UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2264

Theoretical and Biochemical

Advancing Protein Structure Investigations with Complementing Computations

MAXIM N. BRODMERKEL

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023

ISSN 1651-6214
ISBN 978-91-513-1797-7
URN urn:nbn:se:uu:diva-500274

Dissertation presented at Uppsala University to be publicly examined in Room B41, Biomedicinska Centrum, Husargatan 3, Uppsala, Friday, 2 June 2023 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Helmut Grubmüller.

Abstract
Brodmerkel, M. N. 2023. Theoretical and Biochemical. Advancing Protein Structure Investigations with Complementing Computations. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2264. 96 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-1797-7.

Life as we know it today would not exist without proteins. The functions of proteins for us and other organisms are linked to their three-dimensional structures. As such, protein structure investigations are a crucial contribution for understanding proteins and the molecular basis of life. Some methods probe the structure of proteins in the gas phase, which brings various advantages as well as complications. Amongst them is mass spectrometry, a powerful method that provides a multitude of information on gaseous protein structures. Whilst mass spectrometry shines in obtaining data of the higher-order structures, atomistic details are out of reach. Molecular dynamics simulations on the other hand allow the interrogation of proteins in high-resolution, which makes it an ideal method for their structural research, be it in or out of solution.

This thesis aims to advance the understanding of protein structures and the methods for their study utilising classic molecular dynamics simulations. The research presented in this thesis can be divided into two themes, comprising the rehydration of vacuum-exposed structures and the interrogation of the induced unfolding process of proteins. Out of their native environment, proteins undergo structural changes when exposed to vacuum. Investigating the ability to revert those potential vacuum-induced structural changes by means of computational rehydration provided detailed information on the underlying protein dynamics and how much of the structure revert back to their solution norm. We have further shown through rehydration simulations that applying an external electric field for dipole-orientation purposes does not induce irreversible changes to the protein structures. Our investigations on the induced unfolding of protein structures allowed a detailed look into the process of unfolding, accurately pinpointing areas within the proteins that unfolded first. The details provided by our simulations enabled us to describe potential mechanisms of the unfolding processes of different proteins on an atomistic level. The obtained results thus provide a potent theoretical basis for current and future experiments, where it will be very interesting to see MD compared with or complemented to experiments.

Keywords: Molecular dynamics simulation, Protein structure, Structural biology, Protein hydration, Electric dipole, Collision Induced Unfolding

Maxim N. Brodmerkel, Department of Chemistry - BMC, Biochemistry, Box 576, Uppsala University, SE-75123 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-513-1797-7
URN urn:nbn:se:uu:diva-500274 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-500274)

Dedicated to my family and friends,
and my teacher Dr. Fredi Engelbrecht,
who encouraged this journey.

'It's a long way to the top, if you wanna Rock'n'Roll.'

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I **Stability and conformational memory of electrosprayed and rehydrated bacteriophage MS2 virus coat proteins**
M. N. Brodmerkel, E. De Santis, C. Uetrecht, C. Caleman, E. G. Marklund.
Current Research in Structural Biology, **4**, 338-348 (2022).

II **Rehydration Post-orientation: Investigating Field-induced Structural Changes via Computational Rehydration**
M. N. Brodmerkel, E. De Santis, C. Caleman, E. G. Marklund.
The Protein Journal, in press.

III **Simulated collision induced unfolding of norovirus capsid dimers reveal strain-specific stability profiles**
M. N. Brodmerkel, E. De Santis, C. Uetrecht, C. Caleman, E. G. Marklund.
Submitted.

IV **Molecular dynamics simulations reveal barrel opening during the unfolding of the outer membrane protein FhaC**
M. N. Brodmerkel, E. De Santis, A. Konijnenberg, F. Sobott, E. G. Marklund.
Manuscript.

The author's contribution to the aforementioned scientific work is described in the following, based on the CRediT authorship contribution statement taxonomy:

Conceptualisation, Methodology, Formal analysis, Investigation, Validation, Writing—Original Draft, Writing—Review & Editing, Visualisation.

List of additional Papers

The following publications are not included in this thesis.

V **Structural Heterogeneity in Single Particle Imaging Using X-ray Lasers**
T. Mandl. C. Östlin, I. E. Dawod, M. N. Brodmerkel, E. G. Marklund, A. V. Martin, N. Timneanu, C. Caleman.
The Journal of Physical Chemistry Letters, **11**, 6077-6083 (2020).

VI **Reproducibility in the unfolding process of protein induced by an external electric field**
A. Sinelnikova, T. Mandl, C. Östlin, O. Grånäs, M. N. Brodmerkel, E. G. Marklund,C. Caleman.
Chemical Science, **12**, 2030-2038 (2021).

VII **Glycan-Induced Protein Dynamics in Human Norovirus P Dimers Depend on Virus Strain and Deamidation Status**
J. Dülfer, H. Yan, M. N. Brodmerkel, R. Creutznacher, A. Mallagaray, T. Peters, C. Caleman, E. G. Marklund, C. Uetrecht.
Molecules, **26**, 2125 (2021).

VIII **Coherent diffractive imaging of proteins and viral capsids: simulating MS SPIDOC**
T. Kierspel, A. Kadek, P. Barran, B. Bellina, A. Bijedic, M. N. Brodmerkel, J. Commandeur, C. Caleman, T. Damjanović, I. E. Dawod, E. De Santis, A. Lekkas, K. Lorenzen, L. L. Morillo, T. Mandl, E. G. Marklund, D. Papanastasiou, L. AI Ramakers, L. Schweikhard, F. Simke, A. Sinelnikova, A. Smyrnakis, N. Timneanu, C. Uetrecht.
Analytical and Bioanalytical Chemistry, 1-12 (2023).

List of Abbreviations

bMS2	Bacteriophage MS2
C_α	α-Carbon
CASP	Critical Assessment of Structure Prediction
CCS	Collision Cross Section
D	Dextrorotary
DNA	Deoxyribonucleic Acid
HDX-MS	Hydrogen-Deuterium Exchange Mass Spectrometry
ICTV	International Committee on Taxonomy of Viruses
IM-MS	Ion Mobility Mass Spectrometry
KS	Kawasaki Strain
L	Levorotary
MD	Molecular Dynamics
μs	Microsecond
MS	Mass Spectrometry
MS SPIDOC	Mass Spectrometry for Single-Particle Imaging of Dipole Oriented Protein Complexes
mRNA	Messenger Ribonucleic Acid
ns	Nanosecond
NMR	Nuclear Magnetic Resonance
NS	Norwalk Strain
P	Protruding
POTRA	Polypeptide-Transport-Associated
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
RNA	Ribonucleic Acid
S	Shell
SPI	Single Particle Imaging
XFEL	X-Ray Free-Electron Laser

Contents

Introduction

Proteins Under a Computational Microscope

Revealing, investigating and understanding the secrets of nature, the universal laws we all obey, are the fundamental pillars of science. Questioning our environment, observing and studying our surroundings drives our pursuit of knowledge. Whereas some people look up in the night sky amazed by the sheer endlessness of the universe and ponder about humanity's place amongst the stars, we still have to uncover all truths that makes us *us*. In a literal sense: what are we made of, how do we exist? A question that different research fields might answer distinctively. For a mathematician: equations. For a physicist: fermions and bosons. For a chemist: elements—hydrogen, carbon, nitrogen, oxygen, here and there a phosphorus and some calcium, with many others. For a biochemist: lipids, sugars, nucleic acids and proteins. We could go further and further like that, and obviously just scratching the bare surface of all the interpretations of life for those research areas, yet we came across the fundamental purpose of this thesis: advancing the understanding of proteins.

The common theme throughout all projects presented in this thesis, and those that are not part of it, is employing molecular dynamic (MD) simulations to interrogate the dynamic nature of proteins. The atomistic details MD simulations are able to provide are incredibly insightful not only about the general dynamics of proteins, but also how they react to changes of the simulation environment and applied parameters. One feature is consistent throughout all scientific publications this thesis is based on, namely simulating proteins in the gas phase. Out of their native milieu, proteins *feel* the presence of the hydrophobic environment of vacuum and consequently adjust their confirmation accordingly [1]. To what extent depends on a multitude of factors, such as charge state, size and composition—to name a few. Working close with scientists studying the experimental structural research of proteins, we aimed to investigate the magnitude and details of protein *in-vacuo* dynamics under pre-defined conditions.

The Scope of this Thesis

Proteins are the workhorses in our cells. Almost every single reaction or mechanism within our bodies involve at some point a protein [2]. Consequently, the workload of proteins is immense—transportation, catalysing reactions, muscular movement, regulation, signalling, *etcetera*. A crucial step in understanding proteins is to examine their function, which furthermore is often necessary for one to investigate their structure [3]. **Chapter 1**, ***Proteins***, provides an overview of the composition of proteins, how their structure manifests from it and why protein structures are so important.

Technically, not only living things make use of proteins. Viruses are a unique type of pathogen, as its living or non-living nature can be summed up as: kind of both [4]. Living, because they are driven by the replication of their genome, and non-living, as they are unable to store free energy. Surely, this is a simple grasp of the nature of viruses, but shows how interesting viruses are. Especially, when we talk about the structure of viruses, and the proteins that protect the viral genome. Whilst viruses differ widely in size and shape, their genome is encapsulated by a distinctive arrangement of proteins forming a capsid. Understanding the basics of these structural proteins not only provides crucial information of the viral life-cycle, but moreover accounts for an ideal model system to study larger protein complexes in general. A brief introduction to viruses is given in **Chapter 2**, ***Viruses***, featuring the significance of the structure of viral capsid proteins.

So, you might ask now, how does one obtain information and details about protein structures? A large number of methods to study protein structures exist, some potentially more suitable than others, depending on the protein and the motivation behind the research. Biomolecular nuclear magnetic resonance (NMR) spectroscopy [5] or X-ray crystallography [6] might be amongst the go-to methods, but we will put the focus on another powerful method, mass spectrometry (MS). MS enables the separation of particles according to their specific mass-to-charge (m/z) ratios in the gas phase [7]. Especially the development of electrospray ionisation (ESI) allowed to study intact proteins using MS, applying ideal experimental conditions for proteins to preserve their higher order structures close to their native solution environment; hence accrediting that process the term native MS [8]. However, while native MS provides insightful information about the composition and shapes of proteins, a drawback is the limited resolution when it comes to atomistic details about the spatial arrangement of the investigated proteins. Filling this gap can be accounted for using computational methods, particularly MD simulations, where the exact coordinates of all atoms within a well-defined system can be tracked, be it in solution, or, complementary to MS, in the gas phase [9]. Information of the precise atomistic positions obtained from MD simulations allows to examine the dynamics of the simulated system at a level of detail inaccessible for MS experiments, making MD and MS complementary partners for

structural research. MD simulations is further the main method employed for all research presented in this thesis, and is therefore extensively discussed in **Chapter 3**, ***Molecular Dynamics Simulations***. Apart from MD, we occasionally employed other methods or were provided input from MS experiments with the aim to support our simulations. These methods are introduced in order to provide a basic understanding of their capabilities and how they aided our research, briefly outlined in **Chapter 4**, ***Complementary Methods***.

Paper I and **Paper II** explored to what extent the proteins might undergo structural rearrangement in their conformations, and further, if those are reversible. Any information obtained from experiments about gas phase proteins could potentially carry over structural artifacts caused by vacuum exposure. As such, we took gas-phase protein structures and rehydrated them, with the aim to understand if the proteins will revert to their native solution structure. Are there parts that do not, and what are the underlying processes during the rewetting of these proteins?

In **Paper III** and **Paper IV**, we heated up gas-phase proteins to understand the structures' response to an applied heat stress. In vacuum, the absence of any surrounding for the proteins means that their atoms cannot exchange their energy with the environment, resulting in the thermally-induced unfolding of the proteins. So, why would we deliberately disturb protein structures? With the aim to simulate collision-induced unfolding (CIU), a method used during ion mobility MS (IM-MS) experiments to study protein conformations and their stability. Here, particles are purposely collided with inert gas molecules, eventually resulting in the unfolding of the analyte in a characteristic pattern. MD comes into play here being able to simulate CIU by running simulations over a range of increasing temperatures, and capturing their dynamics to understand the underlying unfolding mechanism.

1. *Proteins*

From wild elephants roaming the plains in Africa to cyanobacteria inhabiting thermal springs and geysers—life as we know it can be found in various sizes, evolved and adapted to live and endure in fascinating environments and conditions. Fundamental to the success of life were and are undoubtedly proteins, which, besides lipids, carbohydrates and nucleic acids, make up the fourth class of macromolecules in biology, and are by far the most abundant and diverse. Providing essential functions within our bodies, proteins act as enzymes, catalysing vital chemical reactions, as transporters, moving important molecules to their designated destinations, or as antibodies, supporting our immune systems combating pathogens; protein function is highly complex [2,3]. Certainly, the spatial conformation of proteins are directly linked to their functions, reason why investigating and understanding protein structures is an immensely important endeavour. In this chapter, the fundamentals of protein structures are introduced, how they fold into their specific three dimensional (3D) arrangements, and how those correlate to their function eventually.

1.1 Protein Building Blocks – Amino Acids

Multiple smaller molecules, called amino acids, are connected together to build proteins [2, 3]. Predominately made up of hydrogen, carbon, nitrogen and oxygen atoms, with sulphur and even selenium atoms in specific cases, amino acids are rather simple molecules themselves, yet nevertheless fundamental for our existence. With an exception for the cyclic amino acid proline, at a very basic level, the general structure throughout the remaining 19 standard α-amino acids is the same: a centric carbon atom (the α-carbon, C_α), bonded to a hydrogen atom, a side chain, as well as an amino and a carboxylic group. The α annotation describes here the location of the C-atom those substituents are attached to, being at the initial position of the C-chain [2, 3, 10]. Whilst β and higher amino acids exist, they are non-proteinogenic, and therefore not present in proteins [10, 11]. Hence, α-amino acids will be put into focus here, with an example amino acid structure displayed in figure 1.1.

The type of amino acid, and as such its properties, are defined by the side chain, the fourth bond and commonly designated as R (rest) group [2,3,10]. As an example, an aliphatic side chain would increase the hydrophobic, lipophilic properties of an amino acid, and consequently influence the properties of the protein it is making up. The four non-identical substituents located around the sp^3-hybridised C_α-atom form a chiral centre depending on the spatial config-

Where Do Amino Acids Originate From?

A question that has been studied extensively throughout the years. Worth mentioning is the Miller-Urey experiment conducted in 1953, which aimed to simulate the environment and conditions of a prebiotic earth using simple compounds such as hydrogen, water, ammonia and methane, which revealed the formation of simple amino acids amongst other products [12–14]. Potential reaction sites for the synthesis of amino acids on a primitive earth are hydrothermal vents on the oceanic floor. Experiments were conducted to mimic the correct hydrothermal conditions such as temperature and pH, and proved to be able to synthesise amino acids successfully [15–17]. However, one does not only need to look down to earth to find them: several amino acids were discovered in space or originating from space on meteorites. Alternative scenarios were proposed and investigated the interstellar formation of biomolecules and thus suggest an alternating origin of these amino acids [18–21]. More recently, the European Space Agency ESA was indeed able to confirm the presence of the simplest amino acid glycine on the comet 67P/Churyumov-Gerasimenko during the ROSETTA mission [22]. Independent from the origins, from the depths of earth or amongst the stars, amino acids are immensely important for life as we know it and thus responsible for our existence.

uration of the substituents [23]. The order of the substituents determines, based on a convention first established by Emil Fischer, whether a chiral molecule can either rotate plane-polarised light to the right (dextrorotary, D) or left (levorotary, L) [24, 25]. As such, amino acids are further defined as either D- or L-amino acids. The sole exception is hereby given for glycine with its two

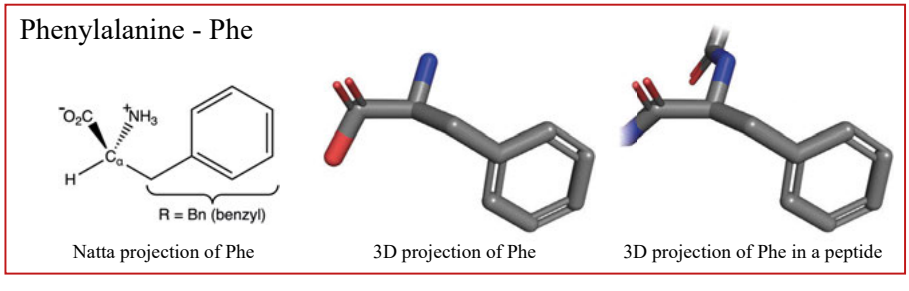

Figure 1.1. **The fundamental structure of an α-amino acid.** An α-amino acid is given by a central carbon atom connected to a hydrogen atom, the side chain R, and an amino and carboxylic group, where the type and properties of the amino acid is defined by R. Phenylalanine (Phe) for example is defined by its benzyl substituent as R, and possesses therefore hydrophobic properties, which further influences the properties of the polypeptide chain it may be part of.

hydrogen substituents; therefore being the only achiral amino acid. Interestingly, proteogenic amino acids for humans are almost exclusively found in the L-form, with the explanation for the evolutionary preference of the L-form over its other enantiomer being subject of various scientific debates [3, 26]. Regardless, research over the past decades has revealed that D-aspartate and D-serine can be found in humans as well, especially elderly individuals, whilst mostly being linked to diseases such as Alzheimers [25–27]. The importance of the correct amino acid stereochemistry was shown for an experiment where both enantiomers of the human immunodeficiency virus (HIV) protease enzyme were chemically synthesised [28]. Whilst the native L-form of the protease was able to process and cleave proteins of the same stereochemistry, its non-native mirror image was unable to do so, only catalysing the lysis of proteins solely made from D-amino acids, thus revealing the influence of the amino acid's chirality on both protein structure and function.

1.2 From Amino Acids to Peptide to Protein

Proteins consist of a series of concatenated amino acids, which are linked together by the formation of peptide bonds [2, 3]. As such, proteins are polymers, and further known as polypeptides. Amino and carboxylic groups are highly reactive species of functional groups, able to react with each other via a condensation reaction under the loss of molecular water [29, 30]. Within the cell, the information about the protein to be synthesised is supplied by the messenger ribonucleic acid (mRNA), providing the template of the amino acid composition for the ribosome to interpret during translation [2, 3, 30]. Whilst non-ribosomal protein synthesis exist, most commonly, the ribosome reads the information about the order of the amino acids of the protein chain, called the protein sequence, and builds up said chain under the formation of peptide bonds between the individual amino acids [30–32]. The template provided by the mRNA is highly important in order to assemble the correct protein with a designated sequence. As an example, for building any small protein of $N = 50$ building blocks length, made up from any combination of the 20 proteogenic amino acids, a total of 20^{50} potential sequences exist, with the possibility of finding the desired sequence being 1 over 20^{50}. Clearly, and luckily so, protein synthesis is no random assembly of polypeptide chains until the required protein is available, but rather a highly ordered and regulated series of mechanisms and processes [33]. Post-translation, the polypeptide chain can undergo further modifications, eventually folding into its predetermined structure [34].

After the polypeptide has been synthesised by the ribosome, the protein chain needs to obtain its native conformation. Driven by thermodynamics to access the most stable conformation, the polypeptide chain folds into a defined spatial conformation [35]. Envision a large, unfolded polypeptide chain; the number of conformations it could arrange in 3D space and fold into are enormous.

Nevertheless, timescales of protein folding range, depending on the size and composition of the polypeptide chain, between nanoseconds (ns) up to microseconds (μs) [36–39]. The process itself is often represented as a funnel aiming to visualise the free energy landscape a polypeptide chain has to navigate [40–42]. Initially covering a broad ensemble in the unfolded state, the polypeptide chain explores the landscape downhill towards the native structure, most often including the peptide to passing through intermediate structures along the way [43, 44]. The folding itself is driven by several factors, for example burying hydrophobic amino acids into a hydrophobic core away from the aqueous environment, where the loss of interactions with the water molecules may be compensated by the formation of internal hydrogen bonds and salt bridges [45]. In the cell, protein folding can in specific cases be initiated parallel to translation, directly in the cytoplasm or after being translocated to a specific part of the cell, for example the endoplasmic reticulum [46, 47]. The process of protein folding is therefore constantly monitored and checked by a plethora of interlinked mechanisms and processes in order to provide a quality control of the folded structure [48]. However, not only the folding process is surveilled; misfolded proteins need to be monitored and regulated just as well [49]. Disruptions of pathways regulating misfolding are often linked to diseases, even leading to apoptotic cell death, further indicating the importance of the quality control of protein structures for the benefit of the organism [49]. Eventually, reaching the bottom of the free energy landscape funnel, the protein has obtained its native conformation and thus is ready to carry out its designated task.

1.3 Proteins in Three Dimensions

Eventually, each protein adopts a specific 3D conformation. Providing further details about proteins, their structures can be separated and classified over four distinct levels [50]. Figure 1.2 demonstrates the connection between these individual levels based on an example for the structure of the alcohol dehydrogenase enzyme from the *Drosophila lebanonensis* fly [51].

The primary structure or level of a protein is described by its sequence, and as such its composition and order of amino acids [2, 3]. As highlighted in section 1.1, the individual amino acids influence the properties of the polypeptide chain, and therefore encode the proteins' biological function. Ergo, the primary structure directly determines the eventual protein structure. Within the first steps of protein folding, secondary structures form that further influence the shape and conformation of the protein [2, 3, 10]. Often termed local folding, secondary structures fundamentally describe well-defined folds, that amino acids adopt as a result of the formation of hydrogen bonds [29]. Most prominent representatives of secondary structures are α-helices and β-sheets, first described by Pauling *et al.* in 1951 [52, 53]. Whilst both α-helices and

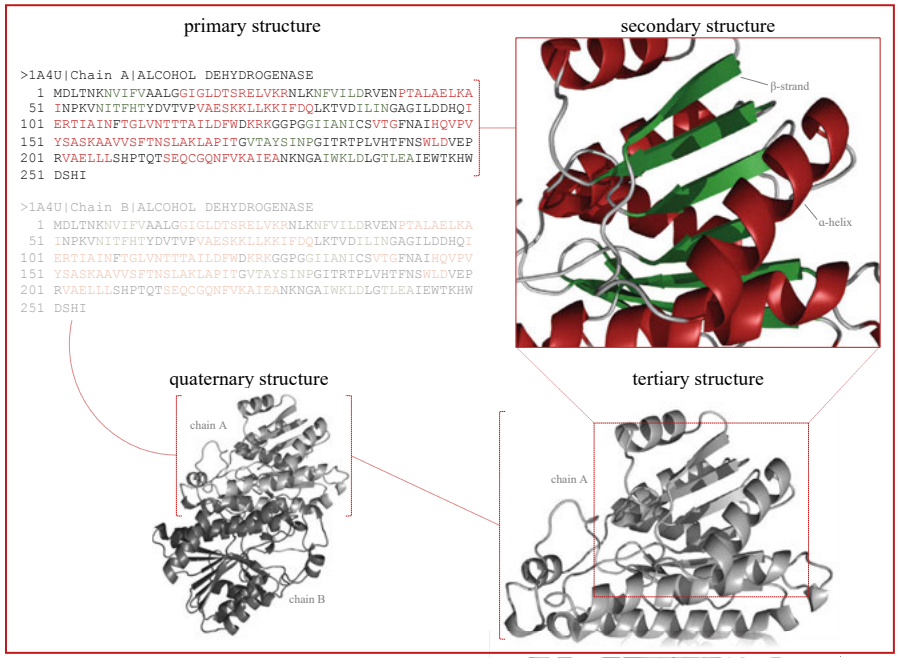

Figure 1.2. **Classifying protein structures.** The one dimensional description of a protein composition is its sequence, also known as primary structure. During the folding process, amongst the first key features the protein chain folds into are secondary structures, comprising predominately α-helices (red) and β-sheets, later made up by several β-strands (green), amongst other lesser common folds. The tertiary structure of the alcohol dehydrogenase protein describes the 3D conformation of the two individual chains A and B, which are taken together represented by the quaternary structure.

β-sheets are formed by hydrogen bonds, α-helices are determined by hydrogen bonds along parts of a single polypeptide chain, whilst β-sheets are formed by hydrogen bonds between adjacent parts (β-strands) of the same polypeptide chain, or other chains entirely. Other secondary structures may occur as well within a protein structure, such as the 3_{10}-helix, but are less common than α-helices or β-sheets. The tertiary level or structure describes the final, 3D spatial configuration of the respective polypeptide chain [54]. Tertiary structures are stabilised by internal and external interactions, including ionic bonds, disulfide bridges, hydrophilic or lipophilic interactions [2, 3, 10]. For monomeric proteins, no higher level exists. However, various proteins are present and only biologically active as aggregates of multiple folded polypeptide chains, refered to as subunits. As such, these proteins possess a quaternary structure [55]. Being the overarching and final level, the quaternary structure for a protein describes the total composition of the biologically active protein, providing information about the number and interconnection of folded polypeptide chains and potential co-factors which make up said protein [2, 3, 10].

The four levels of protein structures provide a basic understanding about their compositions, their folds and spatial arrangements. As the higher levels describe the final conformations of a protein, it being monomeric or consisting of multiple peptide chains, the structural information given might already point towards its biological task.

1.4 Correlating Structure and Function

Within the nanomachinery of the cell, proteins make up the vast majority of biological cogwheels keeping the engine running. Specific proteins have their specific tasks and responsibilities, which are ultimately linked to their 3D conformation [2, 3]. Looking at proteins, one could make the assumption that their structure does not change as the proteins fulfil their function, although the structure of a protein is not static, but rather highly dynamic [56]. As such, understanding how a protein structure is associated with its function ultimately involves understanding protein interactions, either with other proteins, with ligands or other molecules. To emphasise on the correlation between protein structure and their function, let us have a look at two examples.

Not all biochemical reactions are favourable; for the majority, the activation energy is relatively large [23]. Evolution provided the solution: enzymes. With some exceptions, those being RNA-based, almost all enzymes are proteins [57]. Enzymes function as biological catalysts, enhancing the reaction rate of a biochemical reaction by lowering its activation energy [57, 58]. This often includes the enzyme to form favourable, specific interactions with the substrate at a defined area, the active site, of the enzyme protein structure. The active site is therefore made up by amino acids that are able to form preferential interactions with a substrate. Any mutation of the active site can therefore have dramatic effects on the enzyme's catalytic activity and efficiency, proving how important the information encoded in the primary protein structure is for enzymes and their function [59–61]. Moreover, enzyme protein denaturation, meaning the protein is losing its functional fold, or dissociation into its subunits usually signifies a total loss of its catalytic capabilities [62–65]. This indicates further that changes of the secondary, tertiary and quaternary structure of an enzyme directly impact its activity, and how crucially important these are for the enzyme.

However, proteins possess significantly more functions in biology other than catalysing biochemical reactions [2, 3]. Specific proteins ensure the distribution of molecules through the body, or allow the translocation of a solute across membranes [66, 67]. These proteins generally demonstrate an overall highly mobile structure throughout transportation. An excellent example is given by the oxygen transporter hemoglobin. Hemoglobin is a tetrameric protein consisting of two pairs of subunits, termed α- and β-subunits, that differ in length and amino acid composition [67, 68]. Crucial for its task as oxygen

transporter are the heme prosthetic groups with a Fe^{2+} core, which eventually allows for O_2 fixation, and thus oxygen transportation. The iron ion itself is coordinated by the porphyrin rings of the heme group, and a histidine residue from a protein subunit [69, 70]. Upon binding with oxygen, conformational changes occur in hemoglobin that eventually shifts the position of the Fe^{2+} atom further into the plane of the heme group, therefore increasing the affinity towards oxygen for the protein [2, 3]. The importance of the structure of hemoglobin to carry out its function becomes more evident if the structure is disordered [70]. Sickle cell disorder describes a disease where the structure of hemoglobin is affected as a result of a residue point mutation at position 6 of the sequence from glutamic acid to valine within the β-subunits [71–73]. The mutation causes hemoglobin to form a sickle-like conformation instead of the native concave disk-like structure, which can lead to hemoglobin proteins clotting together, eventually causing anemia.

These are just a few examples of how the structure of proteins is directly linked to their function. As such, investigating protein structures is of immense importance in understanding their role and significance, and motivated the research presented in this thesis.

2. *Viruses*

Viruses are non-living microscopic infectious particles, and cause of various known diseases such as gastroenteritis, hepatitis, influenza or severe acute respiratory syndrome [74–77]. Unable to reproduce by themselves, viruses require a host organism for replication [78, 79]. All life forms may be infected by viruses, from microorganisms, including bacteria and archaea, up to plants and animals [80]. After infecting a host cell, these pathogens are able to reprogram and hijack the nanomachinery of the cell in order to produce more viruses. As a result of their rapid reproduction and mutation rate, viruses represent a considerable public health threat, making them subjects of continuous endeavours by researchers worldwide in order to prevent and treat viral infections [81–84]. The recent coronavirus pandemic serves as a painful example why viral research is so highly important [85, 86]. The research presented in this thesis includes investigations of the dynamics of selected viral proteins. Fundamental information about viruses and their biological composition is therefore provided in this chapter, with particular emphasis on the viruses relevant to the scientific work presented in this thesis.

2.1 Classifying Viruses

As more and more different appearances and variations of viruses were discovered, the more obvious it became to develop and assign a fitting nomenclature to gather and categorise the growing information about viruses in a meaningful way. The International Committee on Taxonomy of Viruses (ICTV) is an international organisation within the International Union of Microbiological Societies, providing a globally accepted hierarchical virus taxonomy [87]. This taxonomy defines unique names and suffixes which should serve as identifiers to improve the communication between scientists. The latest[1] published classification scheme for a total number of 10434 viral species is listed in table 2.1. One of the most important criteria for the categorisation of viruses depends on the genome they carry, as the chemical composition and the type of genome determines the mode of viral replication [87, 88]. Generally, the viral genome may be configured as RNA or deoxyribonucleic acid (DNA). Roughly 70 % of all viruses carry RNA as their genetic material [89], where the RNA strand

[1]Based on the latest ICTV report, published in July 2021. Data is available at https://talk.ictvonline.org/taxonomy/. Accessed on the 17^{th} of April 2023, 2:35 pm.

Table 2.1. ***Primary ranks for virus taxonomy proposed by the ICTV.*** *Ranks are ordered from the highest to lowest given for the Norwalk virus.*

Rank	Suffix	Current Taxonomy	Example
Realm	*-vira*	6	*Riboviria*
Kingdom	*-virae*	10	*Orthornavirae*
Phylum	*-viricota*	17	*Pisuviricota*
Class	*-viricetes*	39	*Pisoniviricetes*
Order	*-virales*	65	*Picornavirales*
Family	*-viridae*	233	*Caliciviridae*
Genus	*-virus*	2606	*Norovirus*

may be single- or double-stranded. If a single-stranded RNA can function as messenger RNA, and thus be translated at the host cell ribosome, the RNA is classified as sense strand, denoted with a +-sign (+RNA). If the RNA is complementary to the messenger RNA, the strand cannot directly express viral proteins and thus is considered antisense (-RNA). DNA-containing viruses may comprise either a single- or double-stranded genome, and very few are found with a circular single-stranded DNA. This approach for virus classification based on the type of nucleic acid was proposed by David Baltimore in 1971 and is still used today [90]. Nevertheless, nomenclature for viruses is still a highly debated topic and will likely become more defined and detailed in near future, taking into account other crucial information such as sequencing and the genome structure [91, 92].

2.2 Virus Capsids and Coat Proteins

The general structure of a virus comprises their genome, often together with other important viral proteins, encapsulated by a protein shell—the capsid. The capsid can further be encased by a lipid-based envelope. These compartments together are often summarised and described as virion. Capsids are built up by a defined number of protein building blocks, protomers, which organise themselves to form a specific 3D arrangement. Three examples for such arrangements are depicted in figure 2.1.

Capsid structures are usually highly symmetrical, as icosahedral (*e.g.* Norwalk virus) and helical (*e.g.* Measles morbillivirus) morphologies, yet can manifest as well as seemingly asymmetric, as one can observe for the conical form of the HIV capsid. Icosahedral capsids are, as the name suggests, made up of 20 identical triangular faces. In 1962, Donald Caspar and Aaron Klug found that only a certain number of building blocks make up these triangular faces, and thus are able to form quasi-symmetrical capsids [93]. They published a mathematical approach in order to calculate the arrangement of protomers in

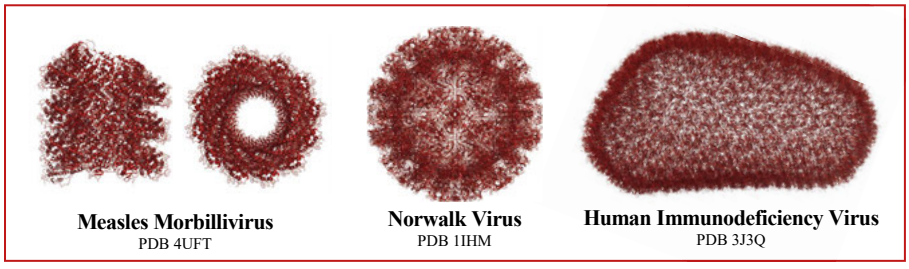

Figure 2.1. **Examples of viral capsids.** Shape and structure of the capsid of viruses differ between different families and strains, yet serve the same function: housing, protection and delivery of the viral genome during its life-cycle. Structures were obtained from the Protein Data Bank (PDB) from the codes 1IHM, 2TMV, and 3J3Q, respectively, and scaled for better visualisation.

viruses, defined as triangulation number T, as

$$T = h^2 + hk + k^2, \tag{2.1}$$

with h and k being any pair of positive integers (including zero). As such, T represents a certain value determining the number of protomers composing a single triangular face. As a result, capsids are often classified according to their triangulation number.

2.3 The Viral Life-cycle

Protection, transportation and delivery of the genome are the main functions of viral capsids. As such, capsids play a critical role during the viral life cycle. How viruses replicate, and the specific steps in between, may differ depending on the type of virus infecting a host [94]. Nevertheless, main stages of the replication cycle can be defined, providing a general overview of the progress of viral infection, replication and the release of new virions. Viruses follow a rather simple task—replication. In more detail, replication of the genetic information. Consequently, the life cycle of a virus focuses on the replication of its genome and the expression of its proteins in order to produce multiple replicas of itself (figure 2.2).

At first, the virion has to attach to and enter the cell, followed by the disassembly of the capsid which ultimately releases the viral genome. The genome gets replicated and serves as template for the expression of viral proteins. Eventually, the capsid reassembles encoating the genome, and as such, new virions are created, which can now exit the cell to further infect and replicate.

Studies have shown that some capsids can self-assemble *in vitro* from isolated genetic material and coat proteins [95]. Even in absence of a viral genome, yet under the right conditions (pH and ionic strength), empty capsids can assemble rapidly *in vitro*. This remarkable feature, together with the importance of

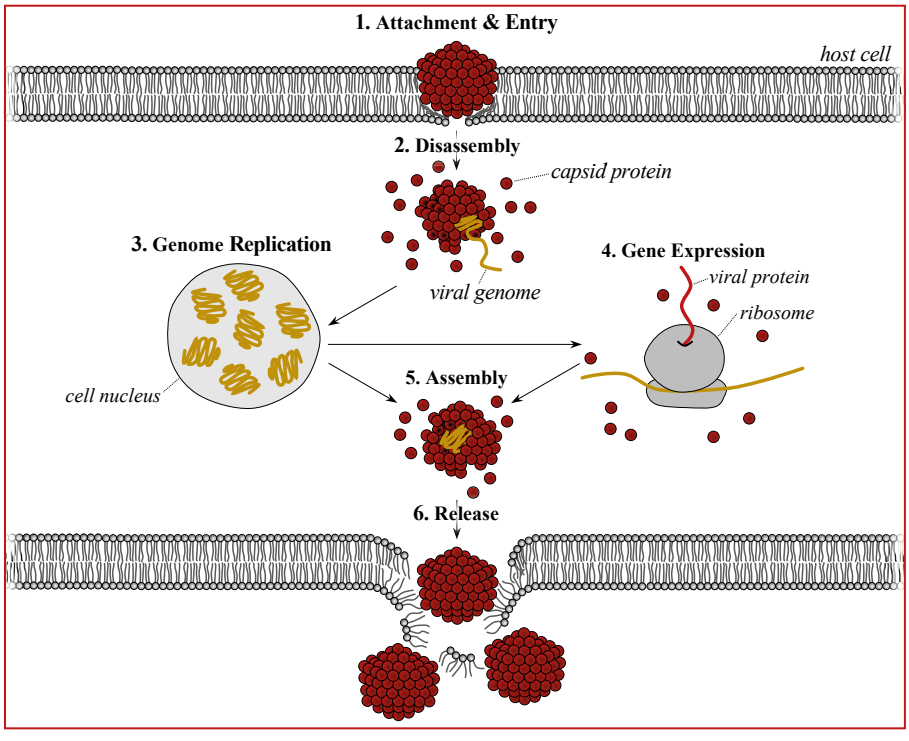

Figure 2.2. **The viral life cycle.** Replication of viruses can be generalised in six basic steps. Attachment and entry of a host cell can be counted together as the initial step of the infection. The capsid disassembles following cell entry and releases the viral genome into the cytosol, where it hijacks the cell's nanomachinery for its replication. The amassed genetic copies of the initial viral genome serve as a template for viral proteins. In this example, the genome is replicated in the nucleus, whereas other viruses might do so in the cytosol as well. Regardless, for the final step, new virus particles (re)assemble and exit the cell in order to spread out and infect other cells.

the capsid throughout the viral life cycle, makes capsids an interesting target in various research areas [96–98]. Correctly assembled capsids are crucial for genome protection, and any errors during the self-assembly process can have a negative impact on the life cycle [99]. Inhibiting capsid assembly is therefore an appropriate antiviral-drug target [100, 101]. Studying the assembly process can give important insights of viruses. A steadily increasing number of approaches to do so are employing theoretical and computational methods in order to understand and investigate capsid assembly protocols [99, 102]. In general, understanding the dynamics of the viral capsid and its subunits may provide substantial information about viruses. MD accounts for one methodology to study this, and will therefore be described in more detail in the following chapter.

3. *Molecular Dynamics Simulations*

Understanding the structure and dynamics of proteins goes hand-in-hand with understanding their function. One toolset providing detailed atomistic information about the motions of proteins is MD simulations, described in this chapter of this thesis. After introducing the underlying theory, an example workflow of setting up an MD simulation is given, and the impact and importance of the simulation environment is discussed.

3.1 Fundamentals of Molecular Dynamics Simulations

MD provides a detailed description of atoms, (bio)molecules, solids and liquids on an atomistic level, and is therefore regularly employed as a powerful method in various research disciplines. It allows the theoretical study of time-dependent dynamics of molecular systems on an atomistic level, at a reasonable computational cost.

The (Hi)story of Molecular Dynamics Simulations

Contrary to the potential perception MD simulations might be a 'novel' computational method, it has been around for several decades. Alder and Wainwright performed the first MD simulations in 1957, performing simulations of the phase transition of simple gases [103,104]. However, it would still take several years until proteins were put into the spotlight for MD simulations. In the late 1970s, McCammon *et al.* reported the first MD study of a protein, bovine pancreatic trypsin inhibitor, revealing the proteins high internal motions at 295 K [105]. As such, MD simulations were introduced to a new audience, most noteworthy in structural biology, starting a success story indicated by the number of MD-related publications since 1977 [106]. Technological advances, especially in high performance computing, accompany and accelerate the growing interest in it, thus progressively marking MD as powerful method for a multitude of research disciplines [107].

Other methods interrogate computational biophysics of proteins embracing quantum effects to model electronic rearrangement, yet along with an excessive demand of resources, which only allows simulations on rather limited

timescales or small systems. MD however is based on a classical mechanics approach, and therefore requires significantly fewer resources whilst providing a good numerical description of the intramolecular dynamics within targeted proteins.

3.1.1 Classic Newton

Atoms and molecular bonds within the system are defined in a classical description as a series of charged particles linked (bonded) by mechanical springs. Stepping through time, MD solves Newtons equations of motion for each atom to accurately describe their positions and velocities. The information obtained from all atoms defines the current state of the system at a specific point in time. Combined molecular frames represent the trajectory of the molecule, containing all information about its dynamics and behaviour throughout the simulation.

Newtons second law of motion defines the force $\vec{F}_i$ acting upon an i^{th} particle of a system being equal to its accelerated ($\vec{a}_i$) mass m_i, given by

$$\vec{F}_i(t) = m_i \vec{a}_i(t), \tag{3.1}$$

with $i = \{1, 2, ..., N\}$. MD follows a deterministic physical model, meaning that future states of the system are in principle clearly defined by its initial conditions. This allows the calculation of $\vec{F}_i$ of the system by the gradient of the potential energy V, a function of the positions $\vec{r}_i$ of all atoms, from

$$\vec{F}_i(t) = -\frac{dV}{d\vec{r}_i}. \tag{3.2}$$

Coordinate files, usually supplied by from experimental observations, such as crystallography data, specify the 3D arrangement of all atoms within a protein, and thus provide their starting positions for MD simulations. Initial velocities $\vec{v}_i$ are derived from the Maxwell-Boltzmann distribution of the atoms, scaled to reflect the correct temperature and pressure [108]. Derived from experimental data and quantum mechanics, force fields provide the parameters and physical models describing all interactions between the atoms of a given system, and further define the potential energy V. From here, all forces are known, allowing the calculation of the acceleration by solving equation 3.1 for $\vec{a}_i$, as given by

$$\vec{a}_i(t) = \frac{\vec{F}_i(t)}{m_i}. \tag{3.3}$$

The acceleration of an i^{th} particle can be derived as the derivative of its velocity $\vec{v}_i$ with respect to time t, or as the second derivate of its position $\vec{r}_i$ with respect to the time,

$$\vec{a}_i(t) = \frac{d\vec{v}_i}{dt} = \frac{d^2\vec{r}_i}{dt^2}, \tag{3.4}$$

implying that the new atomic positions and velocities of the next time step can be determined by integration of $\vec{a}_i$. This is done by various algorithms, utulising different numerical approaches to integrate Newton's equations of motion for all atoms within a system. One important applied numerical algorithm for MD is the *Verlet* integration method, used by its name patron, french physicist Loup Verlet [109, 110]. Since then, the *Verlet* algorithm counts as one of the most prominent integration methods, due to its accuracy and simplicity [111]. The algorithm itself is obtained by the addition of two Taylor expansions for the atomic positions at time t,

$$\vec{r}_i(t+\Delta t) = 2\vec{r}_i(t) - \vec{r}_i(t-\Delta t) + \frac{1}{m}\vec{F}_i(t)(\Delta t)^2. \tag{3.5}$$

Equation 3.5 explains how the algorithm calculates the new position of an atom at time $(t+\Delta t)$, from the position and force at (t) and its previous position at $(t-\Delta t)$. However, velocities are not directly calculated, and thus need to be generated if need be [111], on the cost of additional computational resources. This led the scientific community to develop more efficient algorithms with the goal to accelerate MD [112–115].

3.1.2 Force Fields

The quality of MD simulations strongly depends on the applied physical model chosen as force field. As mentioned above, a force field includes parameters and functions to describe all atomic interactions, defined as the potential V. Moreover, V can be written as the sum of bonded and non-bonded interatomic interactions, V_{bonded} and $V_{non-bonded}$ respectively,

$$V = V_{bonded} + V_{non-bonded}, \tag{3.6}$$

yet may include additional terms to further account for specific interactions and external forces. Nevertheless, the majority of classical force fields rely on five fundamental potential energy terms that together describe the bonded and non-bonded interactions within the simulated system (figure 3.1).

Bonded Interactions

The bonded contribution of V can be split up in three major terms. The first term describes the stretching of a covalent bond between two atoms, denoted V_{bond}. Usually represented as harmonic potential, V_{bond} describes the chemical bond between two atoms, and is given as

$$V_{bond} = \sum_{bond} \frac{k_d}{2}(d-d_0)^2. \tag{3.7}$$

The force constant is given as k_d, whilst d_0 represents the referenced or equilibrium bond length. Approximating V_{bond} as harmonic potential accommodates well when used in MD without bond breaking, as the function would

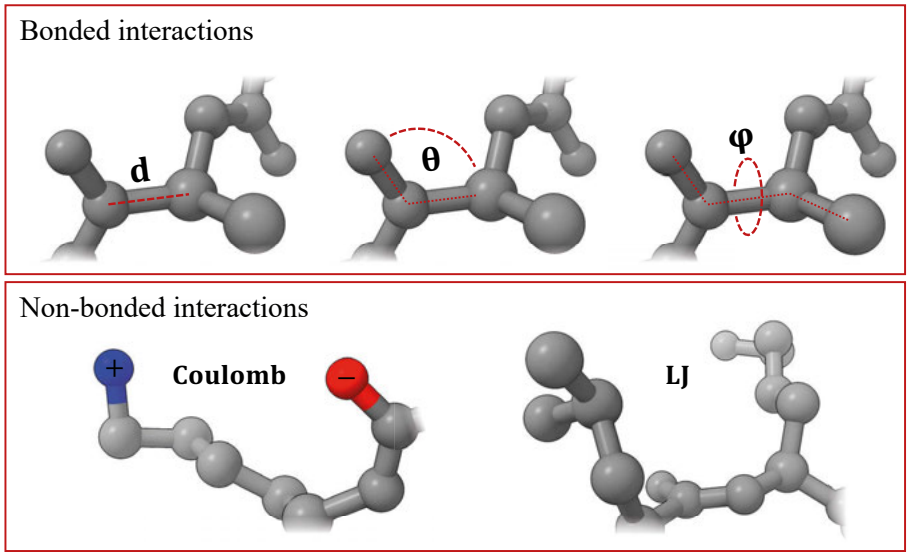

Figure 3.1. **Fundamental atomic interactions defined in force fields.** Representing the bonded and non-bonded interactions within a system, five terms describe the underlying potential energy contributions: bond stretching, angle bending, dihedral rotation, electrostatic (Coulomb) and Lennard-Jones (Van der Waals) interactions.

otherwise grow inaccurate at larger deviations (compression/stretching) from d_0. For simulations, in which bond breaking is necessary, a Morse potential might be the better choice as it provides a higher accuracy of V_{bond} under these circumstances [111].

The second term reflects the contribution of angular deformation from the equilibrium angle θ_0,

$$V_{angles} = \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2, \tag{3.8}$$

with k_θ being the harmonic spring constant. The angle θ itself is formed between three atoms across two bonds. Lastly, one has to take into condsideration the potential energy contribution of atoms seperated by three bonds, and the dihedral angle ϕ which is formed between them. This is done by the third term $V_{dihedrals}$, often referred to as torsional term, and can be defined as

$$V_{dihedrals} = \sum_{dihedrals} \frac{k_\phi}{2}(1 + \cos(n\phi - \phi_0)). \tag{3.9}$$

Here, k_ϕ describes the barrier height, n the multiplicity (number of minima), and ϕ_0, which defines the position of the minima. As such, the bonded contri-

bution for V can be expressed as

$$V_{bonded} = \sum_{bond} \frac{k_d}{2}(d-d_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta-\theta_0)^2 + \sum_{dihedrals} \frac{k_\phi}{2}(1+\cos(n\phi-\phi_0)). \tag{3.10}$$

Non-bonded Interactions

Interactions between atoms that are separated by more than three bonds (dependent on the choice of the force field), or belong to another molecule for that matter, are accounted for in the non-bonded contribution for the potential V, defined as $V_{non-bonded}$. Similar as V_{bonded} can $V_{non-bonded}$ be separated into specific terms which describe the non-bonded interactions. Two terms make up $V_{non-bonded}$, the electrostatic interaction between atoms $V_{Coulomb}$ and V_{LJ}, which models the Van der Waals interactions, as

$$V_{non-bonded} = V_{Coulomb} + V_{LJ}. \tag{3.11}$$

The electrostatic contribution, also known as Coulomb potential, between pairs of atoms i and j at distance r_{ij} of certain charge, q_i and q_j respectively, can be calculated using $V_{Coulomb}$,

$$V_{Coulomb} = \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}, \tag{3.12}$$

with ε_0 being the permittivity of free space. Van der Waals interactions are often calculated by the Lennard-Jones potential, which is given by

$$V_{LJ} = \sum_{ij} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]. \tag{3.13}$$

Moreover, the Lennard-Jones potential contains two fundamental interactions, an attractive and a repulsive term. Attractive Van der Waals forces are approximated in the $\left(\frac{\sigma_{ij}}{r_{ij}}\right)^6$ term, based on dipole-dipole interactions between the atoms. However, atoms will repel each other below a certain distance as an effect of the Pauli exclusion principle, accounted for in the repulsive term $\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}$ in equation 3.13. Parameters ε represents the depth of the potential well, with σ describing the distance of minimum energy of an interacting atom pair. Together, V_{LJ} represents a simple model to approximate the inter- and intramolecular interactions of atoms within a system. Accordingly, $V_{non-bonded}$ can be written as

$$V_{non-bonded} = \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{ij} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{3.14}$$

which eventually, together with V_{bonded}, outlines the potential V within a force field, defined as

$$V = \sum_{bond} \frac{k_d}{2}(d-d_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta-\theta_0)^2 + \sum_{dihedrals} \frac{k_\phi}{2}(1+\cos(n\phi-\phi_0)) + \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + \sum_{ij} 4\varepsilon_{ij}\left[\left(\frac{\sigma ij}{r_{ij}}\right)^{12} - \left(\frac{\sigma ij}{r_{ij}}\right)^{6}\right]. \quad (3.15)$$

For MD simulations of biological macromolecules, the user can select from a large number of different force fields that are currently available, such as CHARMM, AMBER, OPLS, or GROMOS [116–120]. The aforementioned force fields are non-polarisable, meaning that partial charges are approximated as fixed points on the respective atoms. This lead researchers to develop polarisable force fields, which expand the potential energy functions to allow the charge distribution to adopt to the simulation environment [121, 122].

Regardless, based on the model and system to be investigated, certain force fields might deliver more accurate results than others. Force fields are highly important for MD simulations, providing accurate descriptions of the dynamics of proteins. For that reason, force fields are themselves subject of ongoing research to improve their reliability, steadily advancing the accuracy of the simulations which employ them [123, 124].

3.1.3 The GROMACS Simulation Package

A potential user can select between several software packages to run MD simulations. Some prominent software codes include chemistry at Harvard macromolecular mechanics (CHARMM) [117, 125], Amber [126, 127], or GROMACS, with its name being derived from its former abbreviation for *Groningen machine for chemical simulations* [128–130], with various other packages being available. The GROMACS software package was solely employed to obtain all simulations presented in this thesis, and will therefore be focused on here. Developed in the 1990s at Department of Chemistry at the University of Groningen, the Netherlands, GROMACS became one the most widespread MD simulation software suites today [128, 131]. Its open-source GNU General Public license encourages research groups from all over the world to edit and improve the source code steadily [130–133], further facilitating the success of GROMACS. Nevertheless, any chosen software provides the means to perform a molecular simulation, most importantly due to an underlying algorithm [129, 134], with its basic concept being displayed in figure 3.2.

Initially, a file is supplied containing the positions of the atoms of the system (step 1). GROMACS then calculates the forces acting on each atom (step 2), applies them, and updates the information of the new coordinates of all atoms within system (step 3). This is done for each time point, separated by

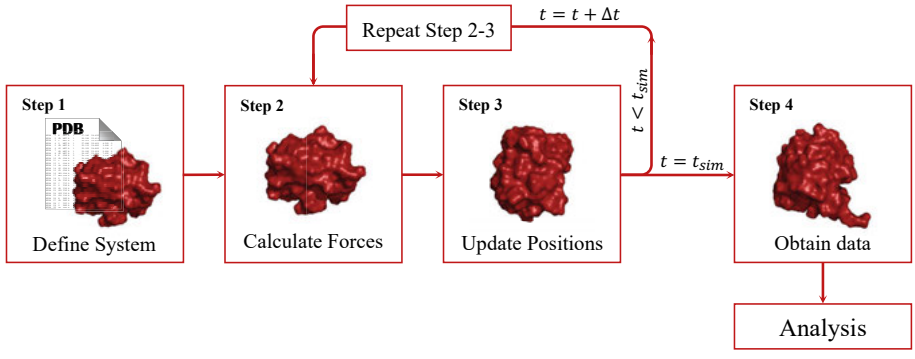

Figure 3.2. **MD algorithm at work.** At the beginning, a file containing the positions of all atoms is given as input to define the system. Subsequently, the algorithm calculates the forces acting on each atom, responsible for the atoms changing their position. For each time step Δt, a defined time window for the simulation, the algorithm updates the positions of each atom again, until the final simulation length t_{sim} is reached.

Δt, until the desired simulation time is reached (step 4). The time step should however be chosen with care to allow for accurate simulations at acceptable computational speed and cost [135]. Too small of a time interval, whilst conserving the energy with high accuracy, might increase the computing duration unnecessarily, and thus waste resources, whereas time steps too large could potentially imperil the energy conservation during the integration process and thus destabilising the system [135, 136]. Nevertheless, after the defined simulation length is reached and the simulation finishes, the obtained data can either be analysed to obtain important information about the dynamics of the system, or can further be provided as input for follow-up simulations.

3.2 The Workflow of a Molecular Dynamics Simulation

Obtaining valuable data is the goal of every MD simulation. Be it to understand the detailed dynamics involved in protein-ligand interaction, the thermal unfolding of a protein or the structural change induced by the change of the simulation environment—MD simulations need to be planned in a clear-defined and motivated manner. Setting up an MD simulation is not complicated, yet requires several steps in order to do so. Eventually, the obtained data of these simulations are analysed with a manifold of methods in order to explore the quantities of interest. The workflow shown here assumes to use the GROMACS simulations package for a protein simulation in solution, if not stated otherwise.

3.2.1 Towards Obtaining Molecular Dynamics Data

As mentioned above, an MD simulation can be split up into specific steps. Whilst these steps might differ from simulation to simulation, a solution simulation can be characterised by six elemental steps from the initial input structure to the final MD simulation production run. An overview of this process for a model protein simulation is briefly outlined in this section, as well as illustrated in figure 3.3.

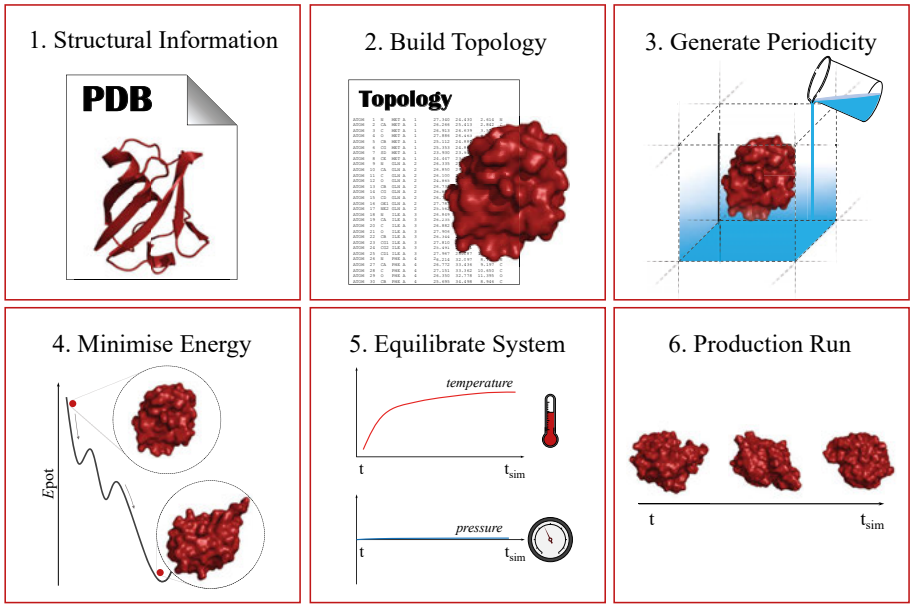

Figure 3.3. **Basic MD solution simulation workflow.** The most important principles of an MD simulation can be split up into six steps. At first, an input file containing the structural information is required, which is used to build the topology of the system, *i.e.* a description of all atoms and their interactions. Next, the molecule of interest is placed in a simulation box of chosen geometry, with periodic boundary conditions applied. If the molecule is to be simulated in solution, the solvent with counter ions is added to the simulation box. For better visualisation, the solvent is shown only in step 3. Initial large forces and structural distortions are cleared by minimising the energy of the system, allowing it to find the nearest local energy minimum. Afterwards, the temperature and pressure of the system are adjusted to a designated value, followed by the final production run that captures the dynamics over a given timespan. For *in vacuo* simulations, step 3 is skipped, and only the temperature of the molecule equilibrated in step 5, whilst the rest of the workflow follows the same schematic.

1. Structural Information

At first, an initial structure of the system to be simulated needs to be provided. Most often, the input is obtained from the PDB [137, 138], a collection of structural data of proteins and other biological macromolecules. Generally,

the entries in the PDB are based on refined structures provided by crystallography experiments, yet can as well be provided by other methods just the same, for example through NMR or cryo-electron microscopy. Eventually, the information is provided as an input file containing the coordinates of each atom within the protein or macromolecule in order to accurately describe the 3D structure.

2. Build Topology

The structural information about the protein needs to be able to be interpreted by the MD simulation software. In order to do so, the GROMACS software translates the input file and builds the topology, a description of all atoms and their interactions in the system [136]. Moreover, the topology is updated if other molecules, like solvent molecules, are added, and as such functions as an archive over all atoms in the system to be simulated. The topology itself is highly important, as it provides the information about which atoms add and parameters apply to the potential energy function, thus supplying the force field with the parameters to be used.

3. Generate Periodicity

Solution simulations need to be prepared in a way that the simulation environment is built and functions as realistically as possible. Simply adding some layers of water around a protein would result in the solvent molecules dispersing into infinite space, evaporating of the solvent. Water constitutes the vast majority of the native environment of proteins, therefore a realistic simulation environment should contain enough water so that the proteins are submerged throughout the whole MD simulation [2, 139, 140]. However, as the water molecules are dynamic as well, and the computational cost grows with the number of atoms in the simulation, simulating a massive body of water with a single protein inside would be computationally inefficient [111, 134]. Thus, specific conditions for the boundaries of the simulation need to be applied.

GROMACS places the protein in a specified simulation box and applies periodic boundary conditions (PBC), aiming to create a unit cell of a macro-scale system [136]. As such, infinitely large systems are represented as infinite replicas of the primary cell under PBC, dramatically reducing the computational time and cost to simulate the system, whilst retaining a realistic simulation environment [111, 134]. The box geometry plays here a highly important role: as the simplest approach of creating periodic images of the primary cell is done using a cubic geometry, a cubic box might potentially be computationally less efficient than other geometries. Assuming that the protein is in the centre of the box, solvent molecules located in the cube corners could be considered *dead-weight*—barely interacting with the protein, yet contributing to the computational workload [136]. Depending on the size of the total system, and the available resources, an octahedral or dodecahedral periodic box might

be the better choice, and thus reducing the effective volume of the simulated box, rendering theses geometries more computationally efficient [136]. For the here presented example MD simulation however, a cubic periodic box was chosen. After GROMACS created the system under PBC, it fills the periodic boxes with the chosen solvent, and is further able to adjust the salt concentration to a pre-defined level, providing all necessary files preparing the system for simulation.

4. Minimise Energy

So far, the system is purely static, and its spatial arrangement has not changed from its initial configuration. However, the system could potentially possess sterical problems, *i.e.* atoms in too close proximity or added solvent molecules, which were not placed accurately enough around the protein [141]. As such, the system might be in a state that is not energetically favourable. An energy minimisation, employing the *steepest decent* algorithm for example, is conducted to allow the system to find the closest local energy minimum. In more detail, during this step, the system is allowed to rearrange slightly in order to remove potential steric problems and obtain a more proper molecular arrangement [141]. In certain cases, this can be complemented with short MD simulations with applied position restraints for a specified selection of atoms [129]. Position restraints are applied to prevent sudden structural shifts and rearrangements, allowing the system to obtain proper and plausible configurations.

5. Equilibrate System

Before the final step, the system is allowed to relax and equilibrate further. As experiments are generally defined at an intended temperature and pressure, it is reasonable to perform MD simulations under similar conditions. Running simulations in specific thermodynamic ensembles, the temperature of the system can be ramped up to a targeted value, and the volume, and therefore the pressure, of the simulation box might be altered if necessary, with the goal of matching experimental conditions [111].

To adjust the temperature of the simulation, the accurate ensemble to select is given where the numbers of particles (N), volume (V) and temperature (T) is kept constant—denoted canonical NVT ensemble. This means that the total energy of the system is subject to change, and therefore allows for adjusting the temperature to a designated value. During an MD simulation, a thermostat either increases or decreases the kinetic energy of the system, thus modulating its temperature, with some prominent thermostats being the Berendsen thermostat [142], the velocity-rescaling [143], or Nosé-Hoover thermostat [144, 145].

The constant-temperature, constant-pressure (P) ensemble (NPT), therefore an isothermal-isobaric ensemble, allows control over both the temperature and

pressure. The unit cell vectors of the periodic box are allowed to change, and the pressure is altered by adjusting the volume. NPT is the ensemble of choice when the correct pressure, volume, and densities are important in the simulation. Modifying the volume is done by a barostat, supplementing the MD simulation by adding an external variable. The first barostat for MD simulations was proposed by Andersen in 1980 [146], with various other barostats available since, such as the Parrinello-Rahman [147, 148] or Berendsen [142] barostat. The NPT ensemble can also be used during equilibration to achieve the desired temperature and pressure before changing to the constant-volume or constant-energy ensemble when data collection starts during the production run.

6. Production Run

Up to this point, the system underwent several steps to prepare for the final simulation, termed production run. The protein was placed in a simulation box under PBC, solvent was added and the saline concentration adjusted. Potential clashes were avoided by energy minimisation, and the temperature and pressure was adjusted to the respective required values. Now, the system is simulated for a pre-defined length, often in a NVT ensemble to account for the correct temperature during the simulation, and data about the dynamics of the system is collected and saved in a specific output [141]. Usually, the MD data is written to a trajectory file, describing all the atomic motion within the system or selected parts of it, and other files, *i.e.* containing all information of the energies. Afterwards, the obtained trajectories are further processed and analysed to extract the underlying information and details about the time-evolved dynamics of the systems, which is described in detail in the next section.

3.2.2 Analysing Molecular Dynamics Simulations

Over the course of an MD simulation, data is obtained describing the time-evolution of a pre-defined system. In the most cases, all information about the simulated system is collected and stored in a trajectory file, including data about the coordinates, velocities and forces of the involved atoms. Naturally, each simulation is computed to gain answers to questions about the investigated system, such as *How similar are the structures to one another?* or *How much does the C-terminus fluctuate?* The obtained trajectories are thus further analysed using selected methods and tools, providing the essential answers to these questions, with the most frequently used methods for analysis conducted for this thesis presented here.

Root Mean Square Deviation

The evolution of molecular structure over time essentially goes hand-in-hand with the structure changing its conformation as it 'evolves'. As such, a structure at time point x would differ from the structure at time point $x+1$, which

can be estimated by computing the root mean square deviation (RMSD). The RMSD is an analysis method used to gain information of how a specific structure deviates from a reference, and to what extent. Given a structure of N atoms in conformation $\vec{r}$, the RMSD can be calculated as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i - \vec{r}_i^{\,ref})^2}, \tag{3.16}$$

where $\vec{r}^{\,ref}$ describes the conformation of the reference structure to compare to. The RMSD is further computed for each i-th atom, and as such able to provide basic information if a system has reached equilibrium, where major structural changes are indicated by a flattening of the RMSD trend, or if substantial conformational changes are occurring during the simulation, which would be displayed by an increase of the deviation.

Root Mean Square Fluctuation

Whilst major conformational changes are able to be detected by the RMSD, it lacks to provide detailed information about what specific areas of the simulated structure are subject or cause of this observation. A highly dynamic part of a protein, as an example, would therefore be reflected by an increase in fluctuation around the positions of the respective atoms. This can be calculated by the root mean square fluctuation (RMSF), which indicates the divergence of an atom or a selection of atoms, such as residues, from their average positions during an MD simulation. For an i-th atom at its position $\vec{r}$, the RMSF is computed as

$$RMSF = \sqrt{\langle(\vec{r}_i - \langle\vec{r}_i\rangle)^2\rangle}, \tag{3.17}$$

with $\langle r\rangle$ reflecting its average position. As such, the RMSF effectively describes the standard deviation of the position of atom i. A large RMSF value would consequently indicate a high divergence from the average position, suggesting increased mobility in that area of the simulated molecule. This allows one to screen the flexibility of residues in a protein, providing useful and more detailed information about the underlying protein dynamics.

Theoretical Collision Cross Section

Theoretical models are often a crucial step of a scientific investigation, providing the foundation of experiments to be conducted and predicting their potential results. Nevertheless, theoretical models, especially those provided by MD simulations, can further be utilised to explain, test and complement already existing experimental data. The collision cross section (CCS) of proteins, provided by IM-MS experiments, is an example of an experimental observation that can conveniently be compared with a theoretical computed CCS

obtained from structures, which in turn can be obtained from MD simulations. The CCS itself describes the area of a molecule, orientation-averaged over all directions, that buffer gas in the drift tube of a IM-MS instrument is able to interact and collide with. Thus, the CCS is a measurement of the shape of a molecule during the experiment projected in 2D; or in layman's terms: the CCS describes the average projected shadow of a molecule in all directions [149]. For instance, the CCS can further be used to track conformational changes of proteins during IM-MS experiments, often seen in CIU experiments, where the impact between molecule and the buffer gas leads to the unfolding of specific parts of the protein structure. Structures and trajectories obtained from MD simulations can be fed into specific software in order to calculate their theoretical CCS values, of which the Ion Mobility Projection Approximation Calculation Tool (IMPACT) software was exclusively used for all CCS calculations presented in this thesis [150].

Atomic Distance and Contact Maps

The clear advantage of MD simulations is given by the provided atomistic information in a level of detail that is often inaccessible for other methods, being able to pinpoint, follow and visualise the motion of atomic positions over time. As the atoms change their position at each time step during the simulation, so will the distances between them, which presents a great opportunity for analysis; distance and contact maps. Calculating the pairwise distances between all atoms, one can express the 3D structural information of a protein or other molecule as a 2D matrix. Moreover, as a single molecular structure can be interpreted as a single distance map, consequently, a complete trajectory can be expressed in that way in order to visualise conformational changes in detail as they occur during the simulation. However, calculating and presenting all atom-atom distances includes displaying distances that might not possess crucial structural meaning. Whilst this is certainly not the case for protein unfolding studies, where especially the longer distances are of great meaning, potentially indicating unfolding to take place, one can further extract valuable information from a distance map by applying a pre-defined threshold. Screening all distances, whilst accounting for the upper limit of hydrogen bonds by applying a threshold of 3.5 Å, allows one to define and express the structural information as a contact map. Here, two residues are defined to be in contact, when the distance of at least one pair of atoms between them is equal or below the threshold of 3.5 Å. Consequently, all distances above that threshold are unaccounted for, further funnelling and visualising the information with the focus on specific points of interest in close proximity to each other.

3.3 The Influence of the Simulation Environment

H_2O contributes essentially to the natural environment of proteins, influencing protein structure, stability and function [2, 139]. Naturally, protein MD simulations are typically performed in an aqueous solution. However, rebelling against the mainstream use of MD, particular research questions necessitate simulations *in vacuo*, which further requires a different approach to the simulations compared to those in solution.

Described in equation 3.12, the electrostatic interaction between to atoms is calculated using the Coulomb potential. Equation 3.12 can further be written as

$$V_{Coulomb} = \sum_{ij} \frac{q_i q_j}{4\pi\varepsilon\varepsilon_0 r_{ij}}, \tag{3.18}$$

accounting as well for the dielectric constant ε of vacuum. The dielectric constant or relative permittivity of the environment is defined as the ratio of the permittivity of a substance to ε_0, the permittivity of free space [151]. For vacuum, the dielectric constant is consequently 1, thus the expression for the Coulomb potential as shown in equation 3.12. As can be seen however in equation 3.18 including the dielectric constant ε, the Coulomb potential is inversely proportional to ε—an increase of ε equals a decrease of the Coulomb potential. H_2O has a permittivity of approximately 80 at room temperature [152], which means that the electrostatic interaction between two atoms in an aqueous environment is greatly reduced compared to in vacuum. How does that influence MD simulations? Long-range interactions become consecutively less strong in general, whilst counting towards the computation workload of the simulation. Therefore, cut-offs are implemented for the interaction calculation, a radius around an atom within which all long-range interactions are calculated as usual, but approximated outside the cut-off [153–155]. For gas-phase simulations however, no cut-off is applied, and all interactions are calculated. Consequently, the workload of an MD simulation scales with the number of particles N in the system squared, as all atom-atom pairs need to be calculated. This means, that the overall time to simulate in the gas phase might be longer than in solution, whilst further potentially requiring more computational resources.

Furthermore, as the Coulomb potential calculations change noticeable for gas-phase MD simulations, the steps taken to the final production run of the simulation differ as well compared to an MS simulation in solution. In contrast to the workflow presented in 3.2.1, *Towards obtaining Molecular Dynamics data*, and shown in figure 3.3, specific steps are skipped or adjusted for gas-phase simulations. The initial structural information input step remains the same, as well as the building of the topology of the system. In order to simulated ideal vacuum conditions however, no PBC conditions are applied, and as such the volume, density and pressure are not defined. For a better visualisation,

the simulation box can be envisioned as infinitely large, and thus preventing periodicity. The energy of the system is minimised, and the temperature is controlled via a thermostat to a target temperature value. However, the temperature coupling simulation is therefore distinctly shorter. As the temperature is modulated by adjusting the energy of the system, in vacuum, energy transfer to the environment is not given. Ergo, infusing the system with energy for a longer period of time could therefore eventually lead to unintended and unexpected artefacts, rendering the simulation data incorrect. Still, the gas-phase MD data is collected and obtained during the final production run, generating results to investigate the vacuum dynamics of proteins.

4. *Complementing Methods*

My research focused on examining the dynamics of protein structures using MD simulation, predominately under vacuum exposure. However, as is common practice in science, we utilised and pulled resources and information from other methods to complement our data and placed our findings within the broader scientific context. For the most part, we were working together with MS, a powerful method to study gaseous proteins and their complexes. MD and MS possess a complementing relationship, supporting each other with information that is inaccessible or scarcely sourced for one of them [9]. Moreover, the field of protein research was shaken in recent years with the publishing of AlphaFold and similar (*e.g.* RoseTTAFold) neural network-based protein structure-prediction methods able to provide high-confidence structures from their respective sequence alone [156]. Those methods therefore enable MD simulations of proteins for which no structures were available before.

In this chapter, a basic introduction of both MS and AlphaFold is given for the reader with the aim to provide an understanding of the impact and importance of these methods for the here presented research.

4.1 Proteins Put on the Molecular Scale

The early history of MS was dominated by physics, exerting fundamental aspects of atoms [157, 158]. Accredited foundation of the technique was laid by J. J. Thomson and his work on the measuring the mass and charge of the electron, for which he received the Nobel prize in 1906 [159]. The principle of an MS experiment can be summed up over three crucial steps [160]. First, the analyte, the substance to be analysed, must be brought into the gas phase by an ion source. The thus created ions are eventually exposed to an electric or magnetic field, which separates the ions by their respective mass. Basically, depending on the analyte's charge, the more or less it interacts with the applied field, which consequently affects the ions' flight trajectory. The thus induced separation results in the ions arriving at different time points at the detector, which is eventually responsible for the the actual measurement and data recording. Interestingly, MS does not measure the mass *per se*, but rather determines the mass-to-charge ratio, m/z. The obtained data is plotted as the abundance or intensity of the investigated ions over their corresponding m/z, and thus provides fundamental information about the ions' molecular weight and composition [7].

Mass Spectrometry Outside Structural Biology

In the context of this thesis, discussing MS is linked with respect to the research field of structural biology. However, other research areas steadily employ MS as well to provide answers to their distinct question. Astronomers for example use MS to analyse interstellar dust and meteor grains to gather details about their composition and origin [161, 162]. As an example, the MS instrument onboard the Cassini spacecraft was proven to be essential for collecting evidence about potential hydrothermal activities on Saturn's satellite Enceladus [163, 164]. Other areas ranging from petroleum chemistry [165, 166] to cancer research [167, 168] and beyond [169] benefit greatly from employing MS. Additionally, MS finds more and more interest aside scientific research. Progress in the development of field-able MS instruments is further responsible for an increasing usage of MS outside of science, with particular interests for forensic investigations and security applications [170, 171].

4.1.1 A Gentle Transmission into Vacuum

Essential for any MS experiment is to convert the analyte molecules into the gas phase as ions. Moreover, the choice of the ionisation method alone can impact the experiment and its outcomes [172]. Early methods were certainly able to produce ions, however fragmenting larger molecules in the process and thus precluding the measurement of intact compounds [160, 172]. *Intact* is the keyword here, especially for proteins. It wasn't until the development of two soft ionisation techniques, matrix assisted laser desorption/ionisation and ESI, that measurements of larger, intact biomolecules came within the scope of MS [173, 174]. We will be putting the focus here onto ESI.

During ESI, an electric potential of several kilovolts is applied to a volatile solution resulting in the electrostatic spraying of the liquid from the tip of the injection capillary, which ultimately generates charged droplets and aerosols [175, 176]. Due to the vacuum exposure within the MS instrument, residual solvent molecules are eventually separated from the analytes. The process of isolating the ion from remaining solvent can follow different mechanisms, all resulting in leaving only the ionised intact analyte [177]. Most importantly, the applied conditions during ESI experiments allow for proteins to retain their higher order protein structures. This means the investigated proteins maintain their inter- and intramolecular interactions when transitioning to the gas phase, keeping their native structure as much as possible [178]. Hence, ESI-MS applying the correct buffer and experimental conditions were fittingly termed native MS [179,180]. Native MS accelerated the studying of biological

macromolecules and provided high-resolution information about their compositions, allowing to study *i.e.* protein complex formation and protein-ligand interactions in a detail that was not feasible before [181].

4.1.2 Coupling Mass and Mobility Measurements

As stated above, MS separates ions based on their mass and effectively measures m/z in order to derive their respective molecular weights. MS was eventually extended by accounting for the mobility of the ionised molecules, further enhancing the resolution of the obtain structural information [182, 183] (figure 4.1). Post-ionisation, ions are subjected to a instrumental chamber filled with inert gas (*e.g.* N_2). An applied electric field provides control over the velocity of the ions, which ultimately collide with the gas molecules as they migrate through the chamber [184]. Based on their size and shape, larger ions are exposed to a higher number of collisions with the inert gas, and are slowed down as a consequence [185]. Smaller ions on the other hand migrate more quickly, as they provide less surface for an effective collision to occur [186]. Ions are thus separated based on their mobility, a method that can efficiently be combined with MS-based mass separation [187]. The drift time describes here the duration for a species of ions to traverse the IM chamber, and the inferred CCS describes the orientally averaged projected area of the analyte molecule able to be subjected to buffer gas-collision [183] (see subchapter *3.2.2*). IM-MS CCS-data therefore allows for an estimation of the overall 3D protein or protein complex structure [149]. As an example, intermediates of the assembly process of the norovirus capsid were able to be probed and detected distinctively using native IM-MS [98].

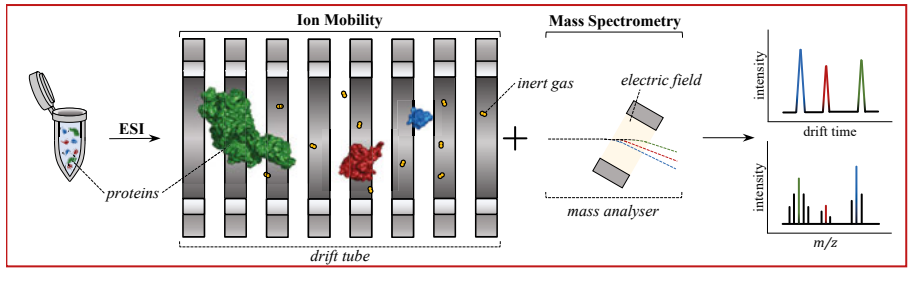

Figure 4.1. **Combining IM with MS.** After ESI, protein ions collide with inert gas molecules within the drift tube, resulting in the separation of the ions dependent on their size, shape and charge. Eventually, m/z is measured for the mobility-separated ions at the mass analyser and detected. The combination of IM with MS thus allows for the recording of high-resolution information of even low-abundance species.

Moreover, by increasing the collision energy controlled by the applied electric field, manifested as decreased IM, researchers actively initiated CIU to examine protein structures and their stability [188, 189]. Larger, more energetic

collisions of the analyte ions with inert gas molecules induce heat stress to the molecule, increasing their internal energies, eventually resulting in the unfolding of parts of their respective structures [190]. The unfolding directly affects the measured CCS, and allows for the generation of unique CIU fingerprints by plotting the drift time or CCS over the applied collision voltages [191], as shown for an example in figure 4.2. The pattern of the fingerprint therefore provides useful information about the underlying unfolding pathways that are often characteristic for the investigated protein [190, 192, 193].

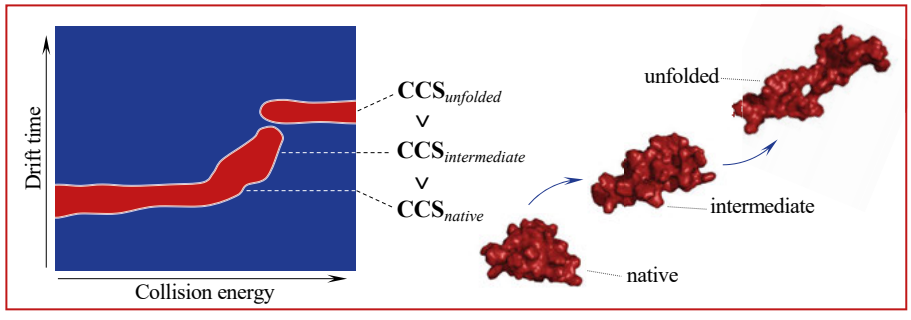

Figure 4.2. **Inducing the unfolding of proteins.** CIU of proteins is a potent method to investigate the structure and stability of proteins. Here, an example is shown of the unfolding of a protein. Initially in the native conformation, increasingly energetic collision with the drift gas molecules result in the unfolding of the protein structure passing through intermediate states.

4.1.3 Complementing Theory with Experiments, and *vice versa*

With native MS, or in combination with IM, detailed information of the tertiary and quaternary structures of proteins can be extracted, often for multiple species over the course of a single experiment. Atomistic details about the 3D structure and dynamics of proteins are however only scarcely interrogated by MS. MD simulations on the other hand provide high-resolution atomistic information of the investigated proteins, yet are restricted by the limit of computational resources to match experimental time scales. As such, both methods present possibilities that benefit the other [9].

Combining MD and MS can be done in several ways. On the one hand, MD can function as interpreter for experimental findings derived from MS. As an example, certain areas of the capsid proteins of different Norovirus strains were seen to show distinct deuterium-uptakes during hydrogen-deuterium exchange MS (HDX-MS) measurements [194]. HDX-MS experiments investigate the exchange of backbone hydrogens with their heavier isotope, and can therefore provide information about regions with large solvent-exposed surfaces [195]. The absence of a fucose ligand saw an increase in the deuterium-uptake for a specific virus strain. This was confirmed by MD simulations,

as residues that are normally interacting with fucose via hydrogen bonds are demonstrating large dynamics in the absence of the ligand. Moreover, **Paper IV** investigated the infolding of β-barrel proteins, where CCS and CIU data was taken from literature to inform our simulations [196]. Apart from MD interpreting MS data, other approaches can see MS informing MD simulations about which parts the protein could be of interest for investigation, which parameters to apply, or which charge states a protein should possess in the gas phase [9]. Charge state information from MS experiments was used for the studies presented in **Paper I** and **Paper III**, where protein net charges for gas-phase simulations was adjusted to those obtained from MS.

Together, MD and MS make a good team; providing each other with information that is scarcely or not available for the other. This thesis proves this, as we frequently employed MS-derived data for our MD simulations. Our research therefore attests to the complementary relationship of MD and MS.

4.1.4 The MS SPIDOC Project

A further promising technique with the purpose of investigating the structures of particles is single particle imaging (SPI) [197]. As the name suggests, rather than measuring multiple particles at the same time, as is commonly practice for NMR or X-ray crystallography, for SPI, a single particle is isolated and measured [198]. SPI uses ultrashort pulses of X-rays that interact with the isolated particle, allowing for high-resolution information about the particle's structure. Due to the intense energy of the X-ray pulses, the particle rapidly gets destroyed and plasmafies, however not before a snapshot of its structure can be taken in the form of a diffraction pattern [199]. High-quality structural information was obtained in a case study of SPI of the Mimivirus, providing convincing evidence of the potential of SPI for structural biology [200]. Reconstructing the 3D structure from collected diffraction patterns is however immensely challenging, as vast numbers of patterns are required to reconstruct a meaningful 3D structure from them. Contributing to this is the random and unknown orientation of the imaged particles, a problem that has been recognised by the SPI community [201].

Specific proteins, as many other molecules, possess a permanent intrinsic dipole. Manipulation of the orientation of their dipole would therefore allow for the manipulation of the orientation of the complete structure. In 2017, Marklund *et al.* proposed a solution to the orientation problem for SPI, based on research published more than three decades ago [202]—orienting the intrinsic dipole of proteins by applying an external electric field [203]. The external electric field is key player here, allowing for the manipulation of the dipole moment orientation to affectively align it along the electric field direction. However, the applied field strength is further highly important; too weak, and the dipole orientation will not be affected resulting in no alignment, whilst

extreme electric fields might be too intense and induce irreversible structural changes, up to unfolding. Using MD, smaller proteins were dipole-oriented in gas-phase simulations under specific electric field strengths, proving that dipole-orientation is indeed possible and controllable for gaseous proteins. Moreover, **Paper II** investigated if an applied external electric field would induce long-lasting structural changes to the proteins, and if those are reversible by means of post-vacuum rehydration.

To put the proposed dipole-orientation to the test in combination with native MS and SPI, the MS SPIDOC project was created [204]. The name is the game; MS SPIDOC is the abbreviation for *Mass Spectrometry for Single-Particle Imaging of Dipole Oriented Protein Complexes*. A multinational and -institutional project that aims to develop a novel instrument for (protein) sample delivery at the European X-ray free-electron laser facility (XFEL). Capable of mass-selecting IM-separated biomolecules via native MS, isolated particles are then dipole-oriented in a specifically designed module of the instrument, eventually followed by SPI of said proteins. Currently within the final steps of its completion, the MS SPIDOC sample delivery device for large-scale facility beamlines such at the European XFEL, has the potential to advance SPI and therefore impact structural biology for the better.

4.2 Input: Sequence – Output: Structure

One of the many mysteries of protein science that drives researchers for years was, and still is, to understand how proteins are able to fold rather quickly into their pre-defined spatial conformations, the latter being encoded in their primary structure—a question that got known as the protein folding problem [43, 205]. From the sheer number of possible folds a polypeptide chain can assume, nature follows its own folding-template in order to obtain the right structure for the right job. Being able to accurately and confidently predict the 3D structure of any amino acid sequence possesses the potential to impact various research areas, ranging from biology to medicine [44, 206].

Experiments employing techniques such as NMR or X-ray crystallography are certainly able to provide amongst the most accurate structural data, yet are often enough linked to high financial costs and being labour-intensive [207]. As a consequence, the gap between determined protein sequences and experimentally-solved corresponding structures steadily increased, which led more and more researches towards employing computational methods for protein structure prediction [208]. Developing convenient and accurate computational approaches for protein modeling was however linked to various challenges [43, 209]. To put this into perspective, let us go back to 2012 and have a look at a small paragraph published by Dill and MacCallum [43]:

A grand challenge has been to develop a computer algorithm that can predict a protein's 3D native structure from its amino acid sequence. On the one hand, knowledge of native structures is a starting point for understanding biological mechanisms and for discovering drugs that can inhibit or activate those proteins. On the other hand, we know 1000-fold more sequences than structures, and this gap is growing because of developments in high throughput sequencing. So, there is considerable value in methods that could accurately predict structures from sequences.

Ken A. Dill and Justin L. MacCallum, **2012**.

With many methods and algorithms being developed and tested over the past decades, certainly noteworthy is the of Critical Assessment of Structure Prediction (CASP) community experiment, which sought out to monitor and evaluate advances of protein structure prediction as biennial competition between various research groups since 1994 [44, 207, 210]. The progress of modern technology, especially that of artificial intelligence, drew the attention of researchers and increased the popularity of training and using neural networks for protein structure prediction [211–213]. The 14^{th} instalment of CASP competition, held from May through August 2020, eventually saw a new algorithm out-performing all other contestants, namely the deep learning algorithm developed by DeepMind: AlphaFold [214]. Trained on experimental data obtained from the PDB, AlphaFold displayed unprecedented accuracy and speed for predicting and generating protein structures from their respective sequences [215, 216]. As of then, AlphaFold was made public and has since been used widely and extensively by the scientific community, eventually leading to deposition of more than 214 million[1] AlphaFold-predicted protein structures in a new-established database [217, 218]. Whilst the protein folding problem might not have been solved with the development and publication of AlphaFold, it is said to herald a new chapter and shift in structural biology, and might just have been the computer algorithm that was foretold by by Dill and MacCallum [43, 219–221].

[1]Obtained from the AlphaFold website, available at https://alphafold.ebi.ac.uk/faq. Accessed on the 19^{th} of April 2023, 6:13 pm.

5. *Summary of the Scientific Work*

The research presented in this thesis has the overarching theme of employing MD simulations in order to investigate the dynamics of protein structures under pre-defined conditions. Concluded over four scientific papers, the conducted scientific work can further be divided into two sub-themes, specified as follows:

- **Rehydrating Vacuum-exposed Protein Structures – Paper I-II**

  Several structural biology methods expose proteins to vacuum, along with potential lasting effects on its conformation [1, 222]. Computational rehydration was investigated to understand the dynamics during rehydration, and to examine if and how vacuum-induced conformational changes might be reverted. At first, we explored rehydration simulations using both capsid proteins of the bacteriophage MS2 (bMS2) virus (**Paper I**), and expanded our investigations to probe how dipole-oriented protein structures react to computational rewetting (**Paper II**).

- **Simulating Collision-induced Unfolding of Gas-phase Proteins – Paper III-IV**

  Inducing the structural unfolding of proteins in IM experiments is used to interrogate the overall structure of proteins, their arrangement and stability [190, 223]. With the unfolding being experimentally initiated by increasing the collision energy, we simulated CIU thermally over separated steps of increasing temperature. Proteins under study were norovirus dimers of the Norwalk and Kawasaki strain (**Paper III**), as well as β-barrel proteins to compare with experimental CIU data (**Paper IV**).

The aforementioned publications are summarised in this chapter, presenting the most important findings and insights of the conducted research. For more details, the complete research papers are attached.

5.1 Rehydrating Vacuum-exposed Protein Structures

The amino acid sequence of a protein dictates the structure it will fold into [2, 3], and the structure in turn is closely connected to the function of the protein, as explained in Chapter 1, *Proteins*. Recent structural prediction methods such as AlphaFold allow to acquire protein structures in high confidence directly from their sequences, whilst other experimental methods are long established as foundation for determining the conformation of proteins, with X-ray crystallography and NMR spectroscopy being amongst the most popular ones [44, 215, 224–229].

Native MS especially allows for probing and examining the structural aspects of gaseous proteins, where detailed information about the higher-order of protein structures can be obtained [230]. Here, under suitable conditions, proteins were shown to retain structures that are close to their native conformation [1, 231]. Nevertheless, taking proteins out of their natural environment into vacuum might still affect their structures [232, 233]. Consequently, as some experimental techniques and methods obtain structural information from vacuum-exposed proteins, the thus 'solved', investigated structures and obtained information might contain structural artefacts that differ from their native state.

The rehydration of previously gaseous protein structures was shown to yield conformations closer to the native norm, employing MD simulations to do so [234]. As such, we conducted investigations into how dimeric protein structures behave during computational rehydration, and to what extent any vacuum-induced changes to the proteins are reverted (**Paper I**). Moreover, we expanded our research into rehydration simulations to include structures of dipole-oriented proteins. In 2017, Marklund *et al.* showed that the dipole of proteins can be controlled by an external electric field (EF) [203]. The applied strength of the EF was shown to play an important role not only for the speed and accuracy of the dipole orientation, but further for the protein structure itself. Too high of an EF strength led to proteins losing their structure, eventually inducing the unfolding of their polypeptide chains. Ergo, we examined if specific EF strengths, that were shown to induce dipole orientation without unfolding, might have affected the protein structures to an extent where they eventually end up with differing structures after rehydration (**Paper II**).

5.1.1 Paper I: Rehydrating Bacteriophage MS2 Dimers

In this paper, we conducted extensive MD simulations of both conformations of the bMS2 coat proteins. Within the capsid of the bMS2 virus, the dimeric building blocks adopt two distinctive conformations: the symmetric or C/C conformation, and the asymmetric A/B conformation [235]. The monomeric chains in conformation A and C show a well-defined, extended β-hairpin between the β-strands F and G, which in contrast is folded onto the main protein

body in the B conformation. Without the genome, the dimers exists predominately in the symmetric conformation, whereas binding to the viral genome facilitates the formation of the A/B dimer [236, 237]. Hence, in absence of the genome, only symmetric dimers should be present. The motivation for this study included how the dimers would behave in the gas phase, and how the structures would subsequently respond to rewetting. This makes the bMS2 virus an excellent model system to test the rehydration of multimeric proteins. Moreover, we were thus able to probe a potential conformational memory, that is, if the dimers would find back to their respective native conformations by means of rehydration.

Simulated initially for 750 ns in solution over ten replicas, a total of 200 structures for each dimer was taken into vacuum, where the net charge of the proteins was adjusted to reflect those published for MS experiments [238]. Subsequently, 200 simulations of 500 ns duration each were started. The thus obtained final structures were then rehydrated, the charge again adjusted to a physiological pH, and simulated as well for 500 ns.

The obtained MD data of the dimers was analysed using several methods in order to extract information about the underlying dynamics during rehydration in comparison with those of the bulk solution and vacuum simulations. The initial RMSD computations show an increasing trend over the course of the rehydration simulations, suggesting that the protein structures are equilibrating towards the solution environment, yet display clear deviations compared to the structures extracted from the bulk solution simulations. Observations of the residue fluctuations during the vacuum and subsequent rehydration simulations revealed similar RMSF values throughout both sets of simulations, where the highest fluctuations are linked to residues forming the extended β-hairpin in the A and C chains. Vacuum exposure can lead proteins and macromolecules to compact in their structure due to the hydrophobic nature of vacuum [232, 233]. Consequently, we calculated the CCS, the total surface area and volume of the proteins to understand the specifics of vacuum exposure for the bMS2 dimers, but as well how these properties might revert back to their norm. Calculating the theoretical CCS from MD simulations further provides an analysis that can be directly compared with experimentally obtained CCS values. The CCS for the dimers in vacuum show a decreasing trend over the 500 ns of simulation, where the respective structures compact in CCS of approximately 7 to 9 %, for the A/B and C/C dimer, as compared to their average CCS during the initial bulk simulations. The rehydration of the vacuum structures on the other hand proves an almost complete recovery of the CCS, a trend which is also shown for the calculations of the total surface area and volume. To delve deeper into the conformations of the bMS2 dimers in the interest of extracting further valuable information, we calculated the distances between all atoms and applied a threshold of 3.5 Å. Thereby filtering for those residues that are in close proximity to each other, we defined

residues within 3.5 Å distance to be in contact of each other, and averaged the thus obtained data over the last 50 ns of all conducted simulations. The results reveal that more contacts are forming due to the vacuum exposure, which are for the most part reverted upon rehydration, with the rehydration contact maps patterns resembling those provided by the initial bulk solution simulations. However, these patterns are not identical, which further suggests that specific differences between the rehydration and initial bulk simulations still persist after 500 ns of rewetting. Regardless, we do observe the proteins to obtain some of their conformational memory, as no crossing between the asymmetric to symmetric conformation can be seen after their rehydration.

5.1.2 Paper II: Rewetting Dipole-oriented Proteins

Continuing on examining the possibilities of rehydration simulations, this work expanded our rehydration studies to include gas-phase proteins which were simulated under the influence of an EF for the purpose if dipole-orientation. Vacuum simulation data of the following proteins was taken from the study of Marklund *et al.*: tryptophan cage, the C-terminal fragment of the L7/L12 ribosomal protein, ubiquitin and lysozyme [203]. Whilst Marklund *et al.* looked at various different EF strengths, the here presented study focused on three data sets comprising non-dipole-oriented (0.0 V/nm) protein simulations, and those that were simulated with an external EF of 0.2 V/nm and 0.4 V/nm strength. Consequently, extending the motivation for **Paper I**, we further include investigating potential long-lasting structural effects on the protein structures induced by an applied EF.

The four proteins were rehydrated over five replicas, each for 200 ns, in an isobaric, isothermal ensemble. In order to obtain a data set for direct comparison, which would reflect the norm of the individual proteins in an solution environment, control solution simulations were conducted with input structures obtained from the corresponding PDB entries of the four proteins, where the MD protocol was kept the same as for rehydration.

The structures of the proteins were compared for each data set, as in derived from control solution, vacuum and rehydration data, by calculating the RMSD. As reference, the initial structure of every solution simulation was used, as well as the final structures of the control solution and the zero-field (0.0 V/nm) data. The proteins display an increasing RMSD trend compared to their initial starting structure during rehydration, a similar trend that can be observed over the control solution simulations as well, however, with higher deviations during rehydration. This indicates that larger structural rearrangements might be occurring during the rehydration process, likely linked to a reversion of the compaction of the protein structure in the gas phase, whilst the proteins are equilibrating to the solvent environment. The movement of the residues was further examined by RMSF calculations, computed for the rehydration

and control solution data sets. The results show obvious larger RMSF values of the residues during the rehydration simulations than during the control simulations. As the structures equilibrated to the solution environment, the large fluctuations are most likely showing additional structural drifts occurring. Moreover, as the fluctuations themselves are occurring around the same residues and parts of the proteins throughout all data sets, this could imply that the rehydration and control simulations of the proteins follow similar underlying dynamics regardless of an applied external EF. The analysis, comprising calculations of the CCS, solvent accessible surface area and total volume, of the simulation data suggests further that rehydration of 200 ns allows the structures to revert closer to their solution average. Averaged over the last 50 ns, the results of the analysis suggest that all data sets, with applied EF and in absence of it, display similar values and therefore comparable dynamics that revert vacuum-induced structural changes. Furthermore, no clear evidence seems to exist of irreversible EF-induced structural changes. Eventually, contact maps were calculated for each data set and each simulation, where the last 50 ns of the rehydration and control solution simulations were averaged. To emphasise on the similarities and dissimilarities of the contact map patterns throughout all data sets, contact data from the vacuum structures and the control simulations was subtracted from the rehydration contact values. The results further suggest that more contacts exist during vacuum, as a consequence of the vacuum compaction, which do not exist as much during rehydration, implying a reversion of aforementioned compaction. Whilst differences in the subtraction contact maps between rehydration and control solution data exist as well, the occupancy of these persisting contacts is lower as shown for the subtraction maps between rehydration and vacuum, indicating again structural similarities between rehydration and the control.

5.2 Simulating Collision-induced Unfolding of Gas-phase Proteins

Where MD shines to provide atomistic information, yet restricted by time scales able to be probed, MS enables the investigation of the tertiary and quaternary structures of proteins and protein complexes at the cost of resolution in terms of atomistic details [9]. The complementary relationship between MD and MS allows for a more detailed view into the structures of proteins, their respective dynamics and stability. For example, MD can predict certain pathways of structures to follow under pre-defined conditions to inform subsequent MS experiments, or MD can be employed investigating the atomic details of MS-derived observations.

A powerful method to study protein stability is given by CIU [190]. Prior to IM-MS, the velocity of the protein molecules is steadily ramped up, resulting in increasing collision energies between the analyte and the inert gas

molecules [188]. The energy eventually leads to the unfolding of specific protein areas, and can be experimentally tracked by the measurement of their respective CCS values. The obtained experimental unfolding pattern may indicate occurring structural shifts, and, based on that, changes in protein stability stability [189].

MD has been employed to probe the atomistic details of CIU [232, 239]. Whilst the collision energy cannot be put as input parameter *per se*, the common practice to induced unfolding is to simulate at increasing temperatures [239]. CIU at higher collision energies induces heat stress to the protein, ultimately leading to unfolding, thus presenting MD simulations at larger temperatures as reasonable approach for inducing *in silico* unfolding. We aimed to build on these studies, and examine the use of MD for CIU-simulation and what information we can extract that could help to guide or explain experimental CIU data. In **Paper III**, we simulated the capsid proteins of two norovirus strains, Norwalk and Kawasaki, and interrogated their response at 300 to 900 K. The two investigated dimers show an overall similar shape, yet only possess a sequence identity of 46.60 % between them. Thus, comparing similarities and dissimilarities of their unfolding behaviour could present interesting results to inform future experimental CIU studies of these dimers. For **Paper IV**, we obtained CIU data from collaborators who unfolded the transmembrane β-barrel structure of the FhaC proteins [196]. The unfolding of FhaC provides interesting insights into its stability, and more so could inform about its function and behaviour as transmembrane transporter protein in Gram-negative bacteria. Hence, we proposed MD simulations as tool to investigate the *in silico* unfolding of FhaC and complement our simulation data with information obtained from Sicoli *et al.* [196].

5.2.1 Paper III: Simulating Thermal Unfolding of Norovirus Dimers

CIU studies of proteins can provide valuable insights into the stability of protein structures [190]. Here, we investigated the dimeric capsid proteins of two strains of the norovirus, namely GI.1 Norwalk (NS) and GII.17 Kawasaki (KS). With a large number of gastroenteritis cases, a severe inflammation of the gastrointestinal tract, being linked to an infection by noroviruses, understanding the stability and dynamics of norovirus proteins may aid in treating gastroenteritis [240]. The structures of norovirus dimers is generally subdivided into two domains, namely the shell (S) domain, which is located on the inside of the full capsid, and the protruding (P) domain, extending outwards from the main capsid body and facilitating cell attachment during the viral infection [241]. The P-domain is commonly further divided into the P1- and P2 subdomain. We simulated both dimers in sets at various temperatures to induce their unfolding, and aimed to track the underlying dynamics during the

simulations in order to define potential similarities and dissimilarities of their unfolding.

Structural information of the NS dimer was obtained from its PDB entry under 1IHM [242], with missing atoms added utulising the UCSF Chimera tool [243]. AlphaFold2 was used to predict the structure of KS from its amino acid sequence, as no dimeric structure was available [215, 244]. Both dimers were initially simulated in solution to acquire equilibrated starting structures, and then taken into vacuum to induce thermal unfolding. This was realised by simulating seven replicas per dimer at temperatures of $\{300, 400, ..., 900\}$ K.

At first, we calculated the time-evolution of the average CCS between the replicas at each temperature. The CCS is an important experimental observation, as an increasing trend most often points towards unfolding to occur during CIU experiments [190]. Moreover, the CCS can be calculated straight from MD simulations, allowing for a direct comparison between MD- and MS-derived CIU data. Up to 600 K, the CCS shows no major changes for both dimers. Larger temperatures of 700+ K however showed an increasing trend for NS dimers, whereas 800 K seems to be the threshold temperature to induce unfolding. To investigate this in more detail, and concentrate the MD information, we defined six time slots, each spanning 5 ns of simulation time, in order to represent states of the unfolding of the dimer proteins. The CCS was averaged per time slot, revealing that indeed at 700 K effectively enough to increase the CCS for NS by 10 % from time slots 1 to 6, and only for 1% for KS. At 800 and 900 K, the increase of CCS is given for both dimers to be at least 17 %, providing first evidence of induced unfolding. Each time slot was assigned a representative structure with the aim to visualise and define the unfolding states, which were then compared to their respective structure for the first 5 ns of simulation. Again, we saw that major structural deviations exist especially at larger temperatures and at later time in the simulations that would indicate unfolding to occur. However, details about the unfolding process are required to pinpoint which parts of the proteins are eventually first to unfold. For that reason, we calculated the residue-based RMSF, complemented with calculations of the residue-residue distances and contacts, averaged per time slot. The thus obtained information may provide evidence to eventually predict the unfolding process on an atomistic level. Where as the data suggest that temperatures up to 600 K show overall similar trends, both in RMSF and atomic distances, we further observed different parts of the proteins demonstrating larger dynamics at 700 K and higher. NS exhibits large fluctuations around the termini, especially for chain A, which is already extending at 600 K and fully unfolding at 700 K. Going higher in temperature, disturbances within the protein structures are more evident, as the RMSF values for each time slot are steadily increasing throughout the protein. Moreover, temperatures of 800 and 900 K are seemingly inducing unfolding so rapidly, that no coherent information may be extracted as to which states are visited step by

step by the NS dimers during unfolding. This becomes more evident when looking at the respective distance and contact maps, where large distances are already present in the beginning of the simulations, and the contact patterns become steadily more disrupted as time progresses. For KS at 700 K, the termini seem to be rather stable, as their RMSF values are overall low, and the distance and contact maps show no obvious unfolding to occur. Simulations at 800 K however reveal largely dynamic areas of the protein which are linked to the P2-domain, which is supported by the distance and contact patterns of the respective maps. In a recent study, we have seen that the same part of the protein showed large dynamical behaviour during MS experiments, which we confirmed by employing MD [194]. Furthermore, this indicates that unfolding seems to occur from different parts of the two dimers, and at different temperatures. However, as was seen for NS, 900 K induces severe unfolding, as the respective structures show and the RMSF and distance values suggest. Apart from the P2-domain unfolding, we saw as well unfolding to occur within he S-domain, yet seemingly in a less chaotic process as was observed for NS.

5.2.2 Paper IV: Complementing Experimental with Theoretical Unfolding Data

In this study, we used IM-MS data of the unfolding of β-barrel proteins such as FhaC to inform out MD simulations, and investigate their unfolding on an atomistic level. FhaC is a transporter protein from the bacterium *Bordetella pertussis*, the causative agent of the contagious respiratory disease pertussis, commonly known as whopping cough [245]. As transporter protein, FhaC is involved in moving designated molecules across the membrane and crucial part of the two-partner secretion pathway. The structure of FhaC composes 16 β-strands forming the membrane-spanning barrel-like structure, two polypeptide-transport-associated (POTRA) domains, and further an α-helix (H1) which is located inside the β-barrel in FhaC's native closed conformation [246]. H1 functions as a plug for the β-barrel, where it vacates the barrel and thus facilitates the binding of the substrate, and ultimately its transport across the membrane [247, 248]. As such, the structure of FhaC is likely to be highly dynamic in order to fulfil its transporter-function. CIU studies of FhaC revealed unfolding fingerprints showing two transitions, understood as originating from the unfolding of both POTRA domains [196]. Hence, we employed CIU data and MD simulations to understand the atomistic details of the unfolding of FhaC, how our simulation might relate to the experimental observations, and eventually how MD can aid in understanding the CIU dynamics of FhaC.

Structures of FhaC were obtained from experiments [196]. In order to obtain equilibrated structures from FhaC, at first, we placed the protein into a membrane of native composition using the CHARMM-GUI Membrane Builder

webtool [249–251], for a short simulation to relax the structure of FhaC. A total of ten replicas were extracted as starting structures for *in vacuo* simulations at temperatures $\{300, 400, ..., 900\}$ K with the aim to induce unfolding.

RMSD calculations of the FhaC structures revealed only minor deviations up to 600 K, which are steadily increasing at 700 K and higher. Computed RMSF trends at all temperatures show a similar behaviour, with residues demonstrating on average RMSF values below 10 Å up to 600 K. At higher temperatures, the residues are increasingly more dynamic, with peak values shown at 900 K, providing potential evidence for unfolding to occur. The theoretical CCS was calculated to track the likely unfolding further, which can moreover be directly compared to those obtained from the CIU experiments [196]. The time-evolution of the averaged CCS again provides evidence for unfolding to take place especially at temperatures 700+ K. Interestingly however, when comparing the theoretical CCS values at 900 K with those obtained from the CIU data from Sicoli *et al.*, both show a remarkable similar trend. To investigate this compelling observation further, we put focus on the 900 K MD data and used the information obtained from MD and IM-MS to define several unfolding states. From information taken from the CIU data, we can see two dominant transitions spanning three major regions, namely region 1 to 3. Moreover, region 2 exhibits two minor transitions, and was therefore subdivided into regions 2a, 2b and 2c. To account for an initial state of the protein after vacuum compaction, we added and additional region 0. The CCS data from the CIU experiments was used to set the CCS-limits for each region, whereas the simulation time-limits was defined based on the MD data. Thus, theoretical and experimental unfolding information was expressed over total of six regions, namely region 0, 1, 2a, 2b, 2c and 3. All structures from the MD simulations belonging to the same region were pooled together and clustered to find a single representative structure per region, and thus visualise the potential unfolding mechanism of FhaC at 900 K. Initially, for region 0, FhaC shows to possess a globular-like structure, with the POTRA domains seemingly being collapsed on the β-barrel. Regions 1, 2a and 2b interestingly provide evidence that the POTRA domains remain collapsed on the protein main body, whilst the C-terminus is exhibiting high dynamics and progressively unfolds. With the C-terminus forming the seam of the β-barrel between β-strand$_1$ and β-strand$_{16}$, the observation of the C-terminus unfolding would indicate a different unfolding mechanism as proposed by Sicoli *et al.* [196]. This is further supported by the representative structure for region 2c, which shows the polypeptide chain to extend from the C-terminus. Region 3 eventually substantiates the evidence provide for FhaC unfolding *in silico* starting from the C-terminus, as the representative structures shows an elongated part of the protein chain extending from the main protein structure.

To further complement the potential unfolding mechanism we calculated distance and contact maps, which allowed us to probe to probe the unfolding in an

atomistic detail. The distance and contact maps were generated for each structures belonging to the same region, and then averaged to present a single map per region. Unfolding is accompanied with an increase of the intramolecular distances between the individual atoms, which makes distance and contact maps and ideal tool to further obtain information about the underlying processes. Eventually, the observation derived from the representative structures per regions in further supported by the distance and contact maps, where an extension of the residue chain from the C-terminus, and thus from the seam of the β-barrel is clearly indicated. The here shown unfolding mechanism is different as proposed by Sicoli *et al.* [196], who predicted the large CCS values to originate from the unfolding of the POTRA domains, rather than starting from the C-terminus. As such, this study demonstrates a novel protocol on approaching MS-informed MD simulations for studying and probing the unfolding of β-barrel and other proteins.

6. *General Conclusion and Future Perspective*

Understanding proteins, their structure, function and dynamics, is immensely important for us. Be it to find patterns in virus-host interactions to influence the viral infection, probe the stability of membrane proteins and examine their conformational changes, or simply follow their dynamics over time under certain conditions—protein research is as multifaceted as it is interesting and necessary. A plethora of methods exist for investigating proteins, each with their own benefits and limitations, which makes it important to choose the right method for the right research task. Often enough, when studying proteins, changes occurring on the atomistic scale cascade into the larger protein dynamics we observe, which makes it all the more crucial to study these seemingly inconspicuous motions.

MD simulations provide the means to do exactly that; investigate protein structures and dynamics with a resolution and in detail that are inaccessible to other methods. Moreover, MD simulations are highly adaptive, and can be adjusted depending on the task at hand. Protein simulations in solution? No problem. Simulating protein embedded in a membrane bilayer? Sure. Transferring proteins from solution to vacuum? Easily done. Adjusting their net charge to mimic ESI whilst doing so? Of course. Simulate CIU using MD? Say no more. Rehydrate proteins *after* vacuum exposure? Yes. This thesis, and the conducted research it is based on, provide reasoning, procedure and impact of using MD to probe protein structures.

In **Paper I** and **Paper II**, we investigated proteins and their rehydration dynamics after vacuum exposure. As this could eventually help to further gain information about native protein structures, rehydration simulations are surely an interesting approach for structural biology and related research areas. As rehydration simulations were a rather unique approach for our research, we first had to find appropriate protein structures to motivate their rehydration. The two existing conformations A/B and C/C of the bacteriophage MS2 virus dimers provided an ideal set of proteins for **Paper I** to investigate rehydration. As the two dimers are predominately alike in structure, they differ around the hairpin between the β-strands F and G, which is extended in protein chains of A and C conformation, yet collapsed onto the main protein body for chain B. Analysing the extensively conducted MD simulations, we found that the majority of the structures do find back in high percentages when compared to their respective norm values in solution. Moreover, we expanded our rehydration studies for **Paper II** to examine the potential effect of an applied EF for

dipole-orientation on proteins. Compared to a control solution simulation, if the applied EF would induce irreversible conformational changes in the protein structures, we would be able to identify those when analysing the MD data. However, we saw that the effect of EF of 0.4 V/nm strength or lower are not inducing lasting effects on the respective protein structures whilst still able to induce dipole-orientation. However, our studies demonstrate as well that an identical conformation of the solution structure was not achieved by means of rehydration, yet still with high structural similarity. Here, especially analysing and comparing the MD data using contact maps was shown to be highly informative, as deviating parts of the protein are thus able to be identified and accurately pinpointed. Longer simulation time could further close the gap between the norm structure in solution and those obtained by rehydration, and improved sampling methods could allow for a broader evaluation of structures across the free energy landscape. Regardless, rehydration MD simulations possess great potential, be it for structural biology, complementing soft-landing experiments, or even stepping into the field of protein drug research, where the dry and wet state of proteins are important. We see rehydration simulations as promising extension of conventional MD simulations.

The second part of our research applied means of unfolding proteins using MD, and how we can relate our findings with CIU data. With CIU experiments providing information about protein stability structure, the detailed underlying unfolding mechanism might just be out of reach of conventional experimental approaches. We thus used MD not to simply induce unfolding, but further to study the effects of the temperature of the protein and its dynamics, and the potential prediction of the process of unfolding as a result of that. In **Paper III**, we approached simulating unfolding with MD for dimeric coat proteins of two Norovirus strains, Norwalk and Kawasaki. Unfolding these structures generally informs us of their stability, as in which parts of the protein might unfold first, and further if we can observe distinct differences between the two dimers. We saw that NS dimers seemed to be rather unstable from the N-termini, as they were shown to unfold first, whereas KS dimers exhibit unfolding occurring from regions of their respective P2-domains. Overall, KS was shown to be more stable than NS, as higher temperatures are needed to exhibit first instances of obvious unfolding as compared to NS. Full KS capsids were shown to be more stable in solution as their NS counterpart, which might indicate thus as well to a higher stability for KS on the dimer-level [241]. The conducted MD simulations thus provide information for future IM-MS experiments of the CIU of NS and KS. Moreover, we used published CIU data about the unfolding of the β-barrel transporter protein FhaC to inform our simulations and guide the reasoning of analysis our obtained MD data for **Paper IV**. FhaC is an interesting protein, not only because it is located in the membrane bilayer of Gram-negative *Bordetella pertussis* bacteria, but more so because its function as transporter protein requires a certain mobility. CIU experiments

conducted by Sicoli *et al.* revealed two major transitions occurring during the unfolding of FhaC, and proposed the unfolding of the POTRA domains as reason for this observation [196]. We thus aimed to simulate the FhaC-unfolding with MD, and found that analysing the MD data at 900 K revealed a striking similarity for our CCS calculations, and those provided by Sicoli *et al*. Hence, we used information from literature and our MD simulations to define potential unfolding states, and saw that the thus gathered data points towards the C-terminus of the β-barrel to unfold along its seam. Interestingly, this indicates a different unfolding mechanism as interpreted by Sicoli *et al.*, as the POTRA domains are shown to collapse onto the main protein body during our simulations, rather than unfold. The conducted MD simulations thus present previously inaccessible atomistic information and details of the unfolding dynamics of FhaC, which could be extended to other β-barrels and proteins in general as well. We further saw that proteins react differently to the applied temperature, which requires a certain screening to understand which temperature is eventually enough to induce unfolding. Moreover, the applied temperature further influences how rapid unfolding occurs, as we saw for example at 900 K for NS and KS. Here, 900 K simulations basically showed overall very chaotic dynamics and thus would indicate 900 K eventually being too high of a temperature to obtain meaningful information about the unfolding. Consequently, lower temperatures and longer time scales could thus be a better combination that high temperature and chaotic dynamics. Investigating this further might be an improvement to the in **Paper III** and **Paper IV** presented unfolding protocol, and thus aid implementing MD simulations as method for complementing CIU experiments.

One could further ask the question, could it be useful to combine CIU and rehydration MD simulations? Obviously, a fully unfolded protein might be to disordered to find back to any native-like structure, or would take a long time to do so. However, we saw in **Paper III** that certain temperatures are just enough to activate a small part of the protein, shown for the extension of certain areas within the NS or KS dimer. It would therefore be interesting to see how a 'mildly' unfolded protein structure might behave during rewetting, and what we could learn from this.

In the end, MD provides a theoretical view on the atomistic dynamics of proteins, yet ultimately needs to be compared to or evaluated by experimental means. The MS SPIDOC project is an excellent example of how impactful an idea based on MD simulations can be, ultimately manifesting idea to become reality with the development of a completely new instrument for SPI-sample delivery. Together with the research presented in this thesis, we show how impactful MD simulations can be for the estimation, prediction, and evaluation of protein dynamics, especially when combined with experimental data.

Popular Summary

Proteins. When we nowadays, in our daily life, hear or read about proteins, we are reminded about their nutritional value and how important proteins are for a healthy lifestyle. But, why is that? And what *are* proteins exactly? Made up by smaller building blocks, so-called amino acids, you can envision proteins consisting of a single or multiple chains of connected amino acids. We need to digest proteins in order to receive amino acids that our bodies cannot synthesise themselves. However, proteins are far more than just a supplement for providing us with essential amino acids. Proteins are biologically highly important macromolecules that enable, manage and regulate the majority of processes within our bodies.

But, before we go ahead, let us pause for a minute and imagine you are told to loosen a tight bolt. Would you use a hammer? Not really, the design of the hammer is not really the correct one for the task at hand. Pliers might already be a better option, potentially even losing the bolt, however accompanied with a lot of effort. No, probably you would choose a spanner with adjustable width, as it provides the ideal design and performance for the assignment. For proteins, this concept is very similar. Nature designed proteins to be remarkably efficient in their designated tasks. Tasks that range from transportation to structural support to mechanical movement or fighting infections, proteins carry out incredibly important functions. The overall structures of proteins are one key aspect for their efficiency. However, structure is not everything—interactions and dynamics are further crucial for proteins. Going back to the tight bolt and the spanner; just as we have seen the structure of the spanner being ideal to loosen the bolt, simply connecting the spanner and the bolt will not loosen the latter. We need to apply dynamics, mobility to loosen it, otherwise both would not interact. And again, we can bridge that observation to proteins. Understanding protein structure and their dynamics and interactions, we can estimate their functions and with that what their specific role is within our body. This become even more evident when we look at proteins that are somehow impaired to fulfil their job, which is often the case or cause for specific diseases or harmful conditions. Hence, understanding how proteins work and what influences them, has the potential to benefit our wellbeing, let alone broadening the horizon of understanding life all around us.

Investigating protein structures can be done using a plethora of methods. There are those that reveal the amino acid composition of proteins, their

structure in high resolution, and even are able to assess their molecular weight, like MS. MS itself is a highly powerful method in use in many laboratories worldwide. Amongst other things, MS studies proteins and their properties, however exclusively with proteins being exposed to the gas phase. That means, proteins are investigated in the absence of any solvent or other molecules. As you know, approximately 70 % of the human body consists of water, which consequentially determines the native environment of proteins. So, why in vacuum and the removal of the solvent? Simply to be able to measure isolated proteins, and thus extract as much information as possible without any potential interferences. The development of gentle conditions for how proteins are transferred into the gas phase allowed further proteins to retain a conformation close to their native structure, which are often applied in experiments using native MS. Detailed information about protein structure and their specific dynamics under distinctive conditions can be traced using native MS. However, where there are benefits, there are limitations as well. Whilst information about the general structure of proteins is provided by (native) MS, high-resolution of their behaviour on an atomistic level is inaccessible for this method. Imagine you would need to fix your car's engine, but you are restricted to opening the hood and can only look at the engine itself. You might be able to spot where the problem could be or originate from, but incapable to obtain the specific details of the problem *itself*. Experimental methods for protein research encounter the same, where atomistic details are often simply not within their capabilities.

To obtain a detailed look into proteins, researchers employ computational methods such as molecular dynamics (MD) simulations. Simulations is here the key word; using MD, we are simulating proteins and gather theoretical data about the dynamics of proteins by applying Newton's second law of motion. As we remember, when a force is acting on a mass, that mass is accelerated. If we would have an apple (= mass) on a table and push it in any direction (= force), the velocity of the apple will change (= acceleration) and the apple itself will end up on a new position on the table. In MD, we basically do the same, however for every atom within the protein, which are bonded and thus move together. From an initial position of the protein, we apply a certain force and update the information for the new position of each atom. Doing that in multiple steps, we see the protein quite literally moving and thus are able to simulate the protein's dynamics.

This thesis aims to provide further atomistic information about the proteins using MD simulations. Over a total of four scientific inquiries (Papers), we simulated proteins in the gas phase or used available gas-phase data to gather additional insights into their underlying dynamics under specific conditions. The here presented research can moreover be subdivided into two major research questions: rehydration and collision-induced unfolding (CIU) simulations of proteins. Rehydration means placing previously vacuum-exposed pro-

teins back into water. When taking proteins out of their native environment, we expose them—quite literally—to an environment they are not made for. This could potentially affect their structure, and thus could have consequences when we use vacuum data of proteins and assume that this data describes their native state. Rehydration is the major research question we investigate in **Paper I** and **Paper II**. For **Paper I**, we chose to simulate the two proteins of the bacteriophage MS2 virus that make up the virus's protective shell. Viruses are pathogens that seek out to infect a host cell, capture and overtake its micromachinery for replication and thus produce new copies of their genomic data. The latter is housed, transported and protected by an encapsulating shell which consists of multiple copies of the same proteins. Those proteins can however differ slightly in their structure, as is the case for the bacteriophage MS2 virus as well. We aimed to understand how vacuum-exposure might affect those two dimers, and further how they behave when placed back into solution afterwards. We saw that the removal of water from the proteins results in an overall compaction of their structure, which is for the most part reverted during rehydration. Using MD, we were able to accurately pinpoint atomistic motions and areas that are eventually reverting back, and also, which areas do not. A similar approach was undertaken for **Paper II**, where we expanded our research to include proteins that were affected by an external electric field whilst in vacuum. Proteins, as many other molecules, possess a natural dipole. That means that the centres of the net positive and net negative charge are located in different areas within the protein structure. An applied electric field is therefore able to interact with said dipole and thus manipulate its orientation along the direction of the electric field. This would mean that the protein will change its orientation as well, just as a boat that set sail will orient itself along the direction of the wind. Manipulating the orientation has the potential to be exceptionally impactful for certain protein structure determination methods; however, if the applied electric field is too strong, the protein structure itself could be distorted or even permanently damaged. As such, we used data of dipole-oriented gas-phase proteins, rehydrated them, and investigated their rehydration dynamics and if the applied electric fields did leave irreversible changes to their respective structures. We observed similar behaviour as in **Paper I**, and also, that proteins oriented with different electric field strengths do not possess structural differences that were induced by the applied electric fields. Rehydration simulations thus are shown to be a potent method to not only investigate the atomistic details of vacuum-compaction, but could generally be used to refine protein structures and bring them closer to their native conformations. The second major research question we studied was to simulate CIU of proteins. In specific MS experiments, proteins are further separated in a small gas-filled compartment of the instrument. The gas is inert, which means that its molecules are highly stable and will not react with the proteins, or other molecules in general. The separation of the proteins is based on their overall shape and size as they travel through the aforementioned com-

partment, where the proteins collide with the gas molecules. The bigger the protein, the more it collides with the inert gas, and thus the longer it needs to migrate through the so-called drift tube. The velocity differences between protein and gas-molecules are increased, and as such more energetic collisions are generated between them. This will eventually result in parts of the protein to unfold, implying that the protein structure is actively disturbed. Why destroying the protein structure, you might ask. CIU experiments are very effective to probe and understand the structure and stability of proteins, which we aimed to simulate via MD in **Paper III** and **Paper IV**. Whereas in CIU experiments the collision energy is increased leading to unfolding, using MD, unfolding is induced by actively increasing and simulating at large temperatures. In **Paper III**, we did so for two dimer proteins of the Norwalk and Kawasaki strains, genetic variants, of the Norovirus pathogen. We saw that the dimers are stable up to 600 K, yet are progressively unfolding at larger temperatures of 700+ K. Moreover, we were able to pinpoint potential areas of the proteins which are most likely to unfold first, and overall revealed a higher stability of the Kawasaki dimer as compared to the Norwalk dimer. Our findings thus provide the basis for follow-up experimental CIU studies of said proteins. For **Paper IV**, we used data derived from CIU experiments of the β-barrel protein FhaC to inform our CIU-MD simulations. FhaC is a transporter protein sitting in the lipid bilayer of bacteria responsible for causing the whopping cough disease. As transporter, FhaC must be able to undergo significant dynamical rearrangements to carry out its job, which was eventually evaluated using CIU. We combined experimental together with our simulation data and were able to predict a detailed unfolding pattern for FhaC. Atomistic information was revealed and allowed for a detailed estimation of FhaC's structure during the unfolding process, presenting a mechanism that contradicts predictions based on experimental data. As such, we show that CIU-MD simulations are highly useful to understand, track and pinpoint important states of the unfolding process of proteins, be it to provide information for future experiments, or to complement existing research data.

Overall, the research presented in this thesis advances the understanding of protein structures using MD simulations. The obtained information from MD simulations shows how important it is to put protein structural research into an atomistic perspective. We have extensively discussed which benefits could be gained using MD, where MD should complement or should be complemented, and what could eventually be improved in future studies. Moreover, our results show the advantages of combining MD simulations with experimental methods, especially MS, pinpointing exact details in atomic resolutions, and thus present the potential of MD becoming further prominent as toolset for protein structures research.

Populärvetenskaplig Sammanfattning

Proteiner. Nu för tiden, när vi hör eller läser om proteiner i våra dagliga liv, tänker vi förmodligen på deras näringsvärde och på hur viktiga de är för en hälsosam livsstil. Men varför är det så? Och vad är proteiner egentligen? De består av mindre byggstenar, så kallade aminosyror, och du kan föreställa dig ett protein som en eller flera kedjor av sammanlänkade aminosyror. Vi behöver äta proteiner för att få i oss de aminosyror som kroppen inte kan syntetisera själv. Men proteiner är mycket mer än bara tillskott för att förse oss med essentiella aminosyror. Proteiner är biologiskt viktiga makromolekyler som möjliggör, driver och reglerar majoriteten av processer i våra kroppar.

Men innan vi går vidare, låt oss stanna till ett ögonblick och föreställ dig att du ska lossa på en hårt åtsittande bult. Hade du valt att använda en hammare? Antagligen inte – hammaren är inte formgiven för att passa uppgiften. En tång är ett något bättre alternativ; kanske kan den få loss bulten något, men det kräver också en hel del ansträngning. Förmodligen hade du istället valt en skiftnyckel, eftersom den har en perfekt form och prestanda för just detta. Konceptet är väldigt likt för proteiner. Naturen formade proteiner för att göra dem anmärkningsvärt effektiva för sina särskilda uppgifter. Dessa uppgifter innefattar allt ifrån transport till strukturellt stöd, mekanisk rörelse och att bekämpa infektioner, med andra ord funktioner som är otroligt viktiga. Proteiners effektivitet beror till stor del på deras struktur, men utöver det spelar deras interaktioner och dynamik även stor roll. Vi återgår till bulten och skiftnyckeln; fastän skiftnyckeln är perfekt utformad för att ta loss bulten, räcker det inte att bara sammankoppla dem. Vi måste också ta till dynamik, rörelse, för att lossa den, annars skulle de inte ha interagerat. Återigen kan vi göra en koppling till proteiner. Genom att förstå proteiners struktur, dynamik och interaktioner kan vi uppskatta deras funktioner och förstå vilken roll de spelar i vår kropp. Det blir ännu tydligare när vi ser på proteiner som på något sätt har en nedsatt förmåga att göra sitt jobb, vilket ofta är orsaken till sjukdomar och skadliga tillstånd. Att förstå hur proteiner fungerar och vad de påverkas av kan därför gynna vårt välmående såväl som bredda horisonten för vår förståelse av livet omkring oss.

För att undersöka proteinstrukturer har vi en mängd metoder till vårt förfogande. Vissa används för att avslöja proteinernas sammansättning av aminosyror, andra deras struktur med hög upplösning, eller så kan de mäta molekylvikten, som med masspektrometri (MS). MS är i sig en mycket kraftfull metod som används i många laboratorier världen över. Bland annat kan man med MS

undersöka proteiner och deras egenskaper, dock endast medan proteinerna är i gasfas. Det innebär att proteinerna studeras utan någon lösning eller andra molekyler. Som du vet så består människokroppen till 70 % av vatten, vilket följaktligen utgör proteiners nativa omgivning. Så varför i gasfas, utan någon lösning? Helt enkelt för att kunna mäta isolerade proteiner och därmed få fram så mycket information som möjligt utan några potentiella störningar. Uppveckningen av milda förhållanden för att överföra proteiner från lösning till gasfas tillåter nu proteiner att behålla sin nativa struktur och de tillämpas därför ofta i s.k. nativa MS-experiment. Detaljrik information om proteiners struktur och deras specifika dynamik under särskilda förhållanden kan tas fram med nativ MS. Men där det finns fördelar finns också begränsningar. Medan information om proteiners generella struktur kan tas fram med (nativ) MS, så är det inte möjligt att beskriva deras beteende på atomnivå med den här metoden. Tänk dig att du behöver laga din bils motor, men att du är begränsad till att enbart öppna motorhuven och kan bara se motorn i sin helhet. Kanske klarar du att hitta var felet sitter, men du kan omöjligen beskriva det i närmre detalj. Samma sak gäller för flertalet experimentella metoder; atomiska detaljer är helt enkelt inte inom räckhåll.

För att få en mer detaljerad inblick i proteiner använder sig forskare av beräkningsmetoder såsom molekylärdynamiska (MD)-simuleringar. Simuleringar är nyckelbegreppet här; vi använder MD för att simulera proteiner och samla teoretiska data om proteiners dynamik genom att tillämpa Newtons andra lag. Som vi minns: om en massa påverkas av en kraft så accelererar massan. Om vi har ett äpple (= massa) på ett bord och puttar det i någon riktning (= kraft) så förändras äpplets hastighet (= acceleration) och äpplet hamnar på en ny plats på bordet. Med MD gör vi i princip samma sak, men för varje atom inom proteinet, vilka är sammanbundna och därmed förflyttas tillsammans. Vi applicerar en särskild kraft på proteinet i dess ursprungsposition och uppdaterar varje atoms nya position. Genom att upprepa detta över flera steg kan vi bokstavligen se hur proteinet rör sig och kan därigenom simulera dess dynamik.

Denna avhandling ämnar att förse ytterligare atomisk information om proteiner genom användning av MD-simuleringar. Över totalt fyra vetenskapliga artiklar (Papers) simulerade vi proteiner i gasfas eller använde tillgänglig gasfas-data för att få fram nya insikter om deras underliggande dynamik under särskilda förhållanden. Forskningen som presenteras här kan vidare delas in i två huvudsakliga forskningsfrågor: proteinsimuleringar av rehydrering och kollision-inducerad uppveckning (Engelska: collision-induced unfolding, CIU). Rehydrering betyder att proteiner som tidigare utsatts för vakuum åter placeras i vatten. När proteiner förflyttas från sin nativa omgivning utsätter vi dem närmast bokstavligen för en omgivning de inte var skapade för. Det kan potentiellt påverka deras struktur och därmed ha konsekvenser när vi använder data från vakuumförhållanden och antar att de beskriver det nativa tillståndet. Rehydrering är den främsta forskningsfrågan som vi undersöker i

Paper I och **Paper II**. I **Paper I** valde vi att simulera de två proteiner som bildar det skyddande skalet kring bakteriofagviruset MS2. Virus är patogener som eftersträvar att infektera en värdcell och överta dess mikromaskineri för replikation och på så vis producera nya kopior av sin genomiska data. Det senare inryms i och transporteras och skyddas av ett inkapslande skal som består av många kopior av samma proteiner. Dessa proteiner kan dock skilja sig lite åt i sina strukturer, vilket är fallet för bakteriofagviruset MS2. Vi hade som mål att förstå hur vakuumexponering påverkar två strukturvarianter av det dimeriska kapsidproteinet i MS2 och även hur de beter sig när de efteråt placeras i lösning igen. Vi såg att när vatten togs bort från proteinerna så blev deras strukturer överlag mer kompakta, men att de under rehydreringen till största del återfick sin ursprungliga struktur. Genom att använda MD kunde vi peka ut exakta atomiska rörelser samt vilka områden som återfick sin struktur och vilka som inte fick det. Ett liknande tillvägagångssätt användes i **Paper II**, där vi breddade vår forskning till att inkludera proteiner som påverkas av ett yttre elektriskt fält under tiden de befinner sig vakuum. Proteiner är, likt många andra molekyler, en naturlig dipol. Det betyder att centren för den nettopositiva och den nettonegativa laddningen är positionerade på olika ställen inom proteinstrukturen. Ett yttre elektriskt fält kan därför interagera med dipolen och därigenom förflytta dess orientering i det elektriska fältets riktning. Detta skulle innebära att även ett protein förändrar sin orientering, precis som att en båt som har satt segel orienterar sig själv i vindens riktning. Att manipulera orienteringen har potential att bli exceptionellt kraftfullt för särskilda metoder som används för att bestämma proteinstrukturer; dock riskerar proteinstrukturen att bli förvrängd eller till och med permanent skadad om det elektriska fältet är för starkt. Därför använde vi data från dipol-orienterade gasfas-proteiner, rehydrerade dem och undersökte rehydreringsdynamiken samt om det elektriska fältet orsakat några irreversibla förändringar på de respektive strukturerna. Vi observerade liknande beteenden som i **Paper I** och kunde dessutom se att proteiner som orienterats med olika starka elektriska fält inte hade några strukturella skillnader som orsakats av fältet. Rehydreringssimuleringar har därmed visats vara en potent metod inte enbart för att undersöka de atomiska detaljerna vid vakuum-kompaktering, utan de kan generellt användas för att förfina proteinstrukturer och föra dem närmre det nativa tillståndet. Den andra huvudsakliga forskningsfrågan som vi studerade var att simulera CIU av proteiner. I särskilda MS-experiment separeras proteiner ytterligare i en liten gasfylld enhet i instrumentet. Gasen är inert, vilket innebär att dess molekyler är mycket stabila och att de inte kommer reagera med proteiner eller andra molekyler. Separationen av proteiner baseras på deras övergripande form och storlek medan de färdas genom den sagda enheten, där proteinerna kolliderar med gasmolekylerna. Ju större proteinet är, desto mer kolliderar det med den interta gasen och desto längre tid tar det för det att migrera genom den så kallade drifttuben. Hastighetsskillnaderna mellan proteinerna och gasmolekylerna ökar, och

därför sker det fler energiska kollisioner mellan dem. Det resulterar till slut i att delar av proteinet vecklas ut, vilket innebär att proteinstrukturen aktivt har rubbats. Kanske undrar du vad det är för mening med att förstöra proteinstrukturen. Experiment med CIU är väldigt effektiva för att undersöka och förstå proteiners struktur och stabilitet, vilket vi avsåg simulera med MD i **Paper III** och **Paper IV**. Medan man i CIU-experiment ökar kollisionsenergin för att proteinet ska vecklas upp, så kan man i MD inducera uppveckning genom att aktivt höja och simulera vid högre temperaturer. I **Paper III** genomförde vi detta för två dimerer från stammarna Norwalk och Kawasaki, genetiska varianter av det patogena noroviruset. Vi såg att dimererna är stabila upp till 600 K, men att de utveckas successivt vid högre temperaturer om 700+ K. Dessutom kunde vi peka ut potentiella områden hos proteinerna som har störst sannolikhet att vecklas ut först, samt fann en högre stabilitet hos Kawasaki-dimeren än hos Norwalk-dimeren. Våra upptäckter ger därmed en utgångspunkt för ytterligare experimentella CIU-studier av dessa proteiner. Till **Paper IV** använde vi data från CIU-experiment på det så kallade "beta-barrel-proteinet" FhaC till våra CIU-MD-simuleringar. FhaC är ett transportprotein som sitter i lipidbilagret hos de bakterier som orsakar kikhosta. Som transportör måste FhaC kunna genomgå betydande restrukturering för att utföra sin uppgift, vilket utvärderades med CIU. Vi kombinerade experimentella data och våra simuleringsdata och kunde förutse ett detaljerat uppveckningsmönster för FhaC. Atomär information kunde erhållas och det tillät en detaljerad uppskattuppveckningning av FhaCs struktur genom uppveckningsprocessen, vilket visade på en mekanism som motsäger modeller baserade på experimentella data. På så vis visade vi att CIU-MD-simuleringar är högst användbara för att förstå, följa och finna viktiga tillstånd under proteiners uppveckningsprocessen, vare sig det gäller att få fram information till framtida experiment eller att komplettera existerande data.

I sin helhet avancerar forskningen som presenteras i denna avhandling förståelsen för proteinstruktur med hjälp av MD-simuleringar. Informationen som framtagits med MD-simuleringar visar på vikten av att sätta proteinstrukturforskningen i ett atomiskt perspektiv. Vi har på ett omfattande sätt diskuterat vilka fördelar som kan uppnås av att använda av MD, när MD ska komplementera eller komplementeras och vad som småningom kan förbättras i framtida studier. Därutöver visar våra resultat på fördelarna med att kombinera MD-simuleringar med experimentella metoder – i synnerhet MS, där exakta detaljer då kan beskrivas med atomisk upplösning och det visar på potentialen att MD kan bli en än mer framträdande verktygslåda för proteinstukturforskningen.

Populaire Samenvatting

Eiwitten. Wanneer we in het dagelijks leven lezen of horen over eiwitten, is dat vaak in de context van hun voedingswaarden en hoe belangrijk ze zijn voor een gezonde levensstijl. Waarom is dat? En wat zijn eiwitten precies? Een eiwit is opgebouwd uit kleinere bouwstenen, zogenaamde aminozuren, die als een keten met elkaar verbonden zijn. Je zou een eiwit eigenlijk kunnen zien als een parelketting, waarin de aminozuren de parels zijn. Om aminozuren op te nemen die ons lichaam niet zelf kan maken, ook wel essentiële aminozuren genoemd, moeten eiwitten worden verteerd. Eiwitten fungeren echter niet alleen als een bron van essentiële aminozuren, maar zijn ook betrokken bij de regulatie van een hoop belangrijke processen in het lichaam. En daar zijn ze in principe erg goed in.

Stel je voor dat je wordt gevraagd om een strakke bout los te draaien. Zou je dan een hamer gebruiken? Waarschijnlijk niet. Hoewel een tang al een betere optie is, en het waarschijnlijk ook best lukt om de bout iets losser te draaien, kost dit waarschijnlijk erg veel moeite. De meest efficiënte manier is om het gereedschap te pakken dat past bij deze taak, namelijk een sleutel of moersleutel met instelbare wijdte. Als we dit concept toepassen op eiwitten, zien we veel gelijkenis. De natuur heeft eiwitten namelijk op een zodanige manier gemaakt, dat zij erg efficiënt zijn in hun aangewezen taak. Dit is belangrijk, aangezien eiwitten, zoals eerder verteld, betrokken zijn bij een hoop belangrijke processen in het lichaam, waaronder transport, structurele ondersteuning, mechanische beweging en het bestrijden van infecties. Hoewel de algemene structuur van eiwitten een belangrijke rol speelt, zijn interactie en dynamiek ook van groot belang voor het functioneren van eiwitten. Terugkomend op het voorbeeld met de bout en de sleutel; net zoals dat we hebben gezien dat de (moer)sleutel ideaal is om de bout los te draaien, zal het simpelweg in contact brengen van de sleutel en de bout niet voldoende zijn om de bout los te krijgen. We moeten dynamiek, beweging, toepassen om de bout ook daadwerkelijk los te krijgen. Ook hier kan weer een vergelijking worden gemaakt met eiwitten. Alleen als we de eiwitstructuur én hun dynamiek én interacties begrijpen, kunnen we de specifieke rol van eiwitten in ons lichaam goed inschatten. Dit is niet alleen van belang voor de algemene kennis, maar ook om ziekten, welke het gevolg kunnen zijn van eiwitten die op de één of andere manier zijn aangetast, beter te begrijpen en/of behandelen.

Eiwitten kunnen worden bestudeerd met allerlei verschillende methodes. Zo zijn er methodes waarmee de aminozuursamenstelling van eiwitten kan wor-

den onderzocht, methodes waarmee hun structuur in hoge resolutie kan worden bestudeerd en zelfs methodes waarmee het molecuulgewicht kan worden bepaald, zoals massaspectrometrie (MS). MS is en veelgebruikte techniek waarmee eiwitten (en hun eigenschappen) worden bestudeerd, terwijl deze zich in de gasfase bevinden. Dat betekent dat eiwitten worden onderzocht in afwezigheid van oplosmiddelen of andere moleculen. Dit is best verrassend, aangezien ongeveer 70 % van het menselijk lichaam uit water bestaat, wat dus ook de natuurlijke omgeving van eiwitten is. De reden dat eiwitten toch in de gasfase worden gebracht, is omdat dit de mogelijkheid biedt om geïsoleerde eiwitten te kunnen bestuderen, met zo min mogelijk invloed van andere variabelen. Tegenwoordig is het mogelijk om eiwitten heel voorzichtig in de gasfase te brengen, wat er voor zorgt dat de eiwitten een conformatie behouden die erg lijkt op hun oorspronkelijk structuur. Hoewel MS een populaire methode is, is het niet mogelijk om eiwitten te bestuderen op atoom-niveau, wat wel erg interessant kan zijn. Stel je voor dat je auto niet wilt starten en dat je erachter probeert te komen wat het probleem is. Als je beperkt bent tot het openen van de motorkap, zul je wellicht zien dat er iets niet klopt en dat er een probleem is met bijvoorbeeld de motor. Dit is echter niet altijd genoeg om er ook daadwerkelijk achter te komen wat de oorzaak van het probleem is, aangezien je geen verdere details kan zien/onderzoeken als je alleen kan kijken naar de motor in zijn geheel. Experimentele methoden voor eiwitonderzoek, zoals dus MS, stuiten op hetzelfde probleem, namelijk dat details op atoom-niveau niet kunnen worden bestudeerd, terwijl deze wel erg relevant kunnen zijn. Om eiwitten in meer detail te kunnen bestuderen, gebruiken onderzoekers computationele methoden zoals moleculaire dynamica (MD) simulaties. ‘Simulaties’ is hier het sleutelwoord; met MD simuleren we eiwitten en verzamelen we theoretische gegevens over de dynamiek van eiwitten. Dit wordt gedaan door de tweede bewegingswet van Newton toe te passen: wanneer een kracht op een massa werkt, wordt die massa versneld. Als we een appel (= massa) op een tafel zouden leggen en deze in een willekeurige richting zouden duwen (= kracht), dan verandert de snelheid van de appel (= versnelling) en komt de appel op een nieuwe positie op de tafel terecht. In MD doen we in principe hetzelfde, maar dan voor elk atoom in het eiwit. Vanuit een beginpositie van het eiwit oefenen we eerst een bepaalde kracht uit. Als de nieuwe positie van elk atoom wordt bijgewerkt, wordt er opnieuw een bepaalde kracht uitgeoefend. Door dit meerdere malen te herhalen, zien we het eiwit letterlijk bewegen, waarmee we de dynamiek van het eiwit kunnen simuleren.

Het doel van dit proefschrift is om met behulp van MD simulaties meer inzicht in (de atomen van) eiwitten te krijgen. Het hier gepresenteerde onderzoek kan worden onderverdeeld in twee onderzoeksthema’s: rehydratie (**Artikel I** en **Artikel II**) en ‘botsingsgeïnduceerde ontvouwing’ (Engels: collision-induced unfolding, CIU) simulaties van eiwitten (**Artikel III** en **Artikel IV**). Rehy-

dratie verwijst naar het terugplaatsen van aan vacuüm blootgestelde eiwitten in water. Wanneer we eiwitten uit hun natuurlijke omgeving halen, stellen we ze letterlijk bloot aan een omgeving waarvoor ze niet gemaakt zijn. Dit kan hun structuur beïnvloeden en daardoor dus de manier waarop zij zich gedragen, wat weer kan leiden tot misleidende resultaten. In Artikel I bestudeerden we twee eiwitten die de beschermende schil van het bacteriofaag MS2-virus vormen. Virussen zijn ziekteverwekkers die cellen kunnen infiltreren en de opdracht kunnen geven om nieuw viraal genetisch materiaal te maken. Dit virale genetische materiaal bevind zich in een omhulsel, wat bestaat uit meerdere kopieën van zo goed als dezelfde eiwitten. Deze eiwitten kunnen wel iets verschillen in structuur, zoals ook het geval is bij het bacteriofaag MS2-virus. Het doel van **Artikel I** was om te begrijpen hoe blootstelling aan vacuüm de twee dimeren zou kunnen beïnvloeden en hoe deze zich gedragen wanneer ze daarna weer in oplossing worden gebracht. We zagen dat het verwijderen van water uit de eiwitten resulteert in een algehele verdichting van hun structuur en dat dit voor het grootste deel wordt teruggedraaid tijdens rehydratie. Met behulp van MD konden we atomistische bewegingen nauwkeurig lokaliseren. Ook konden we zien welke gebieden uiteindelijk terugkeren naar hun oorspronkelijk structuur en welke gebieden dat niet doen. Dit onderzoek werd verder uitgebreid in **Artikel II**, waar eiwitten werden bestudeerd die aan een extern elektrisch veld waren blootgesteld terwijl ze zich in vacuüm bevonden. Net als veel andere moleculen hebben eiwitten een natuurlijke dipool. Dat betekent dat het centrum van de netto positieve en netto negatieve lading zich in verschillende gebieden binnen de eiwitstructuur bevinden. Een extern elektrisch veld kan een interactie aangaan met de dipool, wat de oriëntatie van de dipool kan manipuleren in de richting van het elektrische veld. Dit betekent dat het eiwit ook van oriëntatie zal veranderen, net zoals een boot die uitvaart zich zal oriënteren in de richting van de wind. Hoewel het manipuleren van de oriëntatie bijzonder nuttig kan zijn om de eiwitstructuur beter te begrijpen moet er ook rekening worden gehouden met de sterkteé van het elektrische veld; als het aangelegde elektrische veld te sterk is, is er een kans dat de eiwitstructuur vervormd of zelfs permanent beschadigd. In het vervolg van dit artikel hebben wij dit verder uitgezocht. We gebruikten gegevens van dipool-georiënteerde eiwitten in de gasfase, bekeken de rehydratiedynamiek, waarna we bestudeerden of de toegepaste elektrische velden onomkeerbare veranderingen in de eiwitstructuren achterlieten. We observeerden vergelijkbaar gedrag als in **Artikel I**, en bovendien dat eiwitten die georiënteerd waren met verschillende elektrische veldsterktes geen structurele verschillen hadden die waren veroorzaakt door de toegepaste elektrische velden. Rehydratiesimulaties blijken dus een krachtige methode te zijn om niet alleen de atomistische details van vacuümverdichting te onderzoeken, maar ook om eiwitstructuren te verfijnen en deze dichter bij hun oorspronkelijke conformaties te brengen.

Het tweede onderzoeksthema dat we bestudeerden was het simuleren van CIU van eiwitten. In bepaalde MS-experimenten worden eiwitten verder gescheiden in een klein met gas gevuld compartiment van het instrument. Het gas is inert, wat betekent dat de moleculen zeer stabiel zijn en niet zullen reageren met de eiwitten of andere moleculen. De scheiding van de eiwitten is gebaseerd op hun algehele vorm en grootte. Terwijl de eiwitten door het bovengenoemde compartiment reizen, komen deze in botsing met de gasmoleculen. Hoe groter het eiwit, hoe meer het in aanraking komt met het inerte gas en dus hoe langer het eiwit nodig heeft om te migreren. De snelheidsverschillen tussen eiwit- en gasmoleculen worden vergroot, waardoor er meer energetische botsingen ontstaan. Dit zal er uiteindelijk toe leiden dat delen van het eiwit zich ontvouwen, wat impliceert dat de eiwitstructuur actief wordt verstoord. Dit soort experimenten, waarbij de eiwitstructuur actief wordt verstoord zijn erg effectief om de structuur en stabiliteit van eiwitten te onderzoeken en te begrijpen. In **Artikel III** en IV wilden we dit dan ook simuleren met MD. Waar in CIU-experimenten de botsingsenergie wordt verhoogd, wat leidt tot ontvouwen van eiwitten, wordt ontvouwing bij MD geïnduceerd door de temperatuur te verhogen en te simuleren bij deze hoge temperaturen. In **Artikel III** hebben we dit gedaan voor twee dimeereiwitten van de Norwalk- en Kawasaki-stammen, wat genetische varianten van het Norovirus-pathogeen zijn. We zagen dat de dimeren stabiel zijn tot 600 K, maar zich geleidelijk ontvouwen bij hogere temperaturen van 700+ K. Ook lokaliseerden we potentiële gebieden van de eiwitten die naar waarschijn-lijkheid het eerst zullen ontvouwen en zagen we (in het algemeen) een hogere stabiliteit van het Kawasaki-dimeer in vergelijking met het Norwalk-dimeer. Deze bevindingen vormen de basis voor toekomstig CIU-onderzoek van de genoemde eiwitten. In **Artikel IV** hebben we gegevens van CIU-experimenten van het β-barrel-eiwit FhaC gebruikt voor de CIU-MD-simulaties. FhaC is een transporteiwit dat in de lipidedubbellaag van bacteriën zit en verantwoordelijk is voor het veroorzaken van kinkhoest. Als transporteiwit moet FhaC aanzienlijke dynamische veranderingen kunnen ondergaan om zijn taak te vervullen, die uiteindelijk werden geëvalueerd met behulp van CIU. We combineerden experimentele gegevens met onze simulatiegegevens en konden een gedetailleerd ontvouwingspatroon voor FhaC voorspellen. Ook hebben we laten zien dat de MD simulatie de voorspellingen op basis van experimentele data tegenspreekt. In dit artikel hebben we dus laten zien dat CIU-MD-simulaties zeer nuttig zijn om het ontvouwingsproces van eiwitten te begrijpen, te volgen en te lokaliseren, of het nu gaat om informatie voor toekomstige experimenten of om bestaande onderzoeksgegevens aan te vullen.

De verschillende onderzoeken in dit proefschrift hebben laten zien hoe belangrijk onderzoek naar eiwitten op atoom-niveau is. We hebben uitgebreid laten zien wat voor voordelen het gebruik van MD kan hebben, waar MD een

aanvulling zou moeten zijn of zou moeten worden aangevuld en wat uiteindelijk zou kunnen worden verbeterd in toekomstige studies. Bovendien tonen onze resultaten de meerwaarde van het combineren van MD-simulaties met experimentele methoden, met name MS, waardoor exacte details in atomaire resoluties kunnen worden vastgesteld. Al met al heeft het onderzoek dat in dit proefschrift is gepresenteerd laten zien dat MD een veelbelovende methode is om eiwitten beter te begrijpen en potentie heeft om een prominente research tool te worden binnen dit vakgebied.

Populärwissenschaftliche Zusammenfassung

Proteine. Wenn wir heutzutage in unserem täglichen Leben von Proteinen hören oder lesen, werden wir an ihren Nährwert erinnert und daran, wie wichtig sie für einen gesunden Lebensstil sind. Aber warum ist das so und was sind Proteine genau? Proteine bestehen aus kleinen Bausteinen, den sogenannten Aminosäuren. Ein Protein kann hierbei aus einer einzelnen oder aus mehreren Ketten verbundener Aminosäuren bestehen. Wir müssen Proteine verdauen, um die Aminosäuren zu erhalten, die unser Körper selbst nicht synthetisieren kann. Proteine sind jedoch weit mehr als nur eine Nahrungsergänzung, die uns mit essentiellen Aminosäuren versorgt. Sie stellen auch sehr wichtige biologische Makromoleküle dar, welche uns die Steuerung und Regulierung der meisten Prozesse in unserem Körper ermöglichen.

Stellen Sie sich vor, Sie stehen vor der Herausforderung einen festsitzenden Bolzen lösen zu müssen. Würden Sie hierfür einen Hammer benutzen? Wahrscheinlich nicht, denn die Funktionsweise des Hammers ist nicht wirklich passend für die anstehende Aufgabe. Eine herkömmliche Zange wäre möglicherweise schon eine bessere Option. Mit dieser könnten Sie den Bolzen vielleicht sogar lösen, jedoch wäre dies mit einem enormen Kraftaufwand verbunden. Im Idealfall würden Sie wahrscheinlich eine Rohrzange mit verstellbarer Breite wählen. Diese bietet die ideale Form zur Lösung des Problems bei minimalem Kraftaufwand. Bei Proteinen ist dieses Konzept sehr ähnlich. Diese wurden von der Natur so gestaltet, dass sie ihre zugewiesenen Aufgaben bemerkenswert effizient erfüllen können. Aufgaben, die vom Transport über die strukturelle Unterstützung bis hin zur mechanischen Bewegung oder der Bekämpfung von Infektionen reichen – Proteine erfüllen unterschiedlichste wichtige Funktionen. Die Gesamtstruktur von Proteinen ist ein Schlüsselaspekt für ihre Effizienz. Struktur ist jedoch nicht alles – auch ihre Wechselwirkungen und ihre Dynamik sind für die Proteine entscheidend. Zurück zu dem festen Bolzen und der Rohrzange: wie wir gesehen haben, ist die Form der Rohrzange grundsätzlich ideal zum Lösen des Bolzens, jedoch wird dieser durch lediglliches Auflegen der Zange nicht gelöst. Erst die Anwendung von Dynamik und Mobilität, also die Interaktion, führt zur angestrebten Lösung des Problems. Und auch diese Beobachtung lässt sich auf Proteine übertragen: Durch das Verständnis der Proteinstruktur, ihrer Dynamik und Wechselwirkungen können wir ihre Funktionen und damit ihre spezifische Rolle in unserem Körper abschätzen. Dies wird vor allem bei der Betrachtung von Proteinen, die in ihrer Aufgabenerfüllung gehindert sind, verdeutlicht. Regelmäßig stellen diese die Ursache für bestimmte Krankheiten oder

schädliche Zustände dar. Die Kenntnis über die Funktionsweise von Proteinen sowie deren Beeinflussbarkeit hat das Potential unserem Wohlbefinden zugute zu kommen und unseren Horizont über das Verständnis allen Lebens zu erweitern.

Die Untersuchung von Proteinstrukturen kann mit einer Vielzahl von Methoden durchgeführt werden. Es gibt solche, die die Aminosäurezusammensetzung von Proteinen und ihre Struktur in hoher Auflösung aufzeigen und sogar ihr Molekulargewicht bestimmen können, wie zum Beispiel die Massenspektrometrie (MS). Diese ist eine sehr leistungsfähige Methode, welche in vielen Laboren weltweit zur Untersuchung von Proteinen verwendet wird. Hierbei werden diese in die sogenannte Gasphase überführt und unter der Abwesenheit von Lösungsmitteln und anderen Molekülen auf ihre Eigenschaften untersucht. Der menschliche Körper besteht zu etwa 70 % aus Wasser, was folglich die natürliche Umgebung von Proteinen darstellt. Warum erfolgt die Untersuchung also im Vakuum unter der Entfernung des Lösungsmittels? Ganz einfach, um isolierte Proteine messen und so möglichst viele Informationen ohne potentielle Interferenzen extrahieren zu können. Die Entwicklung schonender Bedingungen für den Transfer von Proteinen in die Gasphase ermöglicht es diesen weiterhin, eine Konformation (d.h. räumliche Anordnung) nahe ihrer nativen Struktur, welche häufig in Experimenten mit nativer MS angewendet wird, beizubehalten. Detaillierte Informationen über die Proteinstruktur und deren spezifische Dynamik unter bestimmten Bedingungen können so mit nativer MS gewonnen werden. Doch wo es Vorteile gibt, gibt es auch Einschränkungen. Während Informationen über die allgemeine Struktur von Proteinen durch (native) MS bereitgestellt werden, ist eine hohe Auflösung ihres Verhaltens auf atomarer Ebene für diese Methode nicht zugänglich. Stellen Sie sich vor, Sie müssten den Motor Ihres Autos reparieren, dürften hierfür aber nur die Motorhaube öffnen und den Motor betrachten. Möglicherweise könnten Sie erahnen, wo das Problem liegen oder woher es stammen könnte. Sie wären jedoch nicht in der Lage, die spezifischen Details des Problems selbst feststellen zu können. Experimentelle Methoden der Proteinforschung stoßen häufig auf gleichartige Probleme, da die Messung atomistischer Details nicht immer in deren Möglichkeiten liegen.

Um einen detaillierten Einblick in Proteine zu erhalten, verwenden Forscher Computermethoden wie Molekulardynamiksimulationen (MD). Simulation ist hier das Schlüsselwort: Mit MD simulieren wir Proteine und sammeln theoretische Daten über deren Dynamik, indem wir Newtons zweites Bewegungsgesetz anwenden. Eine Masse wird durch die Einwirkung einer Kraft beschleunigt. Wenn wir einen Apfel (= Masse) auf einen Tisch legen und diesen in eine beliebige Richtung schieben (= Kraft), ändert sich die Geschwindigkeit des Apfels (= Beschleunigung) und letztendlich zu einer neuen Position des Apfels auf dem Tisch. Bei MD machen wir im Grunde dasselbe, allerdings für alle Atome innerhalb des Proteins, welche miteinander verbunden sind

und sich somit zusammen bewegen. Auf ein Protein in einer bestimmten Anfangsposition wenden wir eine Kraft an und aktualisieren daraufhin die Informationen für die neue Position jedes Atoms. Wiederholen wir diesen Schritt mehrfach, sehen wir das Protein buchstäblich in Bewegung und können so dessen Dynamik simulieren.

Ziel dieser Arbeit ist es, mithilfe von MD weitere atomistische Informationen über Proteine zu erhalten. In insgesamt vier wissenschaftlichen Untersuchungen (Publikationen) haben wir Proteine in der Gasphase simuliert oder verfügbare Gasphasendaten verwendet, um zusätzliche Einblicke in ihre zugrunde liegenden Dynamiken unter bestimmten Bedingungen zu erhalten. Darüber hinaus kann die hier vorgestellte Arbeit in zwei Hauptforschungsbereiche unterteilt werden: Die Simulation von Proteinen durch Rehydrierung und die durch Kollisionsinduzierte Entfaltung (Englisch: collision-induced unfolding, CIU). Rehydrierung bedeutet, Proteine die zuvor einem Vakuum ausgesetzt waren wieder in Wasser zu geben. Wenn wir Proteine aus ihrer natürlichen Umgebung nehmen, setzen wir sie einer neuen Umgebung aus, für die sie nicht gemacht sind. Dies könnte möglicherweise ihre Struktur beeinflussen und folglich Konsequenzen für die Nutzung der Vakuumdaten haben, insofern wir davon ausgehen, dass diese Daten ihren nativen Zustand beschreiben. Rehydrierung ist die Hauptforschungsfrage, welche wir in **Publikation I** und **Publikation II** untersuchen. Für **Publikation I** entschieden wir uns, die beiden Proteine des Bakteriophagen-MS-2-Virus zu simulieren, welche die schützende Hülle des Virus bilden. Viren sind Krankheitserreger, die eine Wirtszelle infizieren und deren Mikromaschinerie für die Replikation übernehmen, um so neue Kopien ihrer genomischen Daten zu produzieren. Letztere wird von einer einkapselnden Hülle beherbergt, transportiert und geschützt, die aus mehreren Kopien derselben Proteine besteht. Diese Proteine können sich jedoch leicht in ihrer Struktur unterscheiden, wie dies auch beim Bakteriophagen-MS-2-Virus der Fall ist. Wir wollten verstehen, wie Vakuumexposition diese beiden Dimere beeinflussen könnte und wie sie sich verhalten, wenn sie danach wieder in Lösung gebracht werden. Wir haben gesehen, dass die Entfernung von Wasser aus den Proteinen zu einer allgemeinen Verdichtung ihrer Struktur führt, die sich während der Rehydrierung zum größten Teil wieder zurückbildet. Mithilfe von MD waren wir in der Lage, atomistische Bewegungen und Bereiche der Proteine zu erkennen, welche zu ihrer Normstruktur zurückkehren. Weiterhin konnten wir diese Bereiche genau lokalisieren und zudem feststellen, welche anderen dies nicht tun. Ein ähnlicher Ansatz wurde für **Publikation II** verfolgt, wo wir unsere Forschung auf Proteine ausweiteten, welche im Vakuum von einem externen elektrischen Feld beeinflusst wurden. Proteine besitzen wie viele andere Moleküle einen natürlichen Dipol. Das bedeutet, dass die Zentren der positiven und negativen Nettoladung in unterschiedlichen Bereichen innerhalb der Proteinstruktur lokalisiert sind. Ein angelegtes

elektrisches Feld ist daher in der Lage, mit dem Dipol zu interagieren und somit seine Ausrichtung entlang der Richtung des elektrischen Feldes zu manipulieren. Das würde bedeuten, dass auch das Protein seine Orientierung ändert, so wie sich ein Boot mit gesetztem Segel in Windrichtung ausrichtet. Die Manipulation der Orientierung hat das Potenzial, für bestimmte Proteinstrukturbestimmungsmethoden außergewöhnlich wirkungsvoll zu sein. Wenn das angelegte elektrische Feld jedoch zu stark ist, könnte die Proteinstruktur selbst verzerrt oder sogar dauerhaft beschädigt werden. Daher haben wir Daten von dipolorientierten Gasphasenproteinen verwendet, diese rehydriert, ihre Dynamiken untersucht und überprüft, ob die angelegten elektrischen Felder irreversible Veränderungen an den jeweiligen Strukturen hinterlassen haben. Wir beobachteten ein ähnliches Verhalten wie in **Publikation I**. Darüberhinausgehend stellten wir fest, dass Proteine, welche mit unterschiedlichen elektrischen Feldstärken orientiert wurden, keine strukturellen Unterschiede aufwiesen, die durch die angelegten elektrischen Felder hätten induziert werden können. Rehydrierungssimulationen erwiesen sich somit als eine wirksame Methode, um nicht nur die atomistischen Details der Vakuumverdichtung untersuchen zu können. Vielmehr können diese allgemein verwendet werden, um Proteinstrukturen zu verbessern und diese näher an ihre nativen Konformationen zu bringen. Die zweite große Forschungsfrage, welche wir untersuchten, war die Simulation von Proteinen durch CIU. In spezifischen MS-Experimenten wurden Proteine in einem kleinen, gasgefüllten Kompartiment des Messinstruments weiter getrennt. Das Gas ist inert, was bedeutet, dass dessen Moleküle sehr stabil sind und nicht mit den Proteinen oder anderen Molekülen im Allgemeinen reagieren. Die Trennung der Proteine basiert auf ihrer Gesamtform und -größe, während sie durch das zuvor erwähnte Kompartiment wandern, wo die Proteine mit den Gasmolekülen kollidieren. Je größer das Protein ist, desto mehr kollidiert es mit dem Gas und desto länger braucht es, um durch die sogenannte Driftröhre zu wandern. Die Geschwindigkeitsunterschiede zwischen Protein- und Gasmolekülen werden erhöht, was energetisch steigende Kollisionen zwischen ihnen erzeugt. Dies führt schließlich dazu, dass sich Teile des Proteins entfalten, wodurch die Proteinstruktur aktiv gestört wird. Vielleicht stellen Sie sich die Frage, warum man die Proteinstruktur zerstört. CIU-Experimente sind sehr effektiv, um die Struktur und Stabilität von Proteinen zu untersuchen und zu verstehen, was wir mittels MD in **Publikation III** und **Publikation IV** simulieren wollten. Während in CIU-Experimenten die Entfaltung durch Erhöhung der Kollisionsenergie herbeigeführt wird, wird diese bei Verwendung von MD durch aktives Erhöhen und Simulieren bei hohen Temperaturen induziert. In **Publikation III** taten wir dies für zwei dimere Proteine der Norwalk- und Kawasaki-Stämme, genetische Varianten, des Norovirus-Pathogens. Wir sahen, dass die Dimere bis zu 600 K stabil sind, sich aber bei höheren Temperaturen von 700+ K zunehmend entfalten. Darüber hinaus konnten wir potenzielle Bereiche der Proteine lokalisieren,

die sich am wahrscheinlichsten zuerst entfalten. Insgesamt zeigte sich eine höhere Stabilität des Kawasaki-Dimers im Vergleich zum Norwalk-Dimer. Unsere Ergebnisse liefern somit die Grundlage für nachfolgende experimentelle CIU-Studien dieser Proteine. Für **Publikation IV** haben wir Daten aus CIU-Experimenten des β-Fass-Proteins FhaC verwendet, um unsere CIU-MD-Simulationen zu informieren. FhaC ist ein Transportprotein, das in der Lipiddoppelschicht von Bakterien sitzt, welche für die Verursachung der Keuchhustenkrankheit verantwortlich sind. Als Transporter muss FhaC erhebliche dynamische Umlagerungen durchlaufen können um seine Aufgabe zu erfüllen, was schließlich mit CIU evaluiert wurde. Wir kombinierten experimentelle Daten mit unseren Simulationsdaten und konnten ein detailliertes Entfaltungsmuster für FhaC vorhersagen. Atomistische Informationen wurden enthüllt und ermöglichten eine detaillierte Abschätzung der Struktur von FhaC während des Entfaltungsprozesses. Dies stellt einen Mechanismus dar, der Vorhersagen auf der Grundlage experimenteller Daten widerspricht. So zeigen wir, dass CIU-MD-Simulationen sehr nützlich sind, um wichtige Zustände des Entfaltungsprozesses von Proteinen zu verstehen, zu verfolgen und zu lokalisieren. Sei es, um Informationen für zukünftige Experimente bereitzustellen oder um bestehende Forschungsdaten zu ergänzen.

Insgesamt erweitert die in dieser Dissertation vorgestellte Forschung das Verständnis von Proteinstrukturen mithilfe von MD-Simulationen. Die gewonnenen Informationen aus MD-Simulationen zeigen, wie wichtig es ist, die Proteinstrukturforschung in eine atomistische Perspektive zu stellen. Wir haben ausführlich diskutiert, welche Vorteile durch die Verwendung von MD erzielt und wo selbiges ergänzt werden könnte oder sollte. Zudem was schließlich in zukünftigen Studien verbessert werden könnte. Darüber hinaus zeigen unsere Ergebnisse die Vorteile der Kombination von MD-Simulationen mit experimentellen Methoden, insbesondere MS. Die hierdurch mögliche Lokalisierung genauer Details atomarer Auflösungen verdeutlicht das Potenzial von MD als Methode für die zukünftige Proteinstrukturforschung.

Acknowledgments

This thesis would not have been possible without the help, be it scientific or any other, of many people, and my gratitude can hardly be expressed in words—I will try anyway.

First and foremost, I would like to thank my supervisors. **Erik**, as my main supervisor, thank you for your constant help, support and guidance throughout these years. A lot that I present in this thesis would simply not have been possible without you. Grabbing a cup of coffee (preferably without cheese inside, *Tack*) and having a quick chat about science or how to make *nice* figures without pouring 30+ hours in it were just as welcome as discussing research with you. Moreover, thank you for your support outside of work, too. **Calle**, you were always interested in my work, but mostly, you were interested how I was doing and for that I would like to thank you especially. Your constant encouragements for me to do my best possible work and make coffee for your Wednesday's visits undoubtedly made me a better researcher. It still amazes me how almost every meeting with you turns at some point into a discussion about beer, and let's be honest, those are the best kind of discussions. **Charlotte**, it is still very impressive to see you delegate a multinational project, and the next moment talk with you about Metal music. And viruses, of course. Thank you for your help and honest feedback, and for allowing me to come to Hamburg and be part of your research group.

To the rest of the infamous Marklund group: *Grazie*, *Tack*, *Obrigado*! **Emiliano**, it is absolutely impressive how much you know about Physics, and your expertise in the field helped me a lot for understanding what the heck I was doing. Our discussions were always helpful, and thanks to you I learned a lot about programming, Italian food, and how to prevent your PC to go into sleep mode. Some of the research presented here would not have been possible without your input. **Louise**, thank you for your help, either for work or during teaching. Your efficiency of getting rid of candy in the office is stunning, and your enthusiasm was definitely a great addition to the team. I wish you all the best for your PhD. The atmosphere in the office was absolutely great, changing from discussions on where protons might migrate to the definition of bread in an instance. But, one person from the Marklund group is still missing that I'd like to thank. **Joana**, a big thank you to you, as you were such a great help and positive factor especially during my first year as PhD.

To all the members of the **MS SPIDOC team** as well: thank you. The regular meetings had such a friendly atmosphere, and made working in such a massive project an absolute joy. I am very happy I was able to be a part of this project. Thanks to the **Biophysics network**, too, for the great discussions, help and feedback that aided me in my journey.

There are many people at this department that I'd like to thank as well, which would end up in this becoming a novel probably. I'd prefer to thank you all in person, yet still would like to express here my thanks to a few very special people. Thank you to **Caroline**, **Susanne**, **Emily** and **Leandro** with all your respective partners—you made B7:3 especially fun! Thanks a lot, **Andrey**. It was an absolute blast teaching, grabbing a coffee or discussing Russian and German customs and traditions with you. **Leonidas**, thank you as well for the help and fun times at work or at the (spray)bar, your incredibly bad jokes, or whilst playing padel. And **Johanna**, thank you for being such a great and supportive friend over the past years. From having a coffee or lunch together so you can update me about the *juice* was just as much welcome as playing beer pong or volleyball with you. *Vielen Dank!* Always nice to have you around, both of you, and you'll always have a seat at my Monopoly table.

Of course there are people at home that I would like to thank as well. From my Bachelor studies, **Charly**. You are literally my *nemesis*, and I am very grateful to have you as a friend. Soon they will put our names in the one and only *Kabuff*, I know it. And obviously, thanks to the legendary **Werkstatt**. To all my friends at home: **Daniel** and **Kaja** (especially for the *very* necessary proof-reading of my German), **Phil** and **Lara**, **Jörg** and **Sabrina**, **Simon** and **Damla**, **Daniel** and **Nadine**, **Fabian** and **Ivonne**, **Maxi**, **Obbi**, **Hannes**, **Marcus**, **Chris**, **Velia** and most importantly **Tobi**—thank you so much for your love, support and help with anything, really. You have no idea how grateful I am that I can call every single one of you my friend. Love you guys, in particular you, **Tobi**.

To my second family, especially **Maureen** and **Wino**. Obviously, that includes all the rest of the aunts, uncles and cousins as well. I am very grateful to have you in my life, and am very much looking forward to soon see you more often and live closer to you. *Wie weet praten we binnenkort wel in het Nederlands!*

To my two older siblings, I love you two *Dumpfbacken*. **Nici**, thank you for being the best sister one can wish for. Our weekly calls to update each other were always amongst the highlights of the week. **Andre**, even with 10+ hours time difference, we still always managed to call and ramble and laugh and play video games (well, you played, I was too busy explaining it to you in the first place). You two gave me great support for which I will always be grateful. And of course, the three T's: **Tizian**, **Tamina** and **Tavio**. I am very proud of

all of you, and I cannot wait to see what is becoming of you. You all mean the world to me.

To the people that made it all possible, my parents. **Mama**, **Papa**, you are the reason I am who I am. Your love and encouragement, whilst having absolutely no idea of what I am doing other than chemistry and something with computers, always motivated me to be better and push forward. *Ich habe euch so viel zu Verdanken, was sich nicht in lediglich ein paar Worte und Zeilen niederschreiben lässt. Nur wegen euch habe ich erreichen können, was hier in dieser Arbeit steht - was ich im Allgemeinen bisher in meinem Leben erreicht habe. Hab euch von ganzem Herzen lieb, ihr Oldies.*

And the best for last. **Sarah**, you are my constant in live, my best friend and the person I love the most. What a journey it has been since I asked you to print my papers in Australia just to spend time with you. I am incredibly grateful for all your love and support, and how you keep inspiring and encouraging me to do my best. You were there for the ups and the downs, and I cannot wait to find out and share what our future together will hold for us.

References

[1] A Patriksson, E G Marklund, and D Van der Spoel. Protein structures under electrospray conditions. *Biochemistry*, 46(4):933–945, 2007.

[2] D L Nelson and M M Cox. *Principles of Biochemistry*. Macmillan Higher Education, 2017.

[3] C Mathews, K E Van Holde, and K G Ahern. *Biochemistry*. Addison Wesley Longman, 2000.

[4] S Pennazio. Viruses: are they living entities. In *Theoretical Biology Forum (poprzedni tytuł: Rivista di biologia–Biology Forum)*, volume 1, pages 45–56, 2011.

[5] L E Kay. NMR studies of protein structure and dynamics. *Journal of Magnetic Resonance*, 213(2):477–491, 2011.

[6] A P Turnbull and P Emsley. Studying protein–ligand interactions using X-ray crystallography. *Protein-Ligand Interactions: Methods and Applications*, pages 457–477, 2013.

[7] G L Glish and R W Vachet. The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery*, 2(2):140–150, 2003.

[8] R HH Van den Heuvel and A JR Heck. Native protein mass spectrometry: from intact oligomers to functional machineries. *Current Opinion in Chemical Biology*, 8(5):519–526, 2004.

[9] E G Marklund and J LP Benesch. Weighing-up protein dynamics: the combination of native mass spectrometry and molecular dynamics simulations. *Current Opinion in Structural Biology*, 54:50–58, 2019.

[10] G C Barrett and D T Elmore. *Amino acids and peptides*. Cambridge University Press, 1998.

[11] O W Griffith. β-amino acids: mammalian metabolism and utility as α-amino acid analogues. *Annual Review of Biochemistry*, 55(1):855–878, 1986.

[12] S L Miller. A production of amino acids under possible primitive earth conditions. *Science*, 117(3046):528–529, 1953.

[13] T M McCollom. Miller-urey and beyond: what have we learned about prebiotic organic synthesis reactions in the past 60 years? *Annual Review of Earth and Planetary Sciences*, 41:207–229, 2013.

[14] E T Parker, H J Cleaves, J P Dworkin, D P Glavin, M Callahan, A Aubrey, A Lazcano, and J L Bada. Primordial synthesis of amines and amino acids in a 1958 Miller H2S-rich spark discharge experiment. *Proceedings of the National Academy of Sciences*, 108(14):5526–5531, 2011.

[15] RJ-C Hennet, NG Holm, and MH Engel. Abiotic synthesis of amino acids under hydrothermal conditions and the origin of life: a perpetual phenomenon? *Naturwissenschaften*, 79(8):361–365, 1992.

[16] W L Marshall. Hydrothermal synthesis of amino acids. *Geochimica et Cosmochimica Acta*, 58(9):2099–2106, 1994.

[17] S W Fox. Thermal synthesis of amino acids and the origin of life. *Geochimica et Cosmochimica Acta*, 59(6):1213–1214, 1995.
[18] M H Engel and S A Macko. The stereochemistry of amino acids in the Murchison meteorite. *Precambrian Research*, 106(1-2):35–45, 2001.
[19] M P Bernstein, J P Dworkin, S A Sandford, G W Cooper, and L J Allamandola. Racemic amino acids from the ultraviolet photolysis of interstellar ice analogues. *Nature*, 416(6879):401–403, 2002.
[20] GM Munoz Caro, U J Meierhenrich, W A Schutte, B Barbier, A Arcones Segovia, H Rosenbauer, WH-P Thiemann, A Brack, and J M Greenberg. Amino acids from ultraviolet irradiation of interstellar ice analogues. *Nature*, 416(6879):403–406, 2002.
[21] U Meierhenrich. *Amino acids and the asymmetry of life: caught in the act of formation*. Springer, 2008.
[22] K Altwegg, H Balsiger, A Bar-Nun, JJ Berthelier, A Bieler, P Bochsler, C Briois, U Calmonte, M R Combi, H Cottin, et al. Prebiotic chemicals - amino acid and phosphorus - in the coma of comet 67P/Churyumov-Gerasimenko. *Science Advances*, 2(5):e1600285, 2016.
[23] Kensal E Van Holde, W Curtis Johnson, Pui Shing Ho, et al. *Principles of physical biochemistry*. Pearson/Prentice Hall Upper Saddle River, NJ, 2006.
[24] F W Lichtenthaler. Emil Fischer, his personality, his achievements, and his scientific progeny. *European Journal of Organic Chemistry*, 2002(24):4095–4122, 2002.
[25] G Genchi. An overview on D-amino acids. *Amino Acids*, 49:1521–1533, 2017.
[26] N Fujii, Y Kaji, and N Fujii. D-amino acids in aged proteins: Analysis and biological relevance. *Journal of Chromatography B*, 879(29):3141–3147, 2011.
[27] N Fujii. D-amino acids in living higher organisms. *Origins of Life and Evolution of the Biosphere*, 32:103–127, 2002.
[28] R C deL Milton, S C F Milton, and S B H Kent. Total chemical synthesis of a D-enzyme: The enantiomers of HIV-1 protease show reciprocal chiral substrate specificity. *Science*, 256(5062):1445–1448, 1992.
[29] J S Richardson. The anatomy and taxonomy of protein structure. *Advances in Protein Chemistry*, 34:167–339, 1981.
[30] V R Pattabiraman and J W Bode. Rethinking amide bond synthesis. *Nature*, 480(7378):471–479, 2011.
[31] H Kleinkauf and H von DÖHREN. Nonribosomal biosynthesis of peptide antibiotics. *European Journal of Biochemistry*, 192(1):1–15, 1990.
[32] P Zuber. Non-ribosomal peptide synthesis. *Current Opinion in Cell Biology*, 3(6):1046–1050, 1991.
[33] D Curtis, R Lehmann, and P D Zamore. Translational regulation in development. *Cell*, 81(2):171–178, 1995.
[34] S I Rattan, A Derventzi, and BF Clark. Protein synthesis, posttranslational modifications, and aging. *Annals of the New York Academy of Sciences*, 663:48–62, 1992.
[35] A I Bartlett and S E Radford. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nature Structural & Molecular Biology*, 16(6):582–588, 2009.

[36] W A Eaton, V Munoz, P A Thompson, E R Henry, and J Hofrichter. Kinetics and dynamics of loops, -helices, -hairpins, and fast-folding proteins. *Accounts of Chemical Research*, 31(11):745–754, 1998.
[37] C D Snow, H Nguyen, V S Pande, and M Gruebele. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420(6911):102–106, 2002.
[38] U Mayor, N R Guydosh, C M Johnson, J G Grossmann, S Sato, G S Jas, S MV Freund, D OV Alonso, V Daggett, and A R Fersht. The complete folding pathway of a protein from nanoseconds to microseconds. *Nature*, 421(6925):863–867, 2003.
[39] V Munoz and M Cerminara. When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. *Biochemical Journal*, 473(17):2545–2559, 2016.
[40] C R Matthews. Pathways of protein folding. *Annual Review of Biochemistry*, 62(1):653–683, 1993.
[41] N D Socci, J Nelson Onuchic, and P G Wolynes. Protein folding mechanisms and the multidimensional folding funnel. *Proteins: Structure, Function, and Bioinformatics*, 32(2):136–158, 1998.
[42] C M Dobson, A Šali, and M Karplus. Protein folding: a perspective from theory and experiment. *Angewandte Chemie International Edition*, 37(7):868–893, 1998.
[43] K A Dill and J L MacCallum. The protein-folding problem, 50 years on. *Science*, 338(6110):1042–1046, 2012.
[44] B Kuhlman and P Bradley. Advances in protein structure prediction and design. *Nature Reviews Molecular Cell Biology*, 20(11):681–697, 2019.
[45] K A Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, 1990.
[46] B Hardesty and G Kramer. Folding of a nascent peptide on the ribosome. *Progress in Nucleic Acid Research And Molecular Biology*, 2000.
[47] F U Hartl and M Hayer-Hartl. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295(5561):1852–1858, 2002.
[48] F U Hartl, A Bracher, and M Hayer-Hartl. Molecular chaperones in protein folding and proteostasis. *Nature*, 475(7356):324–332, 2011.
[49] C Hetz, K Zhang, and R J Kaufman. Mechanisms, regulation and functions of the unfolded protein response. *Nature reviews Molecular Cell Biology*, 21(8):421–438, 2020.
[50] P L Kastritis and A MJJ Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79):20120835, 2013.
[51] J Benach, S Atrian, R Gonzàlez-Duarte, and R Ladenstein. The refined crystal structure of Drosophila lebanonensis alcohol dehydrogenase at 1.9 å resolution. *Journal of Molecular Biology*, 282(2):383–399, 1998.
[52] L Pauling, R B Corey, and H R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.
[53] L Pauling and R B Corey. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proceedings of the*

National Academy of Sciences, 37(11):729–740, 1951.
[54] M Bajaj and T Blundell. Evolution and the tertiary structure of proteins. *Annual Review of Biophysics and Bioengineering*, 13(1):453–492, 1984.
[55] I M Klotz, NR Langebman, and DW Dahnall. Quaternary structure of proteins. *Annual Review of Biochemistry*, 39(1):25–62, 1970.
[56] JA McCammon. Protein dynamics. *Reports on Progress in Physics*, 47(1):1, 1984.
[57] M Kuddus. Chapter 1 - Introduction to food enzymes. In *Enzymes in Food Biotechnology*, pages 1–18. Academic Press, 2019.
[58] P K Robinson. Enzymes: principles and biotechnological applications. *Essays in Biochemistry*, 59:1, 2015.
[59] C Impraim, G Wang, and A Yoshida. Structural mutation in a major human aldehyde dehydrogenase gene results in loss of enzyme activity. *The American Journal of Human Genetics*, 34(6):837, 1982.
[60] AP Gimenez-Roqueplo, J Favier, P Rustin, JJ Mourad, PF Plouin, P Corvol, A Rötig, and X Jeunemaitre. The R22X mutation of the SDHD gene in hereditary paraganglioma abolishes the enzymatic activity of complex II in the mitochondrial respiratory chain and activates the hypoxia pathway. *The American Journal of Human Genetics*, 69(6):1186–1197, 2001.
[61] T M Redmond, E Poliakov, S Yu, JY Tsai, Z Lu, and S Gentleman. Mutation of key residues of RPE65 abolishes its enzymatic role as isomerohydrolase in the visual cycle. *Proceedings of the National Academy of Sciences*, 102(38):13658–13663, 2005.
[62] H Neurath, J P Greenstein, F W Putnam, and J A Erickson. The chemistry of protein denaturation. *Chemical Reviews*, 34(2):157–265, 1944.
[63] G Feller. Protein stability and enzyme activity at extreme biological temperatures. *Journal of Physics: Condensed Matter*, 22(32):323101, 2010.
[64] E Appella and C L Markert. Dissociation of lactate dehydrogenase into subunits with guanidine hydrochloride. *Biochemical and Biophysical Research Communications*, 6(3):171–176, 1961.
[65] T W Traut. Dissociation of enzyme oligomers: a mechanism for allosteric regulation. *Critical Reviews in Biochemistry and Molecular biology*, 29(2):125–163, 1994.
[66] S G Dahl, I Sylte, and A Westrheim Ravna. Structures and models of transporter proteins. *Journal of Pharmacology and Experimental Therapeutics*, 309(3):853–860, 2004.
[67] F B Jensen, A Fago, and R E Weber. Hemoglobin structure and function. *Fish Physiology*, 17:1–40, 1998.
[68] E Antonini. Interrelationship between structure and function in hemoglobin and myoglobin. *Physiological Reviews*, 45(1):123–170, 1965.
[69] B Giardina, I Messana, R Scatena, and M Castagnola. The multiple functions of hemoglobin. *Critical Reviews in Biochemistry and Molecular Biology*, 30(3):165–196, 1995.
[70] A N Schechter. Hemoglobin research and the origins of molecular medicine. *Blood, The Journal of the American Society of Hematology*, 112(10):3927–3938, 2008.
[71] B G Forget and H F Bunn. Classification of the disorders of hemoglobin. *Cold*

Spring Harbor Perspectives in Medicine, 3(2):a011684, 2013.
[72] P Sundd, M T Gladwin, and E M Novelli. Pathophysiology of sickle cell disease. *Annual Review of Pathology: Mechanisms of Disease*, 14:263–292, 2019.
[73] G R Serjeant. The natural history of sickle cell disease. *Cold Spring Harbor Perspectives in Medicine*, 3(10):a011783, 2013.
[74] K Bányai, M K Estes, V Martella, and U D Parashar. Viral gastroenteritis. *The Lancet*, 392(10142):175–186, 2018.
[75] C Trépo, H LY Chan, and A Lok. Hepatitis B virus infection. *The Lancet*, 384(9959):2053–2063, 2014.
[76] A J Eisfeld, G Neumann, and Y Kawaoka. At the centre: influenza A virus ribonucleoproteins. *Nature Reviews Microbiology*, 13(1):28–41, 2015.
[77] M M Lamers and B L Haagmans. SARS-CoV-2 pathogenesis. *Nature Reviews Microbiology*, 20(5):270–284, 2022.
[78] R A Harvey. *Microbiology*. Lippincott Williams & Wilkins, 2007.
[79] U F Greber. *Physical Virology*. Springer, 2019.
[80] P Forterre and D Prangishvili. The origin of viruses. *Research in Microbiology*, 160(7):466–472, 2009.
[81] S Smith, H Harmanci, Y Hutin, S Hess, M Bulterys, R Peck, B Rewari, A Mozalevskis, M Shibeshi, M Mumba, et al. Global progress on the elimination of viral hepatitis as a major public health threat: An analysis of WHO Member State responses 2017. *JHEP Reports*, 1(2):81–89, 2019.
[82] B X Wang and E N Fish. Global virus outbreaks: Interferons as 1st responders. In *Seminars in Immunology*, volume 43, page 101300. Elsevier, 2019.
[83] Y Zhou and L Chen. Twenty-year span of global coronavirus research trends: A bibliometric analysis. *International Journal of Environmental Research and Public Health*, 17(9):3082, 2020.
[84] K Ramphul and S G Mejias. Coronavirus disease: a review of a new threat to public health. *Cureus*, 12(3), 2020.
[85] Z Al-Aly, B Bowe, and Y Xie. Long COVID after breakthrough SARS-CoV-2 infection. *Nature Medicine*, 28(7):1461–1467, 2022.
[86] M Sadeghalvad, A H Mansourabadi, M Noori, S A Nejadghaderi, M Masoomikarimi, M Alimohammadi, and N Rezaei. Recent developments in SARS-CoV-2 vaccines: A systematic review of the current studies. *Reviews in Medical Virology*, 33(1):e2359, 2023.
[87] J H Kuhn. Virus taxonomy. *Reference Module in Life Sciences*, 2020.
[88] A MQ King, M J Adams, E B Carstens, and E J Lefkowitz. Virus taxonomy. *Ninth Report of the International Committee on Taxonomy of Viruses*, pages 486–487, 2012.
[89] G R Gelderblom. Structure and classification of viruses. In *Medical MicrobiologY*. University of Texas Medical Branch at Galveston, 1996.
[90] D Baltimore. Expression of animal virus genomes. *Bacteriological Reviews*, 35(3):235, 1971.
[91] M HV Van Regenmortel. Solving the species problem in viral taxonomy: recommendations on non-latinized binomial species names and on abandoning attempts to assign metagenomic viral sequences to species taxa. *Archives of Virology*, pages 1–7, 2019.

[92] S G Siddell, P J Walker, E J Lefkowitz, A R Mushegian, B E Dutilh, B Harrach, R L Harrison, S Junglen, N J Knowles, A M Kropinski, et al. Binomial nomenclature for virus species: a consultation. *Archives of Virology*, 165(2):519–525, 2020.

[93] D LD Caspar and A Klug. Physical principles in the construction of regular viruses. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 27, pages 1–24. Cold Spring Harbor Laboratory Press, 1962.

[94] R R Novoa, G Calderita, R Arranz, J Fontana, H Granzow, and C Risco. Virus factories: associations of cell organelles for viral replication and morphogenesis. *Biology of the Cell*, 97(2):147–172, 2005.

[95] A Zlotnick. Theoretical aspects of virus capsid assembly. *Journal of Molecular Recognition: An Interdisciplinary Journal*, 18(6):479–490, 2005.

[96] E Selivanovitch and T Douglas. Virus capsid assembly across different length scales inspire the development of virus-based biomaterials. *Current Opinion in Virology*, 36:38–46, 2019.

[97] E C Hartman, C M Jakobson, A H Favor, M J Lobba, E Álvarez-Benedicto, M B Francis, and D Tullman-Ercek. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nature Communications*, 9(1):1–11, 2018.

[98] C Uetrecht, I M Barbu, G K Shoemaker, E Van Duijn, and A JR Heck. Interrogating viral capsid assembly with ion mobility–mass spectrometry. *Nature Chemistry*, 3(2):126, 2011.

[99] J R Perilla, J A Hadden, B C Goh, C G Mayne, and K Schulten. All-atom molecular dynamics of virus capsids as drug targets. *The Journal of Physical Chemistry Letters*, 7(10):1836–1844, 2016.

[100] C R Bourne, MG Finn, and A Zlotnick. Global structural changes in hepatitis B virus capsids induced by the assembly effector HAP1. *Journal of Virology*, 80(22):11055–11061, 2006.

[101] Z Ambrose and C Aiken. HIV-1 uncoating: connection to nuclear entry and regulation by host proteins. *Virology*, 454:371–379, 2014.

[102] M F Hagan. Modeling viral capsid assembly. *Advances in Chemical Physics*, 155:1, 2014.

[103] B J Alder and T E Wainwright. Phase transition for a hard sphere system. *The Journal of Chemical Physics*, 27(5):1208–1209, 1957.

[104] W W Wood and JD Jacobson. Preliminary results from a recalculation of the Monte Carlo equation of state of hard spheres. *The Journal of Chemical Physics*, 27(5):1207–1208, 1957.

[105] J A McCammon, B R Gelin, and M Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–590, 1977.

[106] S A Hollingsworth and R O Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.

[107] A Hospital, J R Goñi, M Orozco, and J L Gelpí. Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry: AABC*, 8:37, 2015.

[108] E P Gross and E A Jackson. Kinetic models and the linearized Boltzmann equation. *The Physics of Fluids*, 2(4):432–441, 1959.

[109] L Verlet. Computer experiments on classical fluids. I. Thermodynamical

properties of Lennard-Jones molecules. *Physical Review*, 159(1):98, 1967.
[110] L Verlet. Computer experiments on classical fluids. II. Equilibrium correlation functions. *Physical Review*, 165(1):201, 1968.
[111] L Monticelli and E Salonen. *Biomolecular Simulations: Methods and Protocols*. Springer, 2013.
[112] Q Spreiter and M Walter. Classical molecular dynamics simulation with the Velocity Verlet algorithm at strong external magnetic fields. *Journal of Computational Physics*, 152(1):102–119, 1999.
[113] P Larsson, B Hess, and E Lindahl. Algorithm improvements for molecular dynamics simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(1):93–108, 2011.
[114] W C Swope, H C Andersen, P H Berens, and K R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, 1982.
[115] R W Hockney, SP Goel, and JW Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.
[116] S J Weiner, P A Kollman, D A Case, U C Singh, C Ghio, G Alagona, S Profeta, and P Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.
[117] B R Brooks, R E Bruccoleri, B D Olafson, D J States, S a Swaminathan, and M Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.
[118] W RP Scott, P H Hünenberger, I G Tironi, A E Mark, S R Billeter, J Fennen, A E Torda, T Huber, P Krüger, and W F Van Gunsteren. The GROMOS biomolecular simulation program package. *The Journal of Physical Chemistry A*, 103(19):3596–3607, 1999.
[119] C Oostenbrink, A Villa, A E Mark, and W F Van Gunsteren. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *Journal of Computational Chemistry*, 25(13):1656–1676, 2004.
[120] W L Jorgensen, D S Maxwell, and J Tirado-Rives. 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(45):11225–11236, 1996.
[121] C M Baker. Polarizable force fields for molecular dynamics simulations of biomolecules. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 5(2):241–254, 2015.
[122] Z Jing, C Liu, S Y Cheng, R Qi, B D Walker, JP Piquemal, and P Ren. Polarizable force fields for biomolecular simulations: Recent advances and applications. *Annual Review of Biophysics*, 48:371–394, 2019.
[123] J Huang and A D MacKerell Jr. Force field development and simulations of intrinsically disordered proteins. *Current Opinion in Structural Biology*, 48:40–48, 2018.
[124] P Robustelli, S Piana, and D E Shaw. Developing a molecular dynamics force

field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, 2018.
[125] B R Brooks, C L Brooks III, A D Mackerell Jr, L Nilsson, R J Petrella, B Roux, Y Won, G Archontis, C Bartels, S Boresch, et al. CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614, 2009.
[126] D A Case, T E Cheatham III, T Darden, H Gohlke, R Luo, K M Merz Jr, A Onufriev, C Simmerling, B Wang, and R J Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, 2005.
[127] R Salomon-Ferrer, D A Case, and R C Walker. An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013.
[128] H JC Berendsen, D Van der Spoel, and R Van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995.
[129] D Van Der Spoel, E Lindahl, B Hess, G Groenhof, A E Mark, and H JC Berendsen. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry*, 26(16):1701–1718, 2005.
[130] S Pronk, S Páll, R Schulz, P Larsson, P Bjelkmar, R Apostolov, M R Shirts, J C Smith, P M Kasson, D Van Der Spoel, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, 2013.
[131] M J Abraham, T Murtola, R Schulz, S Páll, J C Smith, B Hess, and E Lindahl. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1:19–25, 2015.
[132] C Kutzner, S Páll, M Fechner, A Esztermann, B L de Groot, and H Grubmüller. Best bang for your buck: GPU nodes for GROMACS biomolecular simulations, 2015.
[133] H Rakhshani, E Dehghanian, and A Rahati. Enhanced GROMACS: toward a better numerical simulation framework. *Journal of Molecular Modeling*, 25(12):1–8, 2019.
[134] A Liwo. *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*. Springer, 2019.
[135] G Ciccotti, M Ferrario, C Schuette, et al. Molecular dynamics simulation. *Entropy*, 16(233):1, 2014.
[136] P Bauer, B Hess, and E Lindahl. GROMACS 2022 manual, February 2022.
[137] Protein Data Bank. Protein Data Bank. *Nature New Biology*, 233:223, 1971.
[138] H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.
[139] P Nicholls. Introduction: the biology of the water molecule. *Cellular and Molecular Life Sciences CMLS*, 57(7):987–992, 2000.
[140] M Karplus and G A Petsko. Molecular dynamics simulations in biology. *Nature*, 347(6294):631–639, 1990.
[141] A Kukol et al. *Molecular Modeling of Proteins*. Humana Press/Springer, 2015.
[142] H JC Berendsen, JPM Postma, W F Van Gunsteren, ARHJ DiNola, and J R

Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, 1984.
[143] G Bussi, D Donadio, and M Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1):014101, 2007.
[144] S Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52(2):255–268, 1984.
[145] W G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A*, 31(3):1695, 1985.
[146] H C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393, 1980.
[147] M Parrinello and A Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.
[148] M Parrinello, A Rahman, and P Vashishta. Structural transitions in superionic conductors. *Physical Review Letters*, 50(14):1073, 1983.
[149] C Uetrecht, R J Rose, E Van Duijn, K Lorenzen, and A JR Heck. Ion mobility mass spectrometry of proteins and protein assemblies. *Chemical Society Reviews*, 39(5):1633–1655, 2010.
[150] E G Marklund, M T Degiacomi, C V Robinson, A J Baldwin, and J LP Benesch. Collision cross sections for structural proteomics. *Structure*, 23(4):791–799, 2015.
[151] S Moldoveanu and V David. Chapter 6 - Solvent extraction. In *Modern Sample Preparation for Chromatography*, pages 191–279. Elsevier, second edition edition, 2021.
[152] G Raabe and R J Sadus. Molecular dynamics simulation of the dielectric constant of water: The effect of bond flexibility. *The Journal of Chemical Physics*, 134(23):234501, 2011.
[153] D Van Der Spoel and P J Van Maaren. The origin of layer structure artifacts in simulations of liquid water. *Journal of Chemical Theory and Computation*, 2(1):1–11, 2006.
[154] T Darden, D York, and L Pedersen. Particle mesh Ewald: An N·log (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.
[155] M Diem and C Oostenbrink. The effect of different cutoff schemes in molecular simulations of proteins. *Journal of Computational Chemistry*, 41(32):2740–2749, 2020.
[156] M Baek, F DiMaio, I Anishchenko, J Dauparas, S Ovchinnikov, G R Lee, J Wang, Q Cong, L N Kinch, R D Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.
[157] J Griffiths. A brief history of mass spectrometry. *Analytical Chemistry*, 80(15):5678–5683, 2008.
[158] J R Yates III. A century of mass spectrometry: from atoms to proteomes. *Nature Methods*, 8(8):633–637, 2011.
[159] J J Thomson. Xl. cathode rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44(269):293–316, 1897.
[160] D A Williams. Fundamentals of mass spectrometry. In *Pharmacochemistry*

Library, volume 26, pages 19–45. Elsevier, 1997.
[161] U Ott. Interstellar grains in meteorites. *Nature*, 364(6432):25–33, 1993.
[162] E Grün, H Krüger, and R Srama. The dawn of dust astronomy. *Space Science Reviews*, 215:1–51, 2019.
[163] C J Hansen, L Esposito, AIF Stewart, J Colwell, A Hendrix, W Pryor, D Shemansky, and R West. Enceladus' water vapor plume. *Science*, 311(5766):1422–1425, 2006.
[164] J H Waite, C R Glein, R S Perryman, B D Teolis, B A Magee, G Miller, J Grimes, M E Perry, K E Miller, A Bouquet, et al. Cassini finds molecular hydrogen in the Enceladus plume: evidence for hydrothermal processes. *Science*, 356(6334):155–159, 2017.
[165] E Niyonsaba, J M Manheim, R Yerabolu, and H I Kenttämaa. Recent advances in petroleum analysis by mass spectrometry. *Analytical Chemistry*, 91(1):156–177, 2018.
[166] RS Borisov, LN Kulikova, and VG Zaikin. Mass spectrometry in petroleum chemistry (petroleomics). *Petroleum Chemistry*, 59:1055–1076, 2019.
[167] D B Liesenfeld, N Habermann, R W Owen, A Scalbert, and C M Ulrich. Review of mass spectrometry–based metabolomics in cancer research. *Cancer Epidemiology, Biomarkers & Prevention*, 22(12):2182–2201, 2013.
[168] A Macklin, S Khan, and T Kislinger. Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clinical Proteomics*, 17:1–25, 2020.
[169] C L Feider, A Krieger, R J DeHoog, and L S Eberlin. Ambient ionization mass spectrometry: recent developments and applications. *Analytical Chemistry*, 91(7):4266–4290, 2019.
[170] K Evans-Nguyen, A R Stelmack, P C Clowser, J M Holtz, and C C Mulligan. Fieldable mass spectrometry for forensic science, homeland security, and defense applications. *Mass Spectrometry Reviews*, 40(5):628–646, 2021.
[171] D T Snyder, C J Pulliam, Z Ouyang, and R G Cooks. Miniature and fieldable mass spectrometers: recent advances. *Analytical Chemistry*, 88(1):2–29, 2016.
[172] H Awad, M M Khamis, and A El-Aneed. Mass spectrometry, review of the basics: ionization. *Applied Spectroscopy Reviews*, 50(2):158–175, 2015.
[173] M Karas and F Hillenkamp. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Daltons. *Analytical Chemistry*, 60(20):2299–2301, 1988.
[174] J B Fenn, M Mann, C K Meng, S F Wong, and C M Whitehouse. Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71, 1989.
[175] R B Cole. *Electrospray ionization mass spectrometry: fundamentals, instrumentation, and applications*. Wiley-Interscience, 1997.
[176] M Wilm. Principles of electrospray ionization. *Molecular & Cellular Proteomics*, 10(7), 2011.
[177] L Konermann, E Ahadi, A D Rodriguez, and S Vahidi. Unraveling the mechanism of electrospray ionization, 2013.
[178] A JR Heck. Native mass spectrometry: a bridge between interactomics and structural biology. *Nature Methods*, 5(11):927–933, 2008.
[179] A C Leney and A JR Heck. Native mass spectrometry: what is in the name?

Journal of the American Society for Mass Spectrometry, 28(1):5–13, 2016.
[180] A D Rolland and J S Prell. Approaches to heterogeneity in native mass spectrometry. *Chemical Reviews*, 122(8):7909–7951, 2021.
[181] S Tamara, M A den Boer, and A JR Heck. High-resolution native mass spectrometry. *Chemical Reviews*, 122(8):7269–7326, 2021.
[182] John A McLean, Brandon T Ruotolo, Kent J Gillig, and David H Russell. Ion mobility–mass spectrometry: a new paradigm for proteomics. *International Journal of Mass Spectrometry*, 240(3):301–315, 2005.
[183] A B Kanu, P Dwivedi, M Tam, L Matz, and H H Hill Jr. Ion mobility–mass spectrometry. *Journal of Mass Spectrometry*, 43(1):1–22, 2008.
[184] B T Ruotolo, J LP Benesch, A M Sandercock, SJ Hyung, and C V Robinson. Ion mobility–mass spectrometry analysis of large protein complexes. *Nature Protocols*, 3(7):1139–1152, 2008.
[185] A Konijnenberg, A Butterer, and F Sobott. Native ion mobility-mass spectrometry and related methods in structural biology. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1834(6):1239–1256, 2013.
[186] M F Bush, I DG Campuzano, and C V Robinson. Ion mobility mass spectrometry of peptide ions: effects of drift gas and calibration strategies. *Analytical Chemistry*, 84(16):7124–7130, 2012.
[187] H Borsdorf, T Mayer, M Zarejousheghani, and G A Eiceman. Recent developments in ion mobility spectrometry. *Applied Spectroscopy Reviews*, 46(6):472–521, 2011.
[188] J LP Benesch, J A Aquilina, B T Ruotolo, F Sobott, and C V Robinson. Tandem mass spectrometry reveals the quaternary organization of macromolecular assemblies. *Chemistry & Biology*, 13(6):597–605, 2006.
[189] SJ Hyung, C V Robinson, and B T Ruotolo. Gas-phase unfolding and disassembly reveals stability differences in ligand-bound multiprotein complexes. *Chemistry & Biology*, 16(4):382–390, 2009.
[190] S M Dixit, D A Polasky, and B T Ruotolo. Collision induced unfolding of isolated proteins in the gas phase: past, present, and future. *Current Opinion in Chemical Biology*, 42:93–100, 2018.
[191] M Göth and K Pagel. Ion mobility–mass spectrometry as a tool to investigate protein–ligand interactions. *Analytical and Bioanalytical Chemistry*, 409:4305–4310, 2017.
[192] J D Eschweiler, J N Rabuck-Gibbons, Y Tian, and B T Ruotolo. CIUSuite: a quantitative analysis package for collision induced unfolding measurements of gas-phase protein ions. *Analytical Chemistry*, 87(22):11516–11522, 2015.
[193] J N Rabuck-Gibbons, J M Lodge, A K Mapp, and B T Ruotolo. Collision-induced unfolding reveals unique fingerprints for remote protein interaction sites in the kix regulation domain. *Journal of The American Society for Mass Spectrometry*, 30(1):94–102, 2018.
[194] J Dülfer, H Yan, M N Brodmerkel, R Creutznacher, A Mallagaray, T Peters, C Caleman, E G Marklund, and C Uetrecht. Glycan-induced protein dynamics in human norovirus P dimers depend on virus strain and deamidation status. *Molecules*, 26(8):2125, 2021.
[195] X Yan and C S Maier. Hydrogen/deuterium exchange mass spectrometry. *Mass Spectrometry of Proteins and Peptides: Methods and Protocols*, pages

255–271, 2009.
[196] G Sicoli, A Konijnenberg, J Guérin, S Hessmann, E Delnero, O Hernandez Alba, S Lecher, G Rouaut, L Müggenburg, H Vezin, et al. Large-scale conformational changes of FhaC provide insights into the two-partner secretion mechanism. *Frontiers in Molecular Biosciences*, page 752, 2022.
[197] R Neutze, R Wouts, D Van der Spoel, E Weckert, and J Hajdu. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature*, 406(6797):752–757, 2000.
[198] M J Bogan, W H Benner, S Boutet, U Rohner, M Frank, A Barty, M M Seibert, F Maia, S Marchesini, S Bajt, et al. Single particle X-ray diffractive imaging. *Nano Letters*, 8(1):310–316, 2008.
[199] H N Chapman, C Caleman, and N Timneanu. Diffraction before destruction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130313, 2014.
[200] M M Seibert, T Ekeberg, F RNC Maia, M Svenda, J Andreasson, O Jönsson, D Odić, B Iwan, A Rocker, D Westphal, et al. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*, 470(7332):78–81, 2011.
[201] A Aquila, A Barty, C Bostedt, S Boutet, G Carini, D DePonte, P Drell, S Doniach, KH Downing, T Earnest, et al. The linac coherent light source single particle imaging road map. *Structural Dynamics*, 2(4):041701, 2015.
[202] B Friedrich and DR Herschbach. On the possibility of orienting rotationally cooled polar molecules in an electric field. *Zeitschrift für Physik D Atoms, Molecules and Clusters*, 18:153–161, 1991.
[203] E G Marklund, T Ekeberg, M Moog, J LP Benesch, and C Caleman. Controlling protein orientation in vacuum using electric fields. *The Journal of Physical Chemistry Letters*, 8(18):4540–4544, 2017.
[204] A Kadek, K Lorenzen, C Uetrecht, et al. In a flash of light: X-ray free electron lasers meet native mass spectrometry. *Drug Discovery Today: Technologies*, 39:89–99, 2021.
[205] K A Dill, S B Ozkan, M S Shell, and T R Weikl. The protein folding problem. *Annual Review of Biophysics*, 37:289–316, 2008.
[206] Y Zhang. Protein structure prediction: when is it useful? *Current Opinion in Structural Biology*, 19(2):145–155, 2009.
[207] D B Kc. Recent advances in sequence-based protein structure prediction. *Briefings in Bioinformatics*, 18(6):1021–1032, 2017.
[208] D J Rigden. *From protein structure to function with bioinformatics*. Springer, 2009.
[209] Y Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.
[210] J Moult. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology*, 15(3):285–289, 2005.
[211] A W Senior, R Evans, J Jumper, J Kirkpatrick, L Sifre, T Green, C Qin, A Žídek, A WR Nelson, A Bridgland, et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins: Structure, Function, and*

Bioinformatics, 87(12):1141–1148, 2019.
[212] A W Senior, R Evans, J Jumper, J Kirkpatrick, L Sifre, T Green, C Qin, A Žídek, A WR Nelson, A Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.
[213] M Schauperl and R A Denny. AI-based protein structure prediction in drug discovery: Impacts and challenges. *Journal of Chemical Information and Modeling*, 62(13):3142–3156, 2022.
[214] L N Kinch, J Pei, A Kryshtafovych, R D Schaeffer, and N V Grishin. Topology evaluation of models for difficult targets in the 14th round of the Critical Assessment of Protein Structure Prediction (CASP14). *Proteins: Structure, Function, and Bioinformatics*, 89(12):1673–1686, 2021.
[215] J Jumper, R Evans, A Pritzel, R Green, M Figurnov, O Ronneberger, K Tunyasuvunakool, R Bates, A Žídek, A Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.
[216] K Tunyasuvunakool, J Adler, Z Wu, T Green, M Zielinski, A Žídek, A Bridgland, A Cowie, C Meyer, A Laydon, et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, 2021.
[217] M Varadi, S Anyango, M Deshpande, S Nair, C Natassia, G Yordanova, D Yuan, O Stroe, G Wood, A Laydon, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.
[218] M Varadi, N Bordin, C Orengo, and S Velankar. The opportunities and challenges posed by the new generation of deep learning-based protein structure predictors. *Current Opinion in Structural Biology*, 79:102543, 2023.
[219] J M Thornton, R A Laskowski, and N Borkakoti. AlphaFold heralds a data-driven revolution in biology and medicine. *Nature Medicine*, 27(10):1666–1669, 2021.
[220] J Jumper and D Hassabis. Protein structure predictions to atomic accuracy with AlphaFold. *Nature methods*, 19(1):11–12, 2022.
[221] R Nussinov, M Zhang, Y Liu, and H Jang. AlphaFold, artificial intelligence (AI), and allostery. *The Journal of Physical Chemistry B*, 126(34):6372–6383, 2022.
[222] B T Ruotolo and C V Robinson. Aspects of native proteins are retained in vacuum. *Current Opinion in Chemical Biology*, 10(5):402–408, 2006.
[223] A N Calabrese and S E Radford. Mass spectrometry-enabled structural biology of membrane proteins. *Methods*, 147:187–205, 2018.
[224] X Roger Liu, M M Zhang, and M L Gross. Mass spectrometry-based protein footprinting for higher-order structure analysis: fundamentals and applications. *Chemical Reviews*, 120(10):4355–4454, 2020.
[225] J Yang, I Anishchenko, H Park, Z Peng, S Ovchinnikov, and D Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.
[226] L Slabinski, L Jaroszewski, A PC Rodrigues, L Rychlewski, I A Wilson, S A

Lesley, and A Godzik. The challenge of protein structure determination – lessons from structural genomics. *Protein Science*, 16(11):2472–2482, 2007.
[227] S Boutet, L Lomb, G J Williams, T RM Barends, A Aquila, R B Doak, U Weierstall, D P DePonte, J Steinbrener, R L Shoeman, et al. High-resolution protein structure determination by serial femtosecond crystallography. *Science*, 337(6092):362–364, 2012.
[228] wwPDB–Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, 47(D1):D520–D528, 2019.
[229] K M Yip, N Fischer, E Paknia, A Chari, and H Stark. Atomic-resolution protein structure determination by cryo-EM. *Nature*, 587(7832):157–161, 2020.
[230] L Konermann, S Vahidi, and M A Sowole. Mass spectrometry methods for studying structure and dynamics of biological macromolecules. *Analytical Chemistry*, 86(1):213–232, 2014.
[231] J LP Benesch and C V Robinson. Mass spectrometry of macromolecular assemblies: preservation and dissociation. *Current Opinion in Structural Biology*, 16(2):245–251, 2006.
[232] Z Hall, A Politis, M F Bush, L J Smith, and C V Robinson. Charge-state dependent compaction and dissociation of protein complexes: insights from ion mobility and molecular dynamics. *Journal of the American Chemical Society*, 134(7):3429–3438, 2012.
[233] S Warnke, G von Helden, and K Pagel. Protein structure in the gas phase: the influence of side-chain microsolvation. *Journal of the American Chemical Society*, 135(4):1177–1180, 2013.
[234] T Meyer, X de la Cruz, and M Orozco. An atomistic view to the gas phase proteome. *Structure*, 17(1):88–95, 2009.
[235] P G Stockley, O Rolfsson, G S Thompson, G Basnak, S Francese, N J Stonehouse, S W Homans, and A E Ashcroft. A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. *Journal of Molecular Biology*, 369(2):541–552, 2007.
[236] V L Morton, E C Dykeman, N J Stonehouse, A E Ashcroft, R Twarock, and P G Stockley. The impact of viral RNA on assembly pathway selection. *Journal of Molecular Biology*, 401(2):298–308, 2010.
[237] K M ElSawy, L SD Caves, and R Twarock. The impact of viral RNA on the association rates of capsid protein assembly: bacteriophage MS2 as a case study. *Journal of Molecular Biology*, 400(4):935–947, 2010.
[238] Tom W Knapman, Victoria L Morton, Nicola J Stonehouse, Peter G Stockley, and Alison E Ashcroft. Determining the topology of virus assembly intermediates using ion mobility spectrometry–mass spectrometry. *Rapid communications in mass spectrometry*, 24(20):3033–3042, 2010.
[239] SH Chen and D H Russell. How closely related are conformations of protein ions sampled by IM-MS to native solution structures? *Journal of The American Society for Mass Spectrometry*, 26(9):1433–1443, 2015.
[240] NC Zachos. Gastrointestinal physiology and pathophysiology. In *Viral Gastroenteritis*, pages 1–21. Elsevier, 2016.
[241] R Pogan, J Dülfer, and C Uetrecht. Norovirus assembly and stability. *Current*

Opinion in Virology, 31:59–65, 2018.
[242] BV V Prasad, M E Hardy, T Dokland, J Bella, M G Rossmann, and M K Estes. X-ray crystallographic structure of the norwalk virus capsid. *Science*, 286(5438):287–290, 1999.
[243] E F Pettersen, T D Goddard, C C Huang, G S Couch, D M Greenblatt, E C Meng, and T E Ferrin. UCSF Chimera – a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.
[244] Y Matsushima, M Ishikawa, T Shimizu, A Komane, S Kasuo, M Shinohara, K Nagasawa, H Kimura, A Ryo, N Okabe, et al. Genetic analyses of GII. 17 norovirus strains in diarrheal disease outbreaks from December 2014 to March 2015 in Japan reveal a novel polymerase sequence and amino acid substitutions in the capsid region. *Eurosurveillance*, 20(26), 2015.
[245] A A Weiss and E L Hewlett. Virulence factors of Bordetella pertussis. *Annual Reviews in Microbiology*, 40(1):661–686, 1986.
[246] B Clantin, AS Delattre, P Rucktooa, N Saint, A C Méli, C Locht, F Jacob-Dubuisson, and V Villeret. Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily. *Science*, 317(5840):957–961, 2007.
[247] J Guérin, C Baud, N Touati, N Saint, E Willery, C Locht, H Vezin, and F Jacob-Dubuisson. Conformational dynamics of protein transporter FhaC: large-scale motions of plug helix. *Molecular Microbiology*, 92(6):1164–1176, 2014.
[248] T Maier, B Clantin, F Gruss, F Dewitte, F Delattre, ASand Jacob-Dubuisson, S Hiller, and V Villeret. Conserved Omp85 lid-lock structure and substrate recognition in FhaC. *Nature Communications*, 6(1):7452, 2015.
[249] Y Kawai and A Moribayashi. Characteristic lipids of Bordetella pertussis: simple fatty acid composition, hydroxy fatty acids, and an ornithine-containing lipid. *Journal of Bacteriology*, 151(2):996–1005, 1982.
[250] S Jo, T Kim, V G Iyer, and W Im. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, 29(11):1859–1865, 2008.
[251] S Jo, T Kim, and W Im. Automated builder and database of protein/membrane complexes for molecular dynamics simulations. *PLOS ONE*, 2(9):e880, 2007.
[252] J Hulshof and C Ponnamperuma. Prebiotic condensation reactions in an aqueous medium: a review of condensing agents. *Origins of Life*, 7:197–224, 1976.
[253] J A Lukin and C Ho. The structure- function relationship of hemoglobin in solution at atomic resolution. *Chemical Reviews*, 104(3):1219–1230, 2004.
[254] D G Archer and P Wang. The dielectric constant of water and Debye-Hückel limiting law slopes. *Journal of [hysical and Chemical Reference Data*, 19(2):371–411, 1990.
[255] R Pogan, C Schneider, R Reimer, G Hansman, and C Uetrecht. Norovirus-like VP1 particles exhibit isolate dependent stability profiles. *Journal of Physics: Condensed Matter*, 30(6):064006, 2018.
[256] H Geng, F Chen, J Ye, and F Jiang. Applications of molecular dynamics simulation in structure prediction of peptides and proteins. *Computational and*

Structural Biotechnology Journal, 17:1162–1170, 2019.
[257] R F Alford, A Leaver-Fay, J R Jeliazkov, M J O'Meara, F P DiMaio, H Park, M V Shapovalov, P D Renfrew, V K Mulligan, K Kappel, et al. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017.
[258] J A Maier, C Martinez, K Kasavajhala, L Wickstrom, K E Hauser, and C Simmerling. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.
[259] H Ode, M Nakashima, S Kitamura, W Sugiura, and H Sato. Molecular dynamics simulation in virus research. *Frontiers in Microbiology*, 3:258, 2012.
[260] DC Rapaport. Molecular dynamics study of T = 3 capsid assembly. *Journal of Biological Physics*, 44(2):147–162, 2018.
[261] P L Freddolino, A S Arkhipov, S B Larson, A McPherson, and K Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, 2006.
[262] E Tarasova, V Farafonov, R Khayat, N Okimoto, T S Komatsu, M Taiji, and D Nerukh. All-atom molecular dynamics simulations of entire virus capsid reveal the role of ion distribution in capsid's stability. *The Journal of Physical Chemistry Letters*, 8(4):779–784, 2017.
[263] R Asor, D Khaykelson, O Ben-nun Shaul, A Oppenheim, and U Raviv. Effect of calcium ions and disulfide bonds on swelling of virus particles. *ACS Omega*, 4(1):58–64, 2019.
[264] M Zink and H Grubmüller. Primary changes of the mechanical properties of southern bean mosaic virus upon calcium removal. *Biophysical Journal*, 98(4):687–695, 2010.
[265] D SD Larsson, L Liljas, and D Van der Spoel. Virus capsid dissolution studied by microsecond molecular dynamics simulations. *PLOS Computational Biology*, 8(5):e1002502, 2012.
[266] K Wołek and M Cieplak. Self-assembly of model proteins into virus capsids. *Journal of Physics: Condensed Matter*, 29(47):474003, 2017.
[267] G Zhao, J R Perilla, E L Yufenyuy, X Meng, B Chen, J Ning, J Ahn, A M Gronenborn, K Schulten, C Aiken, et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature*, 497(7451):643–646, 2013.
[268] R G Huber, J K Marzinek, D A Holdbrook, and P J Bond. Multiscale molecular dynamics simulation approaches to the structure and dynamics of viruses. *Progress in Biophysics and Molecular Biology*, 128:121–132, 2017.
[269] J A Roberts, M J Kuiper, B R Thorley, P M Smooker, and A Hung. Investigation of a predicted N-terminal amphipathic α-helix using atomistic molecular dynamics simulation of a complete prototype poliovirus virion. *Journal of Molecular Graphics and Modelling*, 38:165–173, 2012.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2264

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-500274

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2023