Protein Dynamics in Crowded Systems : Cell Thermal Stability and Protein Complexation

Daniele Di Bari

▶ **To cite this version:**

Daniele Di Bari. Protein Dynamics in Crowded Systems : Cell Thermal Stability and Protein Complexation. Physics [physics]. Université Grenoble Alpes [2020-..]; Università degli studi (Pérouse, Italie), 2022. English. NNT : 2022GRALY037 . tel-03813514

HAL Id: tel-03813514

https://theses.hal.science/tel-03813514

Submitted on 13 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITE GRENOBLE ALPES

préparée dans le cadre d'une cotutelle entre l'*Université Grenoble Alpes* et *l'Università degli Studi di Perugia*

Spécialité : **Physique pour les Sciences du Vivant**

Arrêté ministériel : le 25 mai 2016

Présentée par

Daniele DI BARI

Thèse dirigée par **Judith PETERS**, Université Grenoble Alpes
et **Alessandro PACIARONI**, Università degli Studi di Perugia
codirigée par **Fabio STERPONE**, CNRS

préparée au sein des **Laboratoires :**
Laboratoire Interdisciplinaire de Physique, Univ. Grenoble Alpes
Dipartimento di Fisica e Geologia, Univ. di Perugia
Laboratoire de Biochimie Théorique, CNRS
Institut Laue-Langevin

dans **les Écoles Doctorales : Physique**, Univ. Grenoble Alpes **et Scienza e Tecnologia per la Fisica e la Geologia**, Univ. di Perugia

Dynamique des protéines en milieu encombré : *stabilité thermique des cellules et complexation des protéines*

Thèse soutenue publiquement le **15/07/2022**, devant le jury composé de :
Monsieur, Andrea, ORECCHINI — Présidente du jury
Professeur associé, Università degli Studi di Perugia, Examinateur
Madame, Irina, MIHALCESCU
Professeure, Université Grenoble Alpes, Examinatrice
Monsieur, Hans, GEISELMANN
Professeur, Université Grenoble Alpes, Examinateur
Monsieur, Fabio, BRUNI
Professeur, Università Roma Tre, Examinateur
Monsieur, Cristiano, DE MICHELE
Professeur associé, Sapienza Università di Roma, Examinateur
Madame, Victoria, GARCIA SAKAI
Chargée de recherche HDR, ISIS (STFC), Rapportrice
Monsieur, Alessandro, PACIARONI
Professeur associé, Università degli Studi di Perugia, Directeur de thèse
Madame, Judith, PETERS
Professeure, Université Grenoble Alpes, Directrice de thèse

Natura in minima maxima.

"Nature is the greatest in the smallest things"

- LATIN PROVERB -

Abstract

The work presented in this thesis aims at shedding some microscopic insights into thermal stability of bacterial cells (denaturation and cell growth).

The thesis is divided into six chapters, starting from the first one which contains a presentation of the temperature effects on living matter, introduces the topic of thermal response of bacteria to high temperature.

The second chapter is devoted to the theoretical framework of neutron scattering in condensed matter physics, whose formalism is introduced leading to the main equations for the relevant observables. Elastic (EINS), quasi-elastic (QENS) and inelastic neutron scattering are also defined. In the realm of EINS, the Gaussian approximation and the non-Gaussian behavior due to anharmonicity and dynamical heterogeneity are discussed, and different models for the heterogeneity are considered. In the realm of QENS it is derived how to decompose the spectrum to provide information of motions of atoms at different length and time scales: rigid (global), confined diffusive (as e.g., the 'cage diffusive' methyl groups hydrogen motions) and internal vibrations.

Computational methods used in the thesis are described in the third chapter. It contains the explanation of standard molecular dynamics techniques, with focus on protein simulations. Two scenarios are described: full atomistic simulations and coarse grain simulations, including the approach provided by the Lattice Boltzmann method.

Chapter 4 present the main result of this work: through the use of quasi-elastic neutron scattering combined with molecular dynamics simulations, it has been shown that there is a strong slow down of the global diffusion in the cytoplasm starting just below cell death temperature supporting the idea of protein unfolding as part of the irriversible denaturation process. Surprisingly though the fraction of unfolded proteins is only around 5% but it is the effect this causes of the whole proteome and its behaviour, that affects cell viability. No catastrophic denaturation of the proteome occurs at the cell death.

The last to chapters were devoted to two additional works both dealing with the effects of protein-ligand complexation on protein dynamics, in which neutron scattering techniques and molecular dynamics simulations are always coupled.

Résumé en français

Le travail présenté dans cette thèse vise à apporter des informations microscopiques sur la stabilité thermique des cellules bactériennes (dénaturation et croissance cellulaire).

La thèse est divisée en six chapitres, à partir du premier qui contient une présentation des effets de la température sur la matière vivante, introduit le thème de la réponse thermique des bactéries à haute température.

Le deuxième chapitre est consacré au cadre théorique de la diffusion neutronique en physique de la matière condensée, dont le formalisme est introduit conduisant aux principales équations pour les observables pertinentes. La diffusion élastique (EINS), quasi-élastique (QENS) et inélastique des neutrons est également définie. Dans le domaine de l'EINS, l'approximation gaussienne et le comportement non gaussien dû à l'anharmonicité et à l'hétérogénéité dynamique sont discutés, et différents modèles d'hétérogénéité sont considérés. Dans le domaine de QENS, il est dérivé de la manière de décomposer le spectre pour fournir des informations sur les mouvements des atomes à différentes échelles de longueur et de temps : rigide (global), diffusif confiné (comme par exemple, les mouvements d'hydrogène des groupes méthyle "diffusifs en cage") et vibrations interne.

Les méthodes de calcul utilisées dans la thèse sont décrites dans le troisième chapitre. Il contient l'explication des techniques standard de dynamique moléculaire, en mettant l'accent sur les simulations de protéines. Deux scénarios sont décrits : des simulations atomistiques complètes et des simulations à gros grains, y compris l'approche fournie par la méthode Lattice Boltzmann.

Le chapitre 4 présente le résultat principal de ce travail : grâce à l'utilisation de la diffusion quasi-élastique des neutrons combinée à des simulations de dynamique moléculaire, il a été montré qu'il y a un fort ralentissement de la diffusion globale dans le cytoplasme à partir juste en dessous de la température de mort cellulaire soutenant l'idée du déploiement des protéines dans le cadre du processus de dénaturation irréversible. Étonnamment, la fraction de protéines dépliées n'est que d'environ 5%, mais c'est l'effet que cela provoque sur l'ensemble du protéome et son comportement, qui affecte la viabilité cellulaire. Aucune dénaturation catastrophique du protéome ne se produit à la mort cellulaire.

Les derniers chapitres ont été consacrés à deux travaux complémentaires traitant tous deux des effets de la complexation protéine-ligand sur la dynamique des protéines, dans lesquels techniques de diffusion neutronique et simulations de dynamique moléculaire sont toujours couplées.

Contents

Abstract i

Résumé en français iii

1 Introduction 1
- 1.1 Cells and bacteria . . . 1
- 1.2 The effects of temperature on bacteria . . . 2
  - 1.2.1 Thermal inactivation of bacteria . . . 3
- 1.3 The role of proteins: *current situation* . . . 6
- 1.4 Our approach . . . 7

2 Neutron Scattering 9
- 2.1 General concepts . . . 9
- 2.2 Neutron Scattering - Basics . . . 12
  - 2.2.1 Coherent and Incoherent Scattering . . . 14
  - 2.2.2 Elastic Incoherent Neutron Scattering (EINS) . . . 16
    - 2.2.2.1 Gaussian Approximation (GA) . . . 17
    - 2.2.2.2 Non-Gaussian Scattering . . . 18
    - 2.2.2.3 Dynamical Heterogeneity . . . 19
    - 2.2.2.4 Generalized Mean Squared Displacement . . . 20
  - 2.2.3 Quasi-Elastic Neutron Scattering (QENS) . . . 21
- 2.3 Data reduction . . . 24
  - 2.3.1 Empty Cell Subtraction . . . 24
  - 2.3.2 Calibration: detector efficiency and energy resolution . . . 25

3 Molecular dynamics simulations 27
- 3.1 Introduction to molecular dynamics simulations . . . 27
- 3.2 Molecular dynamics simulations of proteins . . . 30
  - 3.2.1 Calculation of the forces . . . 30
    - 3.2.1.1 The force field . . . 31
    - 3.2.1.2 Boundary condition . . . 34
  - 3.2.2 Numerical Integration . . . 35
  - 3.2.3 NVT Ensemble Simulations . . . 37
  - 3.2.4 NPT Ensemble Simulations . . . 38
  - 3.2.5 Initial state of the system . . . 38
    - 3.2.5.1 Solvation . . . 39
    - 3.2.5.2 Minimization . . . 40
    - 3.2.5.3 Heating . . . 41

3.2.5.4 Equilibration . . . 41
3.2.6 Production and Analysis . . . 42
3.3 Coarse graining . . . 42
3.3.1 Production and Analysis . . . 42
3.3.2 The OPEP force field . . . 44
3.3.3 Coupling LBMD with OPEP . . . 45

4 Characterization of the Dynamical State of the E. Coli Cytoplasm and the Effect of Cell Death 49
4.1 Introduction . . . 49
4.2 Methods . . . 51
4.2.1 Sample Preparation . . . 51
4.2.2 QENS Experiments . . . 51
4.2.3 EINS Experiments . . . 59
4.2.4 Model preparation for the simulations . . . 60
4.2.5 LBMD simulation . . . 63
4.2.6 Sub-volume selection and back-mapping . . . 64
4.2.7 All-atom simulations . . . 65
4.2.8 Calculation of D_t and D_r . . . 69
4.2.9 Correcting D_t and D_r for PBC effects . . . 70
4.2.10 Viscosity calculations . . . 71
4.2.11 Evaluation of the apparent diffusion coefficient . . . 75
4.3 Results . . . 77
4.4 Appendix . . . 88
4.4.1 EINS Results . . . 88
4.4.2 Connection between the apparent and the real unfolded fractions 90
4.4.3 Origins of the nonlinearity of the transfer function . . . 92

Further Results 97

5 Differences between Ca^{2+} reach and depleted α-La investigated by MD simulations and NS experiments 97
5.1 Introduction . . . 98
5.2 Experimental Section . . . 99
5.2.1 Sample preparation for neutron experiments . . . 99
5.2.2 Simulation setup . . . 100
5.3 Neutron Scattering . . . 101
5.4 Analysis of the simulated data . . . 102
5.4.1 Direct calculation of the MSD . . . 102
5.4.2 Indirect calculation of the MSD . . . 103
5.5 Neutron Scattering Results . . . 104
5.6 MD Simulation Results . . . 106
5.7 Comparison of Experimental and Simulated Results . . . 108
5.8 Discussion and Conclusion . . . 110

6 Role of low-frequency vibrational dynamics of protein hydration water for ligand binding **113**
6.1 Introduction . . . 113
6.2 Methods . . . 115
6.2.1 Inelastic Neutron Scattering . . . 115
6.2.2 Normal Mode Analysis . . . 115
6.3 Discussion . . . 116
6.3.1 Calculation of the MSD from the vDOS . . . 119
6.3.2 Normalization on the vDOS . . . 121
6.3.3 vDOS from NMA . . . 122
6.3.4 Domain Opening Angle (DOA) . . . 123
6.3.5 Tetrahedral Order Parameter . . . 123
6.3.6 Local structure index (LSI) . . . 123

Conclusions and discussion **125**

Sources **129**

Chapter 1

Introduction

1.1 Cells and bacteria

In biology, an individual cell is the minimal self-reproducing unit of living matter.

To distinguishe living from non-living systems, we may consider that all living systems metabolize (i.e. take in resources), grow and replicate. However, there are systems that are clearly not alive who fit this definition - e.g. candle flames can take in fuel, oxidize it, grow bigger fires and light new fires; oil droplets can grow and duplicate, and similar processes can occur also in self-replicating computer codes or in human institutions. To exclude these possibilities, here we follow the definition of Dill and Agozzino [1], considering a system alive when it:

- exchanges energy and matter with the environment;
- grows and replicates independently;
- has ancestry and offspring variation;
- is formed by molecules.

The unity and diversity of organisms become apparent at the cellular level. The smallest organisms consist of single cells and are microscopic. Larger, multicellular organisms contain many different types of cells, which vary in size, shape, and specialized function. The cellular volume can vary over four orders of magnitude, from $\approx 1.3\mu\text{m}^3$ in the case of the small *Escherichia coli* bacterium, up to $\approx 10^4\mu\text{m}^3$ for the huge mammalian cell [2]. Despite the impressive diversity of living systems, all cells of the simplest and most complex organisms share certain fundamental properties, which can be observed at the biochemical level. In particular, cells of all kinds share a few of structural features [3] – see Fig. 1.1.

Their boundaries are defined by plasma membranes, separating cells' contents from the surroundings. Plasma membranes are made of lipid and protein molecules, creating a fine hydrophobic barrier around the cell that blocks the free passage of inorganic ions and majority of charged and polar compounds. Transport proteins located in the membrane control the passage of specific ions and molecules – receptor proteins transmit signals into the cell, and membrane enzymes participate in some reaction pathways. Moreover, since the individual proteins and lipids that form the plasma membrane are

not covalently bound, the entire structure of the membrane is considerably flexible, allowing changes of the size and shape of the cell. As a cell grows, newly made lipid and protein molecules are inserted into its plasma membrane; cell division produces two cells, each with its own membrane. This growth and cell division occurs without loss of membrane integrity [3].

In general, cell reproduction involves the transmission of genetic information to their offspring. Each cell stores this information in the same chemical form, i.e. as double-stranded DNA molecules. For the replication, the cell copies its information by splitting up the paired DNA strands and using each strand as a template for polymerization to build a new DNA strand with a complementary sequence of nucleotides. The same strategy of templated polymerization is used by the cells for the synthesis of proteins. This is achieved through the transcription of portions of the genetic information from DNA into the closely related RNA molecules, which in turn guide the protein production by a more complex mechanism of translation that takes place in the ribosomes [4].

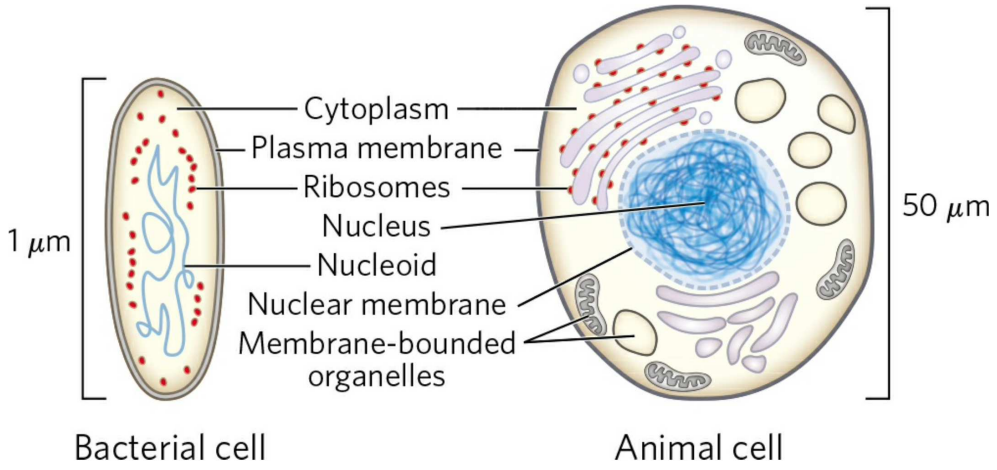

Figure 1.1: Common features of living cells. Every cell has a nucleus or nucleoid storing their DNA, a plasma membrane, and cytoplasm. *Source:* D. L. Nelson and M. M. Cox, "*Lehninger Principles of Biochemistry*" (7th edition, 2017) [3].

These processes of replication and storage of the genetic information, with the associated proteins, take place in a specific region inside the bacteria. For bacteria and archaea, this region is called nucleoid and is quite irregular, meanwhile, in the case of eukaryotic cells, it is enclosed within a double membrane and is called nucleus. The remaining material enclosed between the plasma membrane and the nucleous (or nucleoid) formes the so called cytoplasm. The liquid phase of the cytoplasm in an intact cell is the cytosol, which does not contain any part of the cytoplasm that is inside the organelles. The set of all proteins in the cell is called proteome.

1.2 The effects of temperature on bacteria

Temperature is one of the key environmental factors in microbial life, and it can be used to classify different bacteria depending on the temperature of optimal growth

(T_{OG}) - i.e. the thermal condition at which the bacteria live and thrive [5, 6, 7, 8, 9], see Table 1.1.

Table 1.1: Classification of bacteria according to their optimal growth temperatures.

Classification	Examples	Optimal Growth Temperature
Psychrophiles	*Psychrobacter articus*	$T_{OG} < 24^{\circ}\text{C}$
Mesophiles	*Escherichia coli (E. coli)*	$24^{\circ}\text{C} < T_{OG} < 50^{\circ}\text{C}$
Thermophile	*Thermus thermophilus*	$50^{\circ}\text{C} < T_{OG} < 80^{\circ}\text{C}$
Hyperthermophile	*Aquifex aeolicus*	$T_{OG} > 80^{\circ}\text{C}$

Generally, extreme temperatures are problematic for all living systems. In the case of bacteria, low temperatures can induce a decrease in enzymatic activity in the cell, and the velocity of various biochemical reactions, cell metabolism, and the fluidity of biomass membranes is reduced [10, 11]. If the temperature is further lowered below the freezing point, the water inside the cell condenses forming ice crystals, causing irreversible mechanical damage to the biomass membrane [12]. On the other hand, high temperatures provoke the irreversible denaturation of proteins, affecting the normal cell physiological activities, such as damaging the enzymes involved in the tricarboxylic acid cycle [13]. Besides, high temperatures can lead to a loss of integrity for plasma membrane and to the damage of nucleic acid molecules [14, 15].

In this thesis we will focus on the problem of the thermal response of bacteria at high temperature.

1.2.1 Thermal inactivation of bacteria

Several factors influence bacterial inactivation due to high temperatures such as the growth conditions and composition, and pH of the environment surrounding the cells [16]. Every component of the bacteria (membranes, proteins, DNA, RNA) will be affected to some degree by temperature increase. Thus, it is not an easy task to ascribe a lethal effect to a single alteration within an organism. Nevertheless, some changes are more pronounced than others, some types of damage may be repairable and all depend upon the exact value of the temperature applied.

- **Damage to the membrane.** In the case of Gram-negative bacteria (such as the *E. coli*), the outer membrane is quite sensitive to high temperature. In particular, damage to the membrane already occurs when cells are subjected to moderate heat shocks [17]. Morphological and structural changes altering the permeability of the barrier have been observed [18] – see Figure 1.2 and 1.3. This effect is enhanced in the presence of Tris buffer [19, 20]

  However, the cytoplasmic (inner) membrane, that is responsible for controlling the flux of molecules entering and leaving the cell interior, is more resistant to heat shocks. Damage to this membrane has crucial effects on bacteria, causing leaking of intracellular material [21, 22, 23, 24, 25, 26, 27]. Such injury can be produced by physical processes such as extremely high temperatures (compared with the bacterial T_{OG}) and freezing [28]. Membrane damage can be detected very readily by measuring the extent of intracellular material that leaks from the heated cells.

Overall, even if the rate and extent of leakage increases with the applied heating temperature, the correlation between loss of viability and membrane damage is poor [22, 23, 24, 25]. For this reason, it seems that membrane damage is not the major cellular site responsible for inactivation. This is even more true for Gram-positive bacteria that are relatively more heat resistant than Gram-negative bacteria [29].

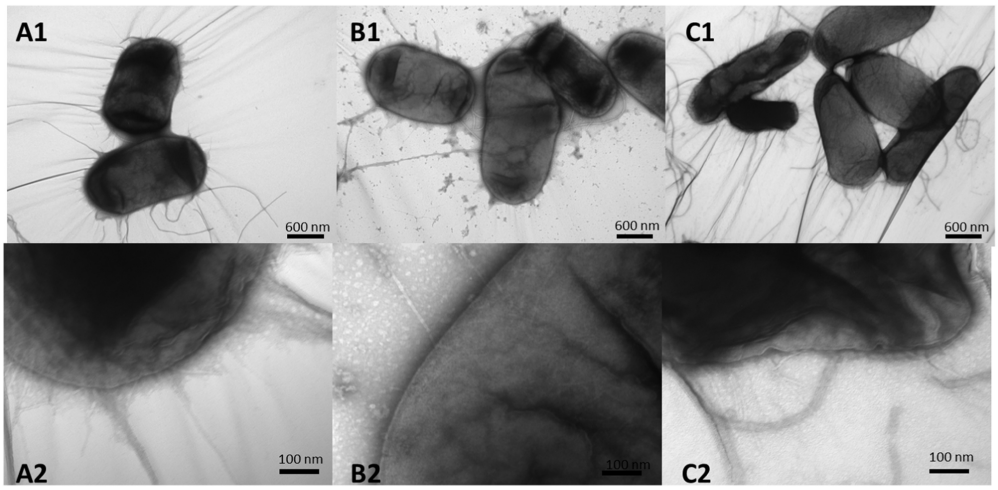

Figure 1.2: Transmission electron microscopy (TEM) images of *E. coli* cells showing the damage of their membranes. A) incubated at 35 °C, and heated to B) 50 °C, and C) 65 °C. The first row images are taken with x13,500 magnification; second row images are taken with x92,000 magnification. *Source:* B. Tonyali et al., "*Evaluation of heating effects on the morphology and membrane structure of Escherichia coli using electron paramagnetic resonance spectroscopy*" (Biophysical Chemistry 2019) [30].

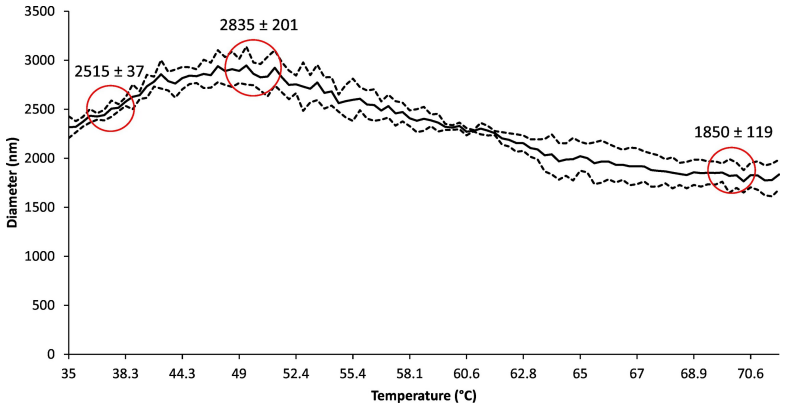

Figure 1.3: Hydrodynamic diameter of *E. coli* cells with heat treatment (35-70 °C), measured by dynamic light scattering (DLS). The solid line represents the mean of individual particle size measurements and the dash lines represent standard deviations. *Source:* B. Tonyali et al., "*Evaluation of heating effects on the morphology and membrane structure of Escherichia coli using electron paramagnetic resonance spectroscopy*" (Biophysical Chemistry 2019) [30].

- **Ribosome and rRNA degradation.** Many studies have been performed on rRNA and ribosome to study their thermal stability in heated bacterial cells [25,

26, 31, 32, 33]. Mild heating produces degradation of rRNA [32], and the degradation of 30S ribosomal subunits seems to be very dependent on the concentration of salt ions in the cytoplasm [34, 35].

However, the rRNA degradation occurs before the loss of the cellular viability [36]. On the other hand, Tomlins & Ordal [26] found that the degradation of ribosomal RNA is reversible. Therefore, these are not considered to be primary causes of heat inactivation.

- **DNA damage.** Due to its thermal resistance, the denaturation of DNA can be considered to be a minor cause of the cell inactivation due to heat shocks [36].

  However, there is a relationship between the bacterial sensitivities to ionizing radiation and mild heat shock [37, 38, 39].

- **Protein denaturation.** Structural and functional proteins are a prime target for heat inactivation. Protein denaturation occurs when cells are thermally stressed, with coagulation occurring at very high temperatures [25, 40]. There is a numerical correlation between the thermodynamic parameters of protein denaturation and the observed death rates of various organisms, suggesting a likely cause of cell death in mesophilic bacteria [41].

  This relation is strengthened by the observation that, in thermophiles and hyperthermophiles, enzymes and proteins have an enhanced thermal stability and appear to function optimally at high temperatures. The heat stability of proteins from thermophiles is increased by the presence of a number of salt bridges and by the densely packed hydrophobic interior nature of the proteins[1].

In conclusion, lethal effects of high temperatures are due to a combination of processes that involve simultaneously different parts of the bacteria. All the major cellular components are impacted, namely the outer cell barrier (especially for Gram-negative bacteria), ribosomal RNA, proteins and DNA. To react to thermal shock, the cells have different active mechanisms, in particular, there is a rapid induction of intracellular heat-shock proteins (HSPs). HSPs behave as molecular chaperones binding to the proteins that are becoming unstable due to the increase in temperature, helping to stabilize their structure and preventing their unfolding [42, 43] – see Figure 1.4.

[1] It must also be observed that, in thermophilic bacteria, not only proteins are less affected by high temperatures, but also ribosomes and membranes are more stable [28].

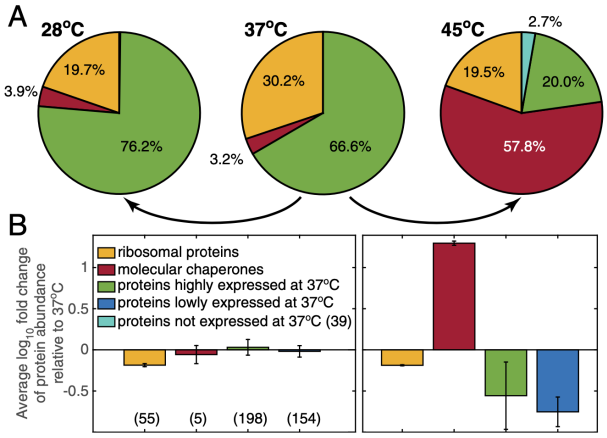

Figure 1.4: Proteome reallocation with change in temperature. The percentage is calculated with FoldME. *Source:* K. Chen et al., "*Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation*" (PNAS 2017) [43].

Moreover, extracellular alarmones may be produced to warn organisms of imminent inactivation. Exposure of cultures to increasing temperatures may enable the cells to adapt to higher, normally rapidly lethal temperatures [14].

1.3 The role of proteins: *current situation*

As described in the previous section, the proteome's thermal sensitivity has to play a key role as a determinant for most of the temperature-dependent whole-organism activities.

Different pictures have been proposed to link the degradation of the proteome to the upper limit of the cellular thermal niche, i.e. the cell's death temperature T_{CD}. A first essential aspect is to quantify the proteome's thermal stability [44, 45, 46]. On one hand a proposed theoretical model [47, 45] finds that cell death is linked to a global catastrophe of the proteome with proteins unfolding in a narrow range of temperatures near the T_{CD}. This picture has being challenged recently by experimental investigations of *E. coli* lysates and cells, and based on different techniques such as limited proteolysis [46] or thermal proteome profile [48] combined with mass spectroscopy. According to these studies only a small set of proteins indeed unfolds at the cell death. Thermal adaptation would result from the preferential stabilization of a homologous subset of proteins, thus indicating that the heat sensitivity of cells can be explained by a small number of proteins that serve critical physiological roles.

Actually, the proteome's thermal stability is not the only physical determinant of the cell's growth rate, which is expected to depend on the rate of protein diffusion throughout the cell, the latter being often the limiting factor of the rates of cellular biochemical processes [49]. Protein diffusion depends in turn on the temperature, especially through the contribution of the intrinsic viscosity in the high-temperature range when biomolecules start to unfold. To date, the relationship between the diffusive dynamics of proteome and the thermal sensitivity of a cell has not yet been investigated, also due to the extremely difficult challenge to represent the motions of proteins in a crowded milieu cell's cytoplasm where local concentration may vary from 200 g/L up to 400

g/L [50]. Here, the protein diffusive dynamics is affected by several factors, such as the presence of steric barriers given by the other macromolecules, hydrodynamic and attractive interactions and spatial heterogeneity.

1.4 Our approach

To address this problem, we combined neutron scattering spectroscopy and multi-scale molecular dynamics simulations to characterized the dynamical state of the *E. coli*'s proteome in the thermal range between 276K and 360K, at increasing and decreasing temperatures to test its reversibility. We focus our studies on *E. coli* because they are the most studied bacteria, and many useful information are available in the literature that can be used to interpret the results. Moreover, this bacteria represent a good model for all bacterial cells, and, since we are mainly interested in the bulk properties of the system, i.e the dynamical behavior of the cytoplasmic proteins which are the most abundant in the cell, *E. coli* bacteria are a good reference also for the cytosol of eukaryotic cells.

Concerning the experimental techniques, in chapter 2, we presented the basics concepts on neutron scattering: introduction to neutron scattering theory, description of the model used for the interpretation of the data, and data reduction. Chapter 3 is dedicated to molecular dynamics simulation where we discussed about: the general aspects of molecular dynamics, the simulation of proteins, and coarse-grained simulations. Finally, in chapter 4 we report our new results and we discuss their interpretations and implications.

In the appendix, we reported also other two original results that we obtained, during the PhD, studying the effects of protein-ligand complexation on protein dynamics.

Chapter 2

Neutron Scattering

The aim of this chapter is to provide a short introduction to scattering theory. The main equations relevant for the scattering techniques used within this thesis are derived to provide a theoretical background. For more detailed explanations, literature e.g. by Lovesey [51], Bée [52], Squires [53], Boothroyd [54], can be consulted.

2.1 General concepts

The neutron is an electrically neutral particle with mass $m_n \approx 939.6\text{MeV}/c^2$, spin $s = \hbar/2$ and a magnetic dipole moment $\mu_n = -1.913\mu_N$, where $\hbar = \frac{h}{2\pi} \approx 6.582\mu\text{eV ns}$ is the reduced Planck constant and $\mu_N \approx 3.1525 \cdot 10^{-2}\mu\text{eV T}^{-1}$ is the nuclear magneton. Neutrons are found in the atomic nuclei together with protons, where they are stable. In order to use them as a probe for neutron scattering experiments they have to be extracted from the nuclei. Free neutrons are not stable, they decay via a β decay process to a proton p^+ by emitting an electron e^- and an electron antineutrino $\overline{\nu}_e$:

$$n \longrightarrow p^+ + e^- + \overline{\nu}_e$$

with a mean lifetime of approximately 880 seconds. Therefore, in order to perform experiments, they have to be produced steadily from a source that, depending on the energy of neutrons required for the measurements, should not be too far from the sample. To date, there are only two ways of producing high fluxes of free neutrons for experiments: through nuclear fission in nuclear reactors or by nuclear spallation in spallation sources.

The characteristics of a neutron flux produced by reactor based sources and spallation sources are different. A reactor produces neutrons at a constant rate and thus the flux of neutrons has no explicit time structure - these are called continuous or steady state neutron sources. Typical examples are the *Institut Laue Langevin* (ILL) reactor in France or the *Forschungs-Neutronenquelle Heinz Maier-Leibnitz* (FRM II) reactor in Germany. In contrast, spallation sources generally work with pulsed neutron beams and are thus called pulsed spallation sources. Typical examples are the ISIS in UK, the *Spallation Neutron Source* (SNS) in USA, and in the near future the *European Spallation Source* (ESS) in Sweden.

The experiments describes in this thesis were all carried out at the ILL, therefore we will briefly focus on nuclear fission being the production method employed at this

facility.

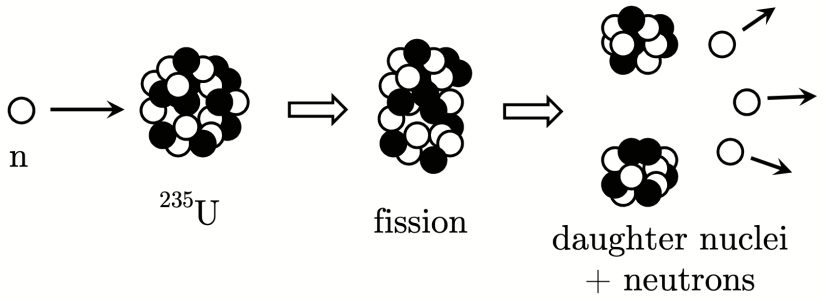

Figure 2.1: Neutron-induced fission of the 235U nucleus. *Source:* [54].

Nuclear fission, specifically neutron-induced fission in a nuclear reactor is the process by which a heavy nucleus like ^{235}U absorbs a neutron and subsequently splits into two or more lighter nuclei accompanied by the release, on average, of 2 or 3 neutrons per fission with typical kinetic energies of about 2 MeV per neutron (Fig. 2.1). These neutrons are released after the fission because the neutron/proton ratio for stable nuclei increases with the increase of their atomic number Z and, as consequence, when heavy nucleus splits into stable lighter nuclei there is an excess of neutrons that will be released with the fission. The core of a nuclear reactor contains the fissile fuel elements and it is settled so that it there is an high probability that neutrons released after the fission of a nucleus induce at least one additional fission by collision with other nuclei. This in turn induces another fission, causing a chain reaction.

In the case of nuclear reactors designed for experiments, like the one of ILL, the core is built so that the majority of the excess neutrons which do not participate in fission reactions are extracted from the reactor through special guides, usually coated with nickel and titanium, in order to guide the neutrons from the source to the instruments. Moreover, since the energies of the neutrons generally needed for scattering experiments are usually lower than few eV, to be suitable for experiments, the neutrons extracted from the reactor must be slowed down to reduce their kinetic energy. This is achieved with moderators, which are mediums that slow the neutrons down by repeated collisions with their nuclei. After many collisions the neutrons reach thermal equilibrium with the moderator and the flux per unit energy $\phi(E)$ of out-coming neutrons can described using a Maxwellian distribution, whose maximum is given by the moderator temperature T_M [54]:

$$\phi(E) \propto E \cdot \exp\left(-\frac{E}{k_B T_M}\right) \tag{2.1}$$

and the mean energy is determined by the temperature of the moderator. Examples of actual moderator substances are liquid hydrogen at 25K, liquid water at 300K, and solid graphite at 2,400K[1]. For obvious reasons, neutrons emerging from these moderators are termed cold, thermal, and hot, respectively.

[1]Light atoms are generally preferred as moderators since they uptake a high amount of the neutron energy during each collision. Often either H_2O or D_2O is used, where mainly the mass of the H or the D atoms accounts for the moderation efficiency and only about 18 (H_2O) or 25 (D_2O) collisions are needed to obtain meV energies [52].

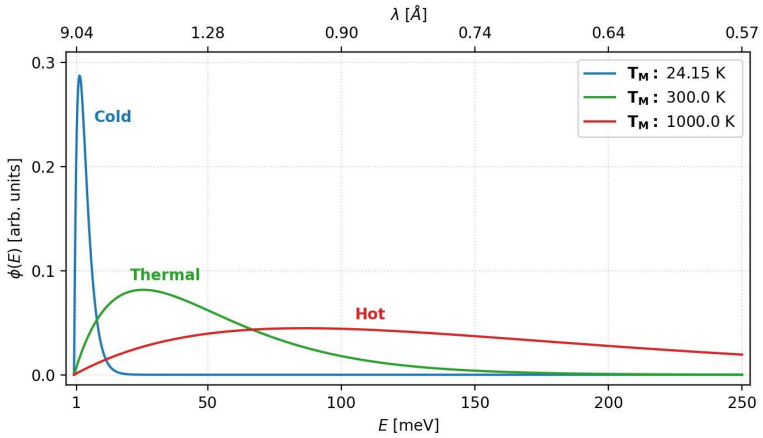

Figure 2.2: The flux per unit energy from a moderator at three temperatures.

In these energy ranges, neutrons can be treated as non-relativistic particles. Thus, given the wave-particle duality, the kinetic energy E of a neutron can either be described as a particle by its mass m_n and momentum $\hbar\mathbf{k}$ or as a wave with a wavelength λ.

$$E = \frac{\hbar^2 k^2}{2m_n} = \frac{h^2}{2m_n\lambda^2} \tag{2.2}$$

In particular, thermal neutrons (e.g. moderated by a moderator with $T_M = 293K$) have a mean velocity of $v \approx 2.20km/s$ and a corresponding wavelength of $\lambda \approx 1.8\text{Å}$ – see Tab. 2.1.

Table 2.1: Temperature, energy, and wavelength ranges for neutrons slowed down by different moderators.

Moderator	**Temperature** [K]	**Energy** [meV]	**Speed** [km/s]	**Wavelength** [Å]
Cold	1 – 120	0.09 – 10.3	0.13 – 1.41	30.8 – 3.98
Thermal	60 – 1000	5.17 – 86.2	0.99 – 4.06	3.98 – 0.97
Hot	1000 – 6000	86.2 – 517	4.06 – 9.95	0.97 – 0.40

Since this wavelength is comparable to the interatomic distances of liquids and solids, thermal neutrons provide an ideal tool for probing the microscopic properties of these materials. In addition, the energy of thermal neutrons is comparable to thermal excitations and can therefore be used to probe molecular vibrations, lattice excitations as well as atomic dynamics.

Combining the average neutron speed with the mean lifetime of the neutron, it is clear that the construction of instruments situated several hundred meters away from the neutron source is feasible even for cold neutrons. This offers the possibility to construct many instruments using one neutron source and to use e.g. curved neutron guides blocking the direct view from the instrument towards the reactor core and therefore reducing the background signal.

A diagram showing the key components of the research reactor at the ILL is provided in Fig. 2.3. The reactor operates at an output power of 56MW and produces a steady flux of neutrons in the moderator region of $1.5 \cdot 10^{15}\,\text{cm}^{-2}\,\text{s}^{-1}$. There are a number of cold, thermal, and hot moderators, and the neutrons from these are delivered to

over forty instruments designed for a wide range of scattering and fundamental physics experiments.

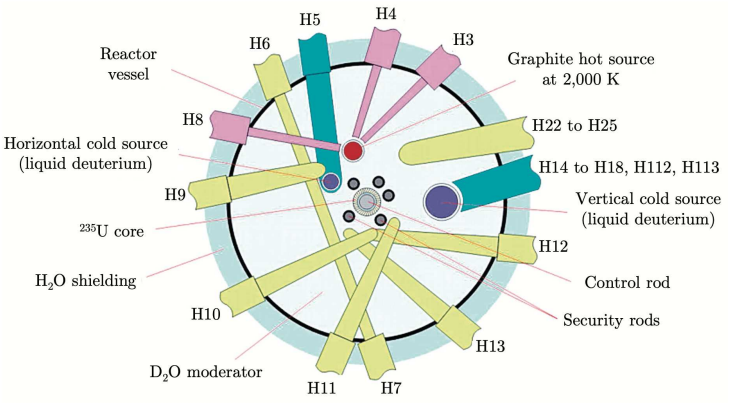

Figure 2.3: Schematic of the high-flux reactor at the Institut Laue–Langevin in Grenoble, France. Labels of the form 'Hn' indicate beam tubes and guides. *Source:* [54].

2.2 Neutron Scattering - Basics

In neutron scattering experiments a beam of neutrons with energy E_i is used to interact with a sample to obtain information on the dynamical structure and the motions of its nuclei. To this end, the flux Φ of incident neutrons is measured as well as the rate J of the scattered neutrons with a final energy between E_f and $E_f + \Delta E_f$ that are revealed, with a certain efficiency η, by a detector placed far from the point of interaction that covers a small solid angle $\Delta\Omega$ in the direction (θ, ϕ).

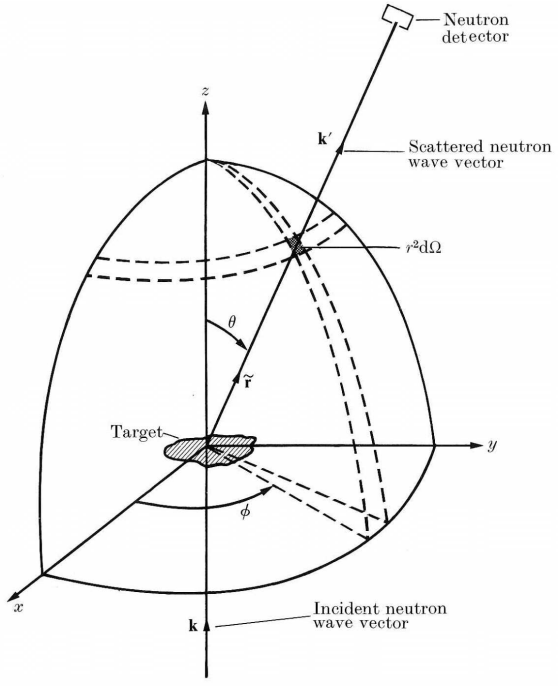

Figure 2.4: Geometry for a scattering process. *Source:* [51].

In general, the scattering is axial symmetric i.e. the rate J does not depend on the ϕ direction, and it can be described as follow

$$J(\theta, E_i, E_f) = \eta(\theta) \cdot \Phi(E_i) \left[\frac{d^2\sigma}{d\Omega\, dE_f}(\theta, E_i, E_f) \right] \cdot \Delta E_f \cdot \Delta\Omega \tag{2.3}$$

where $\frac{d^2\sigma}{d\Omega\, dE_f}$ is the *double differential cross-section*, which represents the probability that a neutron in the initial state $|\psi_i\rangle$ with an energy E_i and wave-vector $\mathbf{k}_i$ passes over into the final state $\left|\psi_f\right\rangle$ characterized with an energy E_f and the wave-vector $\mathbf{k}_f$ that forms an angle θ with $\mathbf{k}_i$. The probability of the transition from $|\psi_i\rangle$ to $\left|\psi_f\right\rangle$ can be described by *Fermi's Golden Rule*.

$$\frac{d^2\sigma}{d\Omega\, dE_f} = \frac{m}{2\pi\hbar^2} \frac{k_f}{k_i} \left| \left\langle \psi_f \right| V_N(\hat{\mathbf{r}}) \left| \psi_i \right\rangle \right|^2 \delta(E + E_i - E_f) \tag{2.4}$$

where $\delta(E + E_i - E_f)$ represents the energy conservation (E is the energy transferred from the neutron to the sample and it should be always equal to $E_i - E_f$) and $\frac{k_f}{k_i}$ is necessary for the normalization of the neutron flux. V_N is the so called pseudo potential with $V_N(\hat{\mathbf{r}}) = \frac{2\pi\hbar^2}{m} \sum_\alpha b_\alpha \delta(\hat{\mathbf{r}} - \hat{\mathbf{r}}_\alpha)$ where N is the number of nuclei in the sample, $\hat{\mathbf{r}}_\alpha$ is the position of the α^{th} particle, and b_α is the corresponding scattering length which represents the strength of the potential[2]. This pseudo potential causes the same scattering as the real potential, but is weak enough to be treated by perturbation expansion derived by Max Born. In first approximation, the Born expansion states that if $|\psi_i\rangle$ can be described by a plane wave with wave vector $\mathbf{k}_i$, the final state $\left|\psi_f\right\rangle$ is as well a plane wave with wave vector $\mathbf{k}_f$, and for the cross-section we obtain the following relation

$$\frac{d^2\sigma}{d\Omega\, dE_f} = \frac{k_f}{k_i} \left\langle \sum_{\alpha,\beta} b_\alpha^* \, \mathrm{e}^{i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\alpha} \cdot b_\beta \, \mathrm{e}^{-i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\beta} \right\rangle \delta(E + E_i - E_f) \tag{2.5}$$

where $\hat{\mathbf{q}} = \hat{\mathbf{k}}_i - \hat{\mathbf{k}}_f$ is the wave-vector transferred from the neutron to the sample.

Now, take into account the integral form of the δ-function:

$$\delta(E + E_i - E_f) = \frac{1}{2\pi\hbar} \int_{-\infty}^{+\infty} dt \cdot \exp\left(-it \cdot \frac{E + E_i - E_f}{\hbar} \right) \tag{2.6}$$

it can be substituted in eq. (2.5) leading to [53]:

$$\frac{d^2\sigma}{d\Omega\, dE_f} = \frac{1}{2\pi\hbar} \frac{k_f}{k_i} \int_{-\infty}^{+\infty} dt \cdot \mathrm{e}^{-itE/\hbar} \sum_{\alpha,\beta} \left\langle b_\alpha^* b_\beta \, \mathrm{e}^{i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\alpha(t)} \cdot \mathrm{e}^{-i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\beta(0)} \right\rangle \tag{2.7}$$

It is important to observe that $\hat{\mathbf{q}}$ and $\hat{\mathbf{r}}_\alpha$ are actually operators and that generally they do not commute. As a consequence, $[\hat{\mathbf{r}}_\alpha, \hat{\mathbf{r}}_\alpha(t)] \neq 0$ and therefore the product of $e^{-i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\alpha(0)}$ and $e^{i\hat{\mathbf{q}}\cdot\hat{\mathbf{r}}_\alpha(t)}$ is not equal to $e^{-i\hat{\mathbf{q}}\cdot(\hat{\mathbf{r}}_\alpha(0) - \mathbf{r}_\alpha(t))}$.

In the classical limit wave-vectors and the atomic positions are not operators, which

[2] b_α is independent of the neutron energy and is a complex number: the real part can be negative or positive depending of the attractive or repulsive nature of the interaction; the imaginary part represents absorption.

means that we replace $\hat{\mathbf{q}} \to \mathbf{q}$ and $\hat{\mathbf{r}}_i \to \mathbf{r}_i$, and consequently:

$$e^{i\mathbf{q}\cdot\mathbf{r}_\alpha(t)}\, e^{-i\mathbf{q}\cdot\mathbf{r}_\beta(0)} \;=\; e^{i\mathbf{q}\cdot[\mathbf{r}_\alpha(t)-\mathbf{r}_\beta(0)]} \tag{2.8}$$

Therefore for the cross-section we obtain

$$\frac{d^2\sigma}{d\Omega\, dE_f} \;=\; \frac{1}{2\pi\hbar}\,\frac{k_f}{k_i}\int_{-\infty}^{+\infty} dt\cdot \mathrm{e}^{-itE/\hbar}\sum_{\alpha,\beta}\Big\langle\, b^*_\alpha b_\beta\; \mathrm{e}^{i\mathbf{q}\cdot\left[\mathbf{r}_\alpha(t)-\mathbf{r}_\beta(0)\right]}\Big\rangle_{\mathrm{cl}} \tag{2.9}$$

where $\langle\ldots\rangle_{\mathrm{cl}}$ means that a classical average is taken. In the following, the index "cl" will be omitted.

2.2.1 Coherent and Incoherent Scattering

The spins of the nuclei in the sample and the spins of the neutrons in the beam are uncorrelated in our experiments. Assuming two scattering atoms α, β of the same isotopes having the scattering lengths b_α and b_β due to their spin states. The evaluation of $\langle b^*_\alpha b_\beta\rangle$ leads to:

$$\begin{aligned} \alpha \neq \beta\ :\quad & \langle b^*_\alpha b_\beta\rangle \;=\; \langle b^*_\alpha\rangle\,\langle b_\beta\rangle \;=\; \langle b\rangle^2 \\ \alpha = \beta\ :\quad & \langle b^*_\alpha b_\beta\rangle \;=\; \langle b^2\rangle \end{aligned} \tag{2.10}$$

Combining the two equations above, $\langle b^*_\alpha b_\beta\rangle$ can be split in two parts, reading

$$\langle b^*_\alpha b_\beta\rangle \;=\; \langle b\rangle^2 + \delta_{\alpha\beta}\left(\langle b\rangle^2 - \langle b^2\rangle\right) \;=\; \frac{1}{4\pi}\cdot\left(\sigma_{\mathrm{coh}} + \delta_{\alpha\beta}\,\sigma_{\mathrm{inc}}\right) \tag{2.11}$$

with $\sigma_{\mathrm{coh}} = 4\pi\langle b^2\rangle$ and $\sigma_{\mathrm{inc}} = 4\pi(\langle b^2\rangle - \langle b\rangle^2)$. With this result the scattering cross-section in eq. (2.9) can be split into two components:

$$\frac{d^2\sigma}{d\Omega\, dE_f} \;=\; \frac{N}{4\pi}\,\frac{k_f}{k_i}\left[\sigma_{\mathrm{coh}}\, S_{\mathrm{coh}}(\mathbf{q},E) + \sigma_{\mathrm{inc}}\, S_{\mathrm{inc}}(\mathbf{q},E)\right] \tag{2.12}$$

where N is the total number of nuclei in the scattering system and

$$S_{\mathrm{coh}}(\mathbf{q},E) \;=\; \frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dt\cdot \mathrm{e}^{-i\frac{tE}{\hbar}}\,\frac{1}{N}\sum_{\alpha,\beta}\Big\langle\, \mathrm{e}^{i\mathbf{q}\cdot\left[\mathbf{r}_\alpha(t)-\mathbf{r}_\beta(0)\right]}\Big\rangle \tag{2.13}$$

$$S_{\mathrm{inc}}(\mathbf{q},E) \;=\; \frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dt\; \mathrm{e}^{-i\frac{tE}{\hbar}}\,\frac{1}{N}\sum_{\alpha}\Big\langle\, \mathrm{e}^{i\mathbf{q}\cdot\left[\mathbf{r}_\alpha(t)-\mathbf{r}_\alpha(0)\right]}\Big\rangle \tag{2.14}$$

$S_{\mathrm{coh}}(\mathbf{q},E)$ and $S_{\mathrm{inc}}(\mathbf{q},E)$ are respectively the coherent and incoherent *dynamic structure factor* (also known as *scattering functions*). Therefore, the coherent scattering depends on the self-correlation between the positions of the same nucleus at different times and on the correlation between the positions of different nuclei at different times, leading to interference effects. In contrast, the incoherent scattering depends only on the self-correlation between the positions of the same nucleus at different times which do not give interference effects. Thus, coherent scattering can give information on structure

and collective motions (correlations between all scatterers), whereas incoherent scattering gives information about the space evolution in time (self- correlation), leading to a probe of the local dynamics of the sample.

The scattering function $S_{\text{coh}}(\mathbf{q},E)$ and $S_{\text{inc}}(\mathbf{q},E)$ can be also expressed as the Fourier transform in time of the coherent and the incoherent *intermediate scattering function* $I(\mathbf{q},t)$ defined as

$$I_{\text{coh}}(\mathbf{q},t) = \frac{1}{N}\sum_{\alpha,\beta}\left\langle e^{i\mathbf{q}\cdot\left[\mathbf{r}_\alpha(t)-\mathbf{r}_\beta(0)\right]}\right\rangle \tag{2.15}$$

$$I_{\text{inc}}(\mathbf{q},t) = \frac{1}{N}\sum_{\alpha}\left\langle e^{i\mathbf{q}\cdot\left[\mathbf{r}_\alpha(t)-\mathbf{r}_\alpha(0)\right]}\right\rangle \tag{2.16}$$

according to which:

$$S_{\text{coh}}(\mathbf{q},E) = \frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dt\cdot I_{\text{coh}}(\mathbf{q},t) \tag{2.17}$$

$$S_{\text{inc}}(\mathbf{q},E) = \frac{1}{2\pi\hbar}\int_{-\infty}^{+\infty} dt\cdot I_{\text{inc}}(\mathbf{q},t) \tag{2.18}$$

It is important to observe that, if $\sigma_{\text{inc}} \ll \sigma_{\text{coh}}$, then the incoherent scattering dominates the scattering signal and we can approximate

$$\frac{d^2\sigma}{d\Omega\, dE_f} \approx \frac{N}{4\pi}\cdot\frac{k_f}{k_i}\cdot\sigma_{\text{inc}}\, S_{\text{inc}}(\mathbf{q},E) \tag{2.19}$$

This is a good approximation for biological samples, like proteins or bacteria. As reported in Tab. 2.2, hydrogen H has a very large incoherent cross section of about 80 barns and in general, biological systems are approximately formed by half of H atoms. The remaining atoms mainly consist of C, N and O which have all very small incoherent cross sections (less than 0.5 barns), thus the majority of the signal comes from the incoherent scattering of H atoms in the sample. In most systems, the surrounding water (H2O) present in the sample will add an important contribution to the signal, thus often heavy water D2O is used as medium since its incoherent scattering cross section is 40 times smaller with 2 barns. Deuterium can also be used as substitute of hydrogen to mask specific regions in the protein.

Therefore, in what follows, we will only consider the incoherent scattering and the index "inc" will be omitted.

Table 2.2: Coherent, incoherent and absorbtion cross sections (σ_{coh}, σ_{inc}, σ_{abs}) of the principal elements present in biological systems. The cross section is given for the isotopes of hydrogen and deuterium separately. For the other elements the isotope average is given. All the σ are measured in barns, i.e. $10^{-24}\,cm^2$.

	H	D	C	O	N	S	Na	Cl	Al	V
σ_{coh}	*1.758*	5.59	5.55	4.23	11.0	1.02	1.66	1.52	1.50	0.02
σ_{inc}	**80.27**	2.05	0.001	0.0008	0.5	0.007	1.62	5.3	0.009	5.08
σ_{abs}	*0.333*	0.001	0.004	0.002	1.9	0.53	0.53	33.5	0.23	5.08

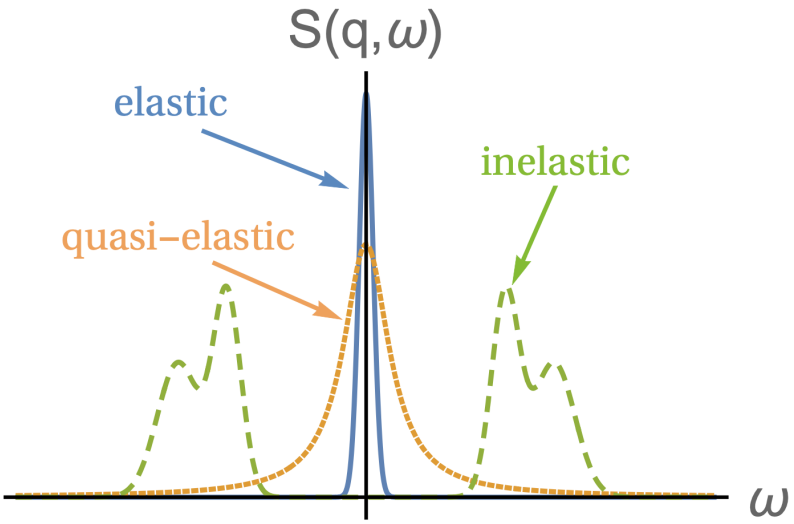

Figure 2.5: Illustration of the EINS, QENS and INS component of $S_{\text{inc}}(q,E)$. *Source:* [55].

Finally, in many neutron scattering experiments it is useful to split the dynamical structure factor in three components, depending on the amount of energy transferred with the scattering:

1. *Elastic incoherent neutron scattering* (EINS): neutrons do not exchange energy with the sample, i.e. $E \approx 0$.

2. *Quasi-elastic neutron scattering* (QENS): only a small amount of energy is exchanged between the neutron and the sample (typically less then few meV), which produce a broadening of the elastic peak.

3. Inelastic neutron scattering (INS) which appears as additional peaks centered at $E \neq 0$ well separated from the elastic peak that correspond to specific modes or excitations within the sample.

A schematic is shown in Figure 2.5.

2.2.2 Elastic Incoherent Neutron Scattering (EINS)

In the case of elastic incoherent neutron scattering (EINS), $E \approx 0$, i.e. the transferred energy is comprised between 0 and ΔE, where ΔE is the accuracy on the measurement of the energy. As it will be described with more detail in section 2.3.2, the measured dynamical structure factor S_{exp} depends on the theoretical scattering function S_{th} (which comes from the sample-neutron interaction), and the resolution function of the instrument as follow:

$$S_{\text{exp}}(q,E=0;\Delta\omega) = S_{\text{th}}(q,0) \otimes R(q,E;\Delta E) = \int_{-\infty}^{+\infty} dE'\, S_{\text{th}}(q,E')\, R(q,E';\Delta E) \tag{2.20}$$

and, considering that in our experiments the resolution of the instruments is generally well represented by a Gaussian where ΔE corresponds to its *Full Width Half Maximum*

(FWHM), the measured scattering function $S_{\text{exp}}(q, E = 0; \Delta E)$ is approximately equal to the intermediate scattering function $I_{\text{exp}}(q, t_R)$ [56]:

$$S_{\text{exp}}(q, E = 0; \Delta E) \approx I_{\text{exp}}(q, t_R) = \frac{1}{N} \sum_{\alpha=1}^{N} \langle e^{i\mathbf{q}\cdot[\mathbf{r}_\alpha(t_R) - \mathbf{r}_\alpha(0)]} \rangle \tag{2.21}$$

where $t_R = 4\hbar\sqrt{\pi^{-1} \ln 2}/\Delta E$ is the time window fixed by the instrumental resolution and determines the experimentally observable motions [57]. Therefore, in elastic neutron scattering the rate of scattered neutrons is proportional to the self intermediate scattering function $I(q,t)$ and, as a consequence, we can focus directly on this function instead of the $S_{\text{exp}}(q, E = 0; \Delta E)$.

If we define the displacement $\Delta\mathbf{r}_\alpha(t_R) = \mathbf{r}_\alpha(0) - \mathbf{r}_\alpha(t_R)$, the self intermediate scattering function can be describe as:

$$I(q, t_R) = \frac{1}{N} \sum_{\alpha=1}^{N} \langle e^{-i\mathbf{q}\cdot\Delta\mathbf{r}_\alpha(t_R)} \rangle \tag{2.22}$$

The index "exp" will be omitted in the following sections.

2.2.2.1 Gaussian Approximation (GA)

Starting from eq. 2.22, we can use the cumulant expansion for each term of the sum [53]:

$$\langle e^{-i\mathbf{q}\cdot\Delta\mathbf{r}_\alpha(t_R)} \rangle = e^{-q^2 \cdot \rho_2(t_R) + q^4 \cdot \rho_4(t_R) \pm \ldots} \tag{2.23}$$

with:

$$\begin{aligned} \rho_2(t) &= \tfrac{1}{2!} \langle \mathbf{n}_q \cdot \Delta\mathbf{r}_\alpha^2(t) \rangle \\ \rho_4(t) &= \tfrac{1}{4!} \left[\langle \mathbf{n}_q \cdot \Delta\mathbf{r}_\alpha^4(t) \rangle - 3 \langle \mathbf{n}_q \cdot \Delta\mathbf{r}_\alpha^2(t) \rangle^2 \right] \end{aligned} \tag{2.24}$$

where $\mathbf{n}_q$ is the direction of $\mathbf{q}$ and $\langle \mathbf{n}_q \cdot \Delta\mathbf{r}_\alpha^n(t) \rangle$ is the mean value of the n-th power of displacements in the direction of $\mathbf{q}$.

Then, neglecting the terms of order higher than q^2 and assuming that, on average, only isotropic displacements take place[3] (such that $\langle \mathbf{n}_q \cdot \Delta\mathbf{r}_\alpha^n \rangle = \langle \Delta r_\alpha^2 \rangle / 3$), the intermediate scattering function can be written as:

$$I(q, t_R) \approx \frac{1}{N} \sum_{\alpha=1}^{N} e^{-\frac{1}{6} q^2 \cdot \langle \Delta r_\alpha^2(t_R) \rangle} \tag{2.25}$$

Now, if we neglect also the motional heterogeneity of the H atoms in the system and we assume that the $\langle \Delta r_\alpha^2 \rangle$ of all the H are the same, i.e. the harmonic potential is equal for all the H atoms, the $\langle \Delta r_\alpha^2 \rangle = \langle \Delta r^2 \rangle$ and we obtain the well known Gaussian Approximation (GA):

$$I(q, t_R) \approx I_0 \cdot e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \tag{2.26}$$

[3] The approximation to the q^2-term is equivalent to assume that the motion of any atoms in the system can be described by diffusive motion in a harmonic potential that we consider isotropic.

where I_0 is the value of $I(q,t_R)$ for $|\mathbf{q}| = 0$ and

$$\ln(S) \approx -\frac{1}{6} q^2 \langle \Delta r^2(t_R) \rangle + \ln(I_0) \tag{2.27}$$

It is important to observe that the GA is generally only valid in a restricted region of q (specifically at low q values) since it neglects any effects from anharmonicity, heterogeneity or anisotropy. Each of these effects alone or any combination can lead to a divergence from the Gaussian behaviour. As shown by Tokuhisa et al. [58] and Vural et al. [59], usually the heterogeneity of motions is the biggest contribution to the non Gaussianity of the self intermediate scattering function [60].

2.2.2.2 Non-Gaussian Scattering

This model tries to correct, separately, the non-Gaussian behavior of the self intermediate scattering function due to single-atom non-Gaussian scattering (anharmonic displacements) and dynamic heterogeneity [61].

Non-Gaussian single-atom dynamics leads to the contribution of higher-order terms in eq. (2.23) being non-negligible, that can be expanded as [61]:

$$\left\langle e^{-i\mathbf{q}\cdot\Delta\mathbf{r}_\alpha(t_R)} \right\rangle = e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \left[1 + \sum_{k=2}^{\infty} \rho_k \, (-q^2)^k \right] \tag{2.28}$$

where we did the approximation of isotropic displacements. Now, retaining only the $k = 2$ term and neglect the motional heterogeneity, with eq. (2.24), $I(q,t_R)$ can be written as [61]:

$$I(q,t_R) = e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \left[1 + \frac{\langle \Delta r^4(t_R) \rangle - \langle \Delta r^2(t_R) \rangle^2}{72} \cdot q^4 \right] \tag{2.29}$$

On the other hand, the dynamical heterogeneity also leads to non-Gaussian behavior and, starting from eq. (2.25), the correction can be made as follows [61]:

$$\begin{aligned} I(q,t_R) &\approx e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \cdot \tfrac{1}{N} \textstyle\sum_{\alpha=1}^{N} e^{-\frac{1}{6} q^2 \cdot \left(\langle \Delta r_\alpha^2(t_R) \rangle - \langle \Delta r^2(t_R) \rangle \right)} \\ &= e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \cdot \tfrac{1}{N} \textstyle\sum_{\alpha=1}^{N} \sum_{\alpha=1}^{\infty} \frac{\mu(m)}{m!} \left(-\frac{q^2}{6} \right)^m \\ &\approx e^{-\frac{1}{6} q^2 \cdot \langle \Delta r^2(t_R) \rangle} \cdot \left(1 + \frac{q^4}{72} \sigma^2 \right) \end{aligned} \tag{2.30}$$

where $\mu(m)$ is the m-th central moment of the distribution of $\langle \Delta r^2(t_R) \rangle$ and σ^2 is the variance:

$$\sigma^2 = \frac{1}{N} \sum_{\alpha=1} \left(\langle \Delta r_\alpha^2(t_R) \rangle - \langle \Delta r^2(t_R) \rangle \right)^2 \tag{2.31}$$

and the last approximation in eq. (2.29) is valid if $(-q^2/6)^m \mu(m) \ll 1$.

Both eq. (2.29) and eq. (2.30) have similar expressions, indicating the general validity of the proposed q^4 model. Moreover, in systems in which dynamical heterogeneity is the dominant contribution to non-Gaussian behavior, the elastic scattering can in principle be used to obtain experimentally the variance and higher statistical moments of the distribution of the mean-square displacements of individual atoms. It

should be noted that this q^4 correction to the GA does not presume any particular form of the distribution of MSD [61].

However, since it does not contain higher order terms further than q^4, its applicability is limited in q [60].

2.2.2.3 Dynamical Heterogeneity - *Distribution function approaches*

In order to incorporate dynamical heterogeneity, the distribution function should be explicitly included for the calculation of the self intermediate scattering function [62]:

$$I(q,t_R) = \int_0^\infty g(s)\cdot e^{-sq^2} ds \tag{2.32}$$

where $g(s)$ is the distribution function of MSD and s is $\langle \Delta r^2(t_R)\rangle/6$. It is useful to underline that, with this approach, we have not a single value for the MSD, but a distribution of values. In particular, the mean value of the MSD, $\overline{\langle \Delta r^2(t_R)\rangle}$, is given by:

$$\bar{s} = \int_0^\infty s\cdot g(s)\, ds \tag{2.33}$$

where the $\overline{\langle \Delta r^2(t_R)\rangle} = 6\cdot\bar{s}$.

Peters and Kneller Model: This model tries to correct the GA model describing the dynamical heterogeneity with a Gamma distribution for the individual MSD.

Following Peters and Kneller [63], we define the dimensionless momentum transfer $\tilde{q} = lq$ where $l > 0$ is a scale variable with the dimension of a length, and substitute s by $\lambda = s/l^2$. Thus, from eq. 2.33, we obtain:

$$I(q,t_R) = \int_0^\infty g(\lambda)\cdot e^{-\lambda q^2} ds = \int_0^\infty \frac{\beta\,(\beta\lambda)^{\beta-1}}{\Gamma(\beta)}\cdot e^{-\beta\lambda}\cdot e^{-\lambda q^2} d\lambda \tag{2.34}$$

where $\Gamma(\beta)$ is the Gamma function and β is a parameter of the distribution such that $0 < \beta < 1$. This integral can be solved analytically and yields a simple analytical form for the model $I(q,t_R)$:

$$I(q,t_R) = \frac{1}{\left(1+\frac{\langle \Delta r^2(t_R)\rangle^2 q^2}{6\beta}\right)^\beta} \tag{2.35}$$

Bimodal: This model tries to correct the GA model describing the dynamical heterogeneity by a bimodal distribution for the individual MSD [62]:

$$g(s) = p\cdot\delta(s-s_F) + (1-p)\cdot\delta(s-s_R) \tag{2.36}$$

where $s_F = \langle \Delta r_F^2(t_R)\rangle/6$ and $s_R = \langle \Delta r_R^2(t_R)\rangle/6$ are two MSD and p is a ratio of the two contributions such that $0 \le p \le 1$. The interpretation of this model is that the system can be roughly separated in two classes of structures, flexible (F) and rigid (R), such that $\langle \Delta r_F^2(t_R)\rangle < \langle \Delta r_R^2(t_R)\rangle$. This leads to:

$$I(q,t_R) = I_0\cdot\left[p\cdot e^{-\frac{1}{6}q^2\langle \Delta r_F^2\rangle} + (1-p)\cdot e^{-\frac{1}{6}q^2\langle \Delta r_R^2\rangle}\right] \tag{2.37}$$

or, simplifying the eq. (2.37) in order to make easier the fit, we have:

$$I(q,t_R) = I_F \cdot e^{-\frac{1}{6}q^2\langle\Delta r_F^2\rangle} + I_R \cdot e^{-\frac{1}{6}q^2\langle\Delta r_R^2\rangle} \tag{2.38}$$

where:

$$p = \frac{I_F}{I_R + I_F} \qquad and \qquad I_0 = I_R + I_F \tag{2.39}$$

The mean value of the MSD is then:

$$\overline{\langle\Delta r^2\rangle} = p \cdot \langle\Delta r_F^2\rangle + (1-p) \cdot \langle\Delta r_R^2\rangle \tag{2.40}$$

Two q-Ranges: It is clear that eq. (2.37) opens the door to another approximation. Indeed, if there exist two q-ranges, q_F and q_R, such that:

$$\begin{array}{lll} I_F\, e^{-\frac{1}{6}q^2\langle\Delta r_F^2\rangle} & >> I_R\, e^{-\frac{1}{6}q^2\langle\Delta r_R^2\rangle} & for \quad q \in \{q\}_F \\ I_F\, e^{-\frac{1}{6}q^2\langle\Delta r_F^2\rangle} & << I_R\, e^{-\frac{1}{6}q^2\langle\Delta r_R^2\rangle} & for \quad q \in \{q\}_R \end{array} \tag{2.41}$$

the eq. (2.37) yields to:

$$S_{el} \approx \begin{cases} I_F\, e^{-\frac{1}{6}q^2\langle\Delta r_F^2\rangle} & for \quad q \in \{q\}_F \\ I_R\, e^{-\frac{1}{6}q^2\langle\Delta r_R^2\rangle} & for \quad q \in \{q\}_R \end{cases} \tag{2.42}$$

and in analogy with eq. (2.27), this results in two linear regimes:

$$\ln(S_{el}) \approx \begin{cases} -\frac{1}{6}q^2\langle\Delta r_F^2\rangle + ln(I_F) & for \quad q \in \{q\}_F \\ -\frac{1}{6}q^2\langle\Delta r_R^2\rangle + ln(I_R) & for \quad q \in \{q\}_R \end{cases} \tag{2.43}$$

The interpretation of this model is that in the system there are two different dynamical processes which are visible at different time scales [64, 65, 60].

2.2.2.4 Generalized Mean Squared Displacement

Hennig et al. in 2011 [66] have further developed the Gaussian approximation by transferring the approach by Rahman et al. [67] to a the elastic scattering function $S(q, E = 0)$, where a general q-dependent mean square displacement $\langle u^2\rangle_q$ is introduced by

$$\mathrm{e}^{-\frac{1}{6}q^2\cdot\langle u^2\rangle_q} := S(q, E = 0) \tag{2.44}$$

thus

$$\langle u^2\rangle_q = -6 \cdot \frac{\ln[S(q, E = 0)]}{q^2} \tag{2.45}$$

We approximate $\langle u^2\rangle_q$ by using a Taylor expansion around $q = 0$ up to the 3rd order

$$\langle u^2\rangle_q = \sum_{n=0}^{3} \frac{a_n}{n!} q^n + \mathscr{O}(q^4) \tag{2.46}$$

with the coefficients

$$a_n = \lim_{q\to 0} \frac{d^n}{dq^n} \left\langle u^2 \right\rangle_q \tag{2.47}$$

Then, using a model for the description of $S(q, E \approx 0)$ (similar to what we will see in next section on QENS), it is possible to show that the coefficients a_1 and a_3 are zero:

$$\left\langle u^2 \right\rangle_q = a_0 + a_2 q^2 + \mathcal{O}(q^4) \tag{2.48}$$

therefore

$$S(q, E=0) = \mathrm{e}^{-\frac{1}{6}\left(a_0 q^2 + a_2 q^4 + \mathcal{O}(q^6)\right)} \tag{2.49}$$

and, if we take the limit $q \to 0$, we obtain a mean square displacement $\left\langle u^2 \right\rangle$ that does not depend on q:

$$\left\langle u^2 \right\rangle = \lim_{q\to 0} \left\{ -6 \cdot \frac{\ln\left[S(q, E=0)\right]}{q^2} \right\} = a_0 \tag{2.50}$$

From this result, it can be proved that it possible to decomposed $\left\langle u^2 \right\rangle$ as follows [66]:

$$\left\langle u^2 \right\rangle = \left\langle u^2_{\mathrm{vib}} \right\rangle + \left\langle u^2_{\mathrm{sub}} \right\rangle + \left\langle u^2_{\mathrm{diff}} \right\rangle \tag{2.51}$$

where $\left\langle u^2_{\mathrm{vib}} \right\rangle$ takes into account the atomic vibrations, meanwhile $\left\langle u^2_{\mathrm{sub}} \right\rangle$ and $\left\langle u^2_{\mathrm{diff}} \right\rangle$ are two diffusive contributions: the first is due to the internal motions of the molecular sub-unit (e.g. conformational changes), the second arises from the roto-translation of the entire molecule described by apparent diffusion coefficient D_{app}. In particular, $\left\langle u^2_{\mathrm{diff}} \right\rangle$ can be calculated from D_{app}:

$$\left\langle u^2_{\mathrm{diff}} \right\rangle = 6 t_R D_{\mathrm{app}} \tag{2.52}$$

and consequently, also the internal mean square displacement $\left\langle u^2_{\mathrm{int}} \right\rangle = \left\langle u^2_{\mathrm{vib}} \right\rangle + \left\langle u^2_{\mathrm{sub}} \right\rangle$ can be obtained as

$$\left\langle u^2_{\mathrm{int}} \right\rangle = \left\langle u^2 \right\rangle - \left\langle u^2_{\mathrm{diff}} \right\rangle \tag{2.53}$$

2.2.3 Quasi-Elastic Neutron Scattering (QENS)

The section is based on references [52, 68], if no other references are mentioned. The incoherent intermediate scattering function in eq. (2.16) can be written as

$$I(\mathbf{q}, t) = \frac{1}{N} \sum_{\alpha} \left\langle \mathrm{e}^{i\mathbf{q}\cdot\mathbf{R}_\alpha(t)} \right\rangle \tag{2.54}$$

where $\mathbf{R}_\alpha(t) = \mathbf{r}_\alpha(t) - \mathbf{r}_\alpha(0)$ is the displacement of the α-th atom covered within the time t. For molecules, it is often possible to decompose the motion of an atom as a superposition of different diffusive modes acting on different time-scales. In particular, in the case of large molecules like proteins, we can rewrite $\mathbf{R}_\alpha(t)$ as a combination of translational ("trn"), rotational ("rot"), internal diffusive ("int") and vibrational ("vib") motions:

$$\mathbf{R}_\alpha(t) = \mathbf{r}_{\mathrm{trn}}(t) + \mathbf{r}_{\mathrm{rot}}(t) + \mathbf{r}_{\mathrm{int}}(t) + \mathbf{r}_{\mathrm{vib}}(t) \tag{2.55}$$

therefore, assuming that those contributions are independent, $I(\mathbf{q},t)$ can then be decomposed likewise in a product with:

$$I(\mathbf{q},t) = I_{\text{trn}}(\mathbf{q},t) \cdot I_{\text{rot}}(\mathbf{q},t) \cdot I_{\text{int}}(\mathbf{q},t) \cdot I_{\text{vib}}(\mathbf{q},t) \tag{2.56}$$

and, consequently, for the incoherent scattering function holds relation:

$$S(\mathbf{q},E) = S_{\text{trn}}(\mathbf{q},E) \otimes S_{\text{rot}}(\mathbf{q},E) \otimes S_{\text{int}}(\mathbf{q},E) \otimes S_{\text{vib}}(\mathbf{q},E) \tag{2.57}$$

where $\otimes$ is the convolution with respect to E.

This equation is the initial point for the evaluation of the QENS spectrum $S(\mathbf{q},E)$. Each component in eq. 2.57 leads to a broadening of the quasielastic peak, which is illustrated by the orange curve in Fig. 2.5. As a consequence, the broadening of this peak gives information about the different types of diffusion within the probed system.

Translational Diffusion: In the simplest case, the component $S_{\text{trn}}(\mathbf{q},E)$ reflects the translational diffusion of the center of mass (CoM) of the molecule which is moving freely and can described by the Brownian motion. The probability of finding the CoM of the molecule at a position $\mathbf{r}_{\text{trn}}(t)$ after a time t is given by the incoherent self-correlation function $G(\mathbf{t},t)$:

$$G(\mathbf{r},t) = (4\pi D_t t)^{-\frac{3}{2}} \cdot \exp\left[-\frac{r^2}{4D_t t}\right] \tag{2.58}$$

where D_t is the translational diffusion coefficient. The intermediate scattering function is obtained by the Fourier transform in space leading to:

$$I(\mathbf{q},t) = \mathrm{e}^{-q^2 D_t t} \tag{2.59}$$

The scattering function is obtained by the Fourier transform of eq. 2.59 in time which leads to a Lorentzian reading [55]:

$$S_{\text{trn}}(q,E) = \frac{1}{\pi} \cdot \frac{\gamma_t(q)}{E^2 + \gamma_t^2(q)} \tag{2.60}$$

where the parameter $\gamma_t(q)$ is the width of the Lorentzian and, in the case of free diffusion, it is equal to $D_t q^2$. It is possible to prove that, even for more complex types of diffusive processes (like jump-diffusion), the scattering function can be described with a single Lorentzian, but the width $\gamma_t(q)$ will have a different dependence on q (in the case of jump diffusion, $\gamma_t(q) = \frac{D_t q^2}{1+D_t q^2 \tau}$) [55].

Global Dynamics: Global diffusion describes the superposition of translational and rotational diffusion . In 1999, Perez et al. [69] showed that the translational and the rotational motions of proteins acts often on similar time-scales, therefore, the contribution to the dynamical structure factor for the rotational and translational diffusion of proteins cannot be separated. On the contrary, they can be approximated by a single

Lorentzian:

$$S_{\text{glb}}(q,E) = \frac{1}{\pi} \cdot \frac{\gamma_g(q)}{E^2 + \gamma_g^2(q)} \tag{2.61}$$

where the width $\gamma_g(q)$ depends on an *apparent diffusion coefficient* D_{app} which, in turn, is a function of both the translational and the rotational diffusion coefficients [70].

Internal Dynamics: Proteins in aqueous solution are not rigid since the peptide chain is able to move within a confined space. The corresponding structure factor can be approximated by

$$S_{\text{int}}(q,E) = A_0(q) \cdot \delta(E) + [1 - A_0(q)]\, \mathscr{L}_\alpha(E) \tag{2.62}$$

with the Kohlrausch-Williams-Watts function [66]:

$$\mathscr{L}_\alpha(E) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} dt \cdot \mathrm{e}^{-itE/\hbar} \cdot \mathrm{e}^{-|t\Gamma|^\alpha} \tag{2.63}$$

This function is a generalized Lorentzian function with a broad range of relaxation times occurring in the dynamics of the protein. $0 \leq \alpha \leq 1$ is a phenomenological parameter, and for $\alpha = 1$, $\mathscr{L}_1$ is a normal Lorentzian with width Γ. $A_0(q)$ is the so called *elastic incoherent structure factor* (EISF) which describes the confined internal dynamics [52].

The EISF describes the internal dynamics of a protein more precisely. These internal dynamics are very complex and therefore difficult to evaluate analytically. Notwithstanding, it is possible to apply a simplified model in which the EISF $A_0(q)$ is assumed to be a simple sum of weighted distributions describing different motions that are not uncorrelated [71]:

$$A_0(q) = p + (1-p)\left[sA_{sph}(q) + (1-s)A_{3JD}(q)\right] \tag{2.64}$$

where p is the fraction of atoms which appears to be fixed within the time scale of the instrument, whereas s is the fraction of hydrogen atoms undergoing a three-site jump diffusion. $A_{3\text{JD(q)}}$ is the intensity caused by the jump diffusion process of the H-atoms in methyl groups (-CH_3). Those atoms are located on a circle at an average distance $a_M \approx 1.715\text{Å}$ from each other and they perform jumps of 120° around the 3-fold axis [52, 69].

$$A_{3JD}(q,T) = \frac{1 + 2j_0(q \cdot a_M)}{3} \qquad \text{where} \qquad j_0(x) = \frac{\sin x}{x} \tag{2.65}$$

The internal motions are additionally described by the model of an atom diffusing in an impermeable spherical volume of radius r. The corresponding scattering intensity $A_{sph}(q)$ is given by

$$A_{sph}(q) = \left(3 \cdot \frac{j_1(q \cdot r)}{q \cdot r}\right)^2 \qquad \text{where} \qquad j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x} \tag{2.66}$$

Atomic Vibration: Atoms in the protein oscillate around their equilibrium position. The structure factor of these vibrations can be decomposed into an elastic and an in-

elastic part

$$S_{\text{vib}}(q,E) = e^{-\frac{1}{6}\langle u^2_{\text{vib}}\rangle \cdot q^2} \tag{2.67}$$

The inelastic component $S_{\text{inel}}(q,E)$ has pronounced peaks at distinct energies E, as illustrated by the green line in Fig. 2.5.

2.3 Data reduction

When performing QENS measurements, very large arrays containing the counts from every detector are recorded together with the time when each neutron was detected - the time information is essential to determine the energy of the neutrons. In order to obtain the scattering function described in the previous sections, before any data analysis can be performed, the data must be reduced to a treatable form: for every count collected from different detectors and time channels is associated with a transferred wave-vector q, and a transferred energy E. Then, the data are grouped and integrated with the desired binning of both energy and wave-vectors. More specifically, the Mantid routines [72] provided by the ILL facility were used to reduce data. The subsequent analysis and fitting was preformed by self-written python available on github: https://github.com/DanieleDiBari/NSAnalysis. In general, QENS spectra present two main issues: 1) there is a contribution to the measured signal coming from the aluminum sample holder; 2) instrument-dependent calibration accounting for different efficiency of different detectors and the instrument resolution.

2.3.1 Empty Cell Subtraction

Even if the scattering cross-sections of aluminum are quite small (see Table 2.2), the contribution to the raw spectra arises from the aluminum cell containing the sample is not negligible. In order to correct such a contribution from the total scattering function, the spectrum of the empty sample holder is measured in addition to the samples.

However, when the gap between the cell walls is filled with the sample, the spectrum of the cell is slightly different from that of the empty sample holder. In fact, both the incident and the scattered beams are subject to absorption not only within the walls of the sample cell, but also within the sample. To account for this effect, if the atomic composition of the sample is known, it is possible to employ the Paalman-Pings correction [73]. In the following, the superscript refers to the scattering object, while the subscript indicates the absorbing object. In both the super- and the subscript, "c" indicates the cell, and "s" the sample. Therefore, the scattering intensity of the bare sample I^{s}, without the contribution of the empty cell is given by

$$I^{\text{s}}(\mathbf{q},E) = \alpha_{\text{sc}} \cdot I^{\text{sc}}_{\text{sc}}(\mathbf{q},E) - \beta_{\text{sc}} \cdot I^{\text{c}}_{\text{c}}(\mathbf{q},E) \tag{2.68}$$

where $I^{\text{sc}}_{\text{sc}}$ indicates the intensity after scattering and absorption from both the sample and the cell, meanwhile I^{c}_{c} is the scattering intensity of the empty cell. Concerning the

parameters α_{sc} and β_{sc}, they are defined as follows [73]:

$$\begin{aligned} \alpha_{sc} &= \frac{1}{A^{s}_{sc}} \\ \beta_{sc} &= \frac{1}{A^{s}_{sc}} \cdot \frac{A^{c}_{sc}}{A^{c}_{c}} \end{aligned} \tag{2.69}$$

where A^{s}_{sc}, A^{c}_{sc}, and A^{c}_{c} are the absorption factors, also known as Paalman-Pings coefficients, which depend on the scattering angle θ, the transferred energy E, the atomic composition of the sample, and the shape of cell.

Nevertheless, for complex samples, like bacteria or other cells, the atomic composition could be difficult to estimate. In these cases, measuring the transmission T_{sc} of the sample and the cell, it is possible to approximate eq. (2.68) as follow:

$$I^{s}(\mathbf{q},t) \approx I^{sc}_{sc}(\mathbf{q},t) - T_{sc} \cdot I^{c}_{c}(\mathbf{q},t) \tag{2.70}$$

2.3.2 Calibration: detector efficiency and energy resolution

The measured intensity of the QENS data depends, practically, on the efficiency of each detector as shown in eq. (2.3). In general we can say that the recorded intensity $S_{exp}(q,E)$ can be described by:

$$S_{exp}(q,E) = \eta(q) \cdot S_{th}(q,E) \tag{2.71}$$

where $S_{th}(q,E)$ is the scattering function theoretically produced by the interaction between the neutron beam and the sample. In the ideal case of perfect detector efficiency $\eta(q)$, $S_{exp}(q,E) = S_{th}(q,E)$, therefore $\eta(q) = 1$. For incoherent QENS, $\eta(q)$ can be estimated by integrating $S^{V}_{exp}(q,E)$ of Vanadium (V):

$$\eta(q) = \int_{-E'}^{+E'} S^{V}_{exp}(q,E) \cdot dE \tag{2.72}$$

whit $(-E', E')$ that is the energy range used for the integration elastic peak of the Vanadium. Vanadium is chosen because of its large incoherent neutron scattering cross-section, and its purely elastic peak (modulated by the Debye-Waller factor, see below) over the observable energy transfer range, meaning that ideally all the incoherent scattering intensity is detected. Therefore, all the measured spectra are corrected for the detector efficiency by the q-wise normalization:

$$S_{th}(q,E) = \frac{S_{exp}(q,E)}{\int_{-E'}^{+E'} S^{V}_{exp}(q,E) \cdot dE} \tag{2.73}$$

However another crucial instrument-dependent feature that often affects every measurements, and that should be included in eq. (2.71), is the resolution function of the instrument. This function filters the signal produced by the scattering accordingly to the limitations of the data acquisition system of the instrument, and it depends generally on

the transferred wave-vector and the time, $R(q,t)$. In particular, it holds the following:

$$I_{\text{exp}}(q,t) = \eta(q) \cdot R(q,t) \cdot I_{\text{th}}(q,t) \tag{2.74}$$

where I_{exp} and I_{th} are, respectively, the measured intermediate scattering function and the theoretical one. To compare this equation with eq. (2.71), we can take the Fourier transform (in time) of eq. (2.74) and we obtain

$$S_{\text{exp}}(q,E) = \eta(q) \cdot R(q,E) \otimes S_{\text{th}}(q,E) \tag{2.75}$$

where $R(q,E)$ Fourier transform of $R(q,t)$.

As for the determination of the detector efficiency, measuring the scattering of Vanadium is ideal for determining the resolution function, since it is a strong elastic incoherent scatterer (Table 2.2) and serves as a standard to determine both the resolution function and the efficiency of the detectors of a neutron scattering instrument. The theoretical incoherent scattering function of vanadium in the μeV energy regime is [52]:

$$S^V_{\text{th}}(q,E) = \mathrm{e}^{-\frac{1}{6}\left\langle u^2\right\rangle_T \cdot q^2} \cdot \delta(E) \tag{2.76}$$

where $\left\langle u^2\right\rangle_T$ is the temperature dependent mean-squared displacement, which for $T =$ 296 K is $\left\langle u^2\right\rangle_T = (6.7 \pm 0.6) \cdot 10^{-3}\text{Å}^2$ [74]. Hence, for the q-range of typical neutron scattering spectrometers, $0.2\text{Å}^{-1} \leq q \leq 2\text{Å}^{-1}$, we can assume that the $\left\langle u^2\right\rangle_T$ exponent in eq. (2.76) is negligible at ambient temperatures. Consequently, the vanadium signal can be approximated by a delta function with a q-independent peak intensity which is exactly what we need to characterize the detectors efficiency and their resolution function. Combining eq. (2.75) and (2.76) we obtain indeed the following relation:

$$\begin{aligned} S^V_{\text{exp}}(q,E) &= \eta(q) \cdot R(q,E) \otimes S^V_{\text{th}}(q,E) \\ &= \eta(q) \cdot R(q,E) \otimes \delta(E) \\ &= \eta(q) \cdot R(q,E) \end{aligned} \tag{2.77}$$

Chapter 3

Molecular dynamics simulations

One technique employed in this thesis is Molecular Dynamics (MD) simulation. It is one of the most extended computer simulation techniques in material science and biophysics. In short, MD is used to calculate the time evolution of a classical many-body system by numerically integrating Newton's equations of motion. It is currently applied to a large spectrum of systems, from nanomaterials to biomolecules. This method provides two key features: it offers high spatial resolution and allows for the calculation of transport properties and relaxation constants since it does not disrupt the kinetics of the system. However, when applied to large systems, the sampling is limited by the current computational resources, e.g. the typical limit for a system of 100,000 atoms is the microsecond timescale [75].

In the first section of this chapter, an introduction to the basic concepts of the MD simulation is presented. In the second section an insight on the atomistic MD simulation of proteins is shown - since proteins are complex systems they require specific techniques developed *ad hoc*. Finally, in the last section, the coarse graining strategy to simulate complex systems is presented.

3.1 Introduction to molecular dynamics simulations

MD simulations are one of the principal tools in the theoretical study of biomolecules providing information on the relative positions of molecules and atoms as a function of time and thus their dynamics. In fact, MD is a computational method which allows to calculate the time dependent behavior of complex molecular systems (e.g. proteins, nucleic acids, etc.) and hence it is generally used to describe such systems in terms of a realistic atomic model, with the aim to understand and predict macroscopic properties based on detailed knowledge on an atomic scale. Indeed, starting from an atomistic level, MD simulations are used to predict and better understand the properties of complex materials. In this way MD provide a direct route from the microscopic details of a system (the masses of the atoms, the interactions between them, etc.) to macroscopic properties of experimental interest (the equation of state, transport coefficients and so on). In particular, usually biomolecular MD simulations are used to gain insight into ligand binding, enzymatic activities, signalling mechanisms and protein folding [76]. Additionally simulations are valuable tools for the refinement of electron microscopy, x-ray, neutron scattering or other spectroscopic data in order to obtain more accurate molecular structures and used to interpret experimental results.

Actually, MD simulations compute the motions of individual molecules for a classical many-body system in order to describe the equilibrium and transport properties of solids, liquids and gasses. Although this modelling of the matter at the microscopic level must be, in principle, based on quantum mechanics, MD generally adopts a classical point of view. In this context, the word classical means that the nuclear motions of the constituent particles obey the laws of classical mechanics (the motions are described by the second Newton's law). This is an excellent approximation for a wide range of materials.

Hence, in MD neither relativistic nor quantum effects are considered:

- *Special relativity* does not allow information to travel faster than light; MD simulations assume forces with an infinite speed of propagation.

- *Quantum mechanics* has at its base the uncertainty principle; MD requires, and provides, complete information about position and momentum at all times.

In practice, the phenomena studied by MD simulations are those where relativistic effects are not observed and quantum effects can, if necessary, be incorporated as semi-classical corrections derived from quantum theory.[1]

MD simulations allow to calculate several properties of many-particle systems. However, not all properties can be directly measured in a simulation. Conversely, most of the quantities that can be measured in a simulation do not correspond to properties that are measured in real experiments. Actually, molecular simulations generate information at the microscopic level (atomic and molecular positions, velocities, etc.) and the conversion of this very detailed information into macroscopic terms (pressure, internal energy, etc.) is the field of the statistical mechanics. Thus, the language of statistical mechanics is necessary to use these simulations as the numerical counterpart of experiments.

In this context, it is useful to see that there is a direct connection between MD simulations and the microcanonical ensemble of statistical mechanics. Indeed, the microcanonical ensemble consists of all microscopic states $(\mathbf{r}^N(t), \mathbf{p}^N(t))$ on the constant energy hypersurface $H(\mathbf{r}^N(t), \mathbf{p}^N(t)) = E$. On the other hand in the classical Hamiltonian mechanics the equations of motion conserve the total energy:

$$\frac{dH}{dt} = 0 \quad \Longrightarrow \quad H(\mathbf{r}^N(t), \mathbf{p}^N(t)) = const. \tag{3.1}$$

This suggests a link between the microcanonical ensemble and classical Hamiltonian mechanics. For a system that evolves according to Hamilton's equations of motion, a trajectory computed with these equations, namely:

$$\dot{\mathbf{r}}_i(t) = \frac{\partial H}{\partial \mathbf{p}_i} \qquad \dot{\mathbf{p}}_i(t) = -\frac{\partial H}{\partial \mathbf{r}_i} \tag{3.2}$$

will generate microscopic configurations belonging to the microcanonical ensemble with the constant energy E. Thus, if after a long time a system with energy E is able

[1] For example, dealing with very light atoms or molecules (e.g. He, H_2, D_2) or with vibrational motions with characteristic energy comparable or larger than $k_B T$, quantum effects became not negligible.

to visit practically all the configurations on the constant energy hypersurface, the dynamical evolution of this system can be used to generate a microcanonical ensemble. A system that has this property is said to be *ergodic*.

In general this dynamical approach, that is at the basis of MD simulations, provides a powerful method for generating an ensemble and its averages.[2] Thus MD simulations have evolved into one of the most widely used techniques for solving statistical mechanical problems.

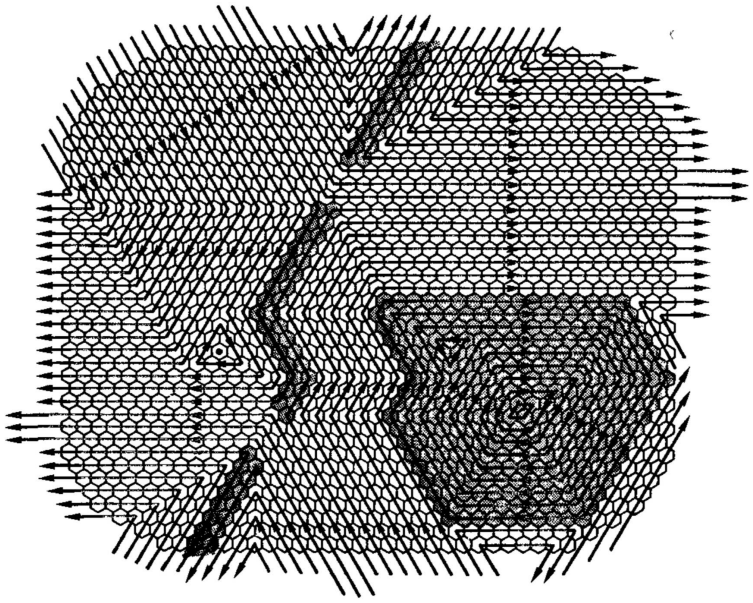

Figure 3.1: A schematic representation of phase space. The hexagonal cells represent state points $(\mathbf{r}^N, \mathbf{p}^N)$. In an ergodic system, all the trajectories in this figure represent different sections of a single long trajectory. Indeed, if the system is ergodic, the single long trajectory would eventually pass through (or arbitrarily near) all states. A substantial region of cyclical trajectories, and a barrier region leading to bottleneck, are shaded. *Source:* Allen and Tildesley, *Computer Simulation of Liquids* (1st edition, 1987) [77].

Given an ergodic trajectory, microcanonical phase space averages can be replaced by time averages over the trajectory according to:

$$\langle A \rangle \equiv \frac{\int dx\, A(x)\, \delta(H(x)-E)}{\int dx\, \delta(H(x)-E)} = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau dt\, A[x(t)] \equiv \bar{A} \tag{3.3}$$

where $x(t)$ is a representative point of the phase space defined as $x(t) = (\mathbf{r}^N(t), \mathbf{p}^N(t))$. This formula can be discretized for MD simulations as follows:

$$\langle A \rangle = \frac{1}{M} \sum_{n=0}^{M} A(x_{n\Delta t}) \tag{3.4}$$

The discretization derives from the fact that the equations of motion are solved numerically using some numerical integrators that generate phase space vectors at discrete

[2] In MD programs the phase of the simulation used to generate the ensemble is usually named: *equilibration*.

times that are multiples of a fundamental time discretization parameter Δt, known as *time step*. Starting from x_0, the vectors $x_{n\Delta t}$ (where $n = 0, \ldots, M$ with M as the total number of integration steps) are generated by applying the integrator iteratively[3] [78].

$$x(t) \xrightarrow[t \to n\Delta t]{} x_{n\Delta t} \tag{3.5}$$

By the way, adopting the classical point of view, doing a MD simulation basically means solving Newton's equations of motion for a system of N interacting atoms:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i[\mathbf{r}^N(t)] \qquad i = 1, \ldots, N \tag{3.6}$$

where the forces are derived as the negative derivatives of the potential energy function $U[\mathbf{r}^N(t)]$:

$$\mathbf{F}_i[\mathbf{r}^N(t)] = -\nabla_{\mathbf{r}_i} U[\mathbf{r}^N(t)] \tag{3.7}$$

The equations (3.6) and (3.7) are solved simultaneously at every time step. The system is supervised for a certain lapse of time, taking care that temperature and pressure remain at the required values and the coordinates are written to an output file at regular intervals. The coordinates, as a function of time, represent a trajectory of the system. After the initial changes, the system will usually reach an equilibrium state. By averaging over an equilibrium trajectory, many macroscopic properties can be extracted from an output file.

3.2 Molecular dynamics simulations of proteins

In this section several useful concepts are presented to outline how MD simulations of macromolecular systems generally work. Firstly, an introduction to the problem of the calculation of the forces is set: classical MD force field and boundary conditions, with a consideration of the special case of the long range Coulomb interactions, are shown. This is followed by a brief introduction to the numerical integration of (3.6) with mention to a simulation strategy for controlling temperature and pressure. Finally, the problem concerning the creation of the initial state is discussed: starting structures, solvation, minimization and equilibration.

3.2.1 Calculation of the forces

The potential energy is one of the most crucial parts of the simulation because it must faithfully represent the interaction between atoms, cast in the form of a simple mathematical function that can be calculated quickly. Indeed, the computation of the forces acting on every particle is the most time-consuming task of almost all MD simulations.

As biological systems involve many atoms of different types, a quantum mechanical treatment of these atoms is not feasible. The usual way to solve them is to use

[3] For biological MD simulations Δt is usually of the order of few femtoseconds $(10^{-15}\, s)$ therefore, in order to obtain a trajectory of few nanoseconds $(10^{-9}\, s)$, one has to perform at least a million of integration steps.

empirical potential energy functions, conventionally called *force fields*, which are computationally less expensive, but involve numerous approximations leading to certain limitations.[4] These functions and parameters have been derived from experimental results and quantum mechanical calculations of small model compounds. They are often refined by the use of computer simulations to compare calculated condensed phase properties with experiment. Current generation force fields provide a reasonable good compromise between accuracy and computational efficiency. Among the most commonly used potential energy functions are the AMBER, CHARMM, GROMOS and OPLS/AMBER force fields. One of the most important limitations of the empirical force fields is that no drastic changes in the electronic structure are allowed. i.e. no events like bond making or breaking can be modeled [79].

3.2.1.1 The force field

As shown in eq. (3.7), the force acting on an atom i is calculated as the negative gradient of a scalar potential energy function U that depends on all atomic positions and, thereby, couples the motion of atoms. For systems of biomolecules, this potential energy function is usually divided into two parts:

$$U = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{3.8}$$

The bonded potential U_{bonded} involves 2 , 3, and 4-body interactions of covalently bonded atoms, with $O(N)$ terms in the summation.[5] The non-bonded potential $U_{\text{non-bonded}}$ involves long-range interactions between all pairs of atoms (usually excluding pairs of atoms already involved in a bonded term), with $O(N^2)$ terms in the summation, although fast evaluation techniques are used to compute good approximations to their contributions to the potential with $O(N)$ or $O(N \log N)$ computational cost. The different terms will be explained in more detail in the following sections.

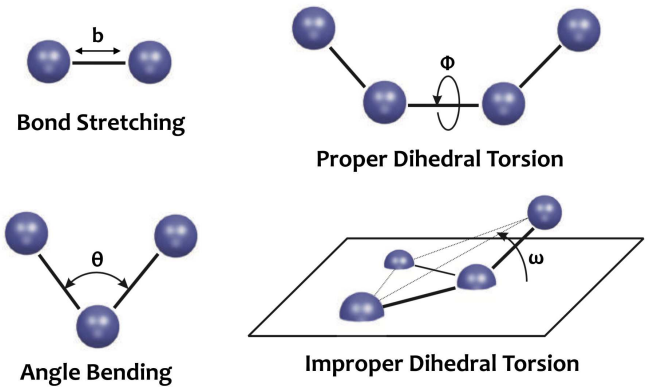

Figure 3.2: Schematic representation of the bonded interaction terms contributing to the force field. *Source:* P. Gkeka and Z. Cournia, *Molecular Dynamics simulations of lysozyme in water* (2015) [79].

[4] A force field, in the context of a computer simulation, refers to the functional forms used to describe the intra-molecular and inter-molecular potential energy of a collection of atoms, and the corresponding parameters that will determine the energy of a given configuration. Thus it is a special case of interatomic potentials and it must not be confused with force field in classical physics.

[5] Indeed, the number of covalent bound is proportional to the number of atoms.

Bonded potential terms

The bonded potential describes the stretching, bending, and torsional of the covalent bonds.

- **Bond stretching:**
  The bond stretching term is a 2-body potential, generally assumed to be harmonic, that describes the vibrational motion between a pair of covalently bonded atoms:

$$U_b = k_b (b - b_0)^2 \tag{3.9}$$

  where b is the distance between the two atoms. Two parameters characterize each bonded interaction: b_0 the average distance between them and a force constant k_b.

- **Angle bending:**
  The angle bending term describes the force originating from the deformation of the valence angles between three covalently bonded atoms (3-body interactions). The angle bending term is described using a harmonic potential:

$$U_\theta = k_\theta (\theta - \theta_0)^2 \tag{3.10}$$

  where θ is the angle between three atoms. Two parameters characterize each angle in the system: the reference angle θ_0 and a force constant k_θ.

- **Torsional terms:**
  The torsional terms are weaker than the bond stretching and angle bending terms. They describe the barriers to rotations existing between four bonded atoms (4-body interaction). There are two types of torsional terms: proper and improper dihedrals. Proper torsional potentials are described by a cosine function:

$$U_\phi = k_\phi [1 + \cos(n\phi - \delta)] \tag{3.11}$$

  where ϕ is the angle between the planes formed by the first and the last three of the four atoms. Three parameters characterize this interaction: δ sets the minimum energy angle, k_ϕ is a force constant and n is the periodicity.

  The improper dihedral term is designed both to maintain chirality about a tetrahedral heavy atom and to maintain planarity about certain atoms. The potential is described by a harmonic function:

$$U_\omega = k_\omega (\omega - \omega_0)^2 \tag{3.12}$$

  where ω is the angle between the plane formed by the central atom and two peripheral atoms and the plane formed by the peripheral atoms (see Fig. 3.2).

Non-bonded potential terms

The non-bonded potential describes the van der Waals forces and the electrostatic interactions between the atoms.

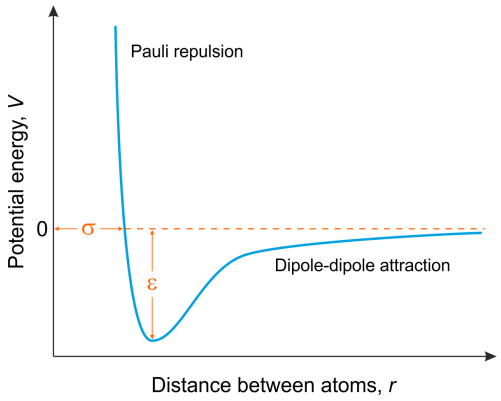

Figure 3.3: Schematic representation of Lennard-Jones potential. The collision parameter, σ, is shown along with the well depth, ε. *Source:* Eni. Generalic, *Lennard-Jones potential* (Croatian-English Chemistry Dictionary & Glossar, 2017) [80].

- **Van der Waals interactions:**
  The van der Waals force acts on atoms in close proximity. It is strongly repulsive at short range and weakly attractive at medium range. The interaction is described by a Lennard-Jones potential:

$$U_{VdW} = 4\varepsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right] \tag{3.13}$$

  where r is the distance between two atoms. It is parametrized by σ: the collision parameter (the separation for which the energy is zero) and ε the depth of the potential well. The Lennard–Jones potential approaches 0 rapidly as r increases, so it is usually truncated (smoothly shifted) to 0 past a cutoff radius, requiring $O(N)$ computational cost.

- **Electrostatic interactions:**
  Finally, the long distance electrostatic interaction between two atoms is described by Coulomb's law:

$$U_{el} = \varepsilon_{1-4} \cdot \frac{q_1 q_2}{4\pi\varepsilon_0 r_{12}} \tag{3.14}$$

  where q_1 and q_2 are the charges of both atoms and r_{12} the distance between them, while ε_0 is the electric susceptibility of vacuum. The parameter ε_{1-4} is a unitless scaling factor whose value is 1, except for a modified $1-4$ interaction, where the pair of atoms is separated by a sequence of three covalent bonds (so that the atoms might also be involved in a torsion angle interaction), in which case $\varepsilon_{1-4} = \varepsilon$, for a fixed constant $0 \leq \varepsilon \leq 1$. Although the electrostatic potential may be computed with a cutoff like the Lennard-Jones potential, the r^{-1} potential approaches 0 much more slowly than the r^{-6} potential, so neglecting the long range electrostatic terms can degrade qualitative results, especially for highly charged systems. There are other fast evaluation methods that approximate the contribution to the long range electrostatic terms that require $O(N)$ or $O(N \log N)$ computational cost, depending on the method.

Potential energy function

Finally, the equation for the potential energy describing the force field can be expressed as:

$$\left.\begin{aligned} U \;=\; & \sum_{bonds\, b} k_b\,(b-b_0)^2 + \sum_{angles\, \theta} k_\theta\,(\theta-\theta_0)^2 + \\ & + \sum_{\substack{proper\\ dihedrals}} k_\phi\,[1+\cos(n\,\phi-\delta)] + \sum_{\substack{improper\\ dihedrals}} k_\omega\,(\omega-\omega_0)^2 + \end{aligned}\right\} \text{BONDED INTERACTIONS} \tag{3.15a}$$

$$+ \sum_{\substack{i,j\\ j<i}} 4\,\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] + \sum_{\substack{i,j\\ j<i}} \varepsilon_{1-4}\cdot\frac{q_1\,q_2}{4\,\pi\,\varepsilon_0\,r_{12}} \quad \Longrightarrow \quad \text{NON-BONDED INTERACTIONS} \tag{3.15b}$$

3.2.1.2 Boundary condition

To avoid surface effects at the boundary of the simulated system, periodic boundary conditions are often used in MD simulations; the particles are enclosed in a cell that is replicated to infinity by periodic translations. A particle that leaves the cell on one side is replaced by a copy entering the cell on the opposite side, and each particle is subject to the potential from all other particles in the system including images in the surrounding cells, thus entirely eliminating surface effects (but not finite-size effects). Because every cell is an identical copy of all the others, all the image particles move together, consequently they should be represented only once inside the molecular dynamics code.

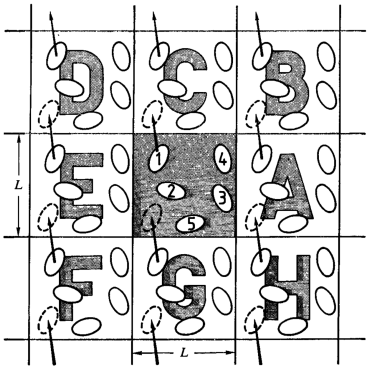

Figure 3.4: A two-dimensional periodic system. Molecules can enter and leave each box across each of the four edges. In a three-dimensional example, molecules would be free to cross any of the six cube faces. *Source:* Allen and Tildesley, *Computer Simulation of Liquids* (1st edition, 1987) [77].

However, because van der Waals and electrostatic interactions exist between every non-bonded pair of atoms in the system (including those in neighboring cells) computing the long-range interaction exactly is unfeasible. To perform this computation, the van der Waals interaction is spatially truncated at a user-specified cutoff distance. For

a simulation using periodic boundary conditions, the system periodicity is exploited to compute the full (non-truncated) electrostatic interaction with minimal additional cost using the Particle-Mesh Ewald (PME) method described in the next paragraph.

Full Electrostatic Computation

Ewald summation is a description of the long-range electrostatic interactions for a spatially limited system with periodic boundary conditions. The infinite sum of charge-charge interactions for a charge-neutral system is conditionally convergent, meaning that the result of the summation depends on the order in which it is taken. Ewald summation specifies the order as follows: sum over each box first, then sum over spheres of boxes of increasingly larger radii. Ewald summation is considered more reliable than a cutoff scheme, although it is noted that the artificial periodicity can lead to bias in free energy, and can artificially stabilize a protein that should have unfolded quickly.

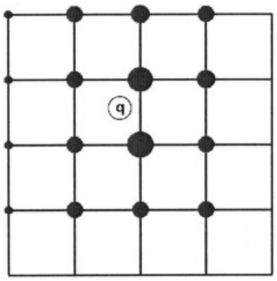

Figure 3.5: In PME, a charge (denoted by an empty circle with label **q** in the figure) is distributed over grid (here a mesh in two dimensions) points with weighting functions chosen according to the distance of the respective grid points to the location of the charge. Positioning all charges on a grid enables the use of the Fast Fourier Transform (FFT) to solve Poisson's equation for the electrostatic potential due to the charge distribution. In real applications, the grid is three-dimensional. *Source:* J. C. Phillips et al., *Scalable Molecular Dynamics with NAMD* (Journal of Computational Chemistry, 2005) [81].

The PME method is a fast numerical method to compute the Ewald sum. The cost of PME is proportional to $N \log N$ and the time reduction is significant even for a small system of several hundred atoms. The strict conservation of energy resulting from the computed force is crucial and is strongly assisted by maintaining the symplecticness of the integrator, as discussed further below. However the PME method does not conserve energy and momentum simultaneously, but momentum conservation can be enforced by subtracting the net force from the reciprocal sum computation, albeit at the cost of a small long-time energy drift.

3.2.2 Numerical Integration

Biomolecular simulations often require millions of time steps. Furthermore, biological systems are chaotic; trajectories starting from slightly different initial conditions diverge exponentially fast and after a few picoseconds are completely uncorrelated. However, highly accurate trajectories are not normally a goal for biomolecular

simulations; more important is a proper sampling of phase space. Therefore, for constant energy (NVE ensemble) simulations, the key features of an integrator are not only how accurate it is locally, but also how efficient it is, and how well it preserves the fundamental dynamical properties, such as energy, momentum, time-reversibility, and symplecticness.

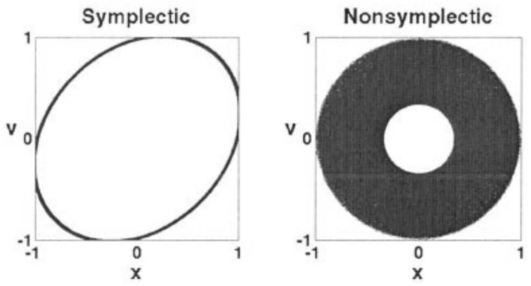

Figure 3.6: Simple example that shows the merit of a symplectic integrator: integration of a one-dimensional harmonic oscillator with an unit circle as exact trajectory. The trajectory of nonsymplectic method initially draw a circle, but after few steps starts to collapse toward the center while the symplectic method maintains a stable orbit, showing a superior long-time stability, even though its trajectory is deformed into an ellipse by a larger local error. *Source:* J. C. Phillips et al., *Scalable Molecular Dynamics with NAMD* (Journal of Computational Chemistry, 2005) [81].

The time evolution of a strict Hamiltonian system is symplectic. A consequence of this is the conservation of phase space volume along the trajectory, that is, the enforcement of the Liouville theorem. To a large extent, the trajectories computed by numerical integrators observing symplecticness represent the solution of a closely related problem that is still Hamiltonian. Because of this, the errors, unavoidably generated by an integrator at each time step, accumulate imperceptibly slowly, resulting in a very small long-time energy drift, if there is any at all. Artificial measures to conserve energy, for example, scaling the velocity at each time step so that the total energy is constant, lead to biased phase space sampling of the constant energy surface; in contrast, there has been no evidence that symplectic integrators have this problem.

One of the most used integratos for NVE simulations is the velocity Verlet [82]. This method obtains the position and velocity at the next time step (r_{n+1}, v_{n+1}) from the current one (r_n, v_n), assuming the force $F_n = F(r_n)$ is already computed, in the following way:

$$\begin{aligned} \textit{half-kick} \rightarrow \quad v_{n+\frac{1}{2}} &= v_n + 0.5 \cdot \Delta t \cdot F_n/m \\ \textit{drift} \rightarrow \quad r_{n+1} &= r_n + \Delta t \cdot v_{n+\frac{1}{2}} \\ \textit{compute force} \rightarrow \quad F_{n+1} &= F(r_{n+1}) \\ \textit{half-kick} \rightarrow \quad v_{n+1} &= v_{n+\frac{1}{2}} + 0.5 \cdot \Delta t \cdot F_{n+1}/m \end{aligned}$$

where m is the mass. The Verlet method is symplectic and time reversible, conserves linear and angular momentum, and requires only one force evaluation for each time step. For a fixed time period, the method exhibits a (global) error proportional to Δt^2.

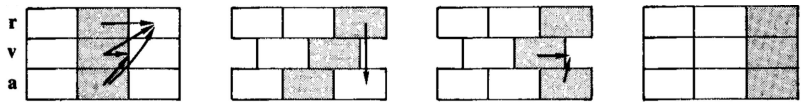

Figure 3.7: Schematic representation of the velocity Verlet algorithm. At each step, the stored variables are in grey boxes. *Source:* Allen and Tildesley, *Computer Simulation of Liquids* (1st edition, 1987) [77].

More accurate (higher order) methods are desirable if they can increase the time step per force evaluation. However, higher order Runge-Kutta type methods, symplectic or not, are not suitable for biomolecular simulations because they require several force evaluations for each time step and force evaluation is by far the most time-consuming task in molecular dynamics simulations. Gear type predictor-corrector methods, or linear multistep methods in general, are not symplectic. Hence, no symplectic method has been found as yet that is both more accurate than the Verlet method and as practical for biomolecular simulations.

On the other hand, it is also possible to employ multiple-time-stepping methods to improve integration efficiency. Because the biomolecular interactions described in eq. (3.15) generally act on different time scales, this allows us to compute the slower-varying forces less frequently than faster ones. This idea can be implemented by three levels of integration loops. The inner loop uses only bonded forces to advance the system, the middle loop uses Lennard–Jones and short-range electrostatic forces, and the outer loop uses long-range electrostatic forces.[6]

Using this multiple time-step approach can increase computational efficiency by a factor of 2, however the longest time step is limited by resonance.[7]

3.2.3 NVT Ensemble Simulations

A fundamental requirement for an integrator is to generate the correct ensemble distribution for the specified temperature and pressure in an appropriate way. For this purpose the Newtonian equations of motion (3.6) should be modified "mildly" so that the computed short-time trajectory can still be interpreted in a conventional way. To generate the correct ensemble distribution, the system is coupled to a reservoir, with the coupling being either deterministic or stochastic. Deterministic couplings generally have some conserved quantities (similar to total energy), the monitoring of which can provide some confidence in the simulation. NAMD uses a stochastic coupling approach because it is easier to implement and the friction terms tend to enhance the dynamical stability.

The (stochastic) Langevin equation is used in NAMD to generate the Boltzmann distribution for canonical (NVT) ensemble simulations. The generic Langevin equation

[6] In the article: *Scalable Molecular Dynamics with NAMD* (Journal of Computational Chemistry, 2005), Phillips and his colleagues note that this method, implemented in the NAMD software, is symplectic and time reversible [81].

[7] When good energy conservation is needed for NVE ensemble simulations, it is recommended to choose $2fs$, $2fs$, and $4fs$ as the inner, middle, and outer time steps if rigid bonds to hydrogen atoms are used; or $1fs$, $1fs$, and $3fs$ if bonds to hydrogen are flexible. More aggressive time steps may be used, instead, for NVT or NPT ensemble simulations - i.e. $2fs$, $2fs$, and $6fs$ with rigid bonds and $1fs$, $2fs$, and $4fs$ without [81].

is:

$$m_i\ddot{\mathbf{r}}_i(t) = \mathbf{F}_i[\mathbf{r}^N(t)] - \gamma\dot{\mathbf{r}}_i(t) + \sqrt{\frac{2\gamma k_B T}{m_i}}\,\mathbf{R}(t) \tag{3.16}$$

where γ is the friction coefficient, k_B is the Boltzmann constant, T is the temperature and $\mathbf{R}(t)$ is a univariate Gaussian random process. Coupling to the reservoir is modeled by adding the fluctuating (the last term) and dissipative ($-\gamma\dot{\mathbf{r}}_i$ term) forces to the Newtonian equations of motion. To integrate the Langevin equation, NAMD uses the Brünger-Brooks-Karplus (BBK) method, a natural extension of the Verlet method for the Langevin equation. The position recurrence relation of the BBK method is:

$$r_{n+1} = r_n + \frac{1 - 0.5\,\gamma\Delta t}{1 + 0.5\,\gamma\Delta t}(r_n - r_{n-1}) + \frac{\Delta t^2}{1 + 0.5\,\gamma\Delta t}\left[\frac{F(r_n)}{m} + \sqrt{\frac{2\gamma k_B T}{m}}\,Z_n\right] \tag{3.17}$$

where Z_n is a set of Gaussian random variables of zero mean and variance 1. The BBK integrator requires only one random number for each degree of freedom. The steady-state distribution generated by the BBK method has an error proportional to Δt^2, although the error in the time correlation function can have an error proportional to Δt.

3.2.4 NPT Ensemble Simulations

Often, it is useful to maintain a simulated system at both constant temperature and pressure. In a thermodynamic sense, systems at constant pressure are the ones that can exchange volume with their surroundings (e.g., by way of a piston). Their volume therefore fluctuates. Likewise, simulated systems at constant pressure involve volume fluctuations.

A number of different barostat techniques exist with the scope of maintaining a target pressure by dynamically adjusting the volume of the system during the simulation. Mainly, the most common barostat techniques are based on:

- ▷ Volume rescaling – the instantaneous pressure is made to equal the target pressure by rescaling the system volume at periodic intervals.
- ▷ Berendsen barostat – the pressure is weakly coupled to a pressure bath and the volume periodically rescaled.
- ▷ Extended ensemble barostat (also known as Andersen barostat) – the system is coupled to a fictitious pressure bath using an extended Lagrangian and the introduction of new degrees of freedom.

A detailed discussion of the algorithms commonly used is beyond the scope of this thesis.

3.2.5 Initial state of the system

In order to perform a simulation for measuring some selected properties of a given system, it is essential, like in a real experiment, to prepare first the initial state of the

system. This means to obtain, in some way, a set of coordinates and velocities for all the atoms that identify an initial configuration of the system compatible with the desired initial state.

Specifically, for the simulation of large biological molecules such as proteins or nucleic acids, and as a consequence of the complex structures of these macromolecules, the initial positions of their atoms are usually obtained from the results of some real experiments, like X-Rays crystallography or nuclear magnetic resonance (NMR) spectroscopy, whereas the initial velocities are typically set pseudorandomly so that the total kinetic energy of the system corresponds to the expected value at the target temperature.

Actually, the sets of coordinates obtained from these experiments often do not represent an initial configuration that is compatible with the desired initial state of the system (e.g. due to the hydration of the sample or to the fact that the set of coordinates usually refer only to one macromolecule, while the real system might be formed by more). However, since these coordinates represent the three-dimensional structure of the macromolecules involved in the simulation, they provide a fundamental starting point for the preparation of the system. Indeed from these coordinates, through several processes that exploit and modify them, it is possible to derive a configuration for the entire system that, been as close as possible to the configuration of the real system that it is intended to study, can be used to simulate its initial state.

Hence, when the initial configuration is achieved, it is possible to make several simulations of the system that can be used to measure the properties to be studied (this phase of the simulation is conventionally named: *production phase* and the main scope of a molecular simulation is to get some interesting results from this phase). The process generally used in MD simulations of biological system, preceding the production phase, is mainly the following: solvation, minimization, heating (set the velocities) and equilibration; that are describe below.

In any case, it is important to point out that the choice of the initial configuration must be done carefully as this can influence the quality of the entire simulation.

3.2.5.1 Solvation

Solvation consists in the process of taking into account the solvent surrounding the biomolecules. Actually, biomolecules are generally in solution with some types of aqueous solvent and therefore solvation is a common process that is used to prepare the initial configuration of the system. This is indeed due to the known fact that solvation effects play a crucial role in determining molecular conformation, electronic properties, binding energies, etc.

There are mainly two ways to solvate a biomolecular system:

- **explicit treatment:** the coordinates of the solvent atoms are directly added to the system with the specification of the structure of the molecules that occur in the solvent (most often water molecules and, sometimes, also salt ions).[8]

- **implicit treatment:** the force field is modified to include also the effect due to the interaction between the biomolecular system and the solvent. Thus this technique eliminates the need of specifying all the atoms of the solvent by including many

[8]Usually, in this phase of the simulation, periodic boundary conditions are particularly important to avoid surface effects.

of their average effects in the inter-atomic force calculation. For example, polar solvent acts as a dielectric and screens (lessens) electrostatic interactions.

The elimination of explicit water accelerates conformational explorations and sometimes increases simulation speed, although at the cost of not modeling the solvent as accurately as explicit models. However, since implicit solvent models represent the solvent in an averaged manner, they are considered less accurate than explicit solvent models. Caution should always be used when implicit solvents are employed for molecular dynamics research.

3.2.5.2 Minimization

In the context of MD simulation, the energy minimization, in short known as minimization, is the process of finding an arrangement in space for a collection of atoms where, according to the force field used, the net inter-atomic force on each atom is acceptably close to zero and the position on the potential energy surface is a stationary point. In general, the collection of atoms might be a single molecule, an ion, a condensed phase, a transition state or even a collection of any of these.

As an example, when optimizing the geometry of a water molecule, one aims to obtain the hydrogen-oxygen bond lengths and the hydrogen-oxygen-hydrogen bond angle which minimize the forces that would otherwise be pulling atoms together or pushing them apart.

The motivation for performing an energy minimization is the physical significance of the obtained structure: even when initial coordinates are available from an experiment, the starting vector may not correspond to a minimum in the potential energy function used, and as such minimization is needed to relax strained contacts. When an experimental structure is not available, a build-up technique may be used to construct a structure on the basis of the known building blocks, and minimization again is required.

For a biomolecular system, the potential energy function is a very complex and multidimensional landscape. It has one deepest point, the global minimum, and a very large number of local minima.

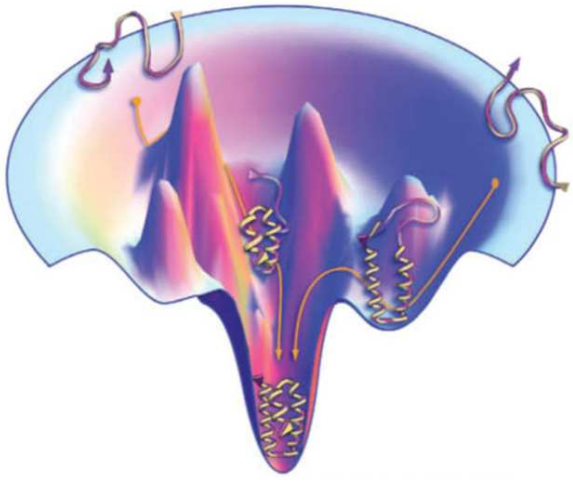

Figure 3.8: Simplified example of the energy landscape (as a function of only two variables). *Source:* Voet D., Voet J.G. and Pratt C.W., "*Fundamentals of Biochemistry*" (5th edition 2016)[83].

The goal of the energy minimization is to find a local minimum. The energy at this local minimum may be much higher than the energy of the global minimum. Performing an energy minimization will guarantee the removal of any strong van der Waals interactions which might otherwise lead to local structural distortion and result in an unstable simulation.

3.2.5.3 Heating

During this phase, the integration of the equations of motion starts. The heating process, usually performed in the NVT ensemble, consists of progressively increasing of the reservoir's temperature to gradually heat the system. This process is performed mainly for two reasons:

- to reach the desired temperature at which the system has to be studied;
- to change the configuration of the system so as to escape from local minima within the energy landscape that are probably obtained with the minimization but that do not represent the real state of the system.[9]

In both cases, the need of a gradual increase of the temperature is due to the fact that, starting from the structures obtained from the X-Ray experiment, typically performed at low temperatures, and after some manipulations and the minimization that modifies these structures, the configuration of the system obtained is built without taking into account the velocities and hence the contribution of the kinetic energy to the total energy. Indeed starting a simulation at high temperature with such artificial configuration of the system that may be considered as "frozen", might be unfeasible because the simulation will be probably numerically unstable.

Therefore, during the heating phase, initial velocities are assigned at a low temperature and the simulation is started with periodically assigning new velocities at a slightly higher temperature and letting the simulation to continue. This step is repeated until the desired temperature is reached.[10]

3.2.5.4 Equilibration

Once the heating process is over and the desired temperature is reached, the simulation is continued and during this phase, properties such as structure, pressure, temperature and the energy are monitored. The point of the equilibration phase is to run the simulation until these properties become stable with respect to time. If in the process, the temperature increases or decreases significantly, the velocities are scaled such that the temperature returns to its desired value.

[9] This might be done also in alternation with some phases of cooling.

[10] Clearly, between the velocities of the atoms and the temperature of the system there is a strong connection. According to the classical equipartition theorem, each normal mode has $\frac{1}{2}k_BT$ energy, on average, at thermal equilibrium. Thus $\langle E \rangle = \frac{1}{2}\sum_i m_i v_i^2 = \frac{1}{2}Nk_BT$. One of the most common ways to set the initial velocities is by generating numbers pseudorandomly and choose them so that the total kinetic energy of the system corresponds to the expected value at the target temperature T.

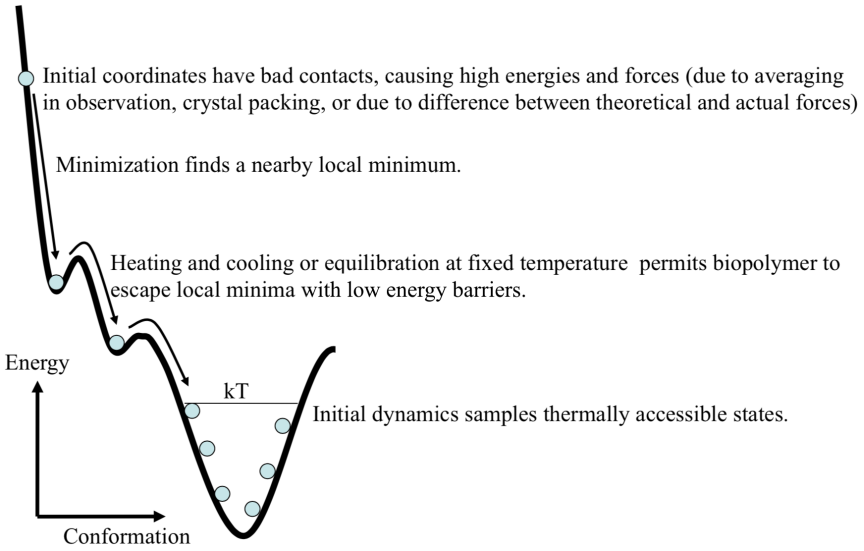

Figure 3.9: From the Mountains to the Valleys: a molecular dynamics fairy tale. *Source:* Theoretical and Computational Biophysics Group, "*Computational Biophysics Workshop*" (Boston, Dec. 5-9, 2004).

3.2.6 Production and Analysis

The last step of the simulation is the production phase, wherein the system is simulated for the time length required, normally from several hundred *ps* to *ns* or more. During this process, coordinates and eventually also velocities, of the system at different times are stored in the form of trajectories. These are then used for calculations of mean energy, mean square distance (MSD) between structures, local mean square atomic fluctuations (MSF) etc. From MD simulations, time dependent properties such as correlation functions can also be calculated and these in turn can be related to spectroscopic measurements.

3.3 Coarse graining

The atomistic simulations described in the previous sections have been successfully employed to investigate many aspects of biological systems like the folding of small proteins.However, biological systems and phenomena exhibit complexities and diversities that spread over a wide and disparate range of spatio-temporal scales. These phenomena include the dynamics of large proteins and the self-assembly of biological molecules. Different space and time resolutions are involved, from the quantum mechanical level, describing the electronic structures, to the atomic scale, and the continuum level of fluid motion at macroscopic scales (Figure 3.10).

3.3.1 Production and Analysis

Despite the growth of computational resources and power, and the emergence of enhanced sampling methodologies, all-atom simulations are still limited to systems containing tens or hundreds of thousands of atoms on a μs time scale, that only rarely

reach the millisecond time-scale [84]. Hence, the deployment of simplified representations still able to capture the essential features of the phenomena is required (Fig. 3.11). One successful strategy is to reduce the number of degrees of freedom by a systematic coarse-graining. Coarse-grained (CG) models present an attractive alternative to the traditional atomistic simulations offering the possibility of investigating complex cellular processes over larger lengths and longer time scales at a reduced level of detail. A coarse-graining operation requires the selection of the level of description. Then, the degrees of freedom are reduced by averaging off the behaviour of the fast ones. In the coarse-graining of biomolecules, the strategy is, for example, to cast together groups of atoms, thus preserving some degrees of molecularity. However, depending on the problem under consideration a more aggressive coarsening can be performed, i.e. a whole protein can be modelled as a single particle.

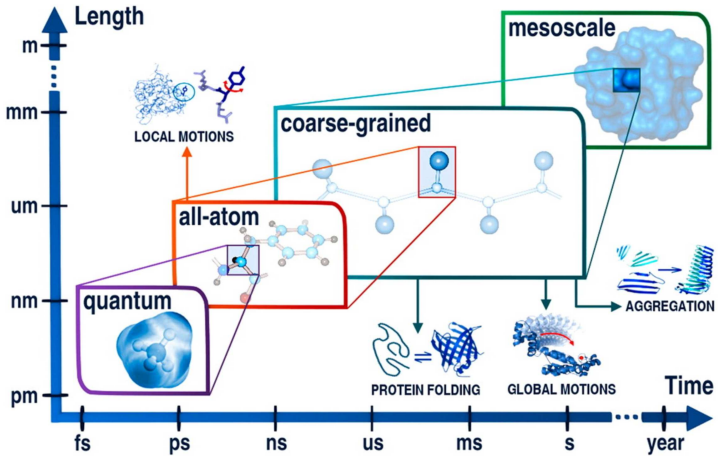

Figure 3.10: Schematic representation of various computational approaches, used to cover different scales of length and time pertinent to different biophysical processes. These methods range from highly accurate, but computational demanding to the highly efficient but very low-detail continuum models. Figure reprinted from [84].

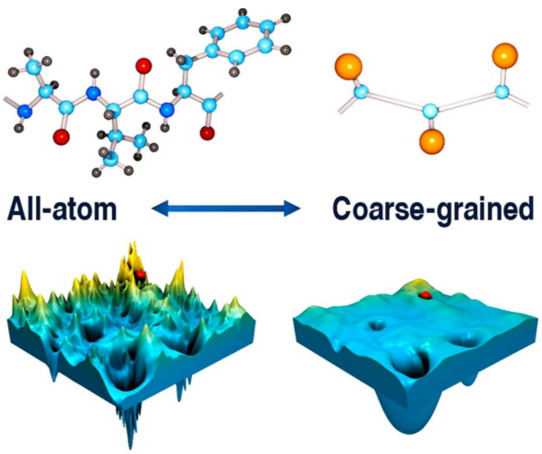

Figure 3.11: All-atom versus coarse-grained energy landscape. The figure illustrates the effect of the smoothening of the energy landscape in a coarse-grained model as compared to an all-atom model. Figure reprinted from [84].

The parametrization of force fields that are both accurate and transferable - that is, capable of describing the general dynamics of systems with different compositions and different configurations - becomes increasingly difficult as the graining becomes 'coarser', because more specific interactions must effectively be included in fewer parameters and functional forms. On other hand, coarse-graining also affects thermodynamic properties of a modelled system, particularly the balance between enthalpy and entropy. Reduction of the degrees of freedom affects the entropy of the simulated system, which is compensated by reduced enthalpic terms. In turn, a coarse-grained model may accurately reproduce free energy differences but contributing enthalpy and entropy values may be inaccurate. Such limitations are typical for the majority of coarse-grained models.

For fluids like water, the molecular CG consists in describing an ensemble of particles as a single entity and inserting important features at the level of the interaction potentials. Stronger simplification can reduce the fluid to essential variables, like density or velocity. For instance, in the Brownian Dynamics, the effects of collisions of the solvent with a large particle are described without representing explicitly the solvent molecules.

It is clear that when the CG is performed only phenomena at the pertinent scales are accessible and the high level information is lost. For instance, when liquid water is described at the mesoscopic level, we can not use this representation to investigate the kinetics of the hydrogen-bond formation. The reason is that the dynamics is averaged out in the model. Many strategies, and rigorous procedures can be devised to perform coarse-graining.

3.3.2 The OPEP force field

In our work, for the description of the proteins in the CG simulations, we have employed the *Optimized Potential for Efficient protein structure Prediction* (OPEP) force field. This model has been developed 21 years ago by Derreumaux and coworkers, [85]. and it consist in an intermediate resolution model where each amino acid is represented with six beads: the backbone is retained in full atomic detail (all N, C_α, C, O and H main-chain atoms are considered), while side chains are represented by a unique bead located at the center of mass of their heavy atoms[11], see Figure 3.12.

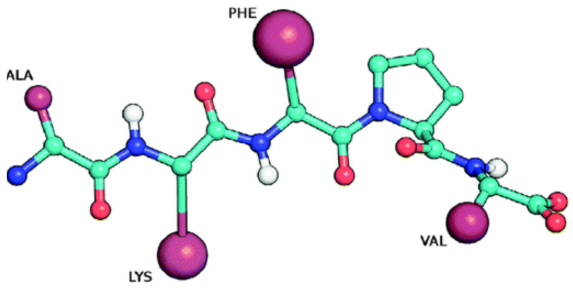

Figure 3.12: The figure show an example of the OPEP model for the peptide Ala-Lys-Phe-Pro-Val in its zwitterion form to show the details of the backbone and the side-chains. Figure reprinted from [85].

[11] Clearly, an exception is made for the amino acid proline, whose the side chain is formed by only one H atom. Therefore it is represented in the OPEP model by all its atomic constituents.

The Van der Waals radius and the positions of the side chain beads were calculated using a database of 2,250 PDB structures with sequence identity lower than 30%. OPEP is highly transferable among different proteins, but it is limited to soluble protein in water. On the other hand, it requires a small integration time step of 1.5–2.0 fs due to the detailed backbone and hydrogen bonds, limiting the amount of sampling that can be accomplished.

3.3.3 Hydrodynamic interactions: *Lattice Boltzmann Molecular Dynamics coupled with OPEP*

Treating the hydrodynamic interactions is complex and computationally demanding, because they are long range interactions, non-linear in nature, and cannot be expressed simply as a sum of two-body terms. Several alternative schemes have been developed along the years. The basic idea is to track the solvent degrees of freedom through a simplified mesoscopic representation, with the local dynamics that satisfies the mass, momentum and energy conservation laws and recovers the solution of hydrodynamic equations in the large-scale limit. In order to simulate a soft matter system, the solvent model must then be coupled with the respective algorithms that models the dynamics of solute particles. One of the most widely used schemes is the Lattice Boltzmann (LB) method.

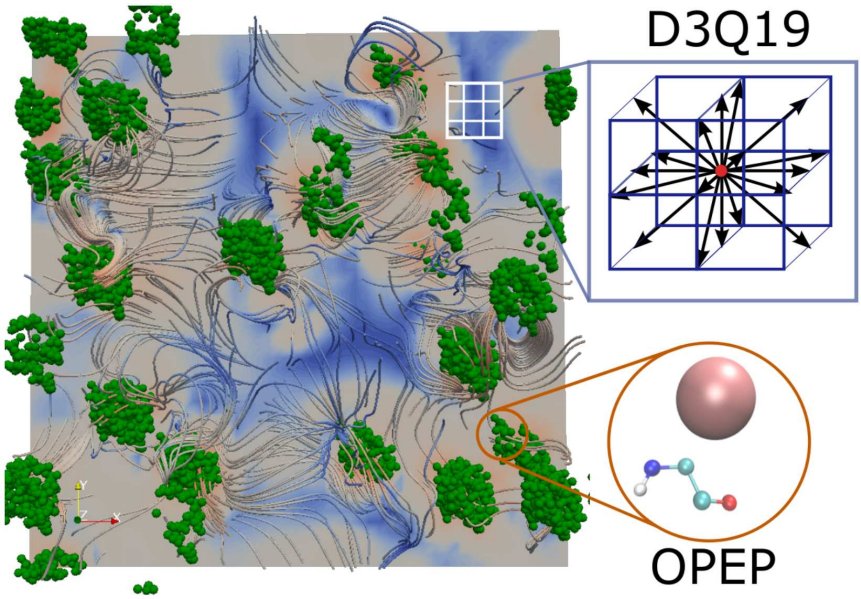

Figure 3.13: Pictorial view of the LBMD multiscale scheme. Proteins are described at microscopic level, interact according to the OPEP coarse-grained force field and move in the continuum. The aqueous solvent is handled by the lattice Boltzmann method, so that fluid populations that reside on a Cartesian mesh are evolved in time and move to neighboring mesh points as connected by a set of discrete speeds (indicated as D3Q19) [86].

In the LB approach, the fluid is represented through particles that reside on a three-dimensional cubic lattice with spacing Δx [87, 88]. Here "fluid particles" do not correspond to single molecular entities but represent instead the collective motion of the

fluid (see Fig. 3.13). In short, we track the solvent kinetic degrees of freedom through a simplified representation, with a local dynamics that satisfies the mass and momentum conservation laws and recovers in the large-scale limit the Navier-Stokes equation of fluid-mechanics.

The lattice Boltzmann equation reads as follows:

$$f_p\left(\mathbf{x}+\mathbf{c}_p\Delta t, t+\Delta t\right) = f_p\left(\mathbf{x}, t\right) - \omega\Delta t \cdot \left(f_p - f_p^{eq}\right)\left(\mathbf{x}, t\right) + g_p(\mathbf{x}, t) \tag{3.18}$$

where $f_p\left(\mathbf{x}, t\right)$ denotes the probability distribution of finding a particle at lattice site $\mathbf{x}$ at time t and moving in lattice space with discrete velocity $\mathbf{c}_p$, Δt is the time-step for the numerical integration. The particle displacement occurs to the first and second lattice neighbors (D3Q19) by using 18 directions plus a null one mimicking particles at rest.

The distribution $f_p\left(\mathbf{x}, t\right)$ evolves in space and time toward the equilibrium target f_p^{eq} with the characteristic relaxation frequency $\omega = 1/\tau$. The term $g_p\left(\mathbf{x}, t\right)$ includes the drag force $\mathbf{F}^D$ and extra contributions $\mathbf{F}^{ext}$ as the random noise encoded in the molecular dynamics. When focusing on a solute particle suspended in the fluid, this term describes essentially the particle-to-fluid back reaction. An accurate expression of $g_p(x,t)$ is given by [89]:

$$g_p\left(\mathbf{x}, t\right) = -w_p\Delta t\left[\frac{\mathbf{G}\cdot\mathbf{c}_p}{c_s^2} + \frac{\left(\mathbf{G}\cdot\mathbf{c}_p\right)\left(\mathbf{u}\cdot\mathbf{c}_p\right) - c_s^2}{c_s^2\,\mathbf{G}\cdot\mathbf{c}_p}c_s^4\right] \tag{3.19}$$

The force $\mathbf{G}$ contains any external force and the exchange of momentum induced by N moving atoms, $\mathbf{G} = \mathbf{F}^D + \mathbf{F}^{ext}$. To lowest order, it can be shown that $g_p\left(\mathbf{x}, t\right) = -w_p\Delta t\,\frac{\mathbf{G}\cdot\mathbf{c}_p}{c_s^2}$, whereas in practice a higher order version is needed to ensure a global second-order accuracy of the lattice Boltzmann solver [90]. Without entering into the analytical details, the method is typically extended to account for local fluctuations at the level of the stress tensor, such that fluctuating hydrodynamics are recovered.

In addition, it can be shown that local mass and momentum of the global particle-fluid elements are preserved, a key condition to obtain the correct fluid dynamic behavior. It should also be remarked that the model can be easily extended to account for thermodynamic forces. The coupling between the motion of a solute particle and the fluid is based on the assumption that momenta exchange in Stokes-like fashion, thus defining a drag force between the i-th particle of mass m_i and the fluid [91]:

$$\mathbf{F}^D\left(\mathbf{R}_i\right) = -m_i\,\gamma_i\left[\mathbf{V}_i - \tilde{\mathbf{u}}\left(\mathbf{R}_i\right)\right] \tag{3.20}$$

where $\mathbf{V}_i$ is the atom velocity, and $\tilde{\mathbf{u}}$ indicates the fluid velocity field distributed over the region occupied by the atom, and γ_i is a friction coefficient that in principle can vary depending on the solute particle type.

In this scheme (LB coupled with the OPEP force field), the dynamics of the particles are governed by the following evolution equations for the positions $\mathbf{R}_i$ and velocities $\mathbf{V}_i$ of the particles:

$$\dot{\mathbf{R}}_i = \mathbf{V}_i \tag{3.21}$$

$$\dot{\mathbf{V}}_i = \frac{\mathbf{F}_i^C + \mathbf{F}_i^D}{m_i} + \mu_i \tag{3.22}$$

where $\mathbf{F}_i^C$ is a conservative force describing the sum of molecular interactions as encoded in the OPEP force field. The drag force is given by eq. (3.20). As anticipated above, $\mathbf{F}_i^D$ represents the mechanical and dissipative friction exerted between a particle and the surrounding fluid. The strength of this dissipation depends on γ_i, a friction parameter that can be tuned in order to alter the response time between fluid and molecular motions. Finally μ_i is a white noise mimicking the effect of the thermal collisions with the molecules of the fluid, with the mean $\langle \mu_i(t) \rangle_t = 0$ and $\langle \mu_i(t)\mu_i(0) \rangle_t = 2\gamma_i k_B T$, being k_B the Boltzmann constant and T the temperature. Equation (3.21) is integrated over the time step Δt_{MD} according to the symplectic position Verlet algorithm [90] and, if $\Delta t_{MD} = \Delta t_{LB}$, the particle and LB dynamics are updated in a synchronous way.

In conclusion, conceptually we adopted Boltzmann kinetic theory, and its numerical representation, in order to describe the solvent as a continuum in a probabilistic sense. Traditionally, in the LB formulation the lattice space Δx which supports the fluid kinetics is defined as a representation (coarse-graining) of the collective kinetic behaviour of a group of solvent molecules. It is also accepted that in order to observe hydrodynamic behavior down to the Δx scale, it is usually considered that the fluid mean free path should not exceed Δx. Since in liquid water, the molecular mean free path is of the order of a few angstroms, a subnanometric lattice space can be supported in LB, allowing for the hydrodynamic behaviour to emerge at distance larger than Δx. In this approach, the lattice grid element must not be viewed as a volumetric entity that contains a fixed number of particles, but instead a numerical support for the probabilistic description of the averaged single particle trajectories. Horbach and Succi [92] have shown that this strategy is effective for the simulation of nanofluids and the obtained results agree very closely with particle-based simulations.

This coupling between the LB method and the OPEP model has been implemented in MUPHY software.

Chapter 4

Characterization of the Dynamical State of the E. Coli Cytoplasm and the Effect of Cell Death

Based on a paper in preparation:

Short-time Diffusive Dynamics of Bacterial Proteome as a Proxy of the Cell Death.

***Daniele Di Bari**, Stepan Timr, Marianne Guiral, Marie-Thérèse Giudici-Orticoni, Tilo Seydel, Christian Beck, Caterina Petrillo, Philippe Derreumaux, Simone Melchionna, Fabio Sterpone, Judith Peters and Alessandro Paciaroni*

(to be submitted to Science)

Temperature is a boost for cellular metabolism, but above a certain threshold, it corrupts functional processes involving proteins and causes cell death. Whether the thermal denaturation involves the whole proteome or just a subset of critical proteins in the cytoplasm is still debated. Here, we attack the problem from a preferential angle by monitoring via QENS and multi-scale simulations the dynamical state of the *E.coli* proteome across the cell death temperature. Above the cell death temperature, the cytoplasm experiences a dynamical slowdown caused by the unfolding of just a small number of proteins. This small fraction is sufficient to induce the gelation of the cytoplasm. From the dynamical properties, the fraction of unfolding is extracted and used to reconstruct successfully the *E. coli* growth rate.

4.1 Introduction

Temperature has a significant impact on cells. Notably, membranes, proteins and nucleic acids suffer in various ways from heat. The membranous integrity can be challenged resulting in the evasion of periplasmic proteins or the entrance of harmful compounds [14]. Proteins are mandatory for good cellular functioning, but high temperature provokes loss of conformations and denaturation. Nucleic acids are the

most stable against thermal stress [93], so that their denaturation can be considered as a minor cause of cell death. Moreover, biological migrations, extinctions, genetic divergence, and speciation can all be triggered by small changes in environmental temperature [94, 95, 96]. A deep understanding of the cell's thermal stability is key to model the impact of climate change on microbial organism growth [97], establish theoretical boundaries for life in extreme environments [98] and optimize thermal based treatments for cancer [99]. Yet, the factors that influence the cell's thermal sensitivity are largely unknown. The proteome's thermal sensitivity has to play a key role as a determinant for most of the temperature-dependent whole-organism activities, as proteins are the most abundant and less stable biomolecules in the cell.

Different pictures have been proposed to link the degradation of the proteome to the upper limit of the cellular thermal niche, i.e. the cell's death temperature T_{CD}. A first essential aspect is to quantify the proteome thermal stability [44, 45, 46]. On one hand a proposed theoretical model [47, 45] finds that the cell death is linked to a global catastrophe of the proteome with proteins unfolding in a narrow range of temperatures near the T_{CD}. This picture has been challenged recently by experimental investigations of *E. coli* lysates and cells, and based on different techniques such as limited proteolysis [46] or thermal proteome profile [48], combined with mass spectroscopy. According to these studies only a small set of proteins indeed unfolds at the cell death. Thermal adaptation would result from the preferential stabilization of a homologous subset of proteins, thus indicating that the heat sensitivity of cells can be explained by a small number of proteins that serve critical physiological roles.

Actually, the proteome's thermal stability is not the only physical determinant of the cell's growth rate, which is expected to depend on the rate of protein diffusion throughout the cell, the latter being often the limiting factor of the rates of cellular biochemical processes [49]. Protein diffusion depends in turn on the temperature, especially through the contribution of the intrinsic viscosity in the high-temperature range when biomolecules start to unfold. To date, the relationship between the diffusive dynamics of proteome and the thermal sensitivity of a cell has not yet been investigated, also due to the extremely difficult challenge to represent the motions of proteins in a crowded milieu cell's cytoplasm where local concentration may vary from 200 g/L up to 400 g/L [50]. Here, the protein diffusive dynamics is affected by several factors, such as the presence of steric barriers given by the other macromolecules, hydrodynamic and attractive interactions and spatial heterogeneity.

On these grounds, here we provide an unprecedented picture of the dynamics of the *E. coli's* proteome in the nanosecond time-scale, based on state-of-the-art neutron scattering spectroscopy and multi-scale molecular dynamics simulation. We show that in *E. coli* the global protein diffusion is a close proxy of the bacterial metabolism, with a linear Stokes-Einstein dependence in the lower temperature range and a striking dynamic slow-down above the thermal death. Combining the results on the proteome dynamics from neutron scattering and simulations we describe the way the unfolded protein fraction progressively increases with temperature, offering an alternative quantification to existing ones [45, 46, 48]. We clearly show that no global proteome unfolding occurs at cell death. Finally, we verify that the derived proteome stability curve and temperature-dependent proteome diffusivity together, allow to excellently reproduce the *E. coli* growth rate profile.

4.2 Methods

4.2.1 Sample Preparation

E. coli BL21 (DE3) was grown overnight in LB medium (made with H2O) at 37°C with shaking (200rpm). 3.3g of *E. coli* cells were collected by centrifugation and washed twice with a buffer made in deuterium oxide (99.9 atom D) as followed: cells were suspended in 36mL of D2O buffer at pD8, spun at 5400g for 18 minutes at 6°C and the supernatant was removed. The pD 8 buffer contains 50mM Tris, 150mM NaCl and 5mM KCl. To obtain a pD of 8, the pH of the buffer was adjusted to 7.6 using HCl.

4.2.2 QENS Experiments

We performed two experiments [100, 101] on the cold neutron backscattering spectrometer IN16B at the Institut Laue-Langevin (ILL) [102], with an energy resolution of $\approx 0.75\mu eV$ FWHM, an energy-range of $|E| \leq 31\mu$eV, and a wave-vector coverage of $0.19\text{Å}^{-1} \leq q \leq 1.9\text{Å}^{-1}$. The samples were measured for 2 hours at each temperature in a temperature range from 275 K to 348 K. The data reduction (i.e. normalization to the monitor, integration over the regions of interest of the vertically position-sensitive detector tubes, calculation of the energy axis, and centering of the elastic line positions using separate Vanadium measurements) where carried out with the built-in module for IN16B of the Mantid program [72]. The subtraction of the sample holder contribution to the signals, the normalization of the detector efficiency, and the fit of the data where performed using an in-house python module that is available on github: *https://github.com/DanieleDiBari/NSAnalysis*.

QENS Model. The fully reduced scattering function measured from the *E. coli* samples, $S_{exp}(q,E)$, can be obtained by the convolution of the theoretical scattering function S_{th}, describing the interaction between the neutrons and the bacteria, and the instrumental resolution, $R(q,E)$, which is determined by a vanadium sample, a completely elastic and incoherent scatterer (see section 2.3.2) [52]:

$$S_{exp}(q,E,T) = e^{-\frac{E}{2k_BT}} \cdot [R(q,E) \otimes S_{th}(q,E,T)] \tag{4.1}$$

where $e^{-\frac{E}{2k_BT}}$ is the detailed balance factor.

Since proteins represent $\approx 55\%$ of the bacterial dry weight (see Table 4.1) and have a percentage of hydrogen higher (50%) than any other type of macromolecules (30%) [103], the major contribution to S_{th} is originating from self-diffusive dynamics of an average protein and of the bulk water present in the sample, in this case D2O (see section 2.2.3):

$$S_{th}(q,E.T) = S_{AP}(q,E,T) + \phi \cdot S_{D2O}(q,E,T) \tag{4.2}$$

where $S_{AP}(q,E,T)$ and $S_{D2O}(q,E,T)$ are, respectively, the scattering functions of the average protein and the D_2O, and ϕ is a scalar factor that weights the contribution of the solvent. ϕ can be estimated by the product of the D_2O amount in the sample, which constitutes about 80% of the sample mass, and the percentage of bulk water molecules in *E. coli*, which is almost 60%.

Table 4.1: Composition of the average *E. coli* bacteria (from Neidhardt et al. [104])

MOLECULES	PERCENTAGE OF DRY WEIGHT
Macromolecules	**96**
Proteins	55
RNA	20.5
Lipids	9
Polysaccharides	5
Lipopolysaccharides	3.4
DNA	3.1
Monomers	**3**
Sugars and precursors	2
Amino acids and precursors	0.5
Nucleotides and precursors	0.5
Inorganic ions	**1**

QENS: Fitting procedure. To mitigate the problem of overfitting due to the complexity of the system and the consequent high number of unknown parameters necessary to describe the resutling QENS signal, we tried to reduce, as much as possible, the number of free parameters that can vary during the fit. To this end, we used the measurements of the PBS-D2O Buffer to fix the q-dependence of the solvent's intensity S_{D2O} in eq. (4.2). Moreover, we employed a three step procedure for the analysis, where at each step we perform a fit of the data and we used the resulting information to improve the model used for the fit and reduce the number of free parameter.

QENS: Energy Resolution. The instrumental resolution function $R(q,E)$ takes several parameters into account and it strongly depends on the set-up of the instrument (section 2.3.2).

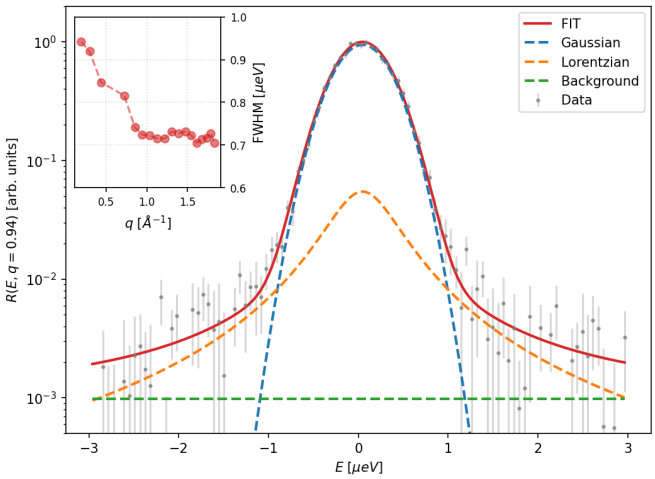

Figure 4.1: Example of the QENS spectrum measured for the Vanadium at $275K$ and $q = 0.94\text{Å}^{-1}$. The red line is the best fit of $R(q,E)$ described by the eq. (4.3), meanwhile the dashed lines are the different components of $R(q,E)$. The inset shows the Full Width Half Maximum (FWHM) of $R(q,E)$ which represents a measure of the energy resolution of the instrument.

In particular, at the IN16B spectrometer the resolution function is well described by a sum weighted by the scalar parameter $F(q)$ of a Gaussian and a Lorentzian function [102]:

$$R(q,E) = A(q) \cdot \left\{ \frac{1}{\pi} \cdot \frac{F(q) \cdot \gamma(q)}{\gamma^2(q) + [E - E_0(q)]^2} + (1 - F(q)) \cdot \frac{\exp\left\{ -\frac{[E-E_0(q)]^2}{2\sigma^2(q)} \right\}}{\sigma(q)\sqrt{2\pi}} \right\} + B(q) \tag{4.3}$$

where $|F(q)| \leq 1$. B(q) reflects a background which can depend on q. Figure 4.1 shows the fit of the Vanadium data with the eq. (4.3) and the measured energy resolution of the instrument, ΔE_{FWHM}, corresponding to the Full Width Half Maximum (FWHM) of eq. (4.3). The resulting parameters averaged over q are reported in table 4.2.

Table 4.2: Average parameters resulting from the fit of the Vanadium data with eq. (4.3).

$\langle F(q) \rangle$	$\langle A(q) \rangle$ [arb. units]	$\langle E_0(q) \rangle$ [μeV]	$\langle B(q) \rangle$ [arb. units]	$\langle \sigma(q) \rangle$ [μeV]	$\langle \gamma(q) \rangle$ [μeV]	$\langle$FWHM$\rangle$ [μeV]
0.077	1.003	0.030	0.00102	0.317	0.41	0.76

QENS: D2O Contribution. To simplify the fit of the *E. coli* data, we first analyzed the QENS data of the PBS-D2O buffer at three temperatures: 276 K, 300 K, and 324 K to obtain an estimation of the q-dependence of the intensity of the signal due to the solvent. The main contribution here comes from D2O bulk water whose scattering function can be well represented as follows [105]:

$$S_{D2O}(q,E) = e^{-\frac{E}{2k_B T}} \cdot R(q,E) \otimes \left[\beta_{D2O}(q) \cdot L_{\gamma_{D2O}}(E) \right] \tag{4.4}$$

where β_{D_2O} is the intensity of the signal, and $L_{\gamma_{D2O}}$ is a Lorentzian function that describes the diffusive translational motions of the D2O molecules. An example of the resulting parameters is reported in Fig. 4.2.

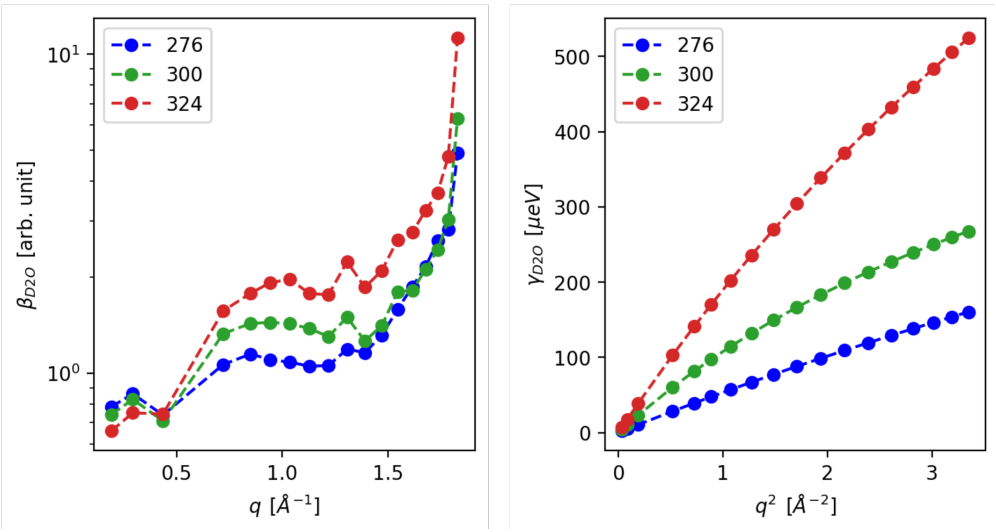

Figure 4.2: PBS-D2O buffer. Resulting parameters of the fit with the eq. (4.4).

The D2O cross section comprises 79% coherent and 21% incoherent scattering and, applying Vineyard's convolution approximation, we have $S_{D2O}(q,E=0) \approx S_{D2O}^{(coh)}(q)$, where $S_{D2O}^{(coh)}$ is the coherent static structure factor [106]. Hence, $S_{D2O}(q,E)$ has as main contribution the coherent scattering of D2O that can not be avoided [107]. It is important to observe that the intensity of $S_{D2O}(q,E)$ that we have measured (Fig. 4.2) reproduces qualitatively well the shape of $S_{D2O}^{(coh)}(q)$ measured by Bosio et al. [108].

QENS: Average Protein Contribution. The dynamic scattering function of the proteins is modeled here by the following expression (see section 2.2.3) [103, 109]:

$$S_{AP}(q,E,T) = I(q,T) \cdot \{L(E;\gamma_G(q,T)) \otimes [A_0(q,T)\,\delta(E) + (1-A_0(q,T))\,L(E;\gamma_L(q,T))]\} \tag{4.5}$$

where $I(q)$ is the intensity of the signal which is related to the vibrational modes of the proteins, meanwhile $L(E;\gamma_G(q,T))$ and $L(E;\gamma_L(q,T))$ are two Lorentzian functions accounting for the diffusive contributions due to the global motions (roto-translations of the entire proteins) and local dynamics (internal motions of sub-parts of the proteins, e.g. conformational changes). Finally, $A_0(q)$, which multiplies the delta function $\delta(E)$, represents the EISF containing information on the geometry of the internal motions that are confined with respect to the space-time window of the spectrometer [52].

QENS: Data analysis. As described in the previous paragraphs, we employed a three step-procedure for the fit of the data.

In the 1st step, we performed a simple fit that treated the measured spectra independently, apart for the weighting factor for the solvent contribution ϕ which was shared among the spectra since it should depend only on the amount of bulk D2O in the samples. On the contrary, all the remaining parameters can vary both with q ant T. As shown in Fig. 4.3 and 4.4, this allows us to study the trends of the Lorentzian widths as function of q^2, and we verify that, both for the global and the local motions, they are well described by a jump-diffusion model as already observed in previous *in vivo* QENS experiments on bacteria (see section 2.2.3) [103, 110, 111]:

$$\gamma_i(q,T) = \frac{q^2 D_i(T)}{1 + q^2 D_i(T)\,\tau_i(T)} \quad \text{for } i \in \{G,L\} \tag{4.6}$$

Within this model, the diffusion is assumed to occur via infinitely small, elementary jumps characterized by a negligible jump time during which the particle diffuses and the residence time τ_i, i.e. the time a proton spends in a given position.

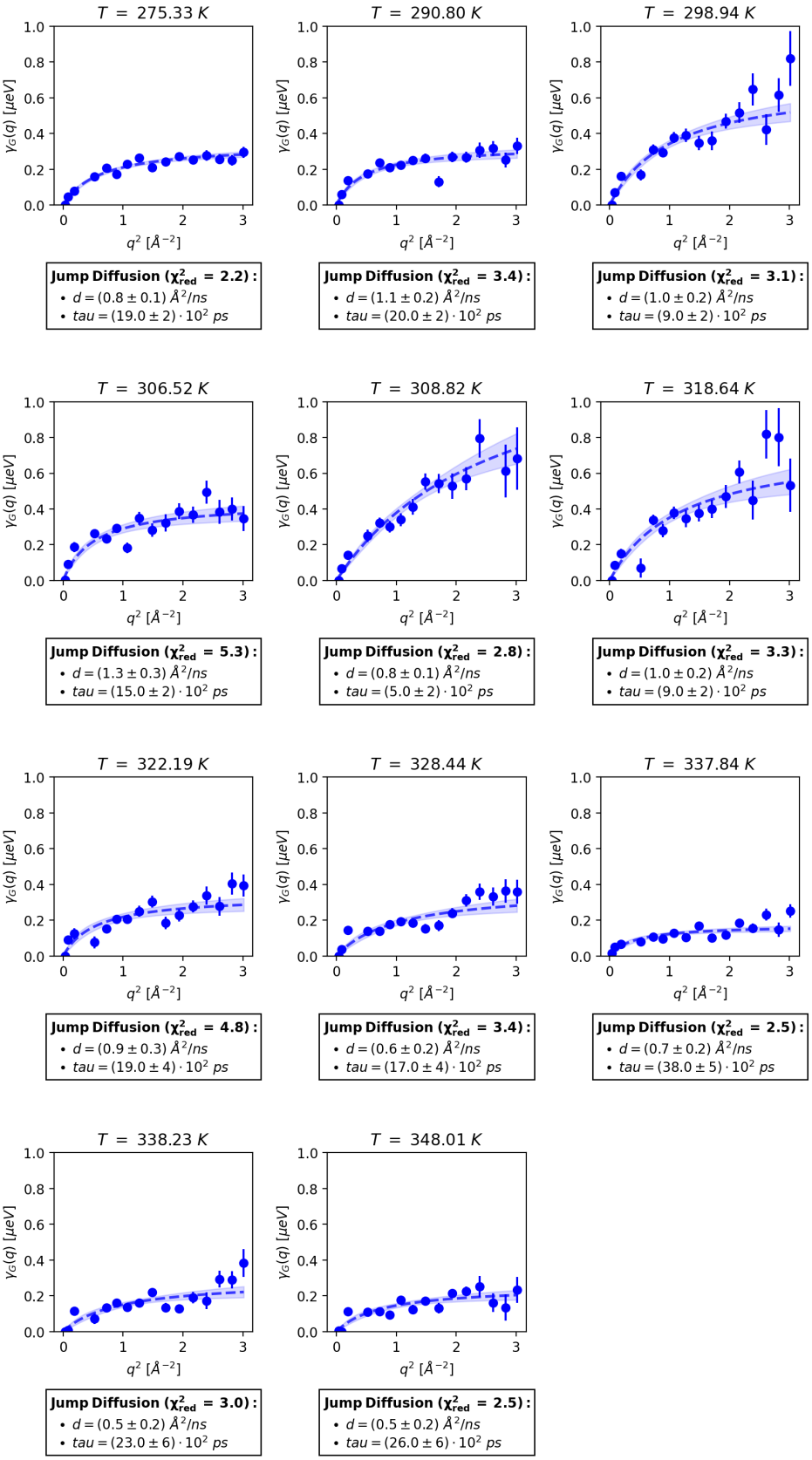

Figure 4.3: Global motions (1st step). Lorentzian widths γ_G vs. q^2 for different temperatures. The dashed lines are fits with the jump-diffusion model described by the eq. (4.6).

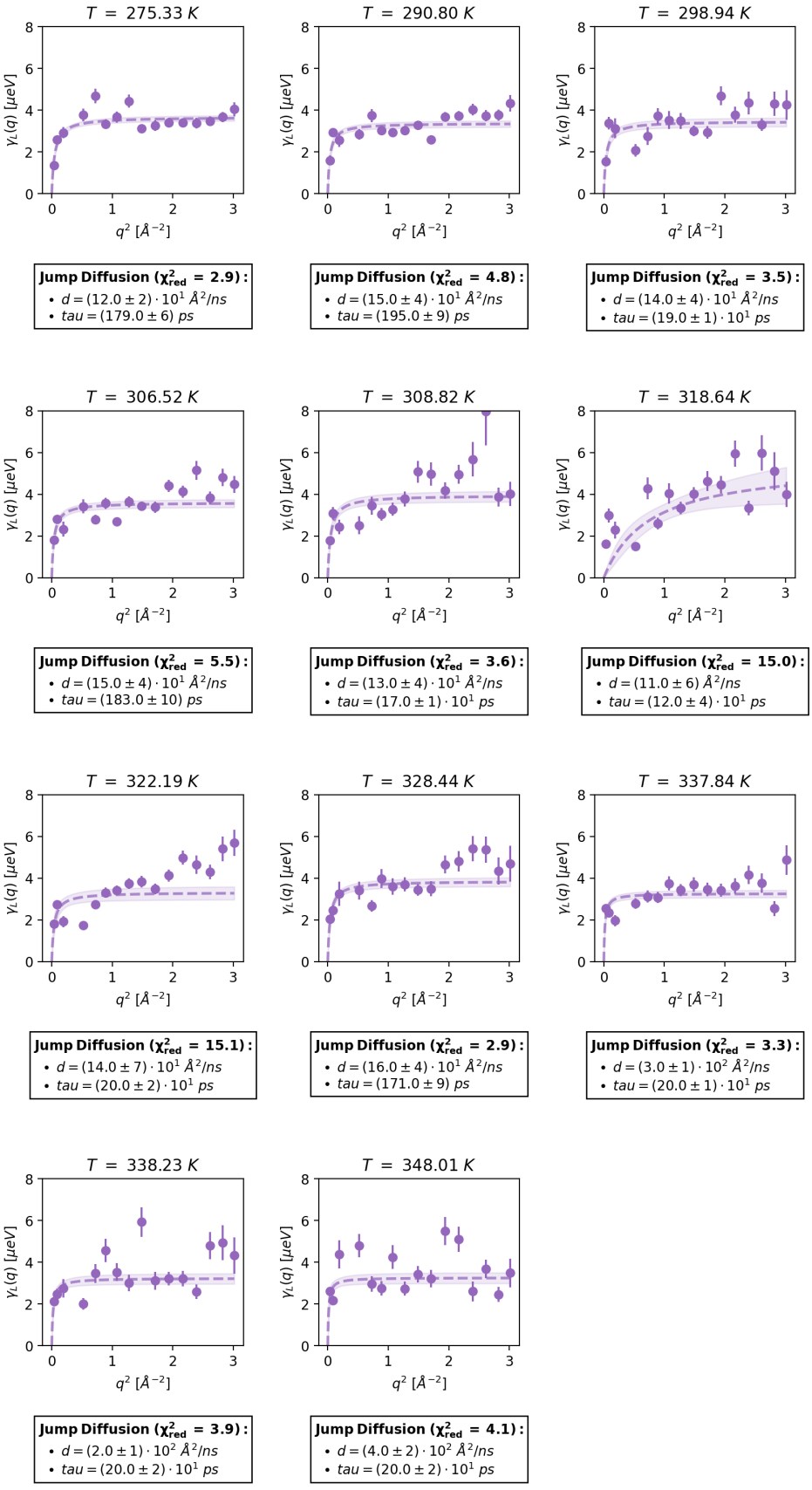

Figure 4.4: Local motions (1st step). Lorentzian widths γ_L vs. q^2 for different temperatures. The dashed lines are fits with the jump-diffusion model described by the eq. (4.6).

In the 2nd step, we performed a simultaneous fit of the spectra measured at different q-values assuming *a-priori* the jump-diffusion model for both the global and the local dynamics, i.e. constraining the q-dependence of γ_G and γ_L with the eq. (4.6). The resulting values of the diffusion coefficients and the residence time for the global and the local dynamics are reported in Fig. 4.5.

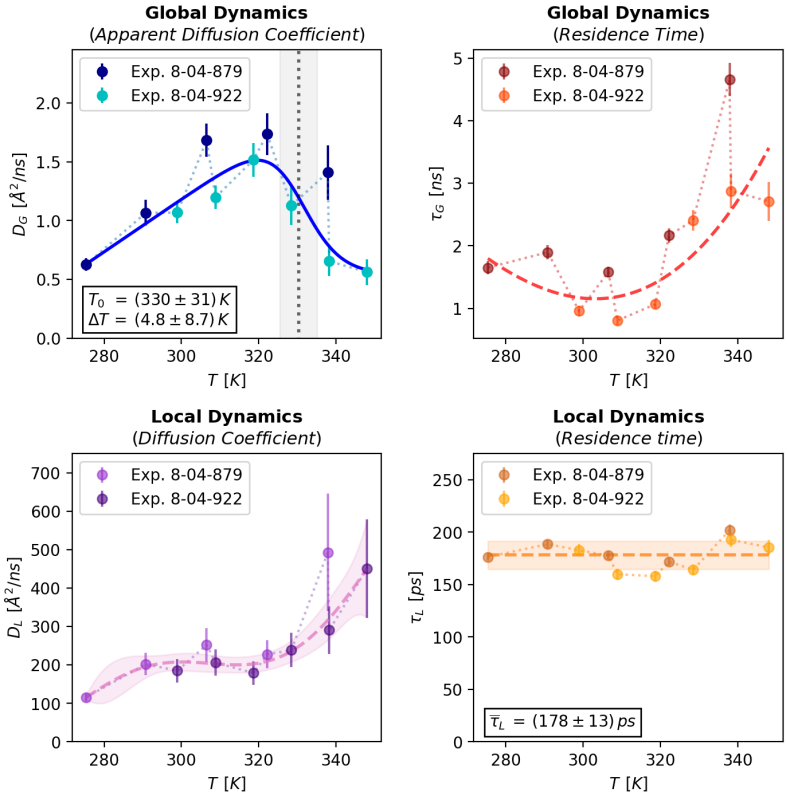

Figure 4.5: Diffusion coefficients and residence times for the global and local motions obtained from the simultaneous fit of the *E. coli* data, assuming *a priori* the jump-diffusion model eq. (4.6). The solid blue line is a fit of D_G with the gelation model described by the eq. (4.7)

Quite remarkably, around the cell-death temperature ($T_{CD} \approx 323.15\,\mathrm{K}$), there is an important reduction of average motion of the entire protein described by D_G. It is possible to model the phenomenon by the following function [66, 109]:

$$D_G(T) = (a_1 T + b_1)\left[1 - a_u^{\mathrm{QENS}}(T)\right] + (a_2 T + b_2)\, a_u^{\mathrm{QENS}}(T) \tag{4.7}$$

with:

$$a_u^{\mathrm{QENS}}(T) = \frac{1}{1 - e^{-(T-T_0)/\Delta T}} \tag{4.8}$$

where a_u^{QENS} is a smeared step function describing the transition between the initial liquid state and the final gel-like state.

In the 3rd and last step, we constrained the temperature dependence of D_G with eq. (4.7), and we performed a simultaneous fit of all the measured spectra. The resulting parameters constant in T and q are reported in table 4.3.

Table 4.3: Parameters shared among all the spectra (i.e. constant in T and q) obtained from the simultaneous fit of the *E. coli* data at all the temperatures with eq. (4.1), taking into account eq. (4.2), eq. (4.4), eq. (4.5), and assuming *a priori* the models described by eq. (4.6) and eq. (4.7).

T_0	(58.7 ± 0.3) C
ΔT	(4.21 ± 0.06) C
a_1	(0.020 ± 0.001) Å$^2/(ns \cdot C)$
b_1	(0.605 ± 0.005) Å$^2/ns$
a_2	$(10^{-10} \pm 0.001)$ Å$^2/(ns \cdot C)$
b_2	(0.48 ± 0.01) Å$^2/ns$
ϕ	(40.1 ± 0.1) %

An example of the resulting parameters for the EISF, $A_0(q)$, at five selected temperatures is shown in Fig. 4.6 (*left*).

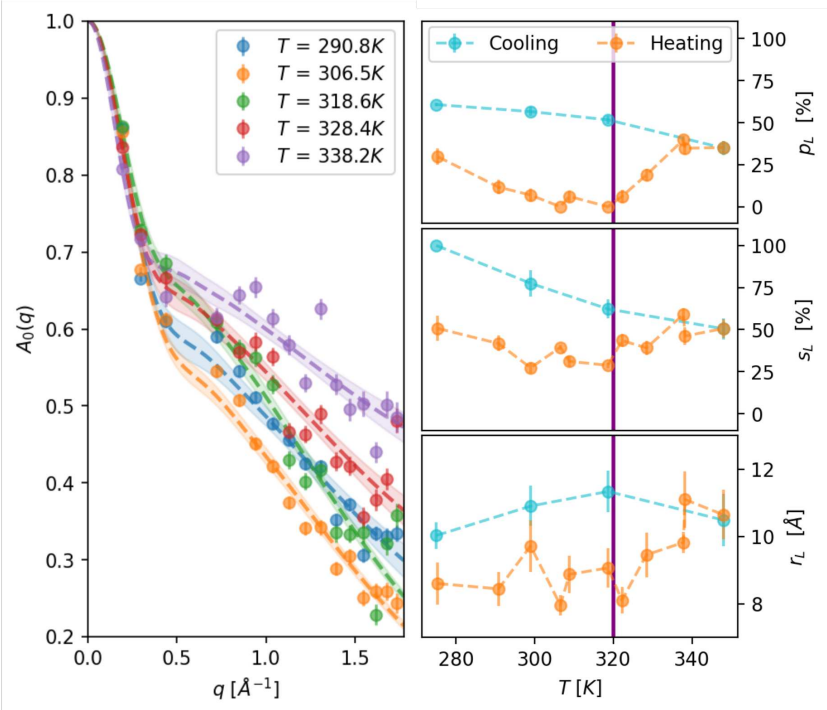

Figure 4.6: *Left*: EISF derived from the global fit of the *E. coli* data. Dashed lines are the best fit of the EISF with the model described by eq. (4.9). *Right*: Parameters resulting from the fit of the EISF. p_L is the fraction of H-atoms appearing fixed on the accessible time scale of the instrument, s_L is the fraction of H-atoms diffusing in a spherical volume with radius r_L.

As it was anticipated in section 2.2.3, the EISF can be described quite well by the model proposed by Grimaldo et al. in previous study on g-globulins [112]:

$$A_0 = A_0(q,T) = p_L(T) + [1 - p_L(T)] \left[s_L(T) A_{\text{sph}}(q,T) + (1 - s_L(T)) A_{\text{3JD}}(q,T) \right] \tag{4.9}$$

where:

$$A_{\text{sph}}(q,T) = \left(3 \cdot \frac{j_1(q \cdot r(T))}{q \cdot r_L(T)} \right)^2 \qquad \text{and} \qquad A_{\text{3JD}}(q,T) = \frac{1 + 2 j_0(q \cdot a_M)}{3} \tag{4.10}$$

p_L is the fraction of H-atoms appearing fixed on the accessible time scale of the instrument. For the remaining H-atoms, two types of motions were taken into consideration: confined diffusion in an impermeable spherical volume of radius r_L, described by the amplitude A_{sph}, and random jump diffusion between three equidistant sites on a circle of radius a_M, arising from the methyl-groups, and described by $A_{\text{3JD}} - a_M$ is fixed to 1.715Å which is the average distance of H-atoms in methyl groups. The relative contribution of these two diffusive processes is measured by s_L, which, more specifically, describes the fraction of H-atoms that, among those atoms that are not appearing fixed (i.e. $1 - p_L$), are undergoing the spherically confined diffusion.

4.2.3 EINS Experiments

As Elastic Incoherent Neutron Scattering (EINS) measurements are much faster, we performed also EINS measurements for much more temperature points on the *E. coli* samples. The elastic data were collected for 30 seconds every 6.5 minutes, meanwhile the sample was first heated from 300K to 350K with a heating rate of 0.25K/min and then cooled down, in 20 minutes, from 350K to 305K [101].

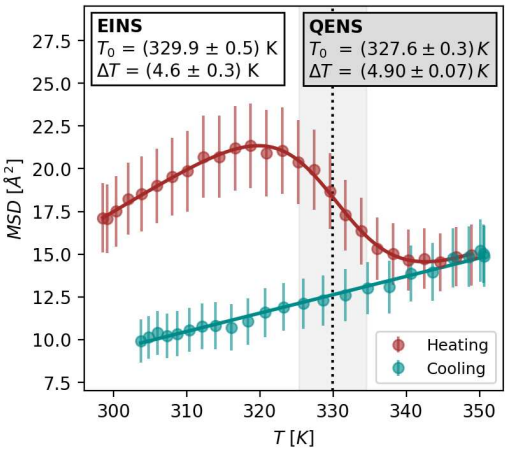

Figure 4.7: EINS MSD fitted with eq. (4.7) describing the gelation of the system. The resulting parameters, T_0 and ΔT, are shown in the box on the upper left. These values are compared with the corresponding parameters obtained from QENS (box on the upper right.

In general, incoherent neutron scattering on a biological sample probes the self-diffusion, therefore the dynamics of hydrogen atoms, which present the highest cross section among the particles contained in living matter [113]. As they are mainly distributed homogeneously in the sample, they allow to probe average molecular dynamics of the H nuclei and of the molecular groups, to which they are bound. Elastic scattering refers to processes within the time limit going to infinity, therefore only short local motions can be resolved. The elastic signal corresponds to the static structure function $S(q, \Delta E)$, which can be written approximately as polynomial in q^2 [66] containing the atomic mean square displacement (MSD) $\langle u^2 \rangle$:

$$S(q, E \approx 0) = e^{-\frac{1}{6} \cdot [a + \langle u^2 \rangle q^2 + c q^4]} \tag{4.11}$$

Taking the logarithm of this expression gives access to $\langle u^2 \rangle$ as function of temper-

ature, which can then be fitted by applying the same gelation model as the one used in eq. (4.7) to fit the global diffusion coefficient. Fig. 4.7 shows the resulting MSD and their fit.

4.2.4 Model preparation for the simulations

Protein composition. To represent the protein composition of the *E. coli* cytoplasm (see section 3.2.4), we built our simulation system on the basis of a previous computational model [114], which was derived from the results of a proteomics study [115] of *E. coli* grown under minimal media conditions. The computational model reported in [114] contained 45 different protein species as well as 5 types of RNA and RNA–protein complexes. For the purposes of the present study and in order to permit back-mapping of smaller sub-volumes into an all-atomistic resolution, we reduced the model in the following aspects. First of all, we considered a smaller cytoplasmic volume than [114], namely a 400 Å cubic box. Second, we focused exclusively on proteins since they are known to form the most abundant macromolecular type in the *E. coli* cytoplasm and since the second most abundant macromolecular type – the RNA – is mainly concentrated in large ribosomal particles, which we did not include in our model. Third, we only considered protein species with molecular weights below 150 kDa to avoid large protein oligomers that would not fit easily in back-mapped sub-volumes. We assumed the same target macromolecular concentration as [114], that is, 275 g/L. Analogously to [114], the number of copies of each individual protein species was based on the relative abundances reported in [115]. From the list of proteins simulated by [114], we omitted those counting less than 0.5 copies per our simulation box. In addition, we did not include the GltD, Hns, Pnp, and Bcp proteins, each of which would only exist in one or two copies in our simulation box, owing to the absence of reliable structures for homology modeling. On the other hand, for computational investigations of protein unfolding that are not reported in this work, we included 10 small monomeric barrels of superoxide dismutase 1 (SOD1) [116], raising the total macromolecular concentration to 279 g/L. Overall, the system comprised 197 proteins of 35 species (see Table 4.4 and Figure 4.8).

Obtaining protein structures. Where available, we used an *E. coli* PDB structure to prepare an atomistic model of each protein species (see section 3.2.4). Unless we found a more recent and higher-quality structure in the PDB database [117], we selected the same PDB structure as was used in the previous work [114]. In four cases, we made use of a homology model stored in the SWISS-MODEL repository [118] as a starting point of the model preparation. Where needed, we subsequently added missing residues and corrected non-proline cis peptide bonds using the MODELLER software, version 9.23 [119]. The quality of the structure was checked with the MolProbity server [120]. When present in the PDB structure, metal ions coordinated to the protein were included in the model; however, larger ligand molecules were omitted. With the exception of TufA and GpmA, the oligomerization state of the protein was taken to be the same as in [114]. While TufA was modeled as a dimer by [114], the prediction of the PDBe PISA server [121] yielded an ambiguous result, and the protein was described as a monomer in a previous experimental work [122]. Therefore, we modeled TufA in a monomeric state. Similarly, while the GpmA protein was simulated in a dimeric form by [114], the

PDBe PISA prediction placed it in a grey zone, and according to the EcoCyc database [123], the ATP-free version of the protein was monomeric. As a result, we considered GpmA in a monomer state.

Table 4.4: Protein composition of the LBMD simulation together with the LBMD coupling coefficient γ for each protein.

	Protein	**PDB-Structure**	**MW** [kDa]	**Count**	γ [10^{-3} fs^{-1}]
1	TufA	1DG1	43.3	67	1.1920
2	MetE	SWISS 4zty.1.A	84.5	39	0.6950
3	CspC	SWISS 3i2z.1.B	7.3	13	3.8150
4	IcdA	1P8F	91.5	8	0.6478
5	Pgk	1ZMR	41.1	5	1.1665
6	FusA	4V9O	77.4	4	0.7346
7	Eno	1E9I	91.0	3	0.6504
8	GlyA	1DFO	90.6	3	0.6526
9	IlvC	1YLR	54.1	3	0.9580
10	Adk	1AKE	23.6	3	1.4770
11	Efp	6ENU	20.6	3	1.8653
12	GapA	1S7C	142.1	2	0.4600
13	Ppa	2EIP	117.4	2	0.5357
14	SerC	1BJN	79.3	2	0.7217
15	DnaK	5NRO	69.0	2	0.8008
16	CysK	5J43	68.7	2	0.8031
17	Mdh	2PWZ	64.7	2	0.8400
18	Tsf	1EFU	60.6	2	0.8815
19	Tig	1W26	48.2	2	1.0413
20	SodA	1D5N	45.9	2	1.0779
21	Upp	2EHJ	45.1	2	1.0928
22	SodB	1ISC	42.3	2	1.1438
23	GpmA	1E59	28.4	2	1.5070
24	Hup	2O97	18.8	2	1.9797
25	Suc	1JLL	142.0	1	0.4628
26	PurA	1ADE	94.4	1	0.6326
27	Asd	1BRM	80.0	1	0.7167
28	Fba	5VJE	78.0	1	0.7305
29	PurC	2GQR	54.0	1	0.9591
30	TpiA	4IOT	53.9	1	0.9597
31	RpiA	1KS2	45.5	1	1.0860
32	UspA	SWISS 1jmv.1.A	31.9	1	1.3938
33	Frr	1EK8	20.6	1	1.8626
34	PpiB	2RS4	18.2	1	2.0213
35	SOD1	4BCZ	11.0	10	2.7740

The atomistic protein structures were hydrogenated by the GROMACS `pdb2gmx` tool [124], placed in rectangular boxes with a minimum distance of 1.5 nm between the protein and the boundary of the box, and solvated in a 150 mM KCl solution, including several additional ions to neutralize the net charge of the system, if needed. Each

system was then subjected to a short energy minimization and equilibration protocol, performed in the GROMACS 2018.7 software [124]. First, an energy minimization was performed to bring the maximum force below 1000 kJ mol^{-1} nm^{-1}, while harmonic restraints were applied to all heavy atoms of the protein (with a force constant k_{BB} = 1000 kJ mol^{-1} nm^{-2} assigned to the backbone atoms and a force constant k_{SC} = 500 kJ mol^{-1} nm^{-2} applied to the side-chain atoms). This energy minimization was followed by a six-step relaxation protocol with gradually decreasing harmonic position restraints, where the first two short simulations were performed in the *NVT* ensemble and were followed by four *NPT* trajectories simulated at a pressure of 1.01 bar (see Table 4.5 for more details). The minimization and equilibration protocol was performed separately for the Amber99SB-disp [125] and CHARMM36m [126] force fields. The same simulation parameters were used as for the respective Amber99SB-disp and CHARMM36m systems described in the subsection 4.2.7, except for the temperature coupling, which was ensured by the Berendsen thermostat [127] with a time constant of 1 ps. The equilibrated protein structures obtained with the Amber99SB-disp force field (Figure 4.8) were converted to the OPEP description (see section 3.3.2) by the OPEP File Generator available on http://opep.galaxy.ibpc.fr.

Initial positions and orientations of the proteins in the simulation box were generated by the Packmol software [128]. To make the spatial distributions of different protein species more uniform, each protein was treated separately in Packmol.

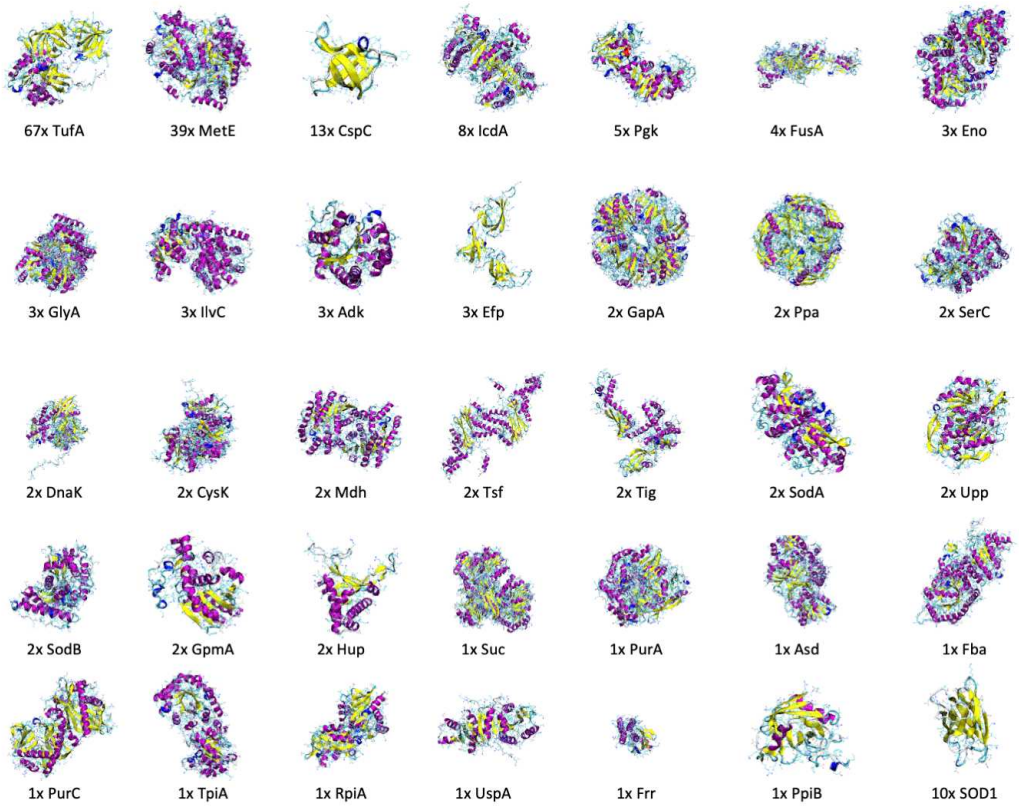

Figure 4.8: Structures and counts of proteins included in the LBMD simulation.

Table 4.5: Overview of an equilibration protocol used for an initial relaxation of isolated protein geometries. k_{BB} and k_{SC} – force constants of harmonic restraint potentials applied to the heavy atoms of the backbone and the side chains, respectively. Δt – time step.

Step	**Length** [ns]	Δt [fs]	**Restraints**	$k_{BB}\left[\frac{kJ}{\text{mol nm}^2}\right]$	$k_{SC}\left[\frac{kJ}{\text{mol nm}^2}\right]$
1	0.1	1	heavy atoms	1000	500
2	0.1	1	heavy atoms	500	250
3	0.1	1	heavy atoms	250	100
4	0.5	2	heavy atoms	100	50
5	0.5	2	heavy atoms	50	0
6	0.5	2	none	0	0

4.2.5 LBMD simulation

We performed a coarse-grained simulation of the large system using the Lattice Boltzmann Molecular Dynamics (LBMD) approach (see section 3.3.3) [129], implemented in-house in the MUPHY software [130]. Our LBMD scheme, which has been successfully applied to a number of biological systems [86, 131, 132, 133], combines a coarse-grained protein model with a lattice-based description of hydrodynamic interactions [129]. The protein model was based on the OPEP v.4 force field [85]. Since the goal of this coarse-grained simulation was to sample different intermolecular arrangements rather than simulate protein conformational changes, the conformations of the proteins were restrained by an elastic network (distance cutoff 6 Å, force constant 5 kcal mol^{-1} Å^{-2}). The primary focus on intermolecular interactions also allowed us to simplify the description of the protein backbone, which was exclusively represented by C_*alpha* beads, the sizes of which were increased by 50 % relative to the standard OPEP v.4 model. Finally, the OPEP sidechain–sidechain non-bonded interactions were rescaled by a factor of 0.857 to mitigate excessive aggregation [134].

The simulation was performed in an NVT ensemble at $T = 300$ K, using a time step of 10 fs for bonded interactions and a time step of 20 fs for non-bonded and hydrodynamic interactions. A trajectory length of 4.3 μs was reached. Hydrodynamic interactions were described using the LB technique [88], employing the BGK (Bhatnagar-Gross-Krook) collisional operator [135], with a lattice grid spacing of 4 Å. The LB kinematic viscosity ν_0 was set to reproduce bulk water behaviour at ambient conditions. The coupling coefficient γ was determined for each protein species as a function of its molecular weight (see Figure 4.9). The dependence of γ on molecular weight was derived by performing separate simulations of six protein species (namely CspC, SOD1, Adk, TufA, MetE, and GapA) in dilute conditions and by adjusting the γ of each species to match the diffusion coefficient of each isolated species with a prediction obtained with the HYDROPRO software [136].

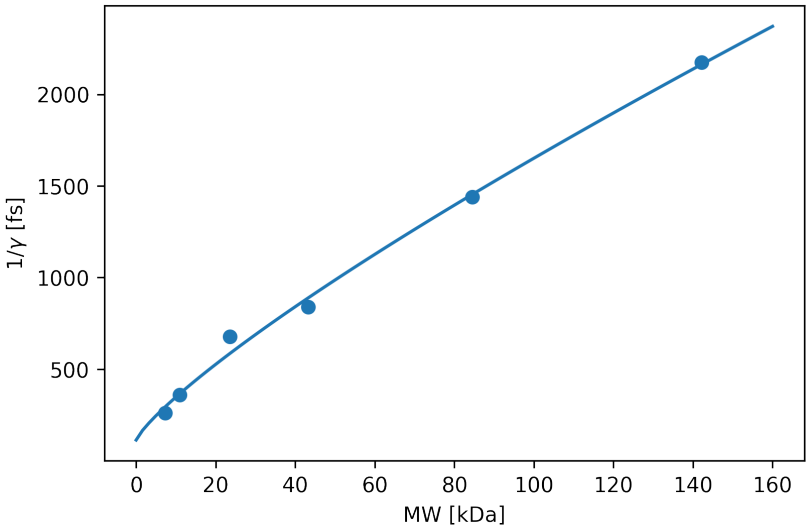

Figure 4.9: Dependence of the LBMD coupling coefficients γ on protein molecular weight. The γ values for six proteins (CspC, SOD1, Adk, TufA, MetE, and GapA) were optimized in order for the diffusion coefficient of the given isolated protein in an LBMD simulation to match the HYDROPRO [137] prediction. To determine the γ values for the remaining 29 protein species, the resulting dependence of $1/\gamma$ on molecular weight (MW) was fitted with a power-law dependence $1/\gamma = a \cdot \mathrm{MW}^b + c$, yielding the parameter values $a = 35.76222504$, $b = 0.81676473$, and $c = 113.05584831$ (for MW expressed in kDa and γ in fs^{-1}).

4.2.6 Sub-volume selection and back-mapping

To explore the diffusion of proteins with an all-atom resolution, allowing the description of conformational flexibility and unfolding, we selected five sub-volumes sampled in the LBMD trajectory and converted them into an atomistic resolution.

The selection of the sub-volumes proceeded in the following way. A cube with an edge length of 170 Å was placed in randomly chosen trajectory frames over the first 3 μs of the LBMD trajectory, with random positions and orientations (60,000 trials). In the dense environment of the cytoplasmic system, it was practically impossible to find sub-volume positions and orientations so as to avoid cutting at least a few proteins by the sub-volume boundary. For each placement, proteins that were fully contained in the sub-volume were counted as well as those that were cut by the sub-volume boundary. Subsequently, we isolated 120 sub-volumes that minimized the ratio of the total mass of proteins cut by the sub-volume boundary versus the total mass of proteins entirely placed inside the sub-volume. If we had only kept those proteins that were not cut by the sub-volume boundary, we would have systematically underestimated the protein concentration in the sub-volumes. Therefore, for each of the 120 sub-volumes, the proteins that were integrally inside the sub-volume were complemented by a subset of proteins that were cut by the boundary. This subset was constructed in the following way. First of all, we only retained proteins the periodic images of which did not overlap with any protein that was integrally placed inside the sub-volume. An overlap was defined as a close contact between two protein beads, with a distance less than 3 Å. Next, we checked for overlaps among the proteins that we retained in the previous step, and we removed, one by one, proteins that exhibited the largest number of overlaps

with the remaining members of the subset. In this way, we ended up with a subset of proteins that had no overlap with any other protein in the sub-volume.

Out of the 120 sub-volumes constructed with the procedure described above, we selected a representative set of five sub-volumes that 1) approximated the distribution of local protein concentrations inside the large volume and that 2) maximized the diversity of protein species represented in the set while approximating the number distribution of protein species in the large volume. The distribution of local protein concentrations (see Figure 4.21B) was obtained by randomly placing a 170 Å cube over the trajectory (2000 trials) and by evaluating the protein concentration inside the cube for each placement.

The protein concentration and species composition of each of the five cubic sub-volumes, sampling the structure and local concentration of the crowded solution, is indicated in Table 4.6.

Table 4.6: Protein concentration and species composition of the five cubic sub-volumes extracted from the LBMD trajectory.

Conc. [g/L]	**Protein List**
134.0	CspC (x4), TufA (x3), MetE, Mdh, FusA, SOD1
197.5	TufA (x4), MetE (x2), CspC, GlyA, CysK, Mdh, SOD1
279.8	TufA (x2), GapA, GpmA, MetE (x3), CspC, CysK, SodA, IlvC, DnaK, Frr, SOD1 (x3)
287.7	TufA (x3), MetE (x4), IcdA (x2), CspC (x2), PpiB, Mdh, IlvC, Efp, GpmA
297.7	TufA (x8), MetE, CspC (x2), IlvC(x2), SOD1 (x2), Ppa, GlyA, Adk, GpmA, RpiA

Each sub-box, with a size of 170 Å and containing between 11 and 27 proteins, was subsequently converted into the all-atom resolution using the following back-mapping protocol. Atomistic structures prepared as described in the paragraph "*Obtaining protein structures*" were overlapped with the geometries from LBMD by aligning the C_*alpha* atoms of the atomistic structures with the corresponding C_*alpha* beads. We checked the back-mapped configuration for the presence of entangled aromatic side chains and corrected this artifact if needed. The boxes were subsequently hydrated, and a 150 mM concentration of K^+ and Cl^- ions was added, including extra ions neutralizing the net charge of the proteins. The systems were then minimized and equilibrated using the protocol described below.

4.2.7 All-atom simulations

Force field and simulation parameters. All-atom molecular dynamics (MD) simulations were performed using the GROMACS 2019.4 software [124]. We employed two distinct sets of force field parameters to describe the proteins (see section 3.2.1): Amber Amber99SB-disp [125] and CHARMM36m [126], which belong to the most recent generation of protein force fields and which were optimized to correctly capture the properties of unfolded ensembles. The Amber99SB-disp protein force field was coupled with the Amber99SB-disp water model, while CHARMM36m used the TIP3P water model [138]. In both cases, K^+ and Cl^- ions were described with the default parameters for the respective force field. A 1.2 nm cutoff was applied to short-range

non-bonded interactions. In addition, van der Waals forces were smoothly switched to zero between 1.0 and 1.2 nm in the CHARMM36m simulations. Long-range electrostatic interactions were evaluated using the particle mesh Ewald method [139]. Newton's equations of motion were propagated using the leap-frog algorithm [140]. The LINCS algorithm [141] was used to constrain the lengths of all protein bonds involving hydrogen, and the SETTLE algorithm [142] was employed to keep water molecules rigid.

The temperature of the system was maintained by the velocity rescaling thermostat with a stochastic term [143], which was coupled separately to the proteins and to the rest of the system with a time constant of 0.1 ps. The pressure was kept at 1.01 bar by the Parrinello-Rahman barostat [144] in all the production runs and by the Berendsen barostat [127] in all the *NPT* equilibration steps, with the time constant equaling 2 ps for both barostats.

Sub-volume equilibration. The procedure that we employed to equilibrate the atomistic sub-volumes at different temperatures and both in folded or unfolded states is detailed in Figure 4.10). First, the configurations back-mapped from LBMD underwent a short initial relaxation (Figure 4.10B). Subsequently, a heating protocol (Figure 4.10C) was used to equilibrate folded sub-volumes at increased temperatures, ranging between 310 and 380 K. In parallel to this, an unfolding protocol (Figure 4.10D) and a subsequent cooling protocol (Figure 4.10E) were employed to produce unfolded sub-volumes and equilibrate them at different temperatures in the range between 300 and 380 K. To efficiently heat up or cool down the system in the heating and cooling protocols, a heating sequence was used [145], enabling us to quickly reach the desired temperature while allowing the system to equilibrate. To follow the unfolding of the proteins during the unfolding protocol, we measured the root-mean-square deviation (RMSD) of each protein's atoms, its radius of gyration (R_g), as well as the amount of secondary structure present in the system, as determined by the DSSP software (Figure 4.11) [146, 147].

Preparation of partially unfolded sub-volumes. The partially unfolded sub-volumes were derived from the same initial configuration as the 287 g/L sub-volume (Table 4.6). The proteins selected for unfolding are listed in Table 4.7. The unfolded fraction r_u was quantified by computing the fraction of hydrogen atoms belong to the unfolded protein structures. The sub-volume preparation followed a protocol equivalent to that employed to obtain fully unfolded sub-volumes (Figure 4.10). In addition, the positions of the atoms in proteins deemed to remain folded were kept frozen during the unfolding protocol (Figure 4.10D) as well as during the three pre-cooling stages of the cooling protocol (Figure 4.10E). Moreover, before the subsequent 100 ns *NPT* equilibration at 360 K, the frozen folded structures were gradually relaxed using the same six-step *NVT* equilibration as in the initial relaxation protocol (Figure 4.10B); however, in contrast to the initial relaxation, the decreasing harmonic position restraints acted on the atoms of the folded structures only.

Production simulations. The production runs following the equilibration procedure (see section 3.3.1) were performed at eight different temperatures (300, 310, 320, 330,

340, 350, 360, and 380 K) both for folded and unfolded sub-volumes with a time step of 2 fs and each reaching a trajectory length of 102.4 ns. For the 288 g/L sub-volume simulated using the CHARMM36m force field, the production trajectories performed at $T = 330$ K were extended to 1 μs for further analysis (see Figure 4.23).

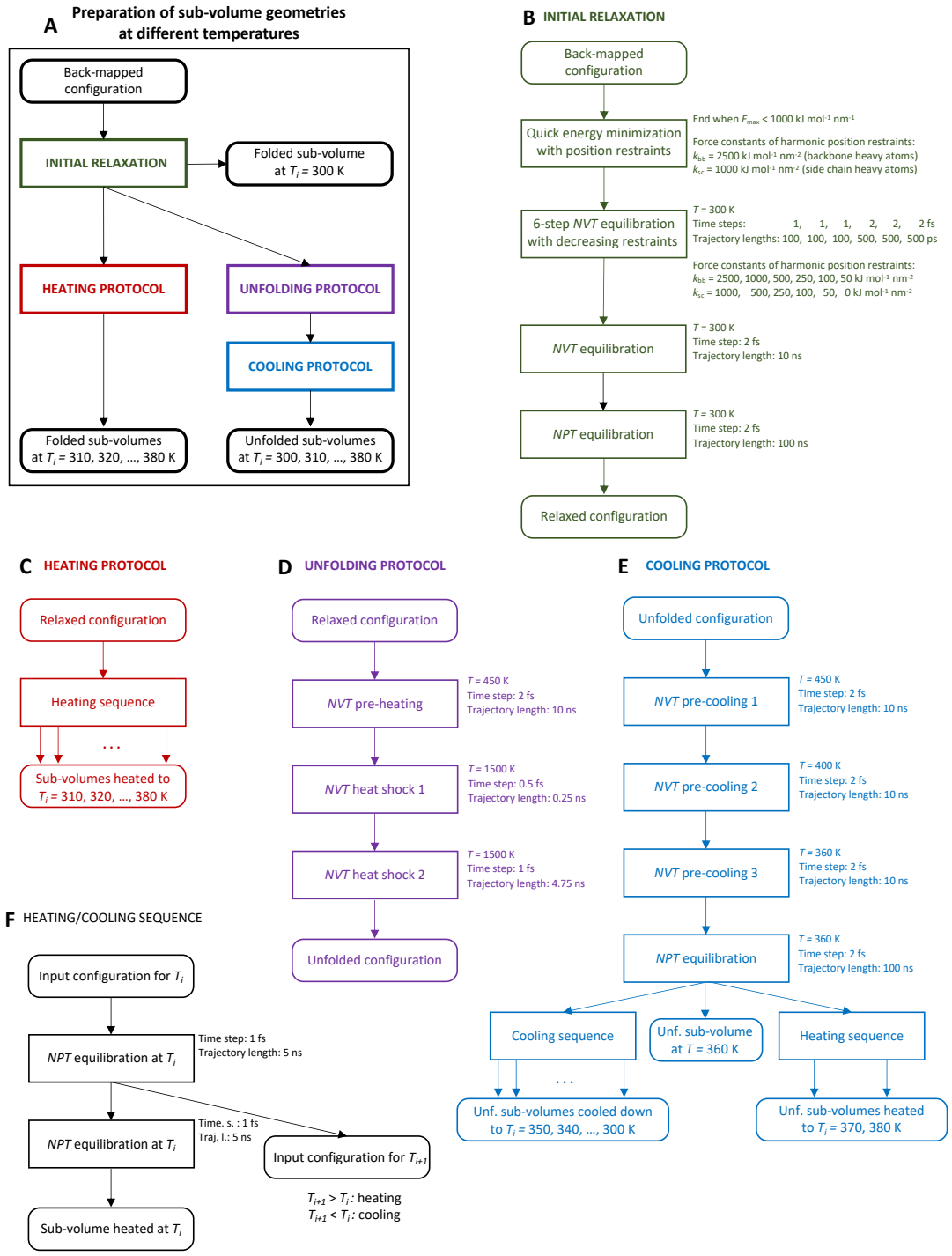

Figure 4.10: Schematic representation of the procedure used to equilibrate atomistic sub-volumes at different temperatures and in different folding states. (A) General overview of the procedure, (B–F) detailed description of its individual parts.

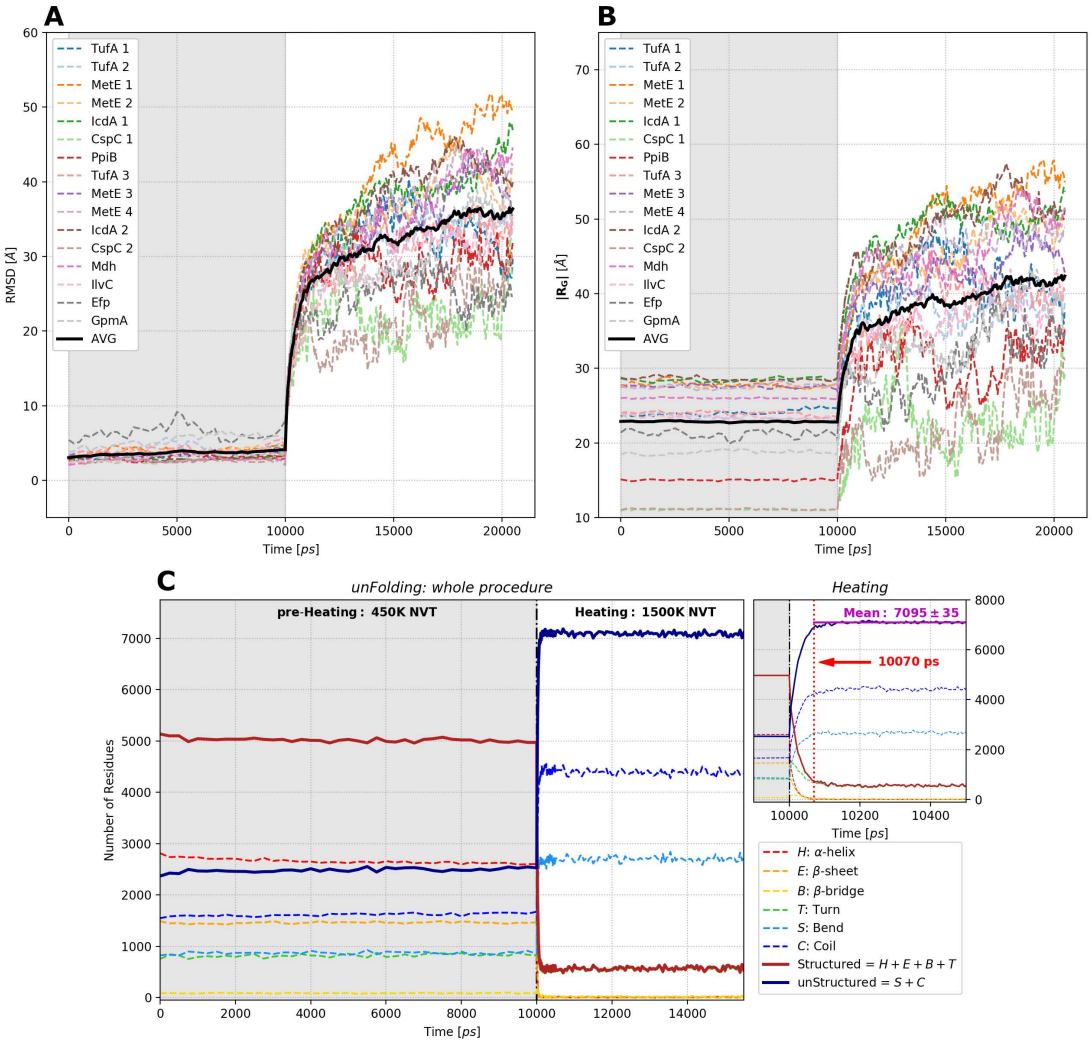

Figure 4.11: Monitoring the unfolding protocol for each protein in the sub-volume with the concentration of 287.7 g/L. (A) Root-mean-square deviation, RMSD, of the proteins. (B) Modulus of radius of gyration. (C) Number of residues per secondary structure, as determined by the DSSP software: the graph in the upper right corner is a magnification of other main plot on the left to focus on the effects of the heat-shock at 1500 K.

Table 4.7: Protein composition of partially unfolded sub-volumes with a varying fraction r_u of unfolded proteins.

$\mathbf{r_u}$ = 25 %
UNFOLDED (5): CspC, IlvC, MetE, PpiB, TufA
FOLDED (11): CspC, Efp, GpmA,2x IcdA, Mdh, 3x MetE 2x TufA

$\mathbf{r_u}$ = 50 %
UNFOLDED (8): CspC, IcdA, IlvC, 2x MetE, PpiB, 2xTufA
FOLDED (8): CspC, Efp, GpmA, IcdA, Mdh, 2x MetE, TufA

$\mathbf{r_u}$ = 75 %
UNFOLDED (11): CspC, 2x IcdA, IlvC, 3x MetE, PpiB, 3x TufA
FOLDED (5): CspC, Efp, GpmA, Mdh, MetE

4.2.8 Calculation of D_t and D_r

Obtaining average D_t per sub-volume. To evaluate the average protein translational diffusion coefficients D_t for a given sub-volume, folding state, and temperature, the production trajectory was divided into 20 ns blocks. In each block, we first calculated the mean squared displacement (MSD) of the center of mass of each protein molecule. We then computed the average of these MSD curves weighted by the numbers of atoms of the proteins (see Figure 4.12). This average MSD curve was subsequently fitted with a straight line in the 0.3–5 ns regime, and D_t was extracted from the slope of the fit. Subsequently, we determined the average of D_t over the 20 ns blocks as well as the corresponding standard error of the mean. For the given sub-volume, folding state, and temperature, the resulting D_t was corrected for the effects of periodic boundary conditions (PBC) (see the section 4.2.9). The estimated error of the PBC correction was then added to the standard error of the mean determined from the block averaging to express the uncertainty of D_t.

To obtain a final value of D_t characterizing the entire cytoplasm, we calculated an average of the PBC-corrected D_t's over all the sub-volumes, with the individual D_t's being weighted by the numbers of protein hydrogen atoms in the sub-volumes. An analogous weighted average of the error bars was performed to estimate the uncertainty of the final value of D_t.

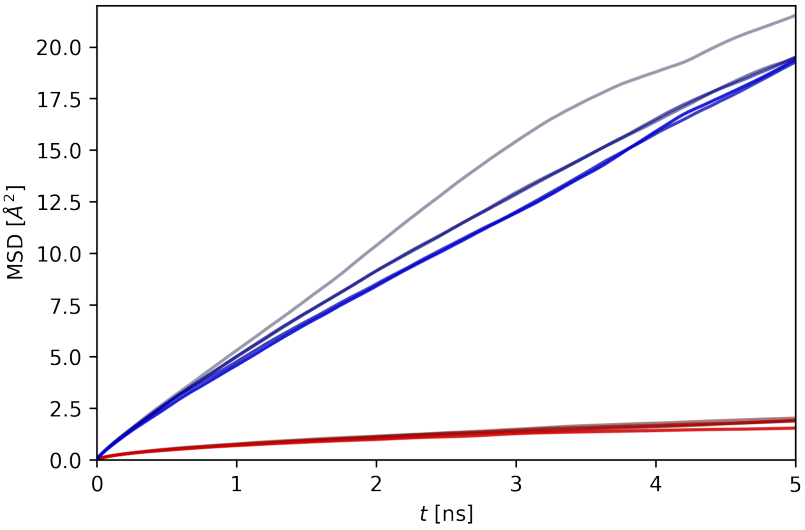

Figure 4.12: Examples of MSD curves in a folded- (blue) and an unfolded (red) system. The MSD curves were computed as a weighted average over all protein center-of-mass MSDs in 20 ns blocks of a production trajectory performed at $T = 300$ K for the 288 g/L sub-volume using the Amber99SB-disp force field.

Calculating D_t and D_r for the evaluation of D_{app}. To compute the apparent diffusion coefficient D_{app} (see section 2.2.3), a quantity directly comparable with the experimental observable, it was necessary to calculate D_t and the rotational diffusion coefficient D_r separately for each individual protein chain. The reason for considering individual chains rather than entire protein molecules lies in the fact that upon unfolding, the different sub-units of a protein may dissociate, and as a consequence, the evaluation of D_r for the entire molecule would no longer be meaningful.

For each protein chain, D_r was computed by using the following approach [148, 149], approximating the chain rotation as being isotropic. One thousand randomly distributed unit vectors were centered and rotated along with the protein chain. This was done – for each frame – by fitting the geometry of the protein chain to a reference structure to which the fixed orientations of the unit vectors had been set. An autocorrelation function of the unit vectors' directions was then calculated by employing the `gmx rotacf` tool [124] and using the second-order Legendre polynomial for evaluating the autocorrelation. The resulting autocorrelation curve was fitted with an exponential function in the 0.3–5 ns regime to obtain the decay time τ which was subsequently used for calculating D_r as $D_r = 1/(6\tau)$.

To achieve the best possible convergence, the single-chain D_t and D_r were calculated from the entire trajectory without cutting it into multiple blocks.

4.2.9 Correcting D_t and D_r for PBC effects

To correct the translational diffusion coefficients for PBC effects, we added the following term [150] to D_t computed from simulations:

$$D_t^{\text{corr}} = \frac{k_{\text{B}} T \xi}{6\pi \cdot \eta(T,c) \cdot L} \tag{4.12}$$

where $\xi = 2.837297$, L is the side length of the cubic box, and $\eta(T,c)$ is an empirical function that we parameterized on the basis of viscosity values computed from simulations (this procedure is described in details in the following section). The uncertainty in the value of $\eta(T,c)$ was propagated to estimate the error of the correction term D_t^{corr}.

Similarly, we computed the correction to the rotational diffusion coefficient as [151]

$$D_r^{\text{corr}} = \frac{k_{\text{B}}T}{6 \cdot \eta(T,c) \cdot L^3} \tag{4.13}$$

4.2.10 Viscosity calculations

We estimated the low-frequency, low-shear viscosity η of the 288 g/L sub-volume (see Table 4.6) at three different temperatures (300 K, 320 K, and 380 K) for both the folded- and the unfolded systems. We also considered the intermediate unfolded systems with 25%, 50%, and 75% of unfolded proteins (see Table 4.7) at $T = 300$ K. All the calculations were repeated for both sets of force field parameters employed in this study.

To calculate η, we followed a similar approach as was used in [152]. For each system and temperature, we extracted 30 snapshots (50 for the unfolded CHARMM36m system at 300 K) in 1 ns intervals from the final part of the respective production run. Each of those snapshots served as a starting point of a 1 ns NPT equilibration followed by a 10–30 ns NVT simulation (10 ns for the folded systems, 20 ns for the rest of the Amber99SB-disp systems, and 30 ns for the remaining CHARMM36m systems). The NVT simulation was used to sample the elements of the pressure tensor, which was saved in 10 fs intervals. The viscosity was calculated from an integral of the autocorrelation functions of the pressure tensor fluctuations. Namely, for each off-diagonal element of the pressure tensor ($P_{ij} = P_{xy}$, P_{xz}, and P_{yz}) and for each of the three combinations $P_{ij} = (P_{xx} - P_{yy})/2$, $(P_{xx} - P_{zz})/2$, and $(P_{yy} - P_{zz})/2$ of its diagonal elements, we computed the running integral

$$\eta_{ij}(t) = \frac{V}{k_B T} \int_0^t \langle P_{ij}(0) P_{ij}(\tau) \rangle \, \mathrm{d}\tau \tag{4.14}$$

where V represents the volume of the simulation box. For each ij, the average of $\eta_{ij}(t)$ over all the NVT simulations was fitted with an analytical function corresponding to a bi-exponential decay of the autocorrelation function:

$$\eta_{ij}(t) = A\,\alpha\,\tau_1(1 - e^{-t/\tau_1}) + B\,(1-\alpha)\,\tau_2(1 - e^{-t/\tau_2}) \tag{4.15}$$

where A, B, α, τ_1, and τ_2 were parameters of the fit. The cutoff time t_{cut} for the fitting was determined as the time where the standard error of the mean of $\eta_{ij}(t)$ for the given ij exceeded 20% of the mean. Examples of $\eta_{ij}(t)$ and the fits can be found in figure 4.13. The η_{ij} was then calculated as the limit of $\eta_{ij}(t)$ at infinity. Finally, the value of η was obtained as the mean of the six η_{ij} values, and the uncertainty of η was determined as the standard error of the mean of the six η_{ij} values.

To obtain a dilute reference, we also calculated the viscosity of an aqueous solution of 150 mM KCl considering the two respective water models—the Amber99SB-disp water model and the TIP3P model—and the same three temperatures as above. In each

case, we performed a 100 ns NVT simulation of a 5.1 nm cubic box after a short energy minimization and a 1 ns NPT equilibration. We split the simulation into 10 ns blocks and computed the autocorrelation functions $\langle P_{ij}(0)P_{ij}(t)\rangle$ separately for each block. Next, we averaged the autocorrelation functions over ij for each block and calculated the running integral $\eta(t)$ from the average. Subsequently, the average of $\eta(t)$ from all the blocks was fitted—using a 10 ps cutoff—with the same analytical function as shown in Eq. (4.15). Analogously to the case of the crowded sub-volumes, η was computed as the limit of the analytical function at infinity.

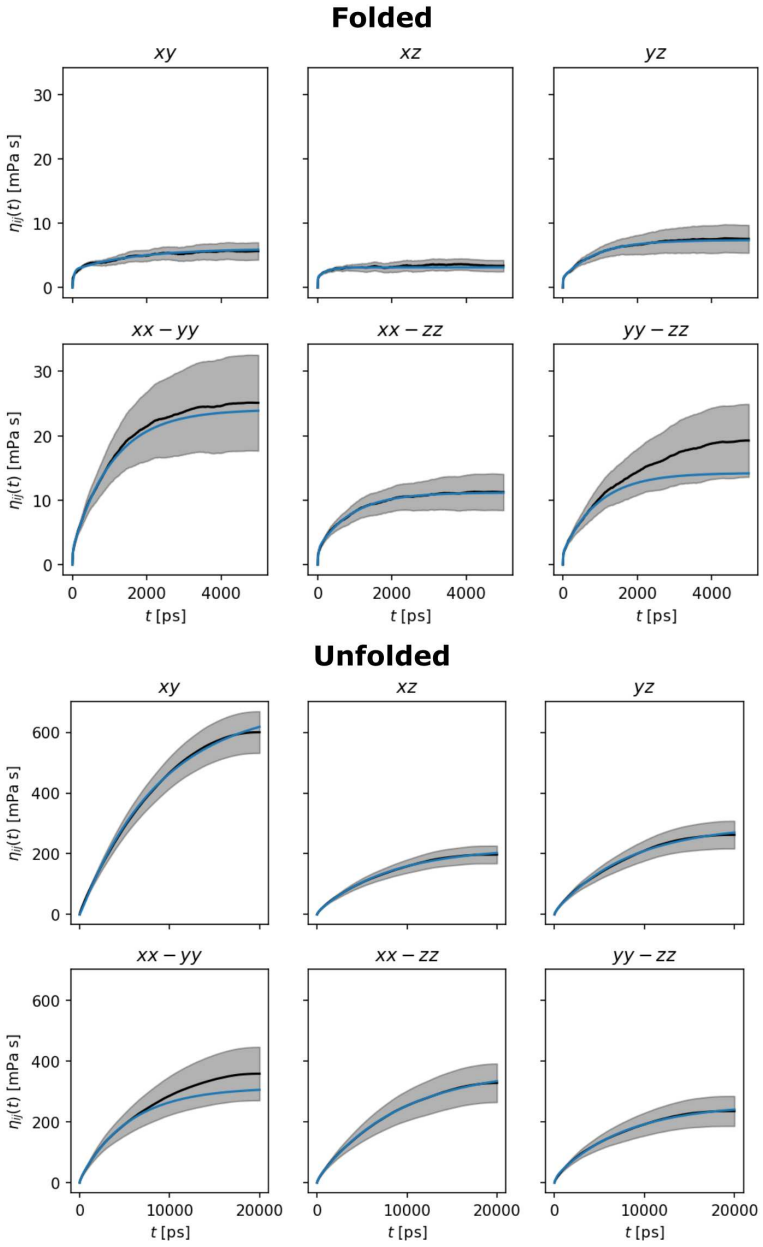

Figure 4.13: The $\eta_{ij}(t)$ running integrals (Eq. (4.14)) obtained for the 288 g/L CHARMM36m sub-volume at $T = 300$ K. The black curve shows the average of $\eta_{ij}(t)$ over all the NVT simulations, the gray stripe represents the standard error of the mean, and the blue curve is a fit using the analytical function (4.15).

To estimate the correction to the diffusion coefficients due to periodic boundary conditions, we interpolated the viscosity to intermediate temperatures and extrapolated it to different protein concentrations by using analytical expressions derived in [153]. First, we interpolated the temperature dependence of the solvent viscosity $\eta_0(T)$ using the following expression [153]:

$$\eta_0(T) = \exp\left(-B_\mathrm{w} + D_\mathrm{w}T + \frac{\Delta E_\mathrm{w}}{RT}\right) \tag{4.16}$$

The curves resulting from these interpolations are shown in Figure 4.14, and the corresponding values of the parameters B_w, D_w, and ΔE_w are listed in Table 4.8.

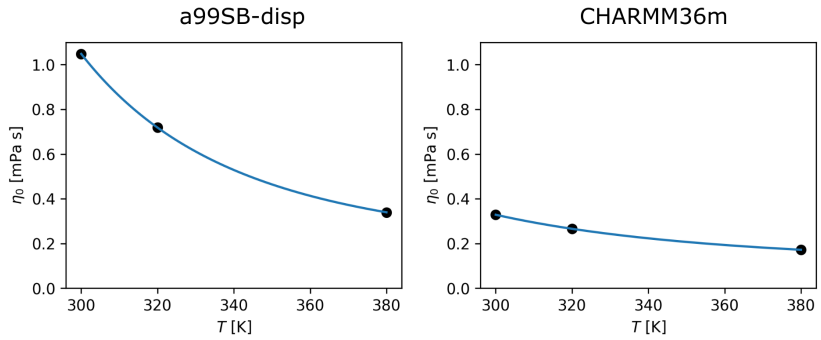

Figure 4.14: Solvent viscosities determined from simulations. The solvent was modeled as a 150 mM aqueous KCl solution, and the respective water model (Amber99SB-disp water, or TIP3P) was used. Intermediate viscosity values (blue line) were interpolated using the expression (4.16).

Table 4.8: Parameters resulting from the interpolation of solvent viscosities.

Force field	B_w [-]	D_w [K^{-1}]	ΔE_w [kJ/mol]
a99SB-disp	20.05	0.0115	24.3
CHARMM36m	14.75	0.0054	12.8

As a next step, we interpolated the viscosities of the crowded system to different temperatures using the relationship [153]:

$$\eta(T,c) = \eta_0(T)\exp\left[\frac{c}{\alpha - \beta c}\left(-B + DT + \frac{\Delta E}{RT}\right)\right] \tag{4.17}$$

In this expression, c denotes the protein concentration while B, D, and ΔE are free parameters. The parameter α is a function of the protein molecular weight M_p, the molecular weight of water M_w, and its density ρ_w:

$$\alpha = \rho_\mathrm{w}\frac{M_\mathrm{p}}{M_\mathrm{w}} \tag{4.18}$$

We set M_p to the average molecular weight of the proteins in our model. The parameter β was defined as:

$$\beta = \alpha v - 1 \tag{4.19}$$

where we set v to a value that was determined previously for BSA, $v = 1.417$ x 10^{-3} m^3/kg [153] The parameters B, D, and ΔE obtained for each force field and for both folded and unfolded systems are listed in Table 4.9 and the resulting $\eta(T,c)$ functions for $c = 288$ g/L are displayed in Fig. 4.15A, and Fig. 4.15B. As illustrated in Figure 4.15C, the analytical function $\eta(T,c)$ defined by Eq. (4.17) allowed us to extrapolate the dependence of viscosity on concentration for a given T.

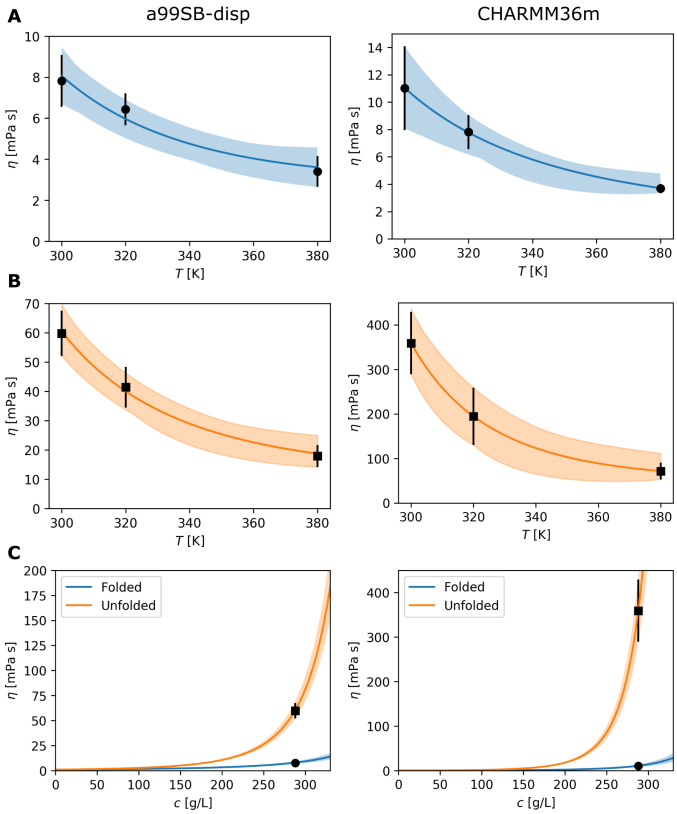

Figure 4.15: Viscosities calculated for the 288 g/L sub-volume using the Amber99SB-disp force field (left column) and the CHARMM36m force field (right column). (A) Folded system. (B) Unfolded system. The solid lines show an interpolation using Eq. (4.17) with bands expressing the uncertainty of this interpolation. (C) Extrapolation to different concentrations using Eq. (4.17).

Table 4.9: Parameters describing an analytical dependence derived from the computed viscosities for the crowded protein solution.

Folded

Force field	B [-]	D [K^{-1}]	ΔE [kJ/mol]
a99SB-disp	-20.05	32.8272	7157.1
CHARMM36m	-14.75	12.9697	44982.1

Unfolded

Force field	B [-]	D [K^{-1}]	ΔE [kJ/mol]
a99SB-disp	-20.05	35.2641	36692.3
CHARMM36m	113942.65	189.9209	251039.8

To estimate the uncertainty of the interpolated/extrapolated viscosity values, we repeated the fitting of Eq. (4.17) with random combinations of viscosities each sampled within the confidence interval at the given temperature.

4.2.11 Evaluation of the apparent diffusion coefficient

In order to directly compare simulations with experiments, we estimated the apparent diffusion coefficient D of each protein, which takes into account not only the translational motions, but also the rotations of the entire protein. To this end, we used the following relation that links D to the translational, D_t, and the rotational. D_r diffusion coefficients [70, 105]:

$$\sum_{n=0}^{N} B_n(q) \cdot \frac{D_r\, n(n+1) + (D_t - D)\, q^2}{\left[D_r\, n(n+1) + (D_t + D)\, q^2\right]^2} \xrightarrow[N\to\infty]{} 0 \tag{4.20}$$

where $B_n(q)$ is:

$$B_n(q) = (2n+1) \int_0^{\infty} \rho_H(r) \cdot j_n^2(qr)\, dr \tag{4.21}$$

and j_n is the spherical Bessel function of order n and ρ_H is the Radial Distribution Function (RDF) of the protein H atoms.

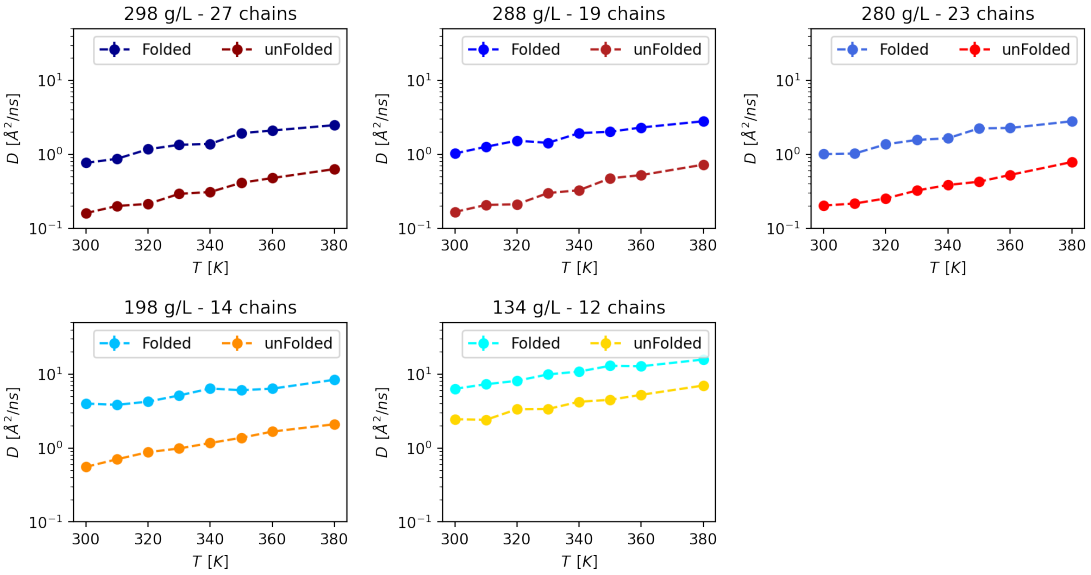

Figure 4.16: Comparison of D for the folded and the unfolded average protein at different temperatures for systems with different concentrations (force field: CHARMM36m).

After the calculation of D_t and D_r for each protein (as described in the previous subsections), we fixed $N = 550$ [105] and we solved numerically the l'eq. (4.20) for three values of q: 0.2Å^{-1}, 1Å^{-1}, and 2Å^{-1}, verifying that, in the q-range explored by the experiment ($0.2\text{Å}^{-1} \leq q \leq 2\text{Å}^{-1}$), the value of D is constant. Indeed, if N and q are big enough, it is possible to show that the result do not depend on them [105].

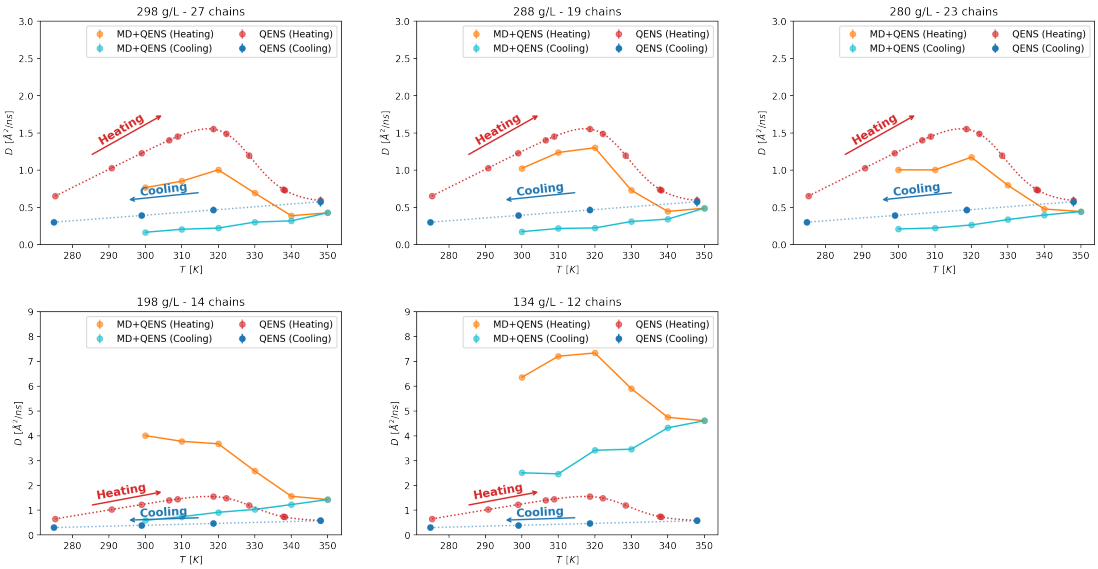

Figure 4.17: D of the average protein at different temperatures for systems with different concentrations (force field: CHARMM36m). The slowdown is reproduced using the smeared step function a_u^{QENS} to weight the prediction obtained from the simulation of the fully folded and fully unfolded systems.

In this way, we obtained the D of each protein in the folded and the unfolded configuration at all the temperatures. Then, aiming to compare these data with the QENS results, for each system we calculated the average value of D weighting each protein with the number of its H atoms. Figure 4.16 shows the results for the 5 selected sub-volumes extracted from the larger cubic box used for the LBMD simulations.

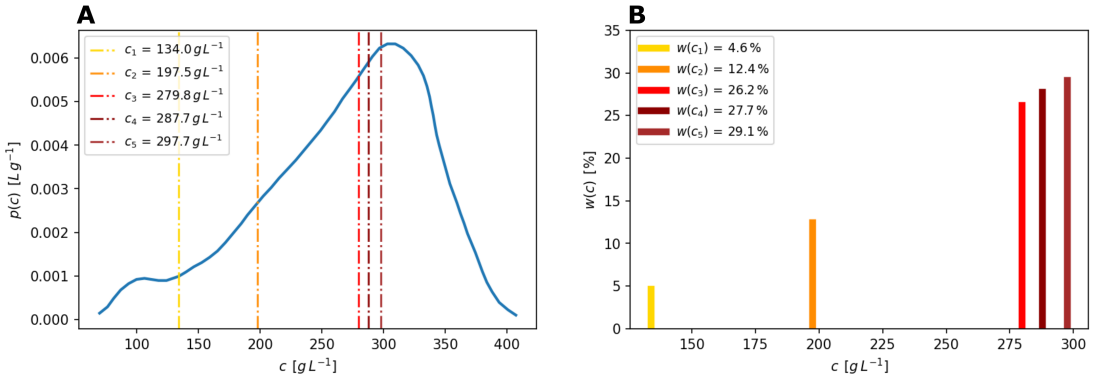

Figure 4.18: A: Probability distribution for the selection of a sub-volume of (17 nm)^3 with a concentration c from the larger system of (40 nm)^3 used for the course-grained representation of the *E. coli* cytoplasm. B: Weight of each sub-volume for the estimation of the D.

To combine the information related to the systems with fully folded proteins with the ones with fully unfolded proteins, we used the following relation:

$$D(T) = D^{(f)} \cdot [1 - a_u^{\mathrm{QENS}}(T)] + D^{(u)} \cdot a_u^{\mathrm{QENS}} \tag{4.22}$$

for further details on the validity of this approach see the section 4.4.2. The resulting values of D, for CHARMM36m, are reported in Fig. 4.17.

Finally, in order to calculate the most representative value of D_{app} for the average protein in the *E. coli* cytoplasm, we considered the probability distribution $p(c)$ for the selection of a sub-volume of (17 nm)^3 and concentration c from the larger system of (40 nm)^3 used in the LBMD simulation as representative of the bacterial cytoplasm - see Fig. 4.18A. The weights of each sub-volume (Fig. 4.18B) was calculated as follow:

$$w(c_i) = \frac{p(c_i)}{\sum_{k=1}^{5} p(c_k)} \tag{4.23}$$

The result, obtained with both the two force field, is shown in Figure 4.28.

4.3 Results

Characterization of the dynamical state of the *E. coli* cytoplasm near cell death. Quasi-elastic Neutron Scattering (QENS) experiments were performed on the backscattering spectrometer IN16b [102] at the Institut Laue Langevin (ILL), France, on *in vivo* *E. coli* samples. The scattering signal of these samples is mainly due to the large incoherent scattering cross section of the hydrogen atoms whose dynamics lying in the nanosecond time- and the nanometer length-scales is accessible to the spectrometer (See Methods in SM). In particular, the predominant contribution is due to the self-diffusive dynamics of an average protein in the bacterial cytoplasm, as more than 75% of the *E. coli* dry weight consists of proteins and ribosomes, themselves made up of about 50% proteins and 50% RNA by mass [104, 103, 110, 154, 111]. In addition, the number of hydrogen atoms in proteins is larger than in nucleic acids, while the remainder biomolecular components, mainly phospholipids, contribute to just about 10 % of the dry weight of the bacterium. QENS allows to directly measure the incoherent dynamic structure factor $S(Q,E)$, which can be in turn modeled to obtain unique information on the microscopic spatial and time correlation functions [52]. The dynamical state of the proteome was sampled at increasing temperatures starting from $276\,K$, where the bacteria can live and thrive, till $350\,K$, i.e. well above the temperature of cell-death ($T_{CD} \approx 323\,K$) [45]. In addition, to test the reversibility of dynamical changes, we performed a few measurements while cooling the bacteria after they underwent thermal death.

Protein dynamics at all the temperatures are well described in terms of two distinct diffusive processes arising from the dynamics of the average protein in the *E. coli* cytoplasm (See Fig. 4.19A), whose narrow and broad Half Width–Half Maximum (HWHM) of the signal γ_G and γ_L correspond, respectively, to the long and the fast characteristic timescales of the probed global (G) and local (L) motions. The wavevector dependence of both γ_G and γ_L follows the Singwi and Sjölander jump-diffusion model [155], as shown in Fig. 4.19B and 4.19C, where the scatterers alternate oscillatory motions around their equilibrium positions for a certain residence time τ and diffusive motions between two equilibrium positions (for further details see section 4.2.2). This is consistent with results of previous *in vivo* QENS experiments on bacteria [103, 110, 111]. We interpret the γ_G component as due to global diffusive dynamics of an average protein while caged in the crowded *E. coli* cytoplasm. The faster γ_L contribution arises from the local internal and inter-domain dynamics occurring within proteins [70].

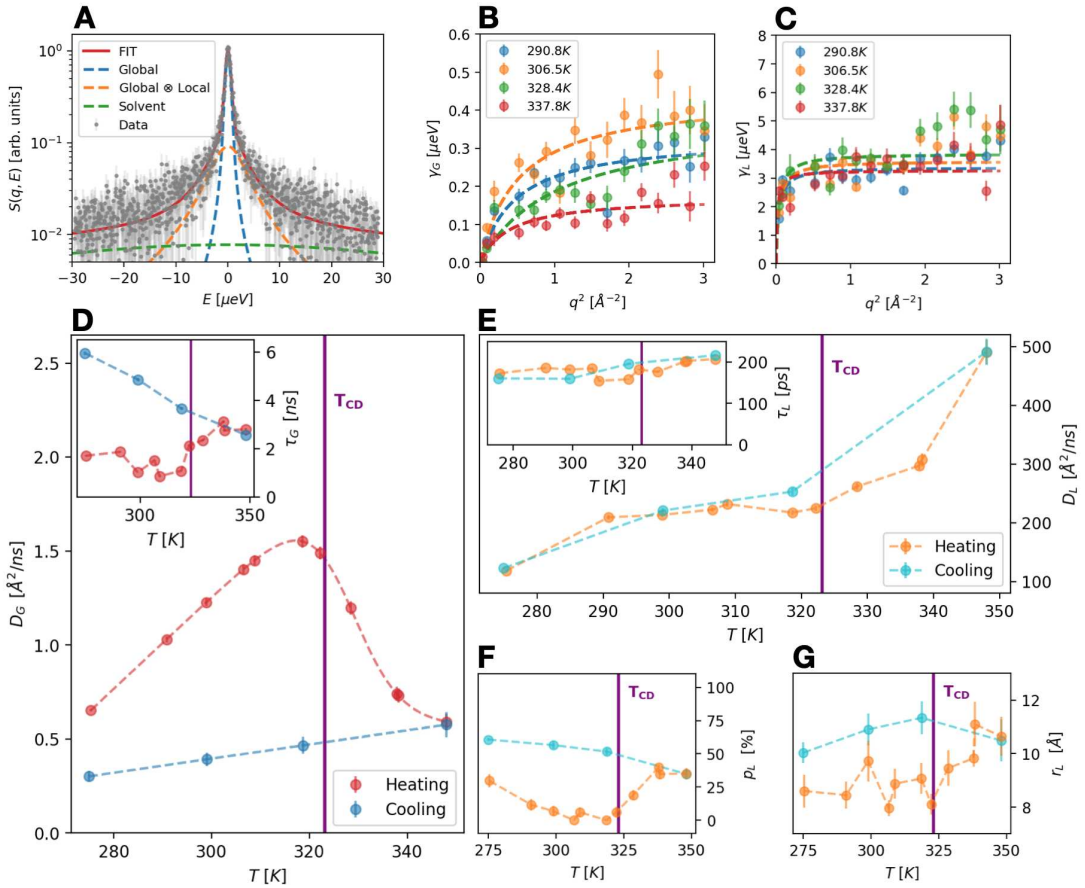

Figure 4.19: (A) Example of a QENS spectrum measured on the neutron backscattering spectrometer IN16b at ILL [102] on *E. coli* cell pellets suspended in D_2O. The displayed data were recorded at 306 K and $q = 0.72 Å^{-1}$. We present the resulting fit of the data assuming three main diffusive contributions accounting for the global and internal motions of proteins, and for the D_2O dynamics (see section 4.2.2 for further details on the model and the fit procedure). (B-C) HWHM of the Lorentzian components accounting for the global (γ_G) and the local diffusive protein motions (γ_L) at four different temperatures. The dashed lines are the fits of the linewidths assuming the Singwi and Sjölander jump-diffusion model [155]. (D) Apparent self-diffusion coefficient, D_G, of an average protein in the *E. coli* cytoplasm as a function of temperature. A clear non reversible slow-down of D_G is visible after the temperature of cell death (T_{CD}) due to the gelation of the cytoplasm. The inset shows the residence time τ_G for the global motions as function of temperature that undergoes an important upturn after T_{CD} and continues to increase during the cooling. (E) Temperature dependence of the diffusion coefficient D_L for the local motions of the side-chains. The transition at T_{CD}, where D_L starts to increase more steeply with the temperature, is reversible. The inset shows the residence time τ_L for the internal dynamics which is nearly constant ($\approx 180 ps$). (F-G) Geometries of the local motions as derived by the elastic incoherent structure factor (EISF) [52, 55]: p_L is the fraction of atoms appearing fixed on the accessible time scale (F), and r_L is the radius of the confinement region for this type of fluctuations (G).

The apparent diffusion coefficient D_G, which combines both the translational and rotational motions of proteins [69] (see section 4.2.2) exhibits a dramatic non-reversible reduction in proximity of the cell-death temperature (Fig. 4.19D). In protein crowded solutions a similar slow-down of the diffusive dynamics was ascribed to the gelation of the system induced by the protein unfolding [66, 109, 156]. The transition to a gel-like

phase is also supported by the slight increase of τ_G above T_{CD} (inset of Fig. 1D), which suggests an increasing localization of proteins in cage-like structures.

Further, the local dynamics is very sensitive to temperature change, but in this case the diffusion coefficient D_L shows a significant increase above T_{CD}, while τ_L is nearly constant at $\approx 180ps$ (Fig. 1E and its inset). This suggests that the changes in the local dynamics are due to a variation in the extent of the explored spatial region. This is confirmed by the increase of the distance between two jumps r_L, that displays a sudden break toward higher values after T_{CD} (Fig. 4.19G). Interestingly, also the number of atoms too slow to be visible by QENS, p_L, shows a similar increasing trend above T_{CD}, as seen in Fig. 4.19F, when temperature is rising. However, comparing the fraction of H-atoms that are fixed with those that diffuse in a sphere or undergo a 3-jump diffusion (see Fig. 4.20), we find that the percentage of H-atoms that are diffusing in a spherical volume is constant in temperature. Therefore, the changes in the fraction of fixed atoms are due to the variation of the fraction of H-atoms undergoing the 3-jump diffusion, only. This behavior is consistent with the progressive unfolding of a part of the *E. coli* proteome. In fact, moving subgroups of denatured proteins can access larger spatial regions, and at the same time their number gradually decreases because of the growing interactions of the gel-like system formed in the cytoplasm.

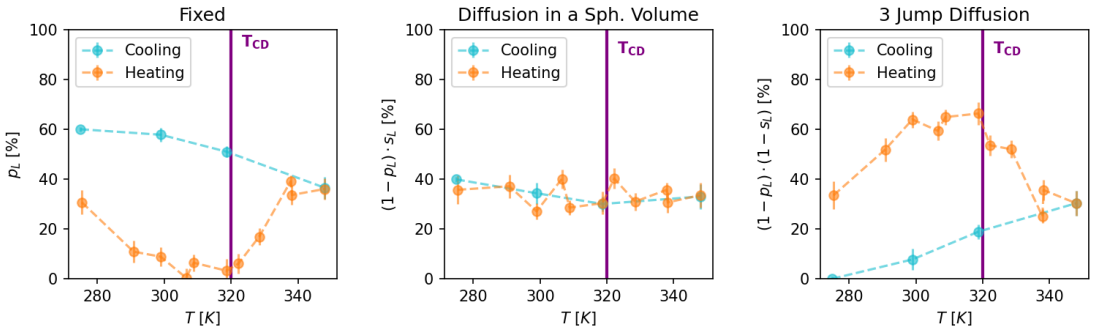

Figure 4.20: Comparison of the percentage of H-atoms that appear fixed with those of H-atoms that diffuse in a spherical volume or undergo a 3-jump diffusion.

The rate at which protein groups locally explore their environment seems to be reversible across the thermal death, as we can see from the trend of both D_L and τ_L in Fig. 4.19E, confirming previous results on crowded protein solutions [109]. On the other hand, the characteristic lengths of the local dynamics, i.e. r_L and p_L, show an irreversible trend.

Simulations show how protein unfolding slows down protein diffusion. To examine how protein motions are affected by heating and thermal denaturation of parts of the *E. coli* interior, we performed multi-scale molecular dynamics simulations. Previous works studied protein diffusion in crowded environments using coarse-grained (CG) [157, 114] and all-atom [158, 152] MD simulations. However, while simplified CG models help sample long time-scales, they either lack solvent mediated correlations [114] – essential for describing the protein mobility – or rely on oversimplified description of the protein lacking chemical features [157]. Here, we combined two levels of description: a CG model for protein (OPEP) owning residue level chemical resolution [85] to sample the local structure of the crowded solution and an all-atom

description to subsequently explore the diffusion of proteins in sub-volumes obtained from the coarse-grained simulation. In our approach we used the lattice Boltzmann molecular dynamics (LBMD) technique that allows to include naturally hydrodynamic interactions in the simulations of implicit solvent molecular models, and already successfully applied to complex biological situations like protein crowded solutions [159] and multi-scale protein aggregation [160].

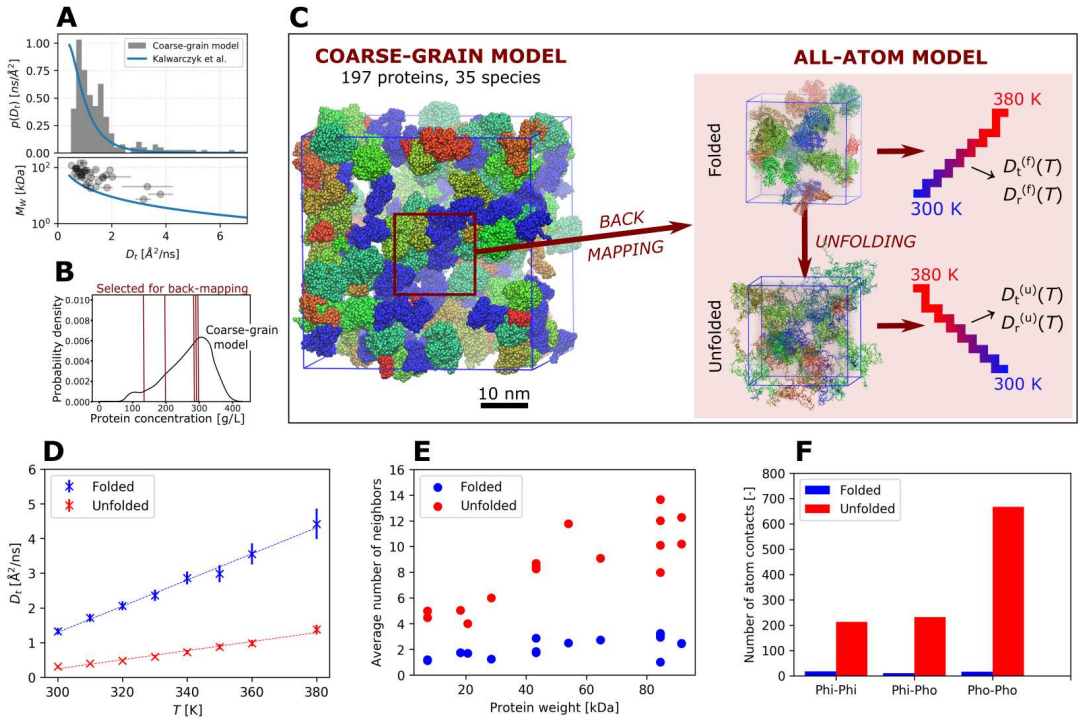

Figure 4.21: (A) Protein diffusion in the CG model cytoplasm simulated by LBMD. Distribution of the traslational diffusion coefficients, and its dependence as a function of the proteins molecular weight. The solid lines represent the fits of the experimental data reported in Ref. [161]. (B) Distribution of local protein concentrations in 17 nm cubic sub-boxes randomly placed in the large simulation system. The vertical lines correspond to the local concentrations of the sub-boxes then back-mapped at the atomistic resolution. (C) Pictorial representation of the CG cytoplasm system (cubic box of side 40 nm) and the schematic strategy of the back-map, and of the temperature scans for the folded and unfolded versions of the atomistic systems. (D) Average protein translational diffusion coefficients, computed in the atomistic systems for the 0.3–5 ns regime and corrected for the effects of periodic boundary conditions, as described in section 4.2.9. The plot shows results obtained with the Amber99SB-disp force field; the results for CHARMM36m are summarized in figure 4.25. (E) Average number of neighboring proteins per protein molecule as a function of its molecular weight. (F) Average number of different classes of protein–protein contacts per protein molecule: hydrophilic–hydrophilic (Phi-Phi), hydrophilic–hydrophobic (Phi-Pho), and hydrophobic–hydrophobic (Pho-Pho). An atom was considered "*hydrophobic*" if its partial charge was less than $0.2 \cdot e$ in magnitude. The plots in (E-F) show results obtained from two Amber99SB-disp trajectories, each extended to $1\,\mu$s, for the back-mapped all-atom systems with a protein concentration of $288\,$g/L and simulated at $T = 330\,$K before and after unfolding. The results for CHARMM36m are reported in figure 4.25.

The $4.3\,\mu$s coarse-grained LBMD [129] simulation contained 197 proteins of 35

different species, mimicking the protein composition of the *E. coli* cytoplasm [114] (see Fig. 4.21C for a pictorial representation of the system, and section 4.2.4 for more details on the protein composition). The length of the trajectory allowed significant reshuffling of the initial positions of the proteins and thus also exploration of different geometries of the crowded system; in fact, only 35% of the initial interaction partners were also in contact at the end of the trajectory (Fig. 4.22). The reshuffling is kinetically meaningful since the computed diffusivity for each molecular species is very close to its experimental estimate. In Fig. 4.21A we report the distribution of the translational diffusion coefficients computed for each protein. The coefficients vary between 0.5 and $7\,\text{Å}^2/\text{ns}$ depending on the molecular weight (Table 4.4), values in excellent agreement with what reported experimentally [161], see blue solid lines in the graphs.

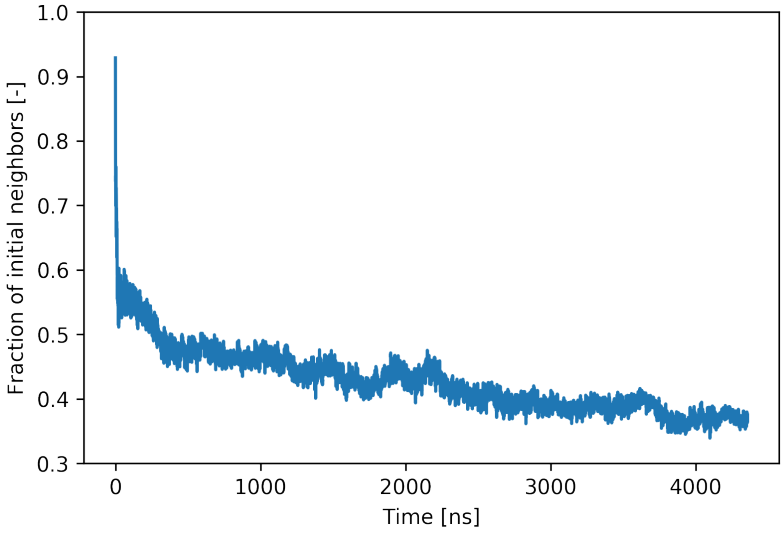

Figure 4.22: Proportion of initial protein neighbors of an average protein that remain in contact with the protein in the course of the LBMD trajectory. Two proteins were considered in contact if the shortest distance between their beads was below 7.34 Å.

From different frames of this large-scale trajectory, we selected five sub-volumes of the whole system containing 11–20 proteins so to reflect the protein composition and concentration heterogeneity of the cytoplasm (Figure 4.21B). Each sub-box was converted to the all-atom resolution (Figure 4.21C) and exposed to a sequence of production simulations ($\sim 100\,\text{ns}$ per run) at increasing temperatures to investigate protein translational diffusion coefficients, probed in the 0.3–5 ns regime. Subsequently, for these atomistic systems we repeated this heating protocol with the same box after exposing it to a simulated rapid heat shock (see section 4.2.7), serving to completely unfold all the proteins within the time scale accessible to the simulation.

Our simulations showed a strong decrease in the average translational diffusion coefficient upon unfolding (Figure 4.21D). The observed decrease quantitatively agreed for two distinct force fields, Amber99SB-disp [125] and CHARMM36m [126], which we used to model the proteins (see Figure 4.25). In addition, for both folded and unfolded proteins the diffusion coefficient scales linearly with temperature. A similar slow-down upon unfolding has been observed in previous theoretical and experimental studies of monocomponent protein solutions [66, 109, 156].

Unfolding changes protein interactions. A detailed look at the protein–protein interactions reveals the cause of the diffusion slow-down. We consider as an example one of the simulated atomistic systems of protein concentration 288 g/L. While each protein had, on average, 2.1 interaction partners in the folded system, this number increased to 8.6 when the proteins are all unfolded (see Figure 4.21E). As a consequence, the average number of atom–atom contacts per protein rose by a factor of 25 upon unfolding. Among the different contact types, this increase was the strongest for "*hydrophobic*" contacts (i.e., between non-polar atoms), which were enhanced by a factor of 41 (see Figure 4.21F). Recently, the diffusion slow-down in crowded solutions of globular proteins at intermediate concentrations ($\leq$ 200 g/L) has been linked to the formation of transient protein clusters [149, 152]. Indeed, we observed that in folded systems with protein concentrations below 200 g/L, the proteins were organized in 1–3 clusters, with a few remaining protein molecules floating freely in solution. At higher concentrations, the folded proteins formed a single large cluster most of the time, while maintaining a degree of dynamical exchange with the bulk (see Figure 4.23). On the contrary, regardless of the protein concentration, unfolding led to the formation of a single cluster, encompassing all the proteins and creating a stable network.

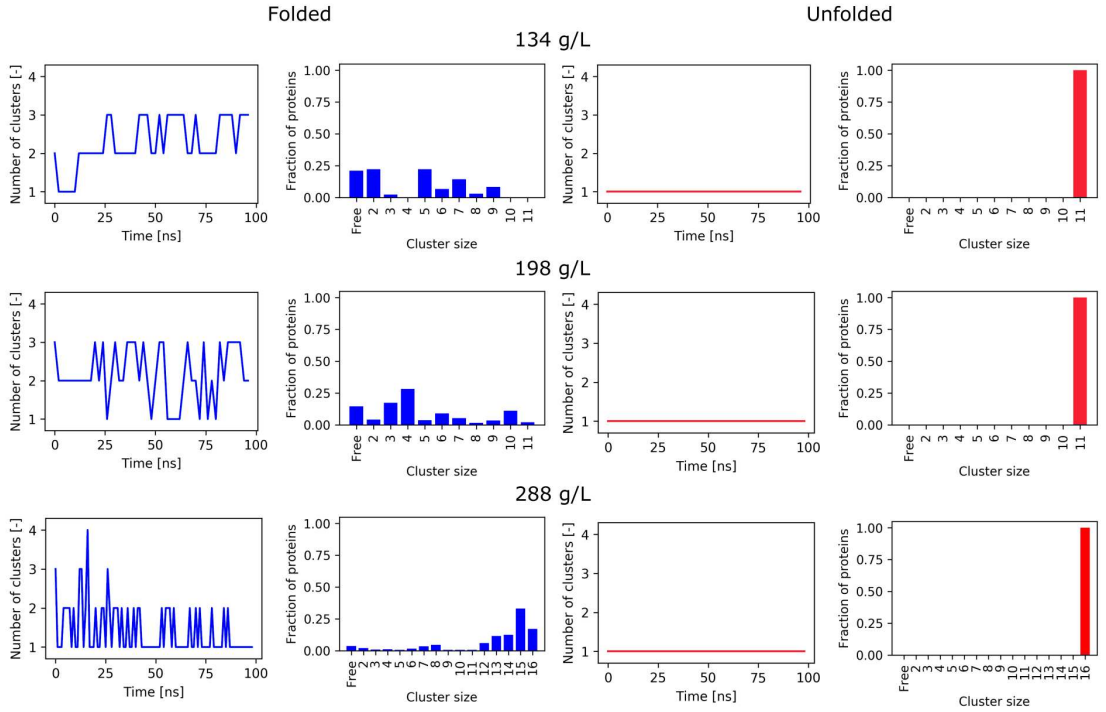

Figure 4.23: Number of protein clusters present in the course of a sub-volume simulation and the resulting partitioning of proteins into clusters of different sizes for three different protein concentrations in folded (left) and unfolded (right) sub-volumes. The trajectories were obtained with the Amber99SB-disp force field at $T = 300$ K. Two proteins were considered in contact if the minimum distance between their atom was lower than 3 Å.

The enhanced stickiness of the unfolded proteins was reflected by a strong rise in the viscosity of the crowded protein solutions. For a 288 g/L protein system at $T = 300$ K, the viscosity increased upon unfolding from $\eta = 8 \pm 1$ mPa·s to 60 ± 8 mPa·s with a99SB-disp and from $\eta = 11 \pm 3$ mPa·s to as much as 360 ± 70 mPa·s

with CHARMM36m (Fig. 4.15). The substantial relative increase in viscosity caused by unfolding persisted even at the highest temperatures, (with η going up from 3.4 $\pm$ 0.8 to 18 $\pm$ 4 mPa·s for Amber99SB-disp and from 3.7 $\pm$ 0.3 to 72 $\pm$ 19 mPa·s for CHARMM36m at $T = 380$ K). Our results are consistent with the trend observed in previous experimental studies [162, 163, 164], reporting an increase in the viscosity of protein solutions undergoing thermal denaturation. Owing to the significantly lower protein concentrations (below 100 g/L) that were examined in those measurements, the unfolded protein viscosities ($\sim$1–40 mPa·s) are lower than our computational estimates. On the other hand, the viscosities of our unfolded systems are smaller than those reported for concentrated antibody solutions [165], reaching up to 1700 mPa·s, and several phase-separated biomolecular condensates [166]. In fact, given the very slow decay of the pressure autocorrelation function (Fig. 4.13), the viscosity values obtained for the unfolded systems likely represent a lower bound for the actual viscosity.

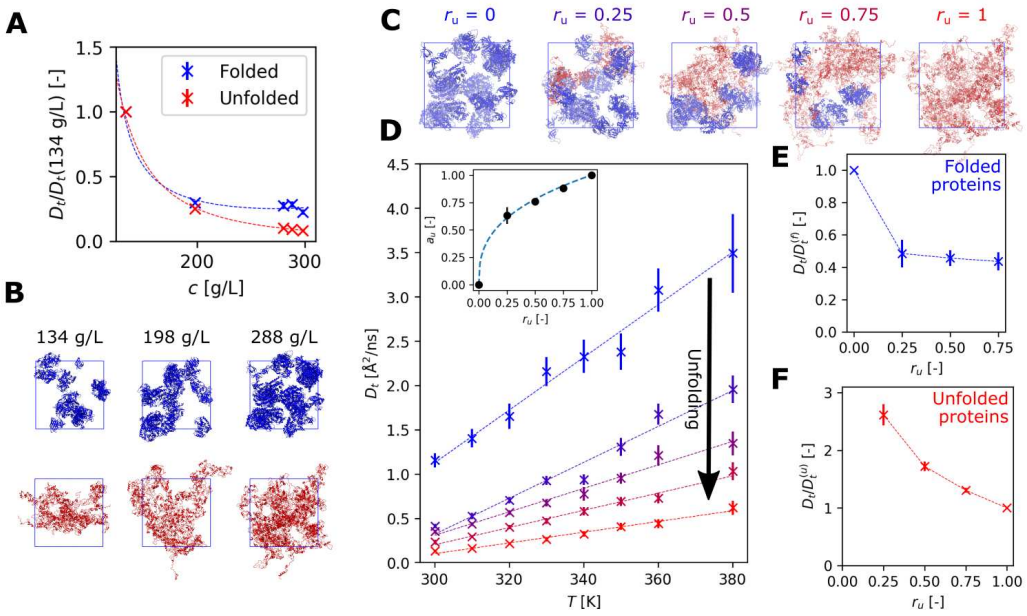

Figure 4.24: (A) Concentration-dependent diffusion slow-down from simulations. The slow-down is expressed relative to the diffusion coefficient for the least concentrated sub-box (134 g/L). The dashed lines represent fits with a second-order polynomial fraction. (B) Snapshots showing folded- (upper row) and unfolded (lower row) simulation boxes at different protein concentrations. (C) Snapshots of sub-boxes (288 g/L) with a progressively increasing unfolded fraction r_u. Folded proteins are blue while unfolded proteins are shown in red. (D) Translational diffusion coefficients in sub-boxes (288 g/L; see panel C) with a varying fraction of unfolded proteins. The insets show the dependence—fitted to a power law—of the apparent unfolded fraction a_u on r_u. (E-F) The decrease in the translational diffusion coefficient of folded and unfolded proteins inside the partially unfolded sub-box (288 g/L) with increasing r_u. The values shown in panels A, E, and F as well as in the inset of panel D are averages across all temperatures, with error bars expressing the standard deviations. The results presented in this figure were obtained with the a99SB-disp force field; analogous plots for CHARMM36m, exhibiting qualitatively the same behavior, can be found in Figure 4.25.

Simulations show concentration dependence of diffusion slow-down. In line with previous results for crowded protein solutions [70, 112] and cell lysates [167, 168], the

translational diffusion coefficients extracted from the atomistic simulations decreased with increasing protein concentration (see Figure 4.24A), and the effect is stronger for the unfolded systems. In both folded and unfolded cases, the sharpest drop occurred below 200 g/L, whereas the concentration sensitivity was predicted to be weaker in the concentration range of 200–400 g/L, characteristic of the *E. coli* cytoplasm [50]. This indicates that variations in the macromolecular concentrations inside the *E. coli* proteome should not affected strongly the comparison of simulations with experiments.

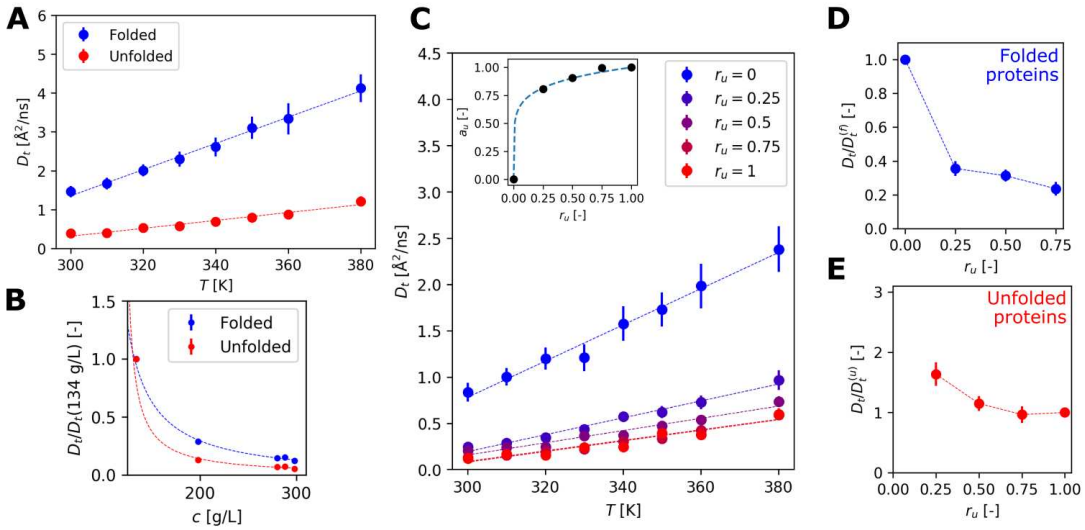

Figure 4.25: Translational diffusion in sub-volumes described using the CHARMM36m force field. (A) Protein translational diffusion coefficients, computed for the 0.3–5 ns regime. (B) Concentration-dependent diffusion slow-down. The slow-down is expressed relative to the diffusion coefficient for the least concentrated sub-box (134 g/L). Averages of the slow-down values for all the temperatures are shown together with their standard deviations. The dashed lines represent fits with a second-order polynomial fraction. (C) Translational diffusion coefficients in sub-boxes (288 g/L) with a varying fraction of unfolded proteins r_u; The inset shows the dependence – fitted with a power law – of the apparent unfolded fraction a_u on r_u. (D-E) Decrease in the translational diffusion coefficient of folded- and unfolded proteins inside the partially unfolded sub-boxes with increasing r_u.

Even partial proteome unfolding causes strong diffusion slow-down. The simulations clearly prove that massive unfolding leads to a strong diffusion slow-down. However, only a fraction of the proteome might be unfolded near the cell-death temperature as suggested by recent experiments [46, 48]. To explore the effect of such partial unfolding on the overall protein diffusivity, we performed additional temperature scans for a selected atomistic system (288 g/L) with a varying fraction r_u of unfolded proteins (25%, 50%, and 75%; see Figure 4.24C). In each case we completely unfolded the chosen amount of proteins while leaving the remainder fully folded. Since the Dill model predicts similar denaturation temperatures ($\approx 328.5K \pm 1.5K$) for all the proteins contained in the system, we selected the proteins to be unfolded randomly, trying to maximize the species' heterogeneity of the chosen sub-set. Our simulations revealed that with the increasing unfolded content, the overall translational diffusion coefficients quickly approached those calculated for the fully unfolded system (see Figure 4.24D

where we report data for Amber99SB-disp force field), this effect is even stronger when using the CHARMM36m force field (Figure 4.25).

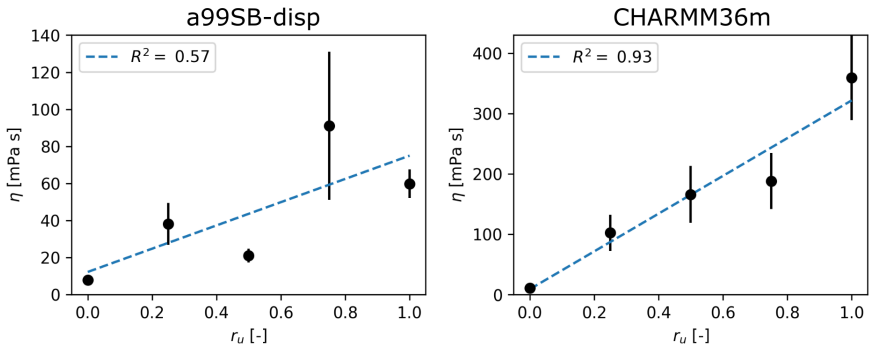

Figure 4.26: Viscosities calculated for the 288 g/L sub-volume at $T = 300\ K$ with a varying fraction of unfolded proteins (see Table 4.7).

This finding demonstrates that the translational diffusion coefficient, experiencing a sharp drop already for small values of r_u, is a non-linear function of r_u. To explain the reason why D_t shows such a rapid drop, we analysed separately the diffusion coefficients of folded and unfolded proteins in the intermediate boxes. We found that the D_t of folded proteins decreased by more than 50% already for the smallest fraction of unfolded proteins (see Figure 4.24E). Thus, the presence of even a small amount of unfolded proteins is able to strongly affect the diffusion of the remaining folded proteins. On the other hand, the diffusion coefficient of the unfolded proteins shows a more gradual decrease (Figure 4.24F). On a quantitative point of view, the translational diffusion coefficient in the partially unfolded proteome can be expressed as $D_t = (1-a_u)\,D_t^{(f)} + a_u D_t^{(u)}$, where a_u is an "apparent" unfolded fraction, weighting the diffusion coefficient of the fully folded $D_t^{(f)}$ and the fully unfolded $D_t^{(u)}$ systems. The non-linear dependence of a_u of the actual proteome unfolded fraction r_u describes the diffusion slow-down due to the unfolding. We found that this dependence can be fitted with the power law $a_u = r_u^p$ (see the insets in Figures 4.24D and 4.25C), with the exponent p equal to 0.411 ± 0.026 for a99SB-disp and 0.142 ± 0.012 for CHARMM36m.

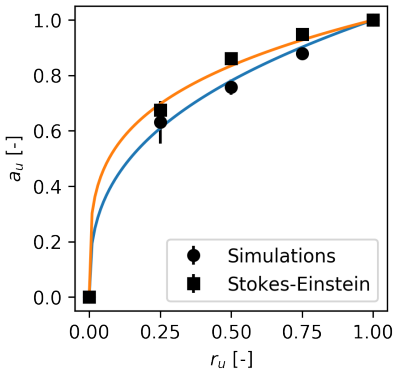

Figure 4.27: Comparison of the transfer function obtained directly from simulations using the Amber99SB-disp force field and a transfer function based on the considerations described in Supplementary Results. Each set of points is fitted with a power law dependence, with the exponents equaling 0.41 and 0.26, respectively.

Unlike the translational diffusion coefficient, the dependence of viscosity on the unfolded content appeared to be nearly linear, within the limits posed by a slow convergence (Figure 4.26). As we describe in section 4.4, a linear unfolding dependence of viscosity combined with the simplified Stokes–Einstein model results in a non-linear dependence of a_u on r_u that is similar to the one recovered directly from simulations (Figure 4.27).

Estimation of the folding state of the cytoplasm and validation of the results. By combining the QENS and MD simulations results, we estimated the temperature dependence of the fraction of unfolded proteins in the *E. coli* cytoplasm.

For this purpose, and inspired by previous work on protein crowded solutions [109], we first consider the temperature dependent profile of the apparent diffusion coefficient measured from QENS experiments (Fig. 4.19D), and we perform a fit using an empirical relation that combines the diffusivity of proteins in the folded and unfolded states, $D(T) = D^{(f)} \cdot [1 - a_u^{QENS}] + D^{(u)} \cdot a_u^{QENS}(T)$, where the function a_u^{QENS}, that measures the changes in the global diffusion, is equivalent to the apparent fraction of unfolded proteins in the system as defined in the previous paragraph.

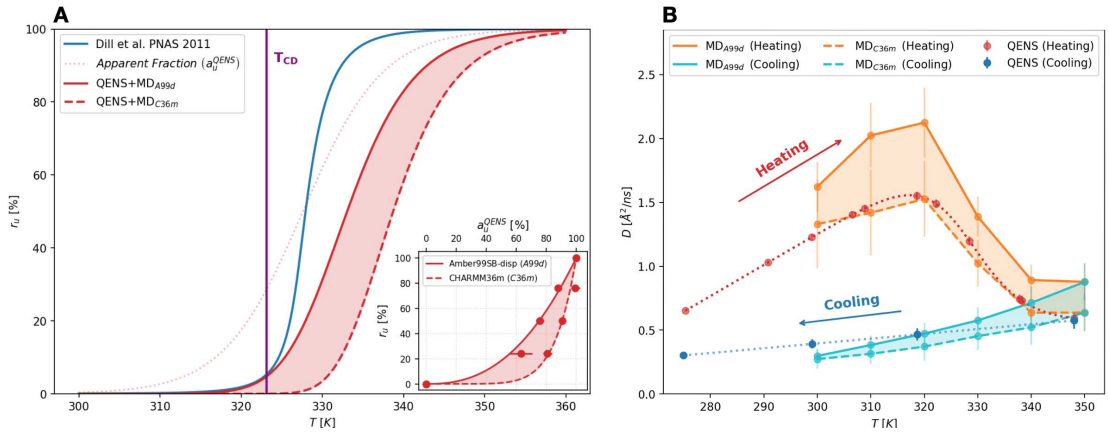

Figure 4.28: (A) Fraction r_u of unfolded proteins in the *E. coli* cytoplasm as a function of temperature. The cell-death temperature of 323.15 K [45], determined from the *E. coli* growth rate, is indicated by a vertical line. The red solid (Amber99SB-disp) and dashed lines (CHARMM36m) show $r_u^{QENS}(T)$ calculated by combining $a_u^{QENS}(T)$, measured by QENS and describing the global diffusion slow-down, with the relationship derived from simulations with the exponent p being force-field dependent. The results are compared with theoretical prediction derived by Dill [45] (blue solid line). (B) Comparison of the apparent diffusion coefficient from QENS experiments (circles) with the apparent diffusion coefficient computed for the two force field from simulations and by combining the contribution from translational and rotational motions, see SM for details.

In the second step we exploited the empirical relationship derived from simulations between the real fraction of unfolded proteins in the solution, r_u, with its apparent counterpart, a_u. We applied at each temperature the empirical relation to the experimentally derived apparent fraction of unfolded proteins[1], $r_u^{\mathrm{QENS}}(T) = \left[a_u^{\mathrm{QENS}}(T)\right]^{1/p}$. With this

[1]Further details on the connection between the apparent and real unfolded fraction are reported in section 4.4.2.

tool in hands, we obtained the experimentally derived fraction of unfolded proteins as a function of the temperature, $r_u^{\text{QENS}}(T)$, see Fig. 4.28A. This quantity is force field dependent via the exponent parameter p, and can be now compared with the equivalent function theoretically derived by Dill and coworkers [45], see Fig. 4.28A.

Strikingly, at variance with the proteome catastrophe scenario derived theoretically [45], the experimentally derived r_u^{QENS}s show a very slow increase of the fraction of unfolded proteins as temperature crosses $T_{CD} = 323.15K$. For sake of clarity, a few degrees above the cell death temperature, the Dill's model predicts more than 50% of the proteins unfolded, while the experimentally derived unfolded fraction is, on average, less than 15% (see Table 4.10). This result is in line with recent experiments on proteome thermal stability [46, 48].

Table 4.10: Different estimations for the fraction of unfolded proteins $r_u(T)$ in the *E. coli* cytoplasm at increasing temperatures and temperature $T_{50\%}$ needed to unfold 50% of the proteins.

Models	$r_u(T_{CD})$	$r_u(T_{CD}+3\text{K})$	$r_u(T_{CD}+5\text{K})$	$T_{50\%}$
Dill et al. PNAS 2011	5.39 %	21.47 %	53.54 %	327.9 *K*
Apparent Fraction	28.75 %	42.65 %	52.72 %	327.6 *K*
QENS+MD$_{A99d}$	4.82 %	12.59 %	21.07 %	333.0 *K*
QENS+MD$_{C36m}$	0.02 %	0.25 %	1.07 %	338.7 *K*
QENS+MD$_{\text{AVG}}$	**0.71 %**	**3.40 %**	**7.86 %**	**335.7 K**

It is worth to recall that our approach is based on two assumptions (see section 4.2). First, we hypothesized that the main contribution to the QENS signal comes from an average protein. This is a quite reasonable approximation, as ribosomes, DNAs and RNAs are too massive and consequently slow to be detected, while the remainder biomolecular components, mainly phospholipids, contribute to just about the 10% of the dry weight of the bacterium [104]. In addition, in our simulations we represented a very simplified version of the *E. coli* cytoplasm, composed of just a small subset of proteins. This simplified representation, however, is able to catch the main dynamic features of the system we investigated, thus strengthening the picture we propose. (see the Discussion for more details).

An additional support to our findings comes from the fact that the experimental apparent diffusion coefficient D_G of the average protein can be correctly described from the simulations just starting from the estimates of the apparent fraction of unfolded proteins and of $D_G^{(f,u)}$. The curves are reported in Fig. 4.28B, showing a very good agreement with the experimental data, thus endorsing the assumptions we made.

Reproducing the growth rate of *E. coli*. The stability curve of the proteome is used now to reconstruct the growth-rate curve of *E. coli*. We follow in spirit the approach described by Dill in [45] where the growth-rate ($g(T)$) is related to the temperature dependent fraction of unfolded protein via an Arrhenius reaction rate term, $g(T) = g_0 e^{-\Delta H^\dagger} \prod_{i=1}^{\Gamma} r_i(T)$; where g_0 is an intrinsic growth-rate parameter, $\Delta H^\dagger$ is the dominant activation barrier, Γ is the number of essential proteins for the bacterium growth, and r_i is the temperature dependent fraction of unfolding for the protein species i. In our approach, r_i is replaced by the average estimated r_u^{QENS}. When using r_u^{QENS} from a99SB-disp we obtained an excellent fit, and values for $\Delta H^\dagger \simeq 45$ kJ/mol and

$\Gamma = 86$. For sake of comparison, we recall that Dill obtained the following values $\Delta H^{\dagger} \simeq 27$ kJ/mol and $\Gamma = 51$ but starting with a very different form of the stability curve.

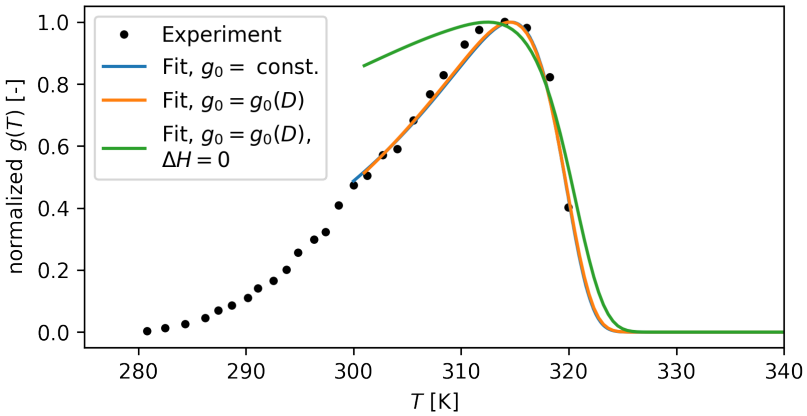

Figure 4.29: Growth-rate of *E. coli* bacteria as a function of temperature. Fit obtained using a kinetic model for the growth-rate based on the temperature dependent fraction of unfolded protein in the cytoplasm deduced by QENS/MD, r_u^{QENS}, and the temperature dependent diffusion constant measured by QENS. The fit is performed using a constant intrinsic growth-rate parameter g_0 (blue line), a temperature dependent $g_0(T) \propto D^{QENS}$ (orange), and assuming a negligible activation barrier $\Delta H^{\dagger}$ (green).

We then introduce into the fit the temperature dependence of the exponential prefactor in terms of a reaction-diffusion model by assuming $g_0 \propto D_G^{\mathrm{QENS}}$ (see Fig. 4.19). The fit, see Fig. 4.29, is comparable to the case where g_0 is constant, but by including the dependence on diffusivity we recover smaller activation barrier and number of essential proteins ($\Delta H^{\dagger} \simeq 35$ kJ/mol and $\Gamma = 81$). A final numerical test is done by assuming the growth rate of the bacterium completely rate-limited by diffusivity ($\Delta H^{\dagger} = 0$). In this case, the *E. coli* growth rate curve cannot be fitted at temperatures below the cell-death.

The obtained results confirm that even without assuming a proteome catastrophe, the growth-rate of *E. coli* can be reproduced very well by a more smooth temperature progress of unfolding of the proteome; and that including the diffusion contribution to reactivity may help tuning the essential parameters of the model, the activation barrier and the estimate of essential proteins for the bacterium growth.

4.4 Appendix

4.4.1 EINS Results

The temperature dependence of the total MSD measured with EINS is similar to the one observed for the apparent diffusion coefficient D_G measured by QENS and the parameters T_0 and ΔT resulting from the fit with eq. (4.7) are close to the ones obtained from D_G (see Figure 4.7). This result becomes evident from Figure 4.30 where we compare a_u^{EINS} and a_u^{QENS}.

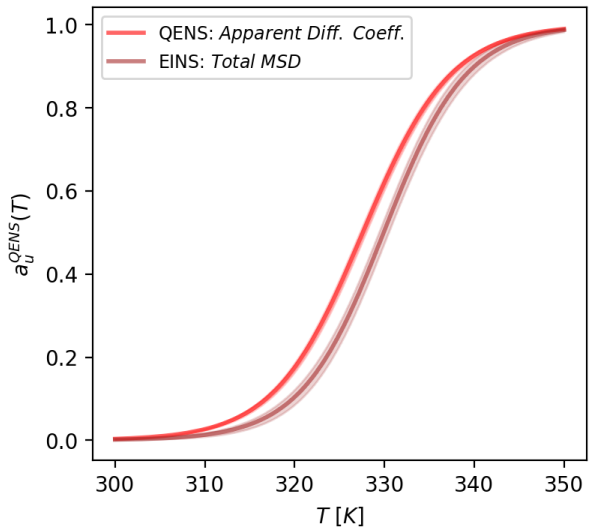

Figure 4.30: Comparison of the smeared step functions $a_u^{\mathrm{QENS}}(T)$, and $a_u^{\mathrm{EINS}}(T)$, obtained respectively from QENS and EINS.

The total MSD represent an average over various motions in the time limit going to infinity. However, as described in sections 2.2.3 and 2.2.2, typical biosystems contain several kind of motions and following [52, 66], we can subdivide the movements into a sum of three contributions: local vibrations, diffusive motions of molecular sub-units and global diffusion of the entire protein, as they correspond to distinct time scales. Consequently, we can rewrite the total MSD, $\langle u^2 \rangle$, as the sum of three different MSD:

$$\langle u^2 \rangle = \langle u^2_{\mathrm{vib}} \rangle + \langle u^2_{\mathrm{sub}} \rangle + \langle u^2_{\mathrm{diff}} \rangle \tag{4.24}$$

where $\langle u^2_{\mathrm{vib}} \rangle$ is due to vibrations, $\langle u^2_{\mathrm{sub}} \rangle$ is arising from the diffusive motions of the protein sub-units, and $\langle u^2_{\mathrm{diff}} \rangle$ is due to the roto-translation of the entire protein.

In general, the first two first contributions $\langle u^2_{\mathrm{vib}} \rangle$ and $\langle u^2_{\mathrm{sub}} \rangle$ correspond to the internal motions, thus we can define $\langle u^2_{\mathrm{int}} \rangle$ as:

$$\langle u^2_{\mathrm{int}} \rangle = \langle u^2_{\mathrm{vib}} \rangle + \langle u^2_{\mathrm{sub}} \rangle = \langle u^2 \rangle - \langle u^2_{\mathrm{diff}} \rangle \tag{4.25}$$

On the other hand, it is possible to show that $\langle u^2_{\mathrm{diff}} \rangle$ can be derived from the global diffusion coefficient [66]:

$$\langle u^2_{\mathrm{diff}} \rangle = 6\, D_G\, t_{\mathrm{res}} \qquad \text{with} \qquad t_{\mathrm{res}} = \frac{4\hbar\sqrt{\pi^{-1}\ln 2}}{\Delta E_{\mathrm{FWHM}}}, \tag{4.26}$$

where t_{res} is the resolution time determined from the energy resolution ΔE_{FWHM} of the instrument [57].

We therefore combined EINS and QENS to measure $\langle u^2 \rangle$, $\langle u^2_{\mathrm{diff}} \rangle$, and $\langle u^2_{\mathrm{int}} \rangle$ at different temperatures, as shown in Figure 4.31. A clear upturn of internal MSD is visible around the temperature of cell-death. This result is similar to what was observed by Henning et al. on proteins *in vitro* [66], and it is consistent with our QENS results for the local dynamics (Fig. 4.19E). A direct comparison between the internal MSD estimated for heating and cooling is reported in Figure 4.32.

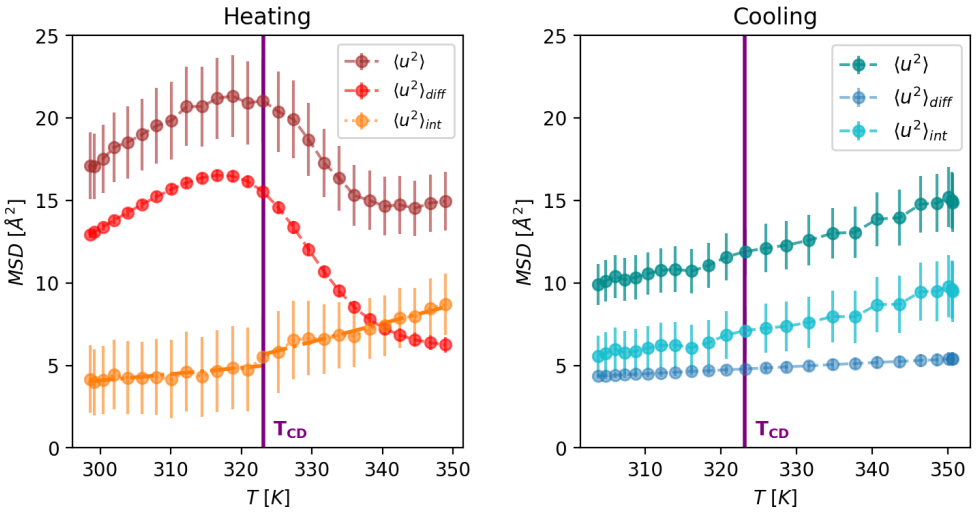

Figure 4.31: Different components of the MSD for the heating and the cooling phases. $\langle u^2 \rangle$ is the total MSD, measured by EINS, taking into account all the types of motions of the average protein in the *E. coli* cytoplasm. $\langle u^2_{\text{diff}} \rangle$ is the MSD due to the roto-tranlational diffusive motions of the entire average protein determined by QENS though the measurement of the apparent diffusion coefficient D_G (see eq. (4.26)). $\langle u^2_{\text{int}} \rangle$ is the MSD related to the internal sub-diffusive and vibrational motions of the proteins, obtained by the subtraction of $\langle u^2 \rangle$ and $\langle u^2_{\text{diff}} \rangle$.

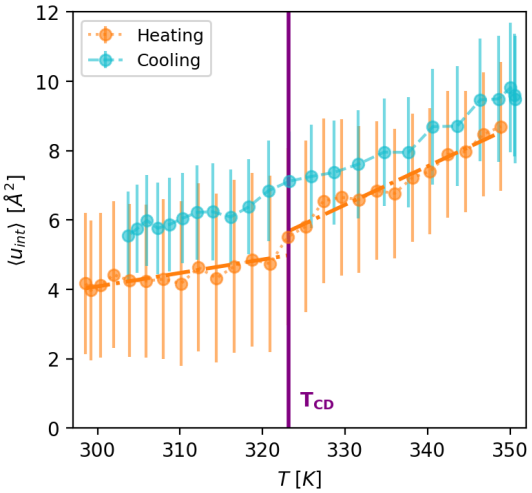

Figure 4.32: Comparison of the internal MSD obtained for the heating with the those calculated for cooling.

4.4.2 Connection between the apparent and the real unfolded fractions

As we detailed in the subsection "*Neutron scattering data analysis*", we found that D_G measured by QENS experiments can be described as:

$$D_G(T) = (T \cdot a_1 + b_1) \cdot [1 - a_u^{\text{QENS}}(T)] + (T \cdot a_2 + b_2) \cdot a_u^{\text{QENS}}(T) \qquad (4.27)$$

where $a_u^{\mathrm{QENS}}(T)$ is a smeared step function parameterized by the temperature of transition T_0 and the width of the transition ΔT:

$$a_u^{\mathrm{QENS}}(T) = \frac{1}{1+e^{-\frac{T-T_0}{\Delta T}}} \tag{4.28}$$

The function $a_u^{\mathrm{QENS}}(T)$ weights the transition between two different dynamical regimes characterized by the two extremal diffusion coefficients $D_1(T)$ and $D_2(T)$, and since $D_2(T) < D_1(T)$, $a_u^{\mathrm{QENS}}(T)$ is a measure of the diffusion slow-down. At low temperatures ($T \ll T_0$), the parameter a_u^{QENS} tends to zero; therefore, $D_G(T) \approx D_1(T) = T \cdot a_1 + b_1$. At high temperatures ($T \gg T_0$), the smeared step function is close to one, and $D_G(T) \approx D_2(T) = T \cdot a_2 + b_2$. However, the observed dynamical transitions is not reversible—after the transition, $D_G(T)$ remains equal to $D_2(T)$, even at low temperatures. The same dynamical behavior was previously found for proteins in solutions and, due to the similarity of the systems, this suggests that the underlying process is the same. In particular, in the first state (state 1), the *E. coli* cytoplasm is a liquid where the majority of proteins are folded. After the transition (state 2), the cytoplasm is in a gel state, and the proteins are mainly unfolded. For simplicity, we will assume that in states 1 and 2, all the proteins are fully folded and fully unfolded, respectively. Therefore, the diffusion coefficient D_1 will describe the global diffusion of the *folded* average protein in the liquid state, termed $D^{(f)}$, whereas D_2 will be related to the global diffusion of the fully *unfolded* average protein in the gel state, termed $D^{(u)}$. Thus, for $D_G(T)$ we have:

$$D_G(T) = D^{(f)}(T) \cdot [1 - a_u^{\mathrm{QENS}}(T)] + D^{(u)}(T) \cdot a_u^{\mathrm{QENS}}(T) \tag{4.29}$$

In this picture, $a_u^{\mathrm{QENS}}(T)$ can be interpreted not only as a measure of the diffusion slow-down during the unfolding, but it also represents the apparent fraction of unfolded proteins that weights $D^{(f)}$ and $D^{(u)}$ and that is necessary to reproduce the values of D_G for systems in a partially unfolded state. In particular, our simulations suggested that the relation between the apparent fraction of unfolded proteins determined dynamically, a_u^{MD}, and the real fraction of unfolded proteins, r_u^{MD}, can be described by a power law:

$$a_u^{\mathrm{MD}} = \left[r_u^{\mathrm{MD}} \right]^p \tag{4.30}$$

Therefore, to estimate the real unfolded fraction as a function of temperature starting from the dynamical QENS results, we can use the inverse relation:

$$r_u(T) = \left[a_u^{\mathrm{QENS}}(T) \right]^{(1/p)} \tag{4.31}$$

In conclusion, comparing QENS experiments and MD simulation, first we verified that the simulations can reproduce the linearity of $D^{(f)}(T) = D_1(T) = T \cdot a_1 + b_1$, and $D^{(u)}(T) = D_2(T) = T \cdot a_2 + b_2$ (Figure 2D and 4.25A). Then, since protein unfolding is computationally demanding for all-atom MD simulations, instead of calculating extremely long trajectories to allow all proteins to unfold, we forced the unfolding (as described in Subsection "*Preparation of partially unfolded sub-volumes*"), and we performed a series of relatively short simulations ($\sim$100 ns), sufficiently long to determine the diffusion coefficient in the timescale explored by the QENS experiments, but brief

enough to ensure that the folding state of the proteins was not affected. This allowed us to verify that if the folding state remains constant, the diffusion coefficient of the average protein scales linearly with the temperature (Figure 3D and Figure 4.25 panel C), and this relation also holds for systems with different concentrations (Figure 4.16).

Comparing the diffusion coefficients calculated for a system simulated in different folding states (Figure 3D and 4.25C), we found a relationship between the amount of unfolded proteins and the consequent diffusion slowdown—eq. (4.30). Finally, with the inverse relation eq. (4.31) and the QENS measurement of the slowdown, we were able to estimate the real fraction of unfolded proteins at different temperatures in the *E. coli* cytoplasm through the study of its dynamical state (see Figure 4A).

4.4.3 Origins of the nonlinearity of the transfer function

The observation that the translation diffusion coefficient D_t cannot be expressed simply as

$$D_t(r_u) = (1 - r_u)D_t^{(f)} + r_u D_t^{(u)} \tag{4.32}$$

where r_u is the unfolded fraction and $D_t^{(f)}$ and $D_t^{(u)}$ are the translational diffusion coefficients corresponding to a fully folded- and a fully unfolded system, respectively, can be rationalized in the following way. If we assume the validity of the Stokes–Einstein relation, the diffusion coefficients $D_t^{(f)}$ and $D_t^{(u)}$, forming two limiting cases, can be written as

$$D_t^{(f)} = \frac{k_B T}{6\pi\eta^{(f)} R^{(f)}} \tag{4.33}$$

and

$$D_t^{(u)} = \frac{k_B T}{6\pi\eta^{(u)} R^{(u)}} \tag{4.34}$$

Here, $R^{(f)}$ and $R^{(u)}$ are protein hydrodynamic radii in the folded and unfolded states, respectively, and $\eta^{(f)}$ and $\eta^{(u)}$ denote the viscosities of the fully folded and fully unfolded volumes.

However, in intermediate boxes with an unfolded fraction r_u, $\eta(r_u)$ is neither identical to $\eta^{(f)}$ nor to $\eta^{(u)}$. Consequently, for the diffusion of folded/unfolded proteins in the intermediate boxes, we can expect

$$D_t^{(f)}(r_u) = \frac{k_B T}{6\pi\eta(r_u) R^{(f)}} \tag{4.35}$$

and

$$D_t^{(u)}(r_u) = \frac{k_B T}{6\pi\eta(r_u) R^{(u)}} \tag{4.36}$$

where the viscosity is a function of r_u. Therefore, rather than assuming Eq. 4.32, one should take into account the effect of the intermediate viscosity:

$$D_t(r_u) = (1 - r_u)D_t^{(f)}(r_u) + r_u D_t^{(u)}(r_u) = (1 - r_u)\frac{\eta^{(f)}}{\eta(r_u)} D_t^{(f)} + r_u \frac{\eta^{(u)}}{\eta(r_u)} D_t^{(u)} \tag{4.37}$$

If we consider a linear dependence of η on r_u,

$$\eta(r_u) = \eta^{(f)} + r_u(\eta^{(u)} - \eta^{(f)}) \tag{4.38}$$

with values of $\eta^{(f)}$ and $\eta^{(u)}$ estimated using simulations, we obtain a similar dependence of a_u on r_u as the one based on the actual diffusion coefficients determined from simulations (see Figure 4.27).

Further Results

scientific results obtained during the Ph.D. studies concerning the effects of protein-ligand complexation on protein dynamics

Chapter 5

Differences between Ca^{2+} reach and depleted α-La investigated by MD simulations and NS experiments

Based the publication:

Differences between calcium rich and depleted alpha-lactalbumin investigated by molecular dynamics simulations and incoherent neutron scattering.

Dominik Zeller, Pan Tan, Liang Hong, ***Daniele Di Bari****, Victoria Garcia Sakai, and Judith Peters*

Physical Review E 101, 032415 – Published 25 March 2020

In this chapter, we present a study comparing atomic motional amplitudes in calcium rich and depleted alpha-lactalbumin. The investigations were performed by elastic incoherent neutron scattering (EINS) and molecular dynamics (MD) simulations. As the variations were expected to be very small, three different hydration levels and timescales (instrumental resolutions) were measured. In addition, we used two models to extract the mean square displacements (MSDs) from the EINS data, one taking into account the motional heterogeneity of the MSD. At a timescale of several nanoseconds, small differences in the amplitudes between the calcium enriched and depleted alpha-lactalbumin are visible, whereas at lower timescales no changes can be concluded within the statistics. The results are compared to MD simulations at 280 and 300 K by extracting the MSDs of the trajectories in two separate ways: first by direct calculation, and second by a virtual neutron experiment using the same models as for the experimental data. We show that the simulated data give qualitatively similar results as the experimental data but quantitatively there are differences. Furthermore, the distribution of the MSDs in the simulations suggests that the inclusion of heterogeneity is reasonable for alpha-lactalbumin, but a bi-or trimodal approach may be sufficient.

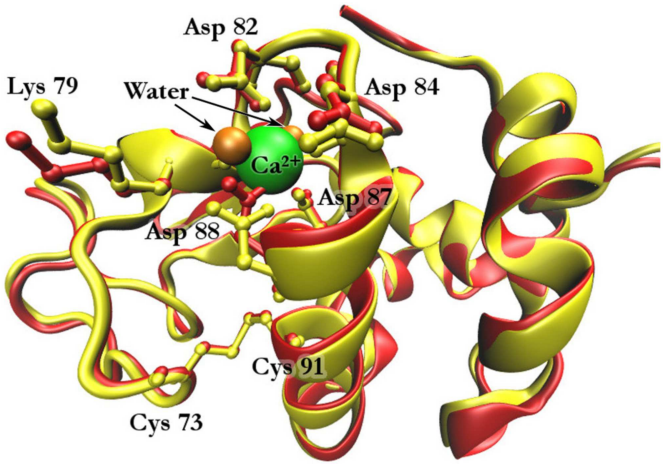

Figure 5.1: Calcium rich (yellow, bright) and depleted (red, dark) forms of α-La and the variations induced on their structures. α-La from most mammals consists of 123 amino acid residues [169] and its molecular 4 weight is $\approx$14.2 kDa.

5.1 Introduction

A number of proteins have the ability of binding ions that may lead to changes in the protein's structure and dynamics at the atomic scale, and subsequently, may affect their functionality. Recent studies have shown that for example, in the case of enzymes the presence of small inhibitors might influence the dynamics in a measurable way compared to the dynamics of the wild type form [169, 170, 171, 172, 173] and single point genetic mutation in proteins can affect collective density fluctuations in hydrating water [174]. Another case is alpha-lactalbumin (α-La), the major whey protein found in the milk of all mammals. It is a simple Ca^{2+} binding milk protein and has a significant role in biosynthesis of lactose in the lactating mammary gland. Together with the enzyme β-1,4-galactosyltransferase (β4GalT) it forms a complex which is responsible for the lactose synthase, i.e., transforming galactose and glucose into lactose. It strongly binds the cation Ca^{2+} and results in changes in the tertiary structure of the protein (see Fig. 5.1). Besides Ca^{2+}, the binding site can also bind Mg^{2+}, Mn^{2+}, Na^{+}, or K^{+}, which induce similar but smaller structural changes than Ca^{2+}. However, the corresponding binding constants are much lower except in the case of Mn^{2+}. In general, the binding of a cation stabilizes α-La and increases its thermal denaturation temperature. Furthermore, recently Shinozaki and Iwaoka [175] showed that Ca^{2+} and Mn^{2+} accelerates folding to the native form of α-La, an effect not seen with the other cations. α–La can also bind Zn^{2+} at several other distinct binding sites, but results in a decrease in the stability of α-La bound to Ca^{2+}. The apo form refers to α-La which is not bound to Ca^{2+}. Owing to the characteristics described above, α-La is often used as a simple model for Ca^{2+} binding proteins.

In addition to structural changes, the binding of α-La to Ca^{2+} may also generate structural rearrangements capable of influencing locally molecular dynamics and therefore varying the functionality of the protein. The task of probing such small effects is not easy and a sophisticated approach is required. Incoherent neutron scattering is a technique used to probe atomic and molecular dynamics on timescales of pico- to nanoseconds, and when combined with molecular dynamics (MD) simulations forms

a powerful partnership, and could indeed offer unique insights into changes occurring in the atomic scale (amplitudes of about a few angstroms). Despite the two techniques accessing very similar times and dimensions, they are not always in full quantitative agreement [176]. To better understand the reasons behind such disagreements, we choose to make a detailed study of the dynamics of α-La, which is commercially available, to do exhaustive neutron experiments, and sufficiently small to permit accurate simulations. We study the dynamics in both the Ca rich and depleted forms, which from hereafter will be referred to as α-La_{ca} and α-La_{dep}, respectively.

The intensities measured using incoherent neutron scattering experiments are commonly used to extract mean square displacements (MSDs) of the protons within the protein, almost exclusively by assuming a harmonic approximation of all possible dynamical contributions, i.e., the Gaussian approximation [67, 53]. Furthermore, it is common to combine results from different neutron spectrometers, since they cover different timescales and length scales. Recently, we applied models that go beyond this approximation and include dynamic heterogeneity, to be able to fully exploit a wider instrumental spatial window [60]. In this work, we combine data from three neutron spectrometers which access different timescales (have different energy resolutions) and length scales (have different momentum transfer coverage). We apply a few models, including the commonly used Gaussian approximation, to the data to investigate to what extent they help to disentangle small effects on the dynamics and make a comparison to results from MD simulations.

5.2 Experimental Section

5.2.1 Sample preparation for neutron experiments

All experiments described use bovine alpha-lactalbumin (α-La), either in its natural form with Ca^{2+} (α-La_{ca}) or Ca^{2+} depleted α-La_{dep}. The protein was purchased from Sigma-Aldrich in lyophilized powder form. Three different hydration levels were prepared for each batch and protein type, hydrated with heavy water, D_2O. This is so the neutron signal is dominated by the incoherent scattering from the protons in the protein (owing to the large incoherent neutron cross section of hydrogen compared to deuterium or other atoms constituting the protein structure [113]). The hydration level was determined from the difference in mass with and without D_2O and is defined as $h =$ grams D_2O/grams dry protein. The different levels of hydration were $h \approx 0$ (dry), $h \approx 0.4$, and $h \approx 0.8$. The dry lyophilized sample represents the case where only harmonic motions are present up to room temperature and $0.4h$ corresponds to around a hydration level of one or two layers of water on the protein surface [177], which is sufficient to allow for localized dynamical motions. Finally, $0.8h$ represents a gel state close to full hydration. The purchased lyophilized protein powder was dried for at least 24 h, after which it was weighed and then loaded into flat aluminum sample holders (standard for neutron spectroscopy experiments) and vacuum sealed with indium wire. The $0.4h$ and $0.8h$ samples were also dried before hydration, and then left in a D_2O rich environment to uptake the water. The samples were weighed periodically until they achieved the desired uptake of D_2O and then sealed with indium in similar flat aluminum holders. Masses of around 100 mg were used to obtain around 10% scattering and ensure there was no significant multiple scattering.

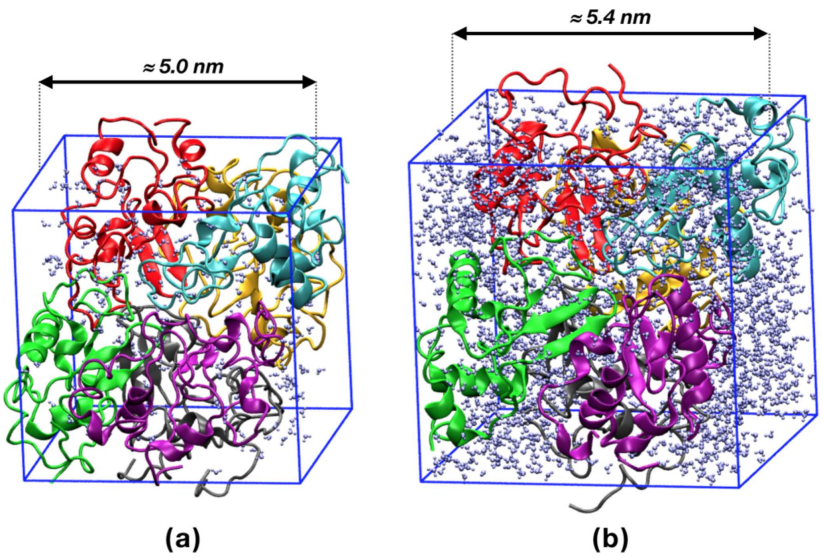

Figure 5.2: Visualization of the dry and hydrated simulations of α-La$_{dep}$. (a) Dry environment ($0.05h$). (b) Hydrated environment ($0.4h$). The lines show the simulation box size and inside the six chains of α-La$_{dep}$ at 300 K are visualized. The little (blue) molecules indicate water. For the sake of better visibility, we represented the proteins as connected molecules, therefore going beyond the limits of the box, and not as distributed within the same box due to the boundary conditions.

5.2.2 Simulation setup

MD simulations of hydrated protein powder were used representing the interactions of proteins with a small amount of water, to be able to compare them directly with experimental data. The simulations were started from two different protein structures which can be found in the Protein Data Bank (PDB) [178]: (1) bovine α-La with calcium (α-La$_{ca}$), PDB ID: 1F6S and (2) bovine α-La without calcium (α-La$_{dep}$), PDB ID: 1F6R. Both structures were published by Chrysina et al. [179]. Each PDB structure consists of six distinct α-La proteins (= chains), allowing to calculate an average dynamics of a single α-La chain. As a matter of fact, Tarek and Tobias [180] pointed out that a single protein covered by a shell of water is not sufficient to describe a powder protein by simulations. Instead, a crystal composed by two proteins or more resulted in a realistic model to reproduce neutron scattering data. This is the reason why, in the present case, we used six chains of proteins placed in each box (see Fig. 5.2).

The protein molecule was centered in a cubic box of size 8.39 nm at first, with the CHARMM27 force field [181, 182], and the TIP4P-EW water model [183], using GROMACS 5.0.7 (GPU version) as the MD engine [184, 124]. The boxes were filled with water molecules to start with, which were then deleted (starting from the outside) until the number of water molecules around the protein met the desired hydration level h. The box for a hydration level of $0.4h$ contained 1824 (dep) / 1834 (Ca) water molecules and 232 (both) for the dry system ($0.05h$). All systems were electrically neutralized by adding NaCl. Van der Waals interaction was truncated at 1.2 nm with the Lennard-Jones potential switched to zero gradually at 1.0 nm. A particle mesh Ewald [185] with a Coulomb cutoff of 1.2 nm was used to calculate electrostatic interaction. All bonds involving hydrogen atoms were constrained with the LINCS [141]

algorithm. The systems were firstly energetically minimized using steepest descent steps with a maximum force of 10.0 kJ mol^{-1} nm^{-1} and a maximum of 50000 steps. Then they were equilibrated in the NVT ensemble at T = 280 K (and 300K) for 300 ps and in the NPT ensemble at p = 1 bar for 50 ns, with a 0.5-fs time step to slowly release the unreasonable atom contact and suppress vacuum. The temperature coupling was performed using the velocity-rescale algorithm with a coupling time of 1 ps [143]. The pressure coupling was performed using the Parrinello-Rahman algorithm with a coupling time of 1 ps [186]. The production MD simulations for hydration level 0.4h were conducted in the NPT ensemble for 100 ns, with a 2-fs time step, while those of the dry systems were conducted for 500 ns. Only the last 20-ns trajectories recorded at every 2 ps were used for the analysis. For such a dense system, the global translation and rotation of the protein molecules was strongly suppressed [187, 188]. A visualization of the difference in box size and hydration level is shown in Fig. 5.2.

5.3 Neutron Scattering

Data were collected on three neutron spectrometers, all so-called inverted geometry spectrometers, covering a wide temporal range, namely, OSIRIS [189] at the ISIS Neutron and Muon Facility, UK; IN13 [190] at the Institut Laue Langevin, (ILL), Grenoble, France; and SPHERES [191] at the MLZ Munich reactor in Germany. The data collected at the ILL can be found under the DOIs in Refs. [192] and [193]. OSIRIS and SPHERES use crystal analyzers that reflect cold neutrons (λ of 6.27 and 6.66 $\AA$, respectively) allowing access to a momentum transfer range, q-range, up to 1.8 $\AA^{-1}$, whereas IN13 uses a thermal neutron crystal analyzer (λ of 2.23 $\AA$) which opens up the accessible q-range to 4.9 $\AA^{-1}$. This permits us to probe dynamics occurring in a variety of length scales, where distance $d = 2\pi/q$. In addition, the three instruments differ in energy resolutions allowing access to motions from a few picoseconds to a few nanoseconds. Specifically, they are 25, 8, and 0.7 μeV, for OSIRIS, IN13, and SPHERES, respectively. Transmission values for all samples were measured on IN13 to be above 90% so that multiple scattering effects were not taken into consideration for the data treatment. The initial data reduction was done with LAMP [194] for IN13, SLAW for SPHERES, and MANTID [72] for OSIRIS. Slab can corrections for a flat sample holder and normalizations providing the relative detector efficiency and the instrumental resolution were done with LAMP for the samples measured on IN13 and SPHERES. The measurements on OSIRIS were corrected using the empty sample holder and normalized in MANTID. All intensity normalizations were done with the lowest available temperature data of each scan. Therefore, the difference between the slab correction algorithm and the subtraction of the empty sample holder alone are negligible.

Incoherent neutron scattering measurements give access to the elastic incoherent structure factor (EISF), S, which is a function of the momentum transfer q at the elastic line, where the energy transfer $\hbar\omega$ that occurs between the neutrons and the scattering atoms (mostly hydrogen), as a result of the scattering event, is approximately zero. The most commonly used approach to analyze this intensity is to assume that the atomic nuclei undergo harmonic motions around their equilibrium positions [67] and thus fit the data to the so-called Gaussian approximation (GA). The intensity can then be expressed

as

$$S(q, 0 \pm \Delta E; \langle r^2 \rangle) \approx S_0 \exp\left(\frac{-q^2 \langle r^2 \rangle_{\mathrm{GA}}}{3}\right) \tag{5.1}$$

where ΔE corresponds to the instrumental energy resolution. From this expression, values for the static mean square displacements of the atoms, $\langle r^2 \rangle_{\mathrm{GA}}$, are obtained at each temperature point measured, by fitting the slope of the logarithm of the scattered intensities plotted vs q^2 according to

$$\langle r^2 \rangle_{\mathrm{GA}} \approx -3 \frac{\partial \ln S(q, 0 \pm \Delta E; \langle r^2 \rangle)}{\partial q^2} \tag{5.2}$$

The Gaussian approximation is strictly valid for $q \to 0$, and it holds up to $q^2 \langle r^2 \rangle_{\mathrm{GA}} \approx 1$, restricting the q-range that can be used for this type of analysis considerably.

A model that imposes no constraints on the q range is that developed by Kneller and Hinsen [195] and applied to experimental data by Peters and Kneller [63]. It differs from the GA in that it takes into account motional heterogeneity of the amino acid side chains and their environment, compared to the Gaussian approximation where only one atomic motion is representative for all hydrogens. The motional heterogeneity of the hydrogen atoms is described by a Gamma distribution and the corresponding elastic intensity can be calculated analytically as

$$S(q; \langle r^2 \rangle, \beta) = \frac{1}{\left(1 + \frac{q^2 \cdot \langle r^2 \rangle_{\mathrm{PK}}}{3\beta}\right)^{\beta}} \tag{5.3}$$

where β is a measure of the homogeneity in the atomic motions; e.g., when $\beta \to 0$ the Gaussian form is retrieved. Fits of the data give, then, access to the corresponding static mean square displacement, $\langle r^2 \rangle_{\mathrm{PK}}$, where PK stands for the Peters-Kneller model hereafter.

An earlier attempt to account for motional heterogeneity in modeling the EISF was suggested by Meinhold et al. [196] by describing the mean square motional amplitudes by a Weibull distribution. However, this approach is not investigated here.

5.4 Analysis of the simulated data

5.4.1 Direct calculation of the MSD

The α-La$_{\mathrm{ca}}$ (α-La$_{\mathrm{dep}}$) proteins used in the simulations consist of a total of 11512 (11457) protein atoms. The number of atoms is different in the two forms, because some chains are missing some amino acids (residues) at the end of the α-La chain since they were not resolved in the PDB structure. In order to compare the simulation data to the experiment, we analyze the H atoms in the protein, which account for the majority of the scattering signal in the neutron experiments. Furthermore, in order to be consistent in the evaluation of the MSD, only H atoms which are in all chains are considered: Every single α-La consists of at least 922 H atoms which are of the same type for all α-La protein chains. Therefore, with six single α-La chains in each

simulation, in total $6 \times 922 = 5532$ H atoms have been evaluated to calculate the MSD and thus to analyze the averaged atomic movements of the protein. The MSD of a single atom α at location $r_\alpha(t)$ at time step t in the simulation is calculated via

$$\mathrm{MSD}_\alpha(t) = \left\langle [r_\alpha(t_0) - r_\alpha(t+t_0)]^2 \right\rangle_{t_0} \tag{5.4}$$

with $\langle \cdots \rangle_{t_0}$ being the average over all t_0 defined by the time steps of the simulation. From these individual atoms, a mean $\mu(t)$ of the MSD can be calculated.

We first calculated the time average of the MSD according to Eq. (5.4) using the complete 20-ns trajectories. Further, to estimate the error of the mean of the MSD due to different conformations, the 20-ns simulations were truncated in four equally time spaced parts of 5 ns. The result of the four independent parts was then averaged to obtain a mean MSD $\overline{\mu}_i(t)$ and its sample standard deviation $s(t)$ taken as an estimation of the error:

$$s(t) = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N=4} [\mu_i(t) - \overline{\mu}_i(t)]^2} \tag{5.5}$$

Finally, we compared the MSDs obtained from a direct calculation with the ones using the fast correlation algorithm proposed by Kneller et al. [197]. See Figs. 5.7 and 5.8 in the supporting information section for the results of the different checks.

In order to compare the dynamic MSD $\overline{\mu}(t)$ of the simulations with the static MSD $\langle r^2 \rangle$ calculated by the models, it has to be divided by 2 since the static MSD $\langle r^2 \rangle$ is defined as a time independent quantity due to the confined motion resulting in [198, 59]

$$2 \langle r^2 \rangle = \mathrm{MSD}(t \to 0) \tag{5.6}$$

For convenience, in the following, the MSDs obtained from the simulations will be labeled as direct MSD $\Delta_{\mathrm{dir}}(t)$ with the following definition:

$$\Delta_{\mathrm{dir}}(t) = \frac{1}{2} \overline{\mu}(t) \tag{5.7}$$

The $\Delta_{\mathrm{dir}}(t)$ can be directly compared to the MSDs extracted from the fits [see Eqs. (5.2) and (5.3)].

5.4.2 Indirect calculation of the MSD

An alternative way to compare the simulation results with the experimental data is to extract the MSD from the convolution of the instrumental resolution $\mathrm{R}(\omega)$ with the theoretical dynamic incoherent structure factor (DISF) $S_{\mathrm{inc}}(q,\omega)$ calculated with the help of the simulation data. The DISF was calculated with the program MDANSE [199] (v.1.1). The resolution function $\mathrm{R}(\omega)$ for each instrument was approximated by a normalized Gaussian function with a full width at half maximum (FWHM) equivalent to the resolution of the instrument:

$$G(\omega,t) = \frac{1}{\sigma_{\mathrm{res}} \sqrt{2\pi}} \cdot \exp\left\{ -\frac{1}{2} \left(\frac{\omega}{\sigma_{\mathrm{res}}} \right)^2 \right\} \tag{5.8}$$

where

$$\mathrm{FWHM} = \sigma_{\mathrm{res}}\sqrt{8\ln 2} \approx 2.35 \cdot \sigma_{\mathrm{res}} \tag{5.9}$$

The FWHM of each instrument was obtained by matching the above-defined Gaussian function to data from vanadium which is used to measure, experimentally, the resolution of neutron spectrometers since it is an isotropic incoherent scatterer. For IN13 and OSIRIS data from a vanadium standard summed over all momentum transfers, q was used; for SPHERES, the resolution function found in the literature for the large angle detectors was used [200] (see Fig. 5.2) (Voigt profile with σ_{res} = 0.244 μeV; $\gamma_{\mathrm{res}} = 0.052\mu$eV). The resolution functions are then convoluted with the DISF which is obtained from the simulation. For each DISF calculated with an absolute momentum transfer q_m, N_q = 50 q-vectors q_i with a randomized direction and an absolute length of $q_i = q_m + \Delta q$, with $\Delta q \leq 0.05$ $\AA^{-1}$, are averaged. In total, the DISF is then calculated in MDANSE as

$$S_{\mathrm{inc}}(q_m,\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} I_{\mathrm{inc}}(q_m,t)\cdot\exp(i\omega t)\,dt \tag{5.10}$$

$$I_{\mathrm{inc}}(q_m,t) = \frac{1}{N_\alpha}\sum_{\alpha}^{N_\alpha}\left\langle \frac{1}{N_q}\sum_{i}^{N_q}\exp\{iq_i r_\alpha(t_0)\}\cdot\exp\{-iq_i r_\alpha(t_0+t)\}\right\rangle_{t_0} \tag{5.11}$$

where N_α is the number of H atoms in the simulation and r_α their location. t and t_0 are defined by the time steps of the trajectory.

From the resolution broadened DISF, $S^R_{\mathrm{inc}}(q,\Delta\omega)$, the elastic tic incoherent structure factor EISF(q_m) is computed by summing up the intensities in the range $\omega = \pm\mathrm{FWHM}/2$ and the resulting EISF(q) is normalized by EISF(q_m = 0). The obtained EISF(q) can be fitted in the same way as the experimental data to calculate the MSD. It is important to mention that for the experimental data the lowest temperature scan was used for the normalization, whereas here the value obtained at q_m = 0 was taken due to the lack of a simulation at very low temperature. The models chosen to analyze the EISF(q) are the same as for the experimental data, namely, the Gaussian approximation [Eq. (5.1)] and the PK model [Eq. (5.3)], over q ranges of $0-1\AA^{-1}$ and $0-4\AA^{-1}$, respectively.

5.5 Neutron Scattering Results

The MSDs were extracted as described above for the three instruments and according to the two models. The PK model fits data over a much wider range of q values, and thus is expected to yield additional information about the motional amplitudes. As, for instance, methyl group rotations are small motions and become particularly visible only at higher q values ($> 2\AA^{-1}$) [201, 202, 203], such treatment is expected to give a more precise description. However, it includes one more fit parameter and gives thus higher error bars for the fitting parameters. No significant differences were observed within statistical error between the dynamics (and MSD) of α-$\mathrm{La_{dep}}$ and α-$\mathrm{La_{ca}}$ on the timescales of OSIRIS or IN13 at any of the three hydrations. An example of this is shown for the OSIRIS data fit using the PK model, in Fig. 5.3 (top). Furthermore, both models show similar trends. Absolutely no differences are appreciable in the dry pro-

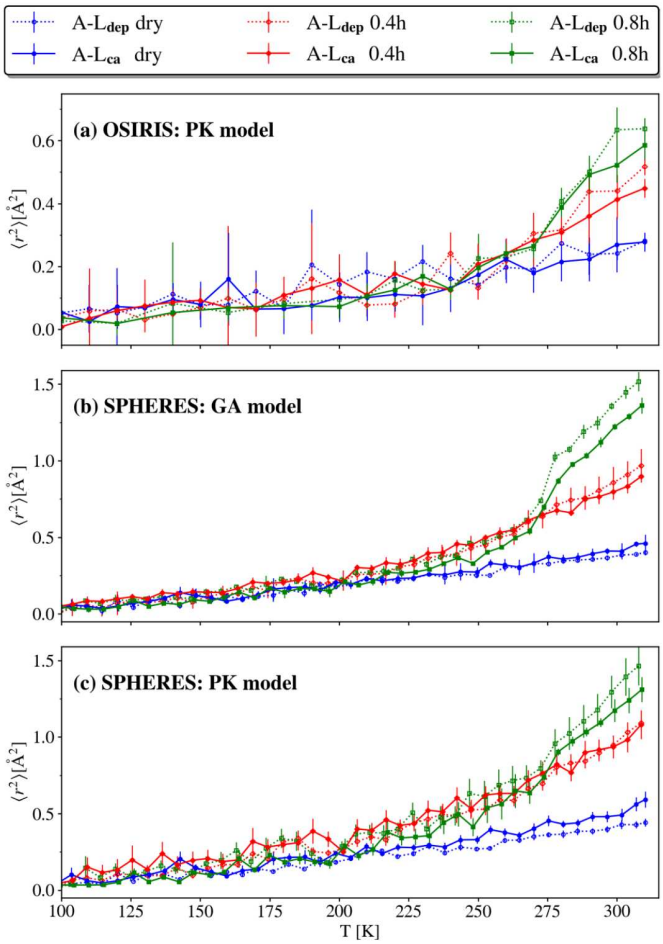

Figure 5.3: Difference of MSD between α-La$_{dep}$ and α-La$_{ca}$ for OSIRIS (a) and SPHERES (b,c). Data are only shown from 100 to 310 K to enhance the differences, since at temperature $<$ 100 K there are no differences within statistical error. The MSD values of the α-La$_{dep}$ are shown as dotted lines and empty symbols, and those for α-La$_{ca}$ as solid lines and filled symbols. The dry samples are blue (lower curves), the 0.4h samples are red (middle curves), and the 0.8h samples are green (upper curves).

teins (blue curves). At $h = 0.4$ (red curves) and $h = 0.8$ (green curves), the dynamics are almost identical except at the higher temperatures (290 - 310 K), where small differences are visible and both models suggest that the α-La$_{ca}$ has a slightly smaller MSD, indicating less dynamics. Given the large error bars, the effect is not conclusive; however, the trend would confirm the findings of Chrysina et al. [179], which suggest that the binding of a protein to a cation stabilizes the protein, irrespective of the hydration level.

Differences between the dynamics of the two samples are visible on the timescale of the SPHERES instrument (Figs. 3(b) and 3(c), slower dynamics up to a couple of nanoseconds). Already in the dry state, the α-La$_{ca}$ sample has a slightly higher MSD than the α-La$_{dep}$ sample above 250 K. This difference is emphasized when using the model that uses a larger q range, the PK model, suggesting that also smaller amplitudes (corresponding to higher q values) have to be included in the analysis to permit such a subtle differentiation.

At h = 0.4 no difference between the samples is observed in the PK model, and

using the GA model gives a small difference where the α-La_{dep} is more mobile than the α-La_{ca}. This could indeed be the case, as for the highly hydrated samples ($h = 0.8$) the same trend is observed, and, more apparent, the MSD is larger for the α-La_{dep} above 270 K. It is in fact the opposite behavior as for the dry, but more in line with the expected scenario of stabilization of the α-La upon binding calcium. A higher resilience of a protein upon binding of a cation indicates an increased free energy including a higher enthalpy arising from bonded interactions [204]. Entropy is likely rather unchanged at the same hydration level.

It appears that the sample hydrated at $0.4h$ presents higher dynamics between 200 and 270 K than the one hydrated at $0.8h$. Similar effects were already observed for the green fluorescent protein (GFP) [205, 206] and interpreted by the authors as a suppression of protein dynamics at lower temperatures by hydration water and an enhancement of it at higher temperatures. Moreover, in the $0.4h$ sample, the water is in a confined or glassy state so that secondary relaxations set in upon heating, whereas in the $0.8h$ sample where water is primarily bulk water, it is in a frozen state. The steplike increase of the MSD around 270 K for the highest hydrated sample corresponds thus to the melting of the surrounding water.

The GA model shows a more pronounced feature at the melting of ice in the higher hydrated sample compared to the PK model. The GA model covers indeed only larger length scales representing more likely the melting of the ice, whereas in the PK model, which also covers local length scales, an average of the larger and smaller length scales slightly smears out such effects.

5.6 MD Simulation Results

The MSD obtained directly and indirectly from the simulations were compared. To relate the time dependent direct MSD $\Delta_{dir}(t)$ to the results of the time independent indirect MSD $\langle r^2 \rangle_{ind}$, Heisenberg's uncertainty principle was used:

with $\tau_{FWHM} = \hbar/\text{FWHM}$. The corresponding times for each instrument are summarized in Table 5.1.

Table 5.1: Instrument resolution FWHM vs time. Relation of the FWHM of Gaussian instrument resolution in energy space to the time τ in time space.

Instrument	**FWHM** (μeV)	$\tau_{\mathbf{FWHM}}$ (ps)
OSIRIS	24.8	30
IN13	10.8	60
SPHERES	0.62	1060

Figure 5.4 shows the results of the MSDs of the two different fitting models GA and PK (blue and red, respectively) and the directly calculated MSDs (black), for the two simulated sam- ples (dry and hydrated) at all three instrumental resolutions. The results obtained with the various methods to calculate the MSDs directly were so close (differences below 0.5%) that it was not possible to represent them individually in Fig. 5.4. For the dry protein, the MSD calculated via the direct method is always larger than the MSD calculated from the models, but the behavior between the simulations is the same. Assuming that the direct calculation represents a result as close as possible

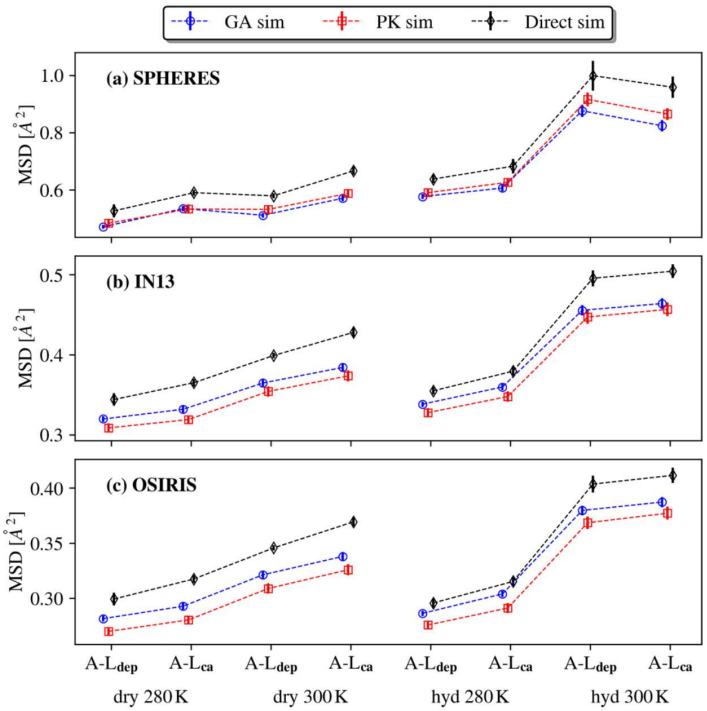

Figure 5.4: MSDs from analyzing data from MD simulations using both indirect (GA approximation: lower curves, PK model: middle curves) and direct calculations (upper curves) for the Gaussian resolution function of SPHERES (a), IN13 (b), and OSIRIS (c). On the left side the dry α-La is shown and on the right side the hydrated protein at $0.4h$. Each side is ordered in the same way by increasing temperature (280 and 300 K) and α-La$_{\text{dep}}$ is next to α-La$_{\text{ca}}$.

to the true MSD, the difference between the values of the direct method and those from the models could indicate the order of magnitude of the error introduced by using models. As anticipated, the MSD increases with increasing temperature and in fact the effect is larger on smaller timescales, i.e., at lower instrumental resolutions (IN13 and OSIRIS). The α-La$_{\text{ca}}$ simulations also have a slightly higher MSD for all instrumental resolutions. When comparing the different models, the GA evaluates to a higher MSD than the PK model for IN13 and OSIRIS. For SPHERES this behavior is inverted.

Similar trends are observed for the hydrated protein, except for three main differences. First, the MSD of the models is much closer to the direct MSD, albeit still smaller. Secondly, the difference between the MSDs at 280 and 300 K is much larger. Thirdly, for SPHERES the MSD for α-La$_{\text{ca}}$ at 300 K is slightly lower than for α-La$_{\text{dep}}$, which is the case for all methods considered.

The simulation also allows us to calculate the distribution of the MSDs for the protons in the protein. This is calculated following the method used by Yi et al. [207] which enables an evaluation of the main contributions to the heterogeneity and of how many populations with different motions are present. Figure 5.5 shows the distribution at $t = 30$ ps (OSIRIS), $t = 60$ ps (IN13), and $t = 1$ns (SPHERES) for all simulations. The curve for each time was obtained by binning the individual direct MSD values in steps of 0.02 $Å^2$ together and normalized by the total number of H atoms. The individual direct MSD values were obtained by averaging the value of the four independent slices of 5 ns for each simulation in the same way as for the direct MSD evaluation. In all simulations and for all three times t , one large peak at around 0.13 $Å^2$ is visible (dashed

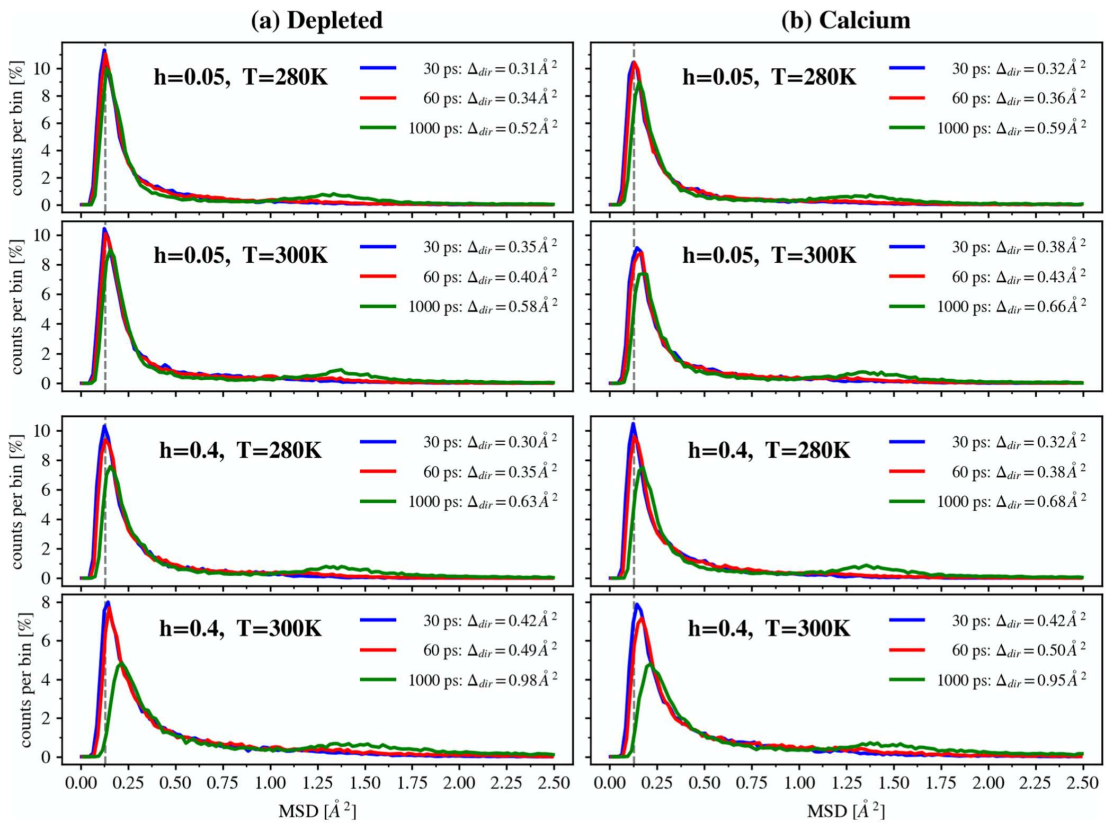

Figure 5.5: Distributions of the direct MSDs. Comparison of distributions of the MSDs at t = 30 ps (OSIRIS, highest curve in the peak), 60 ps (IN13, middle curve in the peak), and 1000 ps (SPHERES, lowest curve in the peak) for α-La_{ca} (b) and α-La_{dep} (a). The MSD values were obtained by the average value of the four independent slices of 5 ns for each simulation. Δ_{dir} in the legend shows the mean value of the distribution as defined in Eq. (5.7).

vertical line). Only for the distribution at 1 ns (green) a small second peak around 1.35 $Å^2$ is visible. The latter peak was identified in the simulations to correspond to methyl group rotations, which exist also within the IN13 data [63, 201], but are then retrieved at much smaller MSD values below 0.5 $Å^2$ and cannot be separated from the motions in the main peak. For the hydrated samples the first peak is shifted slightly to higher MSD values and its peak is significantly smaller than for 30 and 60 ps. This effect is emphasized at 300 K. No significant variation is observed between the α-La_{dep} and α-La_{ca} samples. This seems understandable as the components forming the two samples are extremely similar.

Such analysis helps us to understand if a complete distribution of Gaussian motions is required to describe the MSD or if a bi- or trimodal approach is sufficient. According to our results, a bimodal description seems to be very reasonable, in agreement with recent works of Vural et al. [198] or Doster [201].

5.7 Comparison of Experimental and Simulated Results

To compare the MD simulations with the experimental data, the results of the fitting models from the previous section are plotted together with the experimental results.

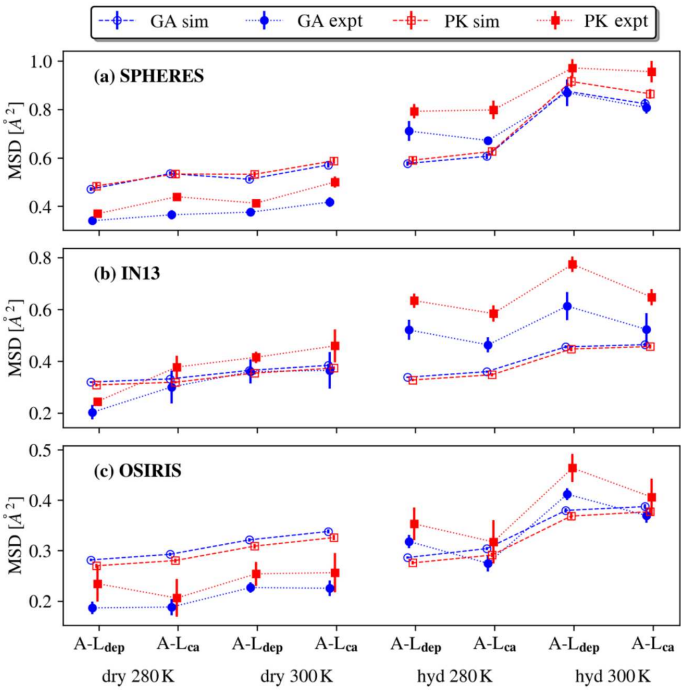

Figure 5.6: Comparison between the three models obtained by fitting the experimental data (open symbols, dashed lines) and the MD simulations (filled symbols, dotted lines). The circles designate the GA approximation and the squares the PK model.

The experimental data were collected at 5-10 K intervals, which unfortunately are not always in coincidence with the two simulation temperatures. To reduce the effects on the results, experimental MSD values were averaged over three temperature values (smoothing average) and then the MSDs between two smoothed temperature data points have been linearly interpolated. This procedure ensured that the simulated and experimental data were at the same temperature as the simulations.

As can be seen in Fig. 5.6, the MSDs extracted from the simulated data agree well with the experimental data and indicate that the different models hardly allow differentiation. Furthermore, the variations between simulated and experimental results may arise mainly from the instrumental limitations. Finally, the differences between the models are larger for the experimental data reflecting the worse statistics.

The experimental MSDs of the depleted hydrated sample seem systematically higher than those of the α-La_{ca} sample, which is hardly visible within the statistics in the simulation results at 280 K and below 1 ns. It indicates slightly enhanced dynamics for the depleted sample in such conditions, which could be expected as calcium has a stabilizing effect [208]. The higher mobility becomes visible only in the simulations at higher temperatures and longer timescales, as the variations in the sample are certainly small. An interesting point is the difference between the models. For the simulations, the PK model mainly evaluates a very slightly smaller MSD values whereas for the experiments they were larger. One has to note that each spectrometer has not only its specific time resolution, but also a characteristic q-range. Both dimensions are important and are related. Therefore, the evaluated results do not only depend on the time resolution, but also on the accessible spatial domain, which permits us to see various behaviors of the samples.

5.8 Discussion and Conclusion

MD simulations are a very powerful tool to understand, in more detail, the dynamics of individual atoms that are measured for a sample in a neutron scattering experiment, as both techniques give access to comparable temporal and spatial scales. Unlike the common simulations run in solution, comparison to elastic incoherent neutron scattering (EINS) measurements, frequently done with hydrated powders, has required the development of approaches to simulate hydrated powders [207, 209] by adapting the setup accordingly.

A direct comparison of neutron data and simulated signals is not always trivial as the absolute values depend significantly on, one hand, data corrections and normalization, and on the other, on the accuracy of force fields and starting structures. It is also common to find that simulations cannot reproduce results extracted from neutron scattering data quantitatively (see [176] or Fig. 5.6), hence, the decision to compare MSDs by extracting them in a very similar way from both experiment and simulation. In addition, we checked that different approaches to calculate the static MSDs from the simulated trajectories gave identical results.

The order of magnitude of the values of experimental MSDs are well reproduced by the simulations, with MSDs of hydrated samples being larger than those of the dry samples. The results indicate that the models describing the simulated EISF (obtained from the DISF) underestimate the simulated directly calculated MSDs (see Fig. 5.4). One might therefore speculate that no model is able to take the whole dynamics into account and that effects due to the limited space and time windows are not negligible. For the hydrated protein, the differences are not as large as for the dry protein. In addition, the difference between the models is not negligible, but the trends are always the same and in agreement with the direct MSDs; i.e., all curves obtained through the different models are mainly parallel. Interestingly, the hierarchy between the models is always the same for an identical instrument resolution (with the exception of SPHERES in a dry environment). One can therefore conclude that the GA gives equally good results as the other models, since the absolute values of the MSDs are unknown.

In comparison to the experimental data, the simulation cannot provide reliable quantitative results (see Fig. 5.6). In general, the experimental MSDs of the PK model are higher than those from the GA model, whereas for the simulations this trend is inverted in most cases. Here, it has to be stressed again, this behavior is highly dependent on the chosen q-range and thus no definitive trend can be concluded. The experimental curves show larger differences and in particular the GA model gives MSDs which are more strikingly different from the MSDs obtained through the PK model. Nevertheless, none of these results favors any one model over another, as the statistics are probably not good enough to discriminate small effects, eventually due to the different q-ranges used.

As shown by Fig. 5.5 the distribution of the MSDs can be mainly described by two different peaks which are independent of hydration. The second peak is most visible above 1 ns, whereas below 60 ps it is not well distinguished. It is mainly the H atoms of the methyl groups (not shown here) that are contributing to this peak, which is in accordance to the findings of Yi et al. [207]. Methyl group rotations indeed contribute to the elastic neutron spectra and the findings here support that they are a major contributor to heterogeneity originating from these motions, which becomes more visible at longer

timescales. Yi et al. [207] simulated the camphor-bound cytochrome P450 at $h = 0.4$ in a way comparable to the simulations here. They also showed that this peak is more dominant at higher temperatures. Furthermore, the second peak at larger amplitudes is also more pronounced at 1 ns. At 100 ps it is closer to the first peak and much broader. In addition, Tokuhisa et al. [58] simulated staphylococcal nuclease (SNase) in a water box at 300 K and also found two distinguishable peaks. The time was not documented but the evaluated simulation time was 1 ns, indicating that the investigated time window was likely smaller than 100 ps.

Overall this leads to the conclusion that the two models give reasonable results in comparison to the direct MSDs from the MD simulations. For a precise data set, the differences between the models are not significant concerning the trends, but the quantitative values are, depending on the evaluated q-range. The PK model gives further insight into the standard deviation of the MSD, but with respect to the MSD it does not give more accurate results. Furthermore, it is also important to state again that in contrast to the experimental data, the simulated EISF was not normalized to the lowest temperature data due to the lack of such simulation data, which could also partly explain the quantitative differences. Doing that, one would more consistently treat experimental and simulated data and eliminate more uncertainties, which might arise.

Supporting Information

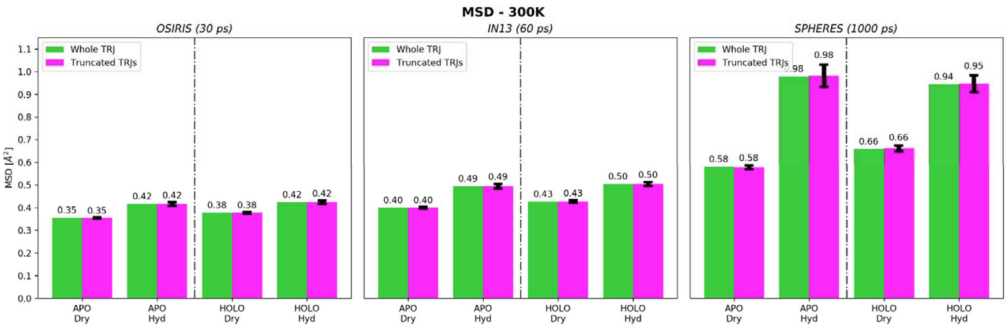

Figure 5.7: Comparison of the time averaged MSD according to eq. (5.4) using the complete 20 ns trajectories with the MSD extracted from truncated trajectories of 5 ns. Error bars were only obtained in the latter case. Both evaluations used the algorithm of Kneller et al. [197].

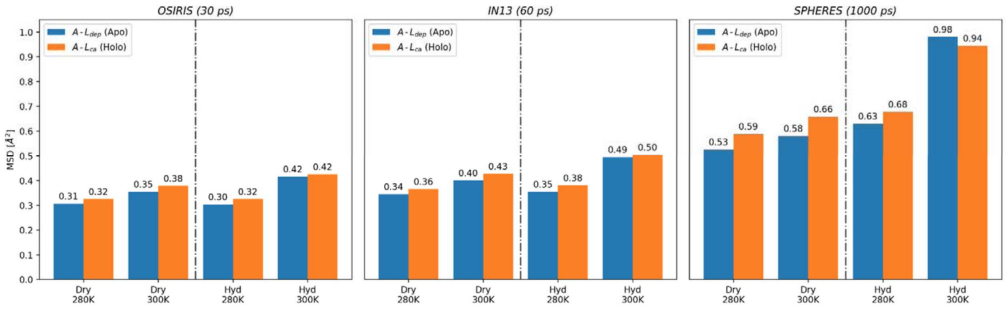

Figure 5.8: Comparison of the MSD obtained from a direct calculation with the ones using the algorithm proposed by Kneller et al. [197]. Both methods use the 20 ns trajectories.

Acknowledgments D.Z. was supported by a Ph.D. scholarship cofunded by the Communauté Université Grenoble Alpes, the STFC Rutherford Appleton Laboratory, and the Institut Laue Langevin. D.D.B. acknowledges support via a scholarship from the da Vinci program of the French-Italian University. J.P. and V.G.S. gratefully acknowledge the support by M. Johnson to partly acquire the Ph.D. grant. The authors gratefully acknowledge the financial support provided by JCNS to perform the neutron scattering measurements at the Heinz Maier-Leibnitz Zentrum (MLZ), Garching, Germany. Experiments at the ISIS Neutron and Muon Source were supported by a beam time allocation from the Science and Technology Facilities Council. We thank the ILL for beam time allocation. D.Z. thankfully appreciates the assistance of M. Zamponi during the experiment at the MLZ and the MDANSE software support of E. Pellegrini and R. Perenon.

Chapter 6

Role of low-frequency vibrational dynamics of protein hydration water for ligand binding

Based on a paper in preparation:

Role of low-frequency vibrational dynamics of protein hydration water for ligand binding.

***Daniele Di Bari**, Judith Peters, Jacques Ollivier, Andrea Orecchini, Caterina Petrillo, Fabio Sterpone and Alessandro Paciaroni*

(to be submitted to Physical Review Letter)

We studied the low-frequency vibrational behavior of the hydration water of the maltose binding protein (MBP), a prototypical biomolecule where a large scale hinge-bending conformational change from the apo/open to the holo/closed state, the so-called Venus-flytrap mechanism, is key for ligand binding. By using inelastic neutron scattering spectroscopy we show that, upon complexation with maltose, not only the MBP, but also its hydration water displays significant low-frequency vibrational changes. The character of the MBP appears substantially softer in the *apo* than in the *holo* state, while the opposite is true for its hydration water. Normal mode analysis supports the experimental results and allows to identify their microscopic origin. The alternative energetic vibrational contributions of protein and protein hydration water support the view of the induced fit mechanism as underlying the MBP conformational switching.

6.1 Introduction

An accurate knowledge of the dynamics of proteins is necessary to deeply understand how they perform their biological activity [210], a question that is fundamental to engineer proteins for specific functions and design next-generation therapeutics. Apart from the dominant contribution of large conformational changes and domain move-

ments, the degree of protein flexibility which is relevant to a given biological process comes also from more subtle dynamical mechanisms [210]. These motions include low-frequency (< 20 meV; < 5 THz) vibrational modes, the spectroscopic signatures of which have been measured in different functional states of proteins by inelastic neutron scattering (INS) [211, 212], optical Kerr effect spectroscopy [213] and anisotropic terahertz microscopy [212, 214]. In the case of complexation, the protein vibrational density of states may either undergo softening [211] or stiffening [212], with significant consequences in the increase or reduction of the complex flexibility. Quite remarkably, these changes affect, mainly through the vibrational entropy term, the stability of the protein–ligand complex and the binding affinity [211, 212]. On the other hand, the flexibility of a protein depends also on the dynamics of the water molecules interacting with its surface, the so-called hydration water (HW) [215]. On the functional point of view, there exists already evidence that the HW diffusive dynamics plays a significant role in assisting enzyme-substrate interactions [216, 217]. On the other hand, the role of the HW vibrational dynamics in protein-ligand recognition processes has been so far largely neglected.

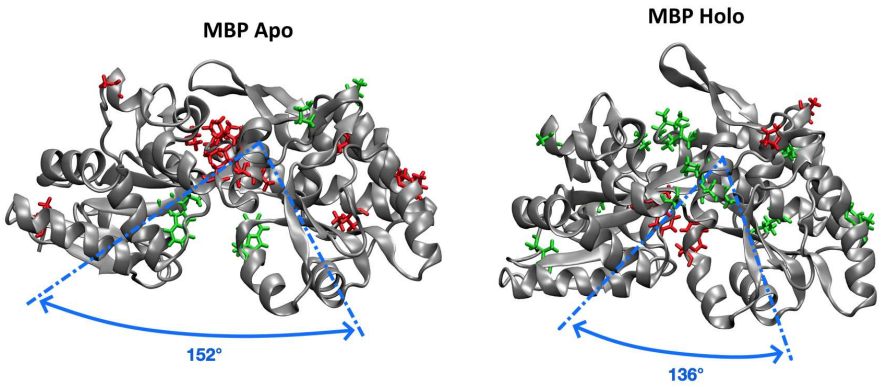

Figure 6.1: Illustration of the APO and HOLO forms of the MBP. The formation of the MBP+MALT complex lead to an important conformational change, that can be measured by the variation of the open anlge.

On these grounds we studied by INS spectroscopy the low-frequency vibrational behavior of the HW of the maltose binding protein (MBP), a prototypical member of a periplasmic-binding protein (PBP) superfamily [218]. Structurally, PBP proteins share a two-domain architecture (N-terminal domain, NTD, and C-terminal domain, CTD) with a central inter-domain cleft where the ligand is trapped following a large scale hinge-bending conformational change from the apo/open to the holo/closed state, the so-called Venus-flytrap mechanism. This conformational change is key for cellular metabolism [219], drug design [220] and biosensor development [221]. It is widely accepted that ligand recognition by MBP proceeds through an induced fit (IF) mechanism, by which the ligand binds the open state and prompts a transition to the closed state [222, 223], however the energetics behind this conformational change is still unclear. Here we investigate the role played by protein HW vibrational changes upon complexation with maltose. We show that not only the very protein, but also its HW displays significant low-frequency vibrational changes, with the character of such changes being opposite in the MBP compared to its HW in terms of softening. The microscopic origin of this effect is dissected by exploiting normal mode analysis (NMA). Biomolecule

and solvent have distinct and opposite contributions to the complexation vibrational free energy change.

6.2 Methods

6.2.1 Inelastic Neutron Scattering

INS experiments were done on the time-of-flight spectrometer IN5 (Institut Laue-Langevin, Grenoble) with an incident wavelength of 5 Å, (energy resolution $\sim$ 0.09 meV) on apo-MBP (uncomplexed) and holo-MBP (complexed with maltose) hydrated powders. The most significant contribution to the measured spectra comes from the incoherent signal of the hydrogen atoms. In the case of the samples hydrated with heavy water, i.e. apo- and holo-MBP at h=0.40 g D_2O/dry protein, the incoherent signal arises mainly from the protein non-exchangeable hydrogen atoms [113] that are copiously and almost uniformly distributed throughout the whole protein, thus allowing for a complete sampling of its molecular vibrational motions, within the time and spatial windows defined by the resolutions and ranges of the experimental energy transfer E and wave-vector transfer q. On the other hand, for apo- and holo-MBP hydrated with normal water, h=0.36 g H_2O/dry protein, also the hydrogen atoms belonging to the HW give a substantial contribution to the measured spectra. The hydration degree h has been chosen to ensure that the average protein was in the presence of approximately the same number of H_2O or D_2O molecules in each sample corresponding to 1 - 2 layers at the surface. The transmission of the samples, whose mass amounted to about 40 mg, ranged from 0.93 to 0.96. Multiple scattering and multiphonon contributions were estimated and considered negligible. The samples, placed in aluminum standard slab cells, were oriented at 135 °. A temperature of 150 K was chosen, in such a way that anharmonic effects can be disregarded. Before any data analysis, the raw spectra were corrected for empty cell contribution and self-absorption, and normalized to a vanadium standard.

6.2.2 Normal Mode Analysis

All the simulations and the normal mode calculations were performed with GROMACS 2019.4 [124] with the CHARMM36m [224] force-field . The crystallographic coordinates of apo-MBP and holo-MBP were taken from the Protein Data Bank, entry 1OMP and 1ANF, respectively. The VMD software [225] and the python module parmed [226] have been used to add the positions of the missing atoms, and generate the topology file for the ligand. Positions the crystal waters were not removed. Following the approach developed by Tarek and Tobias [180], using the PACKMOL software [227], we created two boxes where we placed, at random, eight copies of the MBP proteins in the apo and holo forms, respectively, to reproduce the neutron experiments performed on an amorphous powder. Then, we hydrated the two systems with TIP3P H_2O molecules to a hydration level of 36% (i.e. 36 g of H_2O per 100 g of protein), and we added 64 Na ions in each box to neutralize the two systems.

To obtain the best staring configuration for the NMA, after a short energy minimization to eliminate unfavorable contacts, we performed a series MD simulations to relax the obtained structures. The particle mesh Ewald (PME) method was employed to compute the long-range electrostatic interactions, and a cutoff distance of 1.2 nm

was used for calculation of the short-range electrostatic and van der Waals interactions. For the minimization, the steepest descent algorithm was used to a tolerance of 1000 kJ/mol/nm. The MD simulations were carried out with an integration time step of 1 fs. The systems were first heated to 300 K with a temperature increment of 10 K after each 50 steps. At each step, a temperature equilibration (NVT) was followed by a constant pressure and temperature MD run (NPT) to allow the proteins to interact with their neighbors and periodic images, leading to a contraction of the boxes. After a longer NPT equilibration at 300 K of 100ns, the systems were cooled down to 150 K with a reduction of 10 K every 50 ps. A further long NPT equilibration was performed until the system reached a stationary volume. To keep constant the required temperature in the NVT and the NPT runs, we used the V-Rescale thermostat, coupled with the Parrinello-Rahman barostat to maintain the pressure at 1 bar.

At the end of the last NPT run, the coordinates of each of the eight proteins, both for the apo and holo systems, were taken separately with their closest 791 H_2O molecules (36% hydration level). Each hydrated protein was energy minimized with the steepest descent algorithm [cit] followed by the low-memory Broyden-Fletcher-Goldfarb-Shanno method until the maximum force in the system was smaller than 10^{-4} kJ/mol/nm.

6.3 Discussion

A direct view of the way the low-frequency modes of MBP and its hydration water are affected by ligand binding is provided by Fig. 6.2, where the dynamic structure factor $S(q,E)$ as measured by inelastic neutron scattering is reported. Panel (b) shows that the dynamic structure factor of holo-MBP displays an excess of inelastic signal compared to apo-MBP, This behaviour is in line with that reported for dihydrofolate reductase protein [211], but deviates from the low-frequency trend found for lysozyme [213, 212] and aspartate aminotransferase [228] where ligand binding is related to a rigidified vibrational dynamics. Due to the close interaction of water molecules with the protein surface, one may ask whether there will be a vibrational change upon complexation also in the protein HW dynamic structure factor. The hydration water spectra have been obtained by properly subtracting the contribution of apo-MBP+D_2O (holo-MBP+D_2O) from the signal of MBP-APO-H_2O (MBP-HOLO-H_2O), to obtain the HW $S(q,E)$ for the free (complexed) protein (See the SI for details). Panel (c) of Fig. 6.2 shows that the apo-MBP HW displays an excess of inelastic signal compared to holo-MBP HW, a behavior opposite to that of the protein. Indeed, to quantitatively describe the vibrational changes occurring upon complexation to both the protein and its HW, it is convenient to calculate the corresponding vibrational density of states $g(E)$, which is directly related to the dynamic structure factor by $g(E) = \lim_{q\to 0} \frac{6E}{\hbar q^2}(\exp^{\frac{E}{K_B T}} - 1)S(q,E)$. Actually, the $g(E)$ is mainly the proton weighted vibrational density of states, as it derives from the average over the strong incoherent signal from all the hydrogen atoms. The vibrational modes of the $g(E)$ shown in panel (c) of Fig. 6.2 arise mostly from the collective displacements of both protein groups of atoms and water molecules, while in panel (b) and panel (c) the protein and HW contributions are singled out respectively.

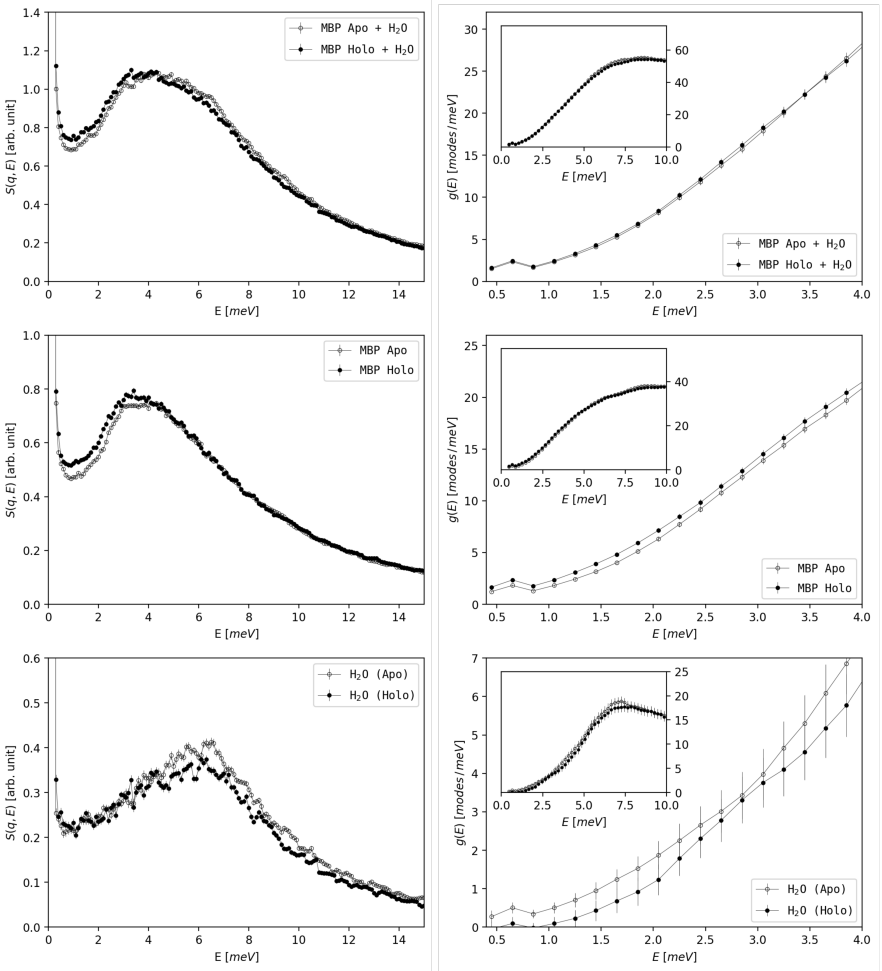

Figure 6.2: Dynamic structure factors (*left*) and vibrational density of states (*right*) for the MBP+HW, MBP, and HW in the APO and HOLO states.

The $g(E)$s have been normalized to their absolute values using the procedure described in Ref. [211] (See the SI for details). A vibrational difference beyond the errorbars between the free and the complexed protein is visible in the low-energy range from 2 meV to 4 meV, with a higher number of low-frequency large-amplitude modes appearing in the holo state. This excess of modes provides complexed MBP with additional flexibility, a feature also known as softening. On the other hand, the protein HW displays a reverse behaviour, with the free system showing an excess of low-frequency modes compared to the case in the presence of maltose. To explain the origin of the antithetical trend shown by the protein and HW $g(E)$ we examined the structural changes occurring to MBP during the apo/open to holo/closed transition. In particular, a key role for this transition is played by protein hinge region (residues K170-D180), which includes a short α-helix and a two-stranded β-sheets NTD [229]. The greater flexibility of this region after complexation, also revealed by the increased average B-factors [230], is related to the excess of low-energy excess of modes displayed by holo-MBP. Indeed, numerical $g(E)$ non only reproduces the trend observed by experiments (Fig. 6.3), but also confirms that the low-frequency vibrational density of states of the atoms in the hinge region is larger in the holo than in the apo conformation and is predominant over the remaining protein contribution (see Fig. 6.4).

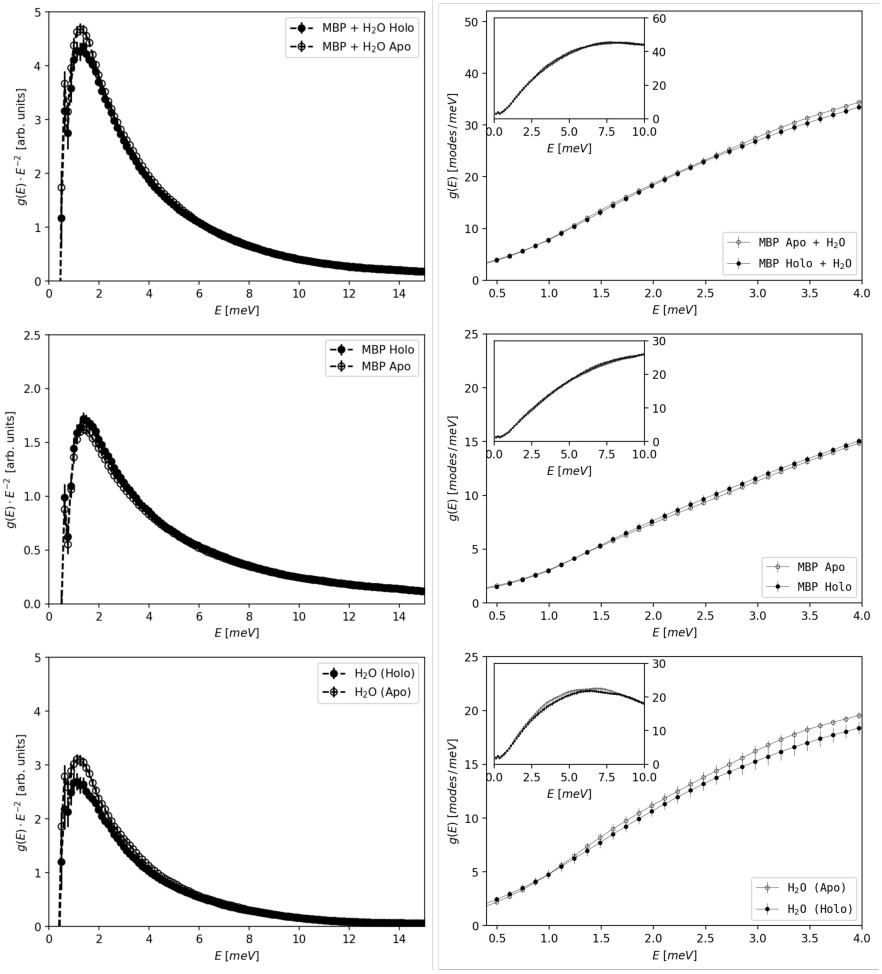

Figure 6.3: Vibrational density of states (*right*) for the MBP+HW, MBP, and HW in the APO and HOLO states, calculated by NMA. (*left*) Estimation of the dynamical structure factor starting as $g(E)/E^2$ by NMA.

On the other hand, Fig. 6.4 shows that the apo-MBP HW is more ordered than the holo-MBP, and this means that it has a lower density [231].

On the thermodynamic point of view, the features in the $g(E)$ reflect directly into the changes of the vibrational free energy A_{vib} of both the protein and its hydration water on passing from the apo/open to the holo/closed conformation. In the harmonic approximation one can easily estimate such changes through the relationship [232]:

$$\begin{aligned}\Delta A^{\alpha}_{vib} &= A^{\alpha}_{vib,holo} - A^{\alpha}_{vib,apo} = \\ &= K_B T \int \ln\left[2\sinh\left(\frac{E}{2K_B T}\right)\right]\left(g^{\alpha}_{holo}(E) - g^{\alpha}_{apo}(E)\right) dE\end{aligned} \tag{6.1}$$

where $\alpha = MBP$ or HW. As for the protein, it turns out that ΔA^{MBP}_{vib} = -(20 ± 3) kJ/mol, i.e. the vibrational change provides a favorable contribution for complexation. On the other hand, and quite surprisingly, the contribution relative to the MBP HW ΔA^{HW}_{vib} = (14 ± 5) kJ/mol promotes the stabilization of the open state. These results are in agreement with the fact that just the MBP intrinsic dynamics is insufficient for a 'selection'

mechanism [233], thus supporting the IF model for this enzyme. The emerging picture is one where substrate-mediated interactions are necessary to induce the protein open-to-closed conformational transition and overcome the energy barrier separating the MBP apo/open and holo/closed conformational state [222] to which the protein HW vibrational dynamics contributes significantly through the ΔA_{vib}^{HW} term. The holo/closed state is further stabilized by the excess of low-frequency protein modes via the contribution ΔA_{vib}^{MBP} value, mainly of entropic nature. Quite remarkably, both ΔA_{vib}^{HW} and ΔA_{vib}^{MBP} are of the same order of the total complexation free energy, which amounts to 35 ± 1 kJ/mol [229], thus indicating that protein and HW vibrational degrees of freedom represent a crucial part in the energetics of the ligand binding processes.

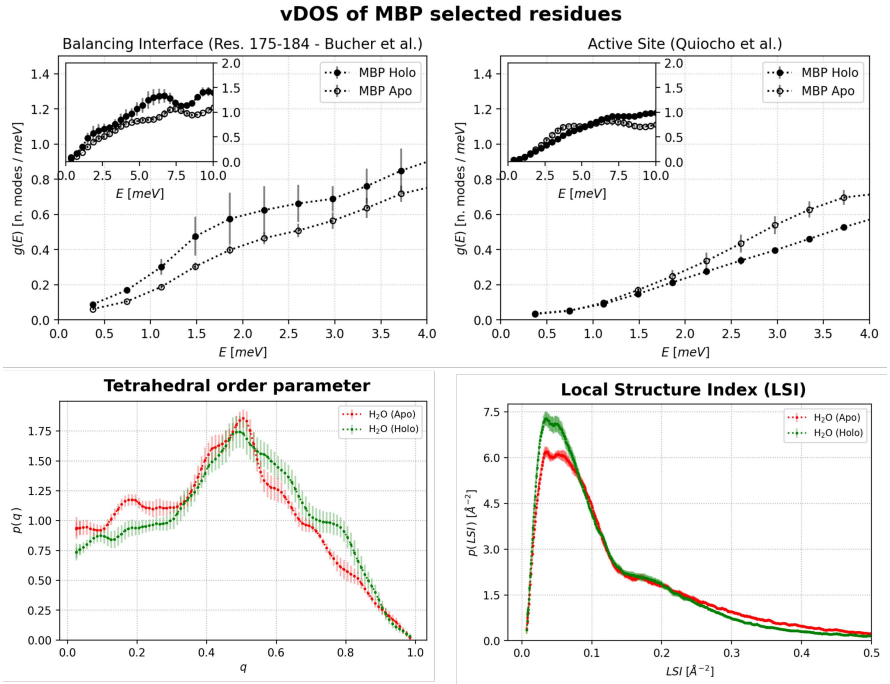

Figure 6.4: (*1st row*) Vibrational density of states (vDOS) of selected residues of the MBP. (*2nd row*) Order parameters of HW – all the Hydrogen-Bonding Donor/Acceptor, i.e the all the Oxygens, Nitrogens and Sulphurs of the systems (water + protein + maltose) which were less than 4.5 Åaway from water Oxygens were counted.

Supporting Information

6.3.1 Calculation of the MSD from the vDOS

From Lovesey eq. (4.40) we can evaluate the Debye-Waller factor, $W(\kappa)$, factor from the vDOS by [51]:

$$W(\kappa) = \frac{3\hbar}{4M}\int_0^{\omega_m} d\omega \frac{g(\omega)}{\omega}\cdot[2n(\omega)+1]\cdot\left\{|\kappa\cdot\sigma|^2\right\}_{avg} \tag{6.2}$$

$$= \frac{3\hbar}{4M}\int_0^{\omega_m} d\omega \frac{g(\omega)}{\omega}\cdot\coth\left(\frac{\hbar\omega}{2k_bT}\right)\cdot\left\{|\kappa\cdot\sigma|^2\right\}_{avg} \tag{6.3}$$

with:

$$\begin{aligned}
\kappa &= \mathbf{k}-\mathbf{k}' \text{ and } \mathbf{k}, \mathbf{k}' \text{ are the initial and final wave vector, respectively} \\
M &= \text{mass of the singol atom} \\
g(\omega) &= \text{vibrational Density of States (vDOS)} \longrightarrow g(\omega) = \tfrac{1}{3N}\sum_{j,\mathbf{q}} \delta\{\omega-\omega_j(\mathbf{q})\} \\
n(\omega) &= \text{Bose-Einstein factor} \longrightarrow g(\omega) = \left[e^{\hbar\omega/k_BT}-1\right]^{-1} \\
\sigma &= \text{polarization vector} \longrightarrow \mathbf{u}(\mathbf{l}) = \tfrac{\sigma}{\sqrt{M}}\, e^{i\,\mathbf{q}\cdot\mathbf{l}} \quad \text{and} \quad |\sigma| = 1
\end{aligned}$$

where $\mathbf{u}$ is the displacement of the atom at the $\mathbf{l}$-site of the lattice - i.e. $\mathbf{u}$ is the fluctuation of the position $\mathbf{r}(t)$ of an atom with respect to its average position $\langle \mathbf{r}(t) \rangle = \mathbf{l}$.

Then, we can consider the relation between the Debye-Waller factor and the atom displacement (see eq. (4.38) - Lovesey [51]):

$$2W(\kappa) = \left\langle (\kappa\cdot\mathbf{u})^2 \right\rangle = \kappa^2 \left\langle u_\kappa^2 \right\rangle \tag{6.4}$$

where $\left\langle u_\kappa^2 \right\rangle$ is the mean value of the projection of the atomic fluctuation into the direction of κ - i.e. if we define the direction of κ as n_κ, u_κ is defined as:

$$u_\kappa = n_\kappa \cdot \mathbf{u}$$

Therefore, combining eq. 6.2 and eq. 6.4 we get:

$$\left\langle u_\kappa^2 \right\rangle = \frac{3\hbar}{2M\kappa^2} \int_0^{\omega_m} d\omega \, \frac{g(\omega)}{\omega} \cdot \coth\left(\frac{\hbar\omega}{2k_bT}\right) \cdot \left\{ |\kappa\cdot\sigma|^2 \right\}_{avg} \tag{6.5}$$

Now, in the case of cubic symmetry (or disordered materials), we have:

$$\left\{ |\kappa\cdot\sigma|^2 \right\}_{avg} = \frac{1}{3}\kappa^2 \tag{6.6}$$

$$\left\langle u_\kappa^2 \right\rangle = \frac{1}{3}\left\langle u^2 \right\rangle \tag{6.7}$$

where $\left\langle u^2 \right\rangle$ correspond to the *Mean-Square atomic Position Fluctuations* (MSPF):

$$\left\langle u^2 \right\rangle = \left\langle \left[\mathbf{r}(t)-\mathbf{l}\right]^2 \right\rangle \equiv \left\langle \left[\mathbf{r}(t)-\langle\mathbf{r}(t)\rangle\right]^2 \right\rangle = \left\langle \mathbf{r}^2 \right\rangle - \langle\mathbf{r}\rangle^2 = \text{MSPF} \tag{6.8}$$

In this case, the Debye-Waller factor and the MSPF are:

$$2W(\kappa) = \frac{1}{3}\kappa^2 \,\text{MSPF} \tag{6.9}$$

$$\text{MSPF} = \frac{3\hbar}{2M} \int_0^{\omega_m} d\omega \, \frac{g(\omega)}{\omega} \cdot \coth\left(\frac{\hbar\omega}{2k_bT}\right) \tag{6.10}$$

Finally, takinkg into account the atomic *Mean-Square Displacements* (MSD) defined as:

$$\text{MSD}(t) = \left\langle \left[\mathbf{r}(t+t_0)-\mathbf{r}(t_0)\right]^2 \right\rangle_{t_0} \tag{6.11}$$

if we assume that the motion is confined in space and stationary such that:

$$\left\langle \mathbf{r}^2(t+t_0)\right\rangle_{t_0} = \left\langle \mathbf{r}^2(t_0)\right\rangle_{t_0} = \left\langle \mathbf{r}^2\right\rangle \tag{6.12}$$

we obtain for the MSD the following expression:

$$\mathrm{MSD}(t) = 2\left(\left\langle \mathbf{r}^2\right\rangle - \left\langle \mathbf{r}(t+t_0)\cdot\mathbf{r}(t_0)\right\rangle_{t_0}\right) \tag{6.13}$$

Now, if we take the limit of $t_0 \to \infty$, we have that $\mathbf{r}(t+t_0)$ and $\mathbf{r}(t_0)$ are uncorrelated:

$$\lim_{t_0\to\infty}\left\langle \mathbf{r}(t+t_0)\cdot\mathbf{r}(t_0)\right\rangle_{t_0} = \left\langle \mathbf{r}(t+t_0)\right\rangle_{t_0}\cdot\left\langle \mathbf{r}(t_0)\right\rangle_{t_0} = \left\langle \mathbf{r}\right\rangle^2$$

and consequently, from eq. 6.13, we obtain:

$$\mathrm{MSD}(t\to\infty) = 2\left(\left\langle \mathbf{r}^2\right\rangle - \left\langle \mathbf{r}\right\rangle^2\right) \equiv 2\,\mathrm{MSPF} \tag{6.14}$$

Therefore, combining eq. 6.9, 6.10, and 6.14 we obtain:

$$2W(\kappa) = \frac{1}{6}\kappa^2\,\mathrm{MSD}(t\to\infty) \tag{6.15}$$

$$\mathrm{MSD}(t\to\infty) = \frac{3\hbar}{M}\int_0^{\omega_m} d\omega\,\frac{g(\omega)}{\omega}\cdot\coth\left(\frac{\hbar\omega}{2k_bT}\right) \tag{6.16}$$

In conclusion, if we want to express eq. 6.10 and 6.16 in function of E instead of ω, we can make the substitutions:

$$\hbar\omega \longrightarrow E$$

$$d\omega\, g(\omega) \longrightarrow dE\, g(E)$$

such that[1]:

$$\mathrm{MSPF} = \frac{3\hbar^2}{2M}\int_0^{E_m} dE\,\frac{g(E)}{E}\cdot\coth\left(\frac{E}{2k_bT}\right) \tag{6.17}$$

$$\mathrm{MSD}(t\to\infty) = \frac{3\hbar^2}{M}\int_0^{E_m} dE\,\frac{g(E)}{E}\cdot\coth\left(\frac{E}{2k_bT}\right) \tag{6.18}$$

6.3.2 Normalization on the vDOS

The vDOS obtained from the incoherent dynamic structure factor (IDSF) is on relative scales. To re-normalize the vDOS we can use the ratio between the experimentally measured mean square MSD, which is on absolute scale, and the one calculated from the vDOS on relative scale. Practically, this means that the vDOS in absolute scale,

[1] **Note:** the equation (3) in the paper of Niessen et al. (http://dx.doi.org/10.1016/j.bpj.2016.12.049) correspond to the projection of the MSPF into an arbitrary direction, i.e.:

$$\frac{\mathrm{MSPF}}{3} = \frac{\hbar^2}{2M}\int_0^{E_m} dE\,\frac{g(E)}{E}\cdot\coth\left(\frac{E}{2k_bT}\right)$$

.

$g_{\text{abs}}(E)$, can be calculated as:

$$g_{\text{abs}}(E) = \frac{\text{MSD}_{\text{exp}}}{\text{MSD}_{\text{calc}}} \cdot g(E) \tag{6.19}$$

where MSD_{exp} is the value of the MSD measured from the elastic peack at low q (i.e. from $S(q, E = 0)$). In the gaussian approximation:

$$S(q, E = 0) = S_0 \cdot \text{e}^{-\frac{1}{6} q^2 \cdot \text{MSD}_{\text{exp}}} \longrightarrow \text{MSD}_{\text{exp}} = -\frac{1}{6} \cdot \frac{\partial \ln S(q, E = 0)}{\partial q^2} \tag{6.20}$$

However, the our IDSF data were collected at IN5 that is not optimized for this type of elastic measurements. Therefore, we decided to take, for the value of the MSD_{exp} of the apo-MBP, the one measured by Wood et al. [234], and in order to re-normalize properly also the vDOS of the holo-MBP, we considered that the differences between in the vDOS of apo-MBP Apo and holo-MBP actually are due to a shift of the vibrational modes due to the complexation [211]. This means that, if there is a softening of the complexed MBP at low energy ($\text{vDOS}_{holo} > \text{vDOS}_{apo}$), there must be a stiffening at higher frequencies ($\text{vDOS}_{holo} < \text{vDOS}_{apo}$) that compensates for the softening. Consequently, if we assume that this compensation takes place before of a certain energy $\overline{E}$, then the integrals of the vDOS between 0 and $\overline{E}$ of the Apo and the Holo systems should be equals and we can use their ratio to re-scale the vDOS.

$$g_{\text{abs}}^{(holo)}(E) = \frac{\text{MSD}_{exp}}{\text{MSD}_{calc}} \cdot \frac{\int_0^{\overline{E}} g^{(apo)}(E)\, dE}{\int_0^{15\mu\text{eV}} g^{(holo)}(E)\, dE} \cdot g^{(holo)}(E) \tag{6.21}$$

with

$$\text{MSD}_{\text{calc}} = \frac{\hbar^2}{M} \int_0^{E_m} dE\, \frac{g^{(apo)}(E)}{E} \cdot \coth\left(\frac{E}{2 k_b T}\right) \tag{6.22}$$

where we assumed that the compensation take place before 15 meV, i.e. that $\overline{E} = 15$ meV.

Finally, with the vDOS properly normalized, we were able to obtain the vDOS of the HW, for both the apo and the holo systems, simply subtracting the vDOS of measured for the MBP from one measured from MBP+H_2O.

6.3.3 vDOS from NMA

The vibrational density of states of a system is the frequency distribution of the vibrational modes, obtained by solving the eigenvalue problem for the mass-weighted Hessian matrix[235]

$$g_\alpha(E) = \frac{1}{\Delta E} \sum_{i=1}^{3N} |\mathbf{e_i}(\alpha)|^2\, \delta(E - E_i) \tag{6.23}$$

where $\mathbf{e_i}(\alpha)$ is a three-dimensional vector with the x, y, and z components of the mass-weighted displacement of atom α within the normal mode i (i.e. it is the eigen-vector obtained with the NMA), ΔE is the width of the sampling interval and $\delta(E - E_i)$ such that, it is equal to 1 when $-\frac{\Delta E}{2} \leq E - E_i < \frac{\Delta E}{2}$, otherwise it is 0.

6.3.4 Domain Opening Angle (DOA)

The MBP DOA is defined as the angle between the Centers of Mass (CoM) of the C atoms of the Nt domain (i.e. the N-terminal residues 1-108, 263-311), the CoM of the C of central t domain (i.e. the -sheet hinge residues: 109-112, 259-262) and the COM of the C of the Ct domain (i.e. the C-terminal residues 113-258, 320-370).

$$\mathbf{r_N} = \frac{\mathbf{1}}{\mathbf{n_N}} \cdot \sum_{\mathbf{i} \in \mathbf{N_t}} \mathbf{r_i} \quad ; \quad \mathbf{r}_\beta = \frac{\mathbf{1}}{\mathbf{n}_\beta} \cdot \sum_{\mathbf{i} \in \beta_\mathbf{t}} \mathbf{r_i} \quad ; \quad \mathbf{r_C} = \frac{\mathbf{1}}{\mathbf{n_C}} \cdot \sum_{\mathbf{i} \in \mathbf{C_t}} \mathbf{r_i} \tag{6.24}$$

$$\mathrm{DOA} = \left\langle \arccos \left[\frac{\left(\mathbf{r_N} - \mathbf{r}_\beta\right) \cdot \left(\mathbf{r_C} - \mathbf{r}_\beta\right)}{\left|\mathbf{r_N} - \mathbf{r}_\beta\right| \cdot \left|\mathbf{r_C} - \mathbf{r}_\beta\right|} \right] \right\rangle \tag{6.25}$$

An angle of around 160° correspond to the APO form (complexed) of the MBP, meanwhile a value of DOS of approximately 135°, correspond to the HOLO form (uncomplexed) of the MBP. As a reference for this section see (Stockner et. all, 2005).

6.3.5 Tetrahedral Order Parameter

The local tetrahedrality is a many-body property and can be defined (Chau & Hardwick, 1998) by:

$$q_i = 1 - \frac{3}{8} \cdot \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left(\cos\theta_{jik} + \frac{1}{3} \right)^2 \tag{6.26}$$

where the sum is over all angles θ_{jik} formed around a reference molecule i by its four nearest potentially hydrogen-bonding neighbors, although no account is taken of whether the neighbors are actually hydrogen bonded to molecule i. A larger value of q_i indicates a greater local tetrahedrality. In the case of water molecules close to the protein, nearest neighbors are taken to include potential hydrogen-bonding moieties of the protein, such as threonine hydroxyl groups or the amide bonds of the protein backbone. Close to the protein surface, there will undoubtedly be water molecules in a constricted environment which cannot form tetrahedral structures.

6.3.6 Local structure index (LSI)

Let $r_{i,j}$ be the radial distance between the oxygen of the molecule i and the oxygen of molecule j ordered such that $r_1 < r_2 < \ldots < r_{i,j} < \ldots < r_{i,n(i)} < r_{i,n(i)+1}$, where $n(i)$ is chosen so that $r_{i,n(i)} < 3.7\text{Å} < r_{i,n(i)+1}$. Then, the LSI parameter is defined by as (Shiratani & Sasai, 1996 and 1998):

$$\mathrm{LSI}_i = \frac{1}{n(i)} \cdot \sum_{j=1}^{n(i)} \left[\Delta_{i,j} - \overline{\Delta}_i \right] \tag{6.27}$$

where $\Delta_{i,j} = r_{i,j+1} - r_{i,j}$, $\overline{\Delta}_i$ is the average value of $\Delta_{i,j}$ over all the $n(i)+1$ molecules.

The LSI aims at measuring the extent of the gap between the first and the second hydration shells surrounding a water molecule measuring the inhomogeneity in the radial distribution within the sphere of radius 3.7 Å. A high value of LSI_i implies that

the molecule i at time t is characterized by a tetrahedral local order and a low-local density, while on the contrary, values of $\mathrm{LSI}_i \approx 0$ indicate a molecule with defective tetrahedral order and high-local density [236].

Conclusions and discussion

The work presented in this thesis aims at shedding some microscopic insights into thermal stability of bacterial cells (denaturation and cell growth). In particular, we found that the thermal death of *E. coli* is signaled by a distinctive behavior of the proteome short-time dynamics: a strong decrease of the protein global diffusion coefficient starts just below T_{CD}, from $D_G(320\text{ K}) = 1.5$ Å^2/ns down to $D_G(350\text{ K}) = 0.5$ Å^2/ns. The results from MD simulations show that this dynamic slow-down is due to the unfolding of a part of the proteome. This finding is consistent with previous experimental work where an analogous trend for the diffusive dynamics has been observed in concentrated protein solutions across the melting temperature [66, 109, 156]. Therefore, protein unfolding dominates the observed temperature dependence, and in both cases the dynamic slow-down is irreversible.

In this study, we take the analysis a crucial step forward by directly relating the diffusion coefficient to the amount of unfolded proteins in the system. The theoretical description of protein solutions containing a varying ratio of folded and unfolded proteins is challenging. These systems neither fit into the standard picture of solutions formed by globular proteins, usually modeled as rigid colloidal particles [237], nor can they be accurately described by means of simplified polymer models [238]. In particular, limitations of the colloidal model were demonstrated for highly concentrated protein solutions involving changes in protein conformation [239], and the importance of protein–protein interactions was stressed [240, 241].

Our simulations showed that the presence of minor amounts of unfolded proteins causes a substantial slow-down in the protein global diffusion. As we demonstrated, this drop was not only due to the slower diffusion of the fraction of unfolded proteins, but also to a twofold slow-down of the remaining folded proteins as a consequence of the enhanced interactions with their unfolded counterparts. Thus, unfolded proteins form a sticky macromolecular network to which folded proteins associate. This resembles the behavior recently observed in biomolecular condensates, where the interactions of folded lysozyme proteins with a macromolecular network formed by pentameric constructs of SH3 domains, and containing disordered linkers, strongly affected the condensate viscoelastic properties [166].

From the combination of QENS experiments and MD simulations, we estimated the amount of unfolded proteins in the cytoplasm at different temperatures. In the last years, there have been several attempts to connect the thermal death of bacteria to a critical amount of unfolding proteins. The aim was to understand if the death results from a collective unfolding of cell proteins [47, 45] or if it is caused by the denaturation of a subset of proteins controlling key biological functions [242, 46, 48, 243]. Here, we found that a few degrees above the cell-death temperature only a small fraction of proteins, less than 15%, are unfolded (see also Table 4.10). This result supports the

hypothesis firstly put forward by Leuenberger et al. [46] that there is no catastrophic denaturation of the proteome, but instead only an unfolding of a subset of proteins.
It is important to stress that there is no unique definition of T_{CD} which, depending on the growth conditions of bacteria and their environment, can vary by several Celsius degrees. Owing to the predicted slow increase with temperature of the unfolded protein fraction (see Fig. 4.28), this uncertainty does not affect our conclusions concerning the minor amount of unfolded proteins that are present in the cytoplasm at the cell death. Apart from the uncertainty in T_{CD}, the quantitative estimate of the unfolded fraction may also be affected by some limitations of our computational model in terms of molecular composition, which focuses exclusively on proteins as most prevalent type of macromolecules in the cytoplasm and which, moreover, is biased toward structurally well-resolved folded proteins. In addition, the model does not consider thermal adaptations of the proteome, such as evolving populations of heat-shock proteins and chaperons. However, even though the cells in the experimental samples still have a basal metabolism, they will have a reduced capability to tune the proteome composition due to lack of nutrients.

The destabilization of the *E. coli* bacteria starts already at temperatures below the T_{CD}. Temperatures near $315K$ represent already a stress condition for the bacteria - they resist to the increase of the environmental temperature with several active mechanisms, such as the change in the global protein population by increasing the number of molecular chaperones to maintain a properly folded proteome [43] and the variation of internal viscosity by regulating the synthesis of glycogen and trehalose [244]. With a contribution of such mechanisms of thermal adaptation, the average dynamical state of the cytoplasm observed at T_{CD} is still similar to the dynamical state of the cytoplasm near the temperature of optimal growth rate. This leads us to argue that the average dynamics should not play a predominant role in thermal death. However, as our simulations have shown, a low amount of unfolded proteins can trigger an important slow-down in the diffusion of the surrounding macromolecules. Therefore, the increase of local viscosity and the associated dramatically reduced protein diffusion caused by unfolding, may threaten the viability of the cell by affecting localized physiological processes. A pertinent example at molecular scale is the dynamics and substrate channeling in enzymatic assembling [**?**], but also at larger scale the viscoelastic response of the cytoplasm associated to organelles localization [245]. Morever, we have shown that a very good reproduction of *E. coli* growth-rate can be obtained when combining together the fraction of unfolded proteins with the temperature-dependent diffusion within a simple reaction-diffusion model. This supports the idea that at least a part of the cell metabolism is modulated by diffusion. Finally, our findings are consistent with the idea that intrinsically disordered proteins may induce as well a mobility slow-down on the local environment, as it is probed in membrane-less organelles [246].

The approach hereby presented can be extended -and possibly complemented by single-molecule techniques- to investigate the relationship between the dynamics and the proteome unfolding in extremophiles resisting either to cold or hot environments. Further, attention could be enlarged to the peculiarity of the proteins' dynamical response to stress in the functioning of specific networks of interaction like in the unfolding protein response cascade.

In this thesis we presented also two other studies: "*Differences between Ca^{2+} reach and depleted α-La investigated by MD simulations and NS experiments*", and "*Role*

of low-frequency vibrational dynamics of protein hydration water for ligand binding"dealing with the effects of protein-ligand complexation on protein dynamics, and in which neutron scattering techniques and molecular dynamics simulations are always coupled.

In the first one, we show how data from EINS and MD can be combined to probe the dynamics in both protein-depleted and protein calcium rich systems. Two models were exploited to extract the hydrogen atoms MSD from the EINS data, measured at three different levels of hydration and with resolutions to explore different time scales (picosecond to nanosecond, i.e. a time scale range accessible to simulations). In parallel, systems mimicking protein powders at different hydration levels were simulated at the atomistic level, and MSD estimated with different methods, allowing a robust comparison with neutron scattering data.

In the last study presented in this thesis, we describe the behavior of hydration water in a system containing MBP, a protein undergoing a large-scale conformational change (open-to-close) in ligand binding. In this case, INS experiments were performed to probe the low frequency vibration modes of protein and water in both states. Results point out that the protein and its hydration water have an opposite behavior in closing. An interesting thermodynamics scenario, which however deserves further studies, is proposed where the protein-ligand interactions allow the protein to overcome the open-to-close free energy barrier, which is contributed by the hydration water (in a way still to be clarified).

Bibliography

[1] KA Dill and L Agozzino. Driving forces in the origins of life. *Open biology*, 11(2):200324, 2021.

[2] Ron Milo and Rob Phillips. *Cell biology by the numbers*. Garland Science, 2015.

[3] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry*. Macmillan, 7th edition, 2017.

[4] Bruce Alberts. *Molecular biology of the cell*. W. W. Norton & Company, 2022.

[5] Sanjeev Kumar Chandrayan, Satya Prakash, Shubbir Ahmed, and Purnananda Guptasarma. Hyperthermophile protein behavior: partially-structured conformations of Pyrococcus furiosus rubredoxin monomers generated through forced cold-denaturation and refolding. *PloS one*, 9(3):e80014, 2014.

[6] Nicolas Galtier and JR Lobry. Relationships between genomic G+ C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, 44(6):632–636, 1997.

[7] Héctor Musto, Hugo Naya, Alejandro Zavala, Héctor Romero, Fernando Alvarez-Valín, and Giorgio Bernardi. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. *Biochemical and biophysical research communications*, 347(1):1–3, 2006.

[8] Hiroshi Nakashima, Satoshi Fukuchi, and Ken Nishikawa. Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. *Journal of biochemistry*, 133(4):507–513, 2003.

[9] Naoto Ohtani, Masaru Tomita, and Mitsuhiro Itaya. An extreme thermophile, Thermus thermophilus, is a polyploid bacterium. *Journal of bacteriology*, 192(20):5499–5505, 2010.

[10] Michel Hébraud and Patrick Potier. Cold shock response and low temperature adaptation in psychrotrophic bacteria. *Journal of molecular microbiology and biotechnology*, 1(2):211–219, 1999.

[11] S Shivaji and Jogadhenu SS Prakash. How do bacteria sense and respond to low temperature? *Archives of microbiology*, 192(2):85–95, 2010.

[12] Sandra L Wilson and Virginia K Walker. Selection of low-temperature resistance in bacteria and potential applications. *Environmental technology*, 31(8-9):943–956, 2010.

[13] Marco Ventura, Carlos Canchaya, Ziding Zhang, Valentina Bernini, Gerald F Fitzgerald, and Douwe Van Sinderen. How high G+ C Gram-positive bacteria and in particular bifidobacteria cope with heat stress: protein players and regulators. *FEMS microbiology reviews*, 30(5):734–759, 2006.

[14] A. D. Russel. Lethal Effects of Heat on Bacterial Physiology and Structure. *Science Progress*, 86:115–137, 2003.

[15] P Teixeira, H Castro, Cs Mohácsi-Farkas, and R Kirby. Identification of sites of injury in Lactobacillus bulgaricus during heat stress. *Journal of Applied Microbiology*, 83(2):219–226, 1997.

[16] SC Stringer, SM George, and MW Peck. Thermal inactivation of Escherichia coli O157: H7. *Journal of Applied Microbiology*, 88(S1):79S–89S, 2000.

[17] Beverley J Hitchener and Aubrey F Egan. Outer-membrane damage in sublethally heated Escherichia coli K-12. *Canadian Journal of Microbiology*, 23(3):311–318, 1977.

[18] N Katsui, T Tsuchido, R Hiramatsu, S Fujikawa, M Takano, and I Shibasaki. Heat-induced blebbing and vesiculation of the outer membrane of Escherichia coli. *Journal of bacteriology*, 151(3):1523–1531, 1982.

[19] T Tsuchido, N Katsui, A Takeuchi, M Takano, and I Shibasaki. Destruction of the outer membrane permeability barrier of Escherichia coli by heat treatment. *Applied and Environmental microbiology*, 50(2):298–303, 1985.

[20] Tetsuaki Tsuchido, Isao Aoki, and Mitsuo Takano. Interaction of the fluorescent dye lN-phenylnaphthylamine with Escherichia coli cells during heat stress and recovery from heat stress. *Microbiology*, 135(7):1941–1947, 1989.

[21] AD Russell. Potential sites of damage in micro-organisms exposed to chemical or physical agents. In *Society for Applied Bacteriology symposium series*, number 12, pages 1–18, 1984.

[22] AD Russell and Diann Harries. Some aspects of thermal injury in Escherichia coli. *Applied Microbiology*, 15(2):407–410, 1967.

[23] Ao Do Russell and Diann Harries. Damage to Escherichia coli on exposure to moist heat. *Applied Microbiology*, 16(9):1394–1399, 1968.

[24] MC Allwood and AD Russell. Mechanism of thermal injury in Staphylococcus aureus: I. relationship between viability and leakage. *Applied microbiology*, 15(6):1266–1269, 1967.

[25] MC Allwood and AD Russell. Mechanisms of thermal injury in nonsporulating bacteria. *Advances in Applied Microbiology*, 12:89–119, 1970.

[26] RI Tomlins and Z Ji Ordal. Thermal injury and inactivation in vegetative bacteria. *Inhibition and inactivation of vegetative microbes*, 5:153–191, 1976.

[27] LR Beuchat. Injury and repair of gram-negative bacteria, with special consideration of the involvement of the cytoplasmic membrane. *Advances in applied microbiology*, 23:219–243, 1978.

[28] A Denver Russell. Microbial susceptibility and resistance to chemical and physical agents. *Topley & Wilson's Microbiology and Microbial Infections*, 2010.

[29] James M Jay, Martin J Loessner, and David A Golden. Taxonomy, role, and significance of microorganisms in foods. *Modern food microbiology*, pages 13–37, 2005.

[30] Bade Tonyali, Austin McDaniel, Valentina Trinetta, and Umut Yucel. Evaluation of heating effects on the morphology and membrane structure of Escherichia coli using electron paramagnetic resonance spectroscopy. *Biophysical Chemistry*, 252:106191, 2019.

[31] John J Iandolo and Z John Ordal. Repair of thermal injury of Staphylococcus aureus. *Journal of Bacteriology*, 91(1):134–142, 1966.

[32] MC Allwood and AD Russell. Thermally induced ribonucleic acid degradation and leakage of substances from the metabolic pool in Staphylococcus aureus. *Journal of Bacteriology*, 95(2):345–349, 1968.

[33] Richard I Tomlins and Z John Ordal. Precursor ribosomal ribonucleic acid and ribosome accumulation in vivo during the recovery of Salmonella typhimurium from thermal injury. *Journal of Bacteriology*, 107(1):134–142, 1971.

[34] A Hurst and A Hughes. Stability of ribosomes of Staphylococcus aureus S6 sublethally heated in different buffers. *Journal of Bacteriology*, 133(2):564–568, 1978.

[35] A Hurst. Revival of vegetative bacteria after sublethal heating. In *Society for Applied Bacteriology symposium series*, number 12, pages 77–103, 1984.

[36] NE Welker. Microbial endurance and resistance to heat stress, 1976.

[37] BA Bridges, MJ Ashwood-Smith, and RJ Munson. Correlation of bacterial sensitivities to ionizing radiation and mild heating. *Microbiology*, 58(1):115–124, 1969.

[38] C Pauling and LA Beck. Role of DNA ligase in the repair of single strand breaks induced in DNA by mild heating of Escherichia coli. *Microbiology*, 87(1):181–184, 1975.

[39] JOHN R Battista, ASHLEE M Earl, and OWEN White. The stress responses of Deinococcus radiodurans. *Bacterial Stress Responses*, pages 383–391, 2000.

[40] DIANN HARRIES and AD Russell. A note on some changes in the physical properties of Escherichia coli after heat treatment. *Journal of Pharmacy and Pharmacology*, 19(11):740–743, 1967.

[41] Barnett Rosenberg, Gabor Kemeny, Robert C Switzer, and Thomas C Hamilton. Quantitative evidence for protein denaturation as the cause of thermal death. *Nature*, 232(5311):471–473, 1971.

[42] Costa Georgopoulos. Properties of heat shock proteins of Escherichia coli and autoregulation of the heat shock response. *The Biology of Heat Shock Proeins and Molecular Chaperones*, pages 209–249, 1994.

[43] Ke Chen, Ye Gao, Nathan Mih, Edward J O'Brien, Laurence Yang, and Bernhard O Palsson. Thermosensitivity of growth is determined by chaperone-mediated proteome reallocation. *Proceedings of the National Academy of Sciences*, 114(43):11548–11553, 2017.

[44] Konstantin B Zeldovich, Peiqiu Chen, and Eugene I Shakhnovich. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proceedings of the National Academy of Sciences*, 104(41):16152–16157, 2007.

[45] Ken A Dill, Kingshuk Ghosh, and Jeremy D Schmit. Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences*, 108(44):17876–17882, 2011.

[46] Pascal Leuenberger, Stefan Ganscha, Abdullah Kahraman, Valentina Cappelletti, Paul J Boersema, Christian von Mering, Manfred Claassen, and Paola Picotti. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science*, 355(6327), 2017.

[47] Kingshuk Ghosh and Ken Dill. Cellular proteomes have broad distributions of protein stability. *Biophysical journal*, 99(12):3996–4002, 2010.

[48] A. Mateus, J. Bobonis, N. Kurzawa, F. Stein, D. Helm, J. Hevler, A. Typas, and M. M. Savitski. Thermal Proteome Profiling in Bacteria: Probing Protein State in vivo. *Molecular systems biology*, 14:e8242 —, 2018.

[49] Paul E Schavemaker, Arnold J Boersma, and Bert Poolman. How important is protein diffusion in prokaryotes? *Frontiers in molecular biosciences*, 5:93, 2018.

[50] R John Ellis. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Current opinion in structural biology*, 11(1):114–119, 2001.

[51] Stephen W Lovesey. Theory of neutron scattering from condensed matter. 1984.

[52] Marc Bée. *Quasielastic neutron scattering*. United Kingdom: Adam Hilger, 1988.

[53] Gordon Leslie Squires. *Introduction to the theory of thermal neutron scattering*. Courier Corporation, 1996.

[54] Andrew T Boothroyd. *Principles of Neutron Scattering from Condensed Matter*. Oxford University Press, 2020.

[55] Marco Grimaldo, Felix Roosen-Runge, Fajun Zhang, Frank Schreiber, and Tilo Seydel. Dynamics of proteins in solution. *Quarterly Reviews of Biophysics*, 52, 2019.

[56] W Doster, M Diehl, R Gebhardt, RE Lechner, and J Pieper. TOF-elastic resolution spectroscopy: time domain analysis of weakly scattering (biological) samples. *Chemical physics*, 292(2-3):487–494, 2003.

[57] Reiner Zorn. On the evaluation of neutron scattering elastic scan data. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 603(3):439–445, 2009.

[58] Atsushi Tokuhisa, Yasumasa Joti, Hiroshi Nakagawa, Akio Kitao, and Mikio Kataoka. Non-Gaussian behavior of elastic incoherent neutron scattering profiles of proteins studied by molecular dynamics simulation. *Physical Review E*, 75(4):041912, 2007.

[59] Derya Vural, Liang Hong, Jeremy C Smith, and Henry R Glyde. Motional displacements in proteins: The origin of wave-vector-dependent values. *Physical Review E*, 91(5):052705, 2015.

[60] D Zeller, MTF Telling, M Zamponi, V Garcia Sakai, and Judith Peters. Analysis of elastic incoherent neutron scattering data beyond the Gaussian approximation. *The Journal of chemical physics*, 149(23):234908, 2018.

[61] Zheng Yi, Yinglong Miao, Jerome Baudry, Nitin Jain, and Jeremy C Smith. Derivation of mean-square displacements for protein dynamics from elastic incoherent neutron scattering. *The Journal of Physical Chemistry B*, 116(16):5028–5036, 2012.

[62] Hiroshi Nakagawa, Hironari Kamikubo, Itaru Tsukushi, Toshiji Kanaya, and Mikio Kataoka. Protein dynamical heterogeneity derived from neutron incoherent elastic scattering. *Journal of the Physical Society of Japan*, 73(2):491–495, 2004.

[63] Judith Peters and Gerald R Kneller. Motional heterogeneity in human acetylcholinesterase revealed by a non-Gaussian model for elastic incoherent neutron scattering. *The Journal of chemical physics*, 139(16):10B620_1, 2013.

[64] Ursula Lehnert, Valérie Réat, Martin Weik, Giuseppe Zaccai, and Claude Pfister. Thermal Motions in Bacteriorhodopsin at Different Hydration Levels Studied by Neutron Scattering: Correlation with Kinetics and Light-Induced Conformational Changes. *Biophysical journal*, 75(4):1945–1952, 1998.

[65] J-M Zanotti, M-C Bellissent-Funel, and S-H Chen. Experimental evidence of a liquid-liquid transition in interfacial water. *EPL (Europhysics Letters)*, 71(1):91, 2005.

[66] Marcus Hennig, Felix Roosen-Runge, Fajun Zhang, Stefan Zorn, Maximilian WA Skoda, Robert MJ Jacobs, Tilo Seydel, and Frank Schreiber. Dynamics of highly concentrated protein solutions around the denaturing transition. *Soft Matter*, 8(5):1628–1633, 2012.

[67] AZMS Rahman, KS Singwi, and A Sjölander. Theory of slow neutron scattering by liquids. I. *Physical Review*, 126(3):986, 1962.

[68] Gerald R Kneller. Quasielastic Neutron Scattering. *Centre de Biophysique Molèculaire, CNRS, France*, 2004.

[69] Javier Pérez, Jean-Marc Zanotti, and Dominique Durand. Evolution of the internal dynamics of two globular proteins from dry powder to solution. *Biophysical journal*, 77(1):454–469, 1999.

[70] Felix Roosen-Runge, Marcus Hennig, Fajun Zhang, Robert MJ Jacobs, Michael Sztucki, Helmut Schober, Tilo Seydel, and Frank Schreiber. Protein self-diffusion in crowded solutions. *Proceedings of the National Academy of Sciences*, 108(29):11815–11820, 2011.

[71] DJ Bicout. Incoherent neutron scattering functions for combined dynamics. In *ILL Millennium Symposium Preface to the proceedings*, page 60, 2001.

[72] Owen Arnold, Jean-Christophe Bilheux, JM Borreguero, Alex Buts, Stuart I Campbell, L Chapon, Mathieu Doucet, N Draper, R Ferraz Leal, MA Gigg, et al. Mantid—Data analysis and visualization package for neutron scattering and μ SR experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 764:156–166, 2014.

[73] HH Paalman and CJ Pings. Numerical evaluation of X-ray absorption factors for cylindrical samples and annular sample cells. *Journal of Applied Physics*, 33(8):2635–2639, 1962.

[74] M Kamal, SS Malik, and D Rorer. Neutron incoherent elastic scattering study of the temperature dependence of the Debye-Waller exponent in vanadium. *Physical Review B*, 18(4):1609, 1978.

[75] Maria Kalimeri. *Are thermophilic proteins rigid or flexible? An in silico investigation*. PhD thesis, Paris Diderot University, 2014.

[76] Andreas Kukol. NAMD-VMD tutorial. *ResearchGate*, 2016.

[77] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, first edition, 1987.

[78] Daniele Di Bari. Molecular Dynamics Simulation - A brief introduction of selected concepts. Also available as https://github.com/DanieleDiBari/Selected_concepts_of_Molecular_Dynamics_Simulation, 2018.

[79] Paraskevi Gkeka and Zoe Cournia. Molecular Dynamics simulations of lysozyme in water. MSc in Bioinformatics and Medical Informatics, 2015/2016.

[80] Eni. Generalic. Lennard-Jones potential, 2017. Croatian-English Chemistry Dictionary & Glossary. Also available on https://glossary.periodni.com.

[81] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable Molecular Dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[82] Loup Verlet. Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.

[83] Donald Voet, Judith G. Voet, and Charlotte W. Pratt. Fundamentals of biochemistry: life at the molecular level. 2016.

[84] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical reviews*, 116(14):7898–7936, 2016.

[85] Fabio Sterpone, Simone Melchionna, Pierre Tuffery, Samuela Pasquali, Normand Mousseau, Tristan Cragnolini, Yassmine Chebaro, Jean-Francois St-Pierre, Maria Kalimeri, Alessandro Barducci, et al. The OPEP protein model: from single molecules, amyloid formation, crowding and hydrodynamics to DNA/RNA systems. *Chemical Society reviews*, 43(13):4871–4893, 2014.

[86] Mara Chiricotto, Simone Melchionna, Philippe Derreumaux, and Fabio Sterpone. Hydrodynamic effects on β-amyloid (16-22) peptide aggregation. *The Journal of chemical physics*, 145(3):035102, 2016.

[87] Shiyi Chen and Gary D Doolen. Lattice boltzmann method for fluid flows. *Annual review of fluid mechanics*, 30(1):329–364, 1998.

[88] Sauro Succi. *The lattice Boltzmann equation: for fluid dynamics and beyond*. Oxford university press, 2001.

[89] Zhaoli Guo, Chuguang Zheng, and Baochang Shi. Discrete lattice effects on the forcing term in the lattice boltzmann method. *Physical review E*, 65(4):046308, 2002.

[90] Simone Melchionna. Design of quasisymplectic propagators for langevin dynamics. *The Journal of chemical physics*, 127(4):044108, 2007.

[91] Patrick Ahlrichs and Burkhard Dünweg. Simulation of a single polymer chain in solution by combining lattice boltzmann and molecular dynamics. *The Journal of chemical physics*, 111(17):8225–8239, 1999.

[92] Jürgen Horbach and Sauro Succi. Lattice boltzmann versus molecular dynamics simulation of nanoscale hydrodynamic flows. *Physical review letters*, 96(22):224503, 2006.

[93] B. M. Mackey, C.A. Miles, S.E. Parsons, and D.A. Seymour. Thermal denaturation of whole cells and cell components of Escherichia coli examined by differential scanning calorimetry. *J.Gen. Microbiol.*, 137:2361–2374, 1991.

[94] James P Kennett and LD Stott. Abrupt deep-sea warming, palaeoceanographic changes and benthic extinctions at the end of the Palaeocene. *Nature*, 353(6341):225–229, 1991.

[95] Linda C Ivany, William P Patterson, and Kyger C Lohmann. Cooler winters as a possible cause of mass extinctions at the Eocene/Oligocene boundary. *Nature*, 407(6806):887–890, 2000.

[96] Andrew P Allen, James F Gillooly, Van M Savage, and James H Brown. Kinetic effects of temperature on rates of genetic divergence and speciation. *Proceedings of the National Academy of Sciences*, 103(24):9130–9135, 2006.

[97] Serita D Frey, Juhwan Lee, Jerry M Melillo, and Johan Six. The temperature response of soil microbial efficiency and its feedback to climate. *Nature Climate Change*, 3(4):395–398, 2013.

[98] Igor N Berezovsky and Eugene I Shakhnovich. Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences*, 102(36):12742–12747, 2005.

[99] Donald S Coffey, Robert H Getzenberg, and Theodore L DeWeese. Hyperthermic biology and cancer therapies: a hypothesis for the "Lance Armstrong effect". *JAMA*, 296(4):445–448, 2006.

[100] J. Peters, D. Di Bari, M. Guiral, M. T. Giudici Orticoni, T. Seydel, F. Sterpone, and A. Paciaroni. Global proteome dynamics as a proxy for cellular thermal stability. *Institut Laue-Langevin (ILL)*, 2020.

[101] D. Di Bari, M. Guiral, M. T. Giudici Orticoni, T. Seydel, F. Sterpone, A. Paciaroni, and J. Peters. Global proteome dynamics as a proxy for cellular thermal stability. *Institut Laue-Langevin (ILL)*, 2021.

[102] Bernhard Frick, Eugene Mamontov, Lambert Van Eijck, and Tilo Seydel. Recent backscattering instrument developments at the ILL and SNS. *Zeitschrift für physikalische Chemie*, 224(1-2):33–60, 2010.

[103] M Jasnin, M Moulin, M Haertlein, G Zaccai, and M Tehei. In vivo measurement of internal and global macromolecular motions in Escherichia coli. *Biophysical journal*, 95(2):857–864, 2008.

[104] Frederick Carl Neidhardt, John L Ingraham, and Moselio Schaechter. *Physiology of the bacterial cell; a molecular approach*. Number 589.901 N397. Sinauer associates, 1990.

[105] Marco Grimaldo, Felix Roosen-Runge, Niina Jalarvo, Michaela Zamponi, Fabio Zanini, Marcus Hennig, Fajun Zhang, Frank Schreiber, and Tilo Seydel. High-resolution neutron spectroscopy on protein solution samples. In *EPJ Web of Conferences*, volume 83, page 02005. EDP Sciences, 2015.

[106] George H Vineyard. Scattering of slow neutrons by a liquid. *Physical Review*, 110(5):999, 1958.

[107] U Wanderlingh, G D'Angelo, C Branca, V Conti Nibali, A Trimarchi, S Rifici, D Finocchiaro, C Crupi, J Ollivier, and HD Middendorf. Multi-component modeling of quasielastic neutron scattering from phospholipid membranes. *The Journal of chemical physics*, 140(17):05B602_1, 2014.

[108] L Bosio, J Teixeira, and M-C Bellissent-Funel. Enhanced density fluctuations in water analyzed by neutron scattering. *Physical Review A*, 39(12):6612, 1989.

[109] Marco Grimaldo, Felix Roosen-Runge, Marcus Hennig, Fabio Zanini, Fajun Zhang, Niina Jalarvo, Michaela Zamponi, Frank Schreiber, and Tilo Seydel. Hierarchical molecular dynamics of bovine serum albumin in concentrated aqueous solution below and above thermal denaturation. *Physical Chemistry Chemical Physics*, 17(6):4645–4655, 2015.

[110] Nicolas Martinez, Gregoire Michoud, A Cario, J Ollivier, Bruno Franzetti, Mohamed Jebbar, P Oger, and Judith Peters. High protein flexibility and reduced hydration water dynamics are key pressure adaptive strategies in prokaryotes. *Scientific reports*, 6(1):1–11, 2016.

[111] Eugene Mamontov. Microscopic diffusion processes measured in living planarians. *Scientific reports*, 8(1):1–8, 2018.

[112] Marco Grimaldo, Felix Roosen-Runge, Fajun Zhang, Tilo Seydel, and Frank Schreiber. Diffusion and dynamics of γ-globulin in crowded aqueous solutions. *The Journal of Physical Chemistry B*, 118(25):7203–7209, 2014.

[113] Varley F Sears. Neutron scattering lengths and cross sections. *Neutron news*, 3(3):26–37, 1992.

[114] Sean R McGuffee and Adrian H Elcock. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS computational biology*, 6(3):e1000694, 2010.

[115] Andrew J Link, Keith Robison, and George M Church. Comparing the predicted and observed properties of proteins encoded in the genome of Escherichia coli K-12. *Electrophoresis*, 18(8):1259–1313, 1997.

[116] Jens Danielsson, Wael Awad, Kadhirvel Saraboji, Martin Kurnik, Lisa Lang, Lina Leinartaitė, Stefan L Marklund, Derek T Logan, and Mikael Oliveberg. Global structural motions from the strain of a single hydrogen bond. *Proceedings of the National Academy of Sciences*, 110(10):3829–3834, 2013.

[117] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide protein data bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.

[118] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303, 2018.

[119] Andrej Šali and Tom L Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815, 1993.

[120] Christopher J Williams, Jeffrey J Headd, Nigel W Moriarty, Michael G Prisant, Lizbeth L Videau, Lindsay N Deis, Vishal Verma, Daniel A Keedy, Bradley J Hintze, Vincent B Chen, et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science*, 27(1):293–315, 2018.

[121] Evgeny Krissinel and Kim Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797, 2007.

[122] Ove Wiborg, Carsten Andersen, Charlotte R Knudsen, Brian FC Clark, and Jens Nyborg. Mapping Escherichia coli elongation factor Tu residues involved in binding of aminoacyl-tRNA. *Journal of Biological Chemistry*, 271(34):20406–20411, 1996.

[123] Ingrid M Keseler, Julio Collado-Vides, Alberto Santos-Zavaleta, Martin Peralta-Gil, Socorro Gama-Castro, Luis Muñiz-Rascado, César Bonavides-Martinez, Suzanne Paley, Markus Krummenacker, Tomer Altman, et al. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic acids research*, 39(suppl_1):D583–D590, 2010.

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

[125] Paul Robustelli, Stefano Piana, and David E Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proceedings of the National Academy of Sciences*, 115(21):E4758–E4766, 2018.

[126] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L De Groot, Helmut Grubmüller, and Alexander D MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature methods*, 14(1):71–73, 2017.

[127] Herman JC Berendsen, JPM van Postma, Wilfred F van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[128] Leandro Martínez, Ricardo Andrade, Ernesto G Birgin, and José Mario Martínez. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry*, 30(13):2157–2164, 2009.

[129] Fabio Sterpone, Philippe Derreumaux, and Simone Melchionna. Protein simulations in fluids: Coupling the OPEP coarse-grained force field with hydrodynamics. *Journal of chemical theory and computation*, 11(4):1843–1853, 2015.

[130] Massimo Bernaschi, Simone Melchionna, Sauro Succi, Maria Fyta, Efthimios Kaxiras, and Joy K Sircar. MUPHY: A parallel MUlti PHYsics/scale code for high performance bio-fluidic simulations. *Computer Physics Communications*, 180(9):1495–1502, 2009.

[131] Mara Chiricotto, Simone Melchionna, Philippe Derreumaux, and Fabio Sterpone. Multiscale aggregation of the amyloid Aβ16–22 peptide: from disordered coagulation and lateral branching to Amorphous prefibrils. *The journal of physical chemistry letters*, 10(7):1594–1599, 2019.

[132] Fabio Sterpone, Philippe Derreumaux, and Simone Melchionna. Molecular mechanism of protein unfolding under shear: A lattice boltzmann molecular dynamics study. *The Journal of Physical Chemistry B*, 122(5):1573–1579, 2018.

[133] Astrid F Brandner, Stepan Timr, Simone Melchionna, Philippe Derreumaux, Marc Baaden, and Fabio Sterpone. Modelling lipid systems in fluid with Lattice Boltzmann Molecular Dynamics simulations and hydrodynamics. *Scientific reports*, 9, 2019.

[134] Stepan Timr and Fabio Sterpone. Stabilizing or Destabilizing: Simulations of Chymotrypsin Inhibitor 2 under Crowding Reveal Existence of a Crossover Temperature. *The Journal of Physical Chemistry Letters*, 12(6):1741–1746, 2021.

[135] Roberto Benzi, Sauro Succi, and Massimo Vergassola. The lattice Boltzmann equation: theory and applications. *Physics Reports*, 222(3):145–197, 1992.

[136] Alvaro Ortega, D Amorós, and J García De La Torre. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic-and residue-level models. *Biophysical journal*, 101(4):892–898, 2011.

[137] Alvaro Ortega, D Amorós, and J García De La Torre. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic-and residue-level models. *Biophysical journal*, 101(4):892–898, 2011.

[138] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[139] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.

[140] Roger Williams Hockney, SP Goel, and JW Eastwood. Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.

[141] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. LINCS: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[142] Shuichi Miyamoto and Peter A Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry*, 13(8):952–962, 1992.

[143] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1):014101, 2007.

[144] Michele Parrinello and Aneesur Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, 52(12):7182–7190, 1981.

[145] Gavin M Seddon and Robert P Bywater. Accelerated simulation of unfolding and refolding of a large single chain globular protein. *Open Biology*, 2(7):120087, 2012.

[146] Wouter G Touw, Coos Baakman, Jon Black, Tim AH Te Beek, Elmar Krieger, Robbie P Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368, 2015.

[147] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[148] Vance Wong and David A. Case. Evaluating rotational diffusion from protein MD simulations. *Journal of Physical Chemistry B*, 112(19):6013–6024, 2008.

[149] Grzegorz Nawrocki, Po-hung Wang, Isseki Yu, Yuji Sugita, and Michael Feig. Slow-Down in Diffusion in Crowded Protein Solutions Correlates with Transient Cluster Formation. *The Journal of Physical Chemistry B*, 121(49):11072–11084, 2017.

[150] In Chul Yeh and Gerhard Hummer. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *Journal of Physical Chemistry B*, 108(40):15873–15879, 2004.

[151] Max Linke, Jürgen Köfinger, and Gerhard Hummer. Rotational Diffusion Depends on Box Size in Molecular Dynamics Simulations. *Journal of Physical Chemistry Letters*, 9(11):2874–2878, 2018.

[152] Sören von Bülow, Marc Siggel, Max Linke, and Gerhard Hummer. Dynamic cluster formation determines viscosity and diffusion in dense protein solutions. *Proceedings of the National Academy of Sciences*, 116(20):9843–9852, 2019.

[153] K. Monkos. Viscosity of bovine serum albumin aqueous solutions as a function of temperature and concentration. *International Journal of Biological Macromolecules*, 18(1-2):61–68, 1996.

[154] Divina B Anunciado, Vyncent P Nyugen, Gregory B Hurst, Mitchel J Doktycz, Volker Urban, Paul Langan, Eugene Mamontov, and Hugh O'Neill. In vivo protein dynamics on the nanometer length scale and nanosecond time scale. *The journal of physical chemistry letters*, 8(8):1899–1904, 2017.

[155] KS Singwi and Alf Sjölander. Diffusive motions in water and cold neutron scattering. *Physical Review*, 119(3):863, 1960.

[156] Olga Matsarskaia, Lena Bühl, Christian Beck, Marco Grimaldo, Ralf Schweins, Fajun Zhang, Tilo Seydel, Frank Schreiber, and Felix Roosen-Runge. Evolution of the structure and dynamics of bovine serum albumin induced by thermal denaturation. *Physical Chemistry Chemical Physics*, 22(33):18507–18517, 2020.

[157] Tadashi Ando and Jeffrey Skolnick. Crowding and hydrodynamic interactions likely dominate in vivo macromolecular motion. *Proceedings of the National Academy of Sciences*, 107(43):18457–18462, 2010.

[158] Isseki Yu, Takaharu Mori, Tadashi Ando, Ryuhei Harada, Jaewoon Jung, Yuji Sugita, and Michael Feig. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *Elife*, 5:e19274, 2016.

[159] Stepan Timr, David Gnutt, Simon Ebbinghaus, and Fabio Sterpone. The Unfolding Journey of Superoxide Dismutase 1 Barrels under Crowding: Atomistic Simulations Shed Light on Intermediate States and Their Interactions with Crowders. *J. Phys. Chem. Lett.*, 11:4206–421., 2020.

[160] M. Chiriccotto, S. Melchionna, P. Derreumaux, and F. Sterpone. Multiscale Aggregation of the Amyloid AB16–22 Peptide: From Disordered Coagulation and Lateral Branching to Amorphous Prefibrils. *J. Phys. Chem. Lett.*, 10:1594–1599, 2019.

[161] T. Kalwarczyk, M. Tabaka, and R. Holyst. Biologistics—Diffusion Coefficients for Complete Proteome of Escherichia coli. *Bioinformatics*, 28:2971–2978, 2012.

[162] ML Anson and AE Mirsky. The effect of denaturation on the viscosity of protein systems. *The Journal of general physiology*, 15(3):341–350, 1932.

[163] Sungyoung Choi and Je-Kyun Park. Microfluidic rheometer for characterization of protein unfolding and aggregation in microflows. *Small*, 6(12):1306–1310, 2010.

[164] Romina Muñoz, Felipe Aguilar-Sandoval, Ludovic Bellon, and Francisco Melo. Detecting protein folding by thermal fluctuations of microcantilevers. *PloS one*, 12(12):e0189979, 2017.

[165] William J Galush, Lan N Le, and Jamie MR Moore. Viscosity behavior of high-concentration protein mixtures. *Journal of pharmaceutical sciences*, 101(3):1012–1020, 2012.

[166] Archishman Ghosh, Divya Kota, and Huan-Xiang Zhou. Shear relaxation governs fusion dynamics of biomolecular condensates. *Nature communications*, 12(1):1–10, 2021.

[167] Marco Grimaldo, Hender Lopez, Christian Beck, Felix Roosen-Runge, Martine Moulin, Juliette M Devos, Valerie Laux, Michael Härtlein, Stefano Da Vela, Ralf Schweins, et al. Protein short-time diffusion in a naturally crowded environment. *The journal of physical chemistry letters*, 10(8):1709–1715, 2019.

[168] Maksym Golub, Nicolas Martinez, Grégoire Michoud, Jacques Ollivier, Mohamed Jebbar, Philippe Oger, and Judith Peters. The effect of crowding on protein stability, rigidity, and high pressure sensitivity in whole cells. *Langmuir*, 34(35):10419–10425, 2018.

[169] Katsutoshi Nitta and Shintaro Sugai. The evolution of lysozyme and α-lactalbumin. *European Journal of Biochemistry*, 182(1):111–118, 1989.

[170] Marcus Trapp, Moeava Tehei, Marie Trovaslet, Florian Nachon, Nicolas Martinez, Marek M Koza, Martin Weik, Patrick Masson, and Judith Peters. Correlation of the dynamics of native human acetylcholinesterase and its inhibited huperzine A counterpart from sub-picoseconds to nanoseconds. *Journal of the Royal Society Interface*, 11(97):20140372, 2014.

[171] Judith Peters, Nicolas Martinez, Marie Trovaslet, Kévin Scannapieco, Michael Marek Koza, Patrick Masson, and Florian Nachon. Dynamics of human acetylcholinesterase bound to non-covalent and covalent inhibitors shedding light on changes to the water network structure. *Physical Chemistry Chemical Physics*, 18(18):12992–13001, 2016.

[172] Melek Saouessi, Judith Peters, and Gerald R Kneller. Asymptotic analysis of quasielastic neutron scattering data from human acetylcholinesterase reveals subtle dynamical changes upon ligand binding. *The Journal of chemical physics*, 150(16):161104, 2019.

[173] Melek Saouessi, Judith Peters, and Gerald R Kneller. Frequency domain modeling of quasielastic neutron scattering from hydrated protein powders: Application to free and inhibited human acetylcholinesterase. *The Journal of Chemical Physics*, 151(12):125103, 2019.

[174] Daniela Russo, Giuseppina Rea, Maya D Lambreva, Michael Haertlein, Martine Moulin, Alessio De Francesco, and Gaetano Campi. Water collective dynamics in whole photosynthetic green algae as affected by protein single mutation. *The Journal of Physical Chemistry Letters*, 7(13):2429–2433, 2016.

[175] Reina Shinozaki and Michio Iwaoka. Effects of metal ions, temperature, and a denaturant on the oxidative folding pathways of bovine α-lactalbumin. *International journal of molecular sciences*, 18(9):1996, 2017.

[176] Bachir Aoun, Eric Pellegrini, Marcus Trapp, Francesca Natali, Laura Cantù, Paola Brocca, Yuri Gerelli, Bruno Demé, Michael Marek Koza, Mark Johnson, et al. Direct comparison of elastic incoherent neutron scattering experiments with molecular dynamics simulations of DMPC phase transitions. *The European Physical Journal E*, 39(4):1–10, 2016.

[177] Stefania Perticaroli, Georg Ehlers, Christopher B Stanley, Eugene Mamontov, Hugh O'Neill, Qiu Zhang, Xiaolin Cheng, Dean AA Myles, John Katsaras, and Jonathan D Nickels. Description of hydration water in protein (green fluorescent protein) solution. *Journal of the American Chemical Society*, 139(3):1098–1105, 2017.

[178] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[179] Evangelia D Chrysina, Keith Brew, and K Ravi Acharya. Crystal structures of apo-and holo-bovine α-lactalbumin at 2.2-Åresolution reveal an effect of calcium on inter-lobe interactions. *Journal of Biological Chemistry*, 275(47):37021–37029, 2000.

[180] Mounir Tarek and Douglas J Tobias. The dynamics of protein hydration water: a quantitative comparison of molecular dynamics simulations and neutron-scattering experiments. *Biophysical journal*, 79(6):3244–3257, 2000.

[181] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616, 1998.

[182] Alexander D Mackerell Jr, Michael Feig, and Charles L Brooks III. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of computational chemistry*, 25(11):1400–1415, 2004.

[183] Hans W Horn, William C Swope, Jed W Pitera, Jeffry D Madura, Thomas J Dick, Greg L Hura, and Teresa Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *The Journal of chemical physics*, 120(20):9665–9678, 2004.

[184] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. GROMACS: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

[185] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh Ewald method. *The Journal of chemical physics*, 103(19):8577–8593, 1995.

[186] Simone Melchionna, Giovanni Ciccotti, and Brad Lee Holian. Hoover NPT dynamics for systems varying in shape and size. *Molecular Physics*, 78(3):533–544, 1993.

[187] JH Roh, VN Novikov, RB Gregory, JE Curtis, Z Chowdhuri, and AP Sokolov. Onsets of anharmonicity in protein dynamics. *Physical review letters*, 95(3):038101, 2005.

[188] Pan Tan, Yihao Liang, Qin Xu, Eugene Mamontov, Jinglai Li, Xiangjun Xing, and Liang Hong. Gradual crossover from subdiffusion to normal diffusion: a many-body effect in protein surface water. *Physical review letters*, 120(24):248101, 2018.

[189] Mark TF Telling and Ken H Andersen. Spectroscopic characteristics of the OSIRIS near-backscattering crystal analyser spectrometer on the ISIS pulsed neutron source. *Physical Chemistry Chemical Physics*, 7(6):1255–1261, 2005.

[190] Natali Francesca, Judith Peters, Daniela Russo, S Barbieri, C Chiapponi, A Cupane, A Deriu, MT Di Bari, Emmanuel Farhi, Y Gerelli, et al. IN13 backscattering spectrometer at ILL: looking for motions in biological macromolecules and organisms. *Neutron News*, 19(4):14–18, 2008.

[191] Joachim Wuttke, Alfred Budwig, Matthias Drochner, Hans Kämmerling, Franz-Joseph Kayser, Harald Kleines, Vladimir Ossovyi, Luis Carlos Pardo, Michael Prager, Dieter Richter, et al. SPHERES, Jülich's high-flux neutron backscattering spectrometer at FRM II. *Review of scientific instruments*, 83(7):075109, 2012.

[192] Dominik Zeller and Judith Peters. Complete data set of different hydration levels of bovine alpha-Lactalbumin. *Institut Laue-Langevin (ILL)*, 2016.

[193] Dominik Zeller, Aline Cisse, Loreto Misuraca, Francesa Natali, and Judith Peters. Testing the Validity of Current Models to Describe the Protein Dynamics from EFWS data. *Institut Laue-Langevin (ILL)*, 2018.

[194] D. Richard, M. Ferrand, and G. J. Kearley. Analysis and Visualisation of Neutron-Scattering Data. *Journal of Neutron Research*, 4(1-4):33–39, 1996.

[195] Gerald R Kneller and Konrad Hinsen. Quantitative model for the heterogeneity of atomic position fluctuations in proteins: a simulation study. *The Journal of chemical physics*, 131(4):07B618, 2009.

[196] Lars Meinhold, David Clement, Moeava Tehei, Roy Daniel, John L Finney, and Jeremy C Smith. Protein dynamics and stability: the distribution of atomic fluctuations in thermophilic and mesophilic dihydrofolate reductase derived using elastic incoherent neutron scattering. *Biophysical journal*, 94(12):4812–4818, 2008.

[197] Gerald R Kneller, Volker Keiner, Meinhard Kneller, and Matthias Schiller. nMOLDYN: a program package for a neutron scattering oriented analysis of molecular dynamics simulations. *Computer physics communications*, 91(1-3):191–214, 1995.

[198] Derya Vural, Jeremy C Smith, and Henry R Glyde. Determination of dynamical heterogeneity from dynamic neutron scattering of proteins. *Biophysical Journal*, 114(10):2397–2407, 2018.

[199] G Goret, B Aoun, and Eric Pellegrini. MDANSE: An interactive analysis environment for molecular dynamics simulations. *Journal of chemical information and modeling*, 57(1):1–5, 2017.

[200] Joachim Wuttke and Michaela Zamponi. Simulation-guided optimization of small-angle analyzer geometry in the neutron backscattering spectrometer SPHERES. *Review of Scientific Instruments*, 84(11):115108, 2013.

[201] Wolfgang Doster. Are proteins dynamically heterogeneous? Neutron scattering analysis of hydrogen displacement distributions. *Int. J. Mol. Theor. Phys*, 2:1–14, 2018.

[202] W Doster, H Nakagawa, and MS Appavou. Scaling analysis of bio-molecular dynamics derived from elastic incoherent neutron scattering experiments. *The Journal of Chemical Physics*, 139(4):07B624_1, 2013.

[203] Wolfgang Doster and Marcus Settles. Protein-water displacement distributions. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1749(2):173–186, 2005.

[204] Moeava Tehei, Dominique Madern, Claude Pfister, and Giuseppe Zaccai. Fast dynamics of halophilic malate dehydrogenase and BSA measured by neutron scattering under various solvent conditions influencing protein stability. *Proceedings of the National Academy of Sciences*, 98(25):14356–14361, 2001.

[205] Liang Hong, Dennis C. Glass, Jonathan D. Nickels, Stefania Perticaroli, Zheng Yi, Madhusudan Tyagi, Hugh O'Neill, Qiu Zhang, Alexei P. Sokolov, and Jeremy C. Smith. Elastic and Conformational Softness of a Globular Protein. *Phys. Rev. Lett.*, 110:028104, 2013.

[206] Jonathan D. Nickels, Hugh O'Neill, Liang Hong, Madhusudan Tyagi, Georg Ehlers, Kevin L. Weiss, Qiu Zhang, Zheng Yi, Eugene Mamontov, Jeremy C. Smith, and Alexei P. Sokolov. Dynamics of Protein and its Hydration Water: Neutron Scattering Studies on Fully Deuterated GFP. *Biophysical Journal*, 103(7):1566–1575, 2012.

[207] Zheng Yi, Yinglong Miao, Jerome Baudry, Nitin Jain, and Jeremy C. Smith. Derivation of Mean-Square Displacements for Protein Dynamics from Elastic Incoherent Neutron Scattering. *The Journal of Physical Chemistry B*, 116(16):5028–5036, 2012.

[208] Eugene A Permyakov and Lawrence J Berliner. α-Lactalbumin: structure and function. *FEBS letters*, 473(3):269–274, 2000.

[209] Mounir Tarek and Douglas J Tobias. Environmental dependence of the dynamics of protein hydration water. *Journal of the American Chemical Society*, 121(41):9740–9741, 1999.

[210] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *NATURE*, 450(7172):964–972, DEC 13 2007.

[211] E Balog, T Becker, M Oettl, R Lechner, R Daniel, J Finney, and JC Smith. Direct determination of vibrational density of states change on ligand binding to a protein. *PHYSICAL REVIEW LETTERS*, 93(2), JUL 9 2004.

[212] Katherine A. Niessen, Mengyang Xu, Alessandro Paciaroni, Andrea Orecchini, Edward H. Snell, and Andrea G. Markelz. Moving in the Right Direction: Protein Vibrational Steering Function. *BIOPHYSICAL JOURNAL*, 112(5):933–942, MAR 14 2017.

[213] David A. Turton, Hans Martin Senn, Thomas Harwood, Adrian J. Lapthorn, Elizabeth M. Ellis, and Klaas Wynne. Terahertz underdamped vibrational motion governs protein-ligand binding in solution. *NATURE COMMUNICATIONS*, 5, JUN 2014.

[214] Katherine A. Niessen, Mengyang Xu, Deepu K. George, Michael C. Chen, Adrian R. Ferre-D'Amare, Edward H. Snell, Vivian Cody, James Pace, Marius Schmidt, and Andrea G. Markelz. Protein and RNA dynamical fingerprinting. *NATURE COMMUNICATIONS*, 10, MAR 4 2019.

[215] Paul W Fenimore, Hans Frauenfelder, BH McMahon, and RD Young. Bulk-solvent and hydration-shell fluctuations, similar to α-and β-fluctuations in glasses, control protein motions and functions. *Proceedings of the National Academy of Sciences*, 101(40):14408–14413, 2004.

[216] Moran Grossman, Benjamin Born, Matthias Heyden, Dmitry Tworowski, Gregg B Fields, Irit Sagi, and Martina Havenith. Correlated structural kinetics and retarded solvent dynamics at the metalloprotease active site. *Nature structural & molecular biology*, 18(10):1102–1108, 2011.

[217] Jessica Dielmann-Gessner, Moran Grossman, Valeria Conti Nibali, Benjamin Born, Inna Solomonov, Gregg B Fields, Martina Havenith, and Irit Sagi. Enzymatic turnover of macromolecules generates long-lasting protein–water-coupled motions beyond reaction steady state. *Proceedings of the National Academy of Sciences*, 111(50):17857–17862, 2014.

[218] CF HIGGINS. ABC TRANSPORTERS - FROM MICROORGANISMS TO MAN. *ANNUAL REVIEW OF CELL BIOLOGY*, 8:67–113, 1992.

[219] FA Quiocho and PS Ledvina. Atomic structure and specificity of bacterial periplasmic receptors for active transport and chemotaxis: Variation of common themes. *MOLECULAR MICROBIOLOGY*, 20(1):17–25, APR 1996.

[220] CB Felder, RC Graul, AY Lee, HP Merkle, and W Sadee. The venus flytrap of periplasmic binding proteins: An ancient protein module present in multiple drug receptors. *AAPS PHARMSCI*, 1(2), 1999.

[221] Lucas F. Ribeiro, Vanesa Amarelle, Liliane F. C. Ribeiro, and Maria-Eugenia Guazzaroni. Converting a Periplasmic Binding Protein into a Synthetic Biosensing Switch through Domain Insertion. *BIOMED RESEARCH INTERNATIONAL*, 2019, 2019.

[222] Lakshmi P Jayanthi, Nahren Manuel Mascarenhas, and Shachi Gosavi. Structure dictates the mechanism of ligand recognition in the histidine and maltose binding proteins. *Current Research in Structural Biology*, 2:180–190, 2020.

[223] Marco van den Noort, Marijn de Boer, and Bert Poolman. Stability of ligand-induced protein conformation influences affinity in maltose-binding protein. *Journal of molecular biology*, 433(15):167036, 2021.

[224] Wenbo Yu, Xibing He, Kenno Vanommeslaeghe, and Alexander D MacKerell Jr. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of computational chemistry*, 33(31):2451–2468, 2012.

[225] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[226] Michael R Shirts, Christoph Klein, Jason M Swails, Jian Yin, Michael K Gilson, David L Mobley, David A Case, and Ellen D Zhong. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *Journal of computer-aided molecular design*, 31(1):147–161, 2017.

[227] Leandro Martínez, Ricardo Andrade, Ernesto G Birgin, and José Mario Martínez. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry*, 30(13):2157–2164, 2009.

[228] Steven Dajnowicz, Yongqiang Cheng, Luke L Daemen, Kevin L Weiss, Oksana Gerlits, Timothy C Mueser, and Andrey Kovalevsky. Substrate Binding Stiffens Aspartate Aminotransferase by Altering the Enzyme Picosecond Vibrational Dynamics. *ACS omega*, 5(30):18787–18797, 2020.

[229] Oscar Millet, Rhea P Hudson, and Lewis E Kay. The energetic cost of domain reorientation in maltose-binding protein as studied by NMR and fluorescence spectroscopy. *Proceedings of the National Academy of Sciences*, 100(22):12700–12705, 2003.

[230] Patrick G Telmer and Brian H Shilton. Insights into the conformational equilibria of maltose-binding protein by analysis of high affinity mutants. *Journal of Biological Chemistry*, 278(36):34555–34567, 2003.

[231] KT Wikfeldt, Anders Nilsson, and Lars GM Pettersson. Spatially inhomogeneous bimodal inherent structure of simulated liquid water. *Physical Chemistry Chemical Physics*, 13(44):19918–19924, 2011.

[232] Alexei A Maradudin, Elliott Waters Montroll, George Herbert Weiss, and IP Ipatova. *Theory of lattice dynamics in the harmonic approximation*, volume 3. Academic press New York, 1963.

[233] Eunkyung Kim, Sanghwa Lee, Aram Jeon, Jung Min Choi, Hee-Seung Lee, Sungchul Hohng, and Hak-Sung Kim. A single-molecule dissection of ligand binding to a protein with intrinsic dynamics. *Nature chemical biology*, 9(5):313–318, 2013.

[234] Kathleen Wood, Andreas Frölich, Alessandro Paciaroni, Martine Moulin, Michael Härtlein, Giuseppe Zaccai, Douglas J Tobias, and Martin Weik. Coincidence of dynamical transitions in a soluble protein and its hydration water: direct measurements by neutron scattering and md simulations. *Journal of the American Chemical Society*, 130(14):4586–4587, 2008.

[235] Erika Balog, David Perahia, Jeremy C Smith, and Franci Merzel. Vibrational softening of a protein on ligand binding. *The Journal of Physical Chemistry B*, 115(21):6811–6817, 2011.

[236] Gustavo Adrian Appignanesi, Jorge Ariel Rodriguez Fris, and Francesco Sciortino. Evidence of a two-state picture for supercooled water and its connections with glassy dynamics. *The European Physical Journal E*, 29(3):305–310, 2009.

[237] Anna Stradner and Peter Schurtenberger. Potential and limits of a colloid approach to protein solutions. *Soft Matter*, 16(2):307–323, 2020.

[238] Ralph H Colby. Structure and linear viscoelasticity of flexible polymer solutions: comparison of polyelectrolyte and neutral polymer solutions. *Rheologica acta*, 49(5):425–442, 2010.

[239] Prasad S Sarangapani, Steven D Hudson, Ronald L Jones, Jack F Douglas, and Jai A Pathak. Critical examination of the colloidal particle model of globular proteins. *Biophysical journal*, 108(3):724–737, 2015.

[240] P Douglas Godfrin, Néstor E Valadez-Pérez, Ramon Castaneda-Priego, Norman J Wagner, and Yun Liu. Generalized phase behavior of cluster formation in colloidal dispersions with competing interactions. *Soft matter*, 10(28):5061–5071, 2014.

[241] Zhenhuan Zhang and Yun Liu. Recent progresses of understanding the viscosity of concentrated protein solutions. *Current opinion in chemical engineering*, 16:48–55, 2017.

[242] Roger L Chang, Kathleen Andrews, Donghyuk Kim, Zhanwen Li, Adam Godzik, and Bernhard O Palsson. Structural systems biology evaluation of metabolic thermotolerance in Escherichia coli. *Science*, 340(6137):1220–1223, 2013.

[243] Anna Jarzab, Nils Kurzawa, Thomas Hopf, Matthias Moerch, Jana Zecha, Niels Leijten, Yangyang Bian, Eva Musiol, Melanie Maschberger, Gabriele Stoehr, et al. Meltome atlas—thermal proteome stability across the tree of life. *Nature methods*, 17(5):495–503, 2020.

[244] Laura B Persson, Vardhaan S Ambati, and Onn Brandman. Cellular control of viscosity counters changes in temperature and energy availability. *Cell*, 183(6):1572–1585, 2020.

[245] J. Xie, J. Najafi, R. Le Borgne, J-M. Verbavatz, C. Durieu, J. Sallé, and N. Minc. Contribution of cytoplasm viscoelastic properties to mitotic spindle positioning. *Proceedings of the National Academy of Sciences*, 119:e2115593119, 2022.

[246] Nicole O. Taylor, Ming Tzo Wei, Howard A. Stone, and Clifford P. Brangwynne. Quantifying Dynamics in Phase-Separated Condensates Using Fluorescence Recovery after Photobleaching. *Biophysical Journal*, 117(7):1285–1300, 2019.