipids due to favorable polar interactions (hydrogen bonds). In total, the number of polar interactions between the SAM and vesicle is higher as compared to the tetherless SAM. However, visual inspection of the interface suggests that the spatial density of such

interactions is similar and the increased number results from the larger interface area. Thickness and continuity of the interfacial water layer appear slightly decreased as compared to the tetherless SAM.

The $EG_6AC_{16}D$ alkyl chain tethers penetrate much deeper into the phospholipid bilayer and locally interfere with both the outer and the inner layer. The interface area is of comparable size as of the vesicle adsorbed onto the EG_6AC_8D tether SAM. The water layer is of similar thickness but more continuous. Insertion of the $EG_6AC_{16}D$ tether clusters was only observed when a 1000 kJ/mol nm external force constant is applied. A spontaneous insertion was not observed. An explanation for such an outcome is that for the insertion of a tether cluster, a large number of lipids must be displaced in the membrane. This is confirmed by the QCM-D data that clearly shows stable tethering of the bilayer on the clustered EG6AC16D SAM.

In the case of randomly (homogeneously) distributed assemblies containing EG_6AC_8D tethers, the MD simulations show their insertion into the vesicle upon contact. The numerous insertions observed during the simulation are suggesting that the energy of the thermal motion is sufficient for the insertion to happen. It is noteworthy that in biological membranes the docking step of the palmitoylated proteins is guided by enzymes and requires energy. For example, experimental studies showed that each methylene group of a lipidated peptide contributes 3.3 kJ/mol to the free energy of membrane binding [436]. Thus, palmitoylated peptides would exhibit a free energy of binding of approximately 53 kJ/mol. In comparison, a geranylgeranyl peptide chain showed a free energy difference of 50 kJ/mol [437]. Previous full atomistic MD simulations of a single and double geranylgeranylated peptide yielded bilayer extraction free energies of 69 kJ/mol and 119 kJ/mol [337]. Our corresponding coarse-grained model slightly underestimated the experimental results and yielded 30 kJ/mol for the single geranylgeranyl peptide. Thus, for the EG_6AC_8D tethers, a binding free energy of 26 kJ/mol and for the $EG6AC_{16}D$ of 52 kJ/mol can be expected. However, in the coarse-grained model, such energies might be slightly underestimated. The accuracy of partition energies in MARTINI is discussed in several articles, e.g. [438-440].

Deformation of the SAM-tethered vesicles was quantified as a reduction of sphericity from 0.98 (untethered, loosely attached vesicle) down to 0.93. The increasing deformation of the vesicles can be seen in the dissipation signals of the QCM-D sensograms. At the EG_6AC_8D SAMs, the tethering i.e., the insertion of membrane anchor acyl chains into the bilayer, appears spontaneously in the MD which is reflected by simultaneous peaks of both frequency and dissipation signals. At the $EG_6AC_{16}D$ SAMs, the peak of the dissipation signal comes before the peak of the frequency signal. This suggests that the insertion of the clustered tethered happens only on the time scale of seconds. This is reflected by the absence of tether insertions in the equilibrium parts of the simulations. One

contribution of the frequency and dissipation signals in QCM-D sensogram is attributed to water molecules. Trapped water molecules are characterized by increasing frequency and decreasing dissipation signals [441]. The trapped water layer can be clearly seen in the tethered portion of the vesicle in the MD density profiles.

8.4.1 Conclusion

In this chapter, we have investigated specific interactions of a small unilammellar vesicle on differently composed SAMs using a non-equilibrium MD approach combined with QCM-D measurements. To our knowledge, such a methodic tandem is reported for the first time. Thus, we had to identify (semi-)quantitative intersection points between MD and QCM-D. These are found in the tether insertion kinetics and subsequent deformation of the vesicle, which is reflected on the temporal shift between frequency and dissipation signals. Furthermore, the number of trapped atoms between tethered bilayer and SAM is quantitatively accessible by MD and QCM-D. However more structural insight for example by neutron reflectometry is necessary to fully understand how the QCM-D signals and the layer thickness correlate.

The MD models reveal different adsorption mechanism depending on the SAM composition. The key driver for the tethering is the incorporation of tether acyl chains into the SAMs. This happens quickly and spontaneously for the EG_6AC_8D tethers upon vesicle-SAM contact. The insertion of the longer and clustered $EG_6AC_{16}D$ tethers affords overcoming a much higher energy barrier. An increase in tether density reduces the critical vesicle density which is based on the increasing deformation of the vesicles. Curiously, the results show that stable membrane tethering can be achieved with only 2-3 mol% of the EG6AC16D tethers.

The novel, coarse-grained, non-equilibrium workflow is well suited to model tethered bilayer systems. A stronger initial attachment force is favored because it generates a more realistic number of inserted tethers. However, in future attempts, the succeeding relaxation phase sampling should be extended to at least 1 μs. We note again that the simulation of attachment and anchor insertion is necessary to avoid artifacts by deleting lipids from the bilayer as commonly done in other approaches. Further development of the approach could include the truncation of the vesicles to induce membrane fusion. Alternatively, it might be possible to drag a bilayer patch onto the SAM. In the global context of the dissertation thesis, the work shows the proficiency of the MARTINI force field to reliable sample layered systems with large portions of linear acyls over long time scales. The differences of the interaction modes and tethering energies between the vesicle and the different SAMs can clearly be seen and agrees with experimental observations.

9 CONCLUSION

In this work, we accomplished to yield valuable insights into the fields of infection biology (norovirus bile acid binding and Legionella protease RavD), protein stability (deamidation and glycosylation), and surface nanotechnology (SAMs and tethered bilayer membranes) and thereby assessed the limitations of the molecular dynamics simulation method and suggested approaches for their overcoming. Thanks to the application of MD to a range of different molecular systems, we were able to identify generally feasible methods for trajectory analysis. Furthermore, as all the studied problems revolve around the pairwise comparison of multiple similar molecular systems, we put great store in the development and automatable, algorithmic, modeling, and quantitative analysis.

9.1 Contributions to the fundamental sciences

The bile acid binding study revealed the weakness of small molecule docking to rigid receptor structures. Especially, when the receptor conformation does not include a ligand or substrate in the proposed binding site, docking and virtual screening approaches are doomed to fail. A variety of methods to select few conformations from a large ensemble such as conformational or pocket shape-based clustering do exist, yet in our case did not yield the best docking results. In the case of the weakly binding bile acid molecules, the best ranking poses were identified A. after multiple 100s of nanoseconds of sampling, and B. in rather inconspicuous protein conformations. This led to the conclusion that the ligand may not only select metastable conformations of higher energy than the native state but even highly transient states that only become local minima through the ligand-induced remodeling of the conformational energy landscape. Anyway, our developed ensemble-docking workflow based on MD sampling of the receptor, rigid body docking, and MD refinement, predicted a small ensemble of reasonable binding modes, of which two were in agreement with the NMR chemical shift perturbation experiments. Due to the weak affinity, we suggest that binding is on the tip of being non-selective and the binding itself is conformationally and orientationally dynamic. A second explanation is binding of two or more bile acid molecules at the time, because bile acid micelle formation can occur at the observed concentrations.

In the purely computational study on the two highly selective isopeptidases RavD (bacterial) and OTULIN (human), mechanisms for selectivity and activation were compared based on dynamics molecular descriptors. The results confirmed earlier statements on OTULIN yet challenged the conclusions regarding RavD. In particular, we designed a detailed survey to quantitatively compare the two recognition sites for the identical substrate (linear di-ubiquitin) of RavD and OTULIN, based on orientational fluctuations, buried surfaces areas, intermolecular residue-residue

interactions, and induction of catalytic competency. We found that RavD and OTULIN employ largely different primary sequences to achieve identical binding modes for di-ubiquitin. In terms of binding stability, interface area and interaction residue composition, one binding site was almost identical between the two proteins from the different phylogenetic domains. The second binding site was shown to be highly disjunct. As the RavD interface area was smaller, it showed high fluctuations of the substrate binding mode and the composition of inter-residue interactions was less favorable and specific. It also appeared that in OTULIN, substrate binding induced a transition of the catalytic triad towards a competent state (substrate-assisted catalysis). In the crystal structures of RavD, such a transition was not apparent. Moreover, even in the substrate -bound structure of RavD, the catalytic triad was incompetent, and the substrate was too far from a reactive distance. We understood that this discrepancy was intentionally induced by the original investigators by introducing a double mutation into the substrate protein to avoid cleavage. In the MD sampling with the re-mutated, wild-type substrate, the transition towards the active state was indicated but could not be completely sampled. Thus, our theoretical considerations reveal an uncommented inconsistency in the original research to RavD, which led to the most likely false conclusion that RavD does not undergo substrate-assisted catalysis. Instead, we have strong indication that the reported high selectivity and activity of RavD can only be rationalized by substrate-induced catalytic priming of the catalytic triad.

In the study on the glycosylation of human erythropoietin (EPO) the reciprocal effects between the protein and its complex N-glycans are investigated. Therefore, at that time, a complex multi-stage modeling workflow based on multiple webservers and parsing steps had to be employed. It revealed the need for automatable glycoprotein modeling tools which only appeared during the compilation of the thesis. Global physical properties of the glycoprotein showed to be rather dependent on the number and not the site of the glycosylation. Such a statement can be generalized to further small globular proteins but would need to be re-evaluated for larger, more polarized proteins. Inter-residue interaction analysis yielded transient contacts between the glycan and the protein mostly mediated by the glycosylation root and the fucose but also the terminal sialic acid moieties. The interaction did not significantly alter protein solvent accessibility and flexibility with the apparent exception of the glycosylated residues. To address the conformational space of the complex N-glycans, an embedded clustering workflow had to be developed and benchmarked with an artificial dataset. The workflow only became successful with the release of UMAP embedding and HDBSCAN clustering. In its final, optimized form, it was able to reliably separate and cluster up to 256 Gaussian populations in a 32-dimensional space. This outcome was considered sufficient for the workflow to be employed on the MD-generated conformational data set of glycosidic torsion angles. It showed that one major and distinct contributor for the conformational flexibility

of N-Glycans is the asparagine sidechain. Conformational dynamics in this area are reflected on the orientation of the entire N-Glycan. On the other hand, the conformational space of the glycan itself is relatively similar and identical conformational clusters are identified independent on the glycosylation site. Also, their individual contributions to the total conformational space are remarkably comparable. The conformational clusters of the asparagine sidechains, however, are to some extent site-specific. Thus, the well-accepted opinion that proteins and glycan do not mutually affect their conformations in a significant manner is right and wrong at the time. Apparently, the intrinsic conformational dynamics of the glycan are not affected by the protein, however the global dynamics of the complex N-glycan are, and this even in a site-specific manner. The results might be of interest for the design principles of novel glycoconjugate drugs, even though it has to be decided in case-specific manner if a folded glycan with many protein interactions, or a more extended, solvent accessible glycan is to be preferred.

The combined NMR and MD study on site-selective deamidation of asparagine 373 (N373) of the norovirus P-dimer yielded exceptionally deep insight into the distortion of the conformational space of asparagine residues within the structurally constrained environment of a protein. Interestingly, even though we could identify and quantify specific and distinct combined backbone and sidechain conformational populations, the processing of molecular conformation to nucleophilic attack geometry did not reveal a clear association between attack probability and deamidation rate. Instead, only with the additional assessment of the conformations and inter-residue interactions of the succeeding N+1 residue, a tendency for the preferential deamidation of N373 becomes apparent. A few earlier studies indicated the effect of the backbone conformation on the acidity of the backbone amide hydrogen. Strikingly, the N+1 residue of N373 is among the few residues that can adopt such a *syn* backbone conformation. Furthermore, the backbone amide at the 374 position undergoes extensive hydrogen bonding with neighboring hydrogen bond acceptors, which further facilitates hydrogen dissociation. Our MD study shows that only the combined examination of solvent accessibility, preferential attack geometry and backbone amide acidity allows the rationalization of fast deamidation rates. We evaluated if our results are feasible for predictions, by determining the same criteria for the P-dimer of a closely related strain. In this protein, the approach indicated a significant likeliness for N373 to be deamidated but did not identify this residue as the most likely. Such an outcome is not surprising, though, because assessment of one molecule alone does not allow tuning of the relative impotencies (weights) of the different involved factors. Additionally, the model is based on quantiles of multivariate, joint probability distributions (e.g., attack distance < 0.4 nm and attack angle $> 90°$). The definition of such plays a significant role for the predicted deamidation probability and must be carefully

optimized. This can only be achieved when a larger number of proteins has been assessed. Nevertheless, the MD model clearly rationalizes how the fast deamidation of N373 is possible.

The combined full-atomistic and coarse-grained simulation study on the membrane anchored Rab5 peptide led by E. Münzberg formed the basis for the subsequent studies on mixed self-assembled monolayers and tethered lipid bilayer membranes. Especially the simplicity of the employed method for the automated packing of lipid bilayers as well as the low computational costs of the coarse-grained simulations paired with the reasonable accuracy of the MARTINI force field were pivotal for the decision to further develop into the field of molecular layers. Additionally, the simulations revealed an aggregation of PI3P signaling lipid in proximity to the Rab5 C-terminal peptide due to electrostatic interactions, which are selective of membrane-anchored proteins of the Ras and Rab family. This outcome was meaningful for the interpretation of other full-atomistic results and the suggestion of an improved Rab5 signaling model.

The tethering lipid bilayer membrane (tBLM) to functionalized self-assembled monolayers (SAMs) is an emerging approach to investigate lipid bilayers and associated or inserted membrane proteins. The technique has tremendous potential for surface nanotechnology and drug research. Hence, a thorough understanding on the molecular processes taking place during the preparation of SAM-tBLMs is crucial. In the mixed SAM study, we developed a novel multi-scale modeling and simulation approach for the mentioned molecular systems. The simulations yielded new insight in the structure and dynamics of membrane anchors (also called tethers) when incorporated into a SAM. The simulations show a conformational dependence on the anchor acyl chain lengths, which is based on unexpected intermolecular interactions. The sampled alkyl chain and oligoethylenglycol conformations are in good agreement with experimental investigations from infrared reflection absorption spectroscopy (IRRAS). However, the anchor density dependent conformational transition (phase transition) as seen in IRRAS, afford a clustered spatial distribution of a one anchor species and an equal spatial distribution of the other. If the clusters form already in the solvated phase, during the adsorption or even within the SAM awaits further studies. However, it is most likely that they anchoring compounds selectively associate with each other during adsorption because lateral diffusion within SAMs is extremely limited and the ethanol-water mixture solvent should be able to solvate such amphiphilic molecules. Hence, the simulations suggest a preferential, clustered adsorption of the anchoring compounds based on specific intermolecular interactions. Such an observation is important because it allows an optimized preparation protocol for tethered bilayers depending on target properties.

The above protocol was employed to study the initial step of tBLM formation by vesicle fusion in a combined molecular dynamics and quartz crystal microbalance with dissipation monitoring

(QCM-D) study. The non-equilibrium simulations show the intermolecular interactions between different SAMs and a small unilammellar vesicle (SUV). It shows that a SUV on tetherless SAM is weakly adsorbed via hydrogen bonds and its shape is undistorted. On a SAM decorated with short tethering compounds in uniform spatial distribution, the SUV becomes increasingly distorted through spontaneous insertion of the tether acyl chains. On SAMs with clustered, long-acyl chain tethered, spontaneous insertion does not occur on the timescale of the simulations. Thus, the energy barrier must be orders of magnitude higher. The short tethers only penetrate one layer, whereas the long tethers penetrate both. The interface water layer is slightly more ordered when the bilayer is tethered to longer, clustered tethers. The results are line with the QCM-D results, which show an earlier vesicle fusion for the longer tether acyl chains. Additionally, it shows that with an unexpectedly small density of clustered, long tethers, stable adsorption can be achieved. Yet, it remains an open question if this will be advantageous for the insertion of membrane proteins.

9.2 Considerations on molecular dynamics

In this thesis, we have employed classical molecular dynamics to a range of different molecular and are thus in the position to make general statements on practical and technical issues and on limitations. First and foremost, it is crucial to understand that the quality and validity of an MD simulation depends on the initial configuration (packing). The packing is guided by experimental insight and/or other complementary methods. In some applications, an experimentally determined structural model is available. In these cases, the initial structural model must be carefully assessed and possible sources for artifacts must be identified. Especially in proteins, these are conformational alteration induced through mutations or the presence or absence of ligands or cofactors. When no structural model is available, it must be generated using protein structural modeling. Here, the output of the MD is heavily dependent on the model quality. For layered systems, such as bilayers or monolayers, the initial packing is mostly achieved by software. However, the software is not aware of preferential intermolecular interactions and the possibility of clusters. Thus, random uniform packing is already biased. The better the initial model, the more accurate the MD sampling becomes. However, unphysical starting structures might be overcome by extensive sampling. In this work we encountered plenty of initial configuration bias problems. The docking models of the bile acids mostly yielded unstable complexes and dissociated from their docked mode. This was overcome by using an ensemble of initial configurations from which we started a set of MD simulations with different initial velocity distributions. Only through such extended sampling, we were able to achieve stable complexes. In the RavD-di-ubiquitin simulations, a mutation in the substrate led to an artificial conformation of the substrate close to

the binding site, which was not able to induce the necessary activation of the catalytic triad. It was so stable, that extensive equilibrium MD simulation in the microsecond scale were not able sample the transition to the true binding mode. In the case of the SAMs, the true molecular structure was not known. Hence, different models were generated, sampled and compared to experiments. It shows that in practice, the initial model is often flawed, and MD is explicitly used to transition the configuration to a more physical representation. So, the aim of molecular dynamics is actually twofold: A. sampling of other conformations in local minima besides the global minimum (e.g., protein dynamics based on X-ray structure) and B. recovery of the native state starting from an artificial configuration (e.g. refinement of a docking pose).

The most discussed limitations of MD are force field inaccuracies and incomplete sampling. Both limitations are in a way interdependent because a higher-accuracy description of the systems affords more computational resources. In fact, force field inaccuracies can often be minimized, only by choice of the correct force field for the application in question. Sometimes, it will thus be necessary to use polarizable force fields, or a hybrid quantum-chemistry based description. On the other hand, when large systems are to be investigated, a coarse-grained force-field might suffice. Recent force fields offer a large array of parameters for most common bonded interactions and offer methods to generate topologies for new molecules based on analogies. De-novo parameterization is rarely necessary and is mostly limited to a few specific parameters. For investigators, the most urgent task is to find the optimum balance between force field accuracy and necessary time scales for the molecular problem. If the preferred combination is suitable for the available hardware, the simulations can be performed. If not, classical MD is not the method of choice and enhanced sampling MD must be performed. In this thesis, we mostly stuck to the CHARMM36 force field which proved to be accurate for proteins, carbohydrate and bilayers and achieved a satisfactory performance and usability with the free and open-source MD software suites GROMACS and OpenMM.

Only in the case of the SAMs and tethered bilayer, we employed the coarse-grained MARTINI force field for which a set of new bonded parameters were developed. The MARTINI force field was initially designed for phospholipid bilayer membranes but was later extended to basically all kinds of molecules – with varying success. Naturally, MARTINI works well for polymers. The extension to proteins is widely used even though it requires the user to fixa the secondary structure using an elastic network model. Thus, a key intrinsic attribute of proteins - conformational dynamics – is neutralized. Yet, MARTINI still shines as a tool to model large scale polymeric and alkyl-chain heavy molecular systems. Thanks to its flat energy landscape and the long possible

integration step, it is highly feasible to bring a large-scale artificial model to a near-physical state. The model can then be subjected to resolution transformation for further full-atomistic sampling.

Considering that the MD user had reasonable starting structure, chose a suitable force field, and has simulated long enough to sample the meaningful transitions and conformational states, they ultimately face the issue on how to analyze their MD trajectories. This question is even more significant in recent times when multi-microsecond sampling in multiple replica simulations is generated. Nevertheless, initial trajectory visualization should under no circumstances be avoided. However, it is often sufficient so reduce the trajectory to a few equidistant snapshots of the important solutes. The visual inspection should focus on the box dimensions, unexpected association, or dissociation, and overly stable of dynamic regions. The main analysis is of course of quantitative nature.

For proteins, the obligatorily calculated root mean squared deviation gives initial insight on stability and conformational sampling as well as convergence. It is implemented in all MD analysis software and should always be elucidated. However, it is lacking details. More insight is given by the root-mean square fluctuation, which allows localization of flexible regions. It is especially helpful when comparing several proteins with only minor differences (mutations, ligands). Its downside is, that the RMSF mostly confirms the secondary structure and rarely uncovers interesting dynamics. The relative solvent accessibility is another standard analysis means and, in light of this thesis, is a highly recommended and valuable observable. It shows which residues are at the outer boundary of the protein and indicates if they are in extended or folded conformation. It can reveal the competency of the catalytic triad as well as interacting residues in bimolecular systems. Assessing the residue type of solvent exposed residues can be used to identify hydrophobic surface patches. Finally, it can be used to quantify interface areas.

In our opinion, one of the most detailed and worthwhile trajectory analysis means is the pairwise inter-residue contact occupancy (contribution) matrix. It is a non-standard method and is not readily implemented in most analysis suites. The matrix allows quick assessment of all contacts and their probability, which of course can be translated into free energy. The data can be quickly projected to either partner of the interacting molecules and can be easily visualized. Furthermore, the interaction pairs can be classified according to the residue types which enables quick comparison of even structurally different complexes. Finally, interesting long-lasting interactions can be further monitored for hydrogen bonds or salt bridge formationss. The contact analysis is more meaningful for intermolecular complexes. Intramolecular contact analysis is less expressive because it often only reveals the secondary structure and interaction of neighboring or bonded residues. However, it can be adapted for both cases by using appropriate distance criteria.

For intramolecular questions, geometric descriptors are often most informative. These are distances between non-bonded atoms as well as torsion angles of the sidechains and backbone. Other descriptors, such as attack angles, do exist. As a protein is usually heavily interconnected, geometric descriptors are interdependent. Furthermore, a full description affords the assessment of an entire range of descriptors. To add to the complexity, interesting regions in molecules are characterized by dynamics. That is the presence of multiple metastable states between which the system transitions. The presence of multiple states implies multimodal distributions of the sampled geometric descriptors. Thus, averaging must be done with extreme care and the distributions must always be examined. All these consideration yield to problems of multivariate statistics, which should not be discouraging. Instead, recent dimensionality reduction methods should be utilized to visualize the multidimensional in a two-dimensional scatter plot or free energy map. Manual selection of automated clustering allows the identification of different conformational states of which molecular representation can be generated.

To conclude on the technique of MD, it must be noted that the method of the atomic position integration itself is already highly optimized and not much further development can be expected. The same is true for the molecular force fields. Besides some minor improvements for specific molecules, the quality of empirical force fields is, in our opinion, exhausted. Novel developments go into the direction of machine-learned force fields, polarizable, or hybrid quantum mechanics-based dynamics. While accessible sampling times as long as milliseconds become accessible, they typically still do not suffice to sample biologically relevant processes such as folding or ligand binding and dissociation. Thus, currently and in the foreseeable future, investigators should focus on improved approaches for initial modeling. This has partially been fulfilled with deep-learning based protein structure prediction methods such as AlphaFold or RoseTTAfold and ongoing studies on the prediction of protein complex binding modes. Such models, followed by extensive MD simulations as a validation and refinement will be at the heart of many future computational structural research projects.

We cannot close this section without commenting on the utmost importance of experimental data for MD simulations. Unclear or unexplainable experimental results often are the motivation for MD simulations. Experimental structures and data must be frequently used to check the validity of the simulation. As a rule of thumb, it can be noted that an MD simulation must first confirm certain experimental observations before any MD based prediction can be taken seriously. A good constellation is a combined and iterative experimental and MD study, in which observations from both worlds can drive hypotheses and adjust the direction of further progression.

9.3 Final remarks

For several fundamentally different molecular questions, we revealed how type, number, longevity and exclusivity of non-bonded interactions shape the global behavior. In all instances, the presence or absence of distinct non-bonded interactions were accompanied by local conformational changes. In turn, such conformational alterations were the reason for experimental observations. Our studies give a general guideline for applied and experiment-oriented MD simulation efforts and their statistical analysis. The recommendation is that analysis of interactions must always be conducted hand-in-hand with analysis of the conformational space. For screening purposes, interactions can easily be expressed as inter-residue contacts maps. The conformational space is most directly described as a set of torsion angles. We showed that the mathematically challenging n-toroidal space can be embedded into a two-dimensional Euclidean plane and conformational clustering in the embedded space is feasible. In some cases, e.g., when inter- or intramolecular chemical reactions are involved, distances and angles of the involved atoms must be included in the analysis.

All in all, in this research thesis, we have gathered comprehensive insight on the pivotal roles on non-bonded interactions for structure and dynamics of molecules and molecular assemblies. The interactions and their effects were assessed from various perspectives and via tailored MD-centered workflows. The thorough description and discussion of the herein developed workflows and analytical method will serve as a basis for future efforts by us and the community. The results in their entirety amplify the significance of the reciprocal effects of conformational dynamics and non-bonded interactions. A complete understanding of a molecular system can only be achieved when both these aspects are thoroughly elucidated. As of today, therefore, classical molecular dynamics simulation, especially when accompanied by physical, chemical, and biological experimentation as well as complementary theoretical means for the pre- and post-production, still is among the most suitable and versatile methods.

1

10 REFERENCES

1. Brown, T.L. et al. (2009) *Chemistry: The Central Science* (11th Edition). Upper Saddle River (NJ), Pearson Education Inc.
2. Saunders, S. and Brown, H.R. (1991) *The Philosophy of Vacuum* (New Edition). Oxford, Clarendon Press
3. Carey, F.A. and Sundberg, R.J. (2007) *Advanced Organic Chemistry: Part A: Structure and Mechanisms* (5th Edition). Berlin and Heidelberg, Springer Science & Business Media
4. Stillinger, F.H. (1980) Water Revisited. *Science* 209, 451-457
5. Černý, J. and Hobza, P. (2007) Non-Covalent Interactions in Biomacromolecules. *Physical Chemistry Chemical Physics* 9, 5291-5303
6. Colovos, C. and Yeates, T.O. (1993) Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions. *Protein Science* 2, 1511-1519
7. Strmcnik, D. et al. (2009) The Role of Non-Covalent Interactions in Electrocatalytic Fuel-Cell Reactions on Platinum. *Nature Chemistry* 1, 466-472
8. Taylor, S.S. et al. (2004) Pka: A Portrait of Protein Kinase Dynamics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1697, 259-269
9. Todd, M., J., Viitanen, P., V., and Lorimer, G., H. (1994) Dynamics of the Chaperonin Atpase Cycle: Implications for Facilitated Protein Folding. *Science* 265, 659-666
10. Mannella, C.A. (1998) Conformational Changes in the Mitochondrial Channel Protein, Vdac, and Their Functional Implications. *Journal of Structural Biology* 121, 207-218
11. Olson, R. et al. (1999) Crystal Structure of Staphylococcal Lukf Delineates Conformational Changes Accompanying Formation of a Transmembrane Channel. *Nature Structural Biology* 6, 134-140
12. May, L.T. et al. (2007) Allosteric Modulation of G Protein–Coupled Receptors. *Annual Review of Pharmacology and Toxicology* 47, 1-51
13. Kruse, A.C. et al. (2013) Activation and Allosteric Modulation of a Muscarinic Acetylcholine Receptor. *Nature* 504, 101-106
14. Frauenfelder, H. et al. (2009) A Unified Model of Protein Dynamics. *Proceedings of the National Academy of Sciences* 106, 5129
15. Henzler-Wildman, K.A. et al. (2007) A Hierarchy of Timescales in Protein Dynamics Is Linked to Enzyme Catalysis. *Nature* 450, 913-916
16. Leckband, D. and Israelachvili, J. (2001) Intermolecular Forces in Biology. *Quarterly Reviews of Biophysics* 34, 105-267
17. Chilkoti, A. and Stayton, P.S. (1995) Molecular Origins of the Slow Streptavidin-Biotin Dissociation Kinetics. *Journal of the American Chemical Society* 117, 10622-10628
18. Smit, B. and Maesen, T.L.M. (2008) Towards a Molecular Understanding of Shape Selectivity. *Nature* 451, 671-678
19. Hsu, C.C., Buehler, M.J., and Tarakanova, A. (2020) The Order-Disorder Continuum: Linking Predictions of Protein Structure and Disorder through Molecular Simulation. *Scientific Reports* 10, 2068
20. Persch, E., Dumele, O., and Diederich, F. (2015) Molecular Recognition in Chemical and Biological Systems. *Angewandte Chemie International Edition* 54, 3290-3327
21. Stoker, H.S. (2015) *General, Organic, and Biological Chemistry* (7th Edition). Boston (USA), Cengage Learning.
22. Benkovic Stephen, J. and Hammes-Schiffer, S. (2003) A Perspective on Enzyme Catalysis. *Science* 301, 1196-1202
23. Fischer, E. (1894) Einfluss Der Configuration Auf Die Wirkung Der Enzyme. *Berichte der Deutschen Chemischen Gesellschaft* 27, 2985-2993
24. Koshland, D.E. (1958) Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences* 44, 98-104

25. Paul, F. and Weikl, T.R. (2016) How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates. *PLOS Computational Biology* 12, e1005067
26. Ma, B. et al. (1999) Folding Funnels and Binding Mechanisms. *Protein Engineering* 12, 713-720
27. López-Otín, C. and Bond, J.S. (2008) Proteases: Multifunctional Enzymes in Life and Disease. *Journal of Biological Chemistry* 283, 30433-30437
28. Schlüter, D. et al. (2022) Ovarian Tumor Domain Proteases in Pathogen Infection. *Trends in Microbiology* 30, 22-33
29. Shen, Z. et al. (2022) Design of Sars-Cov-2 Plpro Inhibitors for Covid-19 Antiviral Therapy Leveraging Binding Cooperativity. *Journal of Medicinal Chemistry* 65, 2940–2955
30. Hartwell, L.H. et al. (1999) From Molecular to Modular Cell Biology. *Nature* 402, C47-C52
31. Petroski, M.D. and Deshaies, R.J. (2005) Function and Regulation of Cullin–Ring Ubiquitin Ligases. *Nature Reviews Molecular Cell Biology* 6, 9-20
32. Komander, D. and Rape, M. (2012) The Ubiquitin Code. *Annual Review of Biochemistry* 81, 203-229
33. Baek, K., Scott, D.C., and Schulman, B.A. (2021) Nedd8 and Ubiquitin Ligation by Cullin-Ring E3 Ligases. *Current Opinion in Structural Biology* 67, 101-109
34. Harper, J.W. and Schulman, B.A. (2021) Cullin-Ring Ubiquitin Ligase Regulatory Circuits: A Quarter Century Beyond the F-Box Hypothesis. *Annual Review of Biochemistry* 90, 403-429
35. Baek, K. et al. (2020) Nedd8 nucleates a Multivalent Cullin–Ring–Ube2d Ubiquitin Ligation Assembly. *Nature* 578, 461-466
36. Jones, S. and Thornton, J.M. (1996) Principles of Protein-Protein Interactions. *Proceedings of the National Academy of Sciences* 93, 13-20
37. Arkin, M.R. and Wells, J.A. (2004) Small-Molecule Inhibitors of Protein–Protein Interactions: Progressing Towards the Dream. *Nature Reviews Drug Discovery* 3, 301-317
38. Lu, H. et al. (2020) Recent Advances in the Development of Protein–Protein Interactions Modulators: Mechanisms and Clinical Trials. *Signal Transduction and Targeted Therapy* 5, 213
39. Arakawa, T. et al. (2007) Suppression of Protein Interactions by Arginine: A Proposed Mechanism of the Arginine Effects. *Biophysical Chemistry* 127, 1-8
40. Li, S. and Hong, M. (2011) Protonation, Tautomerization, and Rotameric Structure of Histidine: A Comprehensive Study by Magic-Angle-Spinning Solid-State Nmr. *Journal of the American Chemical Society* 133, 1534-1544
41. Rust, H.L. and Thompson, P.R. (2011) Kinase Consensus Sequences: A Breeding Ground for Crosstalk. *ACS Chemical Biology* 6, 881-892
42. Marshall, R.D. (1972) Glycoproteins. *Annual Review of Biochemistry* 41, 673-702
43. Farnsworth, C.C. et al. (1994) Rab Geranylgeranyl Transferase Catalyzes the Geranylgeranylation of Adjacent Cysteines in the Small Gtpases Rab1a, Rab3a, and Rab5a. *Proceedings of the National Academy of Sciences* 91, 11963-11967
44. Jimenez-Morales, D. et al. (2014) Lysine Carboxylation: Unveiling a Spontaneous Post-Translational Modification. *Acta crystallographica. Section D, Biological crystallography* 70, 48-57
45. Park, I.-S. and Hausinger Robert, P. (1995) Requirement of Carbon Dioxide for in Vitro Assembly of the Urease Nickel Metallocenter. *Science* 267, 1156-1158
46. Li, J. et al. (2015) A Review on Phospholipids and Their Main Applications in Drug Delivery Systems. *Asian Journal of Pharmaceutical Sciences* 10, 81-98
47. O'Dwyer, C. et al. (2004) The Nature of Alkanethiol Self-Assembled Monolayer Adsorption on Sputtered Gold Substrates. *Langmuir* 20, 8172-8182
48. Woodka, A.C. et al. (2012) Lipid Bilayers and Membrane Dynamics: Insight into Thickness Fluctuations. *Physical Review Letters* 109, 058102
49. Venable, R.M. et al. (2017) Lipid and Peptide Diffusion in Bilayers: The Saffman-Delbrück Model and Periodic Boundary Conditions. *The Journal of Physical Chemistry B* 121, 3443-3457
50. Quinn, P.J. and Wolf, C. (2009) The Liquid-Ordered Phase in Membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1788, 33-46

51. Wang, T.-Y. and Silvius, J.R. (2003) Sphingolipid Partitioning into Ordered Domains in Cholesterol-Free and Cholesterol-Containing Lipid Bilayers. *Biophysical Journal* 84, 367-378
52. Ingólfsson, H.I. et al. (2014) Lipid Organization of the Plasma Membrane. *Journal of the American Chemical Society* 136, 14554-14559
53. McGillivray, D.J. et al. (2007) Molecular-Scale Structural and Functional Characterization of Sparsely Tethered Bilayer Lipid Membranes. *Biointerphases* 2, 21-33
54. Lee, H.H. et al. (2018) Mixed Self-Assembled Monolayers with Terminal Deuterated Anchors: Characterization and Probing of Model Lipid Membrane Formation. *The Journal of Physical Chemistry B* 122, 8201-8210
55. Schulze, E. and Stein, M. (2018) Simulation of Mixed Self-Assembled Monolayers on Gold: Effect of Terminal Alkyl Anchor Chain and Monolayer Composition. *The Journal of Physical Chemistry B* 122, 7699-7710
56. Hill, T.L. (1946) On Steric Effects. *The Journal of Chemical Physics* 14, 465-465
57. Hill, T.L. (1948) Steric Effects. I. Van Der Waals Potential Energy Curves. *The Journal of Chemical Physics* 16, 399-404
58. Galli, S. (2014) X-Ray Crystallography: One Century of Nobel Prizes. *Journal of Chemical Education* 91, 2009-2012
59. Hodgkin, D.C. (1965) The X-Ray Analysis of Complicated Molecules. *Science* 150, 979-988
60. Hodgkin, D.C. (1971) X Rays and the Structures of Insulin. *British Medical Journal* 4, 447
61. Milne, J.L.S. et al. (2013) Cryo-Electron Microscopy – a Primer for the Non-Microscopist. *The FEBS Journal* 280, 28-45
62. Murata, K. and Wolf, M. (2018) Cryo-Electron Microscopy for Structural Analysis of Dynamic Biological Macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1862, 324-334
63. Guerry, P. and Herrmann, T. (2011) Advances in Automated Nmr Protein Structure Determination. *Quarterly Reviews of Biophysics* 44, 257-309
64. Johnson, K.H. (1975) Quantum Chemistry. *Annual Review of Physical Chemistry* 26, 39-57
65. Casewit, C.J., Colwell, K.S., and Rappe, A.K. (1992) Application of a Universal Force Field to Organic Molecules. *Journal of the American Chemical Society* 114, 10035-10046
66. Halgren, T.A. (1996) Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of Mmff94. *The Journal of Computational Chemistry* 17, 490-519
67. Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996) Development and Testing of the Opls All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* 118, 11225-11236
68. MacKerell, A.D. et al. (1998) All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B* 102, 3586-3616
69. Cornell, W.D. et al. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* 117, 5179-5197
70. Berger, O., Edholm, O., and Jähnig, F. (1997) Molecular Dynamics Simulations of a Fluid Bilayer of Dipalmitoylphosphatidylcholine at Full Hydration, Constant Pressure, and Constant Temperature. *Biophysical Journal* 72, 2002-2013
71. Woods, R.J. et al. (1995) Molecular Mechanical and Molecular Dynamic Simulations of Glycoproteins and Oligosaccharides. 1. Glycam_93 Parameter Development. *The Journal of Physical Chemistry* 99, 3832-3846
72. Shi, Y. et al. (2013) Polarizable Atomic Multipole-Based Amoeba Force Field for Proteins. *Journal of Chemical Theory and Computation* 9, 4046-4063
73. Scott, W.R.P. et al. (1999) The Gromos Biomolecular Simulation Program Package. *The Journal of Physical Chemistry A* 103, 3596-3607
74. Marrink, S.J., de Vries, A.H., and Mark, A.E. (2004) Coarse Grained Model for Semiquantitative Lipid Simulations. *The Journal of Physical Chemistry B* 108, 750-760
75. Abraham, M.J. et al. (2015) Gromacs: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* 1-2, 19-25

76. Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. (1995) Gromacs: A Message-Passing Parallel Molecular Dynamics Implementation. *Computer Physics Communications* 91, 43-56
77. Hess, B. et al. (2008) Gromacs 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *Journal of Chemical Theory and Computation* 4, 435-447
78. Kim, S. et al. (2017) Charmm-Gui Ligand Reader and Modeler for Charmm Force Field Generation of Small Molecules. *The Journal of Computational Chemistry* 38, 1879-1886
79. Lee, J. et al. (2016) Charmm-Gui Input Generator for Namd, Gromacs, Amber, Openmm, and Charmm/Openmm Simulations Using the Charmm36 Additive Force Field. *Journal of Chemical Theory and Computation* 12, 405-413
80. Lindahl, E., Hess, B., and van der Spoel, D. (2001) Gromacs 3.0: A Package for Molecular Simulation and Trajectory Analysis. *Molecular modeling annual* 7, 306-317
81. Pronk, S. et al. (2013) Gromacs 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* 29, 845-854
82. Lyubartsev, A.P. and Rabinovich, A.L. (2016) Force Field Development for Lipid Membrane Simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1858, 2483-2497
83. Huang, J. and MacKerell Jr, A.D. (2013) Charmm36 All-Atom Additive Protein Force Field: Validation Based on Comparison to Nmr Data. *The Journal of Computational Chemistry* 34, 2135-2145
84. Pastor, R.W. and MacKerell, A.D. (2011) Development of the Charmm Force Field for Lipids. *The Journal of Physical Chemistry Letters* 2, 1526-1532
85. Guvench, O. et al. (2011) Charmm Additive All-Atom Force Field for Carbohydrate Derivatives and Its Utility in Polysaccharide and Carbohydrate–Protein Modeling. *Journal of Chemical Theory and Computation* 7, 3162-3180
86. Hart, K. et al. (2012) Optimization of the Charmm Additive Force Field for DNA: Improved Treatment of the Bi/Bii Conformational Equilibrium. *Journal of Chemical Theory and Computation* 8, 348-362
87. Vanommeslaeghe, K. et al. (2010) Charmm General Force Field: A Force Field for Drug-Like Molecules Compatible with the Charmm All-Atom Additive Biological Force Fields. *The Journal of Computational Chemistry* 31, 671-690
88. Nutt, D.R. and Smith, J.C. (2007) Molecular Dynamics Simulations of Proteins: Can the Explicit Water Model Be Varied? *Journal of Chemical Theory and Computation* 3, 1550-1560
89. Kwon, Y. and Lee, J. (2021) Molfinder: An Evolutionary Algorithm for the Global Optimization of Molecular Properties and the Extensive Exploration of Chemical Space Using Smiles. *Journal of Cheminformatics* 13, 24
90. Korovina, K. et al. (2020) Chembo: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations. *International Conference on Artificial Intelligence and Statistics* (Online). PMLR, 3393-3403
91. Wilson, S.R. et al. (1991) Applications of Simulated Annealing to the Conformational Analysis of Flexible Molecules. *The Journal of Computational Chemistry* 12, 342-349
92. Hairer, E., Lubich, C., and Wanner, G. (2003) Geometric Numerical Integration Illustrated by the Störmer–Verlet Method. *Acta Numerica* 12, 399-450
93. Berendsen, H.J.C. et al. (1984) Molecular Dynamics with Coupling to an External Bath. *The Journal of Chemical Physics* 81, 3684-3690
94. Lemak, A. and Balabaev, N. (1994) On the Berendsen Thermostat. *Molecular Simulation* 13, 177-187
95. Rühle, V. (2007) Berendsen and Nose-Hoover Thermostats. Available from: https://www2.mpip-mainz.mpg.de/~andrienk/journal_club/thermostats.pdf (accessed on September 14th 2022)
96. Rühle, V. (2008) Pressure Coupling/Barostats. Available from: https://www2.mpip-mainz.mpg.de/~andrienk/journal_club/barostats.pdf (accessed on September 14th 2022)

97. Essmann, U. et al. (1995) A Smooth Particle Mesh Ewald Method. *The Journal of Chemical Physics* 103, 8577-8593
98. Yeh, I.-C. and Berkowitz, M.L. (1999) Ewald Summation for Systems with Slab Geometry. *The Journal of Chemical Physics* 111, 3155-3162
99. Wells, B.A. and Chaffee, A.L. (2015) Ewald Summation for Molecular Simulations. *Journal of Chemical Theory and Computation* 11, 3684-3695
100. Adcock, S.A. and McCammon, J.A. (2006) Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chemical Reviews* 106, 1589-1615
101. Knapp, B., Ospina, L., and Deane, C.M. (2018) Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *Journal of Chemical Theory and Computation* 14, 6127-6138
102. Hruska, E. et al. (2018) Quantitative Comparison of Adaptive Sampling Methods for Protein Dynamics. *The Journal of Chemical Physics* 149, 244119
103. Bowman, G.R., Ensign, D.L., and Pande, V.S. (2010) Enhanced Modeling Via Network Theory: Adaptive Sampling of Markov State Models. *Journal of Chemical Theory and Computation* 6, 787-794
104. Bolhuis, P.G. et al. (2002) Transition Path Sampling: Throwing Ropes over Rough Mountain Passes, in the Dark. *Annual Review of Physical Chemistry* 53, 291-318
105. Isralewitz, B., Gao, M., and Schulten, K. (2001) Steered Molecular Dynamics and Mechanical Functions of Proteins. *Current Opinion in Structural Biology* 11, 224-230
106. Barducci, A., Bonomi, M., and Parrinello, M. (2011) Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1, 826-843
107. Bruccoleri, R.E. and Karplus, M. (1990) Conformational Sampling Using High-Temperature Molecular Dynamics. *Biopolymers: Original Research on Biomolecules* 29, 1847-1862
108. Sugita, Y. and Okamoto, Y. (1999) Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chemical physics letters* 314, 141-151
109. Ferruz, N. and De Fabritiis, G. (2016) Binding Kinetics in Drug Discovery. *Molecular Informatics* 35, 216-226
110. Camunas-Soler, J., Alemany, A., and Ritort, F. (2017) Experimental Measurement of Binding Energy, Selectivity, and Allostery Using Fluctuation Theorems. *Science* 355, 412-415
111. Phizicky, E.M. and Fields, S. (1995) Protein-Protein Interactions: Methods for Detection and Analysis. *Microbiological reviews* 59, 94-123
112. Vuignier, K. et al. (2010) Drug–Protein Binding: A Critical Review of Analytical Tools. *Analytical and Bioanalytical Chemistry* 398, 53-66
113. Metzler, D.E. (1977) Biochemistry: The Chemical Reactions of Living Cells. *Postępy* 265
114. Corzo, J. (2006) Time, the Forgotten Dimension of Ligand Binding Teaching. *Biochemistry and Molecular Biology Education* 34, 413-416
115. Vijayakumar, M. et al. (1998) Electrostatic Enhancement of Diffusion-Controlled Protein-Protein Association: Comparison of Theory and Experiment on Barnase and Barstar. *The Journal of Molecular Biology* 278, 1015-1024
116. Schreiber, G., Haran, G., and Zhou, H.X. (2009) Fundamental Aspects of Protein−Protein Association Kinetics. *Chemical Reviews* 109, 839-860
117. Pulido, D. et al. (2018) Design of a True Bivalent Ligand with Picomolar Binding Affinity for a G Protein-Coupled Receptor Homodimer. *Journal of Medicinal Chemistry* 61, 9335-9346
118. Mitchell, M.J. and McCammon, J.A. (1991) Free Energy Difference Calculations by Thermodynamic Integration: Difficulties in Obtaining a Precise Value. *The Journal of Computational Chemistry* 12, 271-275
119. Jarzynski, C. (1997) Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters* 78, 2690
120. Zwanzig, R.W. (1954) High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* 22, 1420-1426

121. Torrie, G.M. and Valleau, J.P. (1977) Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *Journal of Computational Physics* 23, 187-199
122. Genheden, S. and Ryde, U. (2015) The Mm/Pbsa and Mm/Gbsa Methods to Estimate Ligand-Binding Affinities. *Expert opinion on drug discovery* 10, 449-461
123. Xue, L.C. et al. (2016) Prodigy: A Web Server for Predicting the Binding Affinity of Protein-Protein Complexes. *Bioinformatics* 32, 3676-3678
124. Elcock, A.H. et al. (1999) Computer Simulation of Protein-Protein Association Kinetics: Acetylcholinesterase-Fasciculin. *The Journal of Molecular Biology* 291, 149-162
125. Ferrenberg, A.M. and Swendsen, R.H. (1989) Optimized Monte Carlo Data Analysis. *Computers in Physics* 3, 101-104
126. Kumar, S. et al. (1992) The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules. I. The Method. *The Journal of Computational Chemistry* 13, 1011-1021
127. Ajay and Murcko, M.A. (1995) Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes. *Journal of Medicinal Chemistry* 38, 4953-4967
128. Morris, G.M. et al. (2009) Autodock4 and Autodocktools4: Automated Docking with Selective Receptor Flexibility. *The Journal of Computational Chemistry* 30, 2785-2791
129. Trott, O. and Olson, A.J. (2010) Autodock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *The Journal of Computational Chemistry* 31, 455-461
130. Friesner, R.A. et al. (2004) Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of medicinal chemistry* 47, 1739-1749
131. Lyskov, S. and Gray, J.J. (2008) The Rosettadock Server for Local Protein–Protein Docking. *Nucleic Acids Research* 36, W233-W238
132. Verdonk, M.L. et al. (2003) Improved Protein–Ligand Docking Using Gold. *Proteins: Structure, Function, and Bioinformatics* 52, 609-623
133. Hwang, H. et al. (2010) Protein–Protein Docking Benchmark Version 4.0. *Proteins: Structure, Function, and Bioinformatics* 78, 3111-3114
134. Chandrika, B.R., Subramanian, J., and Sharma, S.D. (2009) Managing Protein Flexibility in Docking and Its Applications. *Drug Discovery Today* 14, 394-400
135. Amaro, R.E. et al. (2018) Ensemble Docking in Drug Discovery. *Biophysical Journal* 114, 2271-2278
136. Huang, S.Y. and Zou, X. (2007) Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking. *Proteins: Structure, Function, and Bioinformatics* 66, 399-421
137. Kollman, P.A. et al. (2000) Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research* 33, 889-897
138. Fogolari, F., Brigo, A., and Molinari, H. (2002) The Poisson–Boltzmann Equation for Biomolecular Electrostatics: A Tool for Structural Biology. *Journal of Molecular Recognition* 15, 377-392
139. Sitkoff, D., Sharp, K.A., and Honig, B. (1994) Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *The Journal of Physical Chemistry* 98, 1978-1988
140. Srinivasan, J. et al. (1998) Continuum Solvent Studies of the Stability of DNA, Rna, and Phosphoramidate−DNA Helices. *Journal of the American Chemical Society* 120, 9401-9409
141. Hayward, S. and Groot, B.L.d. (2008) Normal Modes and Essential Dynamics. *Molecular Modeling of Proteins*. Springer. 89-106
142. Chen, F. et al. (2016) Assessing the Performance of the Mm/Pbsa and Mm/Gbsa Methods. 6. Capability to Predict Protein–Protein Binding Free Energies and Re-Rank Binding Poses Generated by Protein–Protein Docking. *Physical Chemistry Chemical Physics* 18, 22129-22139

143. Kabsch, W. (1976) A Solution for the Best Rotation to Relate Two Sets of Vectors. *Acta Crystallographica Section A* 32, 922-923
144. Rose, G.D. et al. (1985) Hydrophobicity of Amino Acid Residues in Globular Proteins. *Science* 229, 834-838
145. Miller, S. et al. (1987) Interior and Surface of Monomeric Proteins. *The Journal of Molecular Biology* 196, 641-656
146. Tien, M.Z. et al. (2013) Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE* 8, e80635
147. Ran, X. and Gestwicki, J.E. (2018) Inhibitors of Protein–Protein Interactions (Ppis): An Analysis of Scaffold Choices and Buried Surface Area. *Current Opinion in Chemical Biology* 44, 75-86
148. Rakers, C. et al. (2015) Computational Close up on Protein–Protein Interactions: How to Unravel the Invisible Using Molecular Dynamics Simulations? *WIREs Computational Molecular Science* 5, 345-359
149. Kastritis, P.L. and Bonvin, A.M. (2013) On the Binding Affinity of Macromolecular Interactions: Daring to Ask Why Proteins Interact. *Journal of The Royal Society Interface* 10, 20120835
150. Vangone, A. and Bonvin, A.M. (2015) Contacts-Based Prediction of Binding Affinity in Protein–Protein Complexes. *eLife* 4, e07454
151. Abdel-Azeim, S. et al. (2014) Mdcons: Intermolecular Contact Maps as a Tool to Analyze the Interface of Protein Complexes from Molecular Dynamics Trajectories. *BMC Bioinformatics* 15, S1
152. Feher, V.A. et al. (2014) Computational Approaches to Mapping Allosteric Pathways. *Current Opinion in Structural Biology* 25, 98-103
153. Edgcomb, S.P. and Murphy, K.P. (2002) Variability in the Pka of Histidine Side-Chains Correlates with Burial within Proteins. *Proteins: Structure, Function, and Bioinformatics* 49, 1-6
154. Lin, J., Pozharski, E., and Wilson, M.A. (2017) Short Carboxylic Acid–Carboxylate Hydrogen Bonds Can Have Fully Localized Protons. *Biochemistry* 56, 391-402
155. Wendler, K. et al. (2010) Estimating the Hydrogen Bond Energy. *The Journal of Physical Chemistry A* 114, 9529-9536
156. Steiner, T. (2002) The Hydrogen Bond in the Solid State. *Angewandte Chemie International Edition* 41, 48-76
157. Herschlag, D. and Pinney, M.M. (2018) Hydrogen Bonds: Simple after All? *Biochemistry* 57, 3338-3352
158. Dall'Acqua, W. and Carter, P. (2000) Substrate-Assisted Catalysis: Molecular Basis and Biological Significance. *Protein Science* 9, 1-9
159. Seo, M.-H. et al. (2014) Protein Conformational Dynamics Dictate the Binding Affinity for a Ligand. *Nature Communications* 5, 3724
160. Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963) Stereochemistry of Polypeptide Chain Configurations. *The Journal of Molecular Biology* 7, 95-99
161. Sheather, S.J. and Jones, M.C. (1991) A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* 53, 683-690
162. Mu, Y., Nguyen, P.H., and Stock, G. (2005) Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Structure, Function, and Bioinformatics* 58, 45-52
163. Altis, A. et al. (2007) Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *The Journal of Chemical Physics* 126, 244111
164. Altis, A. et al. (2008) Construction of the Free Energy Landscape of Biomolecules Via Dihedral Angle Principal Component Analysis. *The Journal of Chemical Physics* 128, 245102

165. Jolliffe, I.T. and Cadima, J. (2016) Principal Component Analysis: A Review and Recent Developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 20150202
166. Husic, B.E. and Pande, V.S. (2018) Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* 140, 2386-2396
167. Bonati, L., Piccini, G., and Parrinello, M. (2021) Deep Learning the Slow Modes for Rare Events Sampling. *Proceedings of the National Academy of Sciences* 118, e2113533118
168. Bussi, G. and Laio, A. (2020) Using Metadynamics to Explore Complex Free-Energy Landscapes. *Nature Reviews Physics* 2, 200-212
169. Pérez-Hernández, G. and Noé, F. (2016) Hierarchical Time-Lagged Independent Component Analysis: Computing Slow Modes and Reaction Coordinates for Large Molecular Systems. *Journal of Chemical Theory and Computation* 12, 6118-6129
170. Van der Maaten, L. and Hinton, G. (2008) Visualizing Data Using T-Sne. *Journal of Machine Learning Research* 9,
171. Hershey, J.R. and Olsen, P.A. (2007) Approximating the Kullback Leibler Divergence between Gaussian Mixture Models. *IEEE International Conference on Acoustics, Speech and Signal Processing* (Honolulu, HI, USA). IEEE, 317-320
172. McInnes, L., Healy, J., and Melville, J. (2018) Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*
173. Becht, E. et al. (2019) Dimensionality Reduction for Visualizing Single-Cell Data Using Umap. *Nature Biotechnology* 37, 38-44
174. Kotsiantis, S.B., Zaharakis, I., and Pintelas, P. (2007) Supervised Machine Learning: A Review of Classification Techniques. *Emerging Artificial Intelligence Applications in Computer Engineering* 160, 3-24
175. Maxwell, A.E., Warner, T.A., and Fang, F. (2018) Implementation of Machine-Learning Classification in Remote Sensing: An Applied Review. *International Journal of Remote Sensing* 39, 2784-2817
176. Aggarwal, C.C., Hinneburg, A., and Keim, D.A. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. *Database Theory — ICDT 2001* (Berlin, Heidelberg). Springer Berlin Heidelberg, 420-434
177. Hartigan, J.A. (1975) *Clustering Algorithms*. Hoboken (USA), John Wiley & Sons, Inc.
178. Celebi, M.E., Kingravi, H.A., and Vela, P.A. (2013) A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm. *Expert Systems with Applications* 40, 200-210
179. David, A. and Vassilvitskii, S. (2007) K-Means++: The Advantages of Careful Seeding. *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (New Orleans Louisiana). SIAM, 1027–1035
180. Schubert, E. et al. (2017) Dbscan Revisited, Revisited: Why and How You Should (Still) Use Dbscan. *ACM Transactions on Database Systems (TODS)* 42, 1-21
181. Ankerst, M. et al. (1999) Optics: Ordering Points to Identify the Clustering Structure. *ACM Sigmod record* 28, 49-60
182. McInnes, L., Healy, J., and Astels, S. (2017) Hdbscan: Hierarchical Density Based Clustering. *Journal of Open Source Software* 2, 205
183. Ward Jr, J.H. (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58, 236-244
184. de Planque, M.R.R. et al. (1998) Influence of Lipid/Peptide Hydrophobic Mismatch on the Thickness of Diacylphosphatidylcholine Bilayers. A 2h Nmr and Esr Study Using Designed Transmembrane A-Helical Peptides and Gramicidin A. *Biochemistry* 37, 9333-9345
185. Phan, H.T.M. et al. (2015) Investigation of Bovine Serum Albumin (Bsa) Attachment onto Self-Assembled Monolayers (Sams) Using Combinatorial Quartz Crystal Microbalance with Dissipation (Qcm-D) and Spectroscopic Ellipsometry (Se). *PLOS ONE* 10, e0141282

186. McTaguc, J.P., Nielsen, M., and Passell, L. (1979) Neutron Scattering by Adsorbed Monolayers. *Critical Reviews in Solid State and Material Sciences* 8, 135-155
187. Douliez, J.P., Léonard, A., and Dufourc, E.J. (1995) Restatement of Order Parameters in Biomembranes: Calculation of C-C Bond Order Parameters from C-D Quadrupolar Splittings. *Biophysical Journal* 68, 1727-1739
188. Vermeer, L.S. et al. (2007) Acyl Chain Order Parameter Profiles in Phospholipid Bilayers: Computation from Molecular Dynamics Simulations and Comparison with 2h Nmr Experiments. *European Biophysics Journal* 36, 919-931
189. Ipsen, J.H., Mouritsen, O.G., and Bloom, M. (1990) Relationships between Lipid Membrane Area, Hydrophobic Thickness, and Acyl-Chain Orientational Order. The Effects of Cholesterol. *Biophysical Journal* 57, 405-412
190. Fahey, P.F. et al. (1977) Lateral Diffusion in Planar Lipid Bilayers. *Science* 195, 305-306
191. Creutznacher, R. et al. (2020) Chemical-Shift Perturbations Reflect Bile Acid Binding to Norovirus Coat Protein: Recognition Comes in Different Flavors. *ChemBioChem* 21, 1007-1021
192. van Beek, J. et al. (2018) Molecular Surveillance of Norovirus, 2005–16: An Epidemiological Analysis of Data Collected from the Noronet Network. *The Lancet Infectious Diseases* 18, 545-553
193. Ahmed, S.M. et al. (2014) Global Prevalence of Norovirus in Cases of Gastroenteritis: A Systematic Review and Meta-Analysis. *The Lancet Infectious Diseases* 14, 725-730
194. Bok, K. and Green, K.Y. (2012) Norovirus Gastroenteritis in Immunocompromised Patients. *New England Journal of Medicine* 367, 2126-2132
195. Nguyen, G.T. et al. (2017) A Systematic Review and Meta-Analysis of the Prevalence of Norovirus in Cases of Gastroenteritis in Developing Countries. *Medicine* 96,
196. Taube, S., Mallagaray, A., and Peters, T. (2018) Norovirus, Glycans and Attachment. *Current Opinion in Virology* 31, 33-42
197. Heggelund, J.E. et al. (2017) Histo-Blood Group Antigens as Mediators of Infections. *Current Opinion in Structural Biology* 44, 190-200
198. Nasir, W. et al. (2017) Histo-Blood Group Antigen Presentation Is Critical for Binding of Norovirus Vlp to Glycosphingolipids in Model Membranes. *ACS Chemical Biology* 12, 1288-1296
199. Koromyslova, A. et al. (2017) Human Norovirus Inhibition by a Human Milk Oligosaccharide. *Virology* 508, 81-89
200. Koromyslova, A.D., White, P.A., and Hansman, G.S. (2015) Treatment of Norovirus Particles with Citrate. *Virology* 485, 199-204
201. Zhang, X.-F. et al. (2013) Inhibition of Histo-Blood Group Antigen Binding as a Novel Strategy to Block Norovirus Infections. *PLOS ONE* 8, e69379
202. Oka, T. et al. (2018) Attempts to Grow Human Noroviruses, a Sapovirus, and a Bovine Norovirus in Vitro. *PLOS ONE* 13, e0178157
203. Ettayebi, K. et al. (2016) Replication of Human Noroviruses in Stem Cell–Derived Human Enteroids. *Science* 353, 1387-1393
204. Kilic, T., Koromyslova, A., and Hansman, G.S. (2019) Structural Basis for Human Norovirus Capsid Binding to Bile Acids. *Journal of virology* 93, e01581-01518
205. Bartnicki, E. et al. (2017) Recent Advances in Understanding Noroviruses. *F1000Research* 6(F1000 Faculty Review)
206. Nelson, C.A. et al. (2018) Structural Basis for Murine Norovirus Engagement of Bile Acids and the Cd300lf Receptor. *Proceedings of the National Academy of Sciences* 115, E9201-E9210
207. Brown, S.P. and Hajduk, P.J. (2006) Effects of Conformational Dynamics on Predicted Protein Druggability. *ChemMedChem* 1, 70-72
208. Stank, A. et al. (2016) Protein Binding Pocket Dynamics. *Accounts of Chemical Research* 49, 809-815

209. Seethala, R. and Fernandes, P.B. (2001) Handbook of Drug Screening. *Drugs and the pharmaceutical sciences* 114, 1-596
210. Jhoti, H. and Leach, A.R. (2007) *Structure-Based Drug Discovery* (1st Edition). Dodrecht (NL), Springer.
211. Danley, D.E. (2006) Crystallization to Obtain Protein–Ligand Complexes for Structure-Aided Drug Design. *Acta Crystallographica Section D: Biological Crystallography* 62, 569-575
212. Williamson, M.P. (2013) Using Chemical Shift Perturbation to Characterise Ligand Binding. *Progress in Nuclear Magnetic Resonance Spectroscopy* 73, 1-16
213. Mallagaray, A. et al. (2019) A Post-Translational Modification of Human Norovirus Capsid Protein Attenuates Glycan Binding. *Nature Communications* 10, 1320
214. Feng, Z. et al. (2004) Ligand Depot: A Data Warehouse for Ligands Bound to Macromolecules. *Bioinformatics* 20, 2153-2155
215. Klauda, J.B. et al. (2010) Update of the Charmm All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *The Journal of Physical Chemistry B* 114, 7830-7843
216. Jo, S. et al. (2014) Chapter Eight - Charmm-Gui Pdb Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. *Advances in Protein Chemistry and Structural Biology*. Academic Press. 235-265
217. Singh, B.K. et al. (2015) Human Noroviruses' Fondness for Histo-Blood Group Antigens. *Journal of Virology* 89, 2024-2040
218. Brooks, B.R. et al. (2009) Charmm: The Biomolecular Simulation Program. *The Journal of Computational Chemistry* 30, 1545-1614
219. Jorgensen, W.L. et al. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* 79, 926-935
220. Neria, E., Fischer, S., and Karplus, M. (1996) Simulation of Activation Free Energies in Molecular Systems. *The Journal of Chemical Physics* 105, 1902-1921
221. Gasteiger, J. and Marsili, M. (1980) Iterative Partial Equalization of Orbital Electronegativity—a Rapid Access to Atomic Charges. *Tetrahedron* 36, 3219-3228
222. Páll, S. and Hess, B. (2013) A Flexible Algorithm for Calculating Pair Interactions on Simd Architectures. *Computer Physics Communications* 184, 2641-2650
223. Salomon-Ferrer, R. et al. (2013) Routine Microsecond Molecular Dynamics Simulations with Amber on Gpus. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* 9, 3878-3888
224. Li, P. et al. (2013) Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *Journal of Chemical Theory and Computation* 9, 2733-2748
225. Hoover, W.G. and Holian, B.L. (1996) Kinetic Moments Method for the Canonical Ensemble Distribution. *Physics Letters A* 211, 253-257
226. Parrinello, M. and Rahman, A. (1981) Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *Journal of Applied Physics* 52, 7182-7190
227. Hess, B. (2008) P-Lincs: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* 4, 116-122
228. Hess, B. et al. (1997) Lincs: A Linear Constraint Solver for Molecular Simulations. *The Journal of Computational Chemistry* 18, 1463-1472
229. Wagner, J.R. et al. (2017) Povme 3.0: Software for Mapping Binding Pocket Flexibility. *Journal of Chemical Theory and Computation* 13, 4584-4592
230. McGibbon, Robert T. et al. (2015) Mdtraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* 109, 1528-1532
231. Zhao, X. et al. (2019) Convex Hull Principle for Classification and Phylogeny of Eukaryotic Proteins. *Genomics* 111, 1777-1784
232. Totrov, M. and Abagyan, R. (2008) Flexible Ligand Docking to Multiple Receptor Conformations: A Practical Alternative. *Current Opinion in Structural Biology* 18, 178-184

233. Pagadala, N.S., Syed, K., and Tuszynski, J. (2017) Software for Molecular Docking: A Review. *Biophysical Reviews* 9, 91-102
234. Stank, A. et al. (2017) Trapp Webserver: Predicting Protein Binding Site Flexibility and Detecting Transient Binding Pockets. *Nucleic Acids Research* 45, W325-W330
235. Oleinikovas, V. et al. (2016) Understanding Cryptic Pocket Formation in Protein Targets by Enhanced Sampling Simulations. *Journal of the American Chemical Society* 138, 14257-14263
236. Lazim, R., Suh, D., and Choi, S. (2020) Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems. *International Journal of Molecular Sciences* 21, 6339
237. De Vivo, M. et al. (2016) Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry* 59, 4035-4061
238. Reymond, J.-L. (2015) The Chemical Space Project. *Accounts of Chemical Research* 48, 722-730
239. Rueda, M., Bottegoni, G., and Abagyan, R. (2010) Recipes for the Selection of Experimental Protein Conformations for Virtual Screening. *Journal of Chemical Information and Modeling* 50, 186-193
240. Korb, O. et al. (2012) Potential and Limitations of Ensemble Docking. *Journal of Chemical Information and Modeling* 52, 1262-1274
241. Evangelista Falcon, W. et al. (2019) Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations Are Needed to Reproduce Known Ligand Binding? *The Journal of Physical Chemistry B* 123, 5189-5195
242. Strecker, C. and Meyer, B. (2018) Plasticity of the Binding Site of Renin: Optimized Selection of Protein Structures for Ensemble Docking. *Journal of Chemical Information and Modeling* 58, 1121-1131
243. Schulze-Niemand, E., Naumann, M., and Stein, M. (2021) The Activation and Selectivity of the Legionella Ravd Deubiquitinase. *Frontiers in Molecular Biosciences* 8
244. Schulze-Niemand, E., Naumann, M., and Stein, M. (2022) Substrate-Assisted Activation and Selectivity of the Bacterial Ravd Effector Deubiquitinylase. *Proteins: Structure, Function, and Bioinformatics* 90, 947-958
245. Varshavsky, A. (2017) The Ubiquitin System, Autophagy, and Regulated Protein Degradation. *Annual Review of Biochemistry* 86, 123-128
246. Zinngrebe, J. et al. (2014) Ubiquitin in the Immune System. *EMBO reports* 15, 28-45
247. Ben-Neriah, Y. (2002) Regulatory Functions of Ubiquitination in the Immune System. *Nature immunology* 3, 20-26
248. Fiil, B.K. and Gyrd-Hansen, M. (2021) The Met1-Linked Ubiquitin Machinery in Inflammation and Infection. *Cell Death & Differentiation* 28, 557-569
249. Pickart, C.M. (2001) Mechanisms Underlying Ubiquitination. *Annual Review of Biochemistry* 70,
250. Swatek, K.N. and Komander, D. (2016) Ubiquitin Modifications. *Cell research* 26, 399-422
251. Komander, D. (2009) The Emerging Complexity of Protein Ubiquitination. *Biochemical Society Transactions* 37, 937-953
252. Mevissen, T.E. and Komander, D. (2017) Mechanisms of Deubiquitinase Specificity and Regulation. *Annual Review of Biochemistry* 86, 159-192
253. Komander, D., Clague, M.J., and Urbé, S. (2009) Breaking the Chains: Structure and Function of the Deubiquitinases. *Nature Reviews Molecular Cell Biology* 10, 550-563
254. Clague, M.J., Urbé, S., and Komander, D. (2019) Breaking the Chains: Deubiquitylating Enzyme Specificity Begets Function. *Nature Reviews Molecular Cell Biology* 20, 338-352
255. Mevissen, T.E. et al. (2013) Otu Deubiquitinases Reveal Mechanisms of Linkage Specificity and Enable Ubiquitin Chain Restriction Analysis. *Cell* 154, 169-184
256. Mevissen, T.E. et al. (2016) Molecular Basis of Lys11-Polyubiquitin Specificity in the Deubiquitinase Cezanne. *Nature* 538, 402-405

257. Edelmann, M.J. et al. (2009) Structural Basis and Specificity of Human Otubain 1-Mediated Deubiquitination. *Biochemical Journal* 418, 379-390
258. Fiil, B.K. et al. (2013) Otulin Restricts Met1-Linked Ubiquitination to Control Innate Immune Signaling. *Molecular Cell* 50, 818-830
259. Keusekotten, K. et al. (2013) Otulin Antagonizes Lubac Signaling by Specifically Hydrolyzing Met1-Linked Polyubiquitin. *Cell* 153, 1312-1326
260. Jahan, A.S., Elbæk, C.R., and Damgaard, R.B. (2021) Met1-Linked Ubiquitin Signalling in Health and Disease: Inflammation, Immunity, Cancer, and Beyond. *Cell Death and Differentiation* 28, 473-492
261. Schubert, A.F. et al. (2020) Identification and Characterization of Diverse Otu Deubiquitinases in Bacteria. *The EMBO journal* 39, e105127
262. Bailey-Elkin, B.A. et al. (2014) Viral Otu Deubiquitinases: A Structural and Functional Comparison. *PLOS pathogens* 10, e1003894
263. Franklin, T.G. and Pruneda, J.N. (2021) Bacteria Make Surgical Strikes on Host Ubiquitin Signaling. *PLOS pathogens* 17, e1009341
264. Wan, M. et al. (2019) A Bacterial Effector Deubiquitinase Specifically Hydrolyses Linear Ubiquitin Chains to Inhibit Host Inflammatory Signalling. *Nature Microbiology* 4, 1282-1293
265. Hermanns, T. and Hofmann, K. (2019) Bacterial Dubs: Deubiquitination Beyond the Seven Classes. *Biochemical Society Transactions* 47, 1857-1866
266. Humphrey, W., Dalke, A., and Schulten, K. (1996) Vmd: Visual Molecular Dynamics. *Journal of Molecular Graphics and Modelling* 14, 33-38, 27-38
267. Eastman, P. et al. (2017) Openmm 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLOS Computational Biology* 13, e1005659
268. Izaguirre, J.A. et al. (2001) Langevin Stabilization of Molecular Dynamics. *The Journal of Chemical Physics* 114, 2090-2098
269. Åqvist, J. et al. (2004) Molecular Dynamics Simulations of Water and Biomolecules with a Monte Carlo Constant Pressure Algorithm. *Chemical physics letters* 384, 288-294
270. Gowers, R.J. et al. (2019) *Mdanalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations* Los Alamos, NM (United States), Los Alamos National Lab.(LANL)
271. Michaud-Agrawal, N. et al. (2011) Mdanalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations. *The Journal of Computational Chemistry* 32, 2319-2327
272. Harris, C.R. et al. (2020) Array Programming with Numpy. *Nature* 585, 357-362
273. Virtanen, P. et al. (2020) Scipy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, 261-272
274. Hunter, J.D. (2007) Matplotlib: A 2d Graphics Environment. *Computing in Science & Engineering* 9, 90-95
275. Shrake, A. and Rupley, J.A. (1973) Environment and Exposure to Solvent of Protein Atoms. Lysozyme and Insulin. *The Journal of Molecular Biology* 79, 351-371
276. Vangone, A. and Bonvin, A.M.J.J. (2015) Contacts-Based Prediction of Binding Affinity in Protein–Protein Complexes. *eLife* 4, e07454
277. Chakrabarti, P. and Janin, J. (2002) Dissecting Protein-Protein Recognition Sites. *Proteins: Structure, Function, and Bioinformatics* 47, 334-343
278. Horton, N. and Lewis, M. (1992) Calculation of the Free Energy of Association for Protein Complexes. *Protein Science* 1, 169-181
279. Lo Conte, L., Chothia, C., and Janin, J. (1999) The Atomic Structure of Protein-Protein Recognition Sites. *The Journal of Molecular Biology* 285, 2177-2198
280. Janin, J. and Chothia, C. (1990) The Structure of Protein-Protein Recognition Sites. *Journal of Biological Chemistry* 265, 16027-16030
281. Schwarz, F. and Aebi, M. (2011) Mechanisms and Principles of N-Linked Protein Glycosylation. *Current Opinion in Structural Biology* 21, 576-582
282. Schjoldager, K.T. et al. (2020) Global View of Human Protein Glycosylation Pathways and Functions. *Nature Reviews Molecular Cell Biology* 21, 729-749

283. Steentoft, C. et al. (2013) Precision Mapping of the Human O-Galnac Glycoproteome through Simplecell Technology. *The EMBO journal* 32, 1478-1488
284. Hart, G.W. (2019) Nutrient Regulation of Signaling and Transcription. *Journal of Biological Chemistry* 294, 2211-2231
285. Varki, A. (2017) Biological Roles of Glycans. *Glycobiology* 27, 3-49
286. Lin, B. et al. (2020) Role of Protein Glycosylation in Host-Pathogen Interaction. *Cells* 9, 1022
287. Ohtsubo, K. and Marth, J.D. (2006) Glycosylation in Cellular Mechanisms of Health and Disease. *Cell* 126, 855-867
288. Stowell, S.R., Ju, T., and Cummings, R.D. (2015) Protein Glycosylation in Cancer. *Annual Review of Pathology: Mechanisms of Disease* 10, 473-510
289. Seeling, M., Brückner, C., and Nimmerjahn, F. (2017) Differential Antibody Glycosylation in Autoimmunity: Sweet Biomarker or Modulator of Disease Activity? *Nature Reviews Rheumatology* 13, 621-630
290. Rudd, P.M. and Dwek, R.A. (1997) Glycosylation: Heterogeneity and the 3d Structure of Proteins. *Critical reviews in biochemistry and molecular biology* 32, 1-100
291. Goldman, R. and Sanda, M. (2015) Targeted Methods for Quantitative Analysis of Protein Glycosylation. *PROTEOMICS–Clinical Applications* 9, 17-32
292. Xiao, H. et al. (2019) Global and Site-Specific Analysis of Protein Glycosylation in Complex Biological Systems with Mass Spectrometry. *Mass Spectrometry Reviews* 38, 356-379
293. Pralow, A. et al. (2021) Comprehensive N-Glycosylation Analysis of the Influenza a Virus Proteins Ha and Na from Adherent and Suspension Mdck Cells. *The FEBS Journal*
294. Prestegard, J.H. (2021) A Perspective on the Pdb's Impact on the Field of Glycobiology. *Journal of Biological Chemistry* 296, 100556-100556
295. Frank, M. and Schloissnig, S. (2010) Bioinformatics and Molecular Modeling in Glycobiology. *Cellular and Molecular Life Sciences* 67, 2749-2772
296. Pérez, S., Meyer, C., and Imberty, A. (1995) Practical Tools for Molecular Modeling of Complex Carbohydrates and Their Interactions with Proteins. *Molecular Engineering* 5, 271-300
297. Woods, R.J. (2018) Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chemical Reviews* 118, 8005-8024
298. Omotuyi, I.O. et al. (2020) Atomistic Simulation Reveals Structural Mechanisms Underlying D614g Spike Glycoprotein-Enhanced Fitness in Sars-Cov-2. *The Journal of Computational Chemistry* 41, 2158-2161
299. Grant, O.C. et al. (2020) Analysis of the Sars-Cov-2 Spike Protein Glycan Shield Reveals Implications for Immune Recognition. *Scientific Reports* 10, 1-11
300. Casalino, L. et al. (2020) Beyond Shielding: The Roles of Glycans in the Sars-Cov-2 Spike Protein. *ACS Central Science* 6, 1722-1734
301. Singh, A. (2020) Imaging Single Glycan Molecules. *Nature Methods* 17, 757-757
302. Mucha, E. et al. (2019) In-Depth Structural Analysis of Glycans in the Gas Phase. *Chemical Science* 10, 1272-1284
303. Alibay, I. and Bryce, R.A. (2019) Ring Puckering Landscapes of Glycosaminoglycan-Related Monosaccharides from Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling* 59, 4729-4741
304. Lütteke, T. et al. (2006) Glycosciences.De: An Internet Portal to Support Glycomics and Glycobiology Research. *Glycobiology* 16, 71R-81R
305. Syed, R.S. et al. (1998) Efficiency of Signalling through Cytokine Receptors Depends Critically on Receptor Orientation. *Nature* 395, 511-516
306. Park, S.J. et al. (2019) Charmm-Gui Glycan Modeler for Modeling and Simulation of Carbohydrates and Glycoconjugates. *Glycobiology* 29, 320-331
307. Jo, S. et al. (2008) Charmm-Gui: A Web-Based Graphical User Interface for Charmm. *The Journal of Computational Chemistry* 29, 1859-1865

308. Hatcher, E.R., Guvench, O., and MacKerell, A.D. (2009) Charmm Additive All-Atom Force Field for Acyclic Polyalcohols, Acyclic Carbohydrates, and Inositol. *Journal of Chemical Theory and Computation* 5, 1315-1327
309. Páll, S. et al. (2015) Tackling Exascale Software Challenges in Molecular Dynamics Simulations with Gromacs. *Solving Software Challenges for Exascale* (1st Edition). Basel, Springer International Publishing. 3-27
310. Braga, C. and Travis, K.P. (2005) A Configurational Temperature Nosé-Hoover Thermostat. *The Journal of Chemical Physics* 123, 134101
311. Hess, B. et al. (1997) Lincs: A Linear Constraint Solver for Molecular Simulations. *The Journal of Computational Chemistry* 18, 1463-1472
312. Parrinello, M. and Rahman, A. (1980) Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Physical Review Letters* 45, 1196-1199
313. Martonak, R., Laio, A., and Parrinello, M. (2003) Predicting Crystal Structures: The Parrinello-Rahman Method Revisited. *Physical Review Letters* 90, 075503
314. Johnson, R.A. and Wehrly, T. (1977) Measures and Models for Angular Correlation and Angular-Linear Correlation. *Journal of the Royal Statistical Society. Series B (Methodological)* 39, 222-229
315. Müller, M.M. (2018) Post-Translational Modifications of Protein Backbones: Unique Functions, Mechanisms, and Challenges. *Biochemistry* 57, 177-185
316. Catak, S. et al. (2006) Reaction Mechanism of Deamidation of Asparaginyl Residues in Peptides: Effect of Solvent Molecules. *The Journal of Physical Chemistry A* 110, 8354-8365
317. Chelius, D., Rehder, D.S., and Bondarenko, P.V. (2005) Identification and Characterization of Deamidation Sites in the Conserved Regions of Human Immunoglobulin Gamma Antibodies. *Analytical Chemistry* 77, 6004-6011
318. Nowak, C., Tiwari, A., and Liu, H. (2018) Asparagine Deamidation in a Complementarity Determining Region of a Recombinant Monoclonal Antibody in Complex with Antigen. *Analytical Chemistry* 90, 6998-7003
319. Phillips, J.J. et al. (2017) Rate of Asparagine Deamidation in a Monoclonal Antibody Correlating with Hydrogen Exchange Rate at Adjacent Downstream Residues. *Analytical Chemistry* 89, 2361-2368
320. Curnis, F. et al. (2006) Spontaneous Formation of L-Isoaspartate and Gain of Function in Fibronectin*. *Journal of Biological Chemistry* 281, 36466-36476
321. Zhang, T. et al. (2020) An Unusually Rapid Protein Backbone Modification Stabilizes the Essential Bacterial Enzyme Mura. *Biochemistry* 59, 3683-3695
322. Plotnikov, N.V. et al. (2017) Quantifying the Risks of Asparagine Deamidation and Aspartate Isomerization in Biopharmaceuticals by Computing Reaction Free-Energy Surfaces. *The Journal of Physical Chemistry B* 121, 719-730
323. Robinson, N.E. and Robinson, A.B. (2001) Deamidation of Human Proteins. *Proceedings of the National Academy of Sciences* 98, 12409-12413
324. Lorenzo, J.R., Alonso, L.G., and Sánchez, I.E. (2015) Prediction of Spontaneous Protein Deamidation from Sequence-Derived Secondary Structure and Intrinsic Disorder. *PLOS ONE* 10, e0145186
325. Sydow, J.F. et al. (2014) Structure-Based Prediction of Asparagine and Aspartate Degradation Sites in Antibody Variable Regions. *PLOS ONE* 9, e100736
326. Jia, L. and Sun, Y. (2017) Protein Asparagine Deamidation Prediction Based on Structures with Machine Learning Methods. *PLOS ONE* 12, e0181347
327. Delmar, J.A. et al. (2019) Machine Learning Enables Accurate Prediction of Asparagine Deamidation Probability and Rate. *Molecular Therapy - Methods & Clinical Development* 15, 264-274
328. Radkiewicz, J.L. et al. (1996) Accelerated Racemization of Aspartic Acid and Asparagine Residues Via Succinimide Intermediates: An Ab Initio Theoretical Exploration of Mechanism. *Journal of the American Chemical Society* 118, 9148-9155

329. Ugur, I. et al. (2015) Why Does Asn71 Deamidate Faster Than Asn15 in the Enzyme Triosephosphate Isomerase? Answers from Microsecond Molecular Dynamics Simulation and Qm/Mm Free Energy Calculations. *Biochemistry* 54, 1429-1439
330. Yan, Q. et al. (2018) Structure Based Prediction of Asparagine Deamidation Propensity in Monoclonal Antibodies. *mAbs* 10, 901-912
331. Creutznacher, R. et al. (2021) Nmr Experiments Shed New Light on Glycan Recognition by Human and Murine Norovirus Capsid Proteins. *Viruses* 13, 416
332. Mallory, M.L. et al. (2019) Gii. 4 Human Norovirus: Surveying the Antigenic Landscape. *Viruses* 11, 177
333. Van Der Walt, S., Colbert, S.C., and Varoquaux, G. (2011) The Numpy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 13, 22-30
334. Bürgi, H.B. et al. (1974) Stereochemistry of Reaction Paths at Carbonyl Centres. *Tetrahedron* 30, 1563-1572
335. Heathcock, C.H. and Flippin, L.A. (1983) Acyclic Stereoselection. 16. High Diastereofacial Selectivity in Lewis Acid Mediated Additions of Enol Silanes to Chiral Aldehydes. *Journal of the American Chemical Society* 105, 1667-1668
336. Irudayanathan, F.J. et al. (2021) Divining Deamidation and Isomerization in Therapeutic Proteins: Effect of Neighboring Residue. *bioRxiv* 2021.2007.2026.453885
337. Edler, E., Schulze, E., and Stein, M. (2017) Membrane Localization and Dynamics of Geranylgeranylated Rab5 Hypervariable Region. *Biochim Biophys Acta Biomembr* 1859, 1335-1349
338. Münzberg, E. (2019) *Of Proteins and Lipids : A Molecular Dynamics Study of Membrane-Bound Rab5*
339. Stenmark, H. (2009) Rab Gtpases as Coordinators of Vesicle Traffic. *Nature Reviews Molecular Cell Biology* 10, 513-525
340. Chavrier, P. et al. (1991) Hypervariable C-Termmal Domain of Rab Proteins Acts as a Targeting Signal. *Nature* 353, 769-772
341. Korn, E.D. and Wright, P.L. (1973) Macromolecular Composition of an Amoeba Plasma Membrane. *Journal of Biological Chemistry* 248, 439-447
342. Sampaio, J.L. et al. (2011) Membrane Lipidome of an Epithelial Cell Line. *Proceedings of the National Academy of Sciences* 108, 1903-1907
343. Petiot, A. et al. (2003) Pi3p Signaling Regulates Receptor Sorting but Not Transport in the Endosomal Pathway. *Journal of Cell Biology* 162, 971-979
344. Noda, T. et al. (2010) Regulation of Membrane Biogenesis in Autophagy Via Pi3p Dynamics. *Seminars in cell & developmental biology* Elsevier, 671-676
345. Marrink, S.J. et al. (2007) The Martini Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B* 111, 7812-7824
346. de Jong, D.H., Lopez, C.A., and Marrink, S.J. (2013) Molecular View on Protein Sorting into Liquid-Ordered Membrane Domains Mediated by Gangliosides and Lipid Anchors. *Faraday discussions* 161, 347-363
347. Monticelli, L. et al. (2008) The Martini Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation* 4, 819-834
348. de Jong, D.H. et al. (2013) Improved Parameters for the Martini Coarse-Grained Protein Force Field. *Journal of Chemical Theory and Computation* 9, 687-697
349. Wassenaar, T.A. et al. (2015) Computational Lipidomics with Insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations. *Journal of Chemical Theory and Computation* 11, 2144-2155
350. Bussi, G., Donadio, D., and Parrinello, M. (2007) Canonical Sampling through Velocity Rescaling. *The Journal of Chemical Physics* 126, 014101
351. Haug, E.J., Arora, J.S., and Matsui, K. (1976) A Steepest-Descent Method for Optimization of Mechanical Systems. *Journal of Optimization Theory and Applications* 19, 401–424

352. Hooft, R.W.W., van Eijck, B.P., and Kroon, J. (1992) An Adaptive Umbrella Sampling Procedure in Conformational Analysis Using Molecular Dynamics and Its Application to Glycol. *The Journal of Chemical Physics* 97, 6690-6694
353. Prinetti, A. et al. (2009) Glycosphingolipid Behaviour in Complex Membranes. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1788, 184-193
354. van den Bogaart, G. et al. (2011) Membrane Protein Sequestering by Ionic Protein–Lipid Interactions. *Nature* 479, 552-555
355. Feng, Z. et al. (2019) Recruitment of Vps34 Pi3k and Enrichment of Pi3p Phosphoinositide in the Viral Replication Compartment Is Crucial for Replication of a Positive-Strand Rna Virus. *PLOS pathogens* 15, e1007530-e1007530
356. Lee, H.-H. et al. (2018) Mixed Self-Assembled Monolayers with Terminal Deuterated Anchors: Characterization and Probing of Model Lipid Membrane Formation. *The Journal of Physical Chemistry B* 122, 8201-8210
357. Bigelow, W.C., Pickett, D.L., and Zisman, W.A. (1946) Oleophobic Monolayers: I. Films Adsorbed from Solution in Non-Polar Liquids. *Journal of Colloid Science* 1, 513-538
358. Ulman, A. (1996) Formation and Structure of Self-Assembled Monolayers. *Chemical Reviews* 96, 1533-1554
359. Knoll, W. et al. (2000) Functional Tethered Lipid Bilayers. *Reviews in Molecular Biotechnology* 74, 137-158
360. Naumann, R. et al. (2003) Tethered Lipid Bilayers on Ultraflat Gold Surfaces. *Langmuir* 19, 5435-5443
361. Giess, F. et al. (2004) The Protein-Tethered Lipid Bilayer: A Novel Mimic of the Biological Membrane. *Biophysical Journal* 87, 3213-3220
362. Castellana, E.T. and Cremer, P.S. (2006) Solid Supported Lipid Bilayers: From Biophysical Studies to Sensor Design. *Surface Science Reports* 61, 429-444
363. Andersson, J. and Köper, I. (2016) Tethered and Polymer Supported Bilayer Lipid Membranes: Structure and Function. *Membranes* 6, 30
364. Coutable, A. et al. (2014) Preparation of Tethered-Lipid Bilayers on Gold Surfaces for the Incorporation of Integral Membrane Proteins Synthesized by Cell-Free Expression. *Langmuir* 30, 3132-3141
365. Lamken, P. et al. (2005) Functional Cartography of the Ectodomain of the Type I Interferon Receptor Subunit Ifnar1. *The Journal of Molecular Biology* 350, 476-488
366. Gavutis, M. et al. (2005) Lateral Ligand-Receptor Interactions on Membranes Probed by Simultaneous Fluorescence-Interference Detection. *Biophysical Journal* 88, 4289-4302
367. Lata, S., Gavutis, M., and Piehler, J. (2006) Monitoring the Dynamics of Ligand−Receptor Complexes on Model Membranes. *Journal of the American Chemical Society* 128, 6-7
368. Zharnikov, M. and Grunze, M. (2001) Spectroscopic Characterization of Thiol-Derived Self-Assembling Monolayers. *Journal of Physics: Condensed Matter* 13, 11333
369. Gavutis, M., Lata, S., and Piehler, J. (2006) Probing 2-Dimensional Protein–Protein Interactions on Model Membranes. *Nature protocols* 1, 2091-2103
370. Benesch, J. et al. (2001) Protein Adsorption to Oligo(Ethylene Glycol) Self-Assembled Monolayers: Experiments with Fibrinogen, Heparinized Plasma, and Serum. *Journal of Biomaterials Science, Polymer Edition* 12, 581-597
371. Pale-Grosdemange, C. et al. (1991) Formation of Self-Assembled Monolayers by Chemisorption of Derivatives of Oligo(Ethylene Glycol) of Structure Hs(Ch2)11(Och2ch2)Moh on Gold. *Journal of the American Chemical Society* 113, 12-20
372. Harder, P. et al. (1998) Molecular Conformation in Oligo(Ethylene Glycol)-Terminated Self-Assembled Monolayers on Gold and Silver Surfaces Determines Their Ability to Resist Protein Adsorption. *The Journal of Physical Chemistry B* 102, 426-436
373. Lee, H. et al. (2009) A Coarse-Grained Model for Polyethylene Oxide and Polyethylene Glycol: Conformation and Hydrodynamics. *The Journal of Physical Chemistry B* 113, 13186-13194

374. Hautman, J. et al. (1991) Molecular Dynamics Investigations of Self-Assembled Monolayers. *Journal of the Chemical Society, Faraday Transactions* 87, 2031-2037
375. Hautman, J. and Klein, M.L. (1989) Simulation of a Monolayer of Alkyl Thiol Chains. *The Journal of Chemical Physics* 91, 4994-5001
376. Vemparala, S. et al. (2004) Large-Scale Molecular Dynamics Simulations of Alkanethiol Self-Assembled Monolayers. *The Journal of Chemical Physics* 121, 4323-4330
377. Esteban-Martín, S. and Salgado, J. (2007) Self-Assembling of Peptide/Membrane Complexes by Atomistic Molecular Dynamics Simulations. *Biophysical Journal* 92, 903-912
378. Mori, T. et al. (2016) Molecular Dynamics Simulations of Biological Membranes and Membrane Proteins Using Enhanced Conformational Sampling Algorithms. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1858, 1635-1651
379. Bhadra, P. and Siu, S.W.I. (2017) Comparison of Biomolecular Force Fields for Alkanethiol Self-Assembled Monolayer Simulations. *The Journal of Physical Chemistry C* 121, 26340-26349
380. Marrink, S.J. and Tieleman, D.P. (2013) Perspective on the Martini Model. *Chemical Society Reviews* 42, 6801-6822
381. Lee, H. and Pastor, R.W. (2011) Coarse-Grained Model for Pegylated Lipids: Effect of Pegylation on the Size and Shape of Self-Assembled Structures. *The Journal of Physical Chemistry B* 115, 7830-7837
382. Kyrychenko, A. et al. (2011) Preparation, Structure, and a Coarse-Grained Molecular Dynamics Model for Dodecanethiol-Stabilized Gold Nanoparticles. *Computational and Theoretical Chemistry* 977, 34-39
383. Li, Y. et al. (2007) Contact Angle of Water on Polystyrene Thin Films: Effects of Co2 Environment and Film Thickness. *Langmuir* 23, 9785-9793
384. Good, R.J. and Kotsidas, E.D. (1978) The Contact Angle of Water on Polystyrene: A Study of the Cause of Hysteresis. *Journal of Colloid and Interface Science* 66, 360-362
385. Smith, T. (1980) The Hydrophilic Nature of a Clean Gold Surface. *Journal of Colloid and Interface Science* 75, 51-55
386. Islam, N. et al. (2014) Effects of Composition of Oligo(Ethylene Glycol)-Based Mixed Monolayers on Peptide Grafting and Human Immunoglobulin Detection. *The Journal of Physical Chemistry C* 118, 5361-5373
387. Dettre, R.H. and Johnson, R.E. (1964) Contact Angle Hysteresis. *Contact Angle, Wettability, and Adhesion* Vol. 43. Washington D.C, Aerican Chemical Societiy. 136-144
388. Yesylevskyy, S.O. et al. (2010) Polarizable Water Model for the Coarse-Grained Martini Force Field. *PLOS Computational Biology* 6, e1000810
389. de Jong, D.H. et al. (2016) Martini Straight: Boosting Performance Using a Shorter Cutoff and Gpus. *Computer Physics Communications* 199, 1-7
390. Chialvo, A.A. and Debenedetti, P.G. (1992) An Automated Verlet Neighbor List Algorithm with a Multiple Time-Step Approach for the Simulation of Large Systems. *Computer Physics Communications* 70, 467-477
391. Tironi, I.G. et al. (1995) A Generalized Reaction Field Method for Molecular Dynamics Simulations. *The Journal of Chemical Physics* 102, 5451-5459
392. Wassenaar, T.A. et al. (2014) Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *Journal of Chemical Theory and Computation* 10, 676-690
393. Pedregosa, F. et al. (2011) Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830
394. Oelmeier, S.A., Dismer, F., and Hubbuch, J. (2012) Molecular Dynamics Simulations on Aqueous Two-Phase Systems - Single Peg-Molecules in Solution. *BMC Biophysics* 5, 14
395. Oostenbrink, C. et al. (2004) A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The Gromos Force-Field Parameter Sets 53a5 and 53a6. *The Journal of Computational Chemistry* 25, 1656-1676

396. Ulman, A., Eilers, J.E., and Tillman, N. (1989) Packing and Molecular Orientation of Alkanethiol Monolayers on Gold Surfaces. *Langmuir* 5, 1147-1152
397. Vericat, C. et al. (2010) Self-Assembled Monolayers of Thiols and Dithiols on Gold: New Challenges for a Well-Known System. *Chemical Society Reviews* 39, 1805-1834
398. Porter, M.D. et al. (1987) Spontaneously Organized Molecular Assemblies. 4. Structural Characterization of N-Alkyl Thiol Monolayers on Gold by Optical Ellipsometry, Infrared Spectroscopy, and Electrochemistry. *Journal of the American Chemical Society* 109, 3559-3568
399. Nuzzo, R.G., Zegarski, B.R., and Dubois, L.H. (1987) Fundamental Studies of the Chemisorption of Organosulfur Compounds on Gold(111). Implications for Molecular Self-Assembly on Gold Surfaces. *Journal of the American Chemical Society* 109, 733-740
400. Schwartz, D.K. (2001) Mechanisms and Kinetics of Self-Assembled Monolayer Formation. *Annual Review of Physical Chemistry* 52, 107-137
401. Alessi, M.L. et al. (2005) Helical and Coil Conformations of Poly(Ethylene Glycol) in Isobutyric Acid and Water. *Macromolecules* 38, 9333-9340
402. Azri, A. et al. (2012) Polyethylene Glycol Aggregates in Water Formed through Hydrophobic Helical Structures. *The Journal of Colloid and Interface Science* 379, 14-19
403. Wang, R.L.C., Jürgen Kreuzer, H., and Grunze, M. (2000) The Interaction of Oligo(Ethylene Oxide) with Water: A Quantum Mechanical Study. *Physical Chemistry Chemical Physics* 2, 3613-3622
404. Pichot, R., Watson, R.L., and Norton, I.T. (2013) Phospholipids at the Interface: Current Trends and Challenges. *International Journal of Molecular Sciences* 14,
405. Civjan, N.R. et al. (2003) Direct Solubilization of Heterologously Expressed Membrane Proteins by Incorporation into Nanoscale Lipid Bilayers. *Biotechniques* 35, 556-563
406. Bayburt, T.H., Grinkova, Y.V., and Sligar, S.G. (2006) Assembly of Single Bacteriorhodopsin Trimers in Bilayer Nanodiscs. *Archives of biochemistry and biophysics* 450, 215-222
407. Ritchie, T. et al. (2009) Reconstitution of Membrane Proteins in Phospholipid Bilayer Nanodiscs. *Methods in Enzymology* 464, 211-231
408. MacDonald, R.C. et al. (1991) Small-Volume Extrusion Apparatus for Preparation of Large, Unilamellar Vesicles. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 1061, 297-303
409. Hope, M. et al. (1985) Production of Large Unilamellar Vesicles by a Rapid Extrusion Procedure. Characterization of Size Distribution, Trapped Volume and Ability to Maintain a Membrane Potential. *Biochimica et Biophysica Acta (BBA)-Biomembranes* 812, 55-65
410. Boheim, G. (1974) Statistical Analysis of Alamethicin Channels in Black Lipid Membranes. *The Journal of Membrane Biology* 19, 277-303
411. Eisenberg, M., Hall, J.E., and Mead, C. (1973) The Nature of the Voltage-Dependent Conductance Induced by Alamethicin in Black Lipid Membranes. *The Journal of Membrane Biology* 14, 143-176
412. Glazier, R. and Salaita, K. (2017) Supported Lipid Bilayer Platforms to Probe Cell Mechanobiology. *Biochimica et Biophysica Acta (BBA) - Biomembranes* 1859, 1465-1482
413. Pace, H. et al. (2015) Preserved Transmembrane Protein Mobility in Polymer-Supported Lipid Bilayers Derived from Cell Membranes. *Analytical Chemistry* 87, 9194-9203
414. Mio, K. and Sato, C. (2018) Lipid Environment of Membrane Proteins in Cryo-Em Based Structural Analysis. *Biophysical Reviews* 10, 307-316
415. Günsel, U. and Hagn, F. (2022) Lipid Nanodiscs for High-Resolution Nmr Studies of Membrane Proteins. *Chemical Reviews* 122, 9395–9421
416. Bengtsen, T. et al. (2020) Structure and Dynamics of a Nanodisc by Integrating Nmr, Saxs and Sans Experiments with Molecular Dynamics Simulations. *eLife* 9, e56518
417. Li, W. and Haines, T.H. (1986) Uniform Preparations of Large Unilamellar Vesicles Containing Anionic Lipids. *Biochemistry* 25, 7477-7483
418. Da Costa, G. et al. (2007) Nmr of Molecules Interacting with Lipids in Small Unilamellar Vesicles. *European Biophysics Journal* 36, 933-942

419. Hema Sagar, G., Tiwari, M.D., and Bellare, J.R. (2010) Flow Cytometry as a Novel Tool to Evaluate and Separate Vesicles Using Characteristic Scatter Signatures. *The Journal of Physical Chemistry B* 114, 10010-10016
420. Winterhalter, M. (2000) Black Lipid Membranes. *Current Opinion in Colloid & Interface Science* 5, 250-255
421. Blicher, A. and Heimburg, T. (2013) Voltage-Gated Lipid Ion Channels. *PLOS ONE* 8, e65707
422. Benz, M. et al. (2004) Correlation of Afm and Sfa Measurements Concerning the Stability of Supported Lipid Bilayers. *Biophysical Journal* 86, 870-879
423. Favero, G. et al. (2002) Membrane Supported Bilayer Lipid Membranes Array: Preparation, Stability and Ion-Channel Insertion. *Analytica Chimica Acta* 460, 23-34
424. Alessandrini, A. and Facci, P. (2014) Phase Transitions in Supported Lipid Bilayers Studied by Afm. *Soft Matter* 10, 7145-7164
425. Bin, X. et al. (2005) Electrochemical and Pm-Irras Studies of the Effect of the Static Electric Field on the Structure of the Dmpc Bilayer Supported at a Au(111) Electrode Surface. *Langmuir* 21, 330-347
426. Wang, K.F., Nagarajan, R., and Camesano, T.A. (2015) Differentiating Antimicrobial Peptides Interacting with Lipid Bilayer: Molecular Signatures Derived from Quartz Crystal Microbalance with Dissipation Monitoring. *Biophysical Chemistry* 196, 53-67
427. Heinrich, F. et al. (2009) A New Lipid Anchor for Sparsely Tethered Bilayer Lipid Membranes. *Langmuir* 25, 4219-4229
428. Budvytyte, R. et al. (2013) Structure and Properties of Tethered Bilayer Lipid Membranes with Unsaturated Anchor Molecules. *Langmuir* 29, 8645-8656
429. Mar, W. and Klein, M.L. (1994) Molecular Dynamics Study of the Self-Assembled Monolayer Composed of S (Ch2) 14ch3 Molecules Using an All-Atoms Model. *Langmuir* 10, 188-196
430. Ahn, Y. et al. (2011) Molecular Dynamics Study of the Formation of a Self-Assembled Monolayer on Gold. *The Journal of Physical Chemistry C* 115, 10668-10674
431. Tupper, K.J. and Brenner, D.W. (1994) Molecular Dynamics Simulations of Friction in Self-Assembled Monolayers. *Thin Solid Films* 253, 185-189
432. Sung, I.-H. and Kim, D.-E. (2005) Molecular Dynamics Simulation Study of the Nano-Wear Characteristics of Alkanethiol Self-Assembled Monolayers. *Applied Physics A* 81, 109-114
433. Tonda-Turo, C., Carmagnola, I., and Ciardelli, G. (2018) Quartz Crystal Microbalance with Dissipation Monitoring: A Powerful Method to Predict the in Vivo Behavior of Bioengineered Surfaces. *Frontiers in Bioengineering and Biotechnology* 6,
434. Qi, Y. et al. (2015) Charmm-Gui Martini Maker for Coarse-Grained Simulations with the Martini Force Field. *Journal of Chemical Theory and Computation* 11, 4486-4494
435. Foresman, J.B. et al. (1996) Solvent Effects. 5. Influence of Cavity Shape, Truncation of Electrostatics, and Electron Correlation on Ab Initio Reaction Field Calculations. *The Journal of Physical Chemistry* 100, 16098-16104
436. Peitzsch, R.M. and McLaughlin, S. (1993) Binding of Acylated Peptides and Fatty Acids to Phospholipid Vesicles: Pertinence to Myristoylated Proteins. *Biochemistry* 32, 10436-10443
437. Silvius, J.R. and l'Heureux, F. (1994) Fluorometric Evaluation of the Affinities of Isoprenylated Peptides for Lipid Bilayers. *Biochemistry* 33, 3014-3022
438. Genheden, S. (2017) Solvation Free Energies and Partition Coefficients with the Coarse-Grained and Hybrid All-Atom/Coarse-Grained Martini Models. *Journal of Computer-Aided Molecular Design* 31, 867-876
439. Taddese, T. and Carbone, P. (2017) Effect of Chain Length on the Partition Properties of Poly(Ethylene Oxide): Comparison between Martini Coarse-Grained and Atomistic Models. *The Journal of Physical Chemistry B* 121, 1601-1609

440. Alessandri, R. et al. (2019) Pitfalls of the Martini Model. *Journal of Chemical Theory and Computation* 15, 5448-5460
441. Cho, N.-J. et al. (2010) Quartz Crystal Microbalance with Dissipation Monitoring of Supported Lipid Bilayers on Various Substrates. *Nature Protocols* 5, 1096-1106

APPENDIX

A NOROVIRUS RECOGNITION OF BILE ACIDS

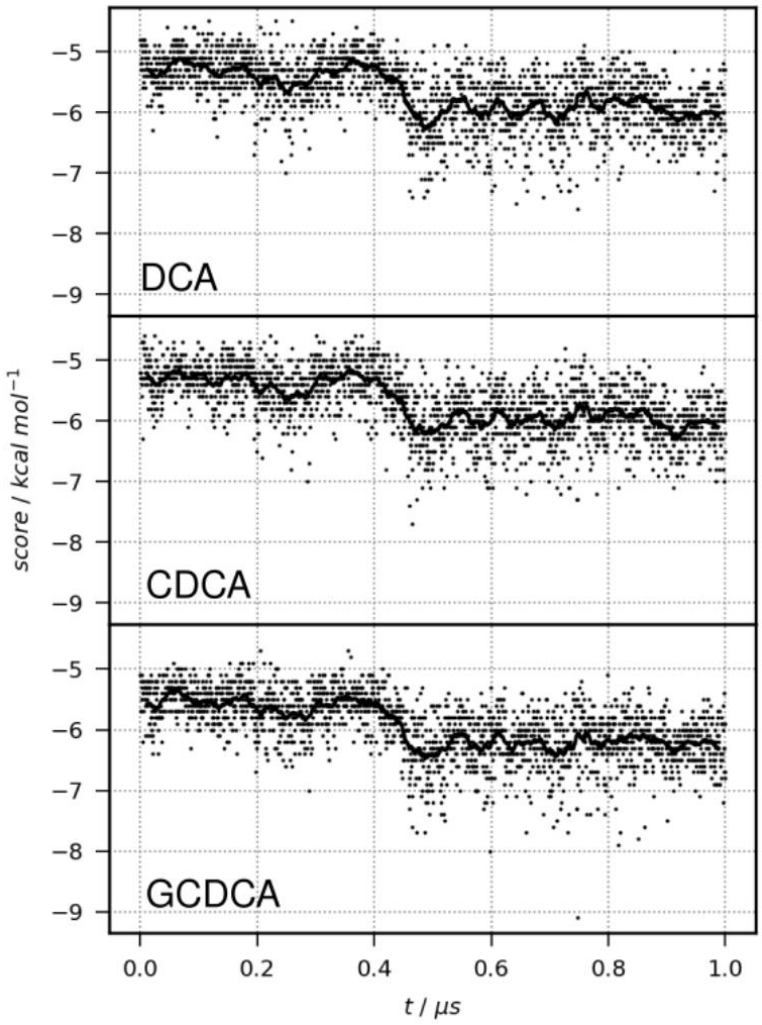

Figure A.1: Docking scores of DCA, CDCA, and GCDCA to 2.000 snapshots of the 1 μs MD trajectory of GII.4 Saga P-dimers. The solid line represents a moving average of each 40 points.

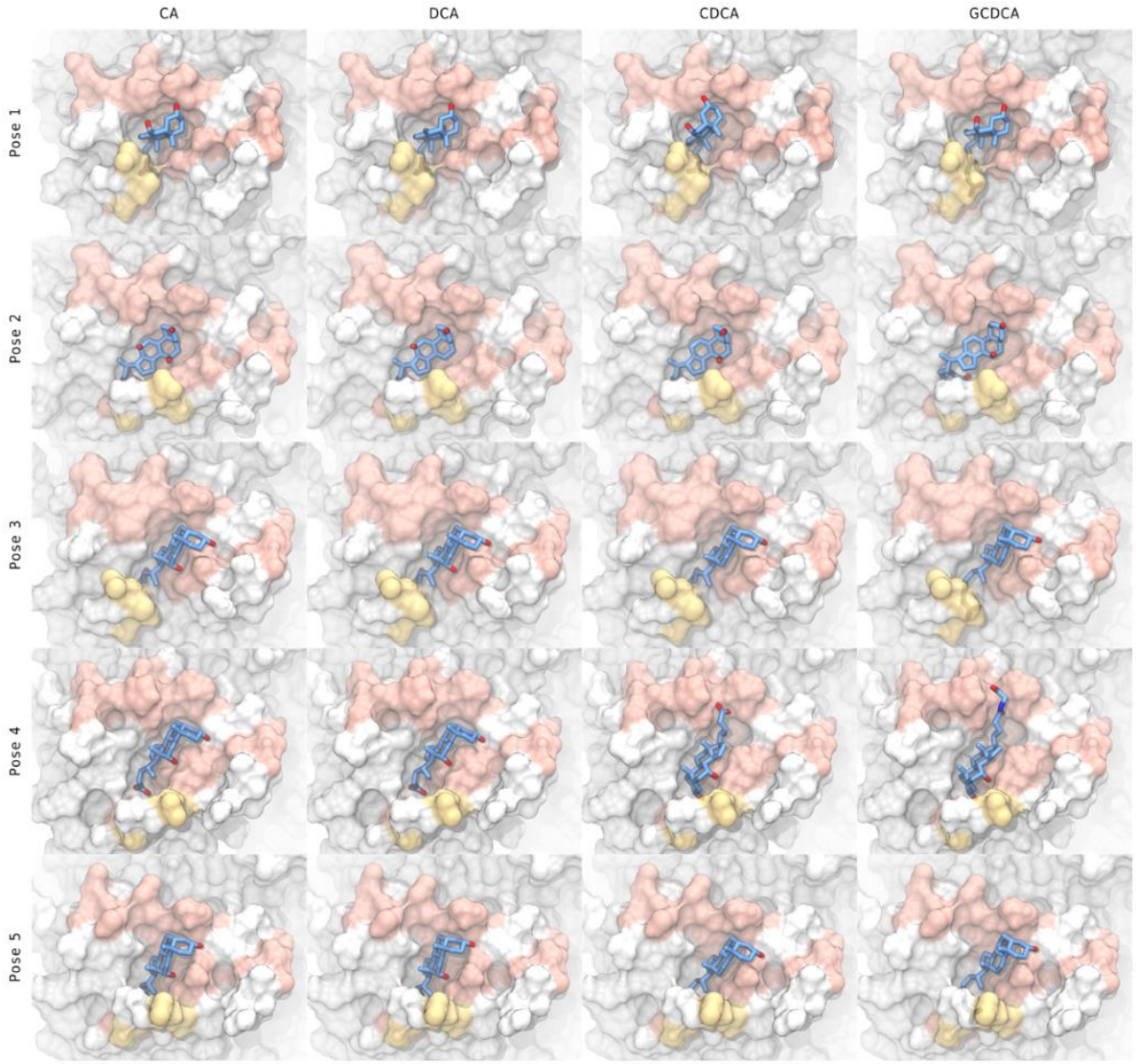

Figure A.2: Snapshots of the top five scoring protein-ligand docking poses for each of the 4 bile acids (CA, DCA, CDCA, GCDCA). Protein is shown as solvent-accessible surface. Ligands are drawn in blue licorice representations with the hydrogens omitted for clarity. Colored surface patches denote amino acids with significant CSPs (pink for 1 H 15 N and yellow for 1 H 13 C CSPs). The surface is slightly translucent so the ligand behind the C-terminus becomes visible.

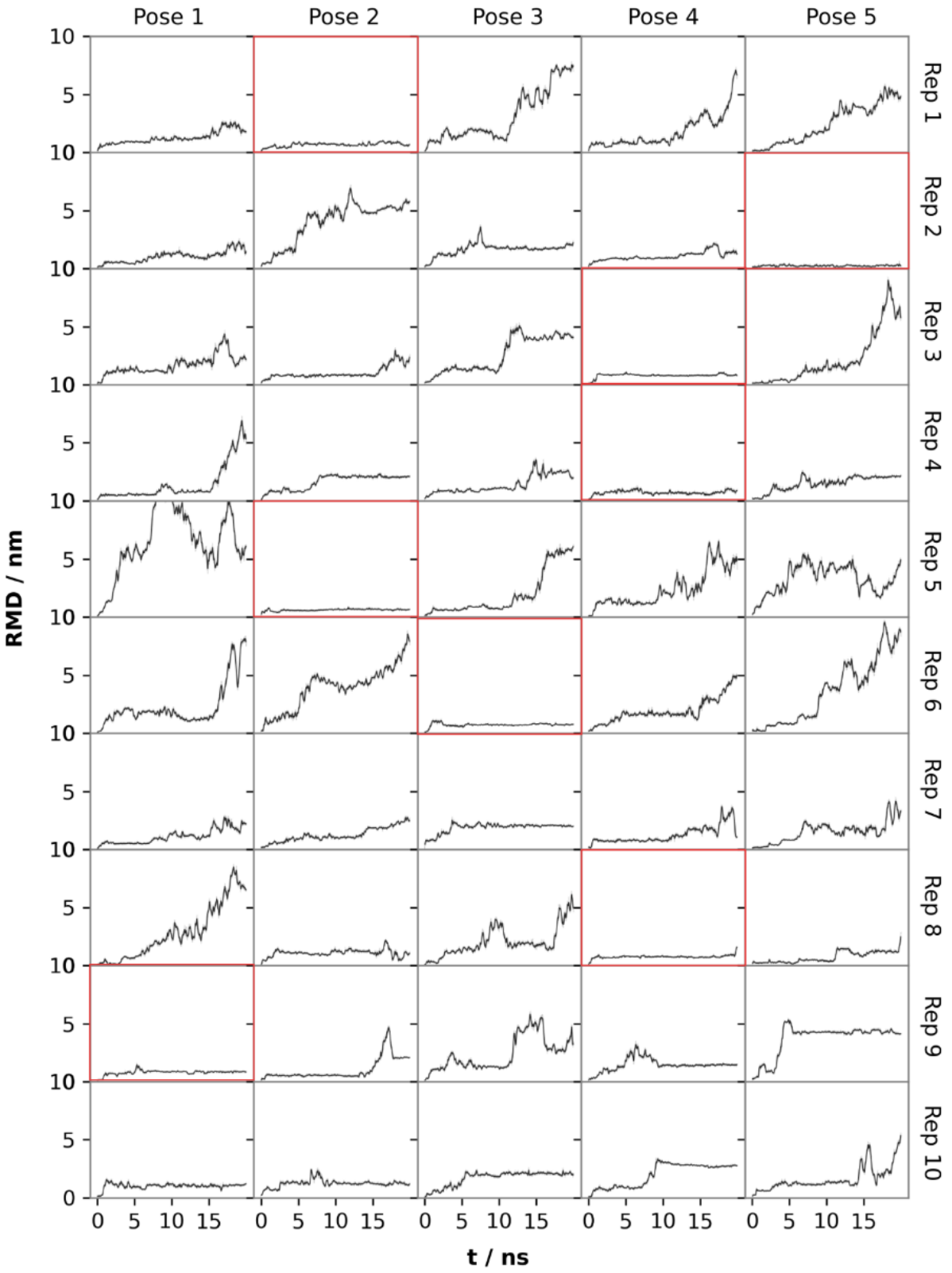

Figure A.3: CA ligand RMSDs in nm over simulation time in ns of all the ten replicate simulations for the five initial geometries (docked complexes). The eight trajectories with lowest averages RMSD (last 10 ns) are framed with a red border. The RMSD of only the ligand is considered, with the trajectory being previously fitted to the protein backbone atoms.

B SELECTIVE CLEAVAGE OF LINEAR POLY-UBIQUITIN

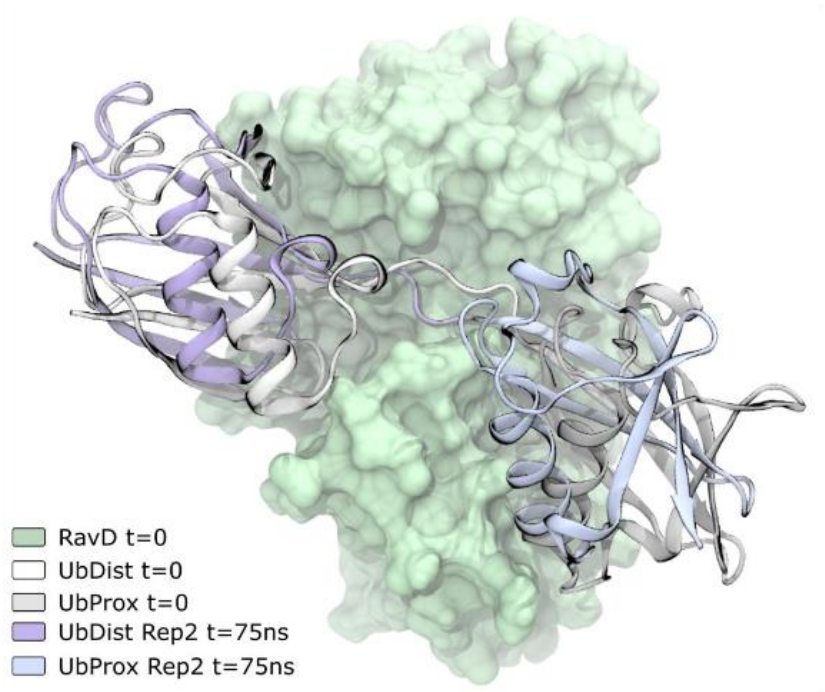

Figure B.1: Snapshots of L. pneumophila RavD-DiUb with high RMSD from crystal structure. Position and orientation of DiUb relative to the initial model.

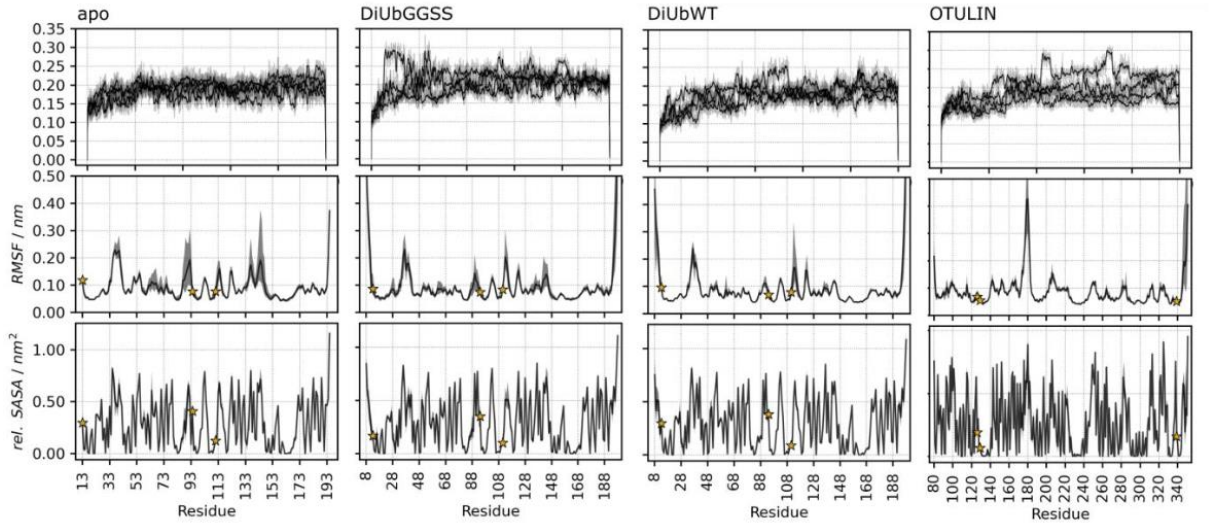

Figure B.2: RMSD / nm, RMSF /nm and relative SASA /nm2 of the simulated DUBs RavD and OTULIN in absence and in complex with

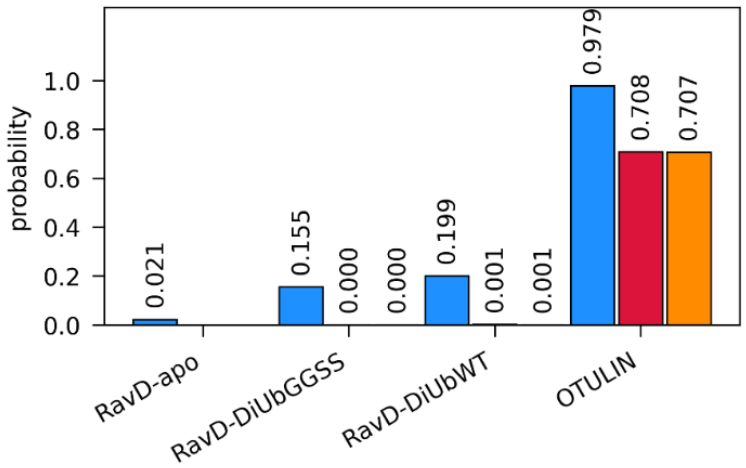

Figure B.3: Structural criteria for a catalytic competent conformation. Probability of occurrence of close contacts (<0.6 nm) between

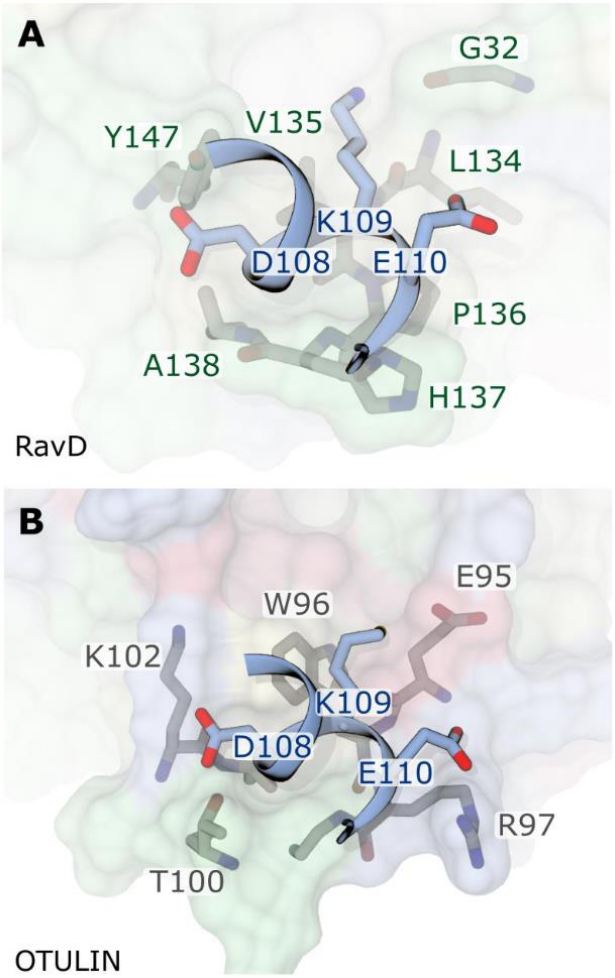

Figure B.4: Structural complementarity of proximal Ub recognition to the Asp108 (D108) patch in RavD (A) and OTULIN (B).

C GLYCAN CONFORMATION DEPENCY ON GLYCOSYLATION SITE

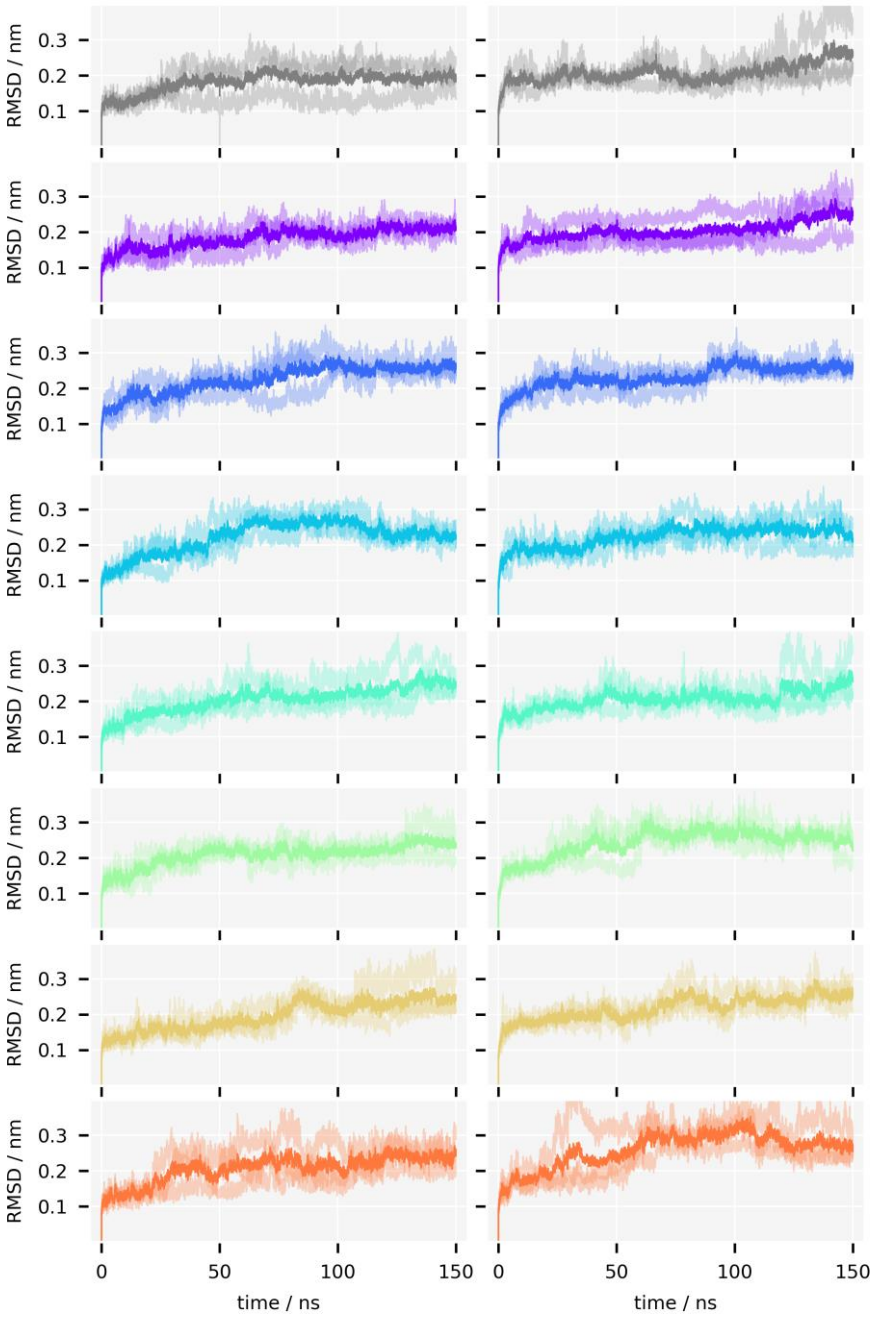

Figure C.1: RMSD over simulation for all simulations. From top to bottom: ngEPO, ASN24, ASN38, ASN83, ASN2438, ASN2483, ASN3883, ASN243883. Left column are only N-glycosylated, right column also carry O-glycosylation at Ser126.

Code 1: UMAP embedding and HDBSCAN clustering

```
neigh_frac = 0.001 # fraction of points considered as neighbors for UMAP
neighs = int(c_all_cossin.shape[0]*neigh_frac) # absolute number of neighboring points

reducer = umap.UMAP(metric='euclidean', n_neighbors=neighs,
                    min_dist=1, n_components=2, random_state=29)
```

D SITE SPECIFICITY OF ASPARAGINE DEAMIDATION

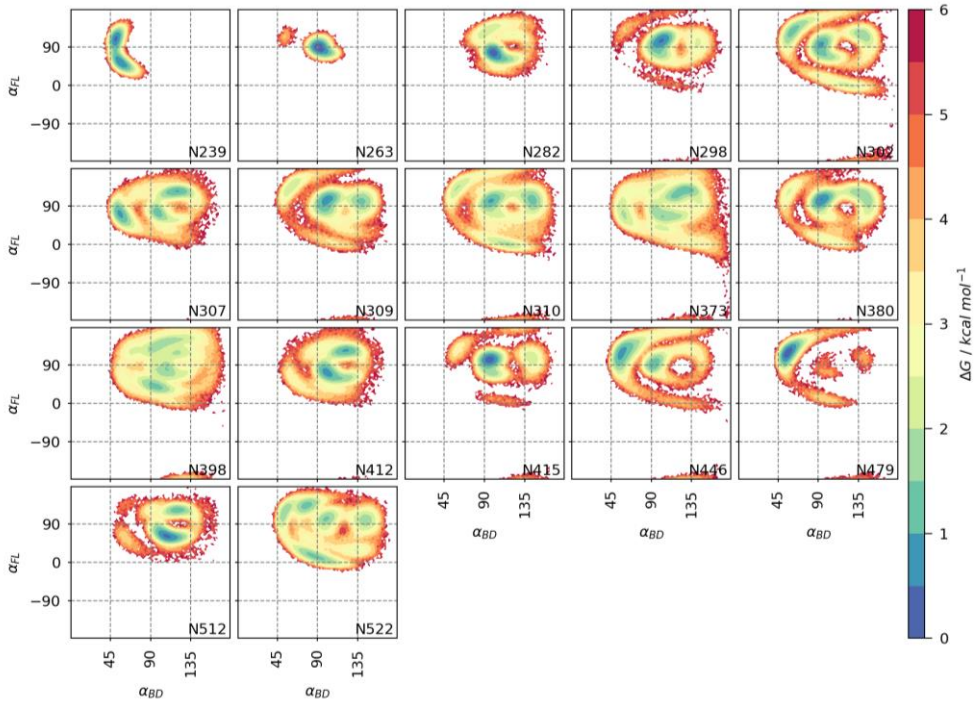

Figure D.1: Free energy maps of the Burgi-Dunitz and Flipping Lodge angles.

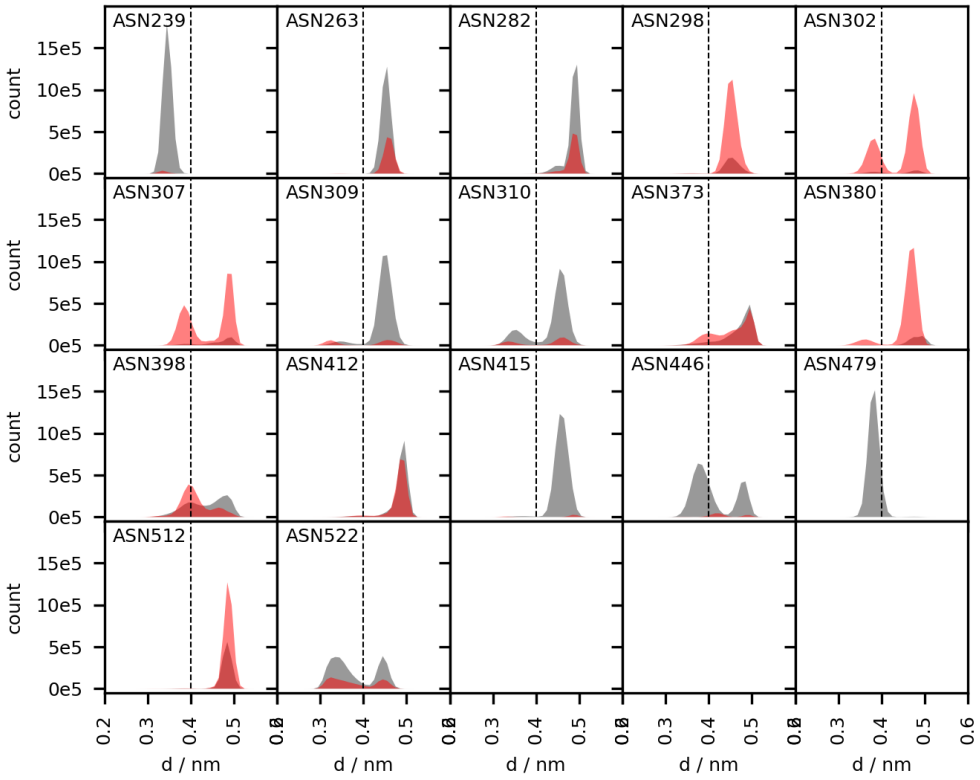

Figure D.2: Distributions of the attack distances in absence (gray) and presence (red) of a N+1 hydrogen backbone hydrogen bond.

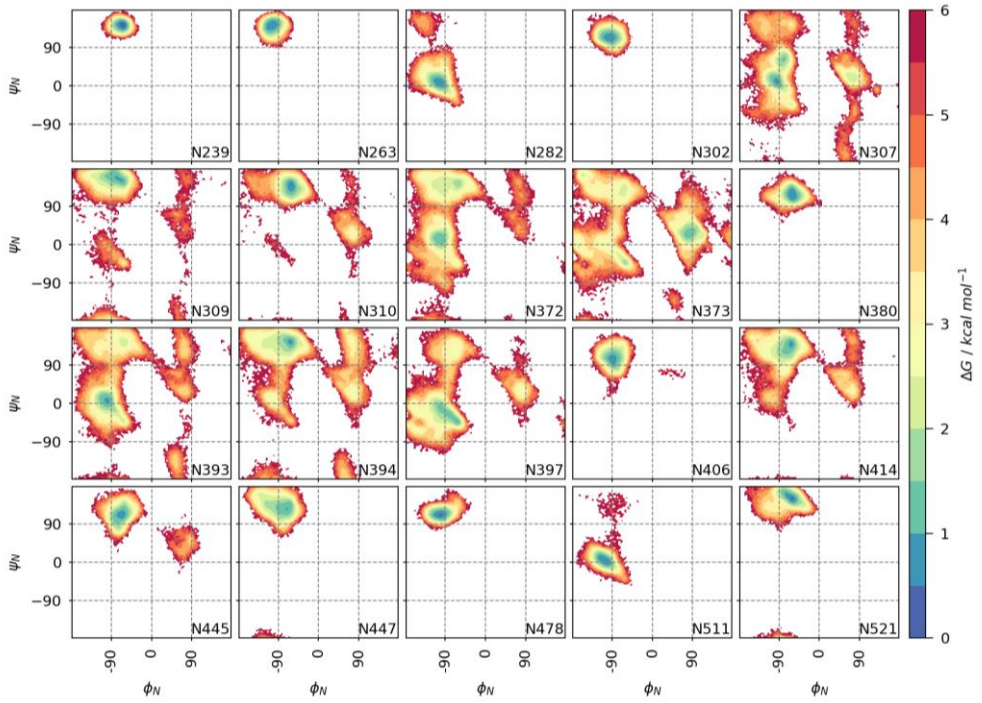

Figure D.3: Backbone torsion angle free energy landscapes of VA387 P-domains

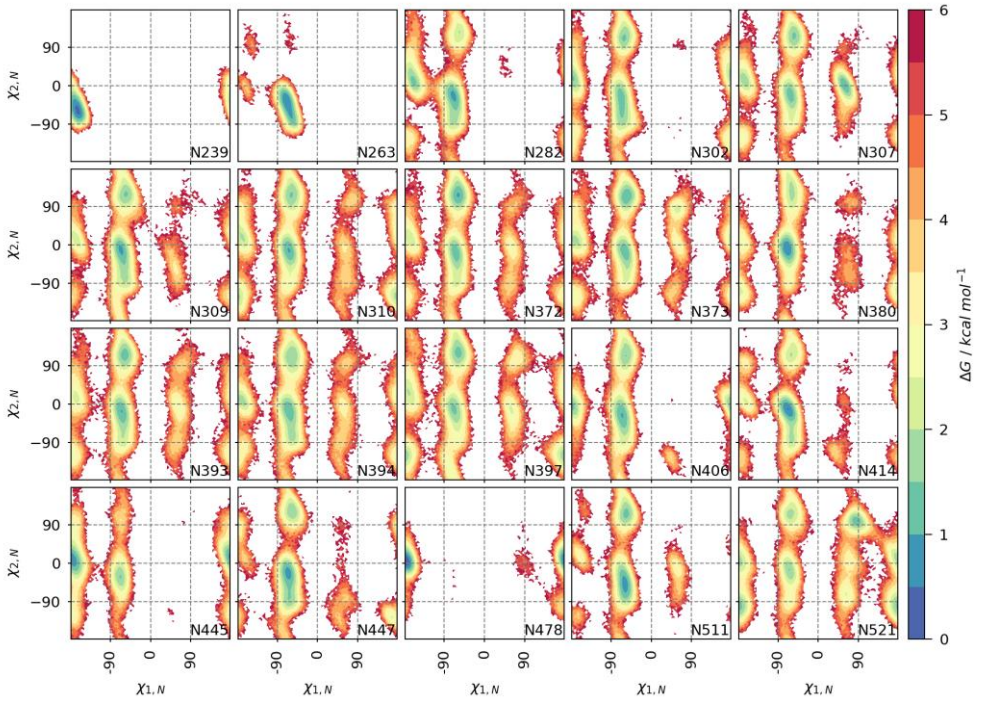

Figure D.4: Sidechain torsion angle angle free energy landscapes of VA387 P-Domains

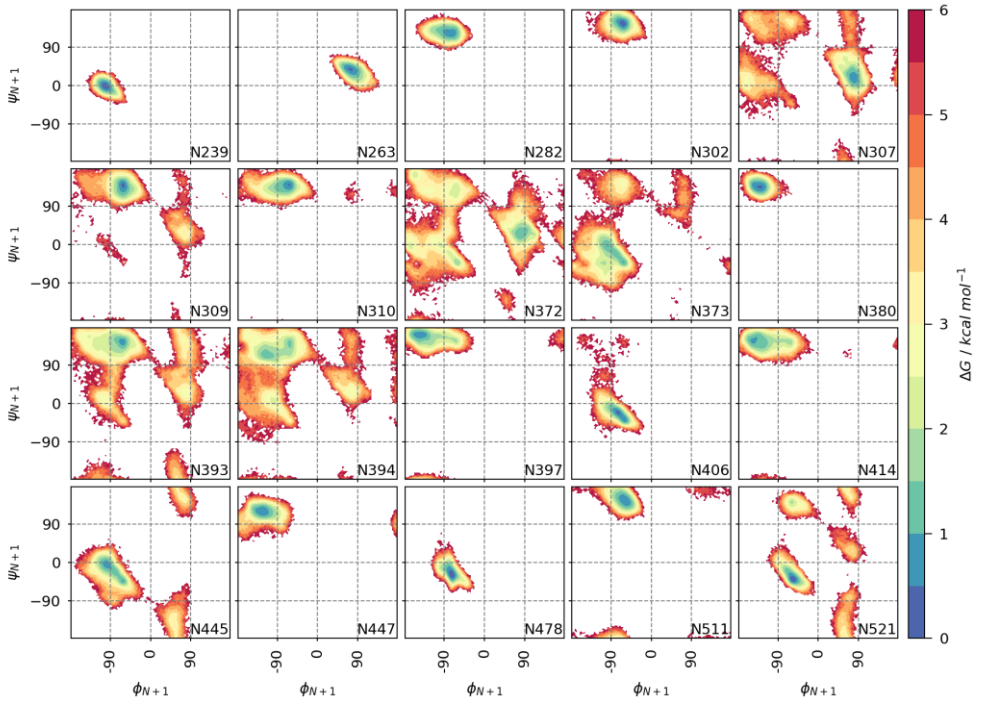

Figure D.5: N+1 Backbone torsion angle free energy landscapes of VA387 P-domains

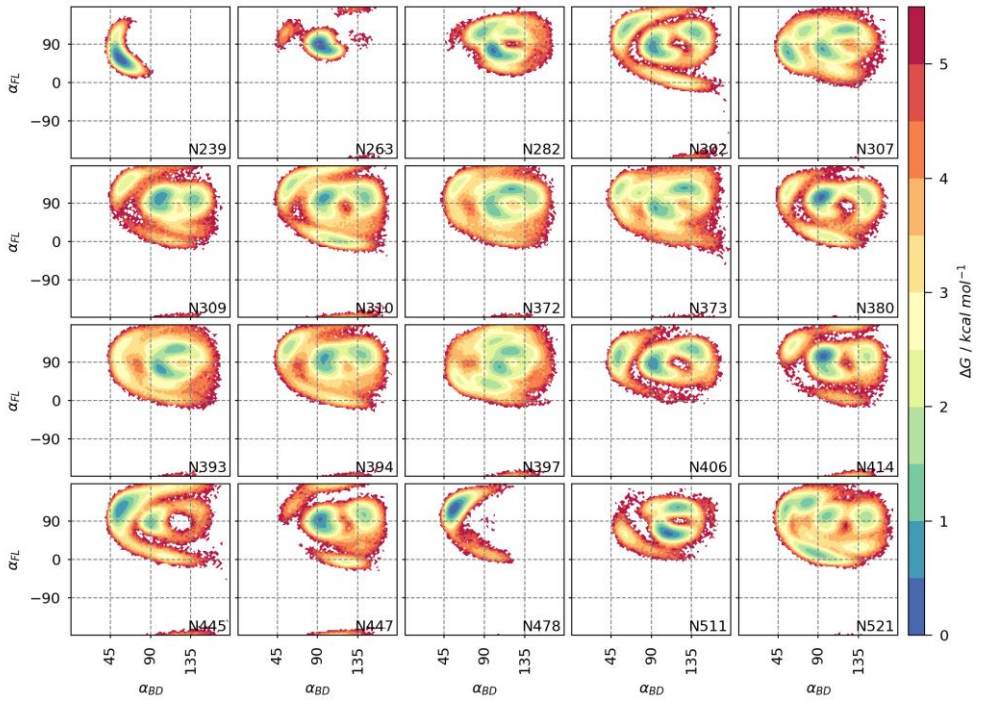

Figure D.6: Attack angle free energy landscapes of VA387 P-domains

E MONOLAYER PHASE SEGREGATION AND TRANSITION

Table E.1: Coarse-grain MD simulation parameters for the matrix compound

Group	Name	Tpye
thiopropyl	THIO	C5
n-butyl 1	C3B	C1
n-butyl 2	C2B	C1
n-butyl 3	C1B	C1
acetamido 1	ACB	P5
hydroxyethyl	PE6	SP2
Bond	**r_0 / nm**	**k_B / kJ mol^{-1} nm^{-2}**
THIO-C3B	0.47	1250
C3B-C2B	0.47	1250
C2B-C1B	0.47	1250
C1B-ACB	0.40	2500
ACB-PE6	0.37	5000
Angle	**θ_0 / °**	**k_B / kJ mol^{-1} rad^{-2}**
THIO-C3B-C2B	180	25
C3B-C2B-C1B	180	25
C2B-C1B-ACB	180	25
C1B-ACB-PE6	160	25

Table E.2: Further and additional coarse-grain parameters for the C8 anchor compound

Group	Name	Tpye
oxyethylen 1	PE6	SN0
oxyethylen 2	PE5	SN0
oxyethylen 3	PE4	SN0
oxyethylen 4	PE3	SN0
oxyethylen 5	PE2	SN0
oxyethylen 6	PE1	SN0
acetamido 2	ACA	P5
n-butyl 4	C1A	C1
n-butyl 5	C2A	C1

Bond	r_0 / nm	k_B / kJ mol^{-1} nm^{-2}
ACB-PE6	0.35	5000
PE6-PE5	0.33	17000
PE5-PE4	0.33	17000
PE4-PE3	0.33	17000
PE3-PE2	0.33	17000
PE2-PE1	0.33	17000
PE1-ACA	0.35	5000
ACA-C1A	0.40	2500
C1A-C2A	0.47	1250

Angle	θ_0 / °	k_B / kJ mol^{-1} rad^{-2}
C1B-ACB-PE6	150	25
ACB-PE6-PE5	140	12.5
PE6-PE5-PE4	130	50
PE5-PE4-PE3	130	50
PE4-PE3-PE2	130	50
PE3-PE2-PE1	130	50
PE2-PE1-ACA	140	12.5
PE1-ACA-C1A	150	25
ACA-C1A-C2A	180	25

Table E.3: Additional coarse-grain parameters for the C16 anchor compound

Group	Name	Type
n-butyl 6	C3A	C1
n-butyl 6	C4A	C1

Bond	r_0 / nm	k_B / kJ mol^{-1} nm^{-2}
C2A-C3A	0.47	1250
C3A-C4A	0.47	1250

Angle	θ_0 / °	k_B / kJ mol^{-1} rad^{-2}
C1A-C2A-C3A	180	25
C2A-C3A-C4A	180	25

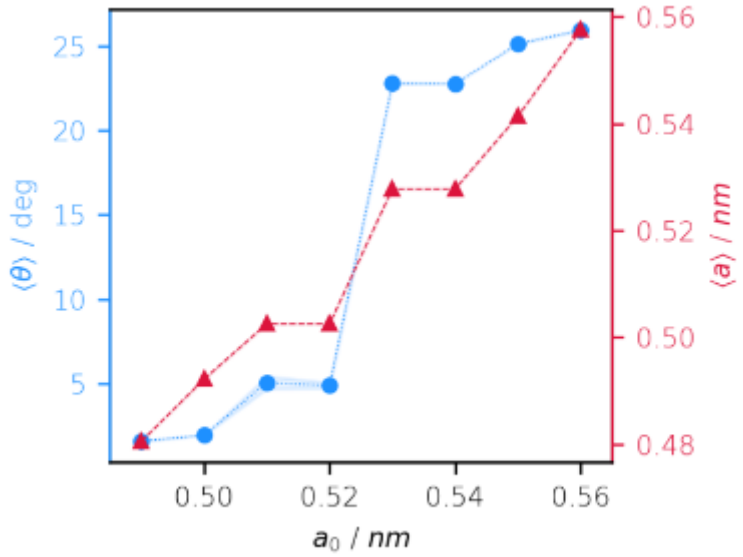

Figure E.1: Tilt angle (left axis, blue circles, dotted line) and lattice parameter (right axis, red triangles, dashed line) dependence on the initial packing. The here considered system is a pure anchor compound monolayer. The simulations were performed in three replicates. The standard deviation is plotted as a shaded area.

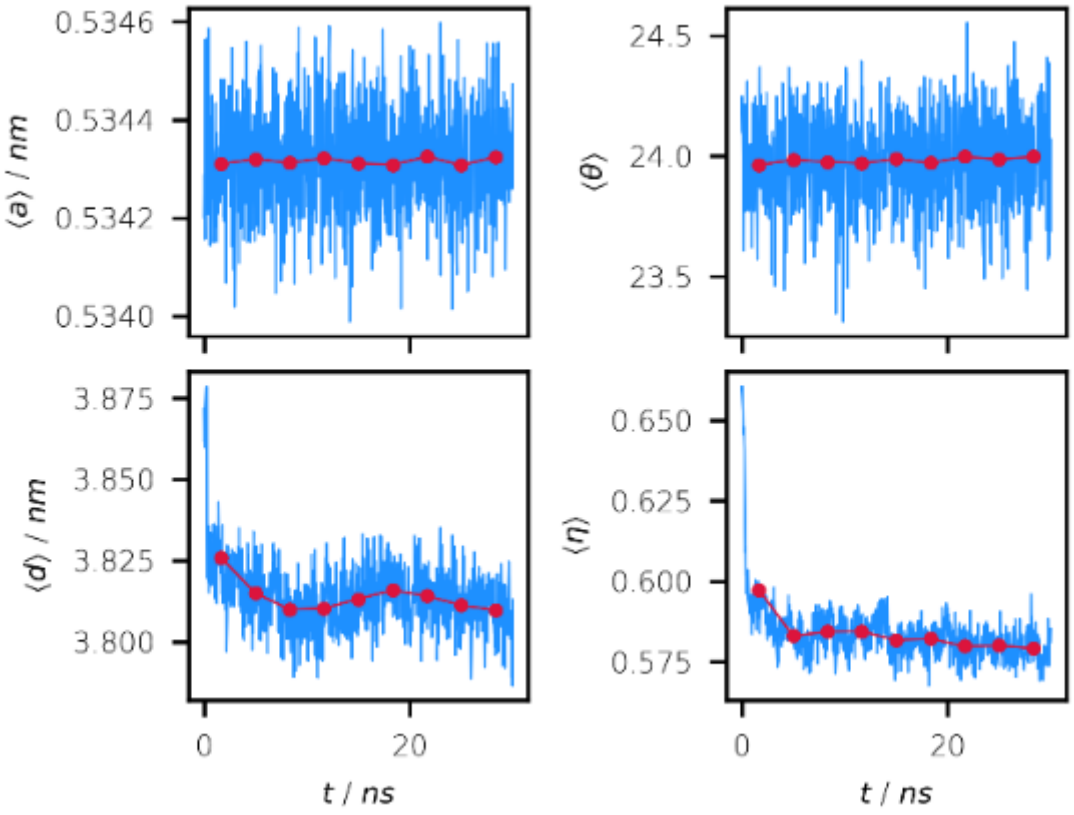

Figure E.2: Time evolution of coarse-grained simulation monolayer properties (blue) together with 10 block averages (red circles) during the production phase

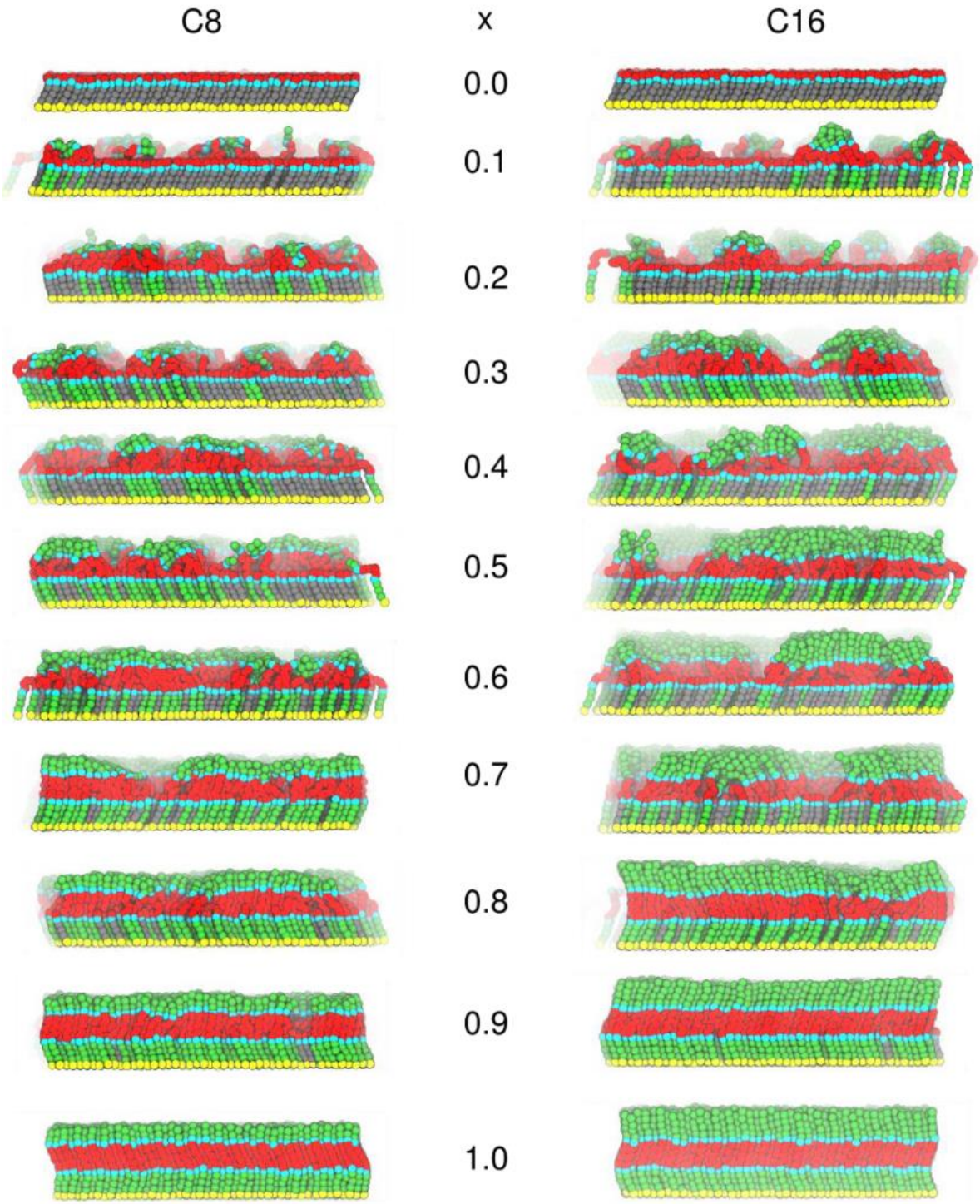

Figure E.3: Final frame snapshots of the coarse-grained production simulations of different SAMs with different molar fractions of anchoring compounds. Matrix compound alkyl chains are depicted in gray, anchoring alkyl chains in green. The OEG part is shown in red, the amide linkages cyan, the thio-propyl group bead is yellow. The molecules were re-constituted for the snapshots.

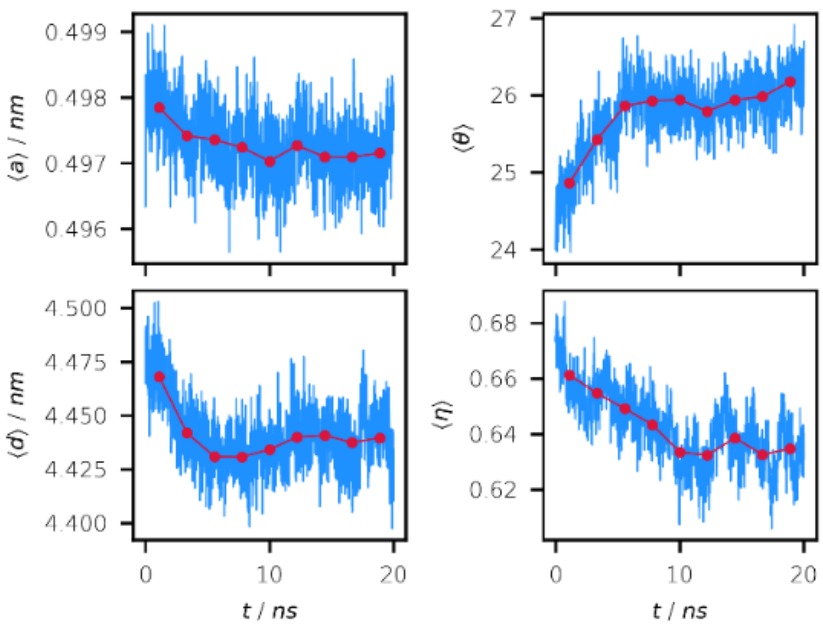

Figure E.4: Time evolution of all-atom MD simulation monolayer properties (blue) together with 10 block averages (red circles) during the production phase.

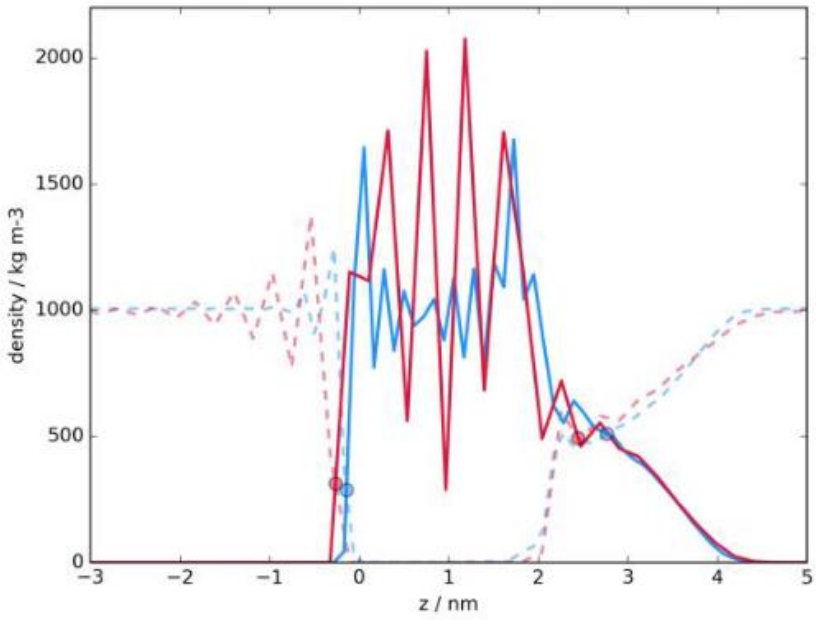

Figure E.5: Exemplarily normal direction density profiles of a full atomistic (blue) and a coarse grained (red) x_{C8} = 0.2 monolayer. Full lines represent the monolayers, dashed lines the solvent molecules. The calculated intercepts between monolayer and solvent profiles are marked with circles. The profiles are normalized to the position of the thio-propyl beads (CG) i.e. the sulfur atoms (AA).