D I S S E R T A T I O N

submitted to the
Combined Faculties for Natural Sciences and Mathematics
of the Ruperto-Carola University of Heidelberg, Germany

for obtaining the degree of
Doctor of Natural Sciences

presented by
Diplom-Physiker Lars Meinhold
born in Karl-Marx-Stadt

Disputation: 2nd November 2005

Crystalline Protein Dynamics

A Simulation Analysis of *Staphylococcal* Nuclease

Referees: Prof. Dr. Jeremy C. Smith
PD Dr. Ilme Schlichting

Z U S A M M E N F A S S U N G

Das Verständnis der Bewegungen in Proteinkristallen verspricht sowohl Einblicke in die funktionsrelevante Proteindynamik als auch eine Verbesserung von Modellen zur Strukturbestimmung mittels Beugungsexperimenten. Diese Arbeit präsentiert Molekulardynamik (MD) Simulationen kristalliner *Staphylococcus* Nuklease und wertet diese in Bezug auf Fluktuationen, Korrelationen, diffuse Röntgenbeugung (RB), Unordnung und Modelle für Proteindynamik aus. Konvergenzeigenschaften dynamischer Observablen werden bestimmt. Aus der logarithmischen Abhängigkeit des R Faktors mit der experimentellen diffusen RB von der Simulationszeit wird Konvergenz innerhalb $\approx$1 μs für die Kovarianz-Matrix abgeschätzt, welche die diffuse RB bestimmt. Der dynamische Ursprung der diffusen RB wird anhand von liquid-artigen und kollektiven Bewegungsmodellen untersucht. Eine gleichförmige, fast isotrope Schale diffuser RB bei q=0,28 Å^{-1} stammt zu gleichen Teilen von Korrelationen in der Dynamik benachbarter Wassermoleküle und interner Proteinbewegung, letztere dominiert von Fluktuationen in der α-Helix-Steigung und dem β-Faltblatt-Abstand. Der Schale überlagert sind intensive Modulationen, welche durch wenige, langsam konvergierende (>10 ns) kollektive Bewegungen verursacht werden. Diese Modulationen werden konkreten kollektiven Proteinbewegungen zugeordnet, welche auch die möglicherweise funtionsrelevante Deformation des aktiven Zentrums beschreiben. Die Dynamik einzelner kollektiver Moden wird mittels Brownscher Bewegung beschrieben. Moden mit Frequenzen unterhalb 0,55 THz sind überdämpft, während die Mehrzahl (98,6%) unterdämpft schwingt. MD Simulationen im Druckbereich von 1 bar bis 15 kbar zeigen eine qualitative Veränderung interner Proteinbewegungen bei $\approx$4 kbar, manifestiert durch zwei lineare Regime der mittleren atomaren Weglänge mit auf die Hälfte vermindertem Gradienten oberhalb 4 kbar. Im wesentlich führt die Druckerhöhung zu einem Verlust kollektiver Bewegungen mit effektiven Frequenzen unter 2 THz.

A B S T R A C T

Understanding motions in protein crystals is likely to furnish insight into functional protein dynamics and will improve models for refinement against diffraction data. In this thesis, molecular dynamics (MD) simulations of crystalline *Staphylococcal* nuclease are reported and analysed in terms of fluctuations, correlations, X-ray diffuse scattering, disorder and models of protein motion. Convergence properties of dynamical quantities are determined. The logarithmic dependence on the simulation length of the R factor with the experimental X-ray diffuse scattering, which is determined by the atomic displacement variance-covariance matrix, is extrapolated to predict a convergence time for the whole variance-covariance matrix of $\approx$1 μs. The dynamical origin of the X-ray diffuse scattering is investigated using models of liquid-like and collective motion. A smooth, nearly-isotropic scattering shell at $q = 0.28$ Å^{-1} originates from equal contributions from correlations in nearest-neighbour water molecule dynamics and from internal protein motions, the latter consisting of α-helix pitch and inter-β-strand fluctuations. Superposed on the shell are intense features that originate from a very small number of slowly-varying (>10 ns) collective motions. The individual features are assigned to specific collective motions in the protein, and some of these describe potentially functional active-site deformations. The dynamics along each collective mode is described using Brownian dynamics. Modes with frequencies below 0.55 THz are overdamped while the majority (98.6%) of modes perfom underdamped vibrations. MD simulations over the pressure range 1 bar to 15 kbar reveal a qualitative change in the internal protein motions at $\approx$4 kbar. This change involves the existence of two linear regimes in the mean-square displacement for internal protein motion with a twofold decrease in the slope above 4 kbar. The major effect of the pressure increase on the dynamics is a loss of large-amplitude, collective protein modes below 2 THz effective frequency.

LIST OF PUBLICATIONS

Those publications marked with an * originate from this thesis.

[11] L Meinhold, D Clement, M Tehei, RM Daniel, JL Finney & JC Smith.
Dynamics of thermophilic and mesophilic dihydrofolate reductase studied using incoherent elastic neutron scattering. *In preparation.*

[10]* L Meinhold & JC Smith.
Collective dynamics in protein crystals: a simulation analysis of X-ray diffuse scattering by *Staphylococcal* nuclease. *To be submitted to Biophysical Journal.*

[9]* L Meinhold & JC Smith.
Correlated dynamics determining X-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation. *Submitted to Physical Review Letters.*

[8]* L Meinhold & JC Smith.
Pressure-dependent transition in protein dynamics at ≈4 kbar revealed by molecular dynamics simulation. *Submitted to Physical Review E.*

[7]* L Meinhold & JC Smith.
Fluctuations and Correlations in Crystalline Protein Dynamics: A Simulation Analysis of *Staphylococcal* Nuclease. *Biophysical Journal* **88**:2554, 2005.

[6] D Sengupta, L Meinhold, D Langosch, GM Ullmann & JC Smith.
Understanding the Energetics of Helical Peptide Orientation in Membranes. *Proteins* **58**:913, 2005.

[5]* L Meinhold, S Lammers, T Becker & JC Smith.
Convergence properties of X-ray scattering calculated from protein crystal molecular dynamics simulations. *Physica B* **350**:127, 2004.

[4] L Meinhold.
Realistic description of single cluster calcium signalling. *Mathematical Modelling & Computing in Biology and Medicine*,V.Capasso (Ed.), The MIRIAM Project Series, pp. 200-205, ESCULAPIO Pub. Co., Bologna, Italy, 2003.

[3] B Lindner, L Meinhold & L Schimansky-Geier.
Two-State Description of Stochastic Excitability. *AIP Conference Proceedings* **665(1)**:443, 2003.

[2] L Meinhold & L Schimansky-Geier.
Analytic description of stochastic calcium-signaling periodicity. *Physical Review E* **66**:050901(R), 2002.

[1] L Meinhold.
Stochastic Oscillations in Cytosolic Calcium Concentration. *Diploma Thesis*, Institute of Physics, Humboldt-University Berlin, Germany, 2002.

Acknowledgements

It is my pleasure to mention and acknowledge the help and contributions of many people to both the progress of my work and the manuscript of this thesis. Of course, I bear the sole responsibility for any shortcomings contained therein.

First and foremost I would like to thank my supervisor Jeremy C. Smith for giving me the opportunity to work on this subject, letting me explore my own ideas and guiding the progress of my work by many discussions. I also thank him and all the present and past members of the Computational Molecular Biophysics Group for providing the friendly atmosphere and stimulating discussions.

Torsten Becker, Alexander L. Tournier and Andrea C. Vaiana introduced me to (much more than) the field of molecular dynamics and, together with G. Matthias Ullmann, aided my first steps of research and towards the cafeteria – for that and the many inspiring thoughts and views I am very grateful.

Durba Sengupta contributed to the progress of my work by being always available for questions and discussions, helping in the preparation of manuscripts and proof-reading this thesis, and collaborating with me. Above all, she has constantly been a good friend with a cheerful laughter and I owe her my sincere thanks.

Bogdan Costescu designed, set up and maintained the computational infra-structure of the lab and this thesis would not have been possible without his patient and extensive help – thank you very much. I also thank Stefan Friedel for support on the Helics cluster and Frank Noé for assistance in programming.

The three nicest graphics of this thesis are noble offerings from Alexander L. Tournier, Andreea D. Gruia and Nicolas Calimet. Cheers. Cheers. Cheers. In the preparation of many other graphics I profited from the counsel of Thomas Splettstößer. Cheers/Cheers!

I gratefully acknowledge Sol Gruner for providing the X-ray diffuse scattering data, and I would like to thank Lutz Schimansky-Geier, Vandana Kurkal, Kei Moritsugu and Zoe Cournia for many helpful discussions.

I thank Edda Kloppmann and Torsten Becker for proof-reading and polishing the final version of this thesis, and for all the many discussions, in particular for those outside the field of science.

I would like to thank my family and friends for support throughout my studies, for helping me grow and for making the not-working hours great fun.

Überreden will ich niemanden, dachte er. Wenn die anderen nur mich nicht überreden. Sie sollen mich fragen und gespannt auf meine Antwort warten. Dahin muß ich es bringen, das ist alles.

Sten Nadolny
Die Entdeckung der Langsamkeit

Contents

Chapter 1

Introduction

Advances in science and technology have often been facilitated by the understanding and utilisation of matter in its condensed state. At macroscopic length scales, the static and dynamic properties of solids and liquids are essentially fully described by the classical theories of thermodynamics, hydrodynamics and elasticity, which were firmly established in the nineteenth century. During the first half of the twentieth century, access to the microscopic, *i.e.*, atomic length scales has been provided by new experimental techniques, such as scattering and spectroscopy, which also laid the foundation for the triumphant success of quantum mechanics [1–3].

As a result of the immense mathematical simplification due to lattice-translational and point-group symmetries, the analyses of condensed matter were often restricted to solids in the (nearly) perfect crystalline state. But condensed matter in disordered phases – such as steel, glass or water – is far more abundant, and of no less technological value, than the idealised single crystal. Therefore, during the second half of the twentieth century, research concerning the structure and dynamics of condensed matter in the disordered phase has been intensified. Although the resulting descriptions are often fragmentary and phenomenological, lacking the mathematical depth and rigour of conventional solid state physics, a variety of valuable concepts has emerged – such as order parameters and the renormalisation group from the investigation of critical phenomena – and has been fruitfully expanded into other fields of physics and science.

At the same time, advances in condensed matter physics have often been stimulated by interdisciplinary approaches which, in many cases, developed soon after into independent, broad research fields. One of these fields, into which the author would like to classify this thesis, is biophysics. The previously not encountered complexity of the living world coupled with its nearly infinite diversity both complicates and vitalises this joint-venture of physics and biology. About 90% of the biological matter is polymeric (polypeptides alias proteins, polynucleotides alias DNA/RNA and polysaccharides alias starch/cellulose) and their molecular constitution, structure and function, interaction with other molecules and molecular properties appear as the central scientific issues in biophysics. The use of computer simulations has been invaluable for these investigations. Nowadays, due to the rapid technological and methodological advances, computer simulations not only assist in understanding experimentally observed phenomena, but also, to an increasing degree, predict new phenomena awaiting their experimental verification.

This thesis contributes to the endeavour of relating protein structures to their function by investigating the dynamics of a model protein, *Staphylococcal* nuclease, in a crystalline environment. Understanding motions in protein crystals is likely to furnish insight into functional protein dynamics and will improve models for refinement against diffraction data. In particular, molecular

dynamics (MD) computer simulations are used to reproduce and interpret a recently performed X-ray diffuse scattering experiment. Furthermore, the dependence of intra-molecular protein dynamics on the applied hydrostatic pressure is investigated. In the following section, the wider context of this thesis, *i.e.* the protein structure–dynamics–function relation, is discussed. Subsequently, the more specific context of X-ray diffuse scattering from protein crystals is introduced and an outline of this thesis is provided.

1.1 Protein Structure–Dynamics–Function Relation

Protein function is commonly believed to depend on both the three-dimensional conformation and flexibility of the polypeptide chain as well as the protein environment [4–6]. However, a comprehensive dynamical description of how the specific fold of the protein, in conjunction with its environment, facilitates its particular function has not yet been achieved. In this section, a brief overview of the relation between protein structure, dynamics and function is given. The selection of topics is inevitably incomplete and reflects the personal view and interests of the author – many more aspects are described in a number of excellent reviews [7–11].

Structure and Motions

Proteins are linear polypeptides with the sequence of residues chosen from a pool of at least 22 genetically encoded amino acids [12] and comprise between approximately 50 and 1000 residues[1], leading to macro-molecules on the length scales of approximately 1 to 10 nm [5, 6]. For any given polypeptide chain, rotations around covalent bonds in the peptide backbone and side chains make possible a merely unlimited variety of potential conformations. The prediction of the (ensemble of) conformations a protein will adopt in its native environment from the primary amino-acid sequence remains one of the fundamental goals of biophysical research. Until this *protein folding problem* is solved, assistance has to be provided by elaborate experiments. The predominant present-day experimental methods for determining high-resolution three-dimensional protein structures are X-ray crystallography and nuclear magnetic resonance (NMR) – the former having been pioneered by Watson and Crick [13][2], Bragg and Perutz [14–20] and Kendrew [21, 22], the latter having been devised by Wütrich [23]. Although both methods investigate proteins in very different non-native environments, *i.e.* either in a highly-ordered crystal or a disordered high-density solution, comparisons of X-ray crystallographic and NMR results indicate only minor structural differences. However, a deficiency of both methods is that they yield the average structure[3] of a macroscopic number of proteins while providing only limited information on the dynamics. For X-ray crystallography, prospects to overcome this limitation are discussed in Section 1.2.

[1] Molecules composed of less than 50 amino acids are commonly referred to as polypeptides. Examples for very small proteins are the extensively studied crambin and bovine pancreatic trypsin inhibitor with 46 and 58 residues, respectively. The largest known protein to date is the giant muscle protein titin with 26,926 amino acids in a single chain. The majority of proteins, however, comprises between 100 to 300 residues.

[2] In fact, the X-ray data collection and processing was done by Franklin and Wilkins. Franklin found, but did not then publish, that the DNA had helical characteristics. Using Franklin's data, Watson and Crick took the crucial conceptual step of suggesting that DNA is composed of two anti-parallel polynucleotide strands.

[3] More precisely: in X-ray crystallography several conformations may be observed and their relative populations (occupancies) are determined; NMR experiments result in an ensemble of structures (usually ≈10) compatible with the observed constraints.

Intrinsic protein motions occur on the length and time scales of 10^{-3} to 1 nm and 10^{-15} to 10^{4} s, respectively [4, 24]. These motions can be decomposed into collective modes [25, 26], ranging from high-frequency, small amplitude local displacements within rather rigid groups of atoms, to low-frequency larger-amplitude relative motions between these rigid groups [27]. The more localised motions are mostly harmonic vibrations not directly related to function. However, including these degrees of freedom significantly decreases energy barriers for the slower, more collective motions, which show significant anharmonic character that may be directly related to functional motions [25, 28, 29]. Determining the exact nature of this coupling between slow and fast motions is the subject of ongoing research [30].

Functional Diversity

Proteins can be categorised as catalytic, structural, transport, storage, motile, regulatory or defence proteins according to their function [31]. Often, this functional diversity is reflected in differences in the protein structure and its flexibility although examples exist, *e.g.* lysozyme and α-lactalbumin[4], where proteins of similar structure undertake very different functions.

In the case of catalytic proteins, or enzymes, the induced-fit model [35] has been shown to adequately describe substrate binding [36, 37]. In the induced-fit model, the binding of a substrate causes a specific change in the enzyme conformation that brings the catalytic groups into the proper orientation required for the reaction. In general, the binding of non-substrates does not cause this specific change.[5] The conformational changes in the induced-fit model are usually local, mostly involving the residues forming the binding pocket and active site. The function of regulatory or transport proteins, however, often requires the collective displacement of large groups of residues being far from the ligand binding site. For example, the conformational changes triggered by a ligand binding to a membrane-bound receptor are relayed across the membrane, *i.e.* over approximately 4 nm [38–40]. Similar long-range collective displacements have been observed in motile proteins, such as kinesin and myosin which, upon the hydrolysis of ATP, generate large-scale collective motion using mechanical amplifiers [41].

The above examples illustrate the importance of collective displacements for protein function. However, the principles that determine the frequencies, forms and damping characteristics of these collective modes are not yet well understood. Moreover, it is important to know whether equilibrium motions in proteins can be correlated over long distances or whether anharmonicity and damping destroy such effects.

Energy Landscape

The concept of a hierarchical protein energy landscape was introduced to explain the temperature dependence of the rebinding kinetics of the ligands carbon monoxide and dioxygen after photodissociation to myoglobin [42, 43]. In particular, the perception that the protein native state

[4] Whereas lysozyme catalyses the hydrolysis of bacterial cell wall polysaccharides [32] α-lactalbumin promotes the conversion of galactosyltransferase to lactose synthase and is essential for milk production [33]. However, the analysis of phylogenetic trees has suggested that lysozyme and α-lactalbumin descend from the same 'ancestor' protein and evolved differently after the divergence of the lineages of birds and mammals [34].

[5] The binding of substrate analogs and competitive inhibitors, however, often cause changes closely similar to those induced by the substrate.

merely represents an ensemble of a large number of almost iso-energetic substates, has provided much insight into the functional mechanisms of proteins [44–47]. The substates are separated by a distribution of barrier heights and transitions may occur due to thermal fluctuations or non-equilibrium processes.

A much-studied phenomenon demonstrating the effect of conformational substates on protein dynamics is the temperature-dependent dynamical transition[6] manifested in a non-linear increase in the atomic mean-square displacement at approximately 200 K, which has been observed experimentally using neutron scattering [48], Mössbauer spectroscopy [49] and X-ray crystallography [50]. Recently, molecular dynamics computer simulations showed that the temperature-dependent dynamical transition involves a solvent-driven activation of anharmonic protein dynamics [51, 52]. For some enzymes it has also been shown that activity ceases below the dynamical transition temperature [50, 53, 54] but whether these observations reflect a general relationship is disputed [55, 56].

Protein–Solvent Interaction

Protein and solvent mutually influence their structural and dynamical properties. Upon dehydration, protein flexibility (*i.e.* the average atomic fluctuation) is significantly reduced [57]. The reduction may, in part, be due to increased interactions between surface side chains (leading to higher energy barriers for transitions between conformational substates) due to the loss in dielectric screening of protein surface charges by water molecules [58]. Conformational flexibility of surface side chains is particularly important for enzyme function. Indeed, enzyme activity has been shown to cease below a minimal hydration of approximately 0.2 g/g (H_2O/protein) [59]. However, very recent results using vapour substrates indicated that enzyme activity is present at hydrations as low as 0.03 g/g (H_2O/protein) [60]. The temperature dependence of overall protein motion is also effected by low hydrations. Some neutron scattering results suggest that the dynamical transition is absent for dry proteins [53, 54, 61]. However, in a study of oxymyoglobin in a frozen aqueous environment the transition was found using Mössbauer spectroscopy [62].

It is only recently, that more attention has been turned to the reverse effect, *i.e.* how the protein influences the solvent. In the vicinity of a protein, *i.e.* within the distance of ≈7 Å of the protein surface, corresponding to two hydration layers, the structural and dynamical properties of water are significantly different from those in the bulk water [63–67]. In particular, the protein surface water is of higher density than the bulk [66, 68], shows retarded rotational and translational relaxations and is in a dispersive diffusion (subdiffusive) regime [64, 67].

Protein Stability

The biosphere is not restricted to the Earth's surface but ranges from the floor of the oceans to the stratosphere, and encompasses also inhospitable areas such as the polar regions, volcanic sites and saturated salt lakes. Organisms living in a particular environment, and thus their proteins, are not only adopted to, but also dependent on the local conditions of temperature,

[6] Often, this effect is also called the glass transition, closely following the notation of condensed matter physics and the class of substances for which it has been observed first.

hydrostatic pressure, salinity and pH. This mesophilic[7] character of proteins is due to a subtle balance of stabilising and destabilising interactions. Even minor changes in the environmental variables may disturb this balance and potentially lead to dynamical, functional and/or structural transitions and, eventually, denaturation or unfolding [7].

1.2 X-Ray Bragg and Diffuse Scattering

The Brookhaven Protein Data Bank (PDB) currently holds about 30,800 (about 10,800 with less than 90% sequence identity) protein structures [69], about 87% of which were determined using X-ray crystallography. In X-ray crystallography, the periodic crystal structure leads scattered X-rays to interfere constructively giving rise to intense peaks in reciprocal space: the Bragg scattering. The distribution and intensities of Bragg peaks can be converted into an average electron density map of the crystal unit cell which can then be interpreted in terms of atomic positions [6, 70–72].

X-ray Bragg diffraction from protein crystals is usually interpreted using one coordinate set that represents the average structure of a macroscopic number of protein molecules. However, a real protein crystal is not perfectly periodic. The disorder present can be static, such as arises from crystal faults, impurities and structural inhomogeneity, or dynamic, *i.e.*, due to motions present at finite temperature. The disorder is represented in a refinement by occupancies (if more than one coordinate set is actually used for refinement) and B factors, the latter describing average fluctuations of the individual atoms around their average positions. B factors contain static and dynamic components [43, 73]. Part of the static disorder may be temperature-independent. However, as the temperature of a protein is lowered some of the dynamic disorder may become static as proteins freeze into structurally inhomogeneous conformational substates. Separation of static from dynamic disorder is therefore non-trivial. Static and dynamics disorder manifest deviations from crystal periodicity, and thus decrease the intensity scattered into the Bragg peaks and distribute it elsewhere in reciprocal space – this latter intensity constitutes the so-called diffuse scattering [27, 74–82]. Diffuse scattering from molecular crystals has been reviewed [83–86]. Whereas Bragg diffraction provides no direct information on the collective motions likely to be related to protein function [87], these can, in principle, be probed using X-ray diffuse scattering, because diffuse scattering directly depends on the relative displacements of different atoms. Thus, diffuse scattering provides information complementary to that obtained by conventional and time-resolved crystallography [88–90].

The goal of time-resolved protein crystallography is to determine the structures of reaction intermediates in the crystal using Bragg diffraction. This goal can, in principle, be achieved by the following two approaches. In the *time-resolved* approach the reaction is initiated, *e.g.* a photolysis by an intense laser flash, and subsequently a time-delayed single X-ray pulse is used to obtain the diffraction pattern [91]. The time resolution is given by the pulse length, thus favouring the polychromatic or Laue method which allows very rapid data acquisition [92]. In the *kinetic* approach, the population of the intermediate is increased either by slowing the reaction by physical or chemical means, or by running the reaction in the steady-state [89]. The structure of the intermediate is then often determined using monochromatic X-rays which provide more reliable data [89]. In both approaches, the observed Bragg diffraction pattern is decomposed

[7] Here, the term mesophilic is not used in the biological terminology, in which it refers to the moderate temperature range, but in its original (greek) meaning of 'liking the inbetween'.

into a number of structural states present, *i.e.* reactant, product and/or intermediate(s), which requires the individual states to reach a population of approximately 25% or more in the crystal. Information on transitions between these states can, in principle, be obtained from the diffuse scattering as it directly measures correlated atomic displacements in those parts of the protein that undergo changes during the reaction.

The length scales over which displacements are correlated in the crystal determine the shape and the spread of the features present in the diffuse scattering pattern. The dynamical processes giving rise to these displacements can be categorised as phonons, rigid-body molecular displacements and intra-molecular motions. Acoustic scattering (also, less accurately, referred to as thermal diffuse scattering) is due to the acoustic modes of crystal vibrations (phonons[8]) and gives rise to diffuse scattering in the vicinity of and centred at the reciprocal lattice points [85]. Although acoustic scattering peaks are less localised than Bragg peaks their integrated intensity is significant and, therefore, Bragg data should be corrected for acoustic scattering prior to refinement. Neglecting the acoustic scattering correction has been shown to lead to significant underestimation of B factors [93, 94]. The form of the phonon-scattering factors can, in principle, be extracted from acoustic scattering [77] and provides information complementary to the phonon dispersion curves obtained from coherent inelastic neutron scattering [95]. However, information of lattice phonons in protein crystals remains scant.

Rigid-body molecular displacements may be correlated within and between unit cells and, in the diffuse scattering pattern, give rise to streaks located on families of reciprocal-lattice planes. Displacement correlations across unit cells have been observed for orthorhombic crystals of lysozyme [75], yeast initiator tRNA [96] and calmodulin [97]. From simulations of rigid-body molecular motions of crystalline fibres of tropomyosin transversal fluctuations of the fibres were determined that did not, however, propagate through adjacent unit cells [74, 78]. Also, for a tetragonal crystal of lysozyme it was demonstrated that independent rigid-body displacements of individual molecules can yield a qualitative description of the diffuse scattering [98].

Intra-molecular protein dynamics is of potential functional importance and has, therefore, been the focus of intense research. Correlated intra-molecular displacements give rise to strongly delocalised features in the diffuse scattering pattern that are not associated with the reciprocal lattice. This 'very diffuse' scattering, also denoted 'variational' scattering [76], is the subject of the present thesis. The variety of intra-molecular motions present in a protein renders an analytical description of the diffuse scattering cumbersome. However, molecular dynamics simulation provides a powerful means of describing the intra-molecular motions leading to diffuse scattering. MD simulations of orthorhombic crystals of lysozyme have been shown to reproduce features of the scattering pattern [27, 80]. Furthermore, it was demonstrated that the motions of groups of rigid bodies, on average composed of five residues, also reproduce the prominent features of the scattering pattern [27]. However, it has also been demonstrated that calculated diffuse scattering does not converge for nanosecond MD simulations [99, 100].

A full analytical description of the diffuse scattering due to intra-molecular motions can be given using the atomic displacement variance-covariance matrix. Unfortunately, this is non-trivial due to the potentially high number of parameters (cross-correlations in displacements), which is of

[8]The diffuse intensity due to one-phonon scattering is proportional to ω^{-2}, where the phonon frequency ω is given by the dispersion relation. Thus, the significant contribution to the diffuse intensity originates from the acoustic modes with small magnitude, q of the scattering vector, whereas the high-frequency optical and large-q acoustic modes give only minor contributions.

the order N^2 where N is the number of atoms in the unit cell. Therefore, the variance-covariance matrix must be approximated using models with significantly smaller numbers of parameters. The difficulty of unambiguously doing so becomes clear when noticing that, depending on the model used for analysis, in earlier research the protein motion determining diffuse scattering was contrastingly found to be either liquid-like (diffusive) or vibrational [76, 80].

A limitation of the early investigations was the rather incomplete experimental sampling in reciprocal space of the X-ray diffuse scattering. Recently, however, the most complete three-dimensional protein crystal X-ray diffuse scattering map yet obtained, using crystals of *Staphylococcal* nuclease, has been reported [101]. In this thesis, MD results are presented and used to directly calculate and interpret the *Staphylococcal* nuclease X-ray data.

1.3 Thesis Outline

The methodological background of this thesis is provided in Chapter 2 which is subdivided into three sections. In Section 2.1 the molecular dynamics simulation technique is introduced. In doing so, the focus is not on a general outline of the technique, as ample textbooks exists on the subject, but rather on the coherent description of the particular methods used in this thesis. In particular, the proper treatment of electrostatic interactions in the crystalline environment using the Ewald method and the coupling to a thermostat-barostat are discussed. Next, in Section 2.2, the setup of the simulation system and all simulations performed are described in detail (complemented by Appendix A). Subsequently, in Section 2.3, the mathematical description of X-ray Bragg and diffuse scattering from molecular crystals is presented. In particular, a model of dynamic disorder is introduced and an equation is derived to directly calculate X-ray diffuse scattering from molecular dynamics trajectories. Furthermore, a brief description of the experimentally obtained X-ray diffuse scattering map for *Staphylococcal* nuclease is given.

A central theme in statistical physics is the ergodic theorem which states that, if a system is ergodic, the expectation value of any phase-space function can be computed either by time (molecular dynamics approach) or ensemble averaging (monte-carlo approach). However, the investigated system may not be ergodic in practice, such as glasses or metastable phases, or even in principle, such as nearly-harmonic solids [102]. Nevertheless, ergodicity is usually assumed when performing computer simulations and the following practical question arises:

> *Did the simulation sufficiently sample the relevant regions of phase space, i.e., did the quantities of interest converge during the simulation?*

For molecular dynamics computer simulations of proteins, the answer is non-trivial as certain quantities converge on the timescales accessible to present-day computational resources whereas others do not. In Chapter 3, therefore, the convergence properties of selected observables relevant for later analyses are investigated. In particular, the atomic fluctuations, which are related to X-ray crystallographic B factors, as well as cross-correlations, which are related to the variance-covariance matrix which determines X-ray diffuse scattering, are examined. Furthermore, the required simulation length for the variance-covariance matrix to converge is estimated. Moreover, the relation between the converged elements of the correlation matrix and the protein topology/structure is determined. The main findings of these analyses also establish a more sound basis for the interpretation of the results of the subsequent chapters.

Chapter 4 addresses several issues regarding the *Staphylococcal* nuclease X-ray diffuse scattering map which can be combined to the following, central question:

What is the origin of the X-ray diffuse scattering intensity from a protein crystal?

Section 4.1 approaches this question from a structural point of view and decomposes the unit-cell scattering into protein and crystal solvent contributions. Subsequently, the protein scattering is further dissected and the influence of the secondary structural elements is determined. A dynamical perspective is then chosen in Section 4.2. In particular, two competing models of protein motion, liquid-like versus vibrational, are investigated quantitatively and their diffuse scattering is compared with that derived directly from the simulations. Furthermore, individual three-dimensional features of the diffuse scattering map are assigned to specific collective motions in the protein, and some of these explicitly involve potentially functional active-site deformations.

In Chapter 5, use is made of the comprehension established in the preceding chapters, and the diffuse scattering is utilised as a tool, amongst others, to tackle the following problem:

How does the application of pressure change the internal protein dynamics?

To answer this question, the effect of pressure on the protein internal motions is studied in the range 1 bar to 15 kbar. After briefly discussing structural relaxations, the dependence of the mean-square displacement on the applied pressure is investigated. Furthermore, the effect of pressure on the protein and solvent diffuse scattering and the protein vibrational density of states are examined.

In each of the Chapters 3 to 5, a certain aspect of crystalline protein dynamics is presented in an approximate self-contained manner. In particular, the last section to each of those chapters discusses the main findings and draws important conclusions. In Chapter 6, the synopsis of these results is given and an outlook is presented.

Chapter 2

Modelling Crystalline Staphylococcal Nuclease

Protein crystallography occupies a prominent position in the field of protein physics, since it is the major source of experimental information on three-dimensional protein structures. However, a complete description of the multi-faceted physics underlying the dynamics of crystalline proteins has not yet been achieved and is the subject of on-going research. The present thesis contributes to this endeavour and this chapter provides the framework on which the studies are based. Due to the multitude of excellent literature and text books, the description is limited to the extent needed for an understanding of the thesis.

The bacterium *Staphylococcus aureus* secretes a thermostable nuclease known as *Staphylococcal* nuclease (SNase, also known as thermonuclease). SNase is a calcium-dependent enzyme that catalyses the hydrolysis of both DNA and RNA. Structurally, SNase is a small globular protein comprising 149 amino-acid residues. It is a common, widely-used model system for the study of protein conformation, dynamics and function. The structure, energetic properties and function of SNase have been reviewed in Refs. [103–108].

In this thesis, the dynamics of a complete unit cell of crystalline SNase, *i.e.* including proteins and crystal solvent, is described using molecular dynamics simulations. Molecular dynamics is a classical, non-quantum technique and is outlined in Section 2.1. The focus here is on the description of inter-atomic forces using the Charmm potential energy function, the representation of the crystalline environment using periodic boundary conditions and the Ewald method for electrostatic interactions, and on the proper implementation of the experimental constant-temperature constant-pressure conditions. General introductions to, and in-depth descriptions of molecular dynamics and related techniques can be found in Refs. [102, 109–112]. Subsequently, Section 2.2 provides a detailed description of the setup and performance of all simulations carried out and analysed in this thesis. In Section 2.3, the principles of X-ray scattering from a crystalline sample are outlined. A model of disorder is introduced, leading to deviations from perfect crystal periodicity and thus a loss in Bragg intensity, and an expression for the resulting X-ray diffuse scattering is derived. For crystals of SNase, three-dimensional X-ray diffuse scattering has been reported [101] which represents the most complete map yet collected for a protein crystal. Details on this experiment, which is in part interpreted in this thesis, are briefly described.

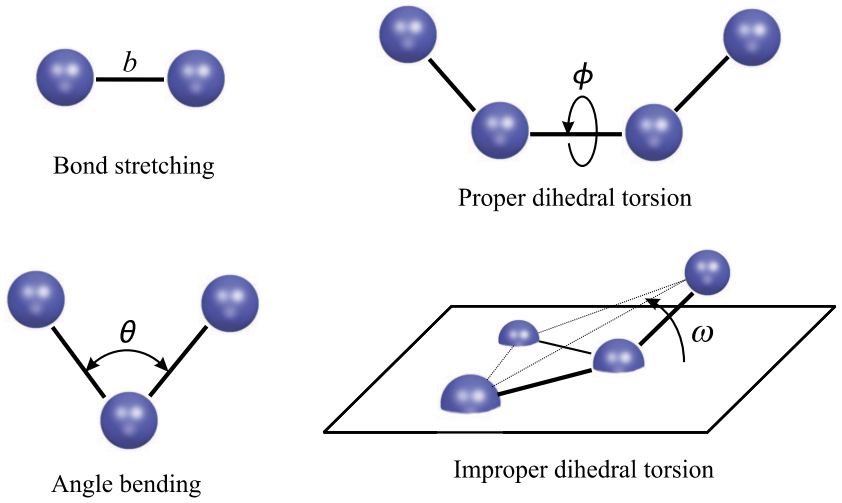

Figure 2.1: Schematic illustration of the bonded terms in the CHARMM force field, cf. Eq. (2.2). (Adapted from AL Tournier: *Dissertation, Heidelberg University, 2003.*)

2.1 Molecular Dynamics

A complete description of the dynamics of any physical system is, in principle, given by the time-dependent Schrödinger equation. However, for many-particle systems this equation cannot be solved analytically and, therefore, approximate methods have to be used which, in practice, involve computer simulations. The choice of which approximation to use depends, on the one hand, on the accuracy required for the particular system under investigation and, on the other hand, on available computational resources. If the electronic dynamics is relevant, *e.g.* for the study of chemical reactions or photon absorption or emission, the choice is restricted to the repertory of quantum mechanical methods. Otherwise, if the focus of interest is for example directed towards protein:ligand binding, conformational changes or folding, methods based on classical mechanics, which allow investigations on significantly larger systems and longer timescales, may be applied.

In this thesis, the computer simulations are performed using *molecular dynamics* which is a non-quantum method combining the classical equations of motion with an empirical model potential, or *force field*. Molecular dynamics is based on the Born-Oppenheimer approximation: due to a large atomic nuclei:electron mass ratio, $\mathcal{O}(10^3 \dots 10^4)$, the electronic dynamics is at least three orders of magnitude faster than the nuclei dynamics and, therefore, can be assumed to instantaneously adopt to the positions of the nuclei. Consequently, in molecular dynamics only the motion of the atomic nuclei is described.

The following section introduces the CHARMM force field, which was used throughout this thesis, and the subsequent sections then outline the treatment of electrostatic interactions in the crystalline environment and the time evolution of a system coupled to an external thermostat-barostat.

2.1.1 CHARMM Force Field

Originally, the CHARMM force field was published 1983 along with a program suite of the same name [113], but since then contributions from different authors resulted in a rather independent development of both. Throughout this thesis, the program version `28b1` was used along with the force field `parameter set 22` [114].

In the CHARMM force field, atoms are represented as charged point masses and the potential energy function, E is separated into internal, or bonded, and pairwise non-bonded terms,

$$E = \underbrace{E_{\text{bonds}} + E_{\text{angles}} + E_{\text{dihedrals}} + E_{\text{impr. dihed.}}}_{\text{bonded terms}} + \underbrace{E_{\text{el}} + E_{\text{vdW}}}_{\text{non-bonded terms}} \tag{2.1}$$

$$\begin{aligned} &= \sum_{\text{bonds}} K_b(b - b_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} K_\phi[1 + \cos(n\phi - \phi_{\text{eq}})] \\ &\quad + \sum_{\text{impr. dihed.}} K_\omega(\omega - \omega_{\text{eq}})^2 + \underbrace{\sum_{i<j} \frac{q_i q_j}{\epsilon r_{ij}}}_{\text{electrostatic}} + \underbrace{\sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]}_{\text{van der Waals}}, \end{aligned} \tag{2.2}$$

where the individual bonded terms describe stretching of the bond length, b, bending of the bond angle, θ, torsion of the proper dihedral angle, ϕ, with multiplicity, n, and bending of the improper dihedral angle, ω, as depicted in Fig. 2.1. K_x and x_{eq} denote the force constants and equilibrium values, respectively. The last two sums in Eq. (2.2) describe the electrostatic and the van der Waals interaction, respectively. $q_{i,j}$ are the charges of, and r_{ij} the distance between atoms i and j, and ϵ is the dielectric susceptibility. The Lenard-Jones 12-6 function is chosen to describe the van der Waals interaction, the parameters being the depth of the potential, ϵ_{ij} and the collision parameter, σ_{ij}. The CHARMM force field parameters form a self-consistent set, being derived from *ab initio* calculations (non-bonded and bonded interactions) and experiments (bonded interactions), and are optimised for use with the TIP3P [115] water model.

The computation of the sums over the non-bonded interactions in Eq. (2.2) requires $\mathcal{O}(N^2)$ operations, N being the number of atoms, which is unfeasible for MD simulations of systems of more then $\mathcal{O}(10^2)$ atoms. Therefore, in practice non-bonded interactions are usually restricted to atom pairs within a certain distance of each other using a *cut-off* method. CHARMM supports a *shift* and *switch* mechanism as cut-off methods, the former shifting the potential (or force) such that its value becomes zero at a specified cut-off distance, r_{cutoff}, the latter smoothly interpolating the potential between a cut-on distance, r_{cuton} and r_{cutoff} such that potential and force become zero at r_{cutoff}. Whereas the short-range van der Waals interactions can be adequately described using any of these cut-off methods (provided that large enough values for r_{cuton} and r_{cutoff} are used), the convergence of the sum over the long-range electrostatic interactions may be poor even for large values of r_{cutoff}.

If the simulation system can be considered as being part of an infinite, periodic system, such as a crystal unit cell, the electrostatic interactions between all atoms in the system may be calculated using the Ewald method, which is outlined in the following section.

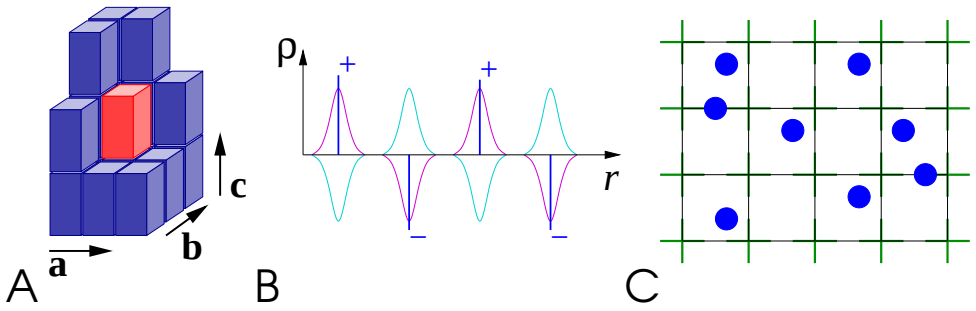

Figure 2.2: A: An ideal crystal is obtained by translations {**a**,**b**,**c**} of a unit cell (red). B: In the Ewald summation, each point charge (blue) is surrounded by a neutralising gaussian charge distribution (cyan), leading to rapidly convergent series in real space; the corresponding cancelling distribution (magenta) is calculated in reciprocal space. C: The irregularly distributed charges (blue) are interpolated onto the vertices (green) of a regular grid.

2.1.2 Ewald Summation

The convergence of the sum over electrostatic interactions, E_{el} in Eqs. (2.1) and (2.2) may be very slow or, for an infinite system such as a crystal, even divergent, depending on the specific charges q_i[1]. For a periodic system, however, convergence can be improved by recasting E_{el} into the sum of two rapidly convergent series, one being evaluated in real space, the other in reciprocal space, as originally proposed by Ewald [116].

In a crystal, illustrated in Fig. 2.2 A, electrostatic interactions between atoms belonging to different unit cells have to be considered,

$$E_{\mathrm{el}} = \sum_{\mathbf{R}}{}' \sum_{i<j} \frac{q_i q_j}{\epsilon \|\mathbf{r}_{ij} + \mathbf{R}\|}, \tag{2.3}$$

where the lattice vector, $\mathbf{R}$ is defined as

$$\mathbf{R} = l\mathbf{a} + m\mathbf{b} + n\mathbf{c}, \tag{2.4}$$

where {**a**,**b**,**c**} is the basis of the crystal lattice, $\{l, m, n\}$ are integers, and the prime in Eq. (2.3) denotes that $\mathbf{R} = \mathbf{0}$ is excluded. Using the identity

$$\frac{1}{x} \equiv \frac{f(x)}{x} + \frac{1 - f(x)}{x}, \tag{2.5}$$

Eq. (2.3) is recast into a sum of two series,

$$E_{\mathrm{el}} = E_{\mathrm{el}}^{\mathrm{R}} + E_{\mathrm{el}}^{\mathrm{Q}}, \tag{2.6}$$

$$= \sum_{\mathbf{R}}{}' \sum_{i<j} \frac{q_i q_j}{\epsilon} \left(\frac{\mathrm{erfc}(\beta \|\mathbf{r}_{ij} + \mathbf{R}\|)}{\|\mathbf{r}_{ij} + \mathbf{R}\|} + \frac{\mathrm{erf}(\beta \|\mathbf{r}_{ij} + \mathbf{R}\|)}{\|\mathbf{r}_{ij} + \mathbf{R}\|} \right), \tag{2.7}$$

where the complementary error function, $\mathrm{erfc}(\cdot) = 1 - \mathrm{erf}(\cdot)$, has been chosen for $f(\cdot)$. The physical picture underlying this choice is illustrated in Fig. 2.2 B: each point charge, q_i is surrounded

[1] As illustrative, one-dimensional, discrete examples consider the divergent harmonic series, $\sum_n {}^1/_n$ and the converging alternating harmonic series, $\sum_n {}^{(-1)^n}/_n = \ln 2$.

by a neutralising gaussian charge distribution of width $(\sqrt{2}\beta)^{-1}$,

$$\rho_i(\mathbf{r}) = -\frac{q_i\beta^3}{\pi^{3/2}} \exp\left(-\beta^2\|\mathbf{r}-\mathbf{r}_i\|^2\right), \tag{2.8}$$

the potential of which is just given by the erf(·) function[2], leading to a short-range, rapidly converging series, E_{el}^{R} in Eq. (2.6). Re-balancing the total charge distribution, using $-\rho_i$, leads to the long-range term, E_{el}^{Q} which is, therefore, computed in reciprocal space using discrete fourier transforms,

$$E_{\text{el}}^{\text{Q}} = \frac{1}{2\pi\epsilon V} \sum_{\mathbf{Q}}{}' \frac{\exp(-\pi^2\|\mathbf{Q}\|^2\beta^{-2})}{\|\mathbf{Q}\|^2} S_i(\mathbf{Q})S_j(\mathbf{Q}), \tag{2.9}$$

with the structure factor,

$$S_j(\mathbf{Q}) = \sum_j q_j \exp(i\mathbf{Q}\cdot\mathbf{r}_j), \tag{2.10}$$

the unit-cell volume,

$$V = \mathbf{a}\cdot(\mathbf{b}\times\mathbf{c}), \tag{2.11}$$

and the reciprocal lattice vector,

$$\mathbf{Q} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*, \tag{2.12}$$

where $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$ is the basis of the reciprocal lattice[3] and $\{h, k, l\}$ are integers not all zero. Convergence of the series E_{el}^{R} and E_{el}^{Q} is determined by the value of β: small or large β improves convergence of E_{el}^{R} or E_{el}^{Q}, respectively, and using an optimised value for β results in an $\mathcal{O}(N^{3/2})$ algorithm.

Further algorithmic improvement can be achieved by using fast fourier transform routines which require interpolation of the charges q_i, located at arbitrary positions $\mathbf{r}_i$, onto a regular grid. This *particle meshing* is illustrated in Fig. 2.2 C. The interpolated charges are then used in the computation resulting in an $\mathcal{O}(N\log N)$ algorithm.

E_{el} in the form of Eq. (2.6) has to be corrected for artefactual interactions of each charge q_i with its associated gaussian charge cloud ρ_i, interactions of those atom pairs which are described using bonded-terms of the force field, and a finite dielectric susceptibility of the outside medium, leading to the corrections

$$E_{\text{el}}^{\text{corr}} = -\frac{\beta}{\sqrt{\pi}}\sum_i q_i^2 - \sum_{i<j}^{\text{bonded}} \frac{q_iq_j\text{erf}(\beta\|\mathbf{r}_{ij}\|)}{\|\mathbf{r}_{ij}\|} + \frac{2\pi}{3V}\left\|\sum_i q_i\mathbf{r}_i\right\|^2. \tag{2.13}$$

Throughout this thesis, periodic boundary conditions were applied to generate the crystal environment. Electrostatic interactions were computed using the particle-mesh Ewald method [117], for which the direct sum cut-off was 13 Å and the reciprocal space structure factors were computed on a $48\times48\times64$ grid using 5^{th} degree cardinal B-splines.

[2]The solution to the Poisson equation, $\Delta\phi = -4\pi\rho_i$ is given by $\phi(\mathbf{r}) = q_i\frac{\text{erf}(\beta\|\mathbf{r}-\mathbf{r}_i\|)}{\|\mathbf{r}-\mathbf{r}_i\|}$.

[3]The reciprocal lattice is defined by the relations $\mathbf{x}_i\cdot\mathbf{x}_j^* = 2\pi\delta_{ij}$, where $\mathbf{x}_i$ and $\mathbf{x}_j^*$ are basis vectors of the direct and reciprocal lattice, respectively.

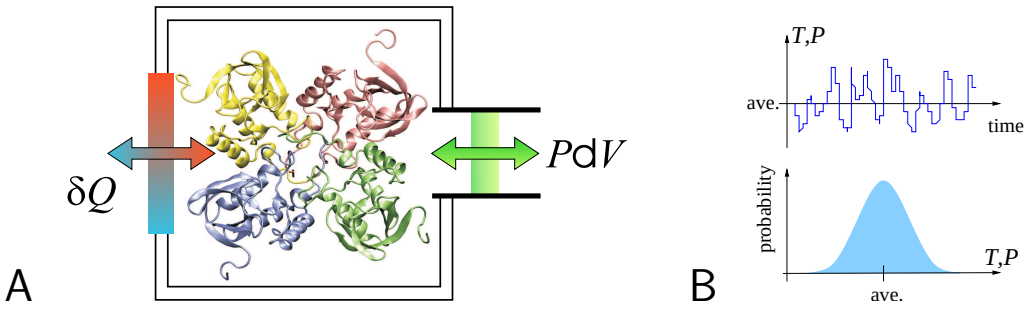

Figure 2.3: Allowing the simulation system to exchange heat, δQ and mechanical work, $P\mathrm{d}V$ with its environment (A) leads to trajectories in the isothermal-isobaric ensemble, *i.e.* with conserved average temperature and pressure (B).

2.1.3 Time Evolution

In molecular dynamics, the equations of motion governing the dynamics of biological macromolecules are formulated within the realm of classical mechanics [118, 119]. In the Hamilton formalism, the equations of motion of the generalised positions, q_l and momenta, p_l are given by

$$\dot{q}_l = \frac{\partial}{\partial p_l} H(q_l, p_l), \tag{2.14}$$

$$\dot{p}_l = -\frac{\partial}{\partial q_l} H(q_l, p_l), \tag{2.15}$$

where the Hamiltonian,

$$H(q_l, p_l) = K(q_l, p_l) + V(q_l, p_l) \tag{2.16}$$

is given by the sum of the kinetic and potential energy, $K(q_l, p_l)$ and $V(q_l, p_l)$, respectively.

Describing the potential energy exclusively by inter-particle interactions, *e.g.* due to the force field E given by Eq. (2.2), integration of Eqs. (2.14) and (2.15) leads to trajectories in the micro-canonical ensemble, *i.e.* with the total energy, volume and number of particles being conserved. However, most physical experiments are performed in an isothermal-isobaric environment, *i.e.* with the temperature, T and pressure, P being constant. Trajectories of the isothermal-isobaric ensemble can be obtained by *extending* the system to contain thermostat and barostat variables, which act as heat and pressure bath coupled to the system, cf. Fig. 2.3.

Extended System Molecular Dynamics

Throughout this thesis, temperature and pressure coupling were enforced with the Nosé-Hoover algorithm [120–122]. However, in its original formulation, this algorithm involves rescaling of the simulation timestep which is inconvenient for molecular dynamics and, therefore, modern implementations are based on non-Hamiltonian schemes. For N particles in d dimensions, a set

of dynamical equations producing the *NPT* ensemble is defined by [109, 123]

$$\dot{\mathbf{r}}_i = \frac{\mathbf{r}_i}{m_i} + \frac{p_\epsilon}{W}\mathbf{r}_i, \tag{2.17}$$

$$\dot{\mathbf{p}}_i = \mathbf{F}_i - \left(1 + \frac{1}{N}\right)\frac{p_\epsilon}{W}\mathbf{p}_i - \frac{p_\eta}{Q}\mathbf{p}_i, \tag{2.18}$$

$$\dot{V} = \frac{dVp_\epsilon}{W}, \tag{2.19}$$

$$\dot{p}_\epsilon = dV(P_{\text{int}} - P_{\text{ext}}) + \frac{1}{N}\sum_i \frac{\mathbf{p}_i^2}{m_i} - \frac{p_\eta}{Q}p_\epsilon, \tag{2.20}$$

$$\dot{\eta} = \frac{p_\eta}{Q} \tag{2.21}$$

and

$$\dot{p}_\eta = \sum_i \frac{\mathbf{p}_i^2}{m_i} + \frac{p_\epsilon^2}{W} - (dN+1)k_{\text{B}}T, \tag{2.22}$$

where $\mathbf{r}_i$ and $\mathbf{p}_i$ are the position and momentum of particle i, respectively, $\mathbf{F}_i = -\nabla_{\mathbf{r}_i}E$ is the internal force acting on particle i, η is the thermostat variable, p_η its conjugate momentum and Q is a parameter determining the coupling timescale[4], $\epsilon = \ln(V/V_0)$ is the barostat variable, p_ϵ its conjugate momentum and W is the associated mass parameter[4], P_{ext} is the external applied pressure, and P_{int} is the instantaneous internal pressure of the system given by

$$P_{\text{int}} = \frac{1}{dV}\left(\sum_i \frac{\mathbf{p}_i^2}{m_i} + \sum_i \mathbf{r}_i \cdot \mathbf{F}_i - dV\frac{\partial U}{\partial V}\right). \tag{2.23}$$

Thus, the variables η and ϵ act as thermostat and barostat, respectively, driving the system towards the steady state, $\langle T_{\text{int}}\rangle = T$ and $\langle P_{\text{int}}\rangle = P_{\text{ext}}$. If $\sum_i \mathbf{F}_i \neq 0$ the only conserved quantity is given by the Hamilton function

$$H' = H(\mathbf{r},\mathbf{p}) + \frac{p_\epsilon^2}{2W} + \frac{p_\eta^2}{2Q} + (dN+1)k_{\text{B}}T + P_{\text{ext}}V, \tag{2.24}$$

where $H(\mathbf{r},\mathbf{p})\widehat{=}H(q_l,p_l)$.

Leap-Frog Integrator

A numerical integration scheme for the above defined equations of motion should generate a trajectory being statistically equivalent to the real trajectory, *i.e.*, assuming ergodicity for the true trajectory,

$$\lim_{\mathcal{T}\to\infty} \overline{O(t)} = \langle O\rangle, \tag{2.25}$$

where O is an observable, $\mathcal{T}$ the simulation length, $\overline{\cdot}$ the time and $\langle\cdot\rangle$ the ensemble average, the ensemble average in Eq. (2.25) can also be computed from a single numerical trajectory, $O(t_n)$.

[4]Throughout this thesis, the temperature coupling constant 2,000 kcal ps^{-2} and the pressure piston mass 500 u were used in the CHARMM program.

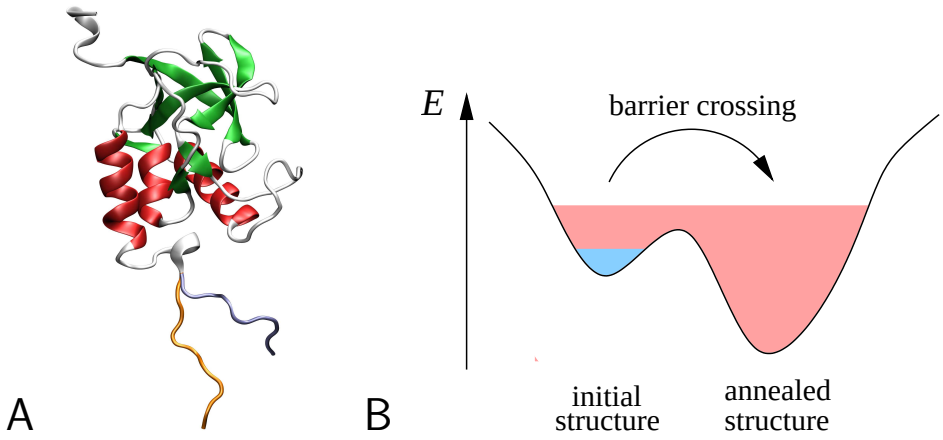

Figure 2.4: A: Initial structure of SNase (cartoon representation with α-helices (red) and β-strands (green), for a detailed description of secondary structural elements see Fig. 3.9 B on pg. 37) with the C terminal residues 142–149 before (orange) and after (iceblue) a simulated annealing MD simulation: the system was heated (20 ps) to and equilibrated (10 ps) at 800 K, then evolved for 200 ps and subsequently cooled during 100 ps to 0 K using the micro-canonical ensemble. B: Simulated annealing is a two-step scheme to sample, and locate the minimum of, a large fraction of the protein energy landscape. First, an MD simulation is performed at high temperature, thus significantly extending the accessible area of the energy landscape (melon) relative to that accessible at lower temperature (navyblue). Then, the temperature is slowly decreased to 0 K, thus freezing the system into the lowest-energy configuration explored.

The constant temperature-pressure module of the CHARMM program uses the leap-frog integrator [124], which is time-reversible and symplectic, *i.e.* it conserves phase-space density and thereby guarantees conservation of H' to within an error, $\triangle H'$ determined by the timestep $\triangle t$. The choice of $\triangle t$ is a trade-off between accuracy ($\triangle H'$ decreases with decreasing $\triangle t$) and computational effort for reaching a certain simulation length, $\mathcal{T}$ – throughout this thesis, $\triangle t = 1$ fs.

The leap-frog algorithm is derived using a Taylor series expansion for the particle position, $\mathbf{r}_i$ at timestep t_n and its velocity, $\mathbf{v}_i$ at an intermediate timestep $t_n + {}^{\triangle t}/_2$, leading to

$$\mathbf{r}_i(t_n + \triangle t) = \mathbf{r}_i(t_n) + \mathbf{v}_i(t_n + {}^{\triangle t}/_2)\ \triangle t \tag{2.26}$$

and

$$\mathbf{v}_i(t_n + {}^{\triangle t}/_2) = \mathbf{v}_i(t_n - {}^{\triangle t}/_2) + \frac{1}{m_i}\mathbf{F}_i\ \triangle t. \tag{2.27}$$

If required, the velocities at timestep t_n are readily computed using

$$\mathbf{v}_i(t_n) = \frac{1}{2}\left[\mathbf{v}_i(t_n - {}^{\triangle t}/_2) + \mathbf{v}_i(t_n + {}^{\triangle t}/_2)\right]. \tag{2.28}$$

2.2 Computer Simulations

Molecular dynamics simulations were performed on the crystal unit cell of *Staphylococcal* nuclease with four protein molecules and explicit solvent with the CHARMM program, cf. Section 2.1.

Their setup and performance will be briefly outlined in this section, whereas the specific analysis tools used will be discussed in the following chapters along with the results. All simulations used the same initial configuration of the unit cell, the setup of which is described in the next section.

2.2.1 System Setup

From the Protein Data Bank (PDB) [69], the entry 2SNS [125], refined at 1.5 Å resolution, was taken as the protein starting structure. Coordinates for hydrogen atoms, which, in general, can only be derived from very-high resolution (<1 Å) X-ray crystallographic data[5] were generated with the `hbuild` algorithm implemented in Charmm. The crystallographically unresolved residues 142-149 were added using Charmm. Initial coordinates for these residues were obtained from an 800 K simulated annealing MD simulation, described in Fig. 2.4, of the protein with residues 1-139 fixed in a TIP3P [115] water box.

The resulting protein structure was used to construct the crystal unit cell, having the initial dimensions $a = b = 48.5$ Å and $c = 63.4$ Å [101], and a symmetry according to the experimental space group $P4_1$ [101, 125], defined by the point operations

$$(x, y, z), \qquad (-y, x, z + 1/4), \qquad (-x, -y, z + 1/2) \qquad \text{and} \qquad (y, -x, z + 3/4), \tag{2.29}$$

describing a tetragonal unit cell containing four protein molecules arranged along a 4-fold screw axis, chosen to be identical with the z-axis. 2115 TIP3P water molecules[6], including 4×83 molecules observed crystallographically in 1STN [126], and 48 chloride counterions were added as solvent, leading to an electrically neutral system of 15,993 atoms, depicted in Fig. 2.5. Periodic boundary conditions were applied to generate the crystal environment and the electrostatic interactions were computed using the particle-mesh Ewald method, cf. Section 2.1.2.

The system was initially energy minimised to a root-mean-square (RMS) force gradient of 10^{-3} kcal mol^{-1}Å^{-1} and then uniformly heated to 300 K during 30 ps and subsequently equilibrated for 100 ps with velocity scaling in the *NVE* ensemble and another 200 ps without velocity scaling in the *NPT* ensemble with $P = 1$ bar and $T = 300$ K. The temperature and pressure coupling were enforced with the Nosé-Hoover algorithm, cf. Section 2.1.3. Subsequently the *NPT* production runs were performed, which are described in the next section.

[5] In crystallography, the resolution usually refers to the length-scale corresponding to the largest-angle Bragg peak observed. In X-ray crystallography the resolution is limited by the X-ray wavelength used and due to the protein static and dynamic disorder as well as solvent scattering. High-resolution experiments are facilitated at high-intensity synchrotron sources with very short wavelengths and at cryogenic temperatures, *i.e.* around or below 100 K. On 24$^{\text{th}}$ June 2005 the PDB held 78 X-ray structures with a resolution of 1 Å or better. In principle, hydrogens can easily be observed in neutron diffraction experiments because their coherent scattering cross-section is comparable to that of other atoms abundant in biomolecules. However, the enormous costs and technical limitations prohibit neutron diffraction to become a widely used method, as is evident from the small number of entries in the PDB, *i.e.* six on 24$^{\text{th}}$ June 2005. However, neutron diffraction is complementary to X-ray studies and the additional information on hydrogens is of significant interest.

[6] The number of water molecules was varied during four preliminary simulation setups to obtain the correct unit-cell density and thus stability for the unit-cell edge parameters, *i.e.* prevent shrinking or expansion of the unit-cell volume.

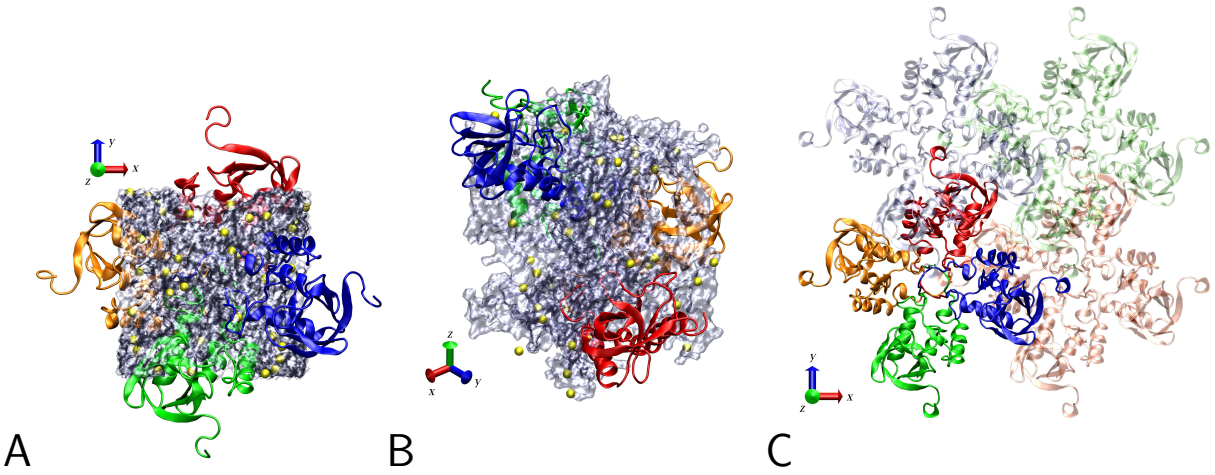

Figure 2.5: The complete simulation system of crystalline *Staphylococcal* nuclease with $P4_1$ symmetry for the protein molecules, cf. Eq. (2.29). A: Single unit cell, including the four proteins (green, blue, red and orange), water molecules (shown as transparent iceblue volume surfaces) and counterions (yellow) viewed along $\mathbf{c} \uparrow\uparrow z$-axis. The rotational symmetry of the proteins around the 4-fold screw axis ($\mathbf{c}$) can be seen. Water molecules and counterions are restricted to volumes inside the primitive unit cell, whereas protein molecules are not truncated at the unit-cell boundary in order to preserve correct topology; periodic boundary conditions assure correct filling of the unit cell (B,C). B: Single unit cell viewed along the (1,1,½)-axis. The translational displacement of the proteins along $\mathbf{c}$ can be seen. Void volumes are occupied by protein molecules of neighbouring unit cells (C). C: Single unit cell (opaque) surrounded by three neighbouring unit cells (transparent). The scales of A and B are approximately the same, being different from that of C.

2.2.2 Production Phases

The two thematically different aspects of crystalline protein dynamics investigated in this thesis were approached using two sets of simulations, differing mainly in the simulation length, $\mathcal{T}$. One aspect, the investigations of fluctuations, correlations and X-ray diffuse scattering, is based on four rather long, $\mathcal{T} = 10\,\text{ns}$, simulations, denoted Set 1, whereas the second aspect, the study of pressure-induced changes in the dynamics, is based on a different set of 22 shorter, $\mathcal{T} = 1\,\text{ns}$, simulations, denoted Set 2. General features of all simulations are discussed in the following paragraphs and the detailed CHARMM inputs defining the energy functions and dynamics parameters are given in Appendix A.

MD Simulations: Set 1

Four 10 ns MD simulations were performed, starting from the equilibrated structure described in the previous section, differing only in the treatment of the crystalline environment. In three simulations (named T1, T2 and T3) a tetragonal constraint was imposed, *i.e.*, the unit-cell dimensions a and b scale identically and independently of c. This corresponds to the $P4_1$ experimental space group symmetry. In the fourth simulation (denoted O1) an orthogonal constraint was used, *i.e.*, a, b and c scale independently, thus allowing for deviations from $P4_1$. Additionally, a 1 ns simulation (named WB) of 4,989 TIP3P water molecules in a tetragonal box (initial dimensions $48.5 \times 48.5 \times 63.4\,\text{Å}^3$) was performed in the *NPT* ensemble. Coordinates were written every 50 fs.

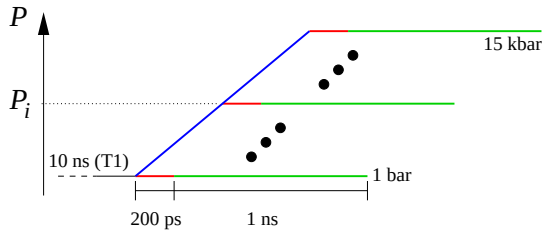

Figure 2.6: Scheme used to generate the high-pressure trajectories forming the MD simulation Set 2.

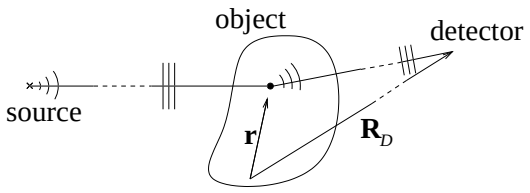

Figure 2.7: Schematic illustration of the X-ray scattering setup in the plane-wave approximation.

MD Simulations: Set 2

Twenty-two 1 ns MD simulations in the pressure range 1 bar to 15 kbar were performed. The final configuration of trajectory T1 served as starting structure, thus granting a very well (10.2 ns) equilibrated configuration of the proteins in the unit cell. This system was then pressurised at a rate of $1\,\mathrm{kbar\,ns^{-1}}$ up to 15 kbar and, at selected pressure values P_i, branched to start a high-pressure trajectory, each of which was equilibrated for 200 ps and then evolved for the 1 ns long production phase, cf. Fig. 2.6. Coordinates were written every 100 fs.

2.3 X-ray Scattering from Molecular Crystals

In this section, those aspects of X-ray crystallography that are used in this thesis are outlined. In particular, the equations for calculating the X-ray diffuse scattering intensity from MD trajectories are derived. The description is limited to the concepts used in the thesis and the notation roughly follows Refs. [127, 128]. More detailed descriptions can be found in Refs. [127–130] and in Ref. [71], the latter being a state-of-the-art description of applied research.

2.3.1 Bragg Scattering

Let

$$A_i(\mathbf{r},t) = A_0 e^{i(\mathbf{k}_i\cdot\mathbf{r}-\omega_0 t)} \tag{2.30}$$

be the amplitude of an incident plane wave, with $\mathbf{k}_i$ being the wave vector, ω_0 the frequency and A_0 the complex amplitude, being scattered at an object with, in general, the complex scattering density $\varrho(\mathbf{r})$, cf. Fig. 2.7. At a detector, located at $\mathbf{R}_D$, the scattered amplitude due to the object point $\mathbf{r}$ is then given by

$$A_D(\mathbf{r},t) = A_i(\mathbf{r},t)\varrho(\mathbf{r})\frac{e^{i\mathbf{k}_s\cdot(\mathbf{R}_D-\mathbf{r})}}{\|\mathbf{R}_D-\mathbf{r}\|}, \tag{2.31}$$

where $\mathbf{k}_s$ is the wave vector of the scattered wave. Then, for sufficiently large distances from the object, $\|\mathbf{R}_D\| \gg \|\mathbf{r}\|$,

$$A_D(\mathbf{r},t) = \frac{A_0}{R_D} e^{i(\mathbf{k}_s\cdot\mathbf{R}_D-\omega_0 t)}\varrho(\mathbf{r})e^{i(\mathbf{k}_i-\mathbf{k}_s)\cdot\mathbf{r}} \tag{2.32}$$

and the total amplitude scattered from the object is given by

$$A_D(t) = \frac{A_0}{R_D} e^{i(\mathbf{k}_s \cdot \mathbf{R}_D - \omega_0 t)} \int \varrho(\mathbf{r}) e^{i(\mathbf{k}_i - \mathbf{k}_s)\cdot \mathbf{r}} \, \mathrm{d}V, \tag{2.33}$$

where the integral is evaluated in the region occupied by the object. For X-rays, not the amplitude but intensity of the scattered radiation is detected,

$$I(\mathbf{q}) \propto |A_D|^2 = \frac{|A_0|^2}{R_D^2} \left| \int \varrho(\mathbf{r}) e^{-i\mathbf{q}\cdot\mathbf{r}} \, \mathrm{d}V \right|^2, \tag{2.34}$$

where the scattering vector $\mathbf{q} = \mathbf{k}_s - \mathbf{k}_i$ was introduced. If the object is a crystal, *i.e.* being periodic in real space with an associated reciprocal lattice, cf. Section 2.1.2, $\varrho(\mathbf{r})$ can be expressed using the Fourier representation,

$$\varrho(\mathbf{r}) = \sum_{\mathbf{Q}} \varrho_{\mathbf{Q}} e^{i\mathbf{Q}\cdot\mathbf{r}}, \tag{2.35}$$

and thus

$$I(\mathbf{q}) \propto \frac{|A_0|^2}{R_D^2} \left| \sum_{\mathbf{Q}} \varrho_{\mathbf{Q}} \int e^{i(\mathbf{Q}-\mathbf{q})\cdot\mathbf{r}} \, \mathrm{d}V \right|^2. \tag{2.36}$$

If the crystal is composed of a large number of unit cells, the integral yields significant values only if [7]

$$\mathbf{q} = \mathbf{Q}, \tag{2.37}$$

which is known as the von Laue condition, leading to peaks located at reciprocal lattice points, the *Bragg scattering*,

$$I(\mathbf{Q}) \propto \frac{V^2 |A_0|^2}{R_D^2} |\varrho_{\mathbf{Q}}|^2 . \tag{2.38}$$

It remains to calculate the Fourier components,

$$\varrho_{\mathbf{Q}} = \frac{1}{V_{\mathrm{uc}}} \int_{\mathrm{uc}} \varrho(\mathbf{r}) e^{-i\mathbf{Q}\cdot\mathbf{r}} \, \mathrm{d}V, \tag{2.39}$$

where the integral is evaluated over the unit-cell volume. In a solid, the majority of electrons are located within small regions around the atoms, allowing to separate the integral into a sum of contributions from individual atoms k,

$$\varrho_{\mathbf{Q}} = \frac{1}{V_{\mathrm{uc}}} \sum_k e^{-i\mathbf{Q}\cdot\mathbf{r}_k} \int_k \varrho_k(\mathbf{r}') e^{-i\mathbf{Q}\cdot\mathbf{r}'} \, \mathrm{d}V', \tag{2.40}$$

where the substitution $\mathbf{r} = \mathbf{R}_L + \mathbf{r}_k + \mathbf{r}'$ has been made, $\mathbf{R}_L$ points to the unit-cell origin, chosen to coincide with a primitive lattice vector, cf. Eq. (2.4), $\mathbf{r}_k$ is the position vector of the k^{th} atom with regard to the unit-cell origin, and $\mathbf{r}'$ originates at the centre of that atom; the integral describes the interference between waves scattered from different points of the atom and is

[7] For an infinite crystal, the value of the integral is $\int e^{i(\mathbf{Q}-\mathbf{q})\cdot\mathbf{r}} \, \mathrm{d}V = V\boldsymbol{\delta}(\mathbf{Q}-\mathbf{q})$, where $\boldsymbol{\delta}(\cdot)$ is the three-dimensional Delta function.

denoted atomic form factor, $f_k(Q)$ – in general, $f_k(q)$. Identifying $\varrho_k(\mathbf{r}')$ with the atomic charge distribution, for coherent scattering $f_k(\mathbf{q})$ may be calculated from the total wave function, ψ of the atom,

$$f_k(\mathbf{q}) = \int \psi^* e^{-i\mathbf{q}\cdot\mathbf{r}_k} \psi \, \mathrm{d}V. \tag{2.41}$$

In general, the charge distribution can be considered to be essentially spherically symmetric[8], thus $f_k(\mathbf{q}) \to f_k(q) \doteq f_k$. Values for f_k are tabulated in the International Tables for X-ray Crystallography and can be well approximated using the analytic expression

$$f_k\left(\frac{\sin\theta}{\lambda}\right) = \sum_{j=1}^{4} a_k^j \exp\left[-b_k^j \left(\frac{\sin\theta}{\lambda}\right)^2\right] + c_k, \tag{2.42}$$

where 2θ is the scattering angle, $\lambda = {}^{2\pi}/_{q}$ and a_k^j, b_k^j and c_k are parameters [131]. Finally, the summation in Eq. (2.40) can be carried out, leading to the unit-cell structure factor,

$$F_L(\mathbf{q}) = \sum_k f_k e^{i\mathbf{q}\cdot\mathbf{r}_k}. \tag{2.43}$$

2.3.2 Diffuse Scattering due to Dynamic Disorder

In a perfect crystal, the structure factor of a single unit cell L is given by Eq. (2.43) and since all unit cells are identical, F_L does not depend on the specific choice of L, hence $F_L \equiv F$. However, a real crystal is not perfectly periodic. The disorder present can be static, such as arises from crystal faults, impurities and structural inhomogeneity, or dynamic, *i.e.*, due to motions present at finite temperature. Here, the focus is on the effect of dynamic disorder, described as time-dependent atomic displacements, on the X-ray scattering. However, the presented analytical tools are also adequate to describe static disorder.

Disorder causes the structure factor F_L of a single unit cell L to deviate from the mean structure factor, F. Furthermore, if the displacements are time dependent so will F_L. The instantaneous position of the k^{th} atom in the L^{th} unit cell can be written as $\mathbf{r}_{kL}(t) = \mathbf{R}_L + \mathbf{r}_k + \mathbf{u}_{kL}(t)$, where $\mathbf{r}_k$ denotes the mean position vector in the unit cell which is, due to the average periodic structure, independent of L, and $\mathbf{u}_{kL}(t)$ is the time-dependent displacement. $\mathbf{R}_L$ and $\mathbf{r}_k$ are chosen such that $\langle \mathbf{u}_{kL}(t) \rangle = \mathbf{0}$. Henceforth, for clarity of notation, the dependence on the variables t and $\mathbf{q}$ is not explicitly written.

Generalisation of Eq. (2.43) to dynamic disorder leads to the time-dependent unit-cell structure factor,

$$F_L = \sum_k f_k e^{i\mathbf{q}\cdot(\mathbf{r}_k + \mathbf{u}_{kL})} \tag{2.44}$$

$$= \sum_k g_k^L e^{i\mathbf{q}\cdot\mathbf{r}_k}, \tag{2.45}$$

where the effective atomic form factor,

$$g_k^L = f_k e^{i\mathbf{q}\cdot\mathbf{u}_{kL}}, \tag{2.46}$$

[8] Although most of the electrons are not in an s-symmetric shell, asymmetries in the charge distribution are almost averaged out.

was introduced which can be written as

$$g_k^L = g_k + \phi_k^L, \tag{2.47}$$

where ϕ_k^L denotes the fluctuation from the mean, $g_k = \langle g_k^L \rangle$,

$$\phi_k^L = g_k^L - \langle g_k^L \rangle. \tag{2.48}$$

The total amplitude A_D scattered by the crystal from an incident plane wave A_i can then be written as the sum over all unit cells,

$$A_D \propto A_i \sum_L F_L e^{i\mathbf{q}\cdot\mathbf{R}_L} = A_i \sum_{kL} g_k^L e^{i\mathbf{q}\cdot(\mathbf{R}_L+\mathbf{r}_k)} \tag{2.49}$$

$$\propto A_i \left(\sum_{kL} g_k e^{i\mathbf{q}\cdot(\mathbf{R}_L+\mathbf{r}_k)} + \sum_{kL} \phi_k^L e^{i\mathbf{q}\cdot(\mathbf{R}_L+\mathbf{r}_k)} \right) \tag{2.50}$$

$$\propto A_i \sum_L e^{i\mathbf{q}\cdot\mathbf{R}_L} \underbrace{\sum_k g_k e^{i\mathbf{q}\cdot\mathbf{r}_k}}_{\langle F_L \rangle \equiv F} + A_i \sum_L e^{i\mathbf{q}\cdot\mathbf{R}_L} \sum_k \phi_k^L e^{i\mathbf{q}\cdot\mathbf{r}_k}. \tag{2.51}$$

The scattered intensity is then given by

$$I \propto \left\langle |A_D|^2 \right\rangle \propto |A_i|^2 (I_1 + I_2), \tag{2.52}$$

where I_1 and I_2 represent the squared magnitude of the first and second term in Eq. (2.51), respectively; the cross term vanishes due to $\langle \phi_k^L \rangle = 0$. I_1 is readily evaluated,

$$I_1 = |F|^2 \sum_{LL'} e^{i\mathbf{q}\cdot(\mathbf{R}_L - \mathbf{R}_{L'})} = |F|^2 N_L \boldsymbol{\delta}(\mathbf{Q} - \mathbf{q}), \tag{2.53}$$

and identified with the Bragg scattering, N_L being the number of unit cells in the crystal. Using Eqs. (2.46) to (2.48), I_2 can be written as

$$I_2 = \sum_{LL'} e^{i\mathbf{q}\cdot(\mathbf{R}_L - \mathbf{R}_{L'})} \sum_{kk'} f_k f_{k'} e^{i\mathbf{q}\cdot(\mathbf{r}_k - \mathbf{r}_{k'})} \times \left\langle \left[e^{i\mathbf{q}\cdot\mathbf{u}_{kL}} - \left\langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}} \right\rangle \right] \left[e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}} - \left\langle e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}} \right\rangle \right] \right\rangle \tag{2.54}$$

$$= \sum_{LL'} e^{i\mathbf{q}\cdot(\mathbf{R}_L - \mathbf{R}_{L'})} \sum_{kk'} f_k f_{k'} e^{i\mathbf{q}\cdot(\mathbf{r}_k - \mathbf{r}_{k'})} \times \left[\left\langle e^{i\mathbf{q}\cdot(\mathbf{u}_{kL} - \mathbf{u}_{k'L'})} \right\rangle - \left\langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}} \right\rangle \left\langle e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}} \right\rangle \right] \tag{2.55}$$

$$= \sum_{LL'} \sum_{kk'} f_k \langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}} \rangle f_{k'} \langle e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}} \rangle \times e^{i\mathbf{q}\cdot(\mathbf{R}_L - \mathbf{R}_{L'} + \mathbf{r}_k - \mathbf{r}_{k'})} \left[\frac{\left\langle e^{i\mathbf{q}\cdot(\mathbf{u}_{kL} - \mathbf{u}_{k'L'})} \right\rangle}{\langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}} \rangle \langle e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}} \rangle} - 1 \right]. \tag{2.56}$$

I_2 is not localised to reciprocal lattice vectors but distributes intensity over all of the reciprocal space and is denoted *X-ray diffuse scattering*. Therefore, the total X-ray scattering from a crystal is given by the sum of Bragg and diffuse scattering,

$$I_{\text{tot}} = I_{\text{Bragg}} + I_{\text{diff}}. \tag{2.57}$$

Thus, the diffuse scattering is given by

$$I_{\text{diff}} = I_{\text{tot}} - I_{\text{Bragg}} \propto \langle |F_L|^2 \rangle - |\langle F_L \rangle|^2. \tag{2.58}$$

2.3.3 Harmonic Approximation

Although exact, Eq. (2.56) is not particularly useful for practical applications since the averages $\langle\cdot\rangle$ need to be computed. To do this, the cumulant expansion for the characteristic function of a stochastic variable X with the realisations x is used [132],

$$f_X(k) = \langle e^{ikx}\rangle = \exp\left(\sum_{n=1}^{\infty}\frac{(ik)^n}{n!}C_n(X)\right), \tag{2.59}$$

where $C_n(X)$ denotes the n^{th}-order cumulant; the first three cumulants being

$$C_1(X) = \langle x\rangle, \tag{2.60}$$
$$C_2(X) = \langle x^2\rangle - \langle x\rangle^2 \tag{2.61}$$

and

$$C_3(X) = \langle x^3\rangle - 3\langle x\rangle\langle x^2\rangle + 2\langle x\rangle^3. \tag{2.62}$$

Identifying $\mathbf{u}_{kL}$ with the realisations of X and bearing in mind $\langle\mathbf{u}_{kL}\rangle = \mathbf{0}$, the first three cumulants reduce to

$$C_1(X) = 0, \qquad C_2(X) = \langle x^2\rangle \qquad \text{and} \qquad C_3(X) = \langle x^3\rangle. \tag{2.63}$$

If the displacements $\mathbf{u}_{kL}$ are harmonic, *i.e.*, $\mathbf{u}_{kL}$ has a gaussian distribution[9], then $C_{n\neq 2}(X) = 0$ and

$$\langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}}\rangle = e^{-W_{kL}(\mathbf{q})}, \tag{2.64}$$

where the Debye-Waller factor was introduced, given by [133, 134]

$$W_{kL}(\mathbf{q}) = \frac{1}{2}\langle(\mathbf{q}\cdot\mathbf{u}_{kL})^2\rangle \tag{2.65}$$
$$= \frac{1}{2}\sum_{\alpha\beta}U_{kL}^{\alpha\beta}q_\alpha q_\beta \tag{2.66}$$
$$= \frac{1}{2}\mathbf{q}^T\langle\mathbf{u}_{kL}\mathbf{u}_{kL}^T\rangle\mathbf{q}, \tag{2.67}$$

where $U_{kL}^{\alpha\beta}$ are the elements of a 3×3 symmetric tensor, $\langle\mathbf{u}_{kL}\mathbf{u}_{kL}^T\rangle$ and $(\cdot)^T$ denotes the transpose of $(\cdot)$. The remaining average is calculated accordingly; it separates into a product of three components, two of which cancel with the denominator in Eq. (2.56), yielding

$$\frac{\langle e^{i\mathbf{q}\cdot(\mathbf{u}_{kL}-\mathbf{u}_{k'L'})}\rangle}{\langle e^{i\mathbf{q}\cdot\mathbf{u}_{kL}}\rangle\langle e^{-i\mathbf{q}\cdot\mathbf{u}_{k'L'}}\rangle} = e^{\mathbf{q}^T\langle\mathbf{u}_{kL}\mathbf{u}_{k'L'}^T\rangle\mathbf{q}}, \tag{2.68}$$

where the displacement *variance-covariance matrix*, $\langle\mathbf{u}_{kL}\mathbf{u}_{k'L'}^T\rangle$ was introduced. The intensity of the X-ray diffuse scattering in the harmonic approximation is then given by

$$I_{\text{diff}} \propto \sum_{LL'}\sum_{kk'} f_k e^{-W_{kL}} f_{k'} e^{-W_{k'L'}} e^{i\mathbf{q}\cdot(\mathbf{R}_L-\mathbf{R}_{L'}+\mathbf{r}_k-\mathbf{r}_{k'})}\left[e^{\mathbf{q}^T\langle\mathbf{u}_{kL}\mathbf{u}_{k'L'}^T\rangle\mathbf{q}}-1\right]. \tag{2.69}$$

[9] Due to the factorial in Eq. (2.59), the equations derived here will approximately hold even if the displacements $\mathbf{u}_{kL}$ are not harmonic, provided the probability distribution of $\mathbf{u}_{kL}$ is approximately symmetric and centred around the origin.

However, the evaluation of the covariances between all atoms in the crystal is in equal measure impossible as it is useless, the former due to the large number of atoms in the crystal, $\mathcal{O}(N_A)$ with N_A being Avogadro's number, leading to $\mathcal{O}(N_A^2)$ covariances, and the latter because the sheer size of such a matrix renders an unambiguous interpretation futile.

Here, the assumption is made that the motions between atoms belonging to different unit cells are independent, leading to vanishing covariances between these atoms and thus Eq. (2.69) simplifies to

$$I_{\text{diff}} \propto \sum_{kk'} f_k e^{-W_k} f_{k'} e^{-W_{k'}} e^{i\mathbf{q}\cdot(\mathbf{r}_k-\mathbf{r}_{k'})} \left[e^{\mathbf{q}^T\langle \mathbf{u}_k\mathbf{u}_{k'}^T\rangle\mathbf{q}} - 1\right] \tag{2.70}$$

$$\begin{aligned} &\propto \sum_k |f_k|^2 \left[1 - e^{-2W_k}\right] \\ &\quad + 2\sum_{k<k'} f_k e^{-W_k} f_{k'} e^{-W_{k'}} \Re\left[e^{i\mathbf{q}\cdot(\mathbf{r}_k-\mathbf{r}_{k'})}\right] \left[e^{\mathbf{q}^T\langle \mathbf{u}_k\mathbf{u}_{k'}^T\rangle\mathbf{q}} - 1\right], \end{aligned} \tag{2.71}$$

where in the last step the contributions of variances and covariances were separated; $\Re[\cdot]$ denotes the real part of a complex number.

Computational Considerations

Estimates for upper bound values of the Debye-Waller factor and the variance-covariance matrix yield

$$\max\left\{\mathbf{q}^T\langle \mathbf{u}_k\mathbf{u}_{k'}^T\rangle\mathbf{q}\right\} \propto \mathcal{O}_{\mathbf{q}^2}(4\pi^2)\mathcal{O}_{\mathbf{u}^2}(10) \sim \mathcal{O}(100). \tag{2.72}$$

Due to the limited accuracy of numerical methods[10] Eq. (2.71) was re-written prior implementation into a computer program to avoid large positive arguments in the exponential, yielding

$$\begin{aligned} I_{\text{diff}} \propto &\sum_k |f_k|^2 \left[1 - e^{-2W_k(\mathbf{q})}\right] \\ &+ 2\sum_{k<k'} f_k f_{k'} \cos\left[\mathbf{q}\cdot(\mathbf{r}_k - \mathbf{r}_{k'})\right] \left[e^{\mathbf{q}^T\langle \mathbf{u}_k\mathbf{u}_{k'}^T\rangle\mathbf{q} - W_k(\mathbf{q}) - W_{k'}(\mathbf{q})} - e^{-W_k(\mathbf{q})-W_{k'}(\mathbf{q})}\right]. \end{aligned} \tag{2.73}$$

2.3.4 Experimental Data

In pioneering work by Gruner and coworkers at the Cornell High-Energy Synchrotron Source images of diffuse X-ray diffraction from *Staphylococcal* nuclease were obtained from crystals at many different orientations relative to the incident beam, and, for the first time, the data were reduced[11] to measurements of diffuse intensity on the reciprocal lattice [101]. The experimentally-obtained map is 99.5% complete in the reciprocal-space resolution range 10 to 2.5 Å, in total comprising 55,691 data points in the range $\|\mathbf{q}\| < 0.62$ Å^{-1} and arises from the diffuse scattering of the complete unit cell, *i.e.*, proteins and crystal solvent, at room temperature.

[10] Numerical accuracy depends mainly on two aspects; firstly, on the specific implementation of mathematical functions, which is, for example, often based on a truncated Taylor series expansion, *e.g.* for functions like $\exp(\cdot)$ or $\cos(\cdot)$, and secondly, on the limited number of digits used to store a number, usually 17 digits for a 64-bit `double` or 31 digits for a 128-bit representation.

[11] A mode-filtering technique was used to remove the Bragg peaks from the series of two-dimensional still exposures.

Chapter 3

Fluctuations and Correlations

An accurate description of the dynamics of protein crystals is an important goal in molecular biophysics. Obtaining a simplified physical description of the motions influencing X-ray scattering from protein crystals should allow improvement of models refined against diffraction data together with a reduction in the number of independent parameters to be adjusted. Furthermore, understanding collective internal protein motions should provide information on dynamic aspects of protein function.

A detailed description of protein crystal dynamics can be obtained using molecular dynamics simulation. In this chapter, molecular dynamics simulations are used to characterise the positional fluctuations of individual atoms and their cross-correlations. Atomic positional fluctuations can be derived from X-ray crystallographic B factors [43], and can be compared with the fluctuations observed in MD simulation. On the sub-nanosecond timescale simulation-derived atomic fluctuations often differ significantly from experimental values. In recent comparisons the inclusion of nanosecond-timescale dynamics has lead to better qualitative agreement but fluctuations larger than the B factors [136–138].

Cross-correlations in the atomic displacements indicate collective motion and are therefore of potential relevance to protein function. In MD simulations correlated motions have been detected [25, 139, 140] and have been used to make deductions concerning dynamical aspects of protein function [10, 141–143] but converge relatively slowly [99, 136]. Correlated motions present in protein crystals can in principle be probed experimentally using X-ray diffuse scattering [27, 76, 80, 83].

Advances in computational resources and methodology constantly improve the timescale and system size accessible to MD simulation. In this chapter, results are presented of four 10 ns MD simulations of crystalline SNase (trajectories T1-3 and O1, cf. Section 2.2). A detailed analysis is made of the atomic fluctuations and cross-correlations, their convergence properties are examined, and their relation to the protein topology is determined. Then, by comparing the calculated with the experimental X-ray diffuse scattering an estimate is made of the simulation time required for the variance-covariance matrix, and thus the correlated motions, to converge.

First, a comparison is made of the MD simulations with experiment on three levels: the crystal parameters and average structure, the atomic fluctuations, and the X-ray diffuse scattering. Subsequently, correlations in C_α-atom motions are investigated.

Results of this chapter have been published in *Meinhold et al.*; Physica B:**350**, 127 (2004), cf. Ref. [135] and *Meinhold & Smith*; Biophys J:**88**, 2554 (2005), cf. Ref. [100].

3.1 Comparing Simulation Results with Experimental Data

3.1.1 Crystal Parameters and Average Structure

The average values for the MD unit-cell sides and volumes are given in Table 3.1. For the trajectories Tx (x=1,2,3) and O1 these values deviate by less than 0.8% and 2.8% from the experimental values, respectively. The unit-cell volume averaged over all simulations is within 0.35±0.23% of the experimental values.

To compare the simulation structure with the experimental coordinates (*i.e.* with the coordinates of the PDB structure 2SNS) the average MD protein structure was computed as the time average $\mathbf{R}_{\rm ave} = \left\langle \bigcup_{i,\rm Tx} \mathbf{R}_i^{\rm Tx}(t) \right\rangle_t$, where $\mathbf{R}_i^{\rm Tx}(t)$ is the coordinate vector of protein i ($i=1,\ldots,4$) in the simulation Tx. $\bigcup$ denotes the union of sets which was performed in such a way that each $\mathbf{R}_i^{\rm Tx}(t)$ was optimally superposed[1] on a reference structure[2]. Thus, for certain analyses the trajectories Tx were merged into a single-protein trajectory with an effective length of 120 ns. The RMSD of $\mathbf{R}_{\rm ave}$ from experiment was computed for the experimentally-resolved residues 1-141. The result for all non-hydrogen atoms is 1.67 Å, and for the C_α-atoms is 1.29 Å.

3.1.2 Fluctuations of Single Atoms

X-ray crystallographic B factors, B_k arise from the fluctuations of the individual atoms k around their space-group symmetric positions as given by

$$\langle \mathbf{u}_k^2 \rangle = \frac{3}{8\pi^2} B_k. \tag{3.1}$$

The fluctuations may arise from internal protein motions, from translation and rotation of the protein molecules in the unit cell, and from motions of the unit cells relative to each other. Experimentally, other effects such as lattice distortion and refinement errors may also contribute. The relative motions of unit cells are suppressed in the simulations due to the imposition of periodic boundary conditions. Thus, the remaining fluctuations contain components from whole-molecule translation and rotation and from internal motion. The protein fluctuations were derived from

[1] The superposition was performed using Charmm by minimising the RMS deviation (RMSD) between the simulation and the reference structure.

[2] The single-protein mean structure, $\langle \mathbf{R}_i^{\rm Tx}(t) \rangle_t$ was used as reference.

Table 3.1: Average values for the MD unit-cell sides and volumes for the simulations (T1-3, O1) and the experimental values [101].

	a [Å]	b [Å]	c [Å]	V [Å^3]
EXP	48.5	48.5	63.4	149132.65
T1	48.3±0.2	48.3±0.2	63.8±0.4	148555±419
T2	48.6±0.2	48.6±0.2	63.0±0.4	148706±413
T3	48.6±0.2	48.6±0.2	62.9±0.5	148695±392
O1	49.9±0.7	47.6±0.3	62.6±0.8	148634±399

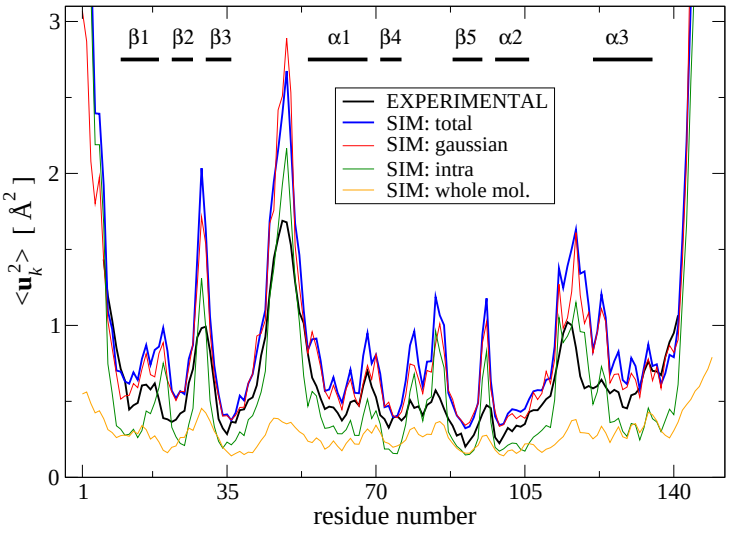

Figure 3.1: Mean-square C_α-atom fluctuations calculated from the MD trajectories and from experiment. The simulation-derived fluctuations are: the total $\langle \mathbf{u}_k^2 \rangle$, including internal motion and whole-molecule translation and rotation (SIM: total); the total $\langle \mathbf{u}_k^2 \rangle$ with a Gaussian fit to the atomic positional distribution (SIM: gaussian, see text); the contribution from internal motion (SIM: intra); and the contribution from whole-molecule translation and rotation (SIM: whole mol.). The decomposition of motions is described in detail in the text, cf. Eqs. (3.2) and (3.3). Experimental values were computed from B factors using Eq. (3.1). B factors were not reported in the starting structure used for the MD simulations (2SNS). Therefore, the experimental B factors shown here are from PDB entry 1STN [126], a room-temperature structure refined at 1.7 Å with an R factor of 0.162 and very similar unit-cell parameters ($a = b = 48.5$ Å and $c = 63.5$ Å). Secondary structural elements are indicated, cf. Fig. 3.9.

the trajectories $\mathbf{R}_i^{\mathrm{Tx}}(t)$, x=1,2,3, and the results are compared with the experimental B factors in Fig. 3.1. The MD simulation fluctuations correlate well with those derived experimentally as supported by a correlation coefficient of 0.89. However, the MD fluctuations tend to be slightly larger, especially in the loop regions. Fig. 3.1 also shows that, on average, whole-molecule translations and rotations contribute $\approx$0.25 Å^2 to the total MS fluctuations.

One possible reason why the simulation-derived fluctuations are slightly larger than experiment may be the assumption made in deriving the experimental B factors that the fluctuations are harmonic. This assumption has been shown to lead to an underestimation of fluctuation magnitudes [144]. To examine the consequences of this assumption, the fluctuations for each atom were re-calculated from Gaussian fits to the simulation-derived atomic positional densities. The results, which correspond to the experimental isotropic B factors, are presented in Fig. 3.1. Use of the Gaussian assumption does indeed significantly reduce the fluctuations, especially in the loop regions. However, the approximation does not account for most of the difference with experiment, the MD fluctuations remaining higher than experiment.

The atomic mean-square fluctuations can be decomposed according to

$$\langle \mathbf{u}_k^2 \rangle = \langle ({}^{\mathrm{I}}\mathbf{u}_k + {}^{\mathrm{T}}\mathbf{u}_k + {}^{\mathrm{R}}\mathbf{u}_k)^2 \rangle \tag{3.2}$$

$$= \langle {}^{\mathrm{I}}\mathbf{u}_k^2 \rangle + \langle {}^{\mathrm{T}}\mathbf{u}_k^2 \rangle + \langle {}^{\mathrm{R}}\mathbf{u}_k^2 \rangle + K_k, \tag{3.3}$$

where $\langle \mathbf{u}_k^2 \rangle$ is the total mean-square fluctuation of atom k, the displacements ${}^{\mathrm{I}}\mathbf{u}_k$, ${}^{\mathrm{T}}\mathbf{u}_k$ and ${}^{\mathrm{R}}\mathbf{u}_k$

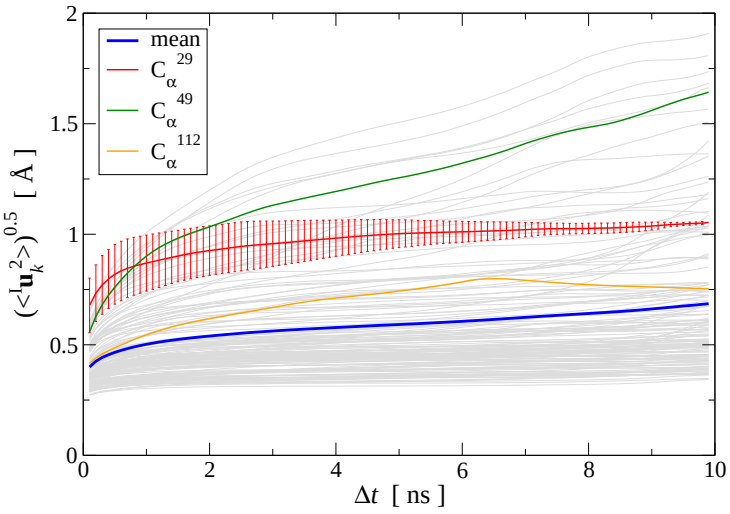

Figure 3.2: Protein internal root-mean-square fluctuations for all C_α-atoms averaged over sub-trajectory lengths $\triangle t$ (grey lines) shown for a single protein of trajectory T1. Highlighted in colour are the average over all C_α-atoms (blue) and three examples of different time dependence (cf. legend and text). Error bars denote the standard deviation. For clarity, only error bars for C_α^{29} are shown.

describe the fluctuations due to the protein internal motion, whole-molecule translation and whole-molecule rotation, respectively, and K_k represents correlations between the displacements ${}^{\rm X}\mathbf{u}_k$, X={I,T,R}. The mean-square fluctuations of the displacements ${}^{\rm X}\mathbf{u}_k$ were calculated from the MD trajectories as follows: to calculate ${}^{\rm I}\mathbf{u}_k$ a new trajectory was created in which the protein whole-molecule motion was removed by superposing all coordinate sets of a single-protein trajectory, $\mathbf{R}_i(t)$, on a single-protein reference structure (the mean structure of $\mathbf{R}_i(t)$). To calculate the whole-molecule rigid-body displacements, ${}^{\rm T}\mathbf{u}_k$ and ${}^{\rm R}\mathbf{u}_k$, a new trajectory was then created by superposing the mean structure of $\mathbf{R}_i(t)$ on each coordinate set of $\mathbf{R}_i(t)$. For protein internal motions the mean-square fluctuation is $\left\langle\langle {}^{\rm I}\mathbf{u}_k^2\rangle\right\rangle_k = 0.74\,\text{Å}^2$. In comparison, whole-molecule translation contributes $\left\langle\langle {}^{\rm T}\mathbf{u}_k^2\rangle\right\rangle_k = 0.13\,\text{Å}^2$ and the whole-molecule rotational contribution is $\left\langle\langle {}^{\rm R}\mathbf{u}_k^2\rangle\right\rangle_k = 0.16\,\text{Å}^2$. The whole-molecule rotational contribution converged on the 1 ns timescale with an average molecule-rotation angle of $1.2 \pm 0.1°$. The translational contribution converged more slowly, on the 10 ns timescale. Thus, the total mean-square fluctuation $\left\langle\langle \mathbf{u}_k^2\rangle\right\rangle_k = 1.07\,\text{Å}^2$ arises mostly (69%) from internal protein motion, with whole-molecule translation and rotation contributing equally to 27%. Although cross-correlations between the atomic displacements ${}^{\rm X}\mathbf{u}_k$, $\langle K_k\rangle_k = 0.04\,\text{Å}^2$ contribute only 3% of the mean-square fluctuations, their average magnitude $\langle |K_k|\rangle_k = 0.09\,\text{Å}^2$ corresponds to 8% of the total fluctuations. This indicates that X={I,T,R} is not an optimal basis for decomposition, due to the fact that, for a non-rigid body, the internal and whole-molecule motion cannot be strictly distinguished, leading to $K_k \neq 0$.

A related question is whether the simulation-derived $\langle {}^{\rm I}\mathbf{u}_k^2\rangle$ values in Fig. 3.1 have converged. To examine this, the following analysis scheme was used.

Convergence of Time Series

The convergence of any given observable or time series, $O(t)$ was investigated as follows. First, the trajectory was grouped into n time-windows of length $\triangle t$ which may overlap. Then, the

observable was calculated on each sub-trajectory yielding n measurements $\{O_i(\triangle t)\}$. The mean and variance of these sets and their dependence on $\triangle t$ were used to study convergence. If, for all time-windows with $\triangle t > \triangle t^*$, the observable has the same value to within $\epsilon > 0$, then $\triangle t^*$ is considered to be the convergence time.

The above analysis scheme will be used throughout this chapter. First, the convergence of protein-internal fluctuations are studied.

The results for all simulations are quantitatively similar. Those for a selected protein from simulation T1 are depicted in Fig. 3.2. For the majority (≈65%) of atoms the fluctuations reach a plateau after 2-4 ns and the error bars gradually reduce during the progression of the simulation. This behaviour is exemplified by C_α^{29} in the figure, which is situated in a turn between two β-strands. For a smaller fraction (≈15%) of atoms, however, the fluctuations significantly increase throughout the simulation, as exemplified in the figure by C_α^{49}. Most of these atoms are located either in a highly-flexible loop region (residues 46-51) or in the very mobile C-terminus that was not visible crystallographically (residues 142-149). As a result, due to this smaller fraction the fluctuation averaged over the whole molecule increases steadily with time. The fluctuations of the remaining 20% of atoms exhibit more complex behaviour[3], one example being C_α^{112} which is situated in a coil region.

3.1.3 X-ray Diffuse Scattering

An improved version of the program Serena [145] was used to calculate three-dimensional scattering intensities according to Eq. (2.58)[4]. These intensities were compared with experimental diffuse scattering data reported in Ref. [101], cf. Section 2.3.4.

The agreement between the theoretical and experimental X-ray diffuse scattering data was investigated using diffuse scattering calculated on several timescales, $\triangle t$, as described in the

[3]For example, if an atom moves on an approximately elliptical orbit the quantity plotted in Fig. 3.2 will shown an oscillatory behaviour for a small number of orbit cycles.

[4]Due to the presence of crystal solvent, Eq. (2.58) has to be used instead of Eq. (2.70) because the assumption of small atomic displacements from an average position is not valid for the very mobile water molecules and counterions, which are homogeneously distributed within the unit cell. Eq. (2.58) is the exact formulation, whereas Eq. (2.70) is the harmonic approximation.

Table 3.2: Agreement factors $R_{\triangle t=10\,\mathrm{ns}}$ between the experimental diffuse X-ray pattern [101] and the scattering obtained from each trajectory. A and B are results of least-squares fits (see text) and the remaining two columns give extrapolated convergence times $\triangle t^*$ for specified values of R.

	$R_{\triangle t=10\,\mathrm{ns}}$	A	B	$\triangle t^*_{R=0.00}$	$\triangle t^*_{R=0.04}$
T1	8.15	9.90e-2	-7.07e-3	1.2 ms	4.2 μs
T2	8.35	9.91e-2	-6.45e-3	4.8 ms	9.6 μs
T3	8.08	9.81e-2	-7.45e-3	0.52 ms	2.4 μs
O1	8.56	9.86e-2	-5.49e-3	63 ms	43 μs

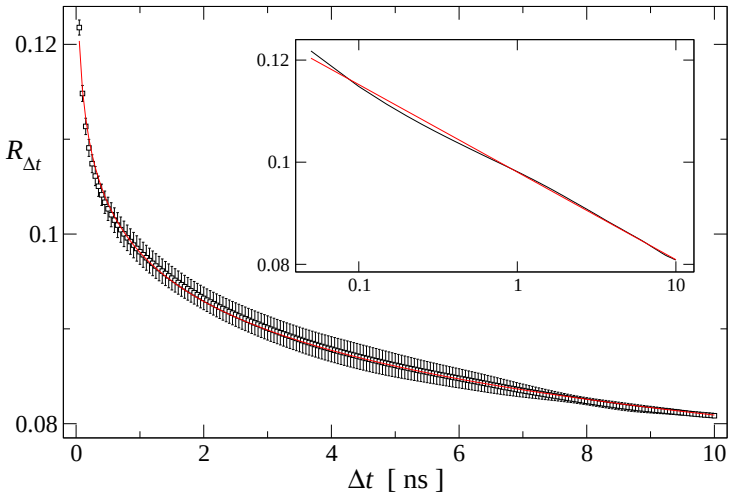

Figure 3.3: Dependence of the X-ray diffuse scattering agreement factor, $R_{\triangle t}$ on the averaging time interval $\triangle t$, the inset shows the data using a logarithmic timescale. The red line represents a least-squares fit (see text and Table (3.2)). The figure shows results for the simulation T3. The results from the other simulations are very similar, cf. Fig. 4.2. Error bars denote the standard deviation.

analysis scheme introduced in the preceding section. The agreement factor, R is defined as

$$R_{\triangle t} = \left\langle \min_{\{s,c\}} \frac{1}{N_q} \sum_{\mathbf{q}} \left| \frac{I^{\mathrm{exp}}(\mathbf{q}) - [sI^{\mathrm{sim}}_{\triangle t}(\mathbf{q}) + c]}{I^{\mathrm{exp}}(\mathbf{q})} \right| \right\rangle , \tag{3.4}$$

where $\langle \cdot \rangle$ denotes the average over all sub-trajectories of length $\triangle t$, N_q is the number of scattering vectors, and s and c are the coefficients for intensity scaling and background intensity, respectively. Optimal agreement with the experimental X-ray diffuse scattering pattern is obtained when the variance-covariance matrix has converged in the simulations.

In Chapter 4, Eq. (3.4) is used to evaluate models of protein motion by comparing the model scattering with that derived directly from the simulation. The notation then changes to $I^{\mathrm{exp}} \rightarrow I^{\mathrm{MD}}$ and $I^{\mathrm{sim}} \rightarrow I^{\mathrm{model}}$. Furthermore, as the reference pattern was then calculated from a simulation, the background intensity was zero and hence $c \equiv 0$.

Fig. 3.3 presents the time-dependence of the agreement factor, $R_{\triangle t}$ between the simulation with the lowest R factor (T3) and the experimental scattering data from Ref. [101]. $R_{\triangle t}$ reduces to $\approx$10% within the first 500 ps and then continues to decrease slowly on the nanosecond timescale, reaching 8.1% at 10 ns. The continued decrease of $R_{\triangle t}$ with time indicates that the scattering has not yet converged in the simulation but that the scattering intensity more closely resembles the experiment when the nanosecond-timescale dynamics are included. The red line shows the least-squares fit of $R(\triangle t; A, B) = A + B \ln(\frac{\triangle t}{[\mathrm{ns}]})$ and demonstrates that the agreement factor decays logarithmically with time. Again, for all simulations, $R_{\triangle t}$ is similar, cf. Fig. 4.2 on pg. 43. The R factors and fit results from each simulation are given in Table 3.2.

3.2 Correlations in C_α-Atom Displacements

The B factor results in Fig. 3.1 and the diffuse scattering R factor plot in Fig. 3.3 suggest that the agreement with crystallographic experiments is sufficient to warrant a more probing analysis of the simulation data. To do this, the correlations in C_α-atom motion are now examined and related to the inter-atomic distances and protein topology. The convergence properties of the C_α-atom displacement correlation matrix are also investigated.

The analysis utilises the following definitions and considerations.

Correlation Matrix

The variance-covariance matrix, $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ characterises the protein collective motions. However, the large number of independent elements makes $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ cumbersome to calculate for the whole unit cell. Instead, the correlation matrix, $C_{kk'}$ for the relative displacements of all C_α-atoms in the unit cell was calculated where $C_{kk'}$ is given by

$$C_{kk'} = \frac{\langle \mathbf{u}_k^T \cdot \mathbf{u}_{k'} \rangle}{\sqrt{\langle \mathbf{u}_k^2 \rangle \langle \mathbf{u}_{k'}^2 \rangle}} \tag{3.5}$$

For an atom pair (kk'), the correlation matrix element, $C_{kk'}$ is determined by the trace of the variance-covariance matrix, *i.e.*, $\langle \mathbf{u}_k^T \cdot \mathbf{u}_{k'} \rangle = \mathrm{Tr} \langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$. The differences between $C_{kk'}$ and $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ are two-fold. Firstly, anisotropic correlations, *i.e.*, off-diagonal elements which describe correlations between the displacements in, *e.g.*, the x and y directions, are not included in $C_{kk'}$. And secondly, $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ is amplitude weighted whereas $C_{kk'}$ is normalised. Alternative measures that overcome these problems are discussed in Ref. [27]. However, for the present purposes $C_{kk'}$ provides a convenient way of determining the correlated motions present.

Absolute Values of $C_{kk'}$

There is no unique way of removing global translation and rotation from MD trajectories [136, 146–148]. Furthermore, MD algorithms that periodically remove global translation and rotation of the simulated system potentially introduce artefactual anticorrelations. To understand this, consider a one-dimensional system composed of two particles at positions $x_1(t) = 0$ and $x_2(t) = x(t)$ where $x(t)$ is an arbitrary displacement. Then Eq. (3.5) yields $C_{x_1 x_2} = 0$. In the centre-of-mass frame, however, $x_1(t) = -\frac{1}{2}x(t)$ and $x_2(t) = +\frac{1}{2}x(t)$ and hence $C_{x_1 x_2} = -1$.

To investigate the amplitude of the effect of removing global protein motion on $C_{kk'}$ a 100 ps segment of the MD simulation T1 without the removal of global translation and rotation was examined. The correlation coefficients, $C_{kk'}$ were directly computed from this trajectory segment using Eq. (3.5). Subsequently, global translation and rotation were removed and again $C_{kk'}$ was computed. Removal of global translation and rotation reduces the correlation coefficient by $\approx$-0.1 for intra- and inter-protein motions almost independently of the inter-atomic distance, cf. Fig. 3.4. Consequently, the average values of $C_{kk'}$ given in Figs. 3.5, 3.8 and 3.9 A are underestimated by about 0.1.

Convergence of $C_{kk'}$

To study the convergence properties of $C_{kk'}$ each trajectory was divided into non-overlapping sub-trajectories as described in Section 3.1.2. For each sub-trajectory the correlation matrix was

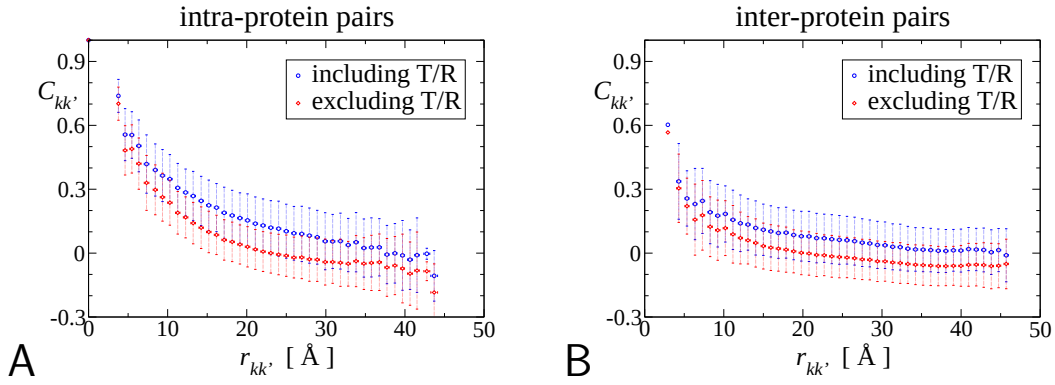

Figure 3.4: Correlation coefficient, $C_{kk'}$ for atom pairs belonging to the same protein (A) or to different proteins (B) calculated from a 100 ps trajectory with the overall unit-cell translation and rotation (T/R) included (blue) or removed (red). Error bars denote the standard deviation.

computed. For each $\triangle t$, the correlation matrices for each sub-trajectory were averaged element-wise to yield the mean, $\langle C_{kk',\triangle t}\rangle$ and standard deviation, $\sigma_{C_{kk',\triangle t}}$. The relative error, $\triangle C_{kk',\triangle t}$ is then

$$\triangle C_{kk',\triangle t} = \frac{\sigma_{C_{kk',\triangle t}}}{|\langle C_{kk',\triangle t}\rangle|}. \tag{3.6}$$

For all quantities, the subscript $\triangle t$ highlights the dependence on the length of the sub-trajectories. The convergence criterion chosen was: $(\sigma_{C_{kk',\triangle t}} < 0.05) \vee (\triangle C_{kk',\triangle t} < 0.2)$, where $\vee$ symbolises the logical *or*. It was checked that the qualitative results do not depend on the specific choices of $\sigma_{C_{kk',\triangle t}}$ or $\triangle C_{kk',\triangle t}$. If the number of sub-trajectories is small ($n \lesssim 10$) the statistics worsen, leading to an overestimation of the number of converged elements due to a small but non-zero probability that matrix elements coincidentally satisfy the convergence criterion. The amplitude of this effect was estimated using a corresponding number of random matrices. The true converged fraction of the correlation matrix, $f_{C^{\text{corr}}_{kk'}}$, was then calculated as the difference $f_{C_{kk'}} - f_{C^{\text{rand}}_{kk'}}$ where $f_{C_{kk'}}$ and $f_{C^{\text{rand}}_{kk'}}$ were calculated from $C_{kk'}$ and the random matrices, respectively.

The above definitions and considerations are now applied to the trajectories T1-3 and O1, *i.e.* those trajectories forming the MD simulation Set 1.

In Fig. 3.5 are shown the displacement correlation matrix elements, $C_{kk'}$ for intra- and inter-protein atom pairs, calculated for trajectory T1 and plotted against their corresponding inter-atomic distances, $r_{kk'}$. The average values of $C_{kk'}$ as a function of $r_{kk'}$ are indicated by the blue lines. Intra-protein motions of atom pairs separated by less than $\approx$20 Å are mostly positively correlated. For larger atomic separations the values of $C_{kk'}$ range approximately from -0.5 to $+0.5$ for both distributions, *i.e.* for intra- and inter-protein atom pairs. For $r_{kk'} < 20$ Å the intra-protein motions exhibit a significantly higher degree of correlation than motions between different proteins. At larger separations intra- and inter-protein motions on average appear slightly anticorrelated. However, as described above, this anticorrelation is an artefact due to the MD algorithm periodically removing the centre-of-mass translation and rotation of the unit cell. Consequently, motions of atoms separated by $r_{kk'} \gtrsim 20$Å are, on average, uncorrelated.

Now, the convergence properties of $C_{kk'}$ are investigated. To address this issue the absolute and relative errors, $\sigma_{C_{kk',\triangle t}}$ and $\triangle C_{kk',\triangle t}$, respectively, were computed for non-overlapping sub-

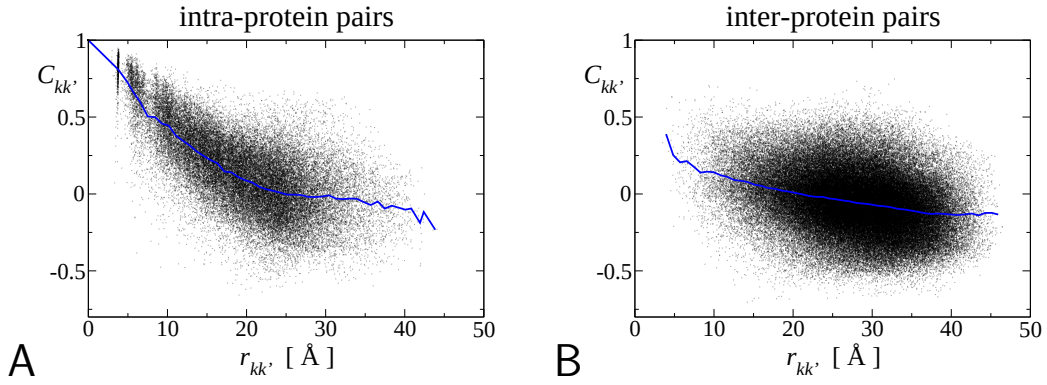

Figure 3.5: Displacement correlation matrix elements for all C_α-atom pairs plotted against their average distance, $r_{kk'}$ for intra-protein motions (A) and inter-protein motions (B). Both graphs were computed for the full trajectory T1 with length 10 ns. The blue lines show the average values for distance intervals of width 1 Å. The atomic distance, $r_{kk'}$ between two atoms k and k' was calculated as $r_{kk'} = \min_{\mathbf{R}}\{\|\mathbf{r}_k - \mathbf{r}_{k'} + \mathbf{R}\|\}$, where $\mathbf{R}$ is a lattice vector, cf. Eq. (2.4), and thus $r_{kk'}$ is the minimum distance between atom k and any crystal image of atom k' created by the periodic boundary conditions.

trajectories with lengths $\triangle t$=0.1,...,5.0 ns and the converged fraction, $f_{C_{kk'}^{\mathrm{corr}}}$ was calculated as described above. Only intra-protein correlations showed a significant degree of convergence. Hence, only these correlations are discussed in the following. Fig. 3.6 presents the dependence of the intra-protein $f_{C_{kk'}^{\mathrm{corr}}}$ on $\triangle t$. Apart from in simulation T2, the converged fraction increases roughly logarithmically with $\triangle t$. Fig. 3.6 also shows that the degree of convergence varies significantly between the simulations. For the equivalent simulations Tx (x=1,2,3) one finds $f_{\mathrm{T2}} > f_{\mathrm{T1}} > f_{\mathrm{T3}}$ indicating that simulation T2 possesses the highest degree of convergence. In contrast, Table 3.2 indicates that the agreement factors R display the reverse order, with simulation T3 in best agreement with experiment. This apparent conflict may be resolved by considering a simulation restricted to a region of phase space significantly smaller than that explored in the experiment. The variance-covariance matrix may appear to approach convergence in this simulation. If, then, the simulation crosses a free-energy barrier into another, previously unexplored region of phase space, then the variance-covariance matrix will appear to be far from convergence while the elements themselves may resemble the experimental values more closely.

It is also of interest whether the matrix elements, $C_{kk'}$ converge to the same value in different simulations. This can be investigated by calculating $f_{C_{kk'}^{\mathrm{corr}}}$ for the merged trajectories Tx and Tx+O1 (x=1+2+3), respectively. If $C_{kk'}$ converges to the same value in all simulations it contributes to $f_{C_{kk'}^{\mathrm{corr}}}$, otherwise it does not. On the other hand, if $C_{kk'}$ does not converge in any single simulation then it is unlikely to converge in the merged trajectories. Therefore, the maximum value of $f_{C_{kk'}^{\mathrm{corr}}}$ for the merged trajectories is determined by the simulation with the lowest degree of convergence *i.e.* T3. $f_{C_{kk'}^{\mathrm{corr}}}$ for the merged trajectories are also shown in Fig. 3.6. For $\triangle t$=2.5 ns the converged fractions of Tx and Tx+O1 differ by 5% and 13% from T3, respectively. Thus, the majority (87%) of the converged matrix elements converged (within the error set by the convergence criterion) to the same value in all four simulations.

In Fig. 3.7 the intra-protein $f_{C_{kk'}}$ is plotted against the inter-atomic distance, the fraction calculated from a random matrix, $f_{C_{kk'}^{\mathrm{rand}}}$, being negligible $(9\pm1)10^{-4}$. For shorter distances,

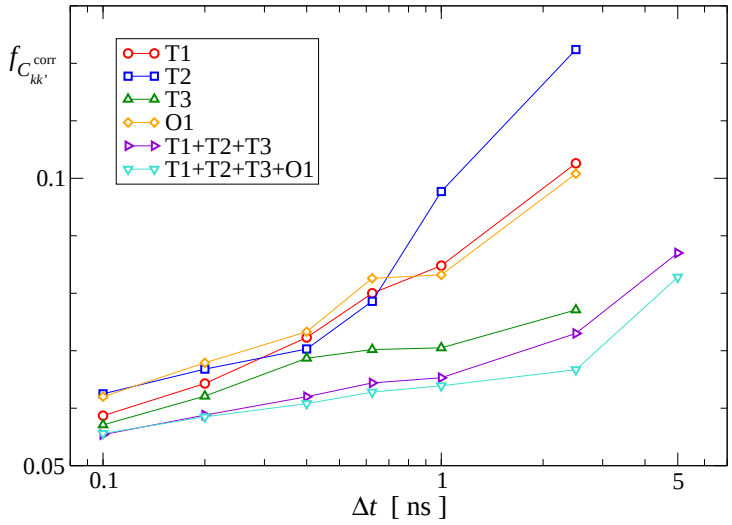

Figure 3.6: Fractions of converged correlation matrix elements, $f_{C_{kk'}^{corr}}$ at different lengths, $\triangle t$ of sub-trajectories for all simulations and two merged trajectories. Results are shown for intra-protein motions.

$r_{kk'} \lesssim 6$ Å, more than 80% of the intra-protein correlation matrix converges on the timescale of $\triangle t$=1.0 ns. This fraction drops to ≈10% for $r_{kk'} = 12$ Å and becomes effectively zero for separations larger than 15 Å.

Another point of interest is the dependence of the correlation matrix on the protein topology. To assess this, only converged elements of the intra-protein correlation matrix are examined. Due to the covalent bonding structure, one might expect the correlations between atoms local in the sequence to be larger than those between neighbouring atoms which are non-local in sequence. The mean values, $\langle C_{kk',\Delta t}\rangle$ for atom pairs that are local in sequence, defined here as $\Delta = k - k' < 10$ (note that for C_α-atoms k and k' are identical with the residue numbers), are compared in Fig. 3.8 with pairs that are non-local in sequence ($\Delta \geq 10$) as a function of $\langle r_{kk'}\rangle$. Indeed, the displacement correlations between atoms that are local in sequence are on average larger than those between atoms that are non-local in sequence. The difference is ≈0.1, independent of $\langle r_{kk'}\rangle$. Furthermore, the inset in Fig. 3.8 clearly demonstrates that the degree of correlation for non-local atoms decreases only slowly with increasing separation. Closely resembling the average distance dependence for all matrix elements shown in Fig. 3.5 A, the magnitude of the converged elements decreases approximately exponentially with $\langle r_{kk'}\rangle$ with a decay length of 10.6 Å for $\langle r_{kk'}\rangle < 25$ Å, beyond which $\langle C_{kk'}\rangle = 0$.

Finally, it is investigated where atoms, for which the intra-protein correlation matrix elements have converged, are located within the protein. In Fig. 3.9 A the converged intra-protein pairs of C_α-atoms in the simulations Tx (x=1,2,3) with $\triangle t$=2.5 ns are shown. 94% of all nearest and 68% of all second-nearest backbone neighbouring pairs converged in all four proteins in the unit cell. Most of the regions in which the correlations converged are located in secondary structural elements. The correlations between atoms located in α-helices converge for sequence distances $\Delta \lesssim 7$. In contrast, the atom pairs within β-strands converge only for $\Delta \lesssim 2$. The off-diagonal elements are converged atom pairs that are non-local in sequence. The majority of these are correlations between secondary structure elements. For almost all the strands in the *Staphylococcal* nuclease β-barrel the inter-strand correlations have converged. Moreover,

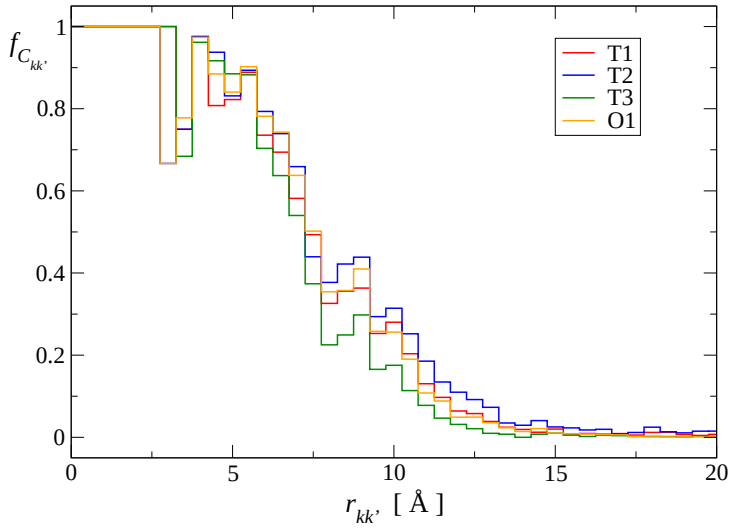

Figure 3.7: Fraction of the intra-protein correlation matrix elements that has converged, $f_{C_{kk'}}$ versus the atomic separation, $r_{kk'}$. Results presented are for sub-trajectory length $\triangle t$=1.0 ns.

there is some convergence between α_1 and α_2, and pronounced convergence between α_2 and α_3. Correlations between the α and β elements in general have not converged. Finally, a comparison between Figs. 3.1 and 3.9 demonstrates that the converged correlations exist in regions of low mean-square fluctuation.

3.3 Discussion and Conclusions

Molecular dynamics simulations (Set 1) of crystalline *Staphylococcal* nuclease have been analysed in terms of B factors, the atomic displacement correlation matrix and X-ray diffuse scattering.

The average protein structures in all simulations are in accordance with the experimental reference structure (1.3 Å average C_α RMSD). The unit-cell edge parameters and volume are also well reproduced, indicating that protein packing is described correctly.

Isotropic B (thermal) factors are widely used to derive atomic fluctuations from X-ray crystallographic protein structures. The question arises, however, as to whether an unambiguous description of the dynamics involved can be derived from B factor data in the absence of detailed additional information. B factors contain static and dynamic components [43, 73], and part of the static disorder may be temperature-independent. However, as the temperature of a protein is lowered some of the dynamic disorder may become static as proteins freeze into structurally inhomogeneous conformational substates. Separation of static from dynamic disorder is therefore non-trivial.

One approach is to fit simplified displacement models directly to the B factor distribution. However, this approach suffers from the drawback that many such qualitatively-different models may yield fits of similar quality. Thus, it has been shown that a rigid-body TLS (translation/libration/screw) displacements model [149], in which the whole protein is considered as rigid and internal dynamics is absent, accurately reproduces most [150] or all of protein B factor

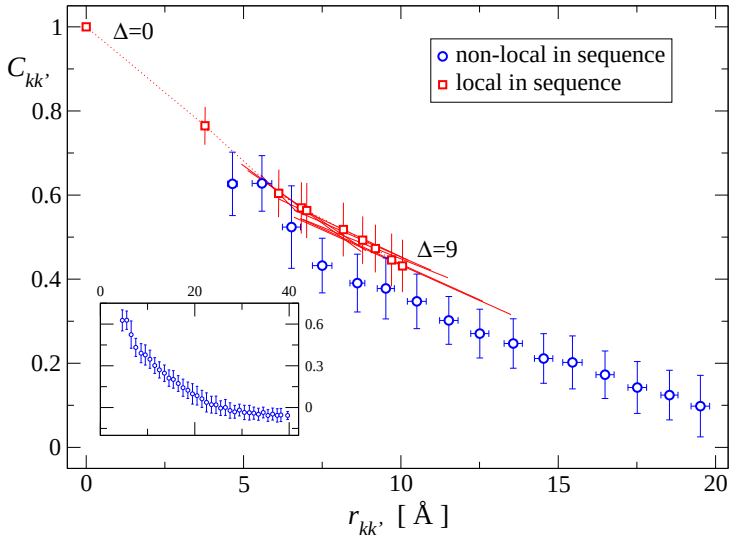

Figure 3.8: $\langle C_{kk',\Delta t}\rangle$ versus $\langle r_{kk'}\rangle$ for atom pairs local or non-local in sequence. For the local-in-sequence data error bars are drawn along the principal axes of each $(C_{kk',\Delta t}, r_{kk'})_\Delta$-distribution. The dotted line connects data points for atom pairs which are local in sequence. The inset shows the non-local-in-sequence data over the full distance range and shows a profile similar to Fig. 3.5 A. Results are shown for trajectory T2 with a sub-trajectory length Δt=1.0 ns. Error bars denote the standard deviation.

distributions [151]. Alternatively, TLS models in which parts of the protein are considered rigid (*e.g.*, aromatic side chains) have also been shown to successfully fit B factor data [152, 153]. An alternative model involves rigid-protein TLS degrees of freedom together with the internal dynamics described by normal mode eigenvectors with refinable amplitude factors [150, 154, 155]. Moreover, a very simplified Gaussian vibrational model, in which the protein is described as an elastic network of locally interconnected C_α-atoms in the absence of whole-molecule motion was also shown to produce B factor distributions that agree very well with experiment [156–158]. Very recently, it has been demonstrated that B factor distributions are closely correlated to local protein packing densities [87].

The availability of the above variety of fundamentally-different models, all of which reproduce experiment, testifies to a lack of information that can be directly extracted from B factor distributions. Thus, additional information must be supplied. For the dynamical contribution, this additional information is conveniently furnished by molecular simulation, in the form of the dynamical equations and the associated atomic model and force field. MD simulation offers a direct way of determining protein internal and whole-molecule motion [27, 136]. However, due to computational limitations only part of the dynamics of the system is explored *i.e.*, in the present simulations of *Staphylococcal* nuclease, fluctuations are sampled occurring on the nanosecond timescale and on lengthscales shorter than the box size of the simulation *i.e.*, the unit cell. That longer-timescale fluctuations exist, which are not sampled in the simulation, can be inferred from the present results. Furthermore, correlated motions between unit cells, such as commonly produce diffuse scattering streaks and rings associated with the reciprocal lattice, are also suppressed by the imposition of periodic boundary conditions.

Although in the simulations the fluctuations derived from the whole-molecule rigid-body external dynamics resemble the distribution of the total mean-square fluctuations their contribution to the

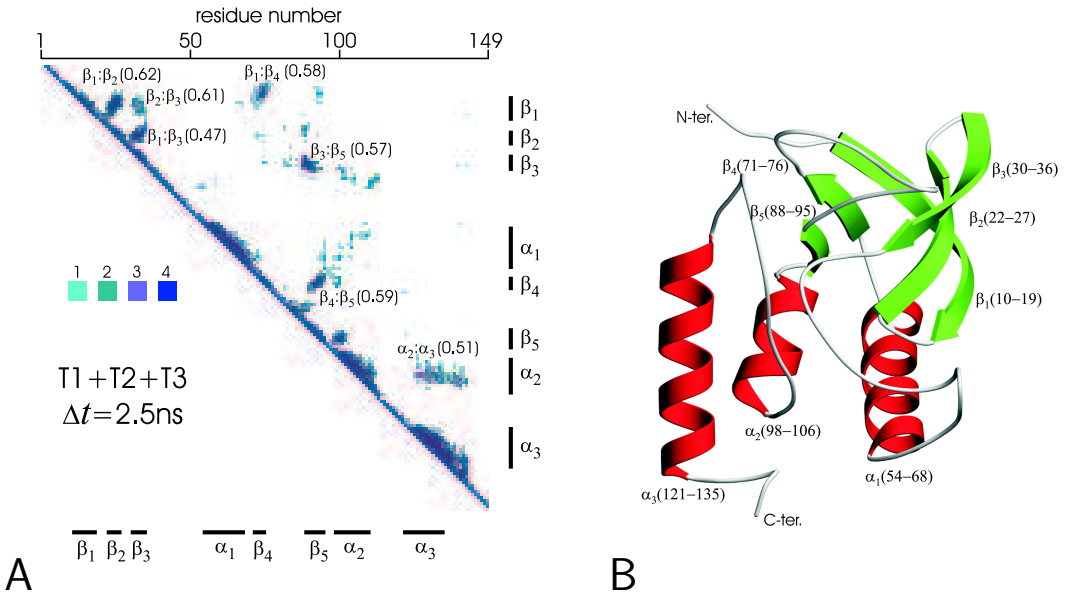

Figure 3.9: A: Pairs of C_α-atoms for which the protein-internal displacement correlations converged. The regions of secondary structural elements are indicated and their average correlation coefficients are given in parentheses. The results shown are for the merged trajectories Tx (x=1+2+3) with sub-trajectory length Δt=2.5 ns. The colour-scale indicates the number of proteins for which $C_{kk'}$ converged. B: Structure of SNase with secondary structural elements indicated. For convenience, the C- and N-termini are truncated, cf. Fig. 2.4 A. (Fig. B adapted from AD Gruia: *Dissertation, Heidelberg University, 2004.*)

total mean-square fluctuations is significantly smaller than that from internal motions. Internal and whole-molecule motions were found to contribute ≈0.74 Å^2 (69%) and ≈0.29 Å^2 (27%) to the total mean-square fluctuations, respectively. Only in the regions of the secondary structural elements are both contributions of comparable magnitude. Furthermore, the present results indicate that the separation of the fluctuations into whole-molecule translation, rotation and internal motion does not provide an adequate basis for describing B factors. Therefore, the TLS-model may not provide a physically consistent description if the protein is represented as a single rigid-body.

The simulation-derived total mean-square fluctuations are significantly larger than those derived from the experimental B factors, cf. Fig. 3.1, as has been observed in previous MD studies [136–138, 144]. Although part of the difference with experiment is due to the harmonic assumption made in deriving the experimental B factors, most of the difference has other origins. One possible contribution is the presence of crystallisation agents, ligands and/or large-size ions, *e.g.* sulfate, in the experiment (which are absent in the simulation) which, due to steric hindrance and/or salt-bridges, could restrict protein motion. Also, the simulation setup and performance (*e.g.* the building of unresolved residues, length of equilibration) or inaccuracies of the force field may have an influence. Furthermore, the uncertainties in experimental B factors may be large. For example, the C_α mean-square fluctuations derived from PDB entry 1SNM [159], a *Staphylococcal* nuclease structure similar to 1STN, are on average 30% larger than those shown in Fig. 3.1.

A large fraction (≈65%) of the fluctuations converges on the nanosecond timescale. The non-converged segments comprise the experimentally-unresolved C-terminus region (residues 142-

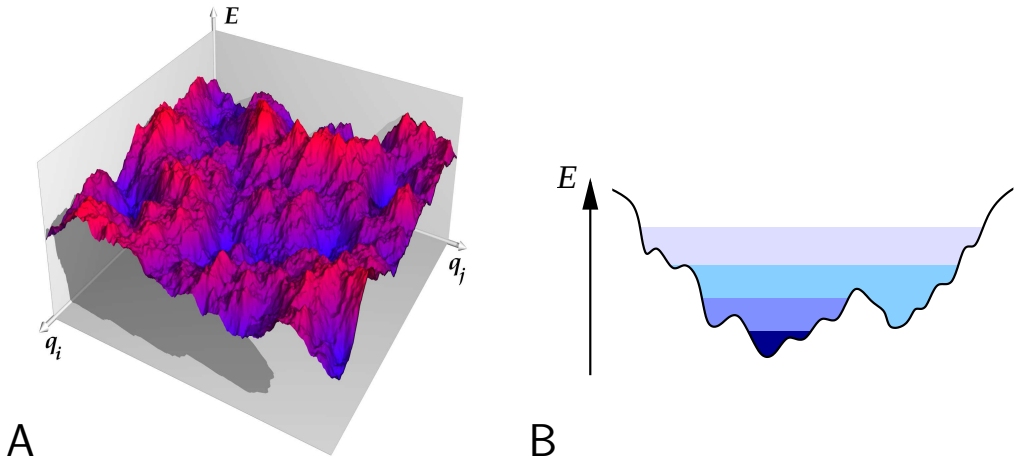

Figure 3.10: A: Graphical representation of the potential energy landscape. The generalised atomic coordinates are indicated and span the whole configuration space upon which the force field E, cf. Eqs. (2.1) and (2.2), is defined. E is hierarchically structured, possessing multiple minima separated by a distribution of barrier heights. B: The sampled fraction of the energy landscape increases with time, indicated by a colour gradient ranging from dark blue symbolising shorter timescales to light blue representing longer timescales. (Fig. A was kindly provided by N Calimet.)

149), a highly flexible loop (residues 46-51), and some coil regions for which the fluctuations vary non-monotonically with the simulation length.

Additionally, the convergence properties of the atomic displacement cross-correlation matrix, $C_{kk'}$ have been determined here as a function of the length of the simulations. In molecular simulations, collective properties often converge on longer timescales than single-particle properties, and poor sampling of the displacement correlation matrix in protein MD has been previously noted [99, 136]. Here, it is found that parts of the protein-internal inter-atomic displacement correlation matrix converge on the nanosecond timescale while other parts do not. A crucial factor is the atomic separation. For atoms separated by $\lesssim$7 Å more than 50% of $C_{kk'}$ converge on the nanosecond timescale in the simulations. The magnitude of the converged matrix elements decays on average approximately exponentially with distance with a decay length of 10.6 Å. Furthermore, convergence is seen between the strands comprising the β-barrel of the protein and between some neighbouring α-helices, but not between the α and β regions of the protein.

For intra-protein motion, the converged fraction of $C_{kk'}$ grows logarithmically with the simulation length, cf. Fig. 3.6, a finding consistent with the logarithmic decay of the diffuse scattering R factor with time (Fig. 3.3), which is determined by the convergence of the variance-covariance matrix. A physical model consistent with this logarithmic time dependence is as follows, cf. Fig. 3.10. For the variance-covariance matrix to converge the system must fully sample the accessible protein energy landscape. For proteins, the energy landscape consists of multiple minima separated by a distribution of barrier heights [8, 46]. The number, N_M of minima sampled is a measure of the fraction of the energy landscape sampled and thus the degree of convergence of the variance-covariance matrix. The configurational sampling within a single minimum is very fast, whereas transition rates between minima are slower and depend exponentially on the barrier height ΔE, occurring with an average transition time, $t_T \sim \exp(\Delta E/k_{\mathrm{B}}T)$. If the neighbouring minima are of higher energy than the original minimum the system will return to the original

minimum within a time shorter than t_T. Therefore, the sampling of neighbouring minima will converge on a timescale t_T given by the maximum barrier height crossed between the original and neighbouring minima. The observation that the variance-covariance matrix converges on a logarithmic timescale therefore implies that N_M increases proportionally to, or as a polynomial function of, the maximum barrier height crossed *i.e.*, $N_M \sim \Delta E \sim \log(t)$.

Finally, the present results allow to estimate how long an MD simulation might have to be in order to fully sample the collective motions within crystalline proteins. The protein collective motions are described by the displacement variance-covariance matrix, which in turn determines the X-ray diffuse scattering. Hence, the comparison with X-ray diffuse scattering experiments allows to estimate the simulation length required for the variance-covariance matrix to converge. Convergence of the diffuse scattering, and thus the variance-covariance matrix, can be defined as occurring when the agreement factor between simulation and experiment reaches $R_{\triangle t} \leq \epsilon$, where ϵ accounts for systematic and experimental errors. The required simulation length depends on the choice of ϵ and the system size. Estimates for the convergence time $\triangle t^*$ are given in Table 3.2 for the entire *Staphylococcal* nuclease unit cell and two ϵ-values. For a realistic value of $\epsilon = 0.04$ the convergence time is of the order $1\,\mu$s for the tetragonally-constrained simulations, *i.e.*, 100 times longer than the present simulations.

Chapter 4

Collective Dynamics

Intra-molecular protein dynamics is of potential functional importance and therefore has been the focus of intense research. Correlated intra-molecular displacements give rise to strongly delocalised features in the diffuse scattering pattern that are not associated with the reciprocal lattice. The variety of intra-molecular motions present in a protein renders an analytical description of the diffuse scattering cumbersome. However, molecular dynamics simulation provides a powerful means of describing the intra-molecular motions leading to diffuse scattering. For example, MD simulations of orthorhombic crystals of lysozyme have been shown to reproduce features of the scattering pattern [27, 80]. Furthermore, it was demonstrated that the motions of groups of rigid bodies, composed on average of five residues, also reproduce the prominent features of the scattering pattern [27]. However, it has also been demonstrated that calculated diffuse scattering does not converge for nanosecond MD simulations [99, 100], cf. Section 3.1.3.

A full analytical description of the diffuse scattering due to intra-molecular motions can be given using the atomic displacement variance-covariance matrix, cf. Section 2.3.2. Unfortunately, this is non-trivial due to the potentially high number of parameters (cross-correlations in displacements), which is of $\mathcal{O}(N^2)$ where N is the number of atoms in the unit cell and, therefore, the variance-covariance matrix must be approximated using models with significantly smaller numbers of parameters. The difficulty in unambiguously doing this becomes clear when noticing that, depending on the model used for analysis, in earlier research the protein motion determining diffuse scattering was contrastingly found to be either liquid-like (diffusive) or vibrational [76, 80].

However, a limitation of the early investigations was the rather incomplete experimental sampling in reciprocal space of the X-ray diffuse scattering. In this thesis, the use of the comprehensive, three-dimensional X-ray diffuse scattering map for SNase reported in Ref. [101], cf. Section 2.3.4, allows for qualitatively and quantitatively more precise tests of protein dynamical models.

In Chapter 3, MD trajectories of crystalline SNase were analysed in terms of the atomic fluctuations and displacement correlations and the convergence properties of these quantities were investigated. The inter-atomic displacement correlations are related to the variance-covariance matrix which determines the X-ray diffuse scattering. In this chapter, the simulation-derived X-ray diffuse scattering is analysed in detail. Firstly, the scattering from the whole unit cell

Results of this chapter will be published as *Meinhold & Smith*; submitted to Phys Rev Lett, cf. Ref. [160] and *Meinhold & Smith*; to be submitted to Biophys J, cf. Ref. [161].

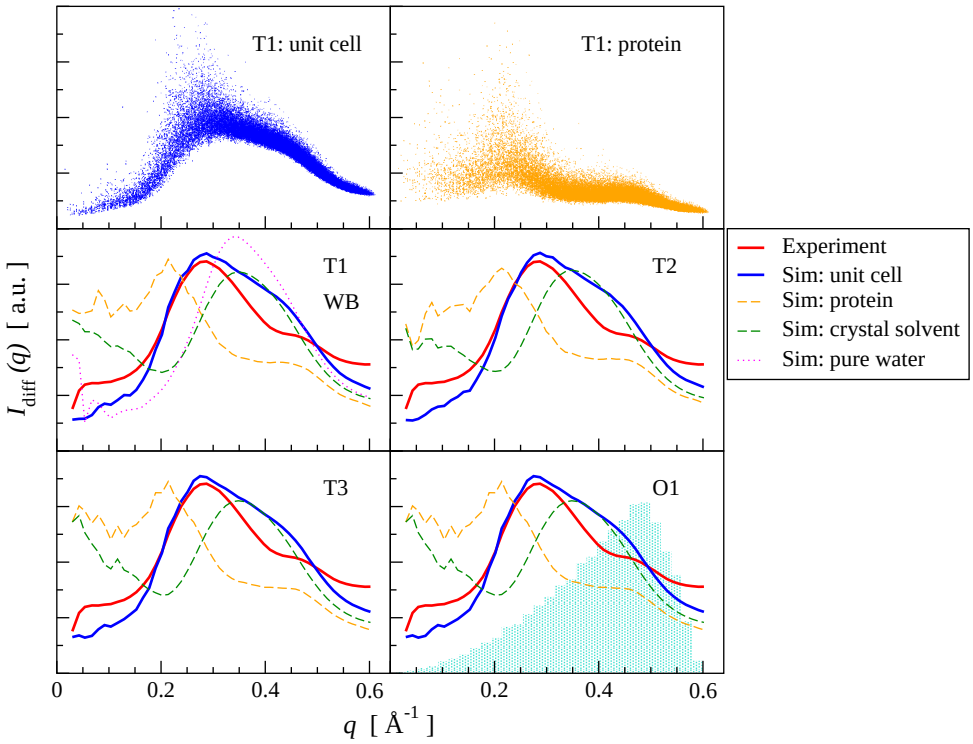

Figure 4.1: X-ray diffuse scattering intensities plotted against the magnitude, q of the scattering vector for all simulations of Set 1 (T1-3 and O1). The top panel shows the unit cell and protein scattering for all **q**-vectors for simulation T1. For clarity, in all other panels intensities are shown as isotropic averages. Shown is the simulation-derived scattering due to the complete unit cell (solid blue lines), the proteins only (orange dashed lines) and the crystal solvent only (green dashed lines) and the experimental scattering from Ref. [101] (red solid lines) which corresponds to scattering of the complete unit cell, *i.e.* proteins and crystal solvent. The scattering calculated from the pure-water simulation (WB) is also shown (magenta dotted line). The turquoise-shaded area shows the distribution of the 55,691 **q**-vectors.

is decomposed into protein and crystal solvent contributions. A qualitative interpretation of the diffuse scattering is presented using the structure factor of the electronic radial distribution functions calculated from the simulations. Then, the protein contribution is further characterised. In particular, two seemingly-contrasting models for the protein motion, one liquid-like and one collective, are compared. The collective motions are investigated using principal component analysis. The potential of mean-force associated with each principal component mode is analysed and the damping coefficient computed. Subsequently, individual three-dimensional features in the protein diffuse scattering are assigned to specific collective motions in the protein, and some of these explicitly involve potentially functional active-site deformations.

4.1 Decomposition of the Unit-Cell Diffuse Scattering

In Fig. 4.1 the isotropically-averaged X-ray diffuse scattering profiles, $I_{\mathrm{diff}}(q)$, computed from four MD simulations (T1-3 and O1) are shown. The experimental profile from the data reported in Ref. [101] is also shown along with the distribution of the scattering vectors. The unit cell

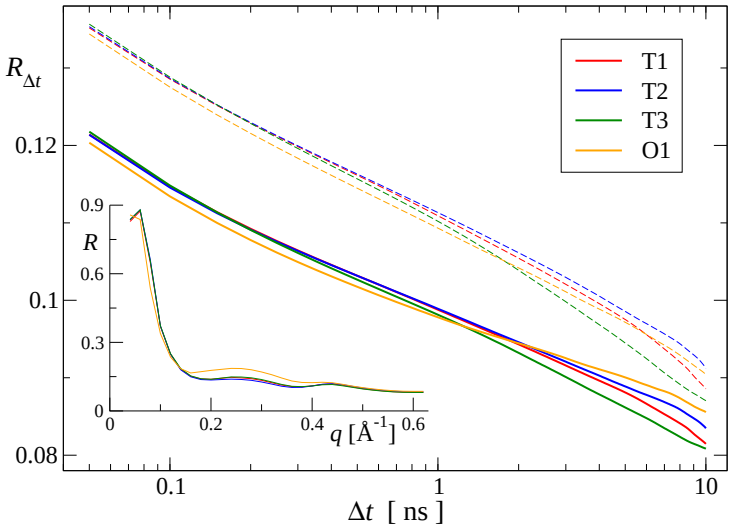

Figure 4.2: Agreement factor, $R_{\Delta t}$ between the simulated and the experimental diffuse scattering reported in Ref. [101] for the complete unit cell, *i.e.* proteins and crystal solvent, including all atoms (solid lines) or only the non-hydrogen atoms (dashed lines). The inset shows Luzzati plots of the R factors for the all-atom scattering and Δt=10 ns.

scattering qualitatively agrees with the experimental profile. In both experiment and simulation $I_{\rm diff}^{\rm protein}$ exhibits a pronounced peak at $q \approx 0.28$ Å^{-1}. At higher q, $I_{\rm diff}^{\rm protein}$ exhibits a shoulder at $q \approx 0.45$ Å^{-1}, which is of somewhat higher intensity in the simulation than in the experiment.

The dependence of the agreement factor, R on the simulation length, Δt is shown in Fig. 4.2 for all four simulations. The agreement with experiment improves with increasing Δt, as indicated by a continuous decrease of the R factor, which reaches values between 8.1% and 8.3% for the tetragonally-constrained simulations. For timescales below 1 ns the orthogonally-constrained simulation shows better agreement with experiment than the Tx-simulations (x=1,2,3), but for longer times its agreement is slightly poorer, due possibly to the asymmetric sampling of the unit-cell shape: instead of $a = b$ the average unit-cell edge parameters are $a = (49.9 \pm 0.7)$ Å and $b = (47.6 \pm 0.3)$ Å, cf. Table (3.1)

Fig. 4.2 also indicates the usefulness of including the scattering from the hydrogen atoms, which improves the R factor by ≈0.7%. Indeed, the hydrogens, due to their large abundance in protein crystals, contribute significantly to the electron density: ≈12% in the present system. Furthermore, hydrogen atoms smooth out the electron density in the unit cell. This has an effect on the fluctuations in the simulation-derived diffuse scattering, which are visible in the top-panel of Fig. 4.1, which shows the non-averaged diffuse scattering profiles. Adding hydrogens reduces the average variations in the diffuse scattering by ≈4% (Fig. 4.3), leading to an improved R factor. The inset to Fig. 4.2 shows that the R factors for all simulations on average improve with increasing resolution. The R factors slightly increase around $q \approx 0.2$ Å^{-1} and $q \approx 0.4$ Å^{-1}, consistent with the results of Fig. 4.1.

In contrast to experiment, in a simulation the scattering of the protein and crystal solvent components can be separately examined by setting the atomic form factors, f_k of the remaining components to zero. The results of this exercise are also shown in Fig. 4.1. It is instructive to compare the crystal solvent scattering, $I_{\rm diff}^{\rm solv}(q)$ with the scattering calculated from the simulation

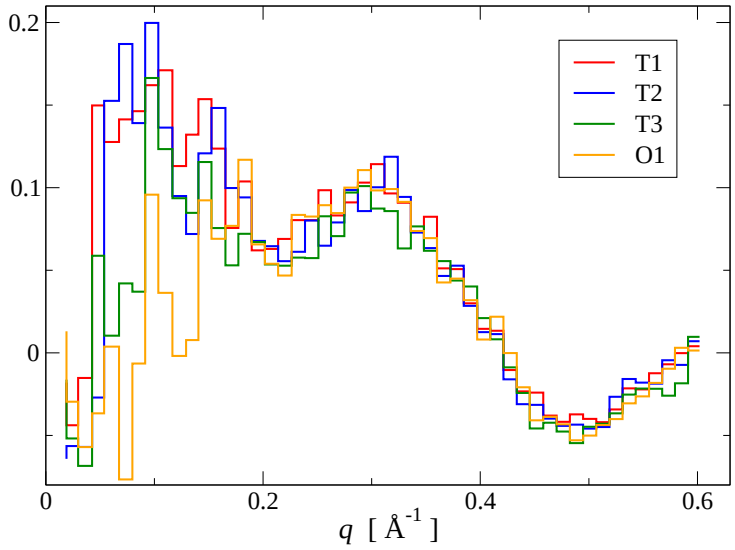

Figure 4.3: Change in the relative variations of the isotropically-averaged unit cell scattering due to the inclusion of hydrogens. The values shown are calculated as $(\triangle I_{\text{diff}}^{\text{nonH}} - \triangle I_{\text{diff}}^{\text{all}})/\triangle I_{\text{diff}}^{\text{all}}$, where $\triangle I_{\text{diff}} = \sigma_{I_{\text{diff}}}/\langle I_{\text{diff}}\rangle$ with σ denoting the standard deviation, and the superscripts `all` and `nonH` refer to the all atom and non-hydrogen atom scattering, respectively; all quantities are q-dependent. In the graph, the q-range was divided into 50 equi-sized bins.

of the pure water-box, $I_{\text{diff}}^{\text{WB}}(q)$. For $q \gtrsim 0.23\,\text{Å}^{-1}$ the two intensities are the same except for a scaling factor. However, for $q \lesssim 0.23\,\text{Å}^{-1}$ the q-dependence of the diffuse intensities is very different: $I_{\text{diff}}^{\text{WB}}(q)$ increases with q while $I_{\text{diff}}^{\text{solv}}(q)$ decreases. This can be understood by considering the scattering arising from a homogeneous disordered system. For such a system the scattered intensity is given by

$$I(q) \propto f^2(q)S(q), \tag{4.1}$$

where $f(q)$ is the atomic form factor and $S(q)$ is the structure factor which is given by

$$S(q) = 1 + n\int_0^\infty \mathrm{d}r\, 4\pi r^2 \frac{\sin(qr)}{qr}\,[g(r) - 1]\,, \tag{4.2}$$

where n is the average density and $g(r)$ is the radial distribution function of the system of scatterers [1]. $g(r)$ can be directly computed from the trajectories and $S(q)$ then calculated using Eq. (4.2). For the calculation of $g(r)$ here the contribution of each atom was weighted by its number of electrons to obtain an $S(q)$ corresponding to X-ray scattering. The resulting $S(q)$ for the trajectories T1 and WB are shown together with $g(r)$ in Fig. 4.4. Comparison with Fig. 4.1 shows that the diffuse scattering intensity profiles for the crystal solvent and pure-water diffuse scattering resemble $S(q)$. In particular, the peak location at $q = 0.35\,\text{Å}^{-1}$ is well reproduced.

The low-q part of the structure factor arises from long-range spatial modulations of $g(r)$, cf. the inset to Fig. 4.4. These originate from the inhomogeneity of the crystal solvent volumes within the unit cell, since the protein molecules act as voids in the crystal solvent scattering density. Therefore, the crystal solvent $g(r)$ exhibits variations on the length scale of the unit cell which give rise to increased low-q scattering intensity.

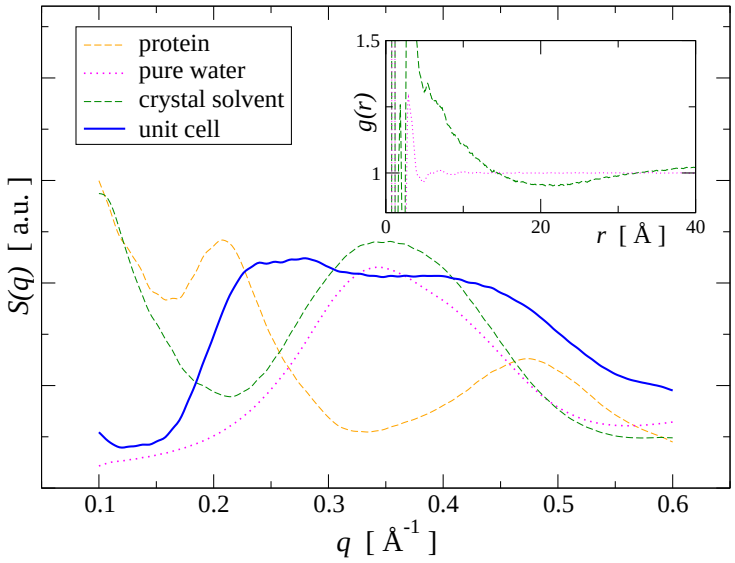

Figure 4.4: Structure factors, $S(q)$ for the unit cell, protein, crystal solvent and pure water scattering calculated from the radial distribution functions, $g(r)$ using Eq. (4.2). The unit cell, protein and crystal solvent scattering were computed from trajectory T1, the pure-water scattering from trajectory WB. The inset shows $g(r)$ for the crystal solvent and pure water.

Fig. 4.4 also shows that the protein structure factor is closely similar to the protein diffuse scattering intensities, $I_{\text{diff}}^{\text{protein}}$ shown in Fig. 4.1, and the peaks at $q = 0.21\,\text{Å}^{-1}$ and $q \approx 0.45\,\text{Å}^{-1}$ are present in both. Also shown in Fig. 4.4 is the unit cell structure factor. Although the unit cell $S(q)$ is similar to $I_{\text{diff}}(q)$, *i.e.*, the peak at $q = 0.28\,\text{Å}^{-1}$ is present in both, the shoulder at $q \approx 0.45\,\text{Å}^{-1}$ is more pronounced in $S(q)$. This is due to the atomic form factor, f_k which relates $I_{\text{diff}}^{\text{protein}}$ to $S(q)$, cf. Eq. (4.1). f_k decreases with increasing q with an approximately Gaussian dependence and, therefore, attenuates the high-q scattering intensity.

To investigate the origins of the peaks in $S(q)$, again the pure-water structure factor is considered first. $S^{\text{WB}}(q)$ exhibits a single peak at $q = 0.35\,\text{Å}^{-1}$ corresponding to a real-space distance of 2.86 Å which equals the radius of the first hydration shell of liquid water at 25°C [162, 163]. $g(r)$ for the pure water-box is shown in Fig. 4.5 and indeed possesses a broad peak centred at ≈ 2.8 Å. A smaller peak, located at 1.8 Å (non-bonded interaction O$\cdots$H), also contributes to $S(q)$ in the high-q range. Also, Fig. 4.5 shows that the short-range $g(r)$ of the crystal solvent is similar to that of pure water, differing by a scaling factor due to the different average densities.

Fig. 4.5 also shows $g(r)$ for the protein and the unit cell. For proteins, due to the presence of different covalent bonds, the short-range order is rather heterogeneous, giving rise to several peaks in $g(r)$ for small r which can be attributed to 1–3 covalent interactions ($r = 1.8$ Å: H-C-H, $r = 2.2$ Å: C-C-H, $r = 2.5$ Å: C-C-C, C-C=C, C-C=O) and lead to the broad peak in $S(q)$ at $q = 0.47\,\text{Å}^{-1}$, cf. Fig. 4.4. However, of particular interest is the peak located at $q = 0.21\,\text{Å}^{-1}$ (in the following denoted $\hat{q}$) because in this q-range the three-dimensional experimental diffuse scattering map exhibits detailed structure, cf. Figs. 4.1 and 4.15, which may provide information about collective protein motions present. This peak corresponds to real-space modulations of $g(r)$ on a length scale of ≈ 5 Å, which corresponds to the inter-strand distance in β-sheets and the pitch of an α-helix. To examine the origin of this peak $S(q)$ for the protein was recomputed

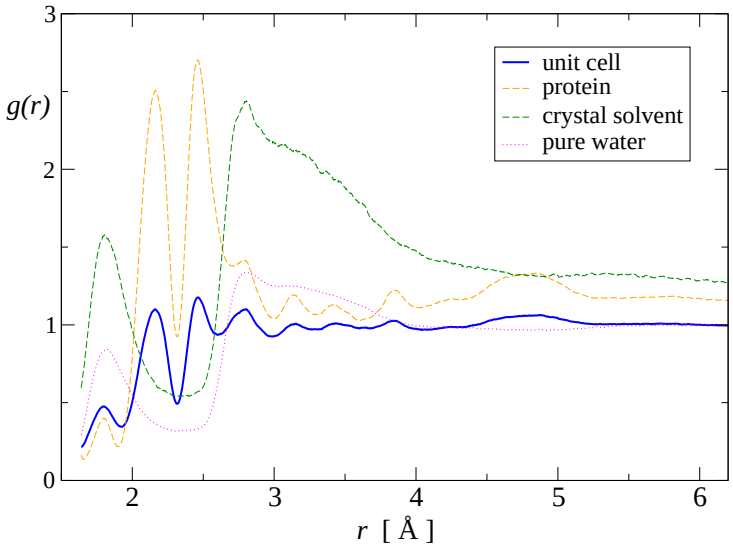

Figure 4.5: Radial distribution function, $g(r)$ for the unit cell, protein, crystal solvent and pure water calculated from the simulations T1 and WB.

with and without only the secondary structural elements. The results of this are shown in Fig. 4.6. If no secondary structural elements are included (Fig. 4.6 A) the $\hat{q}$-peak is absent. If either α-helices or β-strands are excluded, the $\hat{q}$-peak is present but of somewhat smaller intensity due to the increased low-q scattering (which is caused by the additional voids in the protein structure) which also causes a slight shift of the $\hat{q}$-peak towards lower q. $S(q)$ for only the secondary structural elements (Fig. 4.6 B) exhibits the same profile as the protein $S(q)$ but with a significantly increased $\hat{q}$-peak. Therefore, it can be concluded that the $\hat{q}$-peak indeed originates from the secondary structural elements. To assign it to the α-helix pitch and the inter-β-strand distance it remains to show that the peak is present for a single α-helix and a double β-strand, and absent for a single β-strand. That this is indeed the case is shown in Fig. 4.6 C.

To illustrate the relation between $S(q)$ and $I_{\text{diff}}^{\text{protein}}$in Fig. 4.6 the decomposition of the diffuse scattering $\hat{q}$-peak is also shown and re-confirms the above decomposition of $S(q)$. Differences arise from the heterogeneity of the system, *i.e.* from the different atomic form factors, and from dynamical disorder present, reflected in non-zero Debye-Waller factors, as is apparent from a term-wise comparison of Eqs. (2.71) and (4.2), bearing in mind Eq. (4.1).

The results presented in the preceding paragraphs can be utilised to decompose the diffuse scattering of the unit cell. In particular, Figs. 4.1 and 4.4 demonstrate that the pronounced peak at $q \approx 0.28\,\text{Å}^{-1}$, which is a general feature of protein X-ray scattering and is often referred to as the 'solvent ring' (more precisely, the 'solvent shell'), in fact originates from both the protein and the solvent scattering. The position of this peak depends on the relative intensities of the protein and solvent scattering which are determined by the ratio of protein to solvent electrons. In the system studied here, this ratio is 48:52 and therefore the protein and crystal solvent scattering contribute approximately equally to the total intensity. The peak of $I_{\text{diff}}(q)$ is therefore located almost exactly inbetween the peaks of $I_{\text{diff}}^{\text{protein}}$ and $I_{\text{diff}}^{\text{solvent}}$.

Fig. 4.1 also demonstrates that the diffuse scattering intensity of the protein cannot be easily obtained from the unit cell scattering by subtracting the scattering from a sample of pure solvent,

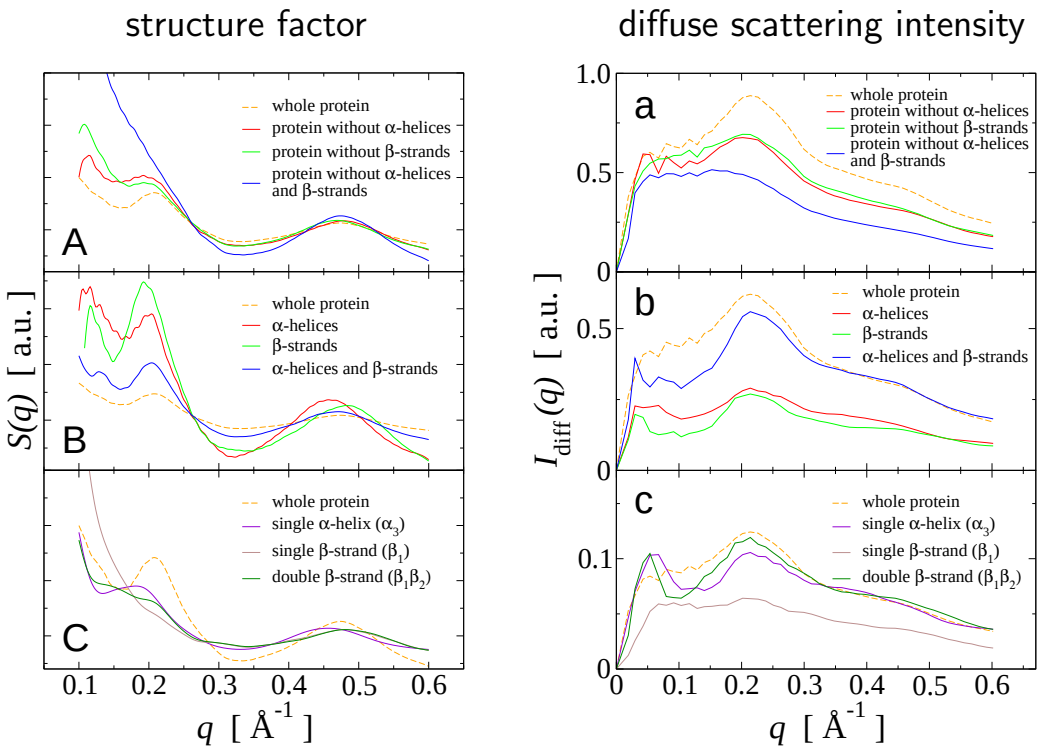

Figure 4.6: Left: Decomposition of the protein structure factor of trajectory T1. The protein $S(q)$ is compared with A: $S(q)$ of the protein without secondary structural elements, B: $S(q)$ of all secondary structural elements, and C: $S(q)$ of selected, single secondary structural elements. Right: Decomposition of the protein X-ray diffuse scattering intensity of trajectory T3. $I_{\mathrm{diff}}^{\mathrm{protein}}$ is compared with a: I_{diff} of the protein without secondary structural elements, b: I_{diff} of secondary structural elements, and c: I_{diff} of selected, single secondary structural elements. For clarity, $I_{\mathrm{diff}}^{\mathrm{protein}}$ has been scaled in b and c.

i.e., here pure water. For the isotropically averaged $I_{\mathrm{diff}}^{\mathrm{protein}}$ this subtraction is valid only in the large-q/small-r range ($q \gtrsim 0.23\,\mathrm{Å}^{-1}$) where the inhomogeneity of the protein and the crystal solvent volumes in the unit cell can be neglected. In the low-q/large-r range, *i.e.*, the range of potentially interesting collective motions, the inhomogeneity of the unit cell must be taken into account and, therefore, the subtraction of pure solvent scattering from the unit cell scattering does not yield the protein scattering.

4.2 Protein X-Ray Diffuse Scattering

Molecular dynamics simulation provides a convenient means to obtain the protein contribution to the diffuse scattering. If the scattering computed for the complete unit cell agrees reasonably well with experiment, then similar agreement is likely for the protein scattering. In this section the types of motion giving rise to the protein X-ray diffuse scattering are investigated. In particular, models of liquid-like [76] and collective [80] dynamics are discussed.

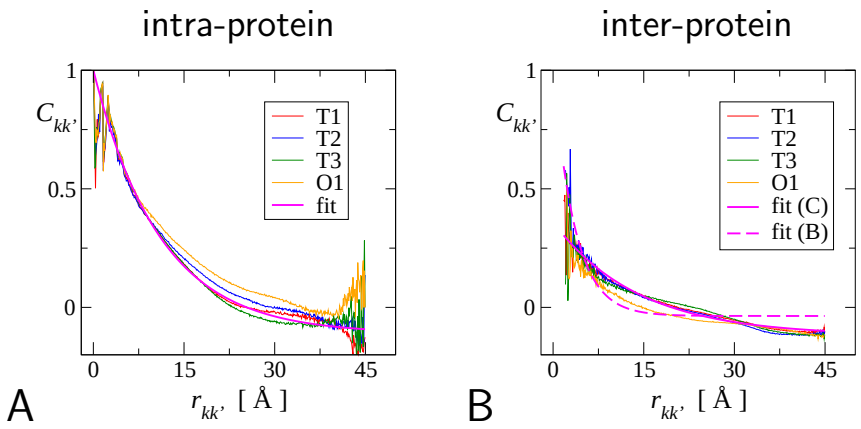

Figure 4.7: Simulation-derived average correlation coefficients $C_{kk'}$ (thin lines) plotted against the inter-atomic distance $r_{kk'}$ for intra-protein (A) and inter-protein (B) atom pairs. The magenta thick lines represent least-squares fits of $C(r_{kk'}; C_0, \lambda, \lambda_0) = C_0(1-\lambda_0)\exp(-r_{kk'}/\lambda) + \lambda_0$ to the data of simulation T1. In panel B the fits are shown for both Models B and C. The parameters C_0 and λ are explained in the text. λ_0 accounts for artefactual anticorrelations due to removal of the centre-of-mass translation and rotation of the unit cell, cf. Fig. 3.4 [100]. For $\lambda_0 = 0$ the equation fitted reduces to Eq. (4.4). For intra- and inter-protein correlations $\lambda_0 = -0.11$ and -0.12, respectively, consistent with the findings in Section 3.2. The values of C_0 and λ for all simulations and all models are given in Table 4.1.

4.2.1 Model of Liquid-Like Protein Motion

First, a model of isotropic, liquid-like motion (LLM) is studied. If the atomic displacements are isotropic the projection of $\mathbf{u}_k$ onto $\mathbf{q}$ is independent of the direction of $\mathbf{q}$ and the term $\mathbf{q}^T\langle \mathbf{u}_k \mathbf{u}_{k'}^T\rangle \mathbf{q}$ in Eq. (2.70) can be simplified to [164]

$$\mathbf{q}^T\langle \mathbf{u}_k \mathbf{u}_{k'}^T\rangle \mathbf{q} = q^2\sqrt{\langle \mathbf{u}_k^2\rangle \langle \mathbf{u}_{k'}^2\rangle}\; C_{kk'}, \tag{4.3}$$

where $C_{kk'}$ denotes the correlation coefficient between the displacements of atoms k and k', cf. Eq. (3.5), and $\langle \mathbf{u}_k^2\rangle$ is the mean-square fluctuation of atom k. To obtain a suitable description of the functional form of $C_{kk'}$ the simulation-derived average $C_{kk'}$ values are plotted against inter-atomic distance, $r_{kk'}$ in Fig. 4.7. This shows that, for both intra-protein and inter-protein atom pairs, $C_{kk'}$ decreases approximately exponentially with increasing $r_{kk'}$. However, Fig. 4.7 also shows that, on average, intra-protein motions are significantly more strongly correlated

Table 4.1: Correlation lengths λ and the scaling factor C_0 derived directly from the MD simulations by least-squares fits, see text and caption to Fig. 4.7.

	Model A	Model B		Model C		
	λ_p [Å]	λ_{int} [Å]	λ_{ext} [Å]	λ_{int} [Å]	λ_{ext} [Å]	C_0
T1	11.02	11.02	3.54	11.02	14.07	0.43
T2	10.54	10.54	4.35	10.54	12.53	0.49
T3	11.29	11.29	3.65	11.29	18.38	0.40
O1	10.87	10.87	3.13	10.87	11.04	0.38

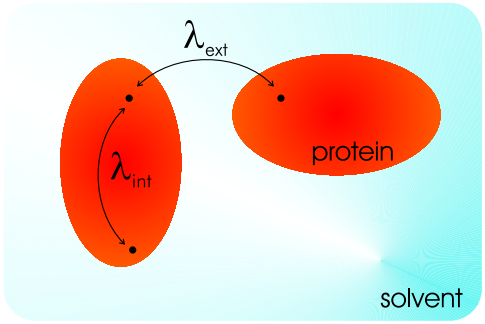

Figure 4.8: Graphical illustration of the liquid-like motion models, see text and Eq. (4.4). The correlation length, λ is different for intra-protein and inter-protein atom pairs. Atoms are symbolised by •, the solvent is not explicitly included in the model but is drawn for convenience.

than inter-protein motions. In particular, while for intra-protein motions $C_{kk'}(r_{kk'} \approx 0\,\text{Å}) \to 1$ for inter-protein motions $C_{kk'}(r_{kk'} \approx 0\,\text{Å}) \to C_0 < 1$. Therefore, the functional form of $C_{kk'}$ was chosen to be

$$C_{kk'}(r_{kk'}) = C_0 \exp\left(-\frac{r_{kk'}}{\lambda}\right), \tag{4.4}$$

with $r_{kk'} = \|\mathbf{r}_k - \mathbf{r}_{k'}\|$, λ being a correlation length and C_0 being a scaling factor for the magnitude of inter-protein correlations[1]. For intra-protein atom pairs $C_0 \equiv 1$. The values of $\langle \mathbf{u}_k^2 \rangle$ were taken from the simulations, although $\langle \mathbf{u}_k^2 \rangle$ could, in principle, also be obtained from X-ray crystallographic B factors.

Three different models based on Eq. (4.4) and illustrated in Fig. 4.8 are examined and the associated parameters directly derived from the simulations. The results are compared with the values of the same parameters obtained by fitting the model-derived X-ray diffuse scattering to that calculated directly from the simulation trajectories using Eq. (2.58) – a similar type of comparison was made in work determining the effect of anisotropy and anharmonicity on protein crystallographic refinement and R factors [165, 166]. The description of intra-protein correlations is the same in all three models, differences existing only in the treatment of the protein:protein correlations. Model A is a minimalistic description, in which only motions within the same protein are correlated and there are no correlations between atoms belonging to different proteins. In this case, the correlation length is $\lambda = \lambda_\text{p}$ if the atoms k and k' belong to the same protein and $\lambda \equiv 0$ otherwise. In Model B the description is made more realistic by using two different correlation lengths, $\lambda = \lambda_\text{int}$ for intra-protein correlations and $\lambda = \lambda_\text{ext}$ for atoms belonging to different proteins. Additionally, in this model the constraint $C_0 \equiv 1$ is set. In Model C λ_int and λ_ext are defined as in Model B but $C_0 \neq 1$ if the atoms belong to different proteins.

The parameters for Models A, B and C were derived directly from the simulation as described in the caption to Fig. 4.7, in which the fitted $C(r)$ functions are shown for the three models. The fitted parameters are given in Table 4.1. The intra-protein correlation lengths are approximately the same for all simulations, the average being (10.9 ± 0.3) Å. For the inter-protein correlation lengths Models B and C yield significantly different results, $\lambda_\text{ext}^\text{B} = (3.7 \pm 0.5)$ Å and $\lambda_\text{ext}^\text{C} = (14.0 \pm 3.2)$ Å. This is due to the fact that Model B does not adequately describe the inter-protein $C_{kk'}$-values, as can be clearly seen in Fig. 4.7 B – this is due to the scaling factor for

[1]The distance $r_{kk'}$ was calculated using the minimum-image convention introduced in the caption to Fig. 3.5.

the correlation strength (C_0). The average value of C_0 from all simulations is (0.43 ± 0.05), indicating that the short-range correlations between atoms belonging to different proteins are approximately 57% weaker than the short-range intra-protein correlations. For larger separations differences in the correlation lengths λ_{int} and λ_{ext} become important.

The parameters for Models A, B and C were also determined directly from the MD X-ray diffuse scattering patterns. To do this the following scheme was used. First, the protein diffuse scattering patterns were calculated directly from the MD trajectory using Eq. (2.58) and subsequently used as the reference in Eq. (3.4), in which the background, c, is in this case zero. The diffuse scattering patterns of the liquid-like models were calculated using Eq. (2.70) with $\mathbf{q}^T\langle\mathbf{u}_k\mathbf{u}_{k'}^T\rangle\mathbf{q}$ given by Eqs. (4.3) and (4.4), $\mathbf{r}_k$ and $\langle\mathbf{u}_k^2\rangle$ being taken from the simulation. The results are shown in Table 4.2.

The correlation lengths obtained from Model A, *i.e.* for intra-protein correlations only, are very similar for the tetragonally constrained simulations Tx (x=1,2,3), with the average being (14.4 ± 0.6) Å. The correlation length obtained from the simulation O1 is somewhat higher, being 17.4 Å. The R factors, representing the goodness-of-fit between the model and the simulation, are approximately 12% for all simulations. Including protein:protein interaction, *i.e.* Model B, the intra-protein correlation lengths (λ_{int}) do not change significantly. The inter-protein correlation lengths (λ_{ext}) are significantly smaller than λ_{int}, ranging from 8.1 Å to 9.5 Å, and show no difference between the simulations Tx and O1. Including the protein:protein interaction via Model B reduces the R factors for all simulations by ≈0.12% and therefore does not significantly improve the description of the diffuse scattering. This may be due to the fact that, for any given $r_{kk'}$, on average intra-protein correlations are larger than inter-protein correlations, cf. Fig. 4.7 and Section 3.2 [100]. This hypothesis was tested using Model C which scales the magnitude of the inter-protein correlations. Using this model, λ_{ext} increases, except for simulation T3, and is significantly different between the simulations, ranging from 7.5 Å to 19.4 Å. The magnitudes of the protein:protein correlations, C_0, decrease relative to intra-protein correlations by a factor between 0.38 and 0.89 for all simulations except T3, for which $C_0 = 1.4$. The average reduction in R factor from Model B to C is smaller than 0.01% for the simulations Tx and is 0.045% for the simulation O1.

The correlation lengths λ and the scaling factor C_0 derived from the LLM model fits to the simulation-derived diffuse scattering patterns can now be compared to the values derived directly

Table 4.2: Correlation lengths λ and the scaling factor C_0 for the models of liquid-like motion, cf. Eq. (4.4). To obtain optimal model parameters, *i.e.* parameters $\{\lambda, C_0\}$ that optimally reproduce the diffuse scattering calculated directly from the simulation using Eq. (2.58), a conjugate-gradient minimisation method [111] was used to optimise the agreement factors R between the model-derived and simulation-derived diffuse scattering patterns. The model parameters were calculated on an adaptive grid with a relative accuracy of $4\cdot10^{-4}$ or better.

	Model A		Model B			Model C			
	λ_{p} [Å]	R [%]	λ_{int} [Å]	λ_{ext} [Å]	R [%]	λ_{int} [Å]	λ_{ext} [Å]	C_0	R [%]
T1	13.80	11.94	13.75	8.11	11.85	13.89	11.06	0.62	11.84
T2	14.90	12.04	14.84	9.29	11.88	14.87	9.99	0.89	11.88
T3	14.56	12.06	14.76	9.56	11.91	14.52	7.54	1.43	11.90
O1	17.35	11.60	17.35	9.44	11.50	17.55	19.43	0.38	11.45

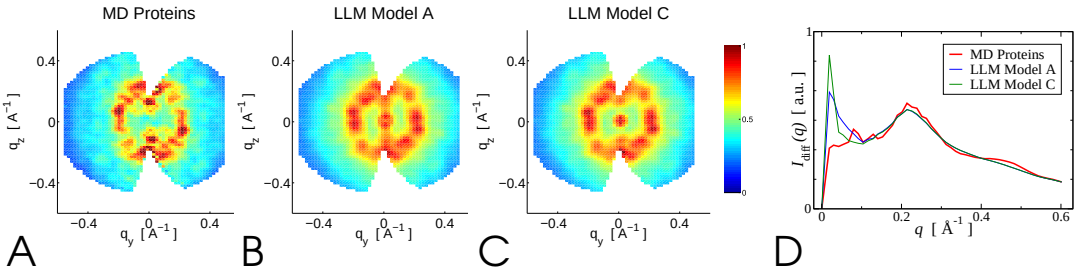

Figure 4.9: Comparison of the LLM model scattering with that calculated directly from the simulation T1. A–C: Cross-sections through the $q_x = 0\,\text{Å}^{-1}$ plane of the three-dimensional protein X-ray diffuse intensities (in [a.u.] on the same scale in each plot) due to the full trajectory (A) and calculated from LLM Model A (graph B) and Model C (graph C). D: Isotropic averages of the complete diffuse map corresponding to A–C.

from the simulations. For all simulations the simulation-derived intra-protein correlation length is significantly shorter than that derived from the model fits. Moreover, the correlation lengths for the inter-protein motions obtained directly from the simulations are also significantly smaller (Model B) or larger (Model C, except simulation O1) than the values obtained from the LLM model fits. Furthermore, the scaling factor C_0 is, on average, overestimated by the LLM model fits. Also, the R factor results for the LLM models show that the description of the protein X-ray diffuse scattering does not significantly improve from Model B to C, although one additional parameter (C_0) was introduced. Together with the small improvement of the R factors from Model A to B these results indicate that the LLM model in the given form does not yield a satisfying description of the diffuse scattering.

In Fig. 4.9 cross-sections of the LLM model-derived diffuse scattering are shown and compared with the scattering pattern derived directly from the simulation. It can be seen that, although the LLM model reproduces the location of the intense protein scattering shell at $q \approx \hat{q}$ (Fig. 4.9 D), the detailed fine structure present in the MD-derived $I_{\text{diff}}^{\text{protein}}$ is not present in the LLM model scattering, cf. Section 4.2.4. In particular, the low-q scattering is significantly overestimated by the LLM model scattering, with the results of Model C being slightly better, *i.e.* less low-q intensity, than those of Model A. Furthermore, the shoulder present in the MD scattering at $q \approx 0.45\,\text{Å}^{-1}$ is absent in the scattering derived using the LLM model, indicating that the small-scale local dynamics cannot be adequately described using only $\langle \mathbf{u}_k^2 \rangle$ (B factors) and an average correlation length.

The disagreement between the LLM models and experiment, *i.e.* MD simulation, might be due to the isotropic assumption made and to the fact that the LLM model neglects long-range, anisotropic correlations, due, for example, to the covalent bonding structure. Furthermore, since the atomic dynamics are included only via $\langle \mathbf{u}_k^2 \rangle$, the LLM model cannot fully describe collective motions because the $\langle \mathbf{u}_k^2 \rangle$ do not provide any information about the relative phases between atomic displacements. This phase information is, however, contained in the principal components of the protein motion.

4.2.2 Model of Vibrational Protein Motion

In this section, a model of protein motion based on principal component analysis is investigated. In particular, the dependence of the diffuse scattering pattern on the number of included principal component modes is determined. Of particular interest is the contribution of the large-amplitude modes, which potentially are related to protein function.

The analysis utilises the following definitions and considerations.

Principal Component Analysis (PCA)

PCA is a method of determining the collective motions present in an MD trajectory [26, 167, 168]. In principle, in PCA the variance-covariance matrix is diagonalised. However, for a physically meaningful description $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ must be mass-weighted prior to diagonalisation, *i.e.*, $V = M^{1/2} \langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle M^{1/2}$ where M is the diagonal matrix with $M_{kk} = m_k$ being the mass of the k^{th} atom. Diagonalisation of V yields the eigenvectors $\mathbf{v}_m$ and their associated eigenvalues λ_m, which are related to the eigenfrequencies by $\nu_m = \left(k_\mathrm{B}T/\lambda_m\right)^{1/2}$, where k_B is the Boltzmann constant. The displacements, $\mathbf{u}_k$ can then be expressed as a sum over collective variables,

$$\mathbf{u}_k = \frac{1}{\sqrt{m_k}} \sum_m \sigma_m \mathbf{v}_{mk}, \tag{4.5}$$

where the projection σ_m is defined as the scalar product

$$\sigma_m = \left(M^{1/2} \mathbf{u}^T \right) \cdot \mathbf{v}_m. \tag{4.6}$$

Note that the σ_m, and thus $\mathbf{u}_k$, depend on the displacements $\mathbf{u}$ of all atoms. The dimension of σ_m is $\texttt{mass}\times\texttt{distance}^2$. The functional form of σ_m is *a priori* unknown. In particular, for low-frequency PCA modes σ_m is anharmonic and possibly correlated with $\sigma_{n\neq m}$. To calculate the scattering due to the PCA-mode interval $[\mathrm{n_s}, \mathrm{n_s} + N_{\text{modes}} - 1]$ for a given trajectory a new trajectory was generated by applying Eqs. (4.5) and (4.6) for each time step. From this new trajectory $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ was calculated and the diffuse scattering was then computed using Eq. (2.70).

The above definitions and considerations are now applied to the trajectories T1-3 and O1, and the results are compared to those of the LLM model.

PCA was performed for all 4732 non-hydrogen protein atoms, giving rise to 14190 vibrational and 6 translational-rotational modes. From the distribution of the PCA eigenfrequencies, ν_m the vibrational densities of states, $g(\nu)$ were calculated for all simulations and are shown in Fig. 4.10. $g(\nu)$ shows a pronounced peak around $\nu = 1.4\,\mathrm{THz}$, which is in accord with previous theoretical studies [169–171] and corresponds to motions on the picosecond timescale ($1\,\mathrm{THz}{=}33.36\,\mathrm{cm}^{-1}$). In all simulations the 15 lowest-frequency modes have frequencies below 0.1 THz and account for ≈80% of the fluctuations present. $g(\nu)$ has a high-frequency cutoff at ≈52 THz corresponding to the maximal frequency of non-hydrogen atom bond vibrations. Several smaller peaks can be assigned to local vibrations in side chains (out-of-plane bending motions at ≈20 THz and ≈25 THz) and to localised bond-stretching vibrations (single-bond vibrations at ≈37 THz and double-bond vibrations at ≈42 THz). A more detailed characterisation of the PCA modes is given later in Section 4.2.3. Recent experiments on globular proteins observed a similar profile of $g(\nu)$ but with the maximum located at ≈2.2 THz [172, 173] or ≈3.7 THz [174].

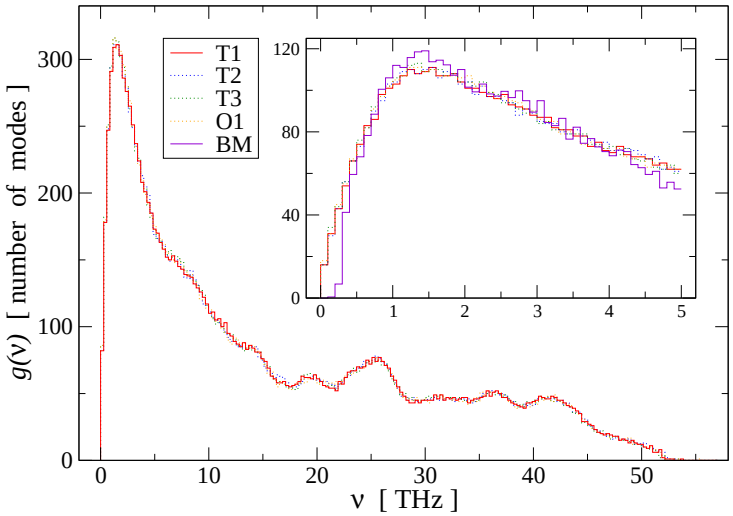

Figure 4.10: Vibrational density of states, $g(\nu)$ directly obtained from PCA for the simulations T1-3 and O1. $g(\nu)$ is almost the same for all simulations; for clarity that for trajectory T1 is highlighted (solid line). The inset shows the low-frequency regime in more detail. Also shown in the inset is the average $g(\nu)$ for the Brownian motion (BM) model given by Eqs. (4.9), cf. Section 4.2.3. Note the conversion factor, 1 THz=33.36 cm^{-1}.

X-ray diffuse scattering due to the PCA modes in the interval $[n_\mathrm{s}, n_\mathrm{s}+N_\mathrm{modes}-1]$ was computed as described above. The dependence of the R factor, calculated with the full simulation non-hydrogen protein scattering as the reference, on the number of PCA modes included is shown for trajectory T1 with $\mathrm{n_s} = 1, \ldots, 10$ and $N_\mathrm{modes} = 1, \ldots, 7000$ in Fig. 4.11. For $\mathrm{n_s} = 1$, *i.e.* for the N_modes lowest-frequency modes, the R factor rapidly reduces, to 10%, when including the twelve lowest-frequency modes and then continues to decrease at a slower rate when more modes are included, reaching 1% for the first 1000 modes. At $N_\mathrm{modes} = 7000$ the R factor is essentially zero. Thus, the scattering is dominated by the lowest-frequency PCA modes, particularly the first ≈10. Therefore, the influence of the first ten modes was further investigated. To do this the scattering was computed for the same number of modes, N_modes but with $\mathrm{n_s} = 2, \ldots, 10$, *i.e.* with the $(\mathrm{n_s}-1)$ lowest-frequency modes excluded. These results are also shown in Fig. 4.11. For all $\mathrm{n_s}$ the dependence of the R factor on N_modes is similar to the $\mathrm{n_s}$=1 curve but the change of slope in R versus N_modes at $N_\mathrm{modes} \approx 10$ is less pronounced, and the plot is shifted to larger values of N_modes. The R factor curves for different $\mathrm{n_s}$ run approximately parallel over the full N_modes range. In particular, for $\mathrm{n_s} > 1$ the R factor does not converge to zero but to a value $R^\mathrm{min}_\mathrm{n_s} > 0$ which increases with increasing $\mathrm{n_s}$. $R^\mathrm{min}_2 = 0.05$ shows that the first PCA mode accounts for 5% of the converged R factor. An additional 5% is contributed by the modes $2, \ldots, 9$ as indicated by $R^\mathrm{min}_{10} = 0.1$.

4.2.3 Principal Components of the Protein Motion

The R factor results of Fig. 4.11 suggest that the motions along protein PCA modes can be successfully used to describe the protein diffuse scattering pattern and thus the collective motions present. Therefore, it is of interest to investigate in more detail the type of motions described by these PCA modes. In this section it is determined which modes are harmonic and then, for the harmonic modes, the damping characteristics are determined.

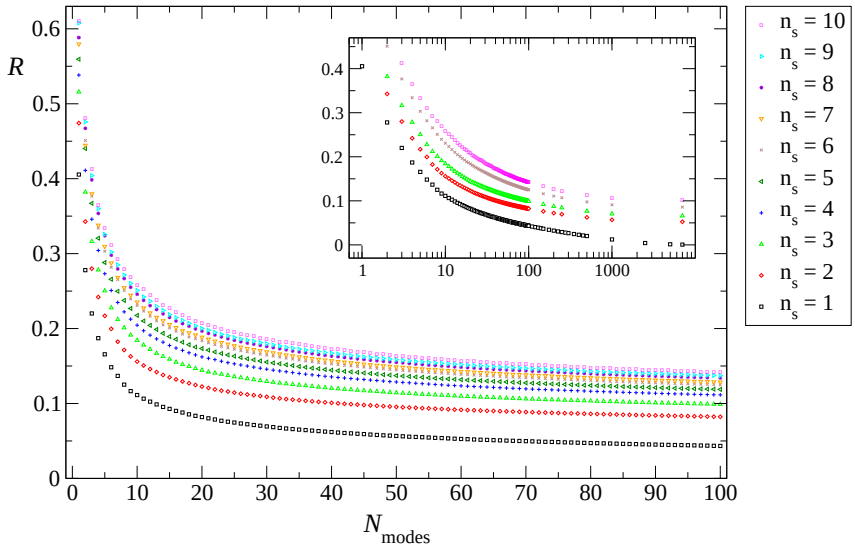

Figure 4.11: Dependence of the agreement factor R on the number N_{modes} of included PCA modes and the first included mode, n_s, *i.e.*, a data point corresponds to the scattering due to a block of PCA modes starting with mode n_s and having N_{modes} modes. The inset presents the results on a larger range of N_{modes} on a logarithmic horizontal scale. Results shown are for trajectory T1 and the reference pattern comprised 5,000 **q**-vectors randomly chosen from the full q-range.

The analysis utilises the following definitions and considerations.

Effective Free Energy Landscape of PCA Modes

The dynamics along the m^{th} PCA mode was investigated by analysing the time series $\sigma_m(t)$, cf. Eq. (4.6). The effective free energy along the mode is given by

$$G_m(\sigma) = -k_\text{B}T \ln p_{\sigma_m}, \tag{4.7}$$

where $p_{\sigma_m}\text{d}\sigma$ is the probability that $\sigma_m(t)$ adopts a value in the interval $[\sigma, \sigma + \text{d}\sigma)$. If the motion is harmonic, then $G_m(\sigma)$ is also harmonic and the probability density, p_{σ_m} is Gaussian. However, the largest-amplitude PCA modes significantly deviate from harmonicity. A measure of the anharmonicity is

$$F_{\text{AH}} = \int_{-\infty}^{+\infty} \text{d}\sigma \left(p_{\sigma_m} - p_{\sigma_m}^{\text{GF}}\right)^2, \tag{4.8}$$

where $p_{\sigma_m}^{\text{GF}}$ is a Gaussian fit to p_{σ_m}. Similar measures for anharmonicity have been used by other authors [52, 167]. However, in Ref. [52] the integral in Eq. (4.8) was calculated only on an interval of three standard deviations around the mean of σ_m.

Brownian Motion along PCA Modes

If $G_m(\sigma)$ is approximately harmonic $\sigma_m(t)$ can be approximated as Brownian motion in a harmonic potential using the set of Langevin equations,

$$\dot{\sigma} = \upsilon, \quad \dot{\upsilon} = -\Gamma\upsilon - \omega_0^2\sigma + \xi, \tag{4.9}$$

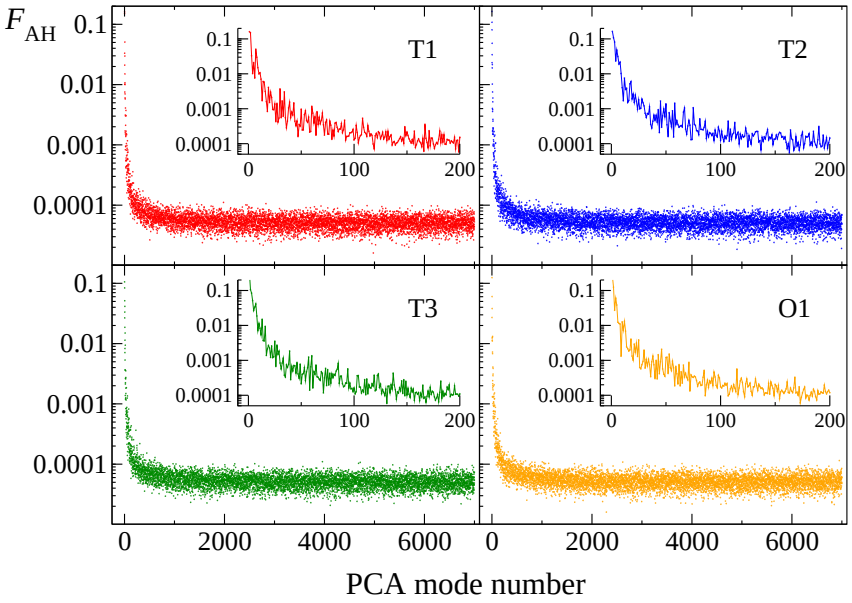

Figure 4.12: Anharmonicity factor, F_{AH} as defined in Eq. (4.8) plotted against the PCA mode number. Note the logarithmic y-scale. Simulation names are indicated and the insets show F_{AH} for the low-frequency modes.

where Γ is the damping frequency, ω_0 the eigenfrequency and ξ is a random force given by a Gaussian white noise process. Γ and ξ are related by the fluctuation-dissipation theorem. The power spectral density (PSD) S_{vv} can be calculated analytically [132],

$$S_{vv}(\omega) = \frac{4\Gamma k_{\mathrm{B}} T \omega^2}{(\omega_0^2 - 2\delta\omega + \omega^2)(\omega_0^2 + 2\delta\omega + \omega^2)}, \tag{4.10}$$

$$\delta = \sqrt{\omega_0^2 - \Gamma^2}, \tag{4.11}$$

and can be compared to the PSD $S_{\sigma\sigma} = \omega^{-2} S_{vv}$ obtained from the simulation. $S_{\sigma\sigma}$ was computed from $\sigma_m(t)$ using the fast Fourier transform routine `fft` of MATLAB [175]. Subsequently the non-linear fitting routine `lsqcurvefit` of MATLAB was used to fit the analytical S_{vv} to the simulation-derived $\omega^2 S_{\sigma\sigma}$ to determine the parameters ω_0 and Γ in Eqs. (4.9).

The above definitions and considerations are now applied to the trajectories T1-3 and O1.

The mode harmonicity was investigated by calculating the anharmonicity factor, F_{AH} given in Eq. (4.8). The results for all simulations are shown in Fig. 4.12. F_{AH} on average rapidly reduces with increasing mode number over the first 500 modes, and then remains approximately constant, fluctuating below 10^{-4}. The effective free energy profiles, G_m of the PCA modes were calculated using Eq. (4.7) and are displayed in Fig. 4.13 for the mode numbers m=1, 10, 50 and 500. Fig. 4.13 A demonstrates that the first, $m = 1$ PCA mode of each trajectory is strongly anharmonic, possessing multiple minima separated by barriers of $2-11\,\mathrm{kJ\,mol^{-1}}$, corresponding to approximately $1-4\,k_{\mathrm{B}}T_{300\,\mathrm{K}}$. Furthermore, the inset to Fig. 4.13 A shows that the first PCA mode is not vibrational but describes a drift of the system through the energy landscape. With increasing m the modes initially remain anharmonic, but the number of minima reduces until only a single minimum remains for $m \gtrsim 5$. Examples are shown in Fig. 4.13 B for $m = 10$. G_{10} possesses a flat energy landscape over a large range of σ. $\sigma_{10}(t)$ for trajectory T1 is shown in the

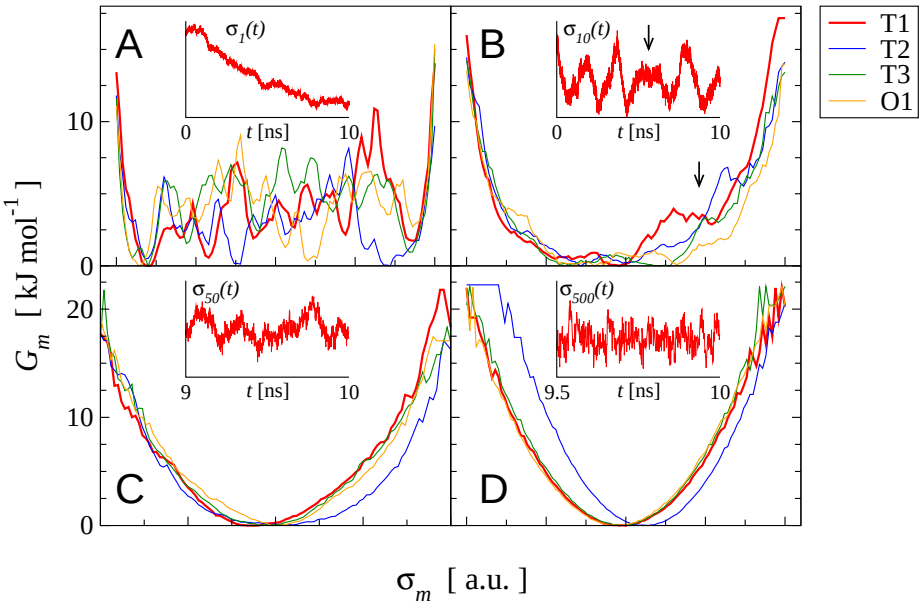

Figure 4.13: Effective free energy profiles, G_m plotted against the projection onto each mode, σ_m. Values of G_m are shifted by a constant such that $\min\{G_m\} = 0\,\mathrm{kJ\,mol^{-1}}$ for each mode. Note that $k_\mathrm{B}T_{300\,\mathrm{K}} = 2.5\,\mathrm{kJ\,mol^{-1}}$. Results are shown for PCA modes 1 (A), 10 (B), 50 (C) and 500 (D) for trajectories T1-3 and O1. The insets show the timeseries $\sigma_m(t)$ ([a.u.]) for trajectory T1 for the full trajectory (A and B), and for the last 1 ns (C) and last 0.5 ns (D) of the trajectory, respectively.

inset to Fig. 4.13 B. In the flat part of G_{10} the system seemingly freely propagates as indicated by the segments of linear slope of $\sigma_{10}(t)$. However, at the boundaries of the potential (*i.e.*, at small and large values of σ) the motion is reflected, as indicated by the inversion of the slope between the linear segments of $\sigma_{10}(t)$. Also, for a 1.4 ns period of the simulation, $\sigma_{10}(t)$ is restricted to a shallow local minimum which is indicated by the arrows in Fig. 4.13 B. In Figs. 4.13 C and D it can be seen that the energy profiles of the higher-frequency PCA modes tend towards more harmonic potentials. In the range $m \approx 30, \ldots, 100$ the PCA modes possess a defined minimum but significantly deviate from harmonicity: these modes are 'quasi-harmonic' [29, 167]. For $m \gtrsim 100$ G_m becomes close to harmonic.

From Figs. 4.12 and 4.13 it can be concluded that, except for the first ≈100 modes, all PCA modes describe motion in a harmonic potential. The motion of the harmonic modes, $\sigma_{m\gtrsim 100}$ is qualitatively closely similar to the motion of $\sigma_{500}(t)$ shown in the inset to Fig. 4.13 D, resembling Brownian motion in a harmonic potential. This type of motion can be well described using Langevin equations, Eq. (4.9). To determine the eigen- and damping frequencies, ω_0 and Γ, respectively, for a given mode the theoretical power spectral density, calculated using Eq. (4.10), was fitted to the simulation-derived power spectral density as explained above. The results are shown in Fig. 4.14. The eigenfrequencies, ω_0 increase with increasing mode number and resemble the $g(\nu)$ obtained directly from PCA, cf. the inset to Fig. 4.10. The damping frequencies decrease for the first ≈250 modes and then slowly increase with increasing mode number. For the low-frequency modes $\omega_0 < \Gamma$ and therefore the motions are overdamped. Although many of the modes in this range are not harmonic ($m \lesssim 100$) and therefore Eq. (4.10) will not strictly hold, estimates for the quasi-harmonic modes ($m = 30, \ldots, 100$) should be reasonable. At around mode number $m \simeq 175$ ω_0 and Γ have approximately the same value and the motions in this

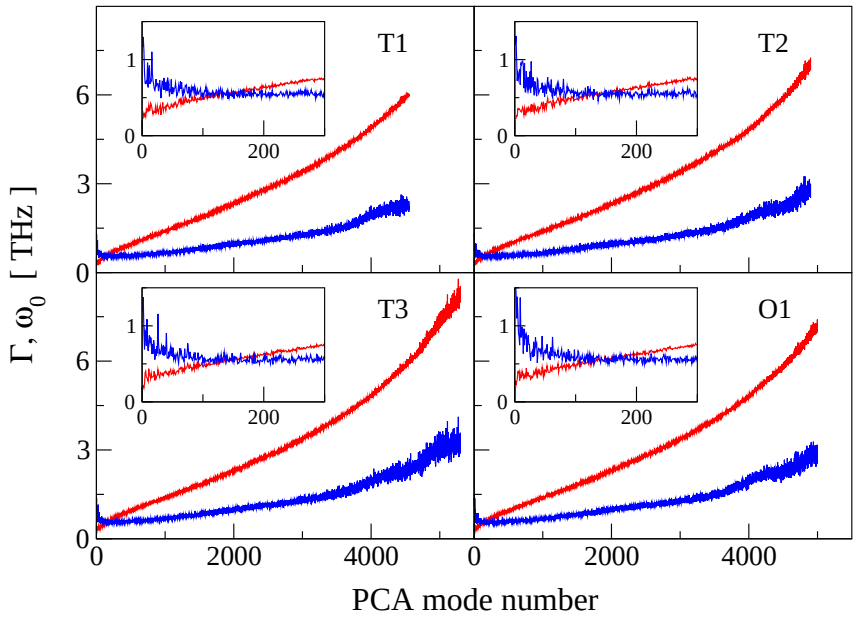

Figure 4.14: Eigenfrequencies ω_0 (red) and damping frequencies Γ (blue) for PCA modes of the simulations T1-3 and O1 obtained from fitting the theoretical PSD given by Eq. (4.10) to the PSD calculated for each mode. The insets show the low-frequency regime in more detail.

range, with frequencies around 0.6 THz, are critically damped. This result is in good agreement with previous theoretical studies on single protein molecules in which overdamped motions have been found for modes with frequencies below 0.7 THz [174, 176]. For higher-frequency PCA modes the eigenfrequencies are significantly larger than the damping frequencies and, therefore, these modes exhibit underdamped vibrations.

4.2.4 Identifying Specific Collective Motions

In this section, three-dimensional (3D) features in the protein diffuse scattering are examined. In Fig. 4.15 A is shown a cross-section through the 3D protein scattering pattern. The high average intensity of the protein shell at scattering vectors $\|\mathbf{q}\| \approx \hat{q}$ is again evident. However, in addition, intense 3D features in $I_{\text{diff}}^{\text{protein}}$ are seen, superposed on the $q \approx \hat{q}$ shell.

To determine which collective motions present in the MD trajectories contribute to the 3D features in $I_{\text{diff}}^{\text{protein}}$ again use was made of principal component analysis and the diffuse scattering was computed as described in Section 4.2.3. The comparison between Fig. 4.15 A and B shows that the intense 3D fine structure in $I_{\text{diff}}^{\text{protein}}$ is dominated by the five largest-amplitude PCA modes. In contrast, the PCA modes 6–7000 yield a less structured, smoother and more delocalised scattering pattern (Fig. 4.15 C). Therefore, it can be concluded that the motions giving rise to the most intense features in $I_{\text{diff}}^{\text{protein}}$ are captured by a very small number of PCA modes.

Visual examination showed that the five lowest-frequency PCA modes all describe collective dynamics involving both inter- and intra-protein motions. In all simulations performed and for all four proteins the intra-protein motion due to the first mode was similar in form and was dominated by a combined screw and hinge-like relative motion of two domains of the protein,

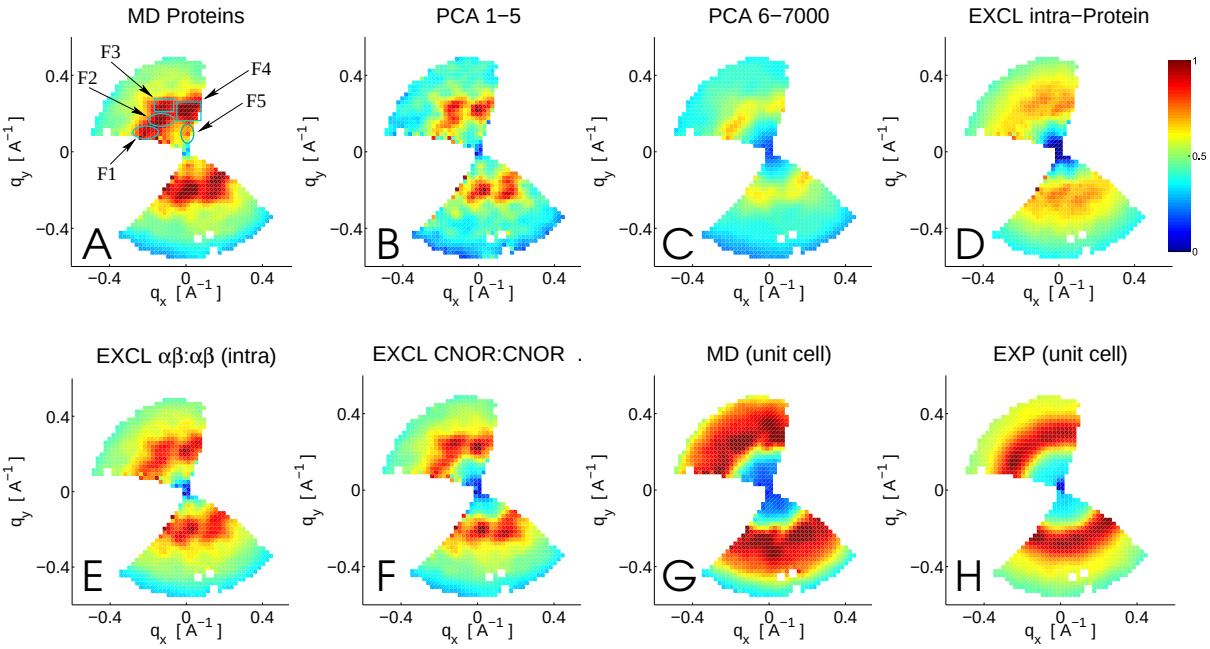

Figure 4.15: A–F: Cross-sections through the $q_z = 0\,\text{Å}^{-1}$ plane of the simulation-derived three-dimensional non-hydrogen atom protein X-ray diffuse intensities (in [a.u.] on the same scale in each plot) due to the full trajectory (A), due to the PCA modes 1–5 (B) and 6–7000 (C), and due to the full trajectory but excluding all intra-protein cross-correlations (D), or excluding only those intra-protein cross-correlations between the α and β secondary structural elements (E), or between the C and N termini and the O and R loops (F). G–H: $q_z = 0\,\text{Å}^{-1}$ cross-sections through the all-atom unit cell scattering of the simulation (F) and experiment (H) on arbitrary intensity scales. All patterns contain inversion symmetry, consistent with the space-group symmetry. Intense 3D features are labelled in A and discussed in the text. The data for graphs A–G are taken from trajectory T3 and that for graph H from Ref. [101].

one containing the β-barrel and α_1-helix, and the other being formed by the α_3-helix and major loops – the α_2-helix is the hinge fulcrum and screw axis, cf. Fig. 3.9 B.

The experimental X-ray diffuse scattering, Fig. 4.15 H, is somewhat smoother than the all-atom unit cell simulation scattering, Fig. 4.15 G. This is consistent with the finding that the 3D fine structure in the scattering is dominated by the five largest-amplitude, slowly-varying, unconverged PCA modes, cf. Figs. 3.3 and 4.2, Table (3.2) and Sections 3.3 and 4.2.3.

It is now determined which elements of $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$, *i.e.* covariances between which parts of the proteins, determine the intense 3D features in $I_{\text{diff}}^{\text{protein}}$ (indicated by F1–5 in Fig. 4.15 A). In Fig. 4.15 D is shown $I_{\text{diff}}^{\text{protein}}$ calculated using Eq. (2.70) with all elements of $\langle \mathbf{u}_k \mathbf{u}_{k'}^T \rangle$ involving intra-protein cross-correlations set to zero (the diagonal elements, $\langle \mathbf{u}_k \mathbf{u}_k^T \rangle$ are kept as setting these to zero would imply $\langle \mathbf{u}_k^2 \rangle = 0$ and thus the absence of any motion for atom k). The comparison between Figs. 4.15 A and D shows that this exclusion removes the intense 3D features and diminishes the intensity of the average, isotropic shell at $q \approx \hat{q}$. Thus, combining the above results demonstrates that the intense 3D features originate almost exclusively from those parts of the first five PCA modes that describe protein intra-molecular cross-correlations. Therefore, these are now analysed in more detail. Intra-protein cross-correlations within or between α-

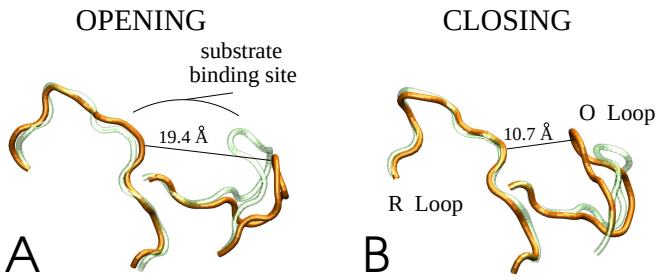

Figure 4.16: Two typical relative displacements between the flexible R and O loops (indicated in B) of different protein molecules due to the five lowest-frequency PCA modes. Shown are the structures after equilibration (green) and after 10 ns (orange).

helices and β-strands do not significantly contribute to the 3D features F1–5 (Fig. 4.15 E). Exclusion of these cross-correlations only results in a reduction of the average intensity of the isotropic shell at $q \approx \hat{q}$, cf. Fig. 4.6[2]. This is consistent with the above finding that the PCA modes 1–5 determining these features describe large-scale domain motions of the secondary structural elements, which scatter into the low-q range, but do not change the α-helix pitch or the inter-β-strand distance.

Specific large-amplitude intra-protein motions give rise to individual intense 3D features. For example, F1 and F3 on Fig. 4.15 A arise from cross-correlations between the C terminus (residues 136–149) and two of the α-helices, *i.e.* α_1 and α_2. Similarly, F4 contains a large contribution from correlations between two groups of atoms, one comprising the C and N (residues 1–9) termini together with two loop regions spanning residues 37–53 (O loop) and 107–120 (R loop), and the other comprising the α, β secondary structural elements. F4 also contains a contribution from cross-correlations between two proteins, which are located at the space-group symmetric positions (x, y, z) and $(-x, -y, z + 0.5)$, respectively, and have a common interface.

The largest mean-square fluctuations, $\langle \mathbf{u}_k^2 \rangle$ in the protein originate from the above mentioned C and N termini and O and R loops. The scattering arising from cross-correlations between these very flexible segments is shown in Fig. 4.15 F. The very intense 3D features F2 and F5 vanish. Thus, these features of the calculated diffuse scattering pattern can be directly assigned to motions involving these distinct parts of the protein. The R loop forms the lining of the substrate binding pocket and is surrounded by the O loop, which contains parts of the enzyme catalytic active site, and the C and N termini, cf. Fig. 4.17. Typical R and O loop displacements are shown in Fig. 4.16. The distance between the two loops significantly increases and decreases, corresponding to active site opening or closing motions: large-scale collective displacements that are of potential functional relevance for substrate binding.

4.3 Discussion and Conclusions

Four 10 ns molecular dynamics simulations of crystalline SNase have been analysed in terms of the X-ray diffuse scattering and models describing protein motion.

[2]Note that in the calculations for Fig. 4.6 the complete secondary structural elements were removed, thus implicitly assuming $\langle \mathbf{u}_k^2 \rangle = 0$ for these elements, whereas in the calculations for Fig. 4.15 only the cross-correlations were set to zero.

The MD simulated X-ray diffuse scattering patterns of the complete unit cell, *i.e.* the proteins and crystal solvent, are qualitatively similar to experiment, cf. Fig. 4.1. Furthermore, the agreement factor R with experiment continuously improves with increasing simulation lengths, cf. Figs. 3.3 and 4.2. Since the diffuse scattering is determined by the variance-covariance matrix, cf. Eq. (2.70), this suggests that MD is capable of properly describing protein collective motions in the crystal unit cell if long-enough timescales are sampled. For the present system, the variance-covariance matrix has been estimated to converge on the $\approx$1 μs timescale, cf. Chapter 3, *i.e.* $\approx$100 times longer than the present simulation lengths [100]. Further lengthening the simulation time may lead to improved agreement with experiment. Furthermore, Fig. 4.2 also demonstrates that the incorporation of hydrogens improves the simulated X-ray diffuse scattering pattern.

The protein and crystal solvent contributions to the unit-cell diffuse scattering pattern have been determined. The isotropically-averaged intensities $I_{\rm diff}^{\rm protein}$ were found to be closely similar to the structure factors $S(q)$ calculated from the radial distribution functions $g(r)$. The α-helix pitch and the β-sheet inter-strand distance, were found to be the origin of the pronounced peak in $I_{\rm diff}^{\rm protein}(q)$ at $q = 0.21$ Å^{-1}. The most-intense three-dimensional features in $I_{\rm diff}^{\rm protein}$ are also located around this q-value (Figs. 4.1 and 4.15) and, therefore, this q-range is a prominent target to determine protein collective motions, cf. Section 4.2.4 [160].

The 'very diffuse' X-ray scattering from protein crystals, which probes mostly intra-molecular motions and is that examined here, has hitherto been interpreted using models of liquid-like motion [76, 79, 96, 97, 101], rigid-body motion [27, 74, 78, 81, 98] or intra-molecular normal mode vibrations [80]. Here, models of liquid-like and collective motion have been investigated. Each of these are discussed in turn.

Previous studies using models of liquid-like motion yielded decay lengths over which short-range intra-molecular motion is correlated to be $\approx$6 Å for insulin [76], $\approx$6 Å for lysozyme in various crystal forms [76, 79], $\approx$3 Å for *yeast* initiator tRNA [96], $\approx$10 Å for *Staphylococcal* nuclease [101] and $\approx$5 Å for calmodulin [97]. These values can be compared with $\lambda_{\rm p}$ or $\lambda_{\rm int}$ given in Tables 4.1 and 4.2. The average intra-protein correlation length obtained directly from the simulations, (10.9 ± 0.3) Å, is in good agreement with the experimental results for *Staphylococcal* nuclease obtained by Wall *et al.* in Ref. [101]. This is likely to be serendipitous as the value obtained by fitting the model scattering to that calculated from the trajectories is significantly higher, being (14.4 ± 0.5) Å averaged over Models A, B and C. This disagreement with the results of Ref. [101] may partially be due to the fact that in Ref. [101] (and in other previous work [76, 79, 97]) the simplifying assumption was made that the atomic root-mean-square fluctuation, $\sqrt{\langle \mathbf{u}^2 \rangle}$, is the same for all atoms. $\sqrt{\langle \mathbf{u}^2 \rangle} = 0.36$ Å was obtained in Ref. [101] whereas the present simulation models incorporate individual atomic fluctuations taken directly from the simulations, the most probable value being 0.82 Å. However, the disagreement may also be due to insufficiencies of the liquid-like model itself. This is indicated by the strong differences between the parameters obtained from the model fits to the simulation-derived scattering (Table 4.2) and those calculated directly from the simulation trajectories (Table 4.1). Furthermore, Model A (modified to also refine $\sqrt{\langle \mathbf{u}^2 \rangle}$) was used to describe the scattering pattern of the pure-water simulation and obtained an exceptionally large λ=15.2 Å, $\sqrt{\langle \mathbf{u}^2 \rangle}$=0.71 Å and a poor R factor, being 12.9%. Moreover, the assumption of isotropic correlations may not be suitable [83]. Although generalisation to anisotropic models, in accord with the space group symmetry, is straight forward, the increased number of parameters may render an unambiguous interpretation of the results difficult. In particular, in the present work an increasing number of LLM fit parameters did not

significantly improve the diffuse scattering, as is indicated by a negligible reduction in the R factor. Therefore it is concluded that the protein dynamics may only be qualitatively described using the liquid-like motion model, and a quantitative description seems unachievable.

A theoretical description of diffuse scattering due to collective motions has been derived using normal mode analysis in Ref. [164] and the diffuse scattering of lysozyme has been interpreted using the 15 lowest-frequency normal modes [80][3] However, the normal-mode model possesses three major problems. The first is technical in that a normal mode analysis requires large amounts of computer memory. Present-day computers enable all-atom normal modes to be computed for proteins of up to approximately 300 amino acids, although approximations have been developed, *e.g.* the block normal-mode approach [177, 178], to enable treatment of much larger systems. Unit cells of protein crystals, in general, contain more than one protein molecule and a variable (but usually large) amount of solvent, all of which contribute to the diffuse scattering. The second difficulty arises from the normal modes themselves which, due to the harmonic approximation, cannot describe barrier crossing. However, as most diffuse scattering patterns are collected at non-cryogenic temperatures diffusive motion exists. As shown in this chapter diffusive motion significantly contributes to the scattering intensity, cf. Figs. 4.11 and 4.13. Third, the variance-covariance matrix, constructed using the displacements $\mathbf{u}_k$ expressed as a sum over collective variables, cf. Eq. (4.5), converges only slowly with the number of normal modes included. This is due to the fact that, for normal modes, $\sigma_m \propto \nu_m^{-1}$ and hence the sum in Eq. (4.5) converges only if the frequencies increase faster than $m^{1/2}$. In the present system, this holds true only for $m \gtrsim 500$. Also, the normal modes are orthogonal only in mass-weighted but not in real space and the scattering intensities due to each mode are not additive, cf. Eqs. (2.70), (4.5) and (4.6).

Here, the problem of harmonicity was circumvented by using the principal components of protein motion for which all non-hydrogen atoms of the four proteins in the unit cell were considered. It was found that the motions due to the ≈10 lowest-frequency modes dominate the protein diffuse scattering. If the lowest-frequency mode is (the nine lowest-frequency modes are) excluded the R factor reaches only 5% (10%), cf. Fig. 4.11. However, for the scattering pattern to fully converge, *i.e.* for the R factor to be essentially zero, the first approximately 7000 modes, corresponding to ≈50% of the total number of modes, must be included, cf. Fig. 4.11. This number is significantly larger than that obtained in Ref. [80] using a normal mode analysis for lysozyme.

The energy landscapes and dynamics of the PCA modes were also investigated. The lowest-frequency modes are strongly anharmonic. The associated effective free energy landscapes are rather flat and, for the first ≈5 modes, possess multiple substates separated by barriers with heights up to $\approx 4\, k_{\mathrm{B}} T_{300\,\mathrm{K}}$. Barriers of this height can be crossed by fluctuations of the kinetic energy contained in each mode. The timeseries of the projection of the trajectory onto the low-frequency modes, *i.e.* $\sigma_{m \lesssim 10}(t)$, is similar to cosines with $m/2$ periods, cf. the inset to Figs. 4.13 A and B – this behaviour was reported for high-dimensional random diffusion in a harmonic potential due to an undersampling of collective motions [168, 179]. Therefore, it can be concluded that the ≈10 lowest-frequency modes are purely diffusive. With increasing mode number the effective free energy profiles of the modes become close to harmonic and the associated timeseries $\sigma_{m \gtrsim 30}(t)$ are vibrational, cf. Figs. 4.13 C and D. The motion along these modes are described by Brownian motion in a harmonic potential and the damping characteristics are

[3] The description in Ref. [80] may in parts be erroneous as the total diffuse scattering intensity was expressed as a sum of contributions from individual normal modes. However, the scattering due to individual modes is not additive, as discussed at the end of the paragraph.

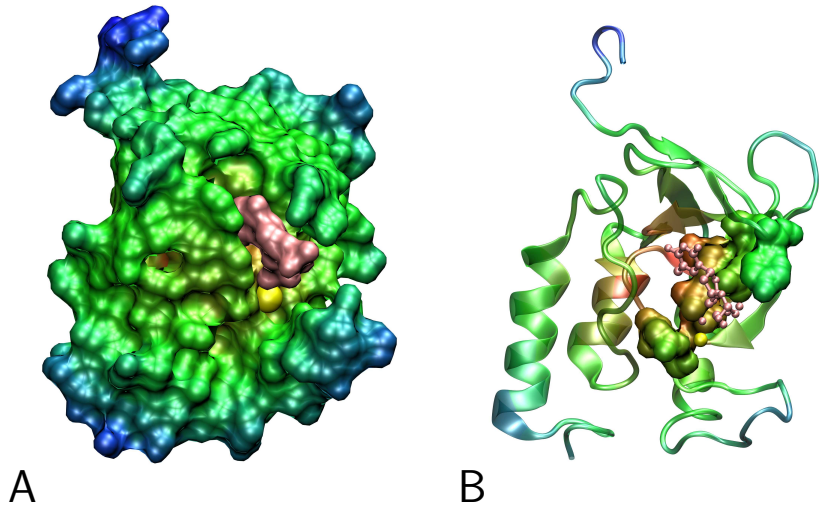

Figure 4.17: Substrate binding site and enzymatically active residues of SNase with the coordinates taken from 2SNS [125]. A: The SNase nucleotide binding site complexed with a small ligand, thymidine 3',5'-biphosphate (pink), and a calcium ion (yellow). The protein surface is shown with the colour coding symbolising the distance from the protein core (ranging from red to blue). B: The active-site residues are highlighted as volume-filling solid surfaces surrounding the inhibitor, which is shown in a ball-stick representation. The protein secondary structure is indicated using the cartoon representation.

determined. For the quasi-harmonic modes ($m = 30, \ldots, 100$) the damping frequency, Γ is larger than the eigenfrequency, ω_0 and therefore these modes perform overdamped vibrations. For the harmonic modes in the range $m \lesssim 200$, Γ and ω_0 are of comparable magnitude $0.5 - 0.6\,\mathrm{THz}$, corresponding to critical damping. The remaining majority of modes ($m \gtrsim 200$) is harmonic and performs underdamped vibrations.

In Section 4.2.4 it was demonstrated that the largest-amplitude, slowly-varying unconverged PCA modes dominate the 3D fine structure of the diffuse scattering and describe motions delocalised over all four protein molecules in the unit cell. Furthermore, individual 3D features in the diffuse scattering were directly assigned to originate from cross-correlations between specific protein segments. In particular, some of the 3D features were found to arise from active-site motions of potential functional relevance. This is illustrated in Fig. 4.17 which shows the substrate binding site and enzymatically active residues of the protein. The opening and closing motions of the R and O loops, cf. Fig. 4.16, potentially restrict access of the substrate to the binding groove (Fig. 4.17 A) and conformational changes in the O loop are likely to affect the accurate steric alignment of the enzymatically active residues (Fig. 4.17 B).

Chapter 5

Pressure-Dependent Dynamical Transition

The temperature dependence of internal protein dynamics has been much studied and has yielded valuable information on the energy landscape underlying protein function [46, 181, 182]. In comparison, relatively few studies have been performed investigating the dependence of protein dynamics on pressure [183–185]. Pressure-induced structural changes have been reported for deoxymyoglobin [186], lysozyme [187–189], BPTI [190], myoglobin [191] and ubiquitin [192], and pressure-induced unfolding in solution has been reported for *Staphylococcal* nuclease [193–196], myoglobin [197], α-lactalbumin [198] and various other proteins [199]. Hydration water has been suggested to play a key role in high-pressure protein unfolding, as indicated by structural and dynamical changes in the protein:water interface and the penetration of water molecules into the hydrophobic core [190, 196, 197, 200, 201]. Pressure-induced dynamical changes have hitherto been relatively neglected, although in early molecular dynamics simulations of small proteins a reduction of positional fluctuations of protein atoms was found upon the application of pressure [190, 200, 202].

In this chapter, results from MD simulations (Set 2, cf. Section 2.2) of crystalline SNase in the pressure range 1 bar to 15 kbar are presented. The advantage of the crystalline state over solution is that steric constraints imposed by the crystalline environment hinder denaturation, as has been demonstrated for orthorhombic crystals of lysozyme, which remain stable up to 10 kbar [203]. SNase was chosen here as it is experimentally well characterised and because crystals of the protein remain structurally stable at pressures at which SNase in solution is partly unfolded [204]. In solution or a close-to-native environment, however, proteins denature upon the application of pressure of a few kbar.

In the following section, the stability of the MD simulations is verified. In particular, the protein and unit-cell compressibilities are compared with experimental data and the root mean-square deviations between the crystallographic, ambient-pressure structure and the high-pressure simulation structures are determined. The solvent structure is also examined. In the subsequent section, the focus is on dynamical properties. The analysis includes the study of the atomic mean-square displacements, which are one-particle properties, and the study of collective dynamics using X-ray diffuse scattering and principal component analysis.

Results of this chapter will be published as *Meinhold & Smith*; submitted to Phys Rev E, cf. Ref. [180].

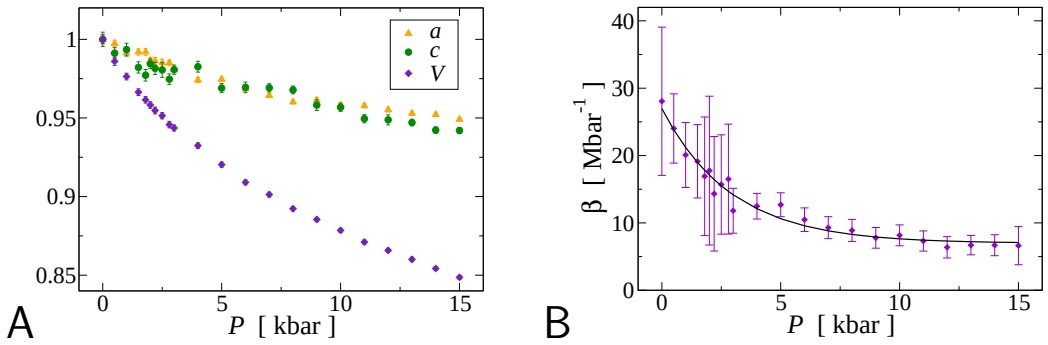

Figure 5.1: A: Pressure-dependence of the average unit cell sides and volume. The data is normalised to the values at ambient pressure, $P = 1\,\text{bar}$. B: Pressure-dependence of the unit-cell compressibility, β. The solid line represents a least-squares fit of the mechanical non-linearity index, $\mu = \partial\beta^{-1}/\partial P$, see text. Error bars denote the standard deviation. Here and henceforth in all graphs, P denotes the reference pressure set in the simulation around which the instantaneous pressure values fluctuate with a standard deviation of $\approx$0.8 kbar.

5.1 Structural Relaxations

MD force fields for protein simulations have been calibrated for ambient pressure and temperature and have been shown to be less accurate in some respects at elevated temperatures and pressures [184, 205, 206]. However, the work on the T-dependent dynamical transition [48, 207] is in quantitative agreement with experiment and has demonstrated that MD provides a qualitative understanding outside the range over which the force field was parameterised.

In this section the structural response of the system to the applied pressure is investigated. This validation of the MD simulations is required as crystalline relaxation processes have been shown to be present on timescales much larger than the present simulation length of $\mathcal{T} = 1\,\text{ns}$. For example, the unit-cell sides in MD simulations of a double unit-cell of ubiquitin were found to deform slowly and reached convergence only after 5–10 ns [140]. This particular difficulty was partially circumvented here by using the final configuration of simulation T1, *i.e.* a structure equilibrated at ambient pressure for 10.2 ns. However, at higher values of the pressure the system may further relax into a configuration different from that adopted at ambient pressure.

In Fig. 5.1 A is shown the pressure-dependence of the unit cell sides and volume. The unit cell sides shrink with increasing pressure, the reduction at 15 kbar relative to 1 bar being $5.1 \pm 0.1\%$ and $5.8 \pm 0.2\%$ for a and c, respectively. The unit-cell volume decreases non-linearly with increasing P, the reduction from 1 bar to 15 kbar being 15%. The pressure-dependence of the unit-cell compressibility,

$$\beta = \frac{\partial(\ln V)}{\partial P}, \tag{5.1}$$

is shown in Fig. 5.1 B. β reduces from $28 \pm 11\,\text{Mbar}^{-1}$ at 1 bar to $6.6 \pm 2.8\,\text{Mbar}^{-1}$ at 15 kbar with the mechanical non-linearity index, defined as

$$\mu = \frac{\partial\beta^{-1}}{\partial P}, \tag{5.2}$$

being $\mu = 9.1 \pm 0.4$. These values are within the experimental ranges of estimates for β (10–20 Mbar^{-1}) and μ (0–10) reported for the P-dependence of the unit-cell volume in lysozyme

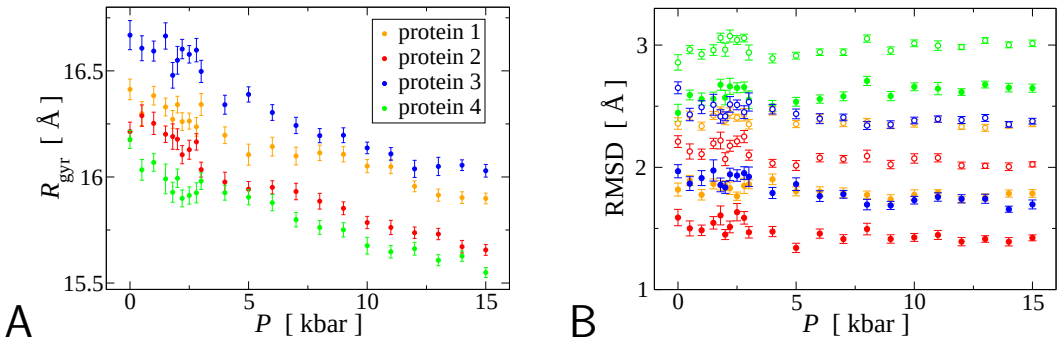

Figure 5.2: A: Pressure-dependence of the radius of gyration, R_{gyr} calculated over the full trajectories for all four protein molecules in the unit cell. B: Pressure-dependence of the root mean-square deviation, RMSD between the ambient-pressure experimental structure 2SNS and the single-protein structures from the simulations. The RMSD is calculated for only the C_α-atoms (filled circles) and for all non-hydrogen atoms (open circles) for the experimentally-resolved residues 1-141, the colour-coding is the same as in graph A. Error bars denote the standard-deviation.

crystals [203, 208, 209]. Furthermore, the ambient-pressure value for the unit-cell compressibility is between estimates for protein compressibilities (2–15 Mbar^{-1}) and that of pure water (45.8 Mbar^{-1}) [183].

The compressibility of the protein can be roughly estimated from the radius of gyration, R_{gyr}. R_{gyr} can be measured experimentally by small-angle X-ray and neutron scattering and provides a convenient measure to detect folding/unfolding transitions [194, 196]. Here, R_{gyr} was calculated directly from the trajectories using

$$R_{\mathrm{gyr}} = \sqrt{\frac{\sum_i m_i \|\mathbf{r}_i - \mathbf{r}_0\|^2}{\sum_i m_i}}, \tag{5.3}$$

where the protein centre of mass, $\mathbf{r}_0$ is given by

$$\mathbf{r}_0 = \frac{\sum_i m_i \mathbf{r}_i}{\sum_i m_i}. \tag{5.4}$$

In Fig. 5.2 A the time average of R_{gyr} of all four proteins is plotted against P. Averaged over all four proteins, R_{gyr} decreases by $3.6 \pm 0.4\,\%$ from 1 bar to 15 kbar, and the average protein volume, estimated using $V_p = {}^{4\pi}/_{3} R_{\mathrm{gyr}}^3$, thus reduces by $10.3 \pm 1.0\,\%$. Therefore, the protein compressibility is significantly smaller than that of the unit cell, cf. Fig. 5.1 A, and thus also the water compressibility, which is again consistent with previous findings [183].

The decrease of protein and solvent volumes is also evident from the radial distribution functions, $g(r)$ shown in Fig. 5.3. For both protein and solvent, at large values of r the high-pressure $g(r)$ is slightly larger than the ambient-pressure $g(r)$, reflecting the increased density at higher P. This effect is smaller for the protein, which is consistent with the above finding of a smaller protein compressibility relative to that of water. Except for the small increase at larger values of r, the functional form of the protein $g(r)$ (Fig. 5.3 A) shows no variation upon the application of pressure, indicating that the local protein structure is not affected at higher values of P.

The decrease of solvent volume is manifested by a shift to shorter distances of the nearest-neighbour peak in the solvent radial distribution function (Fig. 5.3 B), corresponding to a decrease in the average nearest-neighbour distance of 0.7% from 1 bar to 15 kbar. Additionally,

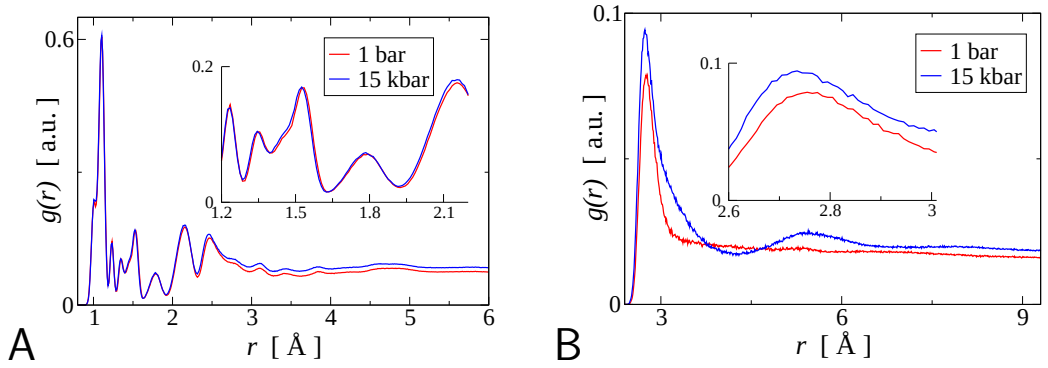

Figure 5.3: Radial distribution functions, $g(r)$ calculated over the full trajectories for the protein (A, all atoms) and the crystal water (B, O–O distance) at two values of P. For convenience, $g(r)$ is not normalised by the density.

upon the application of pressure a second peak occurs at ≈5.4 Å in the solvent $g(r)$, indicating a pressure-induced change in the water structure. The ambient-temperature ambient-pressure water structure possesses a tetrahedral order due to hydrogen bonds between neighbouring water molecules. In liquid water, the hydrogen-bond energy is 23.3 kJ mol^{-1}* [214], being approximately ten times the average thermal energy at 300 K. Using the mechanical work, $-P\mathrm{d}V$, and assuming equi-partition of that energy, the increase of the internal energy due to the application of pressure is estimated to be 0.29 kJ mol^{-1} and 0.8 kJ mol^{-1} for the crystal water at 5 kbar and 10 kbar, respectively[1]. This increase in the internal energy effectively lowers the barrier for hydrogen bond breaking and thus influences the water dynamics, which will be discussed in more detail in the next section, cf. the discussion of Fig. 5.5 B.

At the end of this section it is investigated whether the protein conformations adopted at higher values of P remain close to the native structure. The average root mean-square deviations of the single-protein simulation structures from the experimental (300 K, 1 bar) crystal structure 2SNS [125] are depicted in Fig. 5.2 B and showed little variation with increasing P, the average over all four proteins in the unit cell being 1.95±0.36 Å (2.52±0.29 Å) and 1.89±0.53 Å (2.44±0.41 Å) at 1 bar and 15 kbar, respectively, for the C_α-atoms (all non-hydrogen atoms). Also, the secondary structural elements were conserved throughout the simulations. Furthermore, no penetration of water molecules into the protein hydrophobic core was observed, although in some simulations water molecules did enter or leave cavities located inside the protein. In the crystal structure 1STN [126] most of these cavities were also occupied by water molecules.

The ensemble of the above findings indicate that the protein and simulation system were stable at all pressures and that pressure-related structural features agree within error with experiment. This warrants a more probing analysis of the simulation data in the next section, in which the protein and solvent dynamics are investigated.

*That value corresponds to the energy required for breaking the bond and completely separating the molecules. Just breaking the hydrogen bond and leaving the molecules approximately at the same position requires only 27% of this energy, recently estimated at 6.3 kJ mol^{-1} [210–213].

[1]In a system with an inhomogeneous compressibility the mechanical work, $-P\mathrm{d}V$ is not equi-partitioned. Here, since the water compressibility is larger than that of the proteins, a larger fraction of $-P\mathrm{d}V$ is stored in the water, leading to even larger estimates for the increase in water internal energy upon the application of pressure.

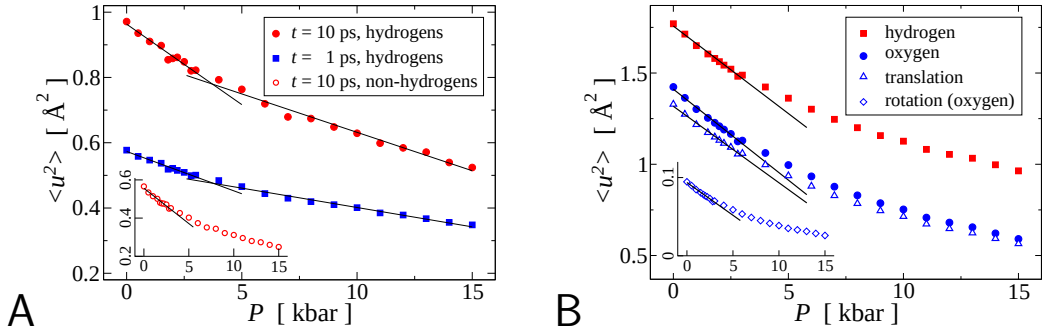

Figure 5.4: A: Mean-square displacement, $\langle u^2 \rangle$ averaged over all protein hydrogen atoms on two timescales, t. The lines show linear fits performed over the ranges $P \leq 3\,\mathrm{kbar}$ and $P \geq 4\,\mathrm{kbar}$. The shape of $\langle u^2 \rangle(P)$ is similar for t=100 ps (data not shown) and for non-hydrogen protein atoms (shown in the inset). B: $\langle u^2 \rangle(P)$ for $t = 1\,\mathrm{ps}$ averaged over all solvent hydrogens and oxygens, respectively. For the oxygens, the decomposition into translation and rotation, $\langle u^2 \rangle = \langle u^2 \rangle_\mathrm{T} + \langle u^2 \rangle_\mathrm{R}$, is also shown. The lines represent linear fits performed in the low-P regime.

5.2 Change in Protein and Solvent Dynamics

A convenient measure for the overall motion present in a protein molecule is provided by the time-dependent mean-square displacement,

$$\langle u^2 \rangle(t) = \left\langle \overline{\left[\mathbf{r}_k(\tau) - \mathbf{r}_k(\tau + t) \right]^2} \right\rangle, \tag{5.5}$$

where $\mathbf{r}_k(\tau)$ is the coordinate vector of atom k at time τ, $\overline{\cdot}$ is the time average and $\langle \cdot \rangle$ the ensemble average, *i.e.* over all protein atoms k. In Fig. 5.4 A $\langle u^2 \rangle(t)$ for the internal motion is plotted against P for two values of t: 1 ps and 10 ps. On both timescales $\langle u^2 \rangle$ significantly decreases with increasing pressure, with the reduction between 1 bar and 15 kbar for hydrogens being 46% and 40% for $t = 1\,\mathrm{ps}$ and 10 ps, respectively. For non-hydrogen atoms, shown in the inset to Fig. 5.4 A, the reduction is larger, being 56% for $t = 10\,\mathrm{ps}$. The slope of $\langle u^2 \rangle(P)$ is linear in two distinct ranges of P, with a broad transition around $P^* \approx 4\,\mathrm{kbar}$. In the regimes below and above P^*, referred to in the following as 'low-P' and 'high-P', respectively, linear regressions were performed and are also shown in Fig. 5.4 A. For both values of t, in the low-P regime the slope is higher by a factor of ≈2 than in the high-P regime. This non-linearity in $\langle u^2 \rangle(P)$ indicates a qualitative change in protein dynamics upon pressurisation and is reminiscent of the much-studied transition in $\langle u^2 \rangle(T)$, which involves a solvent-driven activation of anharmonic protein dynamics with increasing T at ≈200 K, leading to a non-linear increase in the average atomic mean-square displacement, $\langle u^2 \rangle(T)$ [48, 51, 52].

The solvent $\langle u^2 \rangle(P)$ (Fig. 5.4 B) is dominated by translational diffusion. Both the translational and rotational $\langle u^2 \rangle$ decrease linearly with P below P^* and nonlinearly, at a lower rate, above P^*. At all pressures, $\langle u^2 \rangle(t) \propto t^\alpha$ with $\alpha < 1$, indicating subdiffusion [215]. α exhibits no significant P-dependence, the average calculated over all simulations being 0.86 ± 0.01, a value intermediate between that of protein hydrational water, for which $\alpha \approx 0.6$ [67], and bulk water ($\alpha = 1$).

To examine which protein and solvent motions are affected by the pressure increase the P-dependence of the X-ray diffuse scattering intensity, I_diff was calculated, cf. Eqs. (2.58) and (2.70).

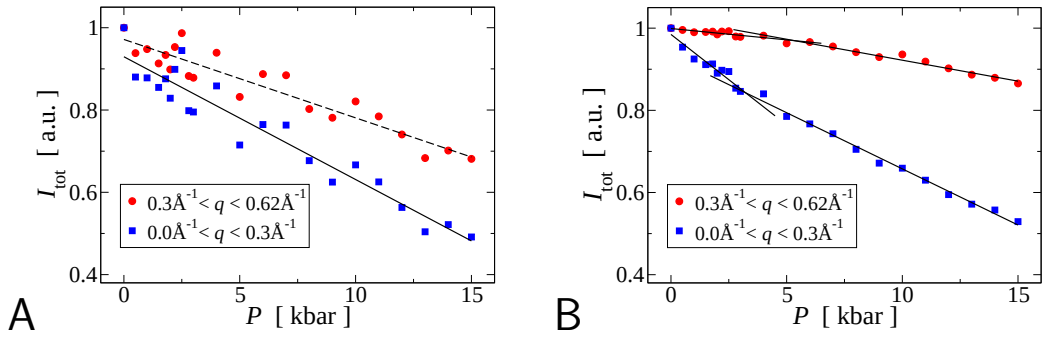

Figure 5.5: Protein (A) and solvent (B) X-ray diffuse scattering intensities, I_{tot} integrated over two ranges of the magnitude, q of the scattering vector. I_{tot} was calculated from the full trajectories (1 ns) and is normalised such that $I_{\mathrm{tot}}(1\,\mathrm{bar})=1$. The fluctuations in the protein $I_{\mathrm{tot}}(P)$ are relatively large due to the fact that calculated protein X-ray diffuse scattering does not converge on the ns-timescale, cf. Section 3.1.3 [100]. A: The solid (dashed) line is a linear fit to the data on the full P-range with $q < 0.3\,\mathrm{Å}^{-1}$ ($q > 0.3\,\mathrm{Å}^{-1}$). B: The lines show linear fits in the low- and high-P regimes, respectively.

I_{diff} indicates the amplitude of collective motion present on the length scale, q^{-1}. Here, I_{diff} was calculated for the same distribution of scattering vectors, $\mathbf{q}$ as in the preceding chapters, *i.e.* for 55,691 values in the range $\|\mathbf{q}\| < 0.62\,\mathrm{Å}^{-1}$. This range was divided into a low-q range, $q < 0.3\,\mathrm{Å}^{-1}$, probing large-scale collective motions, and a high-q range, $q > 0.3\,\mathrm{Å}^{-1}$, probing more local dynamics. I_{diff} was integrated over each range to yield the total intensity, I_{tot} which is plotted against P in Fig. 5.5 A.

$I_{\mathrm{tot}}(P)$ for the protein decreases faster in the low-q range (with the gradient being $-0.030 \pm 0.002\,\mathrm{kbar}^{-1}$) than in the high-$q$ range (with the gradient being $-0.019 \pm 0.002\,\mathrm{kbar}^{-1}$), indicating that, upon the application of pressure, large-scale collective displacements are more strongly suppressed than local dynamics. $I_{\mathrm{tot}}(P)$ calculated for the crystal solvent is shown in Fig. 5.5 B. In the low-q range the solvent $I_{\mathrm{tot}}(P)$ strongly decreases with increasing P, the rate of decrease being larger in the low-P regime. At low q, the large-scale collective displacements, $I_{\mathrm{tot}}(P)$ of the protein and the solvent reduce approximately equally, reaching 0.49 and 0.53, respectively at $P = 15\,\mathrm{kbar}$. This indicates that large-scale collective motions of the protein and collective translational solvent dynamics are similarly affected by high pressure, consistent with previous work suggesting strong coupling between solvent dynamics and large-scale protein motion [64, 66, 216, 217]. In contrast, the P-dependence of the high-q, short length-scale motion differs significantly between the protein and the solvent: whereas the protein I_{tot} reaches 0.68 at $P = 15\,\mathrm{kbar}$ that of the solvent reduces to only 0.86. Furthermore, the change in the solvent I_{tot} is negligible up to $\approx$4 kbar, indicating that the local dynamics of water remains roughly unchanged in the low-P regime. This is consistent with previous work indicating that, at ambient temperature and low P, water mobility is controlled by the tetrahedral ordering due to hydrogen bonds whereas, at higher P, the tetrahedral ordering breaks down and the dynamics is controlled by the van der Waals repulsion of neighbouring molecules [218–220].

The protein collective motions were further dissected using principal component analysis, cf. Section 4.2.2. PCA was performed individually on all four proteins in the unit cell, thus analysing only intra-molecular motions. From the eigenfrequencies, ν_m the protein vibrational density of states, $g(\nu)$ was calculated and averaged over all four proteins in each simulation. The change in $g(\nu)$ upon application of pressure is illustrated in Fig. 5.6. With increasing P the number

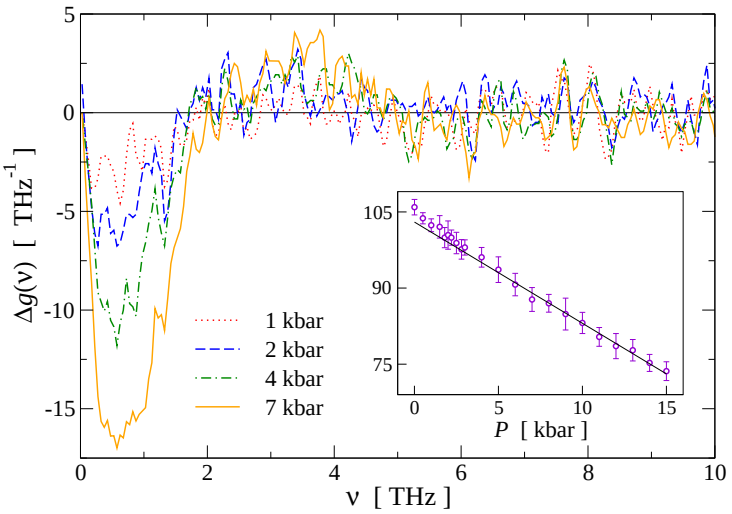

Figure 5.6: Change in the vibrational density of states, $\Delta g(\nu)$ for selected pressure values. $\Delta g(\nu)$ was calculated as the difference $g_P(\nu) - g_{P_0}(\nu)$, with the reference pressure $P_0 = 1\,\text{bar}$. The inset shows the number of modes with $\nu \leq 2\,\text{THz}$ with a linear fit to the high-P regime. The gradient in the low- and high-P regime is $-2.6 \pm 0.2\,\text{kbar}^{-1}$ and $-1.99 \pm 0.06\,\text{kbar}^{-1}$, respectively. Results shown are averaged over all four proteins in each simulation. Error bars denote the standard deviation.

of modes with frequencies $\nu \lesssim 2\,\text{THz}$ strongly decreases, the rate of decrease exhibiting a small non-linearity at P^* (inset to Fig. 5.6). Low-frequency modes are thus shifted into the higher-frequency range. These low-frequency modes have been shown to be mostly collective, anharmonic and distributed over the whole protein [52, 167], cf. Chapter 4.

It is of interest to investigate the form of the effective free energy profiles associated with the PCA modes and, in particular, whether the P-dependent dynamical transition found above is accompanied by a loss of anharmonic motions similar to that seen in the T-dependent dynamical transition. In Fig. 5.7 are shown the effective free energy profiles, G_m for selected PCA modes (mode numbers m=1, 5, 30 and 100) at various pressure values (P=1 bar and 2, 4, 8 and 15 kbar). At all values of P the lowest-frequency, largest-amplitude principal component mode is anharmonic and possesses multiple substates separated by barriers with heights up to approximately $10\,\text{kJ mol}^{-1}$. Also, for all values of P the timeseries of the projection, $\sigma_1(t)$ of the trajectory onto the first principal mode is similar to a half-period cosine (data not shown, similar to that shown in Fig. 4.13), indicating that the lowest-frequency PCA mode is purely diffusive and did not converge in the simulations, cf. Section 4.3 [168, 179]. With increasing mode number m, the change in G_m is similar to that described in Fig. 4.13 for the ambient-pressure simulations, *i.e.* the modes initially remain anharmonic but with the number of substates reducing until G_m becomes quasi-harmonic or harmonic at $m \approx 10$ or $m \approx 30$, respectively. The effect of elevated pressure is similar for all modes. Upon the application of pressure the width of G_m reduces, corresponding to an increase of the effective force constant associated with each mode.

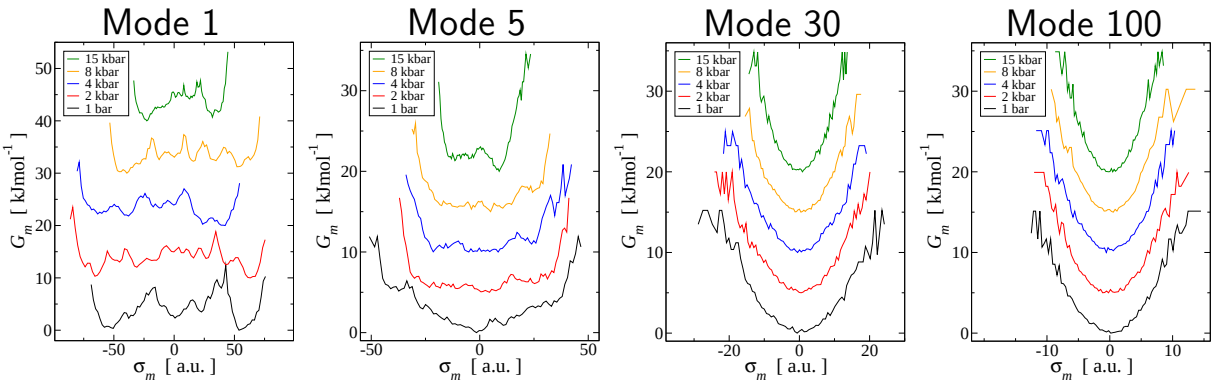

Figure 5.7: Pressure-dependence of the effective free energy profiles, G_m of selected PCA modes at various values of P, indicated in each graph. G_m was calculated using Eq. (4.7) on pg. 54, cf. Fig. 4.13. For clarity, in each plot the profiles are separated by a constant ($10\,\mathrm{kJ\,mol^{-1}}$ for Mode 1 and $5\,\mathrm{kJ\,mol^{-1}}$ for the other modes) along the y-direction.

5.3 Discussion and Conclusions

The present MD simulations of a crystalline protein in the pressure range 1 bar to 15 kbar have revealed the existence of a pressure-dependent transition in internal protein dynamics at ≈4 kbar. The transition is manifested by the existence of two linear regimes in $\langle u^2 \rangle(P)$. The major effect of pressure is a loss, with increasing pressure of large-amplitude, collective modes below 2 THz effective frequency.

The crystalline environment used here may strongly influence the protein dynamics and, in particular, prevents pressure-induced unfolding at medium pressure values. Therefore, further investigations are required to elucidate whether the pressure-induced changes in the protein dynamics found here are also present in a more native environment, *e.g.* a single solvated protein. However, convergence of these simulations may be poor due to slow relaxation processes, *i.e.* protein unfolding, estimated to occur on the timescale of minutes for SNase [194]. Also, as an increasing number of high-pressure non-native protein structures becomes available it is of interest to perform a similar study to investigate whether the pressure-induced changes in protein dynamics found here are limited to a close-to-native region of the energy landscape.

Further characterisation of the dynamics below and above the pressure dynamical transition using a variety of experimental scattering and spectroscopic techniques promises to shed further light on the physics of protein energy landscapes.

Chapter 6

Concluding Remarks and Outlook

In this thesis, various aspects of protein dynamics in a crystalline environment have been investigated using molecular dynamics simulations of *Staphylococcal* nuclease. The following questions have been addressed:

> *Did the simulation sufficiently sample the relevant regions of phase space, i.e., did the quantities of interest converge during the simulation?*
>
> *What is the origin of the X-ray diffuse scattering intensity from a protein crystal?*
>
> *How does the application of pressure change the internal protein dynamics?*

Each of these questions has been discussed in Chapters 3 to 5 and relevant conclusions presented. The synopsis of the major results is presented here along with some concluding remarks and an outlook towards possible directions of future investigations.

Molecular Dynamics Simulations

Computer simulations have developed into an indispensable tool to bridge the gap inbetween theory and experiment. Restricted by computational resources, simulation techniques allow highly-complex theoretical models to be tested against experimental data and processes that are not accessible experimentally to be investigated. However, as access to computational resources is limited by technological constraints there is always a trade-off between the implemented model accuracy and simulation speed.

In molecular dynamics, the simulation technique used in this thesis, speed is gained by replacing an explicit quantum mechanical description of atoms with empirical force fields, rendering possible the treatment of systems comprising more than 100,000 atoms, such as the membrane water-channel aquaporin [221], but prohibiting the description of certain processes relevant for biological protein function, such as chemical reactions and light absorption. However, hybrid quantum/classical methods have been developed and provided important insight into biological processes, such as the photon absorption and proton transport in bacteriorhodopsin [222, 223].

When using the Born-Oppenheimer framework, the accuracy of the approximation of the real energy landscape by the empirical force field must be part of the discussion. In particular, a

commentary is required if non-standard conditions are investigated by the simulations, as is the case in the present high-pressure study. Molecular dynamics force-fields for protein simulations have been calibrated for ambient pressure and temperature and have been shown to be less accurate in some respects at elevated temperatures and pressures [184, 205, 206]. However, the work on the temperature-dependent dynamical transition is in quantitative agreement with experiment and has demonstrated that molecular dynamics provides a qualitative understanding outside the range over which the force-field was parameterised.

X-ray Diffuse Scattering

The central theme of this thesis has been X-ray diffuse scattering. Although it was shown that the protein diffuse scattering did not converge on the 10 ns timescale of the present simulations, it was possible to draw a set of important conclusions concerning the dynamical origin of the X-ray diffuse scattering from protein crystals. The following paragraphs summarise these findings.

A smooth, nearly-isotropic scattering shell originates from equal contributions from correlations in nearest-neighbour water molecule dynamics and from internal protein motions, the latter consisting of α-helix pitch and inter-β-strand fluctuations. Superposed on the shell are intense, three-dimensional scattering features. For the first time, these individual features have been assigned to specific collective motions in the protein, and some of these explicitly involve potentially functional active-site deformations. However, the presence of these deformations in the real crystal cannot be inferred from the present simulations due to the limited convergence reached.

Models of protein motion and their predicted diffuse scattering have also been investigated. The critical analysis of a model of diffuse scattering due to liquid-like protein motions [164], such as commonly used for the analysis of experimental diffuse scattering [76, 79, 97, 101], has shown that this model yields dissatisfying quantitative results. In particular, the length over which atomic motions are correlated is incorrectly estimated by this model. An alternative model, in which the protein motions are decomposed into a set of collective principal component modes, has also been characterised. In this model, a very small number of slowly-varying (>10 ns), large-amplitude collective modes was found to describe most of the intense three-dimensional diffuse scattering features.

The logarithmic time-dependence of the agreement R factor between the simulated and experimental diffuse scattering, which is determined by the displacement variance-covariance matrix, has been extrapolated to predict a convergence time for the diffuse scattering, and thus the whole variance-covariance matrix, on the 1 μs timescale, *i.e.* 100 times longer than the presented simulations. Assuming the validity of Moore's law and algorithmic improvements of molecular dynamics simulations, *e.g.* those allowing the use of larger integration timesteps, for the system size described here the microsecond timescale is expected to be within reach in approximately five years. Currently, in an attempt that is commonly termed brute-force computing, one of the present simulations (T1) is being extended to 100 ns, in total requiring approximately 180,000 CPU hours. Besides the improved reproduction of diffuse scattering such an extended trajectory is essential to verify/falsify the convergence estimates made in thesis. Furthermore, the improved sampling of the low-frequency collective motions is likely to furnish insight into the shape of the essential subspace of the protein energy landscape [25]. Moreover, slow structural rearrangements of protein molecules in the crystal and the coupling between slow and fast protein motions

can be investigated.

As computational resources increase, it is also of interest to perform similar studies on a system composed of several unit cells. Correlated motions between unit cells commonly produce diffuse scattering streaks associated with the reciprocal lattice. In the present simulations of a single unit cell, these correlations are suppressed due to the imposition of periodic boundary conditions. In a recent simulation study of a double unit cell of ubiquitin only little correlation between the motions of the proteins in the two unit cells has been observed [140]. However, this study only determined the correlation coefficient between the protein centre-of-mass displacements. Covariances in the atomic displacements, which determine the diffuse scattering, or other more general correlation measures have not been analysed.

Furthermore, it is important to determine the diffuse scattering due to lattice vibrations which is centred at, and thus contaminates, the Bragg peaks. Experimentally, lattice vibrations have been determined for the heme group in myoglobin using the phonon-assisted Mössbauer effect [224] and for lysozyme using Brillouin scattering [225]. Diffuse scattering due to lattice vibrations has been calculated for crystalline rigid-body naphtalene [226] and the collective dynamics in L-alanine crystals have been determined by both neutron scattering and normal mode calculations in the full configurational space [227]. However, phonons in protein crystals have not yet been determined from dynamical calculations in the full configurational space. In addition to an improved description of diffuse scattering (and similar techniques, *e.g.* neutron and Raman scattering), improved information on phonons in protein crystal is likely to provide insight into thermodynamical and transport properties, phase stability and phase transitions.

X-ray Bragg Scattering

In X-ray crystallography, isotropic B factors are widely used to derive atomic fluctuations during the refinement of a protein structure against diffraction data. Here, the question has been discussed as to whether an unambiguous description of the dynamics involved can be derived from the B factor data in the absence of detailed additional information. As a variety of fundamentally different models reproduce experimental B factor distributions, it has been concluded that additional information must be supplied. In the present thesis, this additional information has been furnished by the molecular dynamical equations of motion and the associated atomic model and force field. In agreement with previous investigations [136–138, 144], the simulation-derived atomic fluctuations qualitatively resemble, but are significantly larger than, those derived from the experimental B factors. Although part of the difference between simulated and experimental fluctuations is due to the harmonic assumption made in deriving B factors, most of the difference has other origin. For example, the presence of crystallisation agents, ligands and/or large-size ions present in the crystal potentially reduce protein fluctuations due to steric hindrance. Molecular dynamics is a suitable technique to investigate the effect of these extra atoms/molecules on crystalline protein fluctuations. Furthermore, the use of anisotropic instead of isotropic B factors provides direction-dependent atomic fluctuations and is likely to improve results for structural refinement. For example, it has been shown that the use of anisotropic B factors significantly decreases R factors [166]. However, whether anisotropy can be used to reduce the ambiguity in the dynamical description of atomic fluctuations derived from B factors remains an open question.

Protein Crystallography

The present-day objective of protein crystallography is to determine the physiological three-dimensional structure of the protein as precisely as possible and to provide insight into its functional mechanism. Increasingly brilliant synchrotron X-ray sources allow both high-resolution and time-resolved structures to be determined [47, 228]. However, in order to prevent radiation damage the crystals have to be held at cryogenic temperatures, with potential temperature-induced structural changes [229, 230]. Furthermore, cryogenic temperatures significantly reduce the fluctuations and, in particular, freeze the protein molecules into an ensemble of conformational substates, thus prohibiting transitions between these substates which may be present at ambient temperature. These considerations argue for additional experiments at ambient temperature to verify the low-temperature structural and dynamical findings and to have direct access to the dynamics at physiological temperatures. Diffuse scattering is likely to contribute to the understanding of the motions within and, at physiological temperatures, transitions between conformational substates as it provides direct information on the collective dynamics involved. Furthermore, the physics underlying diffuse scattering may be helpful in the interpretation of future experiments on (single) biomolecules using fourth generation X-ray sources, *e.g.* free electron lasers with significantly increased brilliance [231].

Collective Dynamics

Collective protein dynamics can be thought of as a voyage over a complex energy landscape. In this thesis, collective motions have been determined by the decomposition of molecular dynamics trajectories into collective modes using principal component analysis. The lowest-frequency, largest-amplitude modes are strongly anharmonic. The associated effective free energy landscapes are rather flat and, for the first five modes, possess multiple substates. Analysis of the timeseries of the projection of the trajectory onto the modes have revealed that approximately the ten lowest-frequency modes are purely diffusive and unconverged. With increasing mode number the effective free energy profiles of the modes become close to harmonic and the associated dynamics vibrational. The eigen and damping frequencies associate with each mode have been determined using a model of Brownian dynamics. Critical damping was found for modes in the range $0.5 - 0.6\,\mathrm{THz}$ effective frequency, approximately 200 modes were overdamped and the rest (98.6%) performed underdamped vibrations.

The vibrational density of states has also been determined here and was found to be maximal around $1.4\,\mathrm{THz}$. Upon protein function, changes in the vibrational partition function may significantly contribute to the thermodynamics of the process, *e.g.* in the binding and unbinding of a substrate or ligand. For example, very recently the change in the vibrational density of states, *i.e.* a change in the collective dynamics, upon the complexation of the protein dihydrofolate reductase with the ligand methotrexate has been determined using incoherent neutron scattering and was found to significantly contribute to the free energy of binding [173]. In particular, this change involved a softening of low-frequency collective modes with frequencies below $0.5\,\mathrm{THz}$. More research is needed to understand how ligands and/or the environment trigger changes in the collective motions and, thus, in the protein energy landscape. Molecular simulation is an excellent technique to study these processes. Furthermore, terahertz spectroscopy [232] has recently been used to examine low-frequency collective vibrations in DNA, bovine serum albumin and collagen [233] and may provide information complementary to that obtained by inelastic neutron scattering.

Proteins Under Pressure

The effect of the environmental variable pressure on the protein dynamics has been investigated in this thesis. Pressure is one of the major thermodynamic variables and its importance is comparable to that of temperature, chemical potential, electric and magnetic fields etc. The finding reported in this thesis of a pressure-induced transition in the protein-internal dynamics opens up a whole new research field in protein physics that may well be as extensively investigated as the temperature-dependent protein dynamical transition has been. Experimental techniques such as NMR spectroscopy, X-ray crystallography and neutron scattering can now be applied to characterise the protein energy landscapes explored in the two dynamical regimes revealed here. It is only recently, that these techniques have been made available for studying protein structural and dynamical changes upon the application of pressure. For neutron scattering experiments, in which dynamical properties such as the mean-square displacement can be directly detected, high-pressure setups suitable for protein studies are currently being installed (for example at the ILL in Grenoble or the FRM2 in Munich).

Of particular interest is the understanding of pressure-induced structural changes and unfolding. Protein hydration water has been suggested to play a key-role in pressure-induced unfolding and, therefore, is a prominent subject of further investigations. These investigations should include the determination of the pressure-dependence of dynamical properties of the hydration water, such as diffusion constants and surface–bulk exchange rates, and protein–water interaction, such as hydrogen bonding and slaving [216]. Furthermore, pressure-induced changes in the protein and solvent volumes, and thus changes in free energy, are likely to provide insight into the ensemble of populated substates on a protein energy landscape.

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.

Thomas Stearns Eliot
Little Gidding

Chapter A

Transcript of Selected Charmm Input

In this appendix, the relevant input to the Charmm program is given to allow full reproduction of all results presented in this thesis. As the information provided here aims at Charmm users, a description of the individual keywords and parameters will not be given.

A.1 MD Simulation Set 1

Crystal Setup

T1-3,WB:
```
crystal define tetragonal 48.5 48.5 63.4 90.0 90.0 90.0
crystal build cutoff 16.5 noper 0
```

O1:
```
crystal define orthogonal 48.5 48.5 63.4 90.0 90.0 90.0
crystal build cutoff 16.5 noper 0
```

Energy Function

T1-3,O1,WB:
```
energy elec atom fswitch vdw vatom vfswitch cdie eps 1.0 -
  CUTNB 14.0 CTONNB 10.0 CTOFNB 13.0 WMIN 1.5 -
  cutim 16.0 imgfrq 50 inbfrq -1 -
  ewald pmewald kappa 0.38 order 6 fftx 48 ffty 48 fftz 64 spline
```

Dynamics Call

T1,O1,WB:
```
DYNAMICS CPT reSTART NSTEP 50000 TIMESTEP 0.001 -
  IHTFRQ 0 IEQFRQ 0 NTRFRQ 1000 -
  IPRFRQ 50 ISVFRQ 1000 NPRINT 50 NSAVC 50 NSAVV 50 -
  IMGFRQ 50 INBFRQ -1 IHBFRQ -1 IXTFRQ 500 -
  IUNREA 30 IUNWRI 31 IUNCRD 32 IUNVEL 33 KUNIT 34 -
  FIRSTT 300.0 FINALT 300.0 TEMINC 0.0 ECHECK 99999.0 -
  IASORS 0 IASVEL 1 ISCVEL 0 ICHECW 0 TWINDH 10.0 TWINDL -10.0 -
  PCONS PINT PMASS 500.0 PREF 1.0 PGAMMA 25.0 -
  HOOVER REFT 300.0 TMASS 2000.0 TBATH 300.0 TSTRUC 300.0
```

T2,T3:
```
DYNAMICS CPT reSTART NSTEP 50000 TIMESTEP 0.001 -
  IHTFRQ 0 IEQFRQ 0 NTRFRQ 1000 -
  IPRFRQ 50 ISVFRQ 1000 NPRINT 50 NSAVC 50 NSAVV 50 -
  IMGFRQ 50 INBFRQ -1 IHBFRQ -1 IXTFRQ 500 -
  IUNREA 30 IUNWRI 31 IUNCRD 32 IUNVEL -1 KUNIT -1 -
  FIRSTT 300.0 FINALT 300.0 TEMINC 0.0 ECHECK 999999.0 -
  IASORS 0 IASVEL 1 ISCVEL 0 ICHECW 0 TWINDH 10.0 TWINDL -10.0 -
  PCONS PINT PMASS 500.0 PREF 1.0 PGAMMA 25.0 -
  HOOVER REFT 300.0 TMASS 2000.0 TBATH 300.0 TSTRUC 300.0 QCOR 0.0
```

A.2 MD Simulation Set 2

All settings (except for the reference pressure) are the same for all trajectories of Set 2.

Crystal Setup

```
crystal define tetragonal 48.5 48.5 63.4 90.0 90.0 90.0
crystal build cutoff 16.5 noper 0
```

Energy Function

```
energy elec atom fshift vdw vatom vshift cdie eps 1.0 -
  CUTNB 14.0 CTONNB 10.0 CTOFNB 13.0 WMIN 1.5 -
  cutim 16.0 imgfrq 50 inbfrq -1 -
  ewald pmewald kappa 0.38 order 6 fftx 48 ffty 48 fftz 64 spline
```

Dynamics Call (Pressure Increase)

```
DYNAMICS CPT reSTART NSTEP 100000 TIMESTEP 0.001 -
  IHTFRQ 0 IEQFRQ 0 NTRFRQ 10000 -
  IPRFRQ 100000 ISVFRQ 100000 NPRINT 100 NSAVC 100 NSAVV 0 -
  IMGFRQ 50 INBFRQ 25 IHBFRQ 25 IXTFRQ 500 -
  IUNREA 30 IUNWRI 31 IUNCRD 32 IUNVEL -1 KUNIT -1 -
  FIRSTT 300.0 FINALT 300.0 TEMINC 0.0 ECHECK 999999.0 -
  IASORS 0 IASVEL 1 ISCVEL 0 ICHECW 0 TWINDH 10.0 TWINDL -10.0 -
  PCONS PINT PMASS 500.0 PREFI Pi PREFF Pi+1 PGAMMA 25.0 -
  HOOVER REFT 300.0 TMASS 2000.0 TBATH 300.0 TSTRUC 300.0 QCOR 0.0
```

P_i and $P_{i+1} = P_i + 1000$ are specified in units of [1 bar]

Dynamics Call (Constant Pressure)

```
DYNAMICS CPT reSTART NSTEP 100000 TIMESTEP 0.001 -
  IHTFRQ 0 IEQFRQ 0 NTRFRQ 111111 -
  IPRFRQ 100000 ISVFRQ 100000 NPRINT 100 NSAVC 100 NSAVV 100 -
  IMGFRQ 50 INBFRQ 25 IHBFRQ 25 IXTFRQ 500 -
  IUNREA 30 IUNWRI 31 IUNCRD 32 IUNVEL 33 KUNIT -1 -
  FIRSTT 300.0 FINALT 300.0 TEMINC 0.0 ECHECK 999999.0 -
  IASORS 0 IASVEL 1 ISCVEL 0 ICHECW 0 TWINDH 10.0 TWINDL -10.0 -
  PCONS PINT PMASS 500.0 PREF P_i PGAMMA 25.0 -
  HOOVER REFT 300.0 TMASS 2000.0 TBATH 300.0 TSTRUC 300.0 QCOR 0.0
```

P_i is specified in units of [1 bar]

A.3 Protonation States

The following protonation states were used throughout (using Charmm nomenclature):

(N terminus, patch `NTER`)
ALA THR SER THR LYS LYS LEU HIS LYS GLU PRO ALA THR LEU ILE LYS ALA ILE ASP GLY ASP THR
VAL LYS LEU MET TYR LYS GLY GLN PRO MET THR PHE ARG LEU LEU LEU VAL ASP THR PRO GLU THR
LYS HSE PRO LYS LYS GLY VAL GLU LYS TYR GLY PRO GLU ALA SER ALA PHE THR LYS LYS MET VAL
GLU ASN ALA LYS LYS ILE GLU VAL GLU PHE ASN LYS GLY GLN ARG THR ASP LYS TYR GLY ARG GLY
LEU ALA TYR ILE TYR ALA ASP GLY LYS MET VAL ASN GLU ALA LEU VAL ARG GLN GLY LEU ALA LYS
VAL ALA TYR VAL TYR LYS PRO ASN ASN THR HIS GLU GLN HIS LEU ARG LYS SER GLU ALA GLN ALA
LYS LYS GLU LYS LEU ASN ILE TRP SER GLU ASP ASN ALA ASP SER GLY GLN
(C terminus, patch `CTER`).

References

[1] J M Ziman. *Models of disorder*. Cambridge University Press, Cambridge, 1982.

[2] P M Chaikin and T C Lubensky. *Principles of condensed matter physics*. Cambridge University Press, Cambridge, UK, 1995.

[3] M Peyrard, editor. *Nonlinear Excitations in Biomolecules*. Springer, Berlin, 1995.

[4] C L Brooks III, M Karplus, and B M Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics*. Wiley, New York, 1988.

[5] T E Creighton. *Protein: Structure and Molecular Properties*. W H Freeman, New York, 1992.

[6] K E van Holde, W C Johnson, and P S Ho. *Principles of Physical Biochemistry*. Prentice Hall, New Jersey, 1998.

[7] R Jaenicke. Protein stability and molecular adaptation to extreme conditions. *Eur J Biochem*, 202:715–728, 1991.

[8] H Frauenfelder, P G Wolynes, and R H Austin. Biological Physics. *Rev Modern Physics*, 71:419–430, 1999.

[9] F H Arnold, P L Wintrode, K Miyazaki, and A Gershenson. How enzymes adapt: lessons from directed evolution. *Trends Biochem Sci*, 26:100–106, 2001.

[10] R M Daniel, R V Dunn, J L Finney, and J C Smith. The Role of Dynamics in Enzyme Activity. *Annu Rev Biophys Biomol Struct*, 32:69–92, 2003.

[11] V Daggett and A R Fersht. Is there a unifying mechanism for protein folding? *Trends Biochem Sci*, 28:18–25, 2003.

[12] J F Atkins and R Gesteland. The 22nd amino acid. *Science*, 296:1409–1410, 2002.

[13] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[14] W L Bragg and M F Perutz. The structure of haemoglobin. *Proc Roy Soc A*, 213:425–435, 1952.

[15] W L Bragg, E R Howells, and M F Perutz. The structure of haemoglobin II. *Proc Roy Soc A*, 222:33–44, 1954.

[16] M F Perutz. The structure of haemoglobin III. direct determination of the molecular transform. *Proc Roy Soc A*, 225:264–286, 1954.

[17] D W Green, V M Ingram, and M F Perutz. The structure of haemoglobin IV. sign determination by the isomorphous replacement method. *Proc Roy Soc A*, 225:287–307, 1954.

[18] E R Howells and M F Perutz. The structure of haemoglobin V. imidazole-methaemoglobin: a further check of the signs. *Proc Roy Soc A*, 225:308–314, 1954.

[19] W L Bragg and M F Perutz. The structure of haemoglobin VI. fourier projections on the 010 plane. *Proc Roy Soc A*, 225:315–329, 1954.

[20] M F Perutz, M G Rossmann, A F Cullis, H Muirhead, G Will, and A C T North. Structure of haemoglobin. a three-dimensional fourier synthesis at 5.5å resolution, obtained by X-ray analysis. *Nature*, 185:416–422, 1960.

[21] J C Kendrew, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, and D C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181:662–666, 1958.

[22] J C Kendrew. Architecture of a protein molecule. *Nature*, 182:764–767, 1958.

[23] K Wüthrich. *NMR of proteins and nucleic acids*. Wiley, New York, 1986.

[24] J A McCammon and S C Harvey. *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, 1988.

[25] A Amadei, A B M Linssen, and H J C Berendsen. Essential Dynamics of Proteins. *Proteins*, 17:412–425, 1993.

[26] B R Brooks, D Janežič, and M Karplus. Harmonic Analyis of Large Systems. *J Comp Chem*, 16:1522–1542, 1995.

[27] S Héry, D Genest, and J C Smith. Rigid-Body Motion and X-ray Diffuse Scattering in Crystalline Lysozyme. *J Mol Biol*, 279:303–318, 1998.

[28] S Hayward, A Kitao, and H J C Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins*, 27:425–437, 1997.

[29] A Kitao and N Gō. Investigating protein dynamics in collective coordinate space. *Curr Opin Struct Biol*, 9:164–169, 1999.

[30] K Moritsugu, O Miyashita, and A Kidera. Vibrational energy transfer in a protein molecule. *Phys Rev Lett*, 85:3970–3973, 2000.

[31] BAlberts, D Bray, J Lewis, M Raff, K Roberts, and J D Watson. *Molecular Biology of the Cell*. Garland Publishing, Inc., New York, 1994.

[32] D M Irwin and A C Wilson. Multiple cDNA sequences and the evolution of bovine stomach lysozyme. *J Biol Chem*, 264:11387–11393, 1989.

[33] J G Shewale, S K Sinha, and K Brew. Evolution of alpha-lactalbumins. the complete amino acid sequence of the alpha-lactalbumin from a marsupial (macropus rufogriseus) and corrections to regions of sequence in bovine and goat alpha-lactalbumins. *J Biol Chem*, 259:4947–4956, 1984.

[34] K Nitta and S Sugai. The evolution of lysozyme and alpha-lactalbumin. *Eur J Biochem*, 182:111–118, 1989.

[35] D E Koshland Jr. Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA*, 44:98–104, 1958.

[36] W N Lipscomb. Structure and catalysis of enzymes. *Annu Rev Biochem*, 52:17–34, 1983.

[37] G E Schulz. Induced-fit movements in adenylate kinases. *Faraday Discuss*, 93:85–93, 1992.

[38] I Schlichting, S C Almo, G Rapp, K Wilson, K Petratos, A Lentfer, A Wittinghofer, W Kabsch, E F Pai, G A Petsko, and R S Goody. Time-resolved X-ray crystallographic study of the conformational change in Ha-Ras p21 protein on GTP hydrolysis. *Nature*, 345:309–15, 1990.

[39] A Miyazawa, Y Fujiyoshi, and N Unwin. Structure and gating mechanism of acetylcholine receptor pore. *Nature*, 423:949–955, 2003.

[40] I Bosanac, T Michikawa, K Mikoshiba, and M Ikura. Structural insights into the regulatory mechanism of ip3 receptor. *Biochim Biophys Acta*, 1742:89–102, 2004.

[41] R D Vale and R A Milligan. The way things move: looking under the hood of molecular motor proteins. *Science*, 288:88–95, 2000.

[42] R H Austin, K W Beeson, L Eisenstein, H Frauenfelder, and I C Gunsalus. Dynamics of ligand binding to myoglobin. *Biochemistry*, 14:5355–73, 1975.

[43] H Frauenfelder, G A Petsko, and D Tsernoglou. Temperature-dependent X-ray diffraction as a probe of protein structural dynamics. *Nature*, 280:558–563, 1979.

[44] A Ansari, J Berendzen, S F Bowne, H Frauenfelder, I E Iben, T B Sauke, E Shyamsunder, and R D Young. Protein states and proteinquakes. *Proc Natl Acad Sci USA*, 82:5000–5004, 1985.

[45] R Elber and M Karplus. Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science*, 235:318, 1987.

[46] H Frauenfelder, S G Sligar, and P G Wolynes. The energy landscapes and motions of proteins. *Science*, 254:1598–1603, 1991.

[47] I Schlichting, J Berendzen, G N Phillips Jr, and R M Sweet. Crystal structure of photolysed carbonmonoxy-myoglobin. *Nature*, 371:808–812, 1994.

[48] W Doster, S Cusack, and W Petry. Dynamical transition of myoglobin revealed by inelastic neutron scattering. *Nature*, 337:754–756, 1989.

[49] F Parak, E N Frolov, A A Kononenko, R L Mossbauer, V I Goldanskii, and A B Rubin. Evidence for a correlation between the photoinduced electron transfer and dynamic properties of the chromatophore membranes from rhodospirillum rubrum. *FEBS Lett*, 117: 368–372, 1980.

[50] B F Rasmussen, A M Stock, D Ringe, and G A Petsko. Crystalline ribonuclease A loses function below the dynamical transition at 220 K. *Nature*, 357:423–424, 1992.

[51] A L Tournier, J Xu, and J C Smith. Translational hydration water dynamics drives the protein glass transition. *Biophys J*, 85:1871–1875, 2003.

[52] A L Tournier and J C Smith. Principal components of the protein dynamical transition. *Phys Rev Lett*, 91:208106, 2003.

[53] M Ferrand, A J Dianoux, W Petry, and G Zaccai. Thermal motions and function of bacteriorhodopsin in purple membranes: effects of temperature and hydration studied by neutron scattering. *Proc Natl Acad Sci USA*, 90:9668–9672, 1993.

[54] U Lehnert, V Réat, M Weik, G Zaccaï, and C Pfister. Thermal Motions in Bacteriorhodopsin at Different Hydration Levels Studied by Neutron Scattering: Correlation with Kinetics and Light-Induced Conformational Changes. *Biophys J*, 75:1945–1952, 1998.

[55] R M Daniel, J C Smith, M Ferrand, S Héry, R Dunn, and J L Finney. Enzyme activity below the dynamical transition at 220 K. *Biophys J*, 75:2504–2507, 1998.

[56] R V Dunn, V Reat, J L Finney, M Ferrand, J C Smith, and R M Daniel. Enzyme activity and dynamics: xylanase activity in the absence of fast anharmonic dynamics. *Biochem J*, 346:355–358, 2000.

[57] J C Smith, S Cusack, P Poole, and J L Finney. Direct measurement of hydration-related dynamic changes in lysozyme using inelastic neutron scattering spectroscopy. *J Biomol Struct Dyn*, 4:583–588, 1987.

[58] S Bone. Time-domain reflectrometry studies of water binding and structural flexibility in chymotrypsin. *Biochim Biophys Acta*, 916:128–134, 1987.

[59] J A Rupley and G Careri. Protein hydration and function. *Adv Protein Chem*, 41:37–172, 1991.

[60] P A Lind, R M Daniel, C Monk, and R V Dunn. Esterase catalysis of substrate vapour: enzyme activity occurs at very low hydration. *Biochim Biophys Acta*, 1702:103–110, 2004.

[61] J Fitter, R E Lechner, and N A Dencher. Picosecond molecular motions in bacteriorhodopsin from neutron scattering. *Biophys J*, 73:2126–2137, 1997.

[62] H Keller and P G Debrunner. Evidence for conformational and diffusional mean square displacements in frozen aqueous solution of oxymyoglobin. *Phys Rev Lett*, 45:68–71, 1980.

[63] M C Bellissent-Funel, J M Zanotti, and S H Chen. Slow dynamics of water molecules on the surface of a globular protein. *Faraday Discuss*, 103:281–294, 1996.

[64] R Abseher, H Schreiber, and O Steinhauser. The influence of a protein on water dynamics in its vicinity investigated by molecular dynamics simulation. *Proteins*, 25:366–378, 1996.

[65] S K Pal, J Peon, and A H Zewail. Biological water at the protein surface: dynamical solvation probed directly with femtosecond resolution. *Proc Natl Acad Sci USA*, 99:1763–1768, 2002.

[66] F Merzel and J C Smith. Is the first hydration shell of lysozyme of higher density than bulk water? *Proc Natl Acad Sci USA*, 99:5378–5383, 2002.

[67] M Marchi, F Sterpone, and M Ceccarelli. Water rotational relaxation and diffusion in hydrated lysozyme. *JACS*, 124:6787–6791, 2002.

[68] D I Svergun, S Richard, M H Koch, Z Sayers, S Kuprin, and G Zaccai. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc Natl Acad Sci USA*, 95:2267–2272, 1998.

[69] 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 Res*, 28:235–242, 2000.

[70] J Drenth. *Principles of protein X-ray crystallography*. Springer, Berlin, 1995.

[71] E Franchon, E Geissler, J-L Hodeau, J-R Regnard, and P A Timmins, editors. *Structure and Dynamics of Biomolecules: Neutron and Synchrotron Radiation for Condensed Matter Studies*. Oxford University Press, Oxford, UK, 2000.

[72] M S Smyth and J H J Martin. x Ray crystallography. *J Clin Pathol: Mol Pathol*, 53:8–14, 2000.

[73] S H Chong, Y Joti, A Kidera, N Gō, A Ostermann, A Gassmann, and F Parak. Dynamical transition of myoglobin in a crystal: comparative studies of X-ray crystallography and mossbauer spectroscopy. *Eur Biophys J*, 30:319–329, 2001.

[74] G N Phillips Jr, J P Fillers, and C Cohen. Motions of tropomyosin. Crystal as metaphor. *Biophys J*, 32:485–502, 1980.

[75] J Doucet and J P Benoit. Molecular dynamics studied by analysis of X-ray diffuse scattering from lysozyme crystals. *Nature*, 325:643–646, 1987.

[76] D L D Caspar, J Clarage, D M Salunke, and M Clarage. Liquid-like movements in crystalline insulin. *Nature*, 332:659–662, 1988.

[77] I D Glover, G W Harris, J R Helliwell, and D S Moss. The Variety of X-ray Diffuse Scattering from Macromolecular Crystals and its Respective Components. *Acta Cryst*, pages 960–968, 1991.

[78] S Chacko and G N Phillips Jr. Diffuse x-ray scattering from tropomyosin crystals. *Biophys J*, 61:1256–1266, 1992.

[79] J B Clarage, M S Clarage, W C Phillips, R M Sweet, and D L Caspar. Correlations of atomic movements in lysozyme crystals. *Proteins*, 12:145–157, 1992.

[80] P Faure, A Micu, D Perahia, J Doucet, J C Smith, and J P Benoit. Correlated Motions and X-ray Diffuse Scattering in Lysozyme. *Nature Struct Biol*, 1:124–128, 1994.

[81] P Faure, J Pérez, J Doucet, and J P Benoit. X-ray diffuse scattering and molecular dynamics in proteins. *J Physique IV*, 4:293–298, 1994.

[82] T Thüne and J Badger. Thermal diffuse X-ray scattering and its contribution to understanding protein dynamics. *Prog Biophys Mol Biol*, 63:251–276, 1995.

[83] J P Benoit and J Doucet. Diffuse scattering in protein crystallography. *Quart Rev Biophys*, 28:131–169, 1995.

[84] J B Clarage and G N Phillips Jr. Analysis of Diffuse Scattering and Relation to Molecular Motion. In C W Carter and R M Sweet, editors, *Methods in Enzymology, Vol. 277*. Academic Press, New York, 1997.

[85] D S Moss, G W Harris, and A Wostrack. Diffuse X-ray scattering from molecular crystals. In E Franchon, E Geissler, J-L Hodeau, J-R Regnard, and P A Timmins, editors, *Structure and Dynamics of Biomolecules: Neutron and Synchrotron Radiation for Condensed Matter Studies*. Oxford University Press, Oxford, UK, 2000.

[86] D S Moss, G W Harris, A Wostrack, and C Sansom. Diffuse X-Ray scattering from molecular crystals. *Cryst Rev*, 9:229–277, 2003.

[87] B Halle. Flexibility and packing in proteins. *Proc Natl Acad Sci USA*, 99:1274–1279, 2002.

[88] J Hajdu and I Andersson:. Fast crystallography and time-resolved structures. *Annu Rev Biophys Biomol Struct*, 22:467–98, 1993.

[89] I Schlichting. Reactions in crystals and time-resolved crystallography. In E Franchon, E Geissler, J-L Hodeau, J-R Regnard, and P A Timmins, editors, *Structure and Dynamics of Biomolecules: Neutron and Synchrotron Radiation for Condensed Matter Studies*. Oxford University Press, Oxford, UK, 2000.

[90] J Hajdu, R Neutze, T Sjögren, K Edman, A Szöke, R C Wilmouth, and C M Wilmot. Analyzing protein functions in four dimensions. *Nature Struct Biol*, 7:1006–1012, 2000.

[91] F Schotte, M Lim, T A Jackson, A V Smirnov, J Soman, J S Olson, G N Phillips Jr, M Wulff M, and P A Anfinrud. Watching a protein as it functions with 150-ps time-resolved X-ray crystallography. *Science*, 300:1944–1947, 2003.

[92] J Hajdu and L N Johnson. Progress with Laue Diffraction Studies on Protein and Virus Crystals. *Biochemistry*, 29:1669–1678, 1990.

[93] S Göttlicher. Der Beitrag der thermisch diffusen Streustrahlung zur Intensität der Röntgeninterferenzen und die Elektronendichteverteilung im NaCl. *Acta Cryst B*, 24:122–129, 1968.

[94] W Cochran. The Correction of Measured Structure Factors for Thermal Diffuse Scattering. *Acta Cryst A*, 25:95–101, 1996.

[95] S W Lovesey. *Theory of Neutron Scattering from Condensed Matter*. Oxford University Press, Oxford, UK, 1984.

[96] A R Kolatkar, J B Clarage, and G N Phillips Jr. Analysis of Diffuse Scattering from Yeast Initiator tRNA Crystals. *Acta Cryst D*, 50:210–218, 1994.

[97] M E Wall, J B Clarage, and G N Phillips Jr. Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering. *Structure*, 5:1599–1612, 1997.

[98] J Pérez, P Faure, and J-P Benoit. Molecular Rigid-Body Displacements in a Tetragonal Lysozyme Crystal Confirmed by X-ray Diffuse Scattering. *Acta Cryst D*, 52:722–729, 1996.

[99] J B Clarage, T Romo, B K Andrews, B M Pettitt, and G N Phillips Jr. A sampling problem in molecular dynamics simulations of macromolecules. *Proc Natl Acad Sci USA*, 92:3288–3292, 1995.

[100] L Meinhold and J C Smith. Fluctuations and correlations in crystalline protein dynamics: A simulation analysis of Staphylococcal nuclease. *Biophys J*, 88:2554–2563, 2005.

[101] M E Wall, S E Ealick, and S M Gruner. Three-dimensional diffuse x-ray scattering from crystals of Staphylococcal nuclease. *Proc Natl Acad Sci USA*, 94:6180–6184, 1997.

[102] D Frenkel and B Smit. *Understanding Molecular Simulation*. Academic Press, San Diego, 2002.

[103] P W Tucker, E E Hazen Jr, and F A Cotton. Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. i. isolation; physical and enzymatic properties. *Mol Cell Biochem*, 22:67–77, 1978.

[104] P W Tucker, E E Hazen Jr, and F A Cotton. Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. ii. solution studies of the nucleotide binding site and the effects of nucleotide binding. *Mol Cell Biochem*, 23:3–16, 1979.

[105] P W Tucker, E E Hazen Jr, and F A Cotton. Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. iii. correlation of the three-dimensional structure with the mechanisms of enzymatic action. *Mol Cell Biochem*, 23:67–86, 1979.

[106] P W Tucker, E E Hazen Jr, and F A Cotton. Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. iv. the nuclease as a model for protein folding. *Mol Cell Biochem*, 23:131–41, 1979.

[107] D Shortle, Y Wang Y, J R Gillespie, and J O Wrabl. Protein folding for realists: a timeless phenomenon. *Protein Sci*, 5:991–1000, 1996.

[108] A D Gruia. *Ionic interactions in protein folding and function: a case study of Staphylococcal nuclease and Halorhodopsin*. PhD thesis, Heidelberg University, 2004.

[109] M E Tuckerman and G J Martyna. Understanding modern molecular dynamics: Techniques and applications. *J Phys Chem B*, 104:159–178, 2000.

[110] M P Allen and D J Tildesley. *Computer simulation of liquids*. Clarendon Press, Oxford, 2001.

[111] A R Leach. *Molecular Modelling: Principles and Applications*. Pearson Education Ltd., Harlow, 2001.

[112] T Schlick. *Molecular Modeling and Simulation*. Springer, New York, 2002.

[113] B R Brooks, R E Bruccoleri, B D Olafson, D J States, S Swaminathan, and M Karplus. `CHARMM`: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comp Chem*, 4:187–217, 1983.

[114] A D MacKerell, D Bashford, M Bellot, J R Dunbrack, R L Evenseck, M J Field, S Fischer, J Gao, H Guo, S Ha, D Joseph, L Kuchnir, K Kuczera, F T K Lau, C Mattos, S Michnick, T Ngo, D T Nguyen, B Prodhom, I W E Reiher, B Roux, M Schlenkrich, J C Smith,

R Stote, J Straub, M Watanabe, J Wiorkiewicz-Kuczera, J Yin, and M Karplus. All-atom empirical potential for molecular modelling and dynamics studies of proteins. *J Phys Chem B*, 102:3586–3616, 1998.

[115] W L Jorgensen, J Chandrasekhar, J D Madura, R W Impey, and M L Klein. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*, 79:926–935, 1983.

[116] P P Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Ann Phys*, 64:253–287, 1921.

[117] U Essmann, L Perera, M L Berkowitz, T Darden, H Lee, and L G Pedersen. A smooth partice mesh Ewald method. *J Chem Phys*, 103:8577–8593, 1995.

[118] W Nolting. *Grundkurs Theoretische Physik*. Zimmermann-Neufang, Ulmen, 1989.

[119] H Goldstein. *Klassische Mechanik*. AULA-Verlag, Wiesbaden, 1991.

[120] H C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys*, 72:2384–2393, 1980.

[121] S Nosé and M L Klein. Constant pressure molecular dynamics for molecular systems. *Mol Physics*, 50:1055–1076, 1983.

[122] W G Hoover. Canonical dynamics: equilibrium phase-space distributions. *Phys Rev A*, 31:1695–1697, 1985.

[123] G J Martyna, D J Tobias, and M L Klein. Constant pressure molecular dynamics algorithms. *J Chem Phys*, 101:4177–4189, 1994.

[124] R W Hockney and J W Eastwood. *Computer Simulation using Particles*. McGraw-Hill, New York, 1981.

[125] M J Legg. *Protein crystallography: new approaches to X-ray data collection, direct space refinement and model building*. PhD thesis, Texas Agricultural and Mechanical University, 1977.

[126] T R Hynes and R O Fox. The crystal structure of Staphylococcal nuclease refined at 1.7Å resolution. *Proteins*, 10:92, 1991.

[127] W H Zachariasen. *Theory of X-Ray Diffraction in Crystals*. John Wiley & Sons, Inc., New York, 1946.

[128] H Ibach and H Lüth. *Festkörperphysik*. Springer, Heidelberg, 1995.

[129] J M Cowley. *Diffraction Physics*. Elsevier, Amsterdam, 1995.

[130] N W Ashcroft and N D Mermin. *Solid State Physics*. Brooks/Coole Thomson Learning, Singapore, 2001.

[131] J A Ibers and W C Hamilton, editors. *International Tables for X-ray Crystallography Vol IV*. The Kynoch Press, Birmingham, 1974.

[132] L E Reichl. *A Modern Course in Statistical Physics*. Wiley-Interscience, New York, 1998.

[133] I Waller. *Theoretische Studien zur Interferenz- und Dispersiontheorie der Röntgenstrahlen.* PhD thesis, Uppsala University, 1925.

[134] W Cochran. Lattice Vibrations. *Rep Prog Phys*, 26:1–45, 1963.

[135] L Meinhold, S Lammers, T Becker, and J C Smith. Convergence properties of x-ray scattering calculated from protein crystal molecular dynamics simulation. *Physica B*, 350: 127–131, 2004.

[136] P H Hünenberger, A E Mark, and W F van Gunsteren. Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J Mol Biol*, 252:492–503, 1995.

[137] L S D Caves, J D Evanseck, and M Karplus. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Sci*, 7:649–666, 1998.

[138] P Eastman, M Pellegrini, and S Doniach. Protein flexibility in solution and in crystals. *J Chem Phys*, 110:10141–10152, 1999.

[139] T Ichiye and M Karplus. Collective Motions in Proteins: A Covariance Analysis of Atomic Fluctuations in Molecular Dynamics and Normal Mode Simulations. *Proteins*, 11:205–217, 1991.

[140] R Walser, P H Hünenberger, and W F van Gunsteren. Molecular Dynamics Simulations of a Double Unit Cell in a Protein Crystal: Volume Relaxation at Constant Pressure and Correlation of Motions Between Two Unit Cells. *Proteins*, 48:327–340, 2002.

[141] A E García and G Hummer. Conformational dynamics of cytochrome c: correlation to hydrogen exchange. *Proteins*, 36:175–191, 1999.

[142] S A Showalter and K B Hall. A functional role for correlated motion in the N-terminal RNA-binding domain of human U1A protein. *J Mol Biol*, 322:533–542, 2002.

[143] C Bossa, M Anselmi, D Roccatano, A Amadei, B Vallone, M Brunori, and A Di Nola. Extended molecular dynamics simulation of the carbon monoxide migration in sperm whale myoglobin. *Biophys J*, 86:3855–3862, 2004.

[144] A E García, J A Krumhansl, and H Frauenfelder. Variations on a theme by Debye and Waller: from simple crystals to proteins. *Proteins*, 29:153–160, 1997.

[145] A M Micu and J C Smith. SERENA: a program for calculating X-ray diffuse scattering intensities from molecular dynamics trajectories. *Comp Phys Comm*, 91:331–338, 1995.

[146] M Karplus and T Ichiye. Comment on a Fluctuation and Cross Correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations. *J Mol Biol*, 263:120–122, 1996.

[147] R Abseher and M Nilges. Are there non-trivial dynamics cross-correlations in proteins? *J Mol Biol*, 279:911–920, 1998.

[148] Y Zhou, M Cook, and M Karplus. Protein Motions at Zero-Total Angular Momentum: The Importance of Long-Range Correlations. *Biophys J*, 79:2902–2908, 2000.

[149] V Schomaker and K N Trueblood. On the rigid-body motion of molecules in crystals. *Acta Cryst B*, 24:63–76, 1968.

[150] R Diamond. On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Cryst A*, 46:425–435, 1990.

[151] J Kuriyan and W I Weis. Rigid protein motion as a model for crystallographic temperature factors. *Proc Natl Acad Sci USA*, 88:2773–2777, 1991.

[152] B Howlin, D S Moss, and G W Harris. Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. *Acta Cryst A*, 45:851–861, 1989.

[153] M D Winn, M N Isupov, and G N Murshudov. Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Cryst D*, 57:122–133, 2001.

[154] A Kidera and N Gō. Refinement of protein dynamic structure: Normal mode refinement. *Proc Natl Acad Sci USA*, 87:3718–3722, 1990.

[155] Y Joti, M Nakasako, A Kidera, and N Gō. Nonlinear temperature dependence of the crystal structure of lysozyme: correlation between coordinate shifts and thermal factors. *Acta Cryst D*, 58:1421–1432, 2002.

[156] M M Tirion. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys Rev Lett*, 77:1905–1908, 1996.

[157] I Bahar, A R Atilgan, and B Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2:173–181, 1997.

[158] S Kundu, J S Melton, , D C Sorensen, and G N Phillips Jr. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys J*, 83:723–732, 2002.

[159] P J Loll and E E Lattman. Active site mutant Glu-43→Asp in Staphylococcal nuclease displays nonlocal structural changes. *Biochemistry*, 29:6866–6873, 1990.

[160] L Meinhold and J C Smith. Correlated dynamics determining x-ray diffuse scattering from a crystalline protein revealed by molecular dynamics simulation. *Phys Rev Lett*, .

[161] L Meinhold and J C Smith. Collective dynamics in protein crystals: a simulation analysis of X-ray diffuse scattering by Staphylococcal nuclease. *to be submitted to Biophys J*, .

[162] A H Narten, M D Danford, and H A Levy. X-ray diffraction study of liquid water in the temperature range 4-200°C. *Disc Faraday Soc*, 43:97–107, 1967.

[163] A K Soper. The radial distribution functions of water and ice from 220 to 673 K and at pressures up to 400 MPa. *Chem Phys*, 258:121–137, 2000.

[164] K Mizuguchi, A Kidera, and N Gō. Collective Motions in Proteins Investigated by X-Ray Diffuse Scattering. *Proteins*, 18:34–48, 1994.

[165] J Kuriyan, G A Petsko, R M Levy, and M Karplus. Effect of anisotropy and anharmonicity on protein crystallographic refinement. an evaluation by molecular dynamics. *J Mol Biol*, 190:227–254, 1986.

[166] D Vitkup, D Ringe, M Karplus, and G A Petsko. Why Protein R-factors Are So Large: A Self-Consistent Analysis. *Proteins*, 46:345–354, 2002.

[167] A Kitao, S Hayward, and N Gō. Energy landscape of a native protein: Jumping-among-minima model. *Proteins*, 33:496–517, 1998.

[168] B Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys Rev E*, 62:8438–8448, 2000.

[169] R Elber and M Karplus. Low-Frequency Modes in Proteins: Use of the Effective-Medium Approximation to Interpret the Fractal Dimension Observed in Electron-Spin Relaxation Measurements. *Phys Rev Lett*, 56:394–397, 1986.

[170] D ben Avraham. Vibrational normal-mode spectrum of globular proteins. *Phys Rev B*, 47:14559–14560, 1993.

[171] M Nöllmann and P Etchegoin. Universal low-frequency vibrations in proteins from a simple interaction potential. *Phys Rev E*, 60:4593–4596, 1999.

[172] G Giraud, J Karolin, and K Wynne. Low-Frequency Modes of Peptides and Globular Proteins in Solution Observed by Ultrafast OHD-RIKES Spectroscopy. *Biophys J*, 85: 1903–1913, 2003.

[173] E Balog, T Becker, M Oettl, R Lechner, R Daniel, J L Finney, and J C Smith. Direct Determination of Vibrational Density of States Change on Ligand Binding to a Protein. *Phys Rev Lett*, 93:028103, 2004.

[174] J C Smith, S Cusack, B Tidor, and M Karplus. Inelastic neutron scattering analysis of low-frequency motions in proteins: Harmonic and damped harmonic models of bovine pancreatic tryspin inhibitor. *J Chem Phys*, 93:2974–2991, 1990.

[175] The MathWorks, Inc. *http://www.mathworks.com/*.

[176] S Hayward, A Kitao, F Hirata, and N Gō. Effect of Solvent on Collective Motions in Globular Protein. *J Mol Biol*, 234:1207–1217, 1993.

[177] P Durand, G Trinquier, and Y-H Sanejouand. A New Approach for Determining Low-Frequency Normal Modes in Macromoleculs. *Biopolymers*, 34:759–771, 1994.

[178] F Tama, F X Gadea, O Marques, and Y H Sanejouand. Building-Block Approach for Determining Low-Frequency Normal Modes of Macromolecules. *Proteins*, 41:1–7, 2000.

[179] B Hess. Convergence of sampling in protein simulations. *Phys Rev E*, 62:031910, 2002.

[180] L Meinhold and J C Smith. Pressure-dependent transition in protein dynamics at ≈4 kbar revealed by molecular dynamics simulation. *submitted to Phys Rev E*, .

[181] F G Parak and G U Nienhaus. Myoglobin, a paradigm in the study of protein dynamics. *Chem Phys Chem*, 3:249–254, 2002.

[182] F G Parak. Physical aspects of protein dynamics. *Rep Prog Phys*, 66:103–129, 2003.

[183] K Heremans and L Smeller. Protein structure and dynamics at high pressure. *Biochim Biophys Acta*, 1386:353–370, 1998.

[184] E Paci. High pressure simulations of biomolecules. *Biochim Biophys Acta*, 1595:185–200, 2002.

[185] S M Gruner. Soft materials and biomaterials under pressure. In A Katrusiak and P McMillan, editors, *High-Pressure Crystallography*. Kluwer Academic Publisher, Netherlands, 2004.

[186] T Yamato, J Higo, Y Seno, and N Gō. Conformational deformation in deoxymyoglobin by hydrostatic pressure. *Proteins*, 16:327–340, 1993.

[187] E Paci and M Marchi. Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure. *Proc Natl Acad Sci USA*, 93: 11609–11614, 1996.

[188] K Akasaka, T Tezuka, and H Yamada. Pressure-induced changes in the folded structure of lysozyme. *J Mol Biol*, 271:671–678, 1997.

[189] M Refaee, T Tezuka, K Akasaka, and M P Williamson. Pressure-dependent changes in the solution structure of hen egg-white lysozyme. *J Mol Biol*, 327:857–865, 2003.

[190] B Wroblowski, J F Diaz, K Heremans, and Y Engelborghs. Molecular mechanisms of pressure induced conformational changes in bpti. *Proteins*, 25:446–455, 1996.

[191] P Urayama, G N Phillips Jr, and S M Gruner. Probing substates in sperm whale myoglobin using high-pressure crystallography. *Structure*, 10:51–60, 2002.

[192] R Kitahara, S Yokoyama, and K Akasaka. NMR snapshots of a fluctuating protein structure: ubiquitin at 30 bar–3 kbar. *J Mol Biol*, 347:277–285, 2005.

[193] M W Lassalle, H Yamada, and K Akasaka. The pressure-temperature free energy-landscape of staphylococcal nuclease monitored by ^{1}h nmr. *J Mol Biol*, 298:293–302, 2000.

[194] J Woenckhaus, R Kohling, P Thiyagarajan, K C Littrell, S Seifert, C A Royer, and R Winter. Pressure-jump small-angle x-ray scattering detected kinetics of staphylococcal nuclease folding. *Biophys J*, 80:1518–1523, 2001.

[195] H Seemann, R Winter, and C A Royer. Volume, expansivity and isothermal compressibility changes associated with temperature and pressure unfolding of Staphylococcal nuclease. *J Mol Biol*, 307:1091–1102, 2001.

[196] A Paliwal, D Asthagiri, D P Bossev, and M E Paulaitis. Pressure denaturation of staphylococcal nuclease studied by neutron small-angle scattering and molecular simulation. *Biophys J*, 87:3479–3492, 2004.

[197] W Doster and R Gebhardt. High pressure - unfolding of myoglobin studied by dynamics neutron scattering. *Chem Phys*, 292:383–387, 2003.

[198] M W Lassalle, H Li, H Yamada, K Akasaka, and C Redfield. Pressure-induced unfolding of the molten globule of all-ala alpha-lactalbumin. *Protein Sci*, 12:66–72, 2003.

[199] Y O Kamatari, R Kitahara, H Yamada, S Yokoyama, and K Akasaka. High-pressure NMR spectroscopy for characterizing folding intermediates and denatured states of proteins. *Methods*, 34:133–143, 2004.

[200] D B Kitchen, L H Reed, and R M Levy. Molecular dynamics simulation of solvated protein at high pressure. *Biochemistry*, 31:10083–10093, 1992.

[201] G Hummer, S Garde, A E García, M E Paulaitis, and L R Pratt. The pressure dependence of hydrophobic interactions is consistent with the observed pressure denaturation of proteins. *Proc Natl Acad Sci USA*, 95:1552–1555, 1998.

[202] R M Brunne and W F van Gunsteren. Dynamical properties of bovine pancreatic trypsin inhibitor from a molecular dynamics simulation at 5000 atm. *FEBS Lett*, 323:215–217, 1993.

[203] A Katrusiak and Z Dauter. Compressibility of lysozyme protein crystals by X-ray diffraction. *Acta Cryst D*, 52:607–608, 1996.

[204] F H O Osterberg. *Induced changes in the diffuse X-ray scattering background from protein crystals (high pressure crystallography).* PhD thesis, Princeton University, 1996.

[205] R Walser, A E Mark, and W F van Gunsteren. On the temperature and pressure dependence of a range of properties of a type of water model commonly used in high-temperature protein unfolding simulations. *Biophys J*, 78:2752–2760, 2000.

[206] M Marchi and K Akasaka. Simulation of hydrated BPTI at high pressure: Changes in hydrogen bonding and its relation with NMR experiments. *J Phys Chem B*, 105:711–714, 2001.

[207] A L Tournier. *The role of the solvent in the protein dynamical transition.* PhD thesis, Heidelberg University, 2003.

[208] C E Kundrot and F M Richards. Crystal structure of hen egg-white lysozyme at a hydrostatic pressure of 1000 atmospheres. *J Mol Biol*, 193:157–170, 1987.

[209] D P Kharakoz. Protein compressibility, dynamics, and pressure. *Biophys J*, 79:511–525, 2000.

[210] J D Smith, C D Cappa, K R Wilson, B M Messer, R C Cohen, and R J Saykally. Energetics of hydrogen bond network rearrangements in liquid water. *Science*, 306:851–853, 2004.

[211] A Nilsson, P Wernet, D Nordlund, U Bergmann, M Cavalleri, M Odelius, H Ogasawara, L A Näslund, T K Hirsch, L Ojamäe, P Glatzel, and L G M Pettersson. Comment on 'Energetics of hydrogen bond network rearrangements in liquid water'. *Science*, 308:793a, 2005.

[212] J D Smith, C D Cappa, K R Wilson, B M Messer, R C Cohen, and R J Saykally. Response to comment on 'Energetics of hydrogen bond network rearrangements in liquid water'. *Science*, 308:793b, 2005.

[213] M Chaplin. *http://www.lsbu.ac.uk/water/.*

[214] S J Suresh and V M Naik. Hydrogen bond thermodynamic properties of water from dielectric constant data. *J Chem Phys*, 113:9727–9732, 2000.

[215] L C Malacarne, R S Mendes, I T Pedron, and E K Lenzi. Nonlinear equation for anomalous diffusion: Unified power-law and stretched exponential exact solution. *Phys Rev E*, 63: 030101(R), 2001.

[216] P W Fenimore, H Frauenfelder, B H McMahon, and F G Parak. Slaving: solvent fluctuations dominate protein dynamics and functions. *Proc Natl Acad Sci USA*, 99:16047–16051, 2002.

[217] G Caliskan, D Mechtani, J H Roh, A Kisliuk, A P Sokolov, S Azzam, M T Cicerone, S Lin-Gibson, and I Peral. Protein and solvent dynamics: How strongly are they coupled? *J Chem Phys*, 121:1978–1983, 2004.

[218] M R Reddy and M Berkowitz. Structure and dynamics of high-pressure TIP4P water. *J Chem Phys*, 87:6682–6686, 1987.

[219] J Brodholt and B Wood. Simulations of the structure and thermodynamic properties of water at high pressures and temperatures. *J Geophys Res*, 98:519–536, 1993.

[220] F W Starr, F Sciortino, and H E Stanley. Dynamics of simulated water under pressure. *Phys Rev E*, 60:6757–6768, 1999.

[221] B L de Groot and H Grubmüller. Water permeation across biological membranes: Mechanism and dynamics of aquaporin-1 and GlpF. *Science*, 294:2353–2357, 2001.

[222] S Hayashi, E Tajkhorshid, and K Schulten. Molecular dynamics simulation of bacteriorhodopsin's photoisomerization using ab initio forces for the excited chromophore. *Biophys J*, 85:1440–1449, 2003.

[223] A N Bondar, M Elstner, S Suhai, J C Smith, and S Fischer. Mechanism of primary proton transfer in bacteriorhodopsin. *Structure*, 12:1281–1288, 2004.

[224] K Achterhold and F G Parak. Protein dynamics: determination of anisotropic vibrations at the haem iron of myoglobin. *J Phys Condens Matter*, 15:1683–1692, 2003.

[225] C L Caylor, S Speziale, S Kriminski, T Duffy, C-S Zha, and R E Thorne. Measuring the elastic properties of protein crystals by brillouin scattering. *J Cryst Grth*, 232:498, 2001.

[226] P A Kroon and A Vos. Thermal diffuse scattering for molecular crystals: errors in x-ray diffraction intensities and atomic parameters. *Acta Cryst A*, 35:675–684, 1979.

[227] A M Micu, D Durand, M Quilichini, M J Field, and J C Smith. Collective vibration in crystalline L-alanine. *J Phys Chem*, 99:5645–5657, 1995.

[228] I Schlichting, J Berendzen, K Chu, A M Stock, S A Maves, D E Benson, R M Sweet, D Ringe, G A Petsko, and S G Sligar. The catalytic pathway of cytochrome p450cam at atomic resolution. *Science*, 287:1615–1622, 2000.

[229] B Halle. Biomolecular cryocrystallography: structural changes during flash-cooling. *Proc Natl Acad Sci USA*, 101:4793–4798, 2004.

[230] K V Dunlop, R T Irvin, and B Hazes. Pros and cons of cryocrystallography: should we also collect a room-temperature data set? *Acta Cryst D*, 61:80–87, 2005.

[231] J Hajdu. Single-molecule X-ray diffraction. *Curr Opin Struct Biol*, 10:569–573, 2000.

[232] S W Smye, J M Chamberlain, A J Fitzgerald, and E Berry. The interaction between terahertz radiation and biological tissue. 46:R101–R112, 2001.

[233] A G Markelz, A Roitberg, and E J Heilweil. Pulsed terahertz spectroscopy of DNA, bovine serum albumin and collagen between 0.1 and 2.0 THz. *Chem Phys Lett*, 320:42–48, 2000.

Herewith I declare that this thesis is my own work and that I cited all references and help used. I consent to this thesis being deposited in and distributed by the Library of the Ruperto Carola University.

Lars Meinhold

Heidelberg
August 2005