UNIVERSITÉ PARIS DIDEROT (PARIS 7) SORBONNE PARIS CITÉ

ÉCOLE DOCTORALE MTCI - ED 563
MÉDICAMENT, TOXICOLOGIE, CHIMIE, IMAGERIES

THÈSE DE DOCTORAT
Spécialité: Modélisation moléculaire

Les protéines thermophiles:

stabilité et fonction

Marina KATAVA

Laboratoire de Biochimie Théorique
Institut de Biologie Physico-Chimique

Thèse dirigée par: Fabio STERPONE

soutenue publiquement le 14 octobre 2016

composition du jury:
Martin WEIK, Rapporteur (DR2, IBS, Grenoble, FR)
Marco CECCHINI, Rapporteur (CR2, ISIS, Strasbourg, FR)
Alesandro PACIARONI, Examinateur (Professeur, Université de Perouse, Perouse, IT)
Anne-Claude CAMPROUX, Examinateur (Professeur, Université Paris Diderot, Paris, FR)
Philippe DERREUMAUX, Examinateur (Professeur, Université Paris Diderot, Paris, FR)
Fabio STERPONE, Directeur de thèse (CR1, CNRS, Paris, FR)

UNIVERSITÉ PARIS DIDEROT (PARIS 7) SORBONNE PARIS CITÉ

Marina KATAVA

Thermophilic proteins:

stability and function

THESIS SUPERVISED BY:
FABIO STERPONE

submitted for the degree of
DOCTOR OF PHILOSOPHY

OCTOBER 2016
PARIS, FRANCE

Abstract

Temperature is one of the major factors governing life as demonstrated by the fine tuning of stability and activity of the molecular machinery, proteins in particular. The structural stability and activity of proteins have been often presented as equivalent. However, the thermophilic proteins are stable at ambient condition, but lack activity, the latter recovered only when the temperature increases to match that of the optimal growth condition for the hosting organism. In discussing the protein stability and activity, mechanical rigidity is often used as a relevant parameter, offering a simple and appealing explanation of both the extreme thermodynamic stability and the lack of activity at low temperature. The reality, however, illustrates the complexity of the rigidity/flexibility trade off in ensuring stability and activity through intricate thermodynamic and molecular mechanisms. Here we investigate the problem by studying three study cases. These are used to relate the thermal effects on mechanical properties and the stability and activity of the proteins. For instance, we have probed the thermal activation of functional modes in EF G-domain and Lactate/Malate Dehydrogenase mesophilic and thermophilic homologues and verified a "universal" scaling of atomistic fluctuation of the Lysozyme approaching the melting in different environmental conditions. Our conclusions largely rest on an *in silico* approach, where Molecular Dynamics and enhanced sampling techniques are utilized, and are often complemented with Neutron Scattering Experiments.

RÉSUMÉ

La température est un paramètre crucial dans le fonctionnement du monde vivant, notamment de la machinerie moléculaire (les protéines) dont la stabilité et l'activité en dépendent sensiblement. Celles-ci sont souvent considérées comme étant équivalentes : si une protéine fonctionne, c'est qu'elle est stable, et vice-versa. Cependant, les protéines des organismes thermophiles, qui prolifèrent dans de températures élevées, sont stables à température ambiante, mais y présentent une faible activité. Cette dernière est optimale à la température de croissance de l'organisme hôte. Lorsqu'on parle de stabilité et d'activité protéique, la rigidité mécanique est souvent utilisée comme paramètre pertinent, offrant une explication simple et attractive à la fois pour la stabilité thermodynamique à haute température et au manque d'activité à des températures plus modérés. La réalité s'avère souvent plus complexe, et les mécanismes moléculaire reliant rigidité/flexibilité avec la stabilité et l'activité sont encore mal compris. Dans ce travail, nous abordons le problème au travers de trois systèmes. Nous avons examiné l'activation thermique des modes fonctionnels du domaine G de la protéine EF ainsi que les homologues mésophiles et thermophiles de la déshydrogénase Lactate/Malate. Par ailleurs, nous avons mis en évidence l'existence d'un paramètre unique (la moyenne des fluctuations atomiques) permettant d'expliquer la dynamique de la protéine lysozyme près de son point de fusion, et ce quelque soit la nature de l'environnement autour de la protéine (qui décale le point de fusion). Nos conclusions se basent principalement sur une approche *in silico* où la dynamique moléculaire et des techniques d'échantillonnage améliorées sont utilisées et sont complémentées par des expériences de diffraction de neutrons.

SOMMAIRE

Il a été suggéré à plusieurs reprises que la vie á des températures élevées serait sans doute apparue il y a de 30,000 á 100,000 ans. Les sources hydrothermales ont été proposées comme un habitat possible où la vie aurait émergé, puisqu'elles sont actuellement habitées par des micro-organismes vivant à des températures allant jusqu'à 122 ^{o}C. Malgré l'attrait de cette théorie, aucune conclusion définitive ne peut être tirée. Par exemple, la chimie prébiotique suggère que le régime à haute température favorise une décomposition rapide des molécules biologiques. Indépendamment de leur origine, les organismes appartenant aux trois domaines du vivant – eucaryotes, archées et bactéries, ont évolué pour habiter la Terre actuelle, y compris dans des environnements extrêmes. Des colonies bactériennes ont été retrouvées à des températures aussi basses que -60^{o}C, et aussi hautes que 113^{o}C pour *Pyrolobus fumarii* ou 122^{o}C pour *Methanopyrus kandleri*. Récemment, un spectre biocinétique de température, représentant les taux de croissance de toutes les souches considérées en fonction de la température, a montré un pic important à 42^{o}C, et un second pic plus faible à 67^{o}C, correspondant respectivement aux conditions dans lesquelles les mésophiles et les thermophiles se développent. Le spectre expérimental a été reconstruit avec un modèle basé sur un compromis entre stabilité et activité des protéines, un sujet important non seulement dans le contexte de l'évolution biologique, mais aussi pour de potentielles applications industrielles visant à optimiser les processus biotechnologiques et la stabilité des médicaments. Dans cette thèse, nous discutons en détail du problème de la stabilité thermique des protéines en examinant à la fois des modèles physiques et les détails moléculaires qui leur sont associés, tout en essayant de comprendre la relation entre la structure et la fonction, sachant que les protéines thermophiles sont inactives à température ambiante malgré la grande stabilité de leur structure.

Afin de comprendre l'effet de la température sur la fonction des protéines et sur leur stabilité, nous considérons trois cas d'étude. Étant donné que la fenêtre de température optimale pour la l'activité d'une protéine est relativement étroite, nous voudrions observer où et comment cette optimisation est réalisée afin de mieux comprendre la relation entre la structure des protéines et leur activité. Pour cela, les protéines thermophiles représentent un modèle idéal puisque leur stabilité à des températures modérées et élevées est accompagnée par une activité dans le régime à haute température uniquement, ce qui suggère que la relation entre un repliement protéique stable et une protéine active n'est pas aussi simple que ce qui est souvent présenté. Dans notre étude, nous avons systématiquement comparé des protéines similaires avec des températures opti-

males de travail différentes afin d'étudier le compromis entre la stabilité et la fonction, ainsi que sa relation avec la flexibilité mécanique de la protéine.

Des simulations numériques et des calculs sont utilisés comme la méthode de choix. La dynamique moléculaire (MD) et son extension pour atteindre un meilleur échantillonnage des configurations moléculaires, REST2, sont largement utilisées pour obtenir des simulations de systèmes sur une échelle de temps de l'ordre de la microseconde. La dynamique moléculaire produit des trajectoires des positions atomiques dans le temps et, en exploitant la mécanique statistique, permet le calcul de la thermodynamique à l'équilibre et des propriétés dynamiques telles les coefficients de diffusion. De plus, la technique permet l'observation d'événements biologiquement pertinents à une résolution atomistique, ce qui est d'une grande valeur pour la recherche moderne. Afin de renforcer davantage nos études, nous complémentons les résultats *in silico* avec des expériences de diffusion de neutrons (Neutron Scattering (NS)) qui sondent des gammes similaires de longueur et d' échelle de temps. La possibilité de déterminer les quantités observées dans les expériences de NS à partir des trajectoires simulées fait de l'utilisation combinée de ces deux techniques un outil solide pour interpréter les études structurales et dynamiques des protéines.

Dans la première étude présentée dans la thèse, nous étudions les changements conformationnels du G-domain se produisant durant le turnover enzymatique d'une paire d'homologues mésophiles et hyperthermophiles. La comparaison d'enzymes homologues adaptées à différents environnements thermiques aide à éclairer le délicat compromis entre stabilité et fonction. La rigidité mécanique des protéines a été proposée comme assurant la stabilité et la fonctionnalité des protéines thermophiles à haute température. Nous avons contesté le principe de cette hypothèse pour une paire d'homologues de domaines GTPase en effectuant de nombreuses simulations de MD, en appliquant des algorithmes de groupement conformationnel et cinétique, tout en exploitant une technique d'échantillonnage amélioré (REST2). Comme il a été auparavant montré que l'amélioration de la flexibilité des protéines et une stabilité à haute température ne peuvent coexister dans la variante hyperthermophile apo, nous nous sommes concentrés sur les états holo des deux homologues en imitant le turnover enzymatique. Nous avons clairement montré que la présence des ligands affecte le paysage conformationnel visité par les protéines, et il s'agit du principe de l'état correspondant nécessaire pour certains modes fonctionnels. Dans les espèces hyperthermophiles, la flexibilité de la région effectrice assurant la communication à longue distance et la boucle P modulant la liaison du ligand ne sont récupérées qu'à haute température.

De plus, nous avons étudié l'activation thermique des modes doux des protéines en combinant des expériences de diffusion de neutrons à écho de spin et des simulations de MD. Le but ultime est de comparer la réponse thermique des modes fonctionnels dans une paire Lactate/Malate Déshydrogénases. La diffusion de neutrons avec écho de spin permet de sonder les mouvements à des longueurs et des échelles de temps de l'ordre du nanomètre et nanoseconde –ce qui est pertinent pour des grandes réorganisations conformationnelles de protéines, tandis que les simulations de MD soutiennent l'interprétation microscopique des spectres expérimentaux. Les résultats obtenus pour les espèces mésophiles, la Lactate Déshydrogénase 5 eucaryote du muscle de lapin (LDH M5), seront

largement présentés. Pour la Lactate Déshydrogénase, nous avons sondé l'activation thermique des modes fonctionnels couvrant les échelles de longueur des séparations entre les domaines, correspondant à la réorganisation allostérique préalablement sondée pour les LDHs bactériennes.

Le dernier système considéré, le Lysozyme à l'état de poudre, nous a permis à évaluer un aspect complémentaire de la relation entre la flexibilité mécanique, la stabilité et la fonction, avec comme résultat principal l'existence d'une loi d'echelle pour les fluctuations atomiques à l'approche du point de fusion de la protéine, concept d'ordinaire utilisé pour décrire la transition de phase dans les solides. Une relation simple, le critère de Lindemann, prédit l'apparition de la fusion une fois que les fluctuations thermiques dépassent une valeur seuil à laquelle la cristal fond. Ici, nous cherchons à vérifier si le concept peut être étendu à la matière biologique non homogène en utilisant comme modèle le Lysozyme et en utilisant la MD et les simulations d'échantillonnage amélioré pour atteindre cet objectif. À cet effet, nous avons considéré la protéine dans trois environnements différents : en solution aqueuse diluée et deux systèmes de poudre, un solvaté avec de l'eau et l'autre avec du glycérol. Un des effet de ces conditions est le déplacement de la température de fusion du Lysozyme. Les systèmes ont été simulés et analysés avec l'objectif final d'élucider si la fusion de protéines est accompagnée d'une mise à l'échelle universelle des fluctuations atomiques. Le travail a été inspiré par les expériences récentes de diffraction de neutrons incohérente élastique réalisées par nos collaborateurs.

En conclusion, notre étude basée sur la MD couplée à des expériences de diffusion des neutrons nous a permis d'étudier la validité de deux paradigmes classiques liés aux effets de température - le principe d'état correspondant de Somero et le critère de Lindemann. Le premier corrèle l'apparition de l'activité enzymatique dans les protéines thermophiles à l'activation thermique de la flexibilité des protéines, tandis que le second définit l'ampleur critique des fluctuations atomiques pour initier la fusion. Les deux principes témoignent du rôle central de la température dans la modulation des adaptations évolutives.

ACKNOWLEDGMENTS

The three years of my PhD have been an incredibly positive experience, which can be greatly credited to Dr. Fabio Sterpone, who supervised this thesis. He has pushed my limits and supported me in equal proportions, and has given me the opportunity to do cutting edge science in a scientifically rigorous manner. This thesis wouldn't have been what it is without him and I couldn't extend enough gratitude in his direction.

The manuscript has been further improved with the help of Dr. Martin Weik and Dr. Marco Cecchini, whom I am thankful for having read and reviewed the thesis in great detail, and for having accepted to be a part of the jury. I also wish to thank other jury members, Prof. Anne-Claude Camproux, Prof. Philippe Derreumaux, and Prof. Alessandro Paciaroni, for deeming this work interesting enough to accept to judge it.

The research presented here was done through several collaborations. I wish to thank Marco Maccarini for the wonderful initiation into the world of Neutron Scattering, which was made easier through his knowledge of the matter and his good will and kindness. Dominique Madern is a scientist of great detail and colossal knowledge on the dehydrogenase protein family and having worked and observed his data interpretation and passion for the subject has contributed to the research a great deal. I am indebted to Alessandro Paciaroni for the opportunity to work on the Lysozyme project which has brought me a lot of joy. Immense thankyous are extended to Guillaume Stirnemann, for always having his doors open for me, for his scientific insight, and method development.

Research is not possible without funding, here provided by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) Grant Agreement no.258748. The financial support for infrastructures is attributed to ANR-11-LABX-0011-01. HPC resources were provided by GENCI [CINES and TGCC] (Grant x201576818), while the experimental part of this work is performed at the J-NSE [1] instrument operated by JCNS and FRM II at the Heinz Maier-Leibnitz Zentrum (MLZ), Garching, Germany.

The Laboratoire de biochimie théorique is generally a fun, diverse, and comfortable place. Cakes are regularly passed around, help is never refused, and the spirits are always high. I feel genuinely lucky to have worked there while pursuing my scientific goals. Many thanks to Victoria Terziyan and Geoffrey Letessier, their work makes the research possible. Maria Kalimeri, Tristan Cragnolini, and Benoist Laurent have pulled me from more first-year-PhD pitfalls than I can acknowledge here. Debashree Chakraboty and Aixiao Li provided the invaluable moral support. I will always remember fondly the repartees with Samuel Murail. I specially wish to acknowledge three lab members -

Thanh Thuy Tran, Dario De Vecchis, and Mara Chiricotto for being PhD chickens with me from the beginning till the end, with all the ups and downs, and for the joint effort of entering the building Lamarck B of the University Paris Diderot.

The journey to this thesis was full of supportive professors and teachers. Paolo de Los Rios gave me the opportunity to do theory through an internship and he was the first scientist whom I heard speaking of molecular simulation. Gerhard Gompper has offered me a practical initiation in the world of simulation and theoretical modeling - I did an internship and a master thesis project with him. Tomica Hrenar is a theoretical chemistry professor who has taught me so much. Hearing professor Vladislav Tomišić's lectures made me fall in love with physical chemistry. Jasna Stublić, highschool teacher, prepared me for math competitions. Ivica Martinjak, whose passion for physics and positivity will always stay with me. Darija Sever, elementary school teacher, encouraged me to follow my dreams. Many thanks to my friends for the fun. It is always so very random, inappropriate, and crazy with you.

Finally, I am most grateful for the support of my family, few are so lucky to be unconditionally supported and believed in as I am.

This thesis is dedicated to my grandparents - Marko Ezgeta, Kata Batarilo, Marko Katava, and Marija Mostarac.

TABLE OF CONTENTS

List of Tables

Table **Page**

LIST OF FIGURES

FIGURE **Page**

CHAPTER 1

INTRODUCTION

It has been repeatedly suggested that life originated at high temperatures presumably between 30,000-100,000 years ago [2, 3, 4]. Hydrothermal vents were proposed as a likely habitat for life to emerge as they are presently inhabited with microorganisms thriving at temperatures up to 122^{o}C [5]. Despite the appeal of the theory, no definite conclusion can be made, e.g. the prebiotic chemistry suggests a quick decomposition of important biological molecules in the high temperature regime and thus favors somewhat lower temperatures [6]. Independent of their origin, organisms belonging to all three domains of life - eukarya, archea, and bacteria, have evolved to inhabit present Earth, including extreme environments. Bacterial colonies have been found in Antarctica at temperatures as low as -60^{o}C [7] and as high as 113^{o}C for *Pyrolobus fumarii* [8] or 122^{o}C for *Methanopyrus kandleri* [9]. Recently, a biokinetic spectrum for temperature was reported, where the growth rates of all considered strains have been presented as a function of temperature, exhibiting a sharp peak at 42^{o}C, and a second lower peak at 67^{o}C [10], corresponding to the conditions in which mesophiles and thermophiles thrive. The experimental spectrum was reconstructed with a model based on protein stability/activity trade off, a topic important not only in the context of biological evolution, but also for potential industrial applications aiming to optimize biotechnological processes and drug stability. In this introduction, we extensively discuss the problem of protein thermal stability by examining both the physical models and molecular details, while trying to grasp the structure to function relationship, knowing that the thermophilic proteins are inactive at moderate temperature despite their stable structural fold.

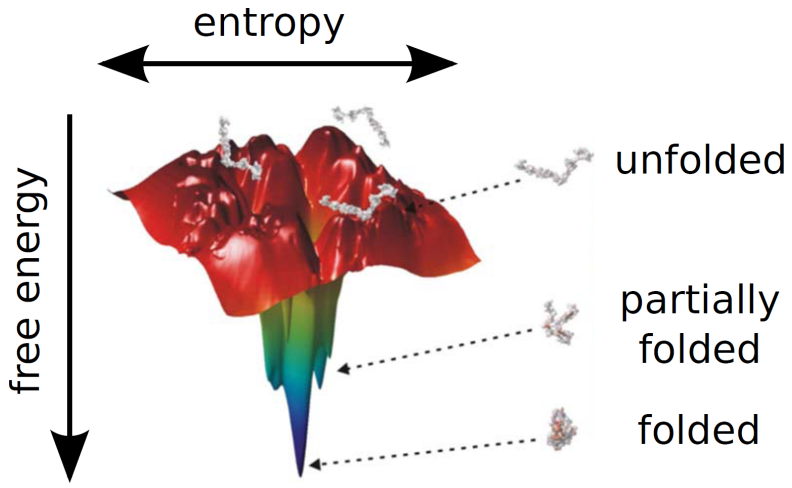

Figure 1.1: The free energy surface of a protein can be schematically represented as a rugged funnel. The top of the funnel contains the unfolded states, while the native state is of significantly lower energy and entropy compared to other states, making the folding a favorable event. Adapted from Ref [15].

1.1 Structural Stability

X-ray crystallography has equipped us with tools to observe proteins at an atomic-level resolution [11]. Knowing that all matter is dynamic due to thermal fluctuations of atoms, the observed static picture emerging from the X-ray crystallography remains powerful due to high level of detail and resolution it offers, but consequently necessitates a dynamic complement still preserving microscopic resolution. Several techniques meet these criteria, most notably Nuclear Magnetic Resonance (NMR) spectroscopy [12], and as of late, cryo-Electron Microscopy (cryo-EM) [13, 14]. On the other hand, *in silico* methods, such as Molecular Dynamics, emerged along the last century as the computational support for the investigation of protein structure, dynamics, and function.

The experimentally resolved structures target biologically relevant, compact states, termed the native state. The characteristic three dimensional structural fold of proteins is the necessary condition for biological activity and, as such, has been the main focus in protein research. Given the 20^N possible combinations to form a protein chain of N amino-acids, and the vast number of possible configurations a polymer of N monomers can theoretically accomplish, the fact that proteins fold on biologically relevant timescales is a remarkable example of evolutionary selection. In fact, a random heteropolymer placed in solution would exhibit Brownian motion and form energetically favorable non-bonded intermolecular contacts, forming a collapsed state resembling a random coil

or a cross-linked gel, depending on the number of contacts formed [16]. These formations lack the necessary configurational specificity to achieve function. Further increase in the number of intermolecular contacts would lead to glass formation, with a number of distinct kinetically trapped states [16]. Any one of these states would be susceptible to mutational or environmental pressure, without the structural robustness attributed to proteins. The efficient protein fold has been instead attributed to an energy landscape resembling a funnel [16, 17, 18], see Figure 1.1, where the native state is of distinctly lower energy than other states as a consequence of 'minimal frustration' [19], i.e. the presence of interactions that would lead to non-native states is minimized as a result of natural selection [16, 19]. The complete elimination of 'frustration' would produce perfectly smooth funnel-shaped energy surfaces [20], in contrast to real proteins, whose funnels are rugged and thus yield a number of non-transient conformational substates corresponding to the local minima on the free energy surface [16, 17, 18]. The distinct conformational substates provided by 'frustration' are a biological necessity as proteins must exist in distinct conformational states to accommodate their role in enzyme activity and signaling. The shape of the conformational landscape further suggests a large number of protein configurations at the top of the funnel and a dramatic decrease in the the number of substates along the path of protein folding, producing a positive entropy contribution to the free energy in the folding pathway. Nonetheless, for the majority of globular proteins, the collapse of the hydrophobic core and the release of the caged water molecules from the immediate surface of these residues offers an entropic compensation accounting for 75% of the free energy of folding [21]. A comparative study has found the globular proteins to be only marginally stable with folding free energies of 5.5 ± 1.6 kcal/mol [22], corresponding to a formation of a few hydrogen bonds, whose energy stabilizing contribution in proteins have been estimated to 0.5-1.5 kcal/mol [23]. This marginal stability is a consequence of enthalpic-entropic compensation in protein folding, where each stabilizing intramolecular interaction, e.g. formation of salt bridges and hydrogen bonds, is accompanied by a destabilizing protein entropy reduction upon folding [24].

The free energy difference reported to characterize protein stability relies on a two state protein model, where all protein substates are merged to either the folded or the unfolded state, valid for many globular proteins [17], but invalid for large and complex proteins [26, 27]. Nonetheless, although the real phase diagram of the protein additionally comprises at least the molten globule and glassy phase [18], the two state model remains widely used as it can be accompanied by simple equilibrium thermodynamic

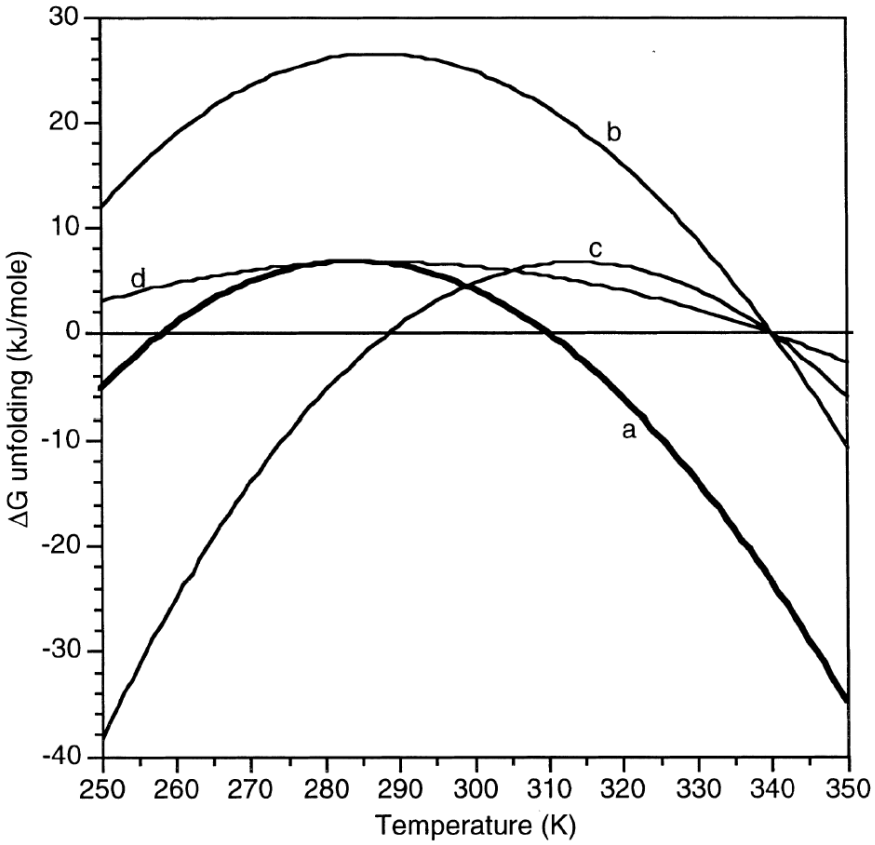

Figure 1.2: Stability curves corresponding to different protein thermal stabilities. Curve (a) represents the stability of a mesophilic protein, while other three curves represent the thermophilic curves in three possible variants: (b) upshifting the stability curve, (c) right-shifting the stability curve, (d) broadening of the stability curve. Taken from Ref [25].

arguments on protein stability [25, 22]. These arguments trace protein stability in terms of its pH, ionic strength, and temperature [28], the latter being of special interest to our studies. The thermal stability of proteins can be understood by outlining the protein stability curve, the free energy of protein unfolding as a function of temperature $\Delta G(T)$, as shown in Figure 1.2. The important feature of the curves is their parabolic shape, producing three characteristic temperatures, the maximum defining the optimal temperature stability and the two temperatures corresponding to $\Delta G=0$, the lower temperature marking the cold unfolding [29], and the higher high-temperature denaturation. We do not inspect the cold unfolding in our studies, rather we concentrate on the unfolding in the high temperature regime, referred to from now on 'melting', and the temperature at which it occurs the melting temperature.

Note that the melting temperature of the protein can be shifted by either of the three mechanisms shown in Figure 1.2; right-shifting, up-shifting, and broadening of the stability curve, or a combination thereof. Importantly, with the exception for the mechanism shown in (c), this does not affect the maximum stability temperature that clusters around ~283 K, pointing to the importance of the hydrophobic interactions in protein stability [21, 30], as its optimum is reached at these values. This agrees with the observation of a stable structural fold at moderate temperatures for both the mesophilic and thermophilic proteins, while still failing to account for the inactivation of thermophilic proteins in the moderate temperature regime. The thermodynamic stabilization of the thermophilic proteins has been granted mostly to the upshift of the stability curve with broadening [31, 32, 33, 25] or, in smaller proportions, shifting it to the right [31]. The molecular mechanisms of the thermal stability can be to some extent devised from the observed attributes of the stability curve, e.g. from the enthalpic elimination of the entropic component at the maximum. The enthalpic contribution is dominated by changes in protein internal energy that can arise by formation of salt bridges and hydrogen bonds, which would in turn make the protein matrix more rigid and increase its thermal stability. Indeed it has been found that the thermophilic proteins contain a surplus of charged amino-acids with respect to their mesophilic homologues [34, 35] and it was additionally suggested that the optimum placement of these amino-acids in the structure is necessary to achieve the enhanced structural stabilization [36, 37, 38]. Additionally, the increased packing of the protein core [22] and loop shortening [39] seem to be important in increasing thermal stability, although other studies on protein packing have also suggested otherwise [40]. In contrast to these conclusions on thermal stability resting on rigidity, an entropic mechanism has been proposed [41], where the increased flexibility aids to dissipate thermal stress, supported by studies that found the flexibility of the thermophilic protein to be as high, and sometimes exceeding their mesophilic counterpart [42, 43, 44, 45]. In addition to the purely thermodynamic considerations of protein stability, it is perfectly plausible that the increased thermal stability could be a consequence of kinetic trapping of the folded state, separated from the unfolded state in this instance by a large energy barrier [31]. The kinetic trapping has been suggested as the mechanism by which rubredoxin from *Pyrococcus furiosus* maintains stability at temperatures above the boiling point of water [46].

Considering the dramatic changes in properties and structure water undergoes in the temperature range 0-100^{o}C, it is reasonable to consider the solvent effect on the protein stability. The number of per-atom water-protein contacts has found to be maximized

in thermophilic species, forming an extended hydration shell around the protein [47] due to large hydrophilic patches on the interface with water [48] and preventing the water to penetrate the hydrophobic core. The notion is further reinforced by showing the disruption of the water network at the interface being a precursor to melting [47]. To inspect the effect of solvent and protein crowders, we have performed a study on the mesophilic protein Lysozyme in the framework of the two-state model. The study is outlined in detail in Chapter 5, and the main result shows the thermoprotective character of glycerol as compared to water, and the increase in the melting temperature in the crowded condition. The effect of crowding is attributed to the reduction of entropy of the unfolded state, observed at the atomic level as the quenching of atomic fluctuations in the presence of crowders.

As can be deduced from our discussion, the thermal stability cannot be conferred to a single mechanism, but rather arises form the superimposition of different factors. An illustration of this is the role of salt bridges in protein stability, which must be put in the context of solvent temperature increase. Raising the temperature decreases the dielectric constant of water, which reduces the 'shielding' effect and makes the electrostatic interactions effectively stronger. This is supported by computational studies showing the desolvation penalty of the salt bridges to decrease in the high temperature regime, producing a stabilizing contribution to protein stability [49]. Another pictorial example is the one of Lysine, a charged amino acid that, in addition to forming salt bridges, increases the entropy of the folded state with its many rotamer substates, thus participating also in entropic protein stabilization [35]. These examples show that in probing the factors affecting protein stability, examinations must be put in the context of the protein environment and biological function.

In addition to the mechanisms of thermal stabilization provided here, we refer the reader to some excellent reviews that cover the topic in more detail [50, 33, 51, 52, 53, 54], while proceeding to examine, in the following section, protein dynamics and activity through the rigidity/flexibility paradigm, and the underlying mechanisms pertaining to the high temperature regime.

1.2 Activity and Dynamics

Assuming different conformational substates is a necessary prerequisite for biological activity, thus the evolutionary retainment of structural 'frustration', as argued in the previous section. In the context of the protein free energy surface, the biologically active

substates correspond to local minima in the well of the native state. The transitions between two conformational substates separated via barrier happen when the relevant degree of freedom, possessing the energy $k_BT/2$ according to the equipartition principle, accumulates the energy (and appropriate entropy) sufficient to achieve the improbable high-energy configuration corresponding to the barrier peak, termed the transition state. In the framework of the Transition State Theory (TST), the rate of transition is associated to the energy barrier via a simple exponential relationship $k = \frac{k_BT}{h}\exp - \frac{\Delta G^\dagger}{k_BT}$, where $\Delta G^\dagger$ corresponds to the free energy difference between the initial (reactant) state and the top of the barrier. The probability of transition therefore depends on the height of the barrier, manifesting in a range of protein activated motions occurring at different timescales [55]. The fastest dynamics corresponds to bond vibrations happening on a femtosecond timescale. These fluctuations happen within valleys separated by energy barriers of k_BT, which result in picosecond (atomic fluctuations, sidechain rotations) or nanosecond (loop motion) dynamics at physiological temperature. Finally, the long dynamics on a micro- and milisecond (matching the characteristic time scale of enzyme catalysis and signaling) corresponds to transitions between valleys separated by energy barriers of several k_BT.

In TST, the rate of conformational transition depends not only on the height of the barrier between the two states, but also on the temperature. Typically, the transitions between states such as those in enzymatic reactions yield a twofold enhancement turnover over a 10^oC temperature increase, e.g. increasing the temperature from 37^oC to 100^oC would increase the efficiency by a factor of $2^{\frac{100-37}{10}} \sim 80$. Based on this notion only, the conformational transitions and reactions obeying the simple exponential kinetics would naturally be faster and more efficient in the thermophiles as these reactions take place in high-temperature habitats, which is not observed in reality [56] as it would bring about an overrepresentation of enzymatic products in the metabolism, an event unfavorable to the homeostasis and survival [57]. The compensation for the temperature effect is identified through different modulations of enzyme parameters and the metabolic tuning of the substrate/product concentrations [57]. The enzyme tuning is achieved for example through the temperature dependence of the chemical equilibria, in particular the effect on the deprotonation of the Lysine, Arginine, and Histidine, which affects the efficiency of chemical reactions and transitions [58].

In the context of the ideas presented above, efficient conformational changes, enzymatic reactions, and signaling would entail protein populating the appropriate minima on the free energy surface, i.e. being flexible, while at the same time maintaining proper structural fold, i.e. being rigid. Thus, the trade off between protein flexibility/rigidity

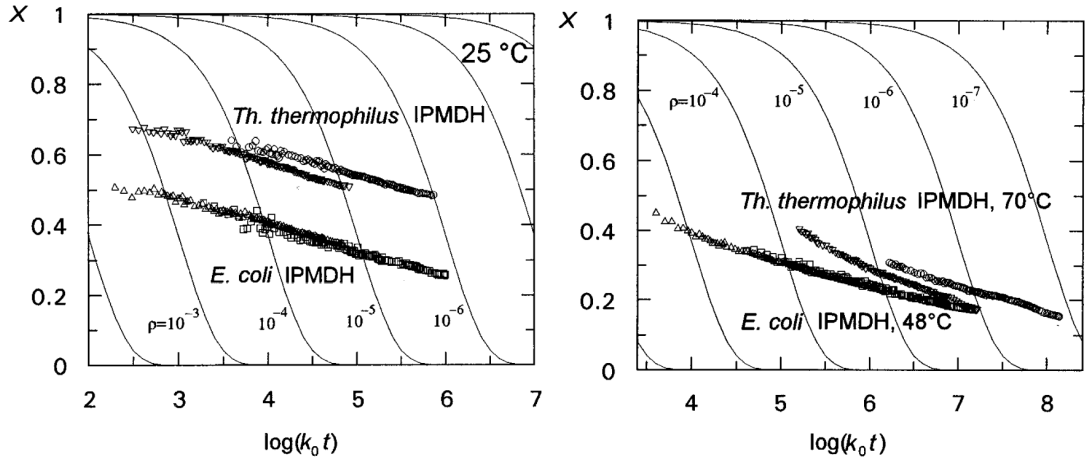

Figure 1.3: The kinetic spectrum of Hydrogen/Deuterium (H/D) exchange for two homologous 3-Ispopropylmalate Dehydrogenases (IPMDH), mesophilic from *E. coli* and thermophilic from *T. thermophilus*. The spectrum shows the ratio of unexchanged peptide Hydrogens (X) as a function of time t, shown here as $log(k_0 t)$ as the exchange rate k_0 is taken into account. The proteins are solvated in D_2O, and a H/D exchange is expected to happen over time, with a proportion depending on the accessibility of the H atoms to the solvent, consequently, the highly flexible proteins will exchange a larger proportions of H, resulting in low X values for flexible, and high X values for rigid peptides. For both proteins, the experiments were performed at two different pH values, and additionally, the thin solid lines in the background show the theoretical prediction of the spectrum that will not be discussed. The left panel shows the experimental results at ambient temperature (25^oC), indicating a lower percentage of unexchanged Hydrogens in the mesophilic protein as compared to its thermophilic counterpart, pointing that the latter is rigid at ambient conditions at any measured experimental time. The right panel shows the convergence of spectra when the experiment is repeated at the optimal working temperatures of the two enzymes, indicating that both proteins exhibit comparable flexibility at their respective optimal working temperatures. Figure is taken from Ref [59].

is directly related to its function and it should be considered in terms of temperature dependence. Somero has proposed a corresponding state principle [57, 56], where the critical degree of protein flexibility is identified as imperative to function, suggesting flexibilities of proteins at their optimal working temperatures to be comparable. This further implies that the shift of thermal stability and activity in thermophiles is due to enhanced mechanical rigidity of the protein matrix recovering the critical flexibility at high temperatures.

The universality of the flexibility/rigidity paradigm was investigated and questioned along the years both experimentally [59, 42, 60] and theoretically [61, 62, 63, 45, 64, 65].

Early Hydrogen/Deuterium (H/D) experiments probing the exposure of amide groups to solvent by local unfolding revealed agreement between the protective factors of mesophilic and thermophilic proteins at their respective optimal working temperatures, as shown in Figure 1.3. [59, 53, 60]. Conversely, the same method did not validate the corresponding state principle for the very thermostable enzyme rubredoxin from *Pyrococcus furiosus* [42]. Similarly, Small Angle Neutron Scattering studies probing fast mode atomistic fluctuations on two Malate Dehydrogenases did not verify the corresponding state as the thermophilic homologue was found to manifest larger atomistic displacements than the mesophilic variant [66]; it should also be mentioned that the microscopic interpretation of the experimental signal was later questioned [67]. In other cases, the principle was proven to be valid for the thermobarophilic protein IF-6 in a combined Neutron Scattering and molecular modeling study [68]. In Chapter 3, we challenge the corresponding state principle by considering the case of two homologous catalytic domains of the elongation factor EF; one from *E. coli* and the other from *S. solfataricus*. The enzymatic turnover is mimicked by simulating the members of the catalytic cycle, and we observe a similar release of conformational flexibility when transitioning from the reactant to the product state of the enzymes at their respective optimal working temperatures, effectively proving the corresponding state principle.

Finally, the issue of the lack of thermophilic activity at moderate temperatures should be addressed. In addition to the lower activity at low temperatures predicted by the exponential kinetics, it has been shown that important functional modes are quenched at ambient temperatures, only to be activated in the high temperature regime [69, 33]. Interestingly, long timescale modes have shown to be dominated by pico- to nanosecond dynamics of loops in an illustrative example for Adenlyate Kinases [70], showing that the thermal activation of functional modes can be accessed in Molecular Dynamics studies. Similar observations on Malate Dehydrogenase have been made, where the propagation of domain interfacial constraints on the nanosecond timescale was observed in the loop gating the substrate binding [71] at ambient conditions, while the simulations at high temperature could not probe the thermal activation of the corresponding mode. The corresponding state scenario related to the activation of modes and chemical reaction was addressed in a number of studies. Hydrogen tunnelling contribution to the chemical step of the enzymatic reaction has found to be comparable at the enzyme working temperatures in the mesophilic and thermophilic Alcohol Dehydrogenase, although the rationale behind the thermal activation of the tunnelling in the thermophilic enzyme still remains elusive [72, 60, 73]. Intense focus has also been placed on homologous Dihydrofolate Re-

ductases [74, 75, 76], where additional complexity in comparing temperature-dependent activities is introduced owing to the fact that the mesophilic enzyme is monomeric, while the active form of the thermophilic variant is dimeric. The dimeric state is suggested to filter accessible conformations by constraining the catalytically important loops at the dimer interface, which is thought to prevent the electrostatic preorganization of the active site [73] and the consequent lowering of the kinetic barrier for the enzymatic reaction [76, 62]. To probe thermal activation of functional modes, in Chapter 4 we present a combined Spin-Echo Neutron Scattering and Molecular Dynamics study on the mesophilic rabbit muscle 5 Lactate Dehydrogenase, where a thermal activation of a mode at lengthscales corresponding to interdomain separations was observed at temperatures corresponding to the optimal working temperatures of the enzyme. The study looks into the allosteric communication network and the implications the activated mode pertains to enzymatic activity. The final goal will be a parallel study on Malate Dehydrogenase, which is structurally homologous to Lactate Dehydrogenase.

The studies on Somero's principle yield results on a case-to-case basis, while the universal conclusion seems to be amiss. In fact, this should come as no surprise taken into account the numerous factors affecting thermal activity and stability. It is also important to note that the timescales and lengthscales of protein dynamics are correlated, making the definition of flexible and rigid vague and dependent on the dynamical range of interest, consequently offering a possible explanation for the diversity of presented results. In this thesis, the concept of flexibility will be mostly inspected in terms of conformational flexibility, while additionally the atomistic fluctuations will also be addressed.

We proceed by first familiarizing the reader with the methods in Chapter 2, and continue, in Chapters 3, 4, and 5, with detailed discussions on the ideas and challenges presented here, only to conclude and pave future paths in Chapter 6.

CHAPTER 2

METHODS

Numerical simulations and calculations are used as the main method of choice. Namely, the Molecular Dynamics (MD) and its extension to achieve better sampling of molecular configurations, Replica Exchange with Solute Scaling (REST2), are employed extensively to obtain simulations of systems on a microsecond timescale. The Molecular Dynamics produces trajectories of atomic positions in time and, by exploiting Statistical Mechanics, ensures the calculation of equilibrium and dynamic properties such as entropy estimates and diffusion coefficients. Moreover, the technique allows the observation of biologically relevant events at an atomistic resolution, making it invaluable to modern research. To further strengthen our studies, we support our *in silico* results with Neutron Scattering (NS) experiments performed on the same systems, as MD and NS probe similar ranges of length and time scales. The possibility of determining quantities observed in NS experiments from the simulated trajectories makes the combined use of the two techniques a solid tool in interpreting the structural and dynamical studies of proteins. The experimental techniques are also covered in the sections below, along with the description of simulation and analysis tools.

2.1 Molecular Dynamics Simulations

Molecular dynamics (MD) is one of the principal computational techniques, the other being the Monte Carlo method, used to investigate condensed matter *in silico*. MD has been especially successful in the study of liquids, soft matter systems, including

biomolecules, and contributed to reinforce the role of computer simulations as the true third pillar in the scientific effort by partnering experiments and theory [77].

The idea of MD is simple, once a molecular model for the system is provided, i.e. the classical potential energy describing the interactions between the constituents of the system, the dynamics of the system is numerically solved by integrating the equations of motion of classical mechanics. These trajectories are useful in that, in the ergodic limit, the calculated time averages correspond to the true thermodynamic average. MD is powerful because it also provides equilibrium transport properties, i.e. diffusion coefficients, as well as monitoring kinetic relaxation upon an out-of-equilibrium perturbation. In the specific context of biomolecular simulations all this has an important impact, having granted for example the quantification of the distribution of the conformational states accessed by a protein, the mobility of ligands toward a target binding site, and the relaxation of the protein matrix upon photo excitation or charge separation processes [78, 79, 80]. Depending on the choice of equations of motion used in simulations, the protein in a MD trajectory samples microstates corresponding to a particular thermodynamic ensemble. The simplest example is the integration of Newton's equations of motion, which produces a trajectory that samples the NVE ensemble, keeping the number of particles N, the volume V, and the energy E of the system constant. Additionally, reservoirs of heat and particles, as well as pistons, can be coupled to the simulated system to keep different quantities constant during simulations, producing trajectories that sample different statistical-mechanical ensembles.

The Molecular Dynamics simulations consists of:

- model describing the interparticle potential (the force-field)
- calculations of energies and forces from the model
- integration of the equations of motions

The method is routinely applied in biomolecular simulations [81, 82] and is used to study proteins, nucleic acids, and drug-like molecules. Due to constant efforts to improve the efficiency of the simulation algorithms and increase in computing power resources, e.g. the use of specialized machines as done by the D.E. Shaw laboratory or GPUs, MD all-atom simulations of very large systems, e.g. the viral capsides, or processes on a long timescale such as the millisecond folding, are now performed [83, 84, 85]. The algorithms and the Statistical Mechanics underlying MD simulations are provided in many classic textbooks [86, 87, 88] to which the reader can refer. We only add that the MD technique

is frequently coupled to enhanced sampling techniques that allow the estimation of the free energy for complex processes (e.g. the binding of a ligand) or exploration of the conformational landscape of a biomolecule in an efficient manner. MD can also be used for systems with a grained resolution, coarse-grained MD, or coupled to explicit quantum-mechanics as in *ab initio* MD.

In our work, we have used the NAMD [89] simulation package, found to be particularly efficient in performing calculations on massively parallel supercomputers. The algorithms, as used in the simulations, are described in the remainder of the section.

2.1.1 Molecular Force Fields

The molecular models used in classical simulations are termed force-fields [90], and they describe the potential energy of a protein as a function of particle coordinates, with coefficients derived from Quantum-Mechanical calculations and experiments. This approach is based on the Born-Oppenheimer approximation, which assumes that the electrons always reach their ground state in the timescale of nuclei movement due to large mass differences between the two (nuclei are 3-4 orders of magnitude heavier), consequently justifying the description of both with a single potential energy function used in the framework of the classical system description.

In describing the interparticle interactions, the potential is decomposed to a sum of terms, each term representing a contribution to the overall protein energy. The energy terms comprise only of pairwise interactions, a crude representation of reality in which many particles interact simultaneously. Nevertheless, the parameters of the pairwise additive potentials are modified so as to account for the multi-body interactions and finally yield a good agreement with different bulk properties used in parameter fitting.

Many models have been developed in attempts to correctly describe different aspects of protein behavior, including models that treat each atom separately, all-atom force-fields, and models treating collection of particles, coarse grained models. In all-atom models, every atom is a center of force and represented with three Cartesian coordinates. Among the all-atom models, which are the most accurate and at the same time the most computationally demanding, different approaches have been developed, the most commonly used in biomolecular protein simulations are CHARMM [91], AMBER [92], and OPLS [93]. Their main differences are in the treatment of planarity and chirality in functional groups, as well as in different ways of obtaining non-bonded atom-pair parameters through different recombination methods.

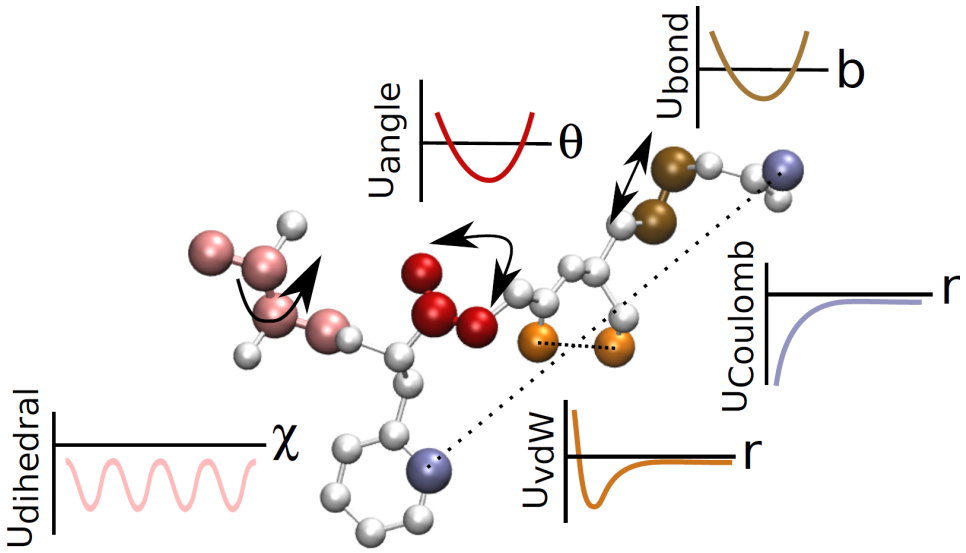

Figure 2.1: Schematic representation of important potential energy terms in a protein and the potentials with which they are modeled.

We generally use the all-atom CHARMM22/CMAP force-field [91, 94] in our simulations because it has been employed on a large number of proteins, and has shown to predict both the structural [91, 95, 96, 97] and the dynamical [98] properties of the proteins in reasonable limits. It has been described in detail in the section below. Occasionally we use a modification of this force-field, the CHARMM36 [99].

2.1.1.1 CHARMM22/CMAP

The CHARMM protein potential is given by two main terms, the bonded and the non-bonded potential energy functions:

$$U = U_{bonded} + U_{nonbonded}. \tag{2.1}$$

Some interactions comprised in the model are depicted in Figure 2.1. The parameters of the bonded terms were parametrized to reproduce X-ray structure geometries, infrared and Raman spectra, and *ab initio* calculations. The non-bonded interactions were fit to reproduce *ab initio* interaction energies and geometries between protein polar residues and water molecules, and empirical data, i.e. heats of vaporization and molecular volumes.

The bonded terms account for bond stretching, angle bending, and dihedrals. Another three terms, Urey-Bradley, improper angle term, and the CMAP term, are added to

obtain good agreement with experimental results:

$$U_{bonded} = U_{bond} + U_{angle} + U_{dihedral} + U_{Urey-Bradley} + U_{improper} + U_{CMAP}. \tag{2.2}$$

The bond stretching between two linked atoms is described by a harmonic potential:

$$U_{bond} = \sum_{bonds} K_b(b - b_0)^2, \tag{2.3}$$

where K_b is the spring constant, b is the bond length, b_0 is the equilibrium bond length, and the sum runs over all the linked pairs of atoms.

An equivalent harmonic potential is used to described the angle bending between three linked atoms:

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta_0)^2, \tag{2.4}$$

where K_θ is the angle bending modulus, θ the angle, θ_0 is the equilibrium angle, and the sum runs over the triplets of atoms.

The dihedral angle contribution is necessary to describe the rotation around a given bond. In order to account for multiple minima, a Fourier expansion is generally used in the form of:

$$U_{dihedral} = \sum_{dihedrals} K_\chi[1 + cos(n\chi - \sigma)], \tag{2.5}$$

where the sum runs over quadruples of atoms and describes the dihedral angle rotation, K_χ is the dihedral rotation constant, n accounts for the cosine multiplicity, χ is the dihedral angle, and σ is the phase.

The three previously described interactions arise naturally from description of chemical connectivity of the molecule. As the force-fields are approximate, additional terms are included to improve accuracy of the model and agreement with experimental data. The Urey-Bradley term is a harmonic correction to distance between the two non-bonded atoms in an angle (atoms 1 and 3), applied to some angles in the force-field to better reproduce the in-plane deformations and to separate symmetric and asymmetric bond-stretching modes in vibrational spectra.

$$U_{Urey-Bradley} = \sum_{UB} K_{UB}(S - S_0)^2, \tag{2.6}$$

where the sum runs over 1, 3 atoms in the angle, K_{UB} is the Urey-Bradley force constant, S is the 1,3 atom distance, and S_0 is the equilibrium 1,3-distance.

The improper dihedral term aids to reproduce the out-of-plane modes in the vibrational spectra, and its role is to restrict protein geometry, e.g. to maintain group

chirality:

$$U_{improper} = \sum_{impropers} K_{imp}(\phi - \phi_0)^2, \tag{2.7}$$

where ϕ is a dihedral angle, ϕ_0 its equilibrium value, and K_{imp} the associated force constant. Finally, the CMAP correction is applied to improve the protein backbone conformation accuracy. It is a cross term applied to dihedral angles:

$$U_{CMAP} = \sum_{residues} u_{CMAP}(\Psi, \Phi), \tag{2.8}$$

where the u_{CMAP} is the energetic correction map (CMAP) term for the dihedral pair Ψ, Φ, applied to correctly reproduce free energy surfaces of dipeptides obtained from QM calculations. Further modifications have been made to improve this CMAP correction (see subsection on CHARMM36 [99] below). Another set of corrections has been applied to all residues other than Gly and Pro in a release termed CHARMM22* [97]. The latter has been successfully used to fold small proteins in all-atom simulation [85], and contains a better description of salt bridges as well as better torsion parameters for Asp.

The non-bonded term in the force-field accounts for the van der Waals and electrostatic interactions between atoms that are separated by at least three bonds:

$$U_{non-bonded} = \sum_{i<j} \left(\epsilon_{ij} \left[\left(\frac{R^0_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R^0_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right). \tag{2.9}$$

The term on the left accounts for van der Waals (vdW) interactions, i.e. induced dipole interactions, and is modeled by the Lennard-Jones potential, while the right term accounts for electrostatic interactions. The short-range Lennard-Jones potential is described with two parameters, ϵ_{ij} determining the potential energy well depth and R^0_{ij} describing the minimum energy interparticle length. The two parameters are generally provided for a homogeneous interaction, that is between atoms of the same type. However, they can be mixed in order to obtain the parameters describing the interactions between atoms of different types, using e.g. the Lorentz-Berthelodt recombination rules; $\epsilon_{ij} = \sqrt{\epsilon_i + \epsilon_j}$, $R^0_{ij} = \frac{R^0_i + R^0_j}{2}$. The Lennard-Jones potential itself contains two terms, the right-side $\left(1/r_{ij}\right)^6$ term represents the attraction between two induced dipoles and dominates at higher interparticle distances, while the left-side $\left(1/r_{ij}\right)^{12}$ term accounts for the repulsion between two atoms coming into close contact, obeying the Pauli exclusion principle. The power 6 comes from analytical theory, while the power 12 is chosen due to computational efficiency, as it is merely a square of the $\left(1/r_{ij}\right)^6$ term; nonetheless it provides a steep energy rise with decreasing r_{ij}. It is important to note that the vdW

term is a short range interaction, which is used to improve the computational efficiency of the method, as shown later in text.

The second term in the non-bonded interaction represents the Coulomb interactions between particles with partial charges q_i and q_j, separated by the distance r_{ij}. As the electronic cloud is not treated explicitly in these force-fields, these charges represent the overall charge distribution of the molecule with effective partial charge values, adding up to a total molecular charge. By using fixed partial charges, the electronic polarization is not accounted for, and specialized force-fields have been recently developed to mend this deficiency [100]. Moreover, the hydrogen bond interactions, essential to structural organization of the biomolecules, arise naturally from the electrostatic and the Lennard-Jones term [101] and are consequently not represented via an explicit term.

In the case of soluble proteins, the biomolecule is simulated in an aqueous solution. The CHARMM force-field was parametrized in combination with a specific water model, the TIP3P model, a 'three point' (3P) model with three interaction sites placed on respective water atoms [102]. Three partial charges are distributed on the respective atoms of the molecule. To simplify the water model with the goal of increasing computational efficiency, the number of degrees of freedom of the molecule has been reduced as the bonds are treated as rigid. This model reproduces first-shell hydration and other important liquid properties, while suffering from a high diffusivity due to inaccuracies in the tetrahedrality of the hydrogen-bond connectivity [102]. Nonetheless, the model presents a good trade-off of computational cost against reproduced thermodynamics accuracy, and was thus used in parametrizing the CHARMM22/CMAP force-field [91]. For the same reason, it presents an ideal water model in our simulations, although more precise water models with different parametrization, e.g. SPC/E [103] or more interaction sites [104] have been developed.

2.1.1.2 CHARMM36

The CHARMM36 force-field [99] is an extension to the CHARMM22/CMAP. In the CHARMM36, the potential energy describing the protein backbone is redefined, as well as the new side chain dihedral parameters. The CMAP is added so as to improve the helical bias of the CHARMM22/CMAP (see Subsection 2.1.1.3). The CMAP was modified for the non-Gly and non-Pro residues in order to fit NMR experiments and QM energy surfaces calculated at a high level of theory. Side chain dihedrals were modified according to the QM energy surfaces. The parameters for aliphatic hydrogens were revised, as were the sidechains of Arginine and Tryptophane. CHARMM36 is supposed to yield a better

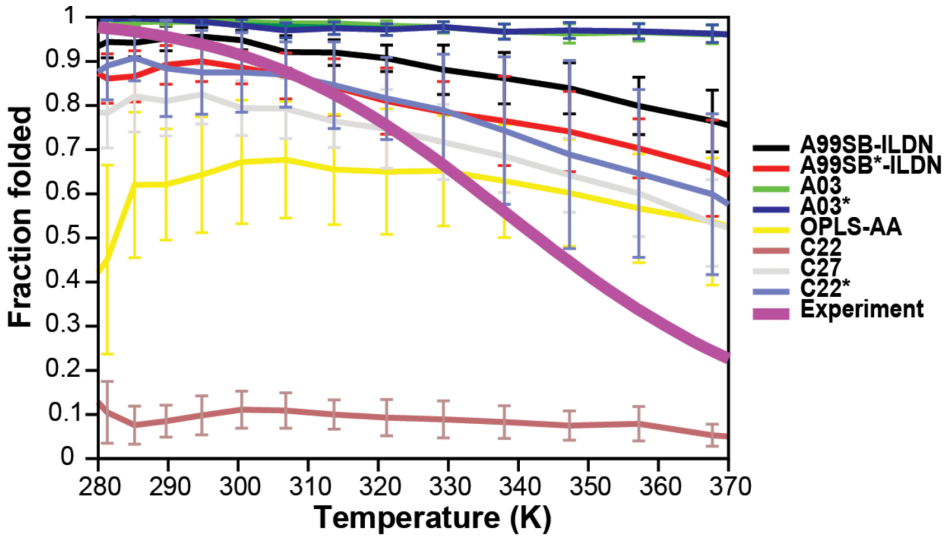

Figure 2.2: The fraction of folded structures for the CLN02 peptide, forming preferentially a hairpin-like structure below T=340 K, as a function of temperature, data show a comparison between the experiment and eight different force-fields. The largest outlier is the CHARMM22 force-field, which possesses a strong helical bias and thus renders the peptide unfolded at all temperatures. Introducing the CMAP correction in CHARMM22/CMAP (colloquially termed CHARMM27) improves the ratio of folded to unfolded state. Nonetheless, the stability curve is broadened in all force-fields, most likely due to lack of proper representation of protein cooperative behavior and many-body interactions. Taken from Ref [96].

description of hydrogen bonds in proteins, in both the α and β secondary structure, and better long chain description due to torsion angle side chain improvements, as well as an improved balance of secondary structure elements. On the other hand, the flexibility of the protein backbone is increased when compared to both CHARMM22/CMAP and the experiments [105]. It has also been suggested that the improved CMAP correction implicitly includes many body effects, witnessed by improved cooperativity in secondary structure transitions [106]. However, this novel version should be tested against a variety of study-cases to demonstrate its performance and possible deficiencies.

2.1.1.3 Force-Field Limitations

A major limitation in MD simulation accuracy is the force-field as it determines all atomic interactions. The limitations arise from attempts to reproduce certain properties e.g. dihedral angles, secondary structure, spectroscopic features, while neglecting or simplifying others such as polarization or structural cooperativity.

Protein behavior in a simulation is necessarily force-field dependent, but artifacts can be avoided by choosing the model most suitable to maintain structural properties of the simulated protein and to better represent the property of interest. For most proteins, the folding free energies are small, and the errors in the force-field development could easily create an artifact of the minimum free energy state. The error comes from both the backbone and the side chain, consequently modifying only the backbone potential is not enough for accurate representation. Moreover, the conformational substates sampled in a typical simulation depend on the force-field [96] as do the kinetic and thermodynamic stabilities of these substates. This results in different convergence kinetics of protein conformational transitions in simulations and different folding/unfolding pathways [96].

Accumulation of small errors and approximations can best be seen in higher-order structures, such as secondary structure elements, where hydrogen bonds and backbone torsions drive the formation of α helices, β sheets, and other elements. Force fields have widely been reported to have a secondary structure propensity; CHARMM22/CMAP, OPLS-AA/LL, and AMBER03 overstabilize the α secondary structure [96, 107], while AMBER99SB-ILDN overstabilizes β sheets [108, 96]. Consequently, modifications to these force-fields; CHARMM36, CHARMM22*, AMBERff03*, and AMBER99SB*-ILDN, have been developed to improve secondary structure sampling, although further efforts need to be made to remove the limitations. The secondary structure stabilization artifacts can lead to shifting the protein stability at different temperatures. Current force-fields cannot reproduce the melting curves of the proteins in classical brute force Molecular Dynamics simulations [96] due to sampling problems and failure in representing co-operative protein behavior from pairwise interactions and poor many-body interaction inclusion, see Figure 2.2.

Convergence of simulations could be ensured by observing multiple transitions on the longest timescale, i.e. that of the protein unfolding. These are still in the microsecond regime, while the pathways themselves are force-field specific [85]. Reversible folding and unfolding at experimental temperatures in brute force simulations has so far been performed for α-helical bundle and β-sheet small proteins only in CHARMM22* and AMBERff03* [96]. The force-field parameters are derived mostly from data on folded structures, which could again introduce an artifact in representing the unfolded state. For example the radius of gyration of the unfolded state has been reported to be systematically lower in computer simulations using AMBER [109], CHARMM22/CMAP, and CHARMM36 force-fields [109, 108] when compared to experimental data. Additionally, the force-field parameters reproduce properties in the temperature range T=300-330 K,

possibly creating artifacts in high-temperature simulations. As the solvent has a central role in protein (un)folding, the temperature dependence of solvent properties is a relevant parameter when considering potential artifacts. Different properties of the TIP3P water were inspected, with the main finding on the maximum density temperature, signature of the anomalous water behavior. While water density at 298 K agrees well with the experimental result, the density of TIP3P water decreases monotonically with temperature increase throughout the simulated temperature interval [-223, 373] K [110], failing to reproduce a maximum density at any temperature, and exhibiting lower densities at temperatures >298 K.

Other artifacts arise from the need to make the potential energy function simple and computationally effective. A good example is omitting the effect of polarization of the electronic cloud arising from interaction with its environment. The effect of polarizability ranges from determining the proper hydrogen bonding geometries to accounting for $10-20\%$ of the ligand-binding energies [111]. Polarizable models developed so far [112, 113, 114] already reproduce protein properties at least as good as their non-polarizable counterparts [100], and are expected to surpass them with increased interest in their use and improved parametrization. The caveats are that a larger number of parameters per atom is necessary, making the parametrization more difficult, and sampling problems, as the simulation times currently do not extend over 1 μs even for small proteins. Naturally, implicitly included hydrogen bonds and non-polarizable isotropic monopoles included in typical protein force-fields make these properties force-field dependent. The variations in the native hydrogen bond content and the number of ionic interactions are reported among different models [108].

2.1.2 Force Evaluation

Once the potential energy function to describe the molecule is chosen, evaluating the forces is theoretically straightforward. For an atom i in the system, the force $\mathbf{F}_i$ is computed as:

$$\mathbf{F}_i(\mathbf{r}_1 ... \mathbf{r}_N) = -\nabla_i U(\mathbf{r}_1 ... \mathbf{r}_N), \tag{2.10}$$

where $\nabla_i = \frac{\partial}{\partial \mathbf{r}_i}$. The simulations are performed in a box, containing protein and solvent (water) molecules. The simulation however, is performed under Periodic Boundary Conditions (PBC), which means that a particle leaving the simulation box re-enters it at the opposite face of the box. This is necessary to eliminate surface effects as the ratio of numbers of surface particles and total particles is large when compared to macroscopic

systems with $N = 10^{23}$ particles. By exploiting the spatial periodicity of the PBC it is also possible to apply efficient algorithms in long-range electrostatic interaction calculations and to maintain constant pressure conditions (see more below).

While the force evaluation is a straightforward procedure, the computational cost associated to it represents the bottle-neck in computational efficiency. The number of force evaluations for the bonded terms are relatively small compared to the $1/2N(N-1)$ number of terms between N particles. To overcome this, several methods have been developed and are routinely made use of in our simulations. They rely on the fact that the Lennard-Jones potential is short-range and can be cut off beyond a certain distance, while the Coulomb interactions have long-range effects that need to be treated, and other solutions need to be found.

The short-ranged van der Waals interactions are cut off typically beyond a distance $r_c = 10$ Å. By using the minimum image convention, each particle interacts with its nearest periodic image under the PBC. This puts restrictions on the value of the cut off, which must be $r_c \leq L/2$, where L is the shortest among the box edges. Instead of making the interactions beyond this distance abruptly zero, which would lead to discontinuities when particles' distance passes through $r = r_c$, a smoothing function is applied, termed 'switching function', that smoothly truncates the potential to zero, starting to take effect at distances $\leq r_c$. Even with applying a cut off, a large number of interparticle distance evaluations need to be made in order to determine which particles are within the cut off distance. This is resolved by using lists of interparticle pairs within a cut off, that are only periodically updated, eliminating the need to evaluate all interparticle distances at every time step of the simulation.

The evaluation of the electrostatic forces is the most time consuming, since the interaction is long-range in nature. An efficient treatment in periodic systems is based on the Ewald scheme, and further improved by the Particle Mesh algorithm (PME). In the Ewald scheme, the electrostatic interactions are separated into a short-range and a long-range contributions. The short-range local contributions are evaluated in the same manner as the van der Waals interactions, at every time step of the simulation within a certain cut off and a switching function to smooth out the truncation. Thus the local electrostatic and van der Waals interactions are computed with a direct calculation:

$$U_{short}(\mathbf{r}_1 ... \mathbf{r}_N) = \sum_{\mathbf{S}} \sum_{i<j}^{*} \left\{ \epsilon_{ij} \left[\left(\frac{R_{ij}}{r_{ij,\mathbf{S}}} \right)^{12} - \left(\frac{R_{ij}}{r_{ij,\mathbf{S}}} \right)^{6} \right] + \frac{q_i q_j \mathrm{erfc}(\alpha r_{\mathrm{ij,S}})}{r_{ij,\mathbf{S}}} \right\}, \tag{2.11}$$

where $\mathbf{S} = \mathbf{hm}$ for a box matrix $\mathbf{h}$ and $\mathbf{m}$ is a vector of integers, α is a constant with units of 1/length, and $\mathrm{erfc}(\alpha r_{ij,\mathbf{S}})$ is the error function that goes to zero when $\alpha r_{ij,\mathbf{S}} \to \infty$.

The sum runs over all the non topologically bonded (*) pair atoms $i < j$.

The long-range electrostatic interactions are evaluated by using the smooth Particle Mesh Ewald algorithm [115], with the computational cost proportional to $NlogN$. As anticipated, the core of this method is the Ewald summation [116]. It performs the calculation in the reciprocal space as the long-range interactions become short-range in the reciprocal space composed of reciprocal-space vectors $\mathbf{g} = 2\pi\mathbf{n}/L$:

$$U_{long} = \frac{1}{V}\sum_{\mathbf{g}\in S}\frac{4\pi}{|\mathbf{g}|^2}e^{-|\mathbf{g}|^2/4\alpha}|S(\mathbf{g})| - \sum_{bonded i,j}\frac{q_i q_j \mathrm{erf}(\alpha r_{ij})}{r_{ij}} - \frac{\alpha}{\sqrt{\pi}}\sum_i q_i^2, \tag{2.12}$$

where the two subtracted terms represent, respectively, the electrostatic interaction between bonded particles already accounted for in the bonded terms of the potential, and the self interaction of the particle with its periodic images. The first term includes these interactions, thus the need to explicitly subtract them. S is a hemisphere of $g \geq 0$ and the Fourier sum is truncated at $|\mathbf{g}| \leq g_{max}$, where g_{max} is chosen so that $e^{-g_{max}^2/4\alpha^2}$ is negligible, and the truncation is possible due to $\mathbf{g} \sim 1/r$. α is a parameter of inverse length, $\mathrm{erf}(\alpha r_{ij})$ is the error function that ensures the condition $r \to \infty, \mathrm{erf}(\alpha r) = 1$. $S(\mathbf{g}) = q_i \sum_i e^{i\mathbf{g}\cdot\mathbf{r}}$ is the structure factor and becomes computationally demanding as the system size increases due to necessary increase in the number of $\mathbf{g}$-vectors that need to be taken into account. The smooth Particle Mesh Ewald simplifies the structure factor term by projecting the charges on a uniform lattice. The lattice-spacing controls the accuracy of the calculations, and for a dense molecular system, its magnitude is in the order of the characteristic atomic length scale, 1 Å. The mathematical properties of the B-splines used for charge interpolation over the lattice ensure the smoothness of the potential and simplify the calculation further. Please refer to Ref [115] for the expression of the structure factor in the smooth PME method. The use of the Ewald sum requires the system to be neutral and all our simulations have counter-ions (Na^+ or Cl^-) added to them so that the total charge of the system is zero.

2.1.3 Integration of Equations of Motion

Having described the calculation of forces acting on individual particles on the basis of the analytical potential energy functions, we proceed with a description of system's time evolution according to the laws of classical mechanics. Prior to presenting the algorithms in use, it is worth mentioning a practical limitation concerning the system initial conditions necessary for the numerical solution of the equations of motion. In simulating liquids, initial configurations can be obtained by generating random states,

while the simulations of proteins in their folded state require the knowledge of their structure. This is taken, when possible, from the databases where the X-ray or NMR resolved structures are publicly available. When the structure is not available, an initial guess can be produced via homology modeling or other empirical methods that predict the folded state on the basis of sequence, however the reliability of these remains uncertain.

Generally, prior to simulating the system, a minimization (conjugate-gradient) procedure is performed so as to remove the steric clashes in the system. The minimization procedure brings the system to a local minimum found close to the initial state of the protein structure. Finally, a set of equations of motions is used to describe the system under particular set of conditions and these are subsequently integrated to produce a trajectory.

In the thermodynamic limit of very large systems, $N \to \infty$, the fluctuations of non-constant quantities in different ensembles become negligible, and their thermodynamic averages converge. In practice, the simulations are run on systems of finite size, with the fluctuation decay scaling as $\sqrt{N}$. For this reason, the choice of the ensemble is important, particularly if the results are to be compared with experiments. Our simulations are performed in the NpT ensemble as we often compare them to experiments performed in equivalent conditions. To sample this ensemble, an appropriate set of equations of motions must be chosen, one that couples a numerical heat bath and a numerical piston to the system. The former ensures an exchange of heat and thus acts as a thermostat, while the latter expands and compresses the volume cell so as to maintain pressure at desired value. The coupling is stochastic, having the advantage of shortening particle correlation times, enhancing sampling efficiency, and stimulating ergodicity. Additionally, the constant pressure control requires Periodic Boundary Conditions.

The coupling between the heat bath and the system is achieved by the Langevin-Hoover method [117], a combination of the Nose-Hoover constant pressure method [118] and the piston fluctuation control adopted from the Langevin Dynamics [119]. To achieve proper NpT sampling, the system volume is an additional degree of freedom in the equations of motions, along with the particle positions and the momenta:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m_i} + \frac{p_\epsilon}{W}\mathbf{r}_i, \tag{2.13}$$

where $\mathbf{r}_i$, $\mathbf{p}_i$, and m_i are atomic positions, momenta, and masses of atom i, respectively. p_ϵ is a momentum associated to strain rate of the volume, $p_\epsilon = \dot{\epsilon} W$, where $\epsilon(t) = \frac{V(t)-V_0}{V_0}$. Initial position in the simulations are the Cartesian coordinates and the initial velocities are obtained by sampling the Maxwell-Boltzmann distribution. W is a fictitious mass

of the piston, defined as $W = 3N\tau^2 k_B T$, where N is the number of particles and τ is the oscillation period. Longer oscillation period slows down the cell volume fluctuations, and it is generally set to ~ 100 fs. The cell vectors oscillate isotropically:

$$\dot{\mathbf{V}} = \frac{3V p_\epsilon}{W}. \tag{2.14}$$

The stochastic coupling is provided by adding a dissipative and a random term to conservative forces derived from the potential:

$$\dot{\mathbf{p}}_i = -\frac{\partial U}{\partial \mathbf{r}_i} - \frac{p_\epsilon}{W}\mathbf{p}_i - \gamma \mathbf{p}_i + \mathbf{R}_i, \tag{2.15}$$

where γ is the damping constant, usually set to 1 ps^{-1}, and $\mathbf{R}_i$ is a stationary Gaussian process with the following properties $\langle R \rangle = 0$, $\sigma^2 = \sqrt{\frac{2k_B T \gamma m}{\Delta t}}$, where k_B is the Boltzmann constant, T is the target temperature, m is the molecular mass, and Δt is the time step. Finally, the momentum associated to the cell volume compressibility evolves in time as:

$$\dot{\mathbf{p}}_\epsilon = 3V(P - P_{target}) - \gamma_p p_\epsilon + R_p, \tag{2.16}$$

where γ_p is the barostat damping coefficient usually set to 0.02 fs^{-1}, R_p is the stochastic force acting on the barostat with zero mean and variance $\sigma^2 = \sqrt{\frac{2k_B \gamma_p W}{\Delta t}}$, and P_{target} is the target pressure set to 1.01325 bar. P is the pressure similar to instantaneous pressure, without the contribution of the Langevin thermostat white noise, $P = \frac{1}{3V}[\sum_{i=1}^{N} \frac{\mathbf{p}_i \cdot \mathbf{p}_i}{m_i} + \sum_{i=1}^{N} \mathbf{r}_i \cdot \mathbf{F}_i] - \frac{\partial U}{\partial V}$.

The positions, momenta, volume, and volume strain momentum integrations schemes are derived from the the equations of motion by using the classical propagator and the Trotter factorization scheme [120], which ensures the multiple-timescale integration:

$$V^{t+1/2\Delta t} = V^t + \frac{\Delta t}{2}\dot{V}(V^t, p_\epsilon^t), \tag{2.17}$$

$$p_\epsilon^{t+1/2\Delta t} = p_\epsilon^t + \frac{\Delta}{2}\dot{p}_\epsilon(\mathbf{r}_i^t, \mathbf{p}_i^t, V^{t+1/2\Delta t}, p_\epsilon^t, \gamma_p p_\epsilon^t, R_p^t), \tag{2.18}$$

$$\mathbf{p}_i^{t+1/2\Delta t} = \mathbf{p}_i^t + \frac{\Delta t}{2}\dot{\mathbf{p}}_i(\mathbf{r}_i^t, \mathbf{p}_i^t, p_\epsilon^{t+1/2\Delta t}, \gamma \mathbf{p}_i^t, \mathbf{R}_i^t), \tag{2.19}$$

$$\mathbf{r}_i{}^{t+\Delta t} = \mathbf{r}_i^t + \Delta t \dot{\mathbf{r}}_i(\mathbf{r}_i^t, \mathbf{p}_i^{t+1/2\Delta t}, p_\epsilon^{t+1/2\Delta t}), \tag{2.20}$$

$$\mathbf{p}_i^{t+\Delta t} = \mathbf{p}_i^{t+1/2\Delta t} + \frac{\Delta t}{2}\dot{\mathbf{p}}_i(\mathbf{r}_i^{t+\Delta t}, \mathbf{p}_i^{t+1/2\Delta t}, p_\epsilon^{t+1/2\Delta t}, \gamma \mathbf{p}_i^{t+\Delta t}, \mathbf{R}^{t+\Delta t}), \tag{2.21}$$

$$p_\epsilon^{t+\Delta t} = p_\epsilon^{t+1/2\Delta t} + \frac{\Delta t}{2}\dot{p}_\epsilon(\mathbf{r}_i^{t+\Delta t}, \mathbf{p}_i^{t+\Delta t}, V^{t+1/2\Delta t}, p_\epsilon^{t+\Delta t}, \gamma_p p_\epsilon^{t+\Delta t}, R_p^{t+\Delta t}) \tag{2.22}$$

$$V^{t+\Delta t} = V^t + \frac{\Delta t}{2}\dot{V}(V^{t+1/2\Delta t}, p_\epsilon^{t+\Delta t}). \tag{2.23}$$

Multiple-timescale integration relies on the fact that protein motions span many different timescales, which is exploited to improve calculation efficiency. The forces derived from the potential will typically be fast-varying for intramolecular terms and slow-varying for the large-scale nonbonded terms. Additionally, the bonded intermolecular terms are computed quicker than the non-bonded terms, because for the latter, the interactions are computed between many particle pairs. Consequently, the integration in NAMD is separated in three parts where different forces are exploited in advancing the system; bonded forces are used with a small integration time step, non-bonded forces and short-range electrostatic forces are integrated with a moderate time-step, and finally long-range electrostatic forces are coupled to the largest integration time-step. This is achieved by using the Reference System Propagator Algorithm (RESPA) [120]. RESPA stems from separating the Liouville classical propagator to its fast and slow component, which yields a numerical integration scheme with separated integration time scales, as shown above. The shortest time step used in the simulations is 2 fs. This time step is relatively large for biomolecular simulations, and can be used on account of fixing the hydrogen bond length and angle to their nominal value defined in the force-field, achieved by use of the SHAKE algorithm [121].

Integrating the equations of motions produces new particle positions, particle velocities, and system volume sizes. The particle positions, potential energies, volume, and pressure are saved in simulation snapshots of the trajectory and are used in subsequent analysis. The use of Statistical Mechanics ensures that many thermodynamic properties can be recovered from the simulation snapshots, making the method a valuable quantitative tool that either complements experiments or stands alone in cases where, due to experimentally unachievable resolutions or conditions, simulations offer an excellent solution. The main drawbacks of the methods is that the equilibrium properties are only obtained under ergodic assumption, which cannot be proved to be valid in biomolecular simulations as many protein conformations are separated by high energy barriers, and the protein most likely does not visit all possible conformational substates in the time course of a trajectory. Additionally, the simulations are highly dependent on the initial conditions, and they diverge quickly even for small differences in initial conditions, which is quantified as the Lyapunov instability. Nevertheless, being familiar with sources of error in simulations and choosing proper conditions ensures the correct interpretations of results and makes MD an invaluable, atomic-level resolution tool.

2.2 REST2 Molecular Dynamics

A major problem in MD simulations is the proper sampling of the potential energy surface, i.e. producing equilibrium trajectories. The problem of sampling arises from the height of energy barriers between many major conformational states. The time needed for a thermal fluctuation of atoms to increase to a threshold value necessary to cross the barrier often surpasses timescales typically accessible to classical MD simulations. Consequently, *enhanced sampling techniques* have been developed to improve the sampling. We use a Replica-Exchange [122] technique termed Replica-Exchange with Solute Scaling (REST2) [123, 124] in order to enhance the conformational sampling in the simulations.

Generally, in Replica-Exchange Molecular Dynamics (REMD) simulations, the system contains n replicas and the simulations of each replica are run simultaneously at different temperatures $T_1 < T_2 < T_3 < ... < T_n$. Occasionally, the replicas are exchanged with a certain probability. The exchange criterion varies depending on the simulated ensemble, but generally depends on the energy differences between the replicas.

The detailed balance condition ensures the reversibility of the process:

$$P(\mathbf{r}_j|\mathbf{r}_i)f(\mathbf{r}_i) = P(\mathbf{r}_i|\mathbf{r}_j)f(\mathbf{r}_j). \tag{2.24}$$

The left side of equation gives the probability of transition from i to j multiplied by the probability that the system is in state i, while the right side gives the probability of system transitioning to i multiplied by the probability that it is in state j, from which it is transitioning. Satisfying the detailed balance condition ensures an unbiased sampling in the REMD. The efficiency of the exchange procedure depends on the overlaps between the distributions of the potential energy sampled by the two replicas, i and j. Since energy fluctuations scale as the inverse of system size, $1/\sqrt{N}$, the application of the methodology to very large system is problematic, in fact in order to obtain a decent exchange (e.g. 30%), a large number of replicas ought to be used. To circumvent this problem, several schemes have been introduced [125], among them the REST2.

In REST2 simulations, schematically shown in Fig 2.3, the replicas are not simulated at different temperatures, but rather the Hamiltonian is rescaled for interactions most dominant in large-scale conformational changes - dihedrals and non-bonded interactions, while the temperature $\beta_{ref} = 1/k_B T_{ref}$ is kept constant in all replica simulations. Replicas evolving with the smoothed potential energy functions at temperature β_{ref} are equivalent to the standard REMD implementation, according to the corresponding state principle. Moreover, these interactions are only rescaled for protein-protein (pp)

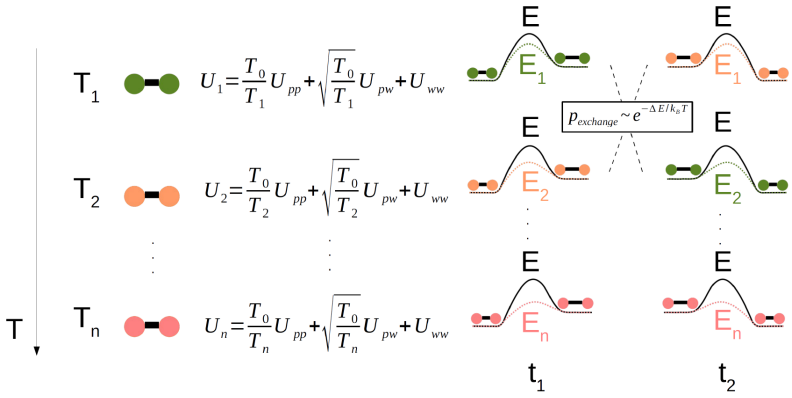

Figure 2.3: Scheme of REST2 simulations. The system is simulated in n copies (replicas). For each copy, the protein-protein and protein-water interactions are rescaled so as to lower the energy barriers with increasing the replica number. Lowering of energy barriers is equivalent to increasing simulation temperatures of the replicas. Replicas occasionally exchange, achieving exchanges between configurations at low effective 'temperatures' and those that are at high effective 'temperatures'. This enables the low temperature replica at ambient conditions to cross high energy barriers and improve configurational sampling in the simulation. Note that T_0 in the Figure corresponds to T_{ref} in the text.

and protein-water (pw) interactions, while the water-water (ww) interactions are not rescaled. This is done so as to decrease the number of replicas that need to be simulated, which grows with number of particles as $\sqrt{N}$. As the water atoms represent the majority of the particles in the system, scaling only the potential energy involving the protein atoms ensures a larger overlap between the potential energy distributions of the replicas. This in turn allows for using a smaller number of replicas while still ensuring a good exchange rate, making the simulations computationally inexpensive when compared to the standard REMD.

The potential of each replica is rescaled according to:

$$U_i(\mathbf{r}) = \lambda_i U_{pp} + \sqrt{\lambda_i} U_{pw} + U_{ww}, \tag{2.25}$$

where $\lambda_i = \frac{\beta_i}{\beta_{ref}} = \frac{T_{ref}}{T_i}$. Furthermore, the dihedrals and Lennard-Jones parameters are rescaled by λ_i, while the Coulomb interactions are rescaled by $\sqrt{\lambda_i}$.

In REST2, the acceptance criterion between replicas i and j is determined by:

$$P_{acceptance} = min[1, e^{-(\lambda_j\beta_{ref}-\lambda_i\beta_{ref})[U_{pp}(\mathbf{r}_i)-U_{pp}(\mathbf{r}_j)]+\frac{\sqrt{\beta_{ref}}}{\sqrt{\lambda_j\beta_{ref}}+\sqrt{\lambda_i\beta_{ref}}}(U_{pw}(\mathbf{r}_i)-U_{pw}(\mathbf{r}_j))}]. \tag{2.26}$$

This acceptance criterion will satisfy the detailed balance condition in the simulation sampling the canonical distribution (NVT). Strictly speaking, the criterion above should also include terms to account for volume fluctuations in order for the process to satisfy the detailed balance condition in the NpT ensemble. These are however negligible when compared to the energy differences and we thus use the expression shown above in NpT simulations. The acceptance criterion implies that good energy overlap is necessary to ensure exchange of replicas, which is satisfied in our simulations given that we achieve exchange ratios of ~ 40%. Once configurations are exchanged, they continue running according to a different Hamiltonian. Eventually, a high-temperature configuration will swap with a low-temperature configuration, enabling the low-temperature replica to explore a portion of the energy surface that would otherwise be inaccessible at lower temperatures. Note that, due to configuration swapping, REST2 cannot recover protein dynamical properties, but is rather used for sampling equilibrium distributions and calculating appropriate properties.

As the solvent-solvent interactions are left unscaled, the effective temperature of the solvent and the protein are different. Therefore, the thermal response of the system can only be obtained in an approximated way. In order to evaluate the temperature dependence of observable averages collected for different replicas, the scaled potential energies used to evolve the system are mapped on an effective temperature scale. A "mean-field" approach was recently introduced [124] to relate the scaled potential energy terms of a replica, E_{pp} and E_{pw}, to an effective temperature $< \beta_i' >$:

$$< \beta_i' >= \beta_i \Big(1 + \left[\sqrt{\frac{\beta_{ref}}{\beta_i}} - 1 \right] \left\langle \frac{E_{pw}}{E_{pp} + E_{pw}} \right\rangle \Big). \tag{2.27}$$

The main benefit of the method is a relatively small number of replicas necessary, decreasing the computational cost and enabling longer per-replica sampling time. The sampling of the configuration space is found to be efficient and the exchange frequencies between replicas are high. Because all replicas are run at the same T_{ref}, exchanging the replicas is easy to implement as only their coordinates are swapped, and velocity rescaling is not necessary. Finally, due to the simplicity of the method that requires only parameter rescaling, it can also be applied to a portion of the system of interest or a single chain in the protein system, further increasing the attractiveness of the method in simulating large systems.

We have rescaled a portion of the protein in REST2 MD simulations to inquire into the conformational transition associated to the catalytic cycle of two homologous GTPase catalytic domains, the details are discussed in the Chapter 3. Alternatively, potential

energy rescaling can be applied to a solution of proteins where only a single protein is rescaled, while others are not. In the latter case, other proteins in the system effectively become the solvent, similarly to the case of a single protein in aqueous solution. This approach has been used to probe the effect of crowding and solvent on the stability of the Lysozyme protein embedded in powder systems, see the Chapter 5.

2.3 Neutron Scattering Experiments

Neutron Scattering (NS) is an important experimental technique in studying biomolecular systems as it captures protein dynamics and relates it to the length scale at which the dynamical processes occur. The basis of the technique, as with any other scattering method, is the collision of the protein sample with the incident neutron beam and subsequent collection of the scattered neutron beam on the detector. The properties of the refracted beam are changed by the interaction with the sample according to the physical properties of the sample. A wide range of NS techniques exists, probing different time and length scales. Common methods include those that use either elastically or inelastically scattered radiation and those that measure the coherent or incoherent contribution to the signal, and their combinations. The coherent contribution arises from cross-correlated atom motions, while the incoherent radiation represents self-correlated particle motions.

Neutrons contain no charge and they interact with the nucleus via short-range, femtometer, nuclear forces, enabling them to penetrate deep into the matter without being absorbed or scattered. The benefit is that the experiments can be made in containers, in liquid or powder samples, and in a wide range of experimental conditions. The weak interaction additionally guarantees that the experiments can be performed on delicate biological samples with virtually no damage. On the other hand, the weak interaction comes at the cost of very low scattering signal. To overcome this limitation, large sample cells and concentrated solutions of samples are often needed.

Neutron scattering can be understood by defining the neutron wave vector traveling in the direction of the sample with magnitude:

$$k = \frac{2\pi}{\lambda}, \tag{2.28}$$

where λ is the neutron wavelength. When the neutron collides with the sample, its interaction at an atomic level can be described by the effective area the atomic nucleus, scattering cross section σ, measured in barns (1 barn=1 m^2). Hydrogen has the largest

scattering cross-section of all atoms in biological materials and a large proportion of the signal comes from scattering on hydrogen atoms. As water contains many hydrogen atoms, to increase the resolution of the scattered signal and to capture only the portion of the radiation scattered from the protein, heavy water (D_2O) is routinely used in neutron scattering experiments instead of water (H_2O).

The nucleus acts as a point scatterer for the neutron, which scatters isotropically around it. The amplitude of the scattered neutron is proportional to the scattering length of the nucleus, related to the scattering cross section by $b = \sqrt{\frac{\sigma}{4\pi}}$. In general, the nuclei are not fixed perfect scatterers, but rather a momentum transfer occurs between the neutron and the nucleus, and the difference between the incident and scattered neutron is described by the scattering vector:

$$\mathbf{Q} = \mathbf{k}_{incident} - \mathbf{k}_{scattered}. \tag{2.29}$$

The neutrons scattered from different atoms are bound to interfere. Constructive interference is achieved upon satisfying Bragg's diffraction law limiting constructive interference only to the case where the distance traveled by waves scattered from different atoms, with positions $\mathbf{r}_i$ and $\mathbf{r}_j$, is an integral multiple of neutron wavelength. This is only achieved when $\mathbf{Q} \cdot (\mathbf{r}_i - \mathbf{r}_j)$ is a multiple of 2π as the waves are then in phase, and when $Q \sim \frac{2\pi}{L}$, where L is the distance between atoms that produced the constructive interference. The last relationship relates a quantity measured in all neutron scattering experiments, the scattering vector Q, with interatomic distances, which is valuable in interpreting the experiments.

Finally, in neutron scattering experiments, the intensity of the scattered neutrons is measured as a function of Q. Van Hove showed that the measured quantities can be written in terms of pairwise spatio-temporal atomic correlations in Fourier space, which forms the basis of interpretation of neutron scattering experiments.

In our investigations, we often compare our simulation results to Neutron Scattering experiments. With the support of our collaborator, M. Maccarini at University Grenoble Alpes, we have performed two Spin-Echo Neutron Scattering Experiments, one at ILL in Grenoble, France, and another one at MZL II in München, Germany. These experiments were carried out on two homologous mesophilic and thermophilic dehydrogenase proteins prepared by our biochemist collaborator D. Madern at IBS, Grenoble. In the context of a separate collaboration with A. Paciaroni at the University of Perugia in Italy, we have complemented experimental data from Elastic Incoherent Neutron Scattering Experiments with our *in silico* study on the effect of crowding and solvent on the

atomistic fluctuations of a protein approaching melting. The quantities and analyses of the two NS techniques of interest will be explained shortly below.

2.3.1 Elastic Incoherent Neutron Scattering

The Elastic Incoherent Neutron Scattering data is obtained in a backscattering device IN13 at ILL in Grenoble [126]. On its way to the sample, the neutron beam first passes through a chopper, selecting neutrons by velocity. Ideally, the distribution of energy in the neutron beam intensity would be a Dirac delta function. In reality, it is a Gaussian distribution due to experimental limitations in velocity selection. The energy resolution of the instrument is derived from the half-width at half-maximum of the energy distribution, in this case $\Gamma_R = 4.5$ μeV, which can be translated to time resolution of atomic motions of $t_R = 150$ ps [127]. The instrument can capture the scattering vector ranges of $Q \in [0.3, 4.7]$ Å, spanning the distances of $[1.3, 20.9]$ Å.

In elastic scattering, no momentum is transferred between the neutron and the nuclei, and the scattering vector is defined as:

$$\mathbf{Q} = \frac{4\pi}{\lambda}\sin\theta. \tag{2.30}$$

As the wavelength is kept constant in the instrument, different Q values are achieved by placing the detector at different angles from the incident beams, thus capturing different Q values.

The quantity derived from the incoherent neutron scattering experiments is the incoherent neutron scattering dynamic structure factor [128]:

$$S(\mathbf{Q},\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} dt e^{i\omega t} I(\mathbf{Q},t), \tag{2.31}$$

the Fourier transform of the intermediate scattering function:

$$I(\mathbf{Q},t) = \frac{1}{N} b_H \sum_{\alpha} \langle e^{i\mathbf{Q}\cdot[\mathbf{r}_\alpha(t)-\mathbf{r}_\alpha(0)]}\rangle, \tag{2.32}$$

where only the hydrogen atoms are taken into account with the index α, b_H are the incoherent scattering lengths of the hydrogen atoms, N is the number of hydrogen atoms, while $\mathbf{r}_\alpha$ are time-dependent atomic positions and the average is intended over an ensemble of configurations.

Moreover, the dynamic structure factor for elastic scattering is roughly equivalent to the intermediate scattering function at t_R:

$$S_{el}(\mathbf{Q},\omega=0) \approx I(\mathbf{Q},t_R) = \frac{1}{N}\sum_{\alpha} \langle e^{i\mathbf{Q}\cdot[\mathbf{r}_\alpha(t_R)-\mathbf{r}_\alpha(0)]}\rangle. \tag{2.33}$$

The structure factor clearly reflects the spatio-temporal atomic correlations over a time-window determined by the temporal instrument resolution. Further quantitative information can be extracted from the elastic incoherent structure factor by adopting a double-well jump model [129], which describes hydrogen atom jumping between two states separated by a free energy difference between the two states separated by a distance d:

$$S_{el}(Q,\omega=0) = e^{-<u^2>_{vib}Q^2[1-2p_1p_2(1-\frac{\sin(Qd)}{Qd})]}, \tag{2.34}$$

where p_1 and p_2 are occupations of the two wells and $< u^2 >_{vib}$ is the vibrational mean-squared displacement, describing the harmonic atomic displacements in the wells. $< u^2 >_{vib}$, d, p_1, and p_2 are fit parameters and they are used in calculating the total mean-squared displacement (MSD):

$$< u^2 >=< u^2 >_{vib} + \frac{p_1p_2d^2}{3}. \tag{2.35}$$

The second term describes the MSD due to jumping between the two potential wells and it is added to the vibrational atomic motion of a particle in a well, given by the first term. MSD is used in characterizing atomic fluctuations over a certain time window and is thus a good quantitative metric of atomic-level protein dynamics.

2.3.2 Spin Echo Neutron Scattering

The resolution of the neutron scattering can be improved by using polarized polychromatic incident radiation, with the energy spread around 20% in the case of Spin Echo Neutron Scattering. The neutron beam polarized along a direction orthogonal to its velocity enters a region with highly homogeneous magnetic field before interacting with the sample. The interaction with the magnetic field causes the neutrons to precess along the direction normal to their velocity with a frequency depending only on the neutron gyromagnetic ratio and on the intensity of the applied magnetic field. After interacting with the sample, the neutron beam progresses through a region having a magnetic field equal and opposite to that present in the region before the sample.

If the interaction is perfectly elastic, the neutrons will make an equal number of precessions with opposite angular velocity in the first and the second magnetic fields, and will recover completely the initial polarization. If however, the neutron interacts inelastically with the sample, their velocity in the two branches of the instrument will be different, as will be the number of precessions. Consequently, the polarization of the

scattered neutrons will be different from that of the incident beam. The measurement of the polarization can be then related to the inelastic interaction between the neutron and the sample, which is in turn related to the dynamical processes occurring in the sample. The details of the method are covered in Ref [130, 131, 132].

The polarized incident neutron beam passes through a flipper that orients the magnetic moments of neutrons in a direction perpendicular to the direction of the beam travel. The magnetic dipoles of the neutron will be therefore oriented orthogonally to the magnetic field, **B**, and this will cause a precession of the neutrons around the direction of the velocity. The precession angle as the neutron proceeds in the magnetic field will be:

$$\phi = \gamma_L \frac{BL}{v}, \tag{2.36}$$

where the gyromagnetic ratio of a neutron is $\gamma_L = 1.832 \cdot 10^8$ rads^{-1}, B is the strength of the magnetic field, L is the length of travel in the field, and v is the speed of neutrons. If the neutrons interact inelastically with the sample, their velocities will change and this will produce a difference in the net precession angle after the second magnetic field:

$$\Delta\phi = \frac{\gamma_L BL\hbar}{mv^3}\omega, \tag{2.37}$$

where m is the neutron mass, ω is the angular frequency of the neutron, and $\hbar = h/2\pi$. Furthermore, we can define the neutron spin-echo time:

$$\tau_{NSE} = \frac{\gamma_L BL\hbar}{mv^3}. \tag{2.38}$$

The change in the precession angle due to interaction with the sample can now be defined in terms of the angular frequency and the neutron spin-echo time only:

$$\Delta\phi = \tau_{NSE}\omega. \tag{2.39}$$

The NSE measures the average change of polarization of a polarized beam upon interaction with a sample. The measurement of the polarization is done by adding another polarizer (parallel to the first one) at the end of the second precessing field. Effectively, the measured quantity is the average value of the projection of the polarization along the direction of the polarizer, $\cos(\omega\tau_{NSE})$. The probability distribution of this quantity is linked to the probability that an inelastic event happens (the $S(Q,\omega)$), therefore the average polarization can be written as

$$P_x(Q,\tau_{NSE}) = \frac{\int d\omega S(Q,\omega)\cos(\omega\tau_{NSE})}{\int d\omega S(Q,\omega)}, \tag{2.40}$$

where $S(Q,\omega)$ is the dynamic structure factor. The numerator and the denominator are the intermediate scattering function and the static structure factor, respectively. The intermediate scattering function is defined in Eq 2.32, while the static structure factor is defined as:

$$I(\mathbf{Q},0) = \frac{1}{N}\sum_{i,j}\langle e^{i\mathbf{Q}\cdot[\mathbf{r}_i-\mathbf{r}_j]}\rangle. \tag{2.41}$$

The structure factor gives a distribution of particles in the molecule, determined by their interatomic correlations. Finally, the quantity measured in the Spin Echo Neutron Scattering Experiment can now be written as:

$$P_x(Q,\tau_{NSE}) = \frac{I(Q,t)}{I(Q,0)}. \tag{2.42}$$

In order to obtain useful information, the P_x is measured as a function of the τ_{NSE} by varying the magnetic field in the coils. Once $\frac{I(Q,t)}{I(Q,0)}$ is obtained, quantitative information on protein structure can be extracted by appropriate fitting. The procedure is equivalent to what was done in the numerical calculations, and will be described in the section Analysis of Molecular Dynamics trajectories, see below.

The main benefit of using Spin Echo Neutron Scattering is that it can access a broad range of time- and lengthscales, making it particularly useful in measuring large-scale motions in proteins. The accessible timescales range from ps to ns, while the accessible lengthscales cover the range from a fraction of Å to tens of Å. The technique measures quasielastic scattering of neutrons and captures both the coherent and the incoherent component of the scattering. The coherent dynamics is dominant at low Q values, while the incoherent dynamics dominates at higher Q.

2.4 Analysis of Molecular Dynamics Trajectories

In this final section, we detail the complete toolbox used for analyzing the MD simulations. Namely, we describe how we addressed the issue of protein flexibility/rigidity, the central challenge in our investigation. Subsequently, the comparison with NS experiments is discussed.

Most of the analysis tools were developed *in house* and were supported/coupled by specific publicly available packages, e.g. the package MDAnalysis providing essential Python libraries for the post processing of MD simulations [133] and the Sassena [134] suite that performs NS-related calculations.

2.4.1 Characterizing the Protein Landscape

In order to quantify the protein flexibility, we used a set of clustering strategies that allow the conformational landscape of a protein to be reduced to an essential ensemble of conformational states. The number of these states and their relative interconversions quantify the protein conformational flexibility. This strategy was complemented by *ad hoc* entropy calculations.

2.4.1.1 Conformational Clustering

Each simulation snapshot provides a different protein configuration visited along the MD trajectory, creating a large quantity of data difficult to interpret. To mine the relevant information from the data, techniques such as clustering are widely used, helping to cast together similar "information" in a single group (also termed cluster). The goal of conformational clustering in particular is to produce only the representative conformational substates visited during the MD simulations or sampled via an enhanced sampling technique such as REST2. As the protein has many degrees of freedom, the number of possible conformational substates is large, as well as the number of collective variables that can be used to distinguish between different conformations. In our work, we have tested three collective variables that help in characterizing the protein structure [45, 78]. The simplest variable measuring the distance between two conformations is the root mean square distance:

$$RMSD(t) = \sqrt{\frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} (r_i(t) - r_i^{cluster})^2}, \tag{2.43}$$

where N_{C_α} is the number of carbon C_α atoms, and $r_i^{cluster}$ are the coordinates of an existing cluster.

Another important variable that can be used to distinguish protein conformations is the difference in the fractions of native contacts:

$$d(t) = \sqrt{\frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} \left(\frac{l_i(t)}{l_i'} - \frac{l_i^{cluster}}{l_i'} \right)}, \tag{2.44}$$

where N_{C_α} is the number of carbon alphas, l_i' is the number of native contacts given by the number of carbon alphas within a spherical cut off from the C_α^i in the reference state, and $l_i(t)$ is the number of native contacts calculated for the configuration at time t, while $\frac{l_i^{cluster}}{l_i'}$ is the fraction of native contact states existing in a particular cluster.

Finally, large-scale rearrangements in a protein can be further explored by considering the backbone torsion angle and defining another clustering variable, the fraction of native torsion angles, thereafter clustering the configurations based on the difference between the fractions of native torsion angles:

$$d(t) = \sqrt{\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\left(\exp[-\frac{(\theta_i(t)-\theta_i')^2}{\sigma^2}] - \exp[-\frac{(\theta_i^{cluster}-\theta_i')^2}{\sigma^2}]\right)}, \tag{2.45}$$

where $|\theta_i(t)-\theta_i'| < 180°$, $\sigma = 60°$, N_θ is the number of torsion angles θ, θ_i' are the torsion angle values in the reference state, $\theta_i(t)$ are the torsion angles in the configuration at time t, and $\theta_i^{cluster}$ are the torsion angles in a conformation representing each cluster. Both ϕ and ψ dihedrals were used in the calculations.

In our work, the clustering is performed with Hartigan's leader algorithm [135], where the first simulation snapshot is set as the first cluster centroid, known as cluster leader. Next snapshot is compared to the first cluster leader by one of the three criteria described above and if it differentiates by more than a cut off value, a new cluster leader is created, else the protein structure is merged to the existing cluster. The centroid of each cluster, i.e. the protein configuration used in comparison with following snapshots, is the first configuration assigned to the cluster – the cluster leader. The procedure is repeated for every trajectory frame, merging similar states and separating those that are vastly different. The algorithm depends on the cut off value used to distinguish the cluster leaders as a large threshold value will create less clusters, while a small one will possibly produce a large number of data. Since the clustering is dependent on the cut off, care must be taken to always use the same cut off value when comparing results. Additionally, number of possible configurations is proportional to protein amino-acid length and comparing the clustering results is only meaningful when the number of amino-acids is similar [45]. Finally, the algorithm also depends on the order of the frame composing the trajectory, although the main output differences are produced by varying the cut off. The main advantage of the algorithm is its speed and the fact that the number of clusters does not need to be provided *a priori*.

When the clustering is performed, the protein conformation of each snapshot is assigned to one cluster leader. Consequently, the interconversions of meaningful configurations can be monitored, as well as the time evolution of the total number of clusters. Assuming the protein landscape is confined, the protein samples a finite number of accessible conformational substates, therefore the total number of clusters should grow

and eventually reach a plateau, which can be described with a simple exponential model:

$$N(t) = N_{\infty} \cdot (1 - \exp(t/\tau)), \tag{2.46}$$

where N_{∞} is the number of cluster leaders at infinite time and τ reflects the rate by which the plateau is reached. However, the growth of the number of cluster depends on temperature and the explored time scale [136], thus care must be taken when results are interpreted.

The number of clusters and the frequencies of their interconversions are further represented as networks of conformational substates. The nodes of networks represent cluster leaders, which are connected with 'edges', i.e. normalized frequencies of interconversion between the leaders. The complex cluster network is projected on a 2D surface by using a force-based algorithm, modeling the network with a set of charges and masses assigned to each node, and spring force constants assigned to edges. The values of these parameters depend on the number of configurations belonging to leaders and the frequencies of transitions between them. The network is subjected to a gravitational, electrostatic, and Hookean force, redistributing the nodes on the surface for a clear representation. The network representation is fully dependent on the choice of the parameters used to calculate the forces and the same parameter set should always be used when comparing different networks to obtain meaningful comparison. The force-based visualization is provided by the software package GEPHI [137].

2.4.1.2 Kinetic Clustering

The output of conformational clustering, a network of conformational substates defined by its nodes and edges, is used as a starting point for Markov chain clustering [138]. The algorithm was discovered empirically and was found to be efficient in further merging clusters that interconvert often, while separating those that rarely interconvert. This is achieved by starting random walks from each cluster leader and is based on the fact that the random walks are unlikely to leave frequently interconverted states.

Technically, a transition matrix is built with normalized frequencies of interconversions, i.e. normalized edge weights are matrix elements. In order to mimic the progress of the random walk, the matrix is subsequently transformed iteratively by applying two matrix operators to convergence - matrix squaring ('expansion') and the Hadamard power ('inflation'), where each element is raised to the power p. After applying the Hadamard power, the matrix has to be rescaled so that the elements belonging to every node add to unity, a condition necessary for the process to be stochastic. The parameter p will

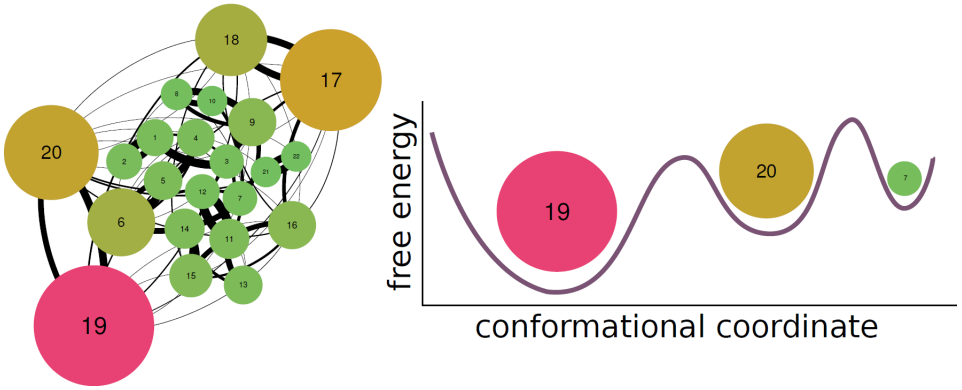

Figure 2.4: The schematic representation of how the network of kinetic substates relates to the conformations sampled from the free energy surface of the protein. The more flexible the protein, the more configurations it is bound to sample, and consequently the larger the number of nodes in the network. Conversely, in the extreme limit for a very rigid protein, the network would produce a single node as the protein would be trapped in a single minimum. The thickness of the edges reflects the frequency of the interconversions between states, and the size of the node reflects the population size of a cluster.

determine the granularity of the network and it represents the height of the kinetic barrier that will trap random walkers to a single cluster centroid. Applying expansion causes the random walkers to dissipate within leaders and inflation eliminates the flow between leaders. While the global convergence of the algorithm is difficult to prove, it is simple, fast, granularity of the final network is easily controllable, and there is evidence that the final output reflects the input cluster structure, making it a useful tool in further reducing the complexity of the protein conformational substates. The final output is a network of kinetic substates with nodes and edges weighted by occupancy and transition frequency, respectively. These networks are visualized with a force-based algorithm in GEPHI (see previous subsection). The final result of kinetic clustering is shown in Figure 2.4.

2.4.1.3 Entropy Estimation

Entropy is the key driving element in protein folding, hydrophobic interaction formation, and ligand binding. In order for the entropy to be calculated exactly, the phase space would need to be sampled to its entirety, which in turn requires an infinitely long simulation time. Advanced methods such the Thermodynamic Integration or other methods can be employed to estimate the entropy of complex processes [139]. When focusing on the

folded state of a protein, the entropy associated to small shifts of the configurations, e.g. caused by the binding of a ligand, can be accessed considering "vibrational" or "rotational" contributions. A routine approach, related to NMR experiments, is used to estimate the entropy variation associated to the backbone fluctuations. The movements of the protein backbone can be described by considering the motion of a rigid bond vector associated to it, e.g. the NH bond vector. The movement of the bond vector is restricted to different degrees depending on the steric interaction with the surroundings. In the simplest case employed here, the motions of all vectors will be described with a single parameter S^2, derived from theory describing NMR relaxation data. The NMR experiment measures spin relaxations and effectively measures Fourier transformed time correlation functions that contain dynamic information. To interpret the obtained results, several models have been adopted, including the model relying on a two-parameter fit of the bond vector time-correlation, presented here. One of the fitted parameters is later used in estimating the entropy.

The second-order Legendre polynomial correlation function describing the orientation of the bond vector $\mathbf{v}$ is calculated as follows:

$$C(t) = (3(\mathbf{v}(t) \cdot \mathbf{v}(0))^2 - 1)/2. \tag{2.47}$$

The correlation function, in the ideal case, has the following properties: $C(0) = 1$ and $C(\infty) = const$. The constant is defined as S^2, the generalized order parameter, a measure of spatial restriction of the N-H vector [140]. It lies between $0 \leq S^2 \leq 1$, where internal isotropic unrestricted motion will produce $S^2 = 0$, while fully restricted motion will produce $S^2 = 1$. To describe a correlation function presented above, the simplest mathematical model of exponential relaxation takes the following form [140]:

$$C(t) = S^2 + (1 - S^2)e^{-t/\tau}. \tag{2.48}$$

Once the S^2 is extracted as a fit parameter, the entropy difference between two protein states, a and b, is estimated according to [141, 142]:

$$\Delta S = -k_B \sum_i ln\{\frac{3 - (1 + 8S_{a,i})^{1/2}}{3 - (1 + 8S_{b,i})^{1/2}}\}, \tag{2.49}$$

where i runs over the residues of the protein and k_B is the Boltzmann constant.

2.4.2 Protein Essential Dynamics

In computer simulation, the protein dynamics spans timescales ranging from picoseconds to microseconds/milliseconds and lengthscales ranging from Ångstroms to nanometers.

In order to extract biologically important motions, methods such as Principal Component Analysis (PCA) and Normal Mode Analysis (NMA) have been developed [143, 144, 145]. They break down protein dynamics to independent modes, where motions of different amplitudes and frequencies are individuated and can be considered separately in terms of their biological function. Both methods are described below. We anticipate that the PCA and NMA were used also to support the interpretation of NS experiments as discussed in the section 2.4.3.

2.4.2.1 Principal Component Analysis

The Principal Component Analysis (PCA) allows the essential dynamics of the protein to be extracted, with taking into account the multiminima nature of the potential energy surface of the protein. The calculation was performed in GROMACS 2.9, where it was implemented according to Ref [145]. The procedure is started with removing the rototranslations by superimposing the trajectory on a single reference structure. The covariance matrix of atomic displacements relative to their time average is subsequently constructed, with the elements of the covariance matrix:

$$C_{ij} = \sqrt{m_i m_j}\langle(\mathbf{r}_i - <\mathbf{r}_i>)(\mathbf{r}_j - <\mathbf{r}_j>)\rangle, \tag{2.50}$$

where $\mathbf{r}$ are atomic positions, m are atomic masses and $\mathbf{C}$ is a $N \times N$ matrix, where N is the number of atoms in the molecule. The covariance matrix is mass weighted so as to account for the inertial effects, as the displacement of heavy atoms will be smaller, while the displacement of the light atoms will be larger. The matrix $\mathbf{C}$ is symmetric and can be diagonalized by an orthogonal transformation $\mathbf{T}$:

$$\mathbf{C} = \mathbf{T}\Lambda\mathbf{T}^T, \tag{2.51}$$

where Λ is the diagonal eigenvalue matrix, and the matrix $\mathbf{T}$ contains as columns the unit eigenvectors $\mathbf{e}$, also called the principal modes or essential modes. The elements of the eigenvalue matrix Λ are eigenvalues λ, representing the variance of data along the orthogonal eigenvectors. Diagonalizing the covariance matrix transforms the atomic displacements to a new orthornormal basis as defined by eigenvectors, where the new basis is constructed so that the variance of the displacement along the first eigenvector will be maximal, and smaller in each subsequent orthogonal direction; $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_{3N}$. The projection of the atomic coordinates along the normal modes yields the principal components:

$$\mathbf{P}(t) = \mathbf{T}^{\tau}\sqrt{M}(\mathbf{r}(t) - <\mathbf{r}>), \tag{2.52}$$

where **M** contains the atomic masses with elements $m_i m_j$, and the transposed **T** contains the eigenvectors as rows. The main result of this procedure is to simplify the multidimensional correlated atomic displacements to a space of reduced dimensionality; from an initial $N \times N$ matrix, a maximum of $3N-6$ eigenvectors with nonzero eigenvalues can be obtained; note that the deducted 6 correspond to the previously removed rototranslations. As the first principal component will account for the majority of dynamics, and each subsequent less, one could assume that considering the first principal components describes the largest displacements in a molecule, potentially corresponding to domain motions and secondary structure elements movements i.e. global protein motions. To visualize these motions in Cartesian coordinates, the trajectory can be projected on one, two, or several principal components:

$$\mathbf{r}^{projected}(t) = <\mathbf{r}> + \sqrt{M}\,T^{\tau}\mathbf{P}(t). \tag{2.53}$$

2.4.2.2 Normal Mode Analysis

Normal Mode Analysis (NMA) is a technique that approximates the protein conformational landscape with a harmonic potential well. Although the reality of the multiminima potential energy surface is staggeringly different, this simplification is valid for examining protein motion in a local minimum. The harmonic approximation breaks down at physiological temperatures [146], and care must be taken when making conclusions based solely on the NMA.

NMA is performed by using the Anisotropic Network Model (ANM) [147] as implemented in the Python package ProDy [148]. To improve computational efficiency, only the C_α atoms are considered, additionally, a cut off is introduced, and only the interactions within a cut off r_c are considered in potential calculations. The C_α atoms are connected with springs of uniform elastic constant γ. Another major advantage of the ANM is that no energy minimization is required in the procedure, which is usually case for the standard NMA.

The harmonic potential function is defined as:

$$U = \frac{\gamma}{2} \sum_{|r_{ij}^0| < r_c} (r_{ij} - r_{ij}^0)^2, \tag{2.54}$$

where the r_{ij} are interatomic distances, r^0_{ij} are the equilibrium atomic distances, and the sum is run over the C_α atoms within a cut off value.

The Hessian matrix H is built of second order derivatives and is of dimensions $3N \times 3N$ for a system with N C_α atoms. Diagonalization of the Hessian $\mathbf{H}$ will yield eigenvectors $\mathbf{e}$ and their eigenvalues ω^2:

$$\mathbf{He} = \omega^2 \mathbf{e}. \tag{2.55}$$

A normal mode Q_j is then specified by:

$$Q_j = \sum_{i=1}^{3N} e_{ij} \mathbf{r}_i. \tag{2.56}$$

The sum runs over the elements of the eigenvector $\mathbf{e}$ and has $3N$ components, where $|\mathbf{e}| = 1$.

In terms of Cartesian coordinates, the modes can be expressed as:

$$r_{ij} = e_{ij} A_j \cos(\omega_j t + \epsilon_j), \tag{2.57}$$

where A_j is the amplitude and ϵ_j is the phase of the harmonic motion, while the square-root eigenvalue ω is its angular frequency. It can be shown that the lower the frequency ω, the larger the amplitude will be, pointing to the lowest-frequency modes as functionally relevant.

Generally speaking, the protein modes can be divided in two classes: anharmonic modes with large amplitudes of fluctuations and multipeaked distributions, captured in PCA, and harmonic modes with small amplitudes and Gaussian-like distributions, captured by NMA [143]. The notion was confirmed on synthetic peptide simulations, where it was found that the fluctuations of the NMA modes accounted for 30% of protein total mean-squared fluctuation (MSF) values extracted from MD simulations, while this number grew to 70% for rigid peptides [149].

2.4.3 Meeting NS Experiments

In this part of the chapter, we introduce the observables used to directly compare the results from MD simulations with NS experiments, and the complementary analyses that were done in order to complement the average experimental signal with a microscopic outlook.

2.4.3.1 Atomistic Fluctuations

In order to compare with Elastic Incoherent NS experiments, see subsection 2.3.1, the atomistic fluctuations of the protein system must be computed. Mean squared fluctuation is a measure of the flexibility of the atomic backbone. It is the variance of atomic positions over the course of the trajectory. For each atom, it is defined as:

$$MSF = \frac{1}{T}\sum_{t}^{T}(\mathbf{r}(t) - <\mathbf{r}>)^2, \tag{2.58}$$

where T is the total number of simulation snapshots and $\mathbf{r}$ contains atomic coordinates. It is contained in experimentally measurable quantities of Neutron Scattering and X-ray crystallography, as shown below.

Alternatively, it is possible to measure the displacement of the atom at a given time interval. Mean squared displacement is a measure of atomic movement over a time window Δt with respect to a reference position:

$$MSD(\Delta t) = <u^2> = \frac{1}{T}\sum_{t}^{T}(\mathbf{r}(t+\Delta t) - \mathbf{r}(t))^2, \tag{2.59}$$

where T is the total number of simulation steps. The mean-squared displacement is experimentally measurable in Neutron Scattering experiments (see section on Elastic Incoherent Neutron Scattering). For a meaningful comparison, the time window Δt should match the resolution of the experimental apparatus. MSD is related to MSF by [128]:

$$MSD = 2MSF. \tag{2.60}$$

Furthermore, the MSD, contains a measure of flexibility in time that is measured in X-ray crystallography in therms of the 'temperature factors', B-factors or Debye-Weller factors:

$$DWF = \exp\left(\frac{-Q^2 < u^2 >}{3}\right), \tag{2.61}$$

where Q is the scattering vector. The atoms with high B-factors belong to disordered protein regions, while those with low B-factors belong to structured regions that are well resolved.

2.4.3.2 Support of the NS Spin Echo

The interpretation of the NS Spin Echo experiments is extremely laborious. The main information we have collected from experiments and simulations concerns the contribu-

tion of protein internal motions to the diffusion spectrum extracted from the measured Intermediate Scattering Function. Our goal, as described extensively in the Chapter 4, was to compare the thermal activation of protein soft modes in two homologous proteins having different optimal working temperatures.

To start with the most general consideration, the atoms in a molecule are subjected to constant kicking and dragging from surrounding atoms, either solvent or intramolecular, which results in random stochastic motion termed diffusion. The theory of random motion in spatial dimension is extended and generalized for any continuous trajectory random variable X with the Markov property, yielding the Fokker-Planck equation that describes the time distribution function of the random variable. In the special case of no external field applied (i.e. zero drift) the Fokker-Planck equation becomes:

$$\frac{\partial p(X,t)}{\partial t} = D\frac{\partial^2 p(X,t)}{\partial X^2}, \tag{2.62}$$

where $p(X,t)$ is the probability density of the random variable X. The solution of the equation is the distribution of the random variable in time, characterized by the diffusion coefficient D, shown here for the simplest case where the diffusion of particle is homogeneous across the random variable phase-space and can be characterized by a constant value. As the diffusion coefficient is a property that characterizes the distribution of any random variable in time, it is a good measure of changes in protein dynamical properties that are governed by random motions. The diffusivity of a protein at different characteristic length scale can be extracted from Spin Echo experiments as detailed below.

The normalized intermediate scattering function, the Fourier transform of the spatio-temporal particle correlation function, which is obtained in experiments, can be calculated directly from the MD trajectories using the same expression as in Eq 2.42. The calculation were done by using the Sassena software package [134]. The function contains information on protein dynamics encoded as a function of the quantity Q, the scattering vector, which is inversely proportional to the the interparticle distances L.

The short time decay of the $I(Q,t)/I(Q,0)$ can be approximated by a cumulant expansion [150]:

$$\ln\frac{I(Q,t)}{I(Q,0)} = -\bar{\Gamma}(Q)t + \frac{1}{2}K_2t^2 + \frac{1}{3!}K_3t^3 + \cdots, \tag{2.63}$$

where K_2 and K_3 are the second and third cumulant, and $\bar{\Gamma}(Q)$ is related to the diffusion coefficient $D_0(Q)$ through:

$$D_0(Q) = \frac{\bar{\Gamma}(Q)}{Q^2}. \tag{2.64}$$

The extracted diffusion coefficient characterizes the atomic correlation at interparticle distances $Q \sim 1/L$. It is assumed that the diffusion coefficient in an MD simulation corresponds to a dilute condition. It is important to note that three different contributions enter in the coefficient D_0, the translational (Q-independent), the rotational, and internal motions. The experimental samples are concentrated and the diffusion coefficient extracted from the experimental $I(Q,t)/I(Q,0)$ must be treated to take into account the concentration and the hydrodynamic effect, see Ref [151].

The calculation of the $D_0(Q)$ spectrum allows comparison to experimental data. However, the important goal is to assess to what extent the internal motion is relevant in shaping the spectrum. This can be done using several strategies. Firstly, in order to evaluate whether the internal motion makes a meaningful contribution to the associated spectrum, the motion of a rigid body can be analytically estimated [152] by projecting the interatomic distances of atomic positions in an X-ray, $\mathbf{r}_{k(j)}$, on the scattering vectors Q, which are spherically distributed:

$$D_0(Q) = \frac{1}{Q^2 F(Q)} \sum_{j,k} \langle b_j b_k (\mathbf{Q} \cdot D^T \cdot \mathbf{Q} + \mathbf{L}_j \cdot D^R \cdot \mathbf{L}_k) e^{i\mathbf{Q}\cdot(\mathbf{r}_j - \mathbf{r}_k)} \rangle. \tag{2.65}$$

D^T and D^R are the translational and rotational diffusion tensors, respectively, and they can be obtained from theoretical calculations performed using the HYDROPRO software [153]. The $b_{k(j)}$ are the scattering lengths, $\mathbf{L}_{k(j)} = \mathbf{r}_{k(j)} \times \mathbf{Q}$ is its angular momentum vector, and $F(Q)$ is the form factor of the protein:

$$F(Q) = \sum_{j,k} b_j b_k \exp[i\mathbf{Q}(\mathbf{r}_j - \mathbf{r}_k)]. \tag{2.66}$$

In past work [151], it was shown that the inclusion of internal motion is necessary in order to reproduce the experimental spectrum, since the simple rigid-body calculations failed to perfectly reproduce the experimental data, see Figure 2.5. As we discuss in Chapter 4, this was also the case in our investigation.

Therefore, the MD simulations must be included to quantify the contribution of internal dynamics by performing an appropriate post-processing of the trajectory before computing the intermediate scattering function $I(Q,t)$. Following the procedure detailed in Ref [154], by removing roto-translation from the MD trajectory, the calculation of $I(Q,t)$ will include internal motion only. The extracted diffusion coefficient will consequently correspond only to the correlated internal displacement of the protein atoms. This contribution can be added to the rigid-body roto-translation in order to match the

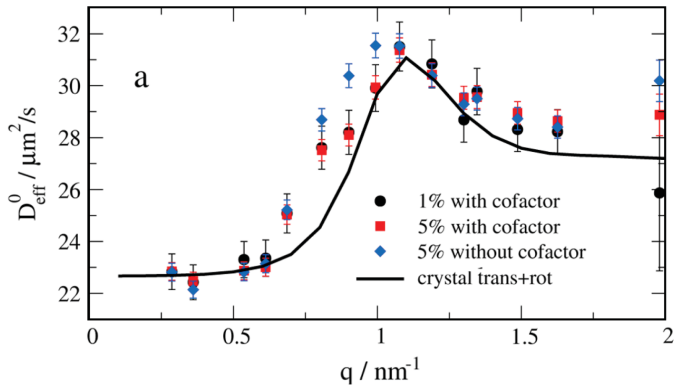

Figure 2.5: The diffusion spectrum of the Alcohol Dehydrogenase with experimental data shown for different cofactor binding conditions, while the solid black line shows the contribution of the calculated spectrum including only the rotational and translational component. Taken from Ref [151].

experiments. Moreover, in order to dissect to which modes the larger diffusivity is associated, a more detailed analysis is adopted by extracting the modes via Normal Mode Analysis and computing the contribution to diffusivity. Alternatively, Principal Component Analysis can be performed, where the projection of the MD trajectory on a finite number of modes is used to quantify their contribution to the internal diffusivity.

Finally, a complementary quantification of internal motion can be provided by the harmonic approximation. In fact, diffusion coefficient characterizing protein dynamics can be simply derived for a harmonic collective variable in terms of its fluctuation δA and time correlation function relaxation τ [155]:

$$D = \frac{\delta A^2}{\tau}, \tag{2.67}$$

where δA^2 is the variance of the fluctuation $\delta A = A(t) - < A >$ and τ is the relaxation time of the correlation function:

$$C(t) = < A(t)A(0) > . \tag{2.68}$$

The relaxation time is obtained by fitting a single exponential, $\exp(-t/\tau)$, to the initial decay of the autocorrelation function. The collective variable can represent for instance the distances among separate secondary structure elements in the protein or any other appropriate variable with harmonic behavior.

CHAPTER 3

Stability and Function at High Temperature for a Mesophilic and Thermophilic GTPase Homologue

Comparing homologous enzymes adapted to different thermal environments aids to shed light on their delicate stability/function trade-off. Protein mechanical rigidity was postulated to secure stability and high-temperature functionality of thermophilic proteins. In this Chapter we challenge the corresponding-state principle for a pair of homologous GTPase domains by performing extensive Molecular Dynamics simulations, applying conformational and kinetic clustering, as well as exploiting an enhanced sampling technique (REST2). While it was formerly shown that enhanced protein flexibility and high temperature stability can coexist in the apo hyperthermophilic variant, here we focus on the holo states of both homologues by mimicking the enzymatic turnover. We clearly show that the presence of the ligands affects the conformational landscape visited by the proteins, and that the corresponding state principle applies for some functional modes. Namely, in the hyperthermophilic species, the flexibility of the effector region ensuring long-range communication and of the P-loop modulating ligand binding are recovered only at high temperature. The work is published in Ref [78].

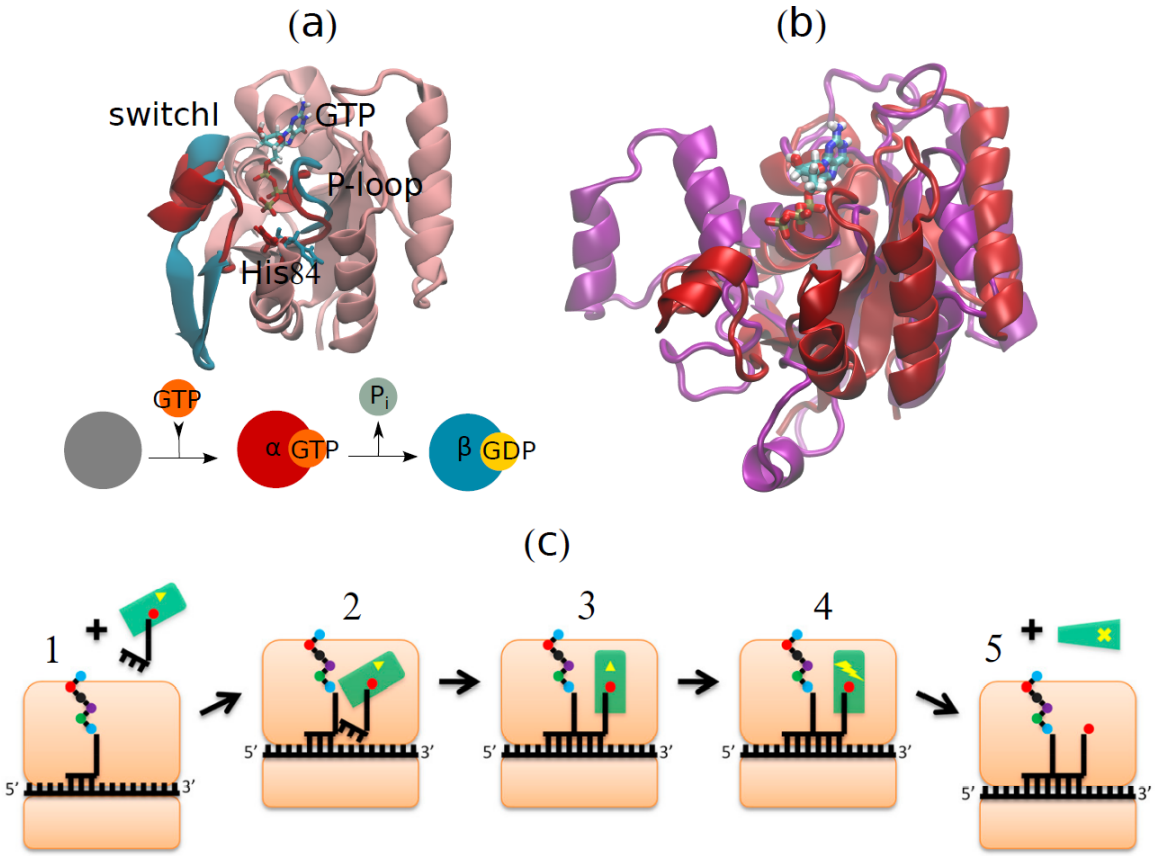

Figure 3.1: Figure (a) contains two panels, the upper shows the superimposition of the G-domain of *E. coli* EF-Tu in the active GTP form, where the switch I region (G40-I62) is in α secondary structure (shown in red), and the inactive GDP form, where the region is partially in the β state (shown in blue). The same color code is used to emphasize other important structural elements, the P-loop (G18-T25) lining the active pocket and the explicitly shown His84, discussed later in text. GTP is shown in the active site. The lower panel in Figure (a) shows the catalytic cycle of the EF-Tu, and the corresponding conformational changes. Figure (b) shows the superimposition of the active forms of EF-Tu *E. coli* in red and EF-1α *S. solfataricus* in purple, aligned by GTP in their active site. The orientations of the proteins in (a) and (b) are similar, and equivalent structural features can be seen in both Figures. The reader is specifically pointed to notice the double helical insertion in the switch I region of EF-1α. Figure (c) is taken from Ref [156] and shows the schematic representation of the EF-Tu in the elongation process. The ribosome, shown in orange, translates a messenger RNA sequence, while EF·GTP·aa-tRNA ternary complex is approaching (1), only to be subsequently bound on the ribosome (2). After the codon-anticodon recognition, the ternary complex undergoes a conformational change on the ribosome (3), whereafter the GTP hydrolysis to GDP follows (4), after which the EF·GDP dissociates from the ribosome (5). Other features shown: amino-acids as small circles, GTP as yellow triangle, yellow lightning bolt represents GTP to GDP hydrolysis, and the yellow cross represents GDP.

3.1 Introduction

Despite extensive research, the direct connection between the thermal activation of protein flexibility and the temperature dependence of the enzymatic activity remains elusive. In fact, it is plausible that mechanical rigidity controls the stability of the protein by ensuring a functional fold at high temperature, while the temperature shift of the activity depends solely on a higher kinetic barrier for the enzymatic chemical step [62, 73]. However, when allostery or conformational changes control substrate binding and unbinding or signal propagation upon catalysis, thermal activation of relevant modes has to be considered in detail.

In this work we tackle the problem by considering a model system, a pair of homologous G-domains from a mesophilic (Elongation-Factor Tu) and a hyperthermophilic (Elongation-Factor 1α) protein. The Elongation-Factor (EF) [157, 158] participates in protein translation, which takes place on the ribosome and is schematically shown in Figure 3.1 (c). The EF carries an aminoacyl-tRNA (aa-tRNA) to the ribosome, where the mRNA is translated to an amino-acid sequence by pairing its nucleotide bases with those of the aa-tRNA. Initially, the EF forms a ternary complex EF·GTP·aa-tRNA, and upon codon-anti codon recognition, the GTP is hydrolyzed to GDP, inducing a conformational change necessary for the release of the aa-tRNA and dissociation from the ribosome in the EF·GDP form. The interested readers can find a substantial body of work where the catalytic boost of the GTPase activity [159] as well as the EF conformational changes [160, 156, 161, 162] induced by ribosome binding were deeply investigated.

The EF-Tu of *E. coli* and EF-1α of *S. solfataricus* are of similar structure [163, 164, 165, 166], both triangular three-domain proteins with a hole in the middle and substantial differences between the reactant GTP (active) and product GDP (inactive) state. The superimposition of the catalytic subunits of the two proteins is shown in Figure 3.1 (b). Despite the role of interdomain interactions ensuring long-range communication, some essential features of the EF activity can be investigated by considering only the catalytic domain, as the isolated domain is catalytically active [167, 168]. The catalytic subunit is slightly less thermostable than the three-domain protein for both the mesophile and the thermophile. The inactivation temperature of the catalytic subunit in the EF-Tu is 41 °C as opposed to 46 °C for the entire protein [169, 170]; for EF-1α the catalytic subunit loses activity at 84 °C and the entire protein at 94 °C [168]. These temperatures are close to the optimum growth temperatures of *E. coli* (37 °C) and *S. Solfataricus* (80 °C), respectively, confirming that the proteins are optimized to function in a narrow

temperature window.

Moreover, the important conformational changes during the enzymatic turn-over [171, 172] occur in some specific regions of this domain. In the mesophilic domain, the crystallographic structures of the reactant and the product states show marked differences in the switch I region (also referred to as effector region, residues G40-I62) and switch II region (G83-T93). The former is reported to undergo a dramatic α to β secondary structure transition between residues P53 and G59 upon GTP hydrolysis (see Figure 3.1 (a)) [173, 174], while the latter is a helix that shifts towards the C-terminus by a single turn [163]. The structure also contains a number of conserved residues, most notably in the P-loop (G18-T25) and in the region between residue N135 and D138, both lining the active site and forming hydrogen bonds with the ligand.

The switch I region of the hyperthermophilic variant contains an insertion of two small α helices, and no conformational change spanning this region was reported in the literature. Computer simulations [45] showed that the early steps of the protein unfolding in the mesophilic G-domain occur at the level of the switch I region, indicating that the structuring effect due to the helical insertions is an essential stabilizing factor for the hyperthermophilic species.

In order to get more precise insight on the rigidity/function relationship, we used computer simulations to investigate the two homologous G-domains in their holo states by virtually mimicking the enzymatic turnover. Molecular dynamics of the protein-substrate complexes were performed at the microsecond time scales. Enhanced sampling of the protein conformations was also performed by employing the Hamiltonian-replica exchange scheme REST2 [123, 124].

Here we verify the validity of the corresponding state principle for some key functional modes of the proteins in their holo state. As stated in the Introduction of the thesis, Somero's corresponding state principle asserts that the protein's conformational flexibility is adjusted to the optimal working temperature of the enzyme, i.e. the flexibilities of proteins should be comparable at their respective optimal working temperatures [57, 56]. Here we show that the magnitude of the changes in the flexibility upon GTP binding and hydrolysis are comparable between the two species at their respective "optimal working" temperatures. Moreover, we confirm that the stability/function trade-off is encoded in the structural motif of the switch I region, which is highly flexible and keen to secondary structure shift in the mesophilic species, while being highly structured and more resistant to temperature in the hyperthermophilic domain.

3.2 Methods

3.2.1 Systems

We have studied the isolated catalytic G-domain of the mesophilic (*E.coli*) EF-Tu and that of the hyperthermophilic (*S.solfataricus*) EF-1α. The mesophile domain was considered in the apo state and with the GTP or GDP molecules bound to its active site. Depending on the nature of the ligand bound to the protein, the switch I region (P53-G59) is α-helical in the crystal structure isolated in the presence of GTP (PDB 1OB2), while it is a β-sheet when GDP is present (PDB 1EFC) [163]. Both conformers have been considered, leading to a total of six systems for the mesophilic domain: the two apo conformers, ecG$^{\alpha}$ and ecG$^{\beta}$, and the two holo states in all their possible conformations, leading to four additional states ecG$^{\alpha}$·GTP, ecG$^{\alpha}$·GDP, ecG$^{\beta}$·GTP, ecG$^{\beta}$·GDP. The catalytic G-domain of the hyperthermophilic EF-1α was extracted from the PDB entry 1SKQ [165], with the missing portion spanning the residues 66 to 76 inserted by homologous modelling [47]. In the crystal structure, the protein was isolated with GDP bound to it, which was either removed, kept intact or replaced by GTP to simulate the apo, the holo ssG·GDP and ssG·GTP states, respectively.

3.2.2 Molecular Dynamics Simulations

The G-domains were capped with COO^{-} and NH^{3+} terminals. Ligands, when not originally present in the crystallographic structures, as in the case of GTP, were aligned to an existing substrate (GDP or GNP). A short minimization was performed to relax structural clashes. The proteins were inserted in a simulation box and solvated with water by surrounding the protein with at least 10 Ålayer of solvent. Ions were added to neutralize the systems. All the simulations have been carried out using the NAMD 2.9 software [89], the CHARMM22/CMAP force field for the protein [91, 94], and the CHARMM-TIP3P model for water. The details of the method are described in Chapter 2. After a 4 ns thermal equilibration, the simulations were propagated in the NpT ensemble using a Langevin thermostat (characteristic time 1 ps) and barostat (dumping time 50 fs). In the simulations we used an integration timestep of 2 fs. All systems were simulated at ambient conditions, T=300 K and p=1 atm, for 0.6 μs. The hyperthermophilic system was also simulated at a higher temperature, T=380 K, mimicking the working condition of the enzyme. The short-range interactions and the real space contribution of electrostatic interactions were cut off at 12 Å, while the long-range contribution of

electrostatic interactions were handled by the PME algorithm [175] with a grid spacing of 1 Å. All bonds involving hydrogens were constrained. The trajectories were recorded with a frequency of 4 ps.

3.2.3 REST2 (Replica Exchange with Solute Scaling)

In order to enhance the conformational sampling of the switch I region, we used a recent implementation of REST2 [123, 124], a Hamiltonian-exchange parallel tempering technique, see Chapter 2 for details. The REST2 algorithm is based on the rescaling of some terms of the potential energy of the system, namely the dihedral potential energy terms of the protein and the non-bonded protein-protein and protein-solvent interactions. This scaling may concern the protein in its entirety or a portion of it. Here we have thermally excited the switch I region of the two systems in the holo state. We used 12 replicas for the mesophilic systems and 16 for the hyperthermophilic ones as the thermophilic system contains both a larger protein and a larger number of water molecules, thus needing a larger number of replicas to achieve similar replica exchange efficiency (40 %). The replicas were allowed to exchange every 10 ps. According to a mean-field rescaling scheme [124], the replicas scanned an effective temperature window $T_{eff} \in [280K, 575K]$ for the fragment. The rest of the system was thermalized at the reference temperature of 300 K. In order to avoid finite-size effect on the sampling of the conformation of the switch I region, the proteins were solvated in a larger box with respect to those used in the unbiased MD set-up. A total number of about 14,000 and 18,000 water molecules were used to solvate the mesophilic and the hyperthermophilic proteins, respectively.

3.2.4 Conformational and Kinetic Clustering

The main analysis of the protein flexibility was based on conformational clustering [135, 45] that allows determining the number of representative conformational states visited by the system and the frequency of transitions between them, described in detail in Chapter 2. Networks of conformational states were built by using the root mean squared distance, fraction of native contacts, and fraction of native torsion angles as the clustering collective variable. The results of the conformational clustering are presented in the text in detail, while the results of the fraction of native torsion and the fraction of native contacts clustering are reported in Tables in the Appendix of this chapter.

The trajectories were saved with a frequency of 20 ps. Conformational networks were further clustered using a Markov Chain Algorithm [138], details described in Chapter 2. The granularity parameter was set to 2. Finally, the results of conformational and kinetic clustering were visualized as network of states by using a force-based algorithm as implemented in GEPHI [137].

3.3 Results and Discussion

3.3.1 Substrate Effects on Protein Conformations: the Mesophilic G-domain.

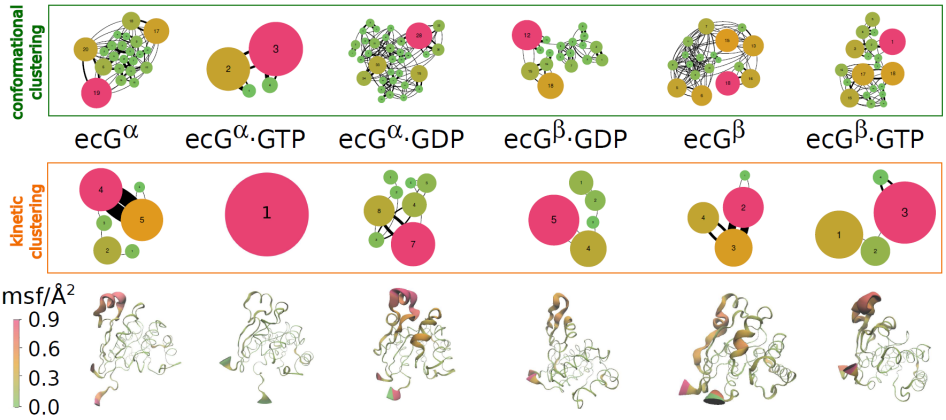

Figure 3.2: Conformational clusters shown in network representations for the protein with and without its ligands, with each node representing conformations with RMSD that differ by 2.5 Å. Each node of the network represents a conformation substate, the size of the node is proportional to its occupancy. The color scale in the network is used to further stress different occupancy of the conformational states, while the numbers inside the nodes reflect the temporal occurence of cluster leaders in a simulation, i.e. the first cluster leader is assigned the label '1', the second cluster leader is assigned label '2' etc. Nodes of kinetic networks show substates that are separated by high energy barriers, while a single node contains states separated by low energy barriers. The lowest panel represents the mean squared fluctuation, a measure of protein flexibility, shown in color and thickness of the protein backbone. Data refer to the MD simulations performed at T=300 K.

In this first section we analyze the conformational changes induced by a ligand bound to the mesophilic ecG. The results are based on long MD simulations extending to the microsecond time scale (0.6 μs). Namely, by following the hypothetical enzymatic

turnover of the protein, we have inquired into the equilibrium shift of protein conformational substates due to binding of the reactant (GTP) and product (GDP) molecules. The sequence of conformations accessed by the ecG during the EF-Tu activity are schematically represented in the bottom panel of Figure 3.1 (a), according to the resolved X-ray structures.

System	N(t_{sim})	N_∞	τ (ns)	Kinetic Clustering
ecG$^\alpha$	22	21.5	189.2	6
ecG$^\alpha$·GTP	4	4.2	228.5	1
ecG$^\alpha$·GDP(*)	14	33.0	1229.1	4
ecG$^\alpha$·GDP	34	37.0	208.7	8
ecG$^\beta$	18	21.9	327.7	4
ecG$^\beta$·GTP	20	19.7	20.3	4
ecG$^\beta$·GDP	18	20.4	249.0	5
ssG	24	23.3	200.0	4
ssG·GTP	12	10.2	114.8	1
ssG·GDP	14	21.5	500.7	4
ssG·GTP (T=380 K)	21	21.3	158.7	2
ssG·GDP (T=380 K)	111	1728	8923.2	14
ssG·GTP‡(T=380 K)	10	9.5	72.2	/
ssG·GDP‡(T=380 K)	40	46.4	316.1	/

Table 3.1: Conformational and kinetic clustering of the MD simulations. The conformational clustering was based on the collective variable RMSD and using a cut off of 2.5 Å. The total number of clusters obtained is indicated in the first column, $N(t_{sim})$. In the third and fourth column, we report the parameters of a simple exponential growth model fitting the data, $N(t) = N_\infty \cdot (1 - exp(t/\tau))$. In the last column we report the number of independent kinetic states as obtained by applying Markov state model based clustering algorithm with a threshold of 2.0. (‡)In the last two lines we report the results for the thermophilic ssG domain simulated at T=380 K but excluding the last 3 and 7 residues at the N- and C-terminals. At this high temperature, the terminals are not anchored to the body of the domain and their random motion gives rise to a linear growth of the number of clusters.

We have performed conformational and kinetic clustering of the long MD trajectories. The network representation of the complex conformational landscape was reconstructed highlighting both the population of the states and the frequencies of their interconversions, see Figure 3.2. We first discuss the results obtained for the "reactant" conformer ecG$^\alpha$. Most notably, the protein gets stiffer when GTP binds to ecG$^\alpha$. In fact, the number of conformational states accessed by the protein drops down by a factor of 5 as compared to the apo state, see also Table 3.1. The region mainly affected by the GTP ligand is

the switch I, which is highlighted in the bottom panel of the Figure 3.2, where we have magnified the portions of the protein matrix exhibiting larger flexibility, measured by the atomistic mean squared fluctuations (msf) of C_α. When the GDP is bound to the same initial ecG$^\alpha$ structure, the sampled conformational landscape is less confined, and the protein preserves its intrinsic flexibility. A very similar behavior was recovered when, as initial state, we considered an equilibrated configuration from the ecG$^\alpha$·GTP simulation and replaced GTP with GDP (simulation denoted ecG$^\alpha$·GDP(*)). Again, when the GDP molecule is bound to the domain, the flexibility of the protein is much higher than in the case of the GTP bound state (data shown in Table 3.1). Apart from the switch I region, the flexibility of the protein is mostly concentrated in the switch II region, similar to what was previously observed in a MD simulation of the entire protein in the apo and holo GDP states [176].

Transition	TΔS [kcal/mol]
ecG$^\alpha$ → ecG$^\alpha$·GTP	-20.8
ecG$^\alpha$·GTP → ecG$^\alpha$·GDP	16.4
ecG$^\alpha$·GTP → ecG$^\beta$·GDP	20.0
ssG → ssG·GTP	-11.0
ssG·GTP → ssG·GDP	8.3
ssG → ssG·GTP (T=380 K)	-13.9
ssG·GTP → ssG·GDP (T=380 K)	10.5

Table 3.2: Difference of backbone entropy estimated by the second order parameter S^2 for the bond vector N-H. The parameter is obtained by considering the time correlation function of the second order Legendre polynomial function $C(t) = (3(\mathbf{v}(t)\cdot\mathbf{v}(0))^2 - 1)/2$ where $\mathbf{v}$ indicates the NH bond vector. The data refer to the simulations at T=300 K, unless otherwise noted. The parameter S^2 is extracted for each residue by fitting $C(t) = S^2 + (1-S^2)e^{-t/\tau}$ [177]. Entropy difference between two states, a and b, is estimated according to the formula $\Delta S = -k_b \sum_i ln\{\frac{3-(1+8S_{a,i})^{1/2}}{3-(1+8S_{b,i})^{1/2}}\}$, where i runs over the residues of the protein [142].

It is important to note that the shift of protein stiffness upon ligand binding is also visible when considering the kinetic clustering of the trajectories on the basis of a Markov state model, see Methods section. This type of clustering helps to visualize protein substates separated by high kinetic barriers (Figure 3.2). The contribution of the backbone flexibility to the entropy changes in the catalytic cycle has been calculated estimating the S^2 order parameter of the NH bond reorientation [142, 177], see Chapter 2 and Table 3.2, and it adds extra weight to the results obtained by conformational

clustering, where the change in the number of cluster leaders implies a change in the probability of microstate occupancy, which is proportional to the overall entropy.

For the EF-Tu, a conversion from the α to the β conformers follows the GTP hydrolysis according to the literature [163, 173]. Therefore it is convenient to examine the reshaping of the conformational landscape of the "product" conformer ecG$^{\beta}$. In this case, for both ligands, the general effect of substrate binding on the global protein flexibility is very weak. In fact, no dramatic changes are observed in the conformational and kinetic clustering of the generated trajectories.

While the exact sequence of conformational inter-conversions is not known, and our data do not provide clues on the selective pathway along the enzymatic turnover, the conformational and kinetic clustering of the independent states suggest that heterogeneous kinetics for GTP hydrolysis could emerge as the effect of the conformational transitions [178]. These transitions would potentially filter substrate binding/unbinding as well as modulate the kinetic barrier of the chemical step. Understanding whether the transitions are specific to the mesophilic homologue or are shared with the more thermally stable ssG will help clarifying the molecular mechanism of the thermal activation of the hyperthermophilic homologue. This will be addressed in the next sections.

3.3.2 In Quest of the Conformational Transition in the Mesophilic G-domain.

We have mentioned that the transition from the α toward the β conformer is expected to occur upon GTP $\rightarrow$ GDP conversion. This conformational change is a biologically relevant signal as it triggers dissociation from the ribosome. Trapping the GDP state in the α configuration prevents this dissociation and thus stops the peptide translation, a fact exploited by some antibiotics [179]. The switch I region should therefore access the β state in the ecG$^{\alpha}$·GDP simulation, and similarly, transition toward the helical state is expected in the same region when GTP replaces GDP, i.e. in the ecG$^{\beta}$·GTP simulation. It should be mentioned that because of the small extension of the fragment of interest (P53-G59), it is difficult to observe extended β-strands. Even in the crystallographic structure, only two cross H-bonds are detected as β linkers. For this reason hereby our definition of β-state casts together both the presence of hairpin-like double strand and β turn-like conformation, similarly to [176].

Despite the high flexibility of switch I, the β secondary structure is poorly sampled in the simulation of the holo ecG$^{\alpha}$·GDP state. Similarly, the α-helical state in the P53-

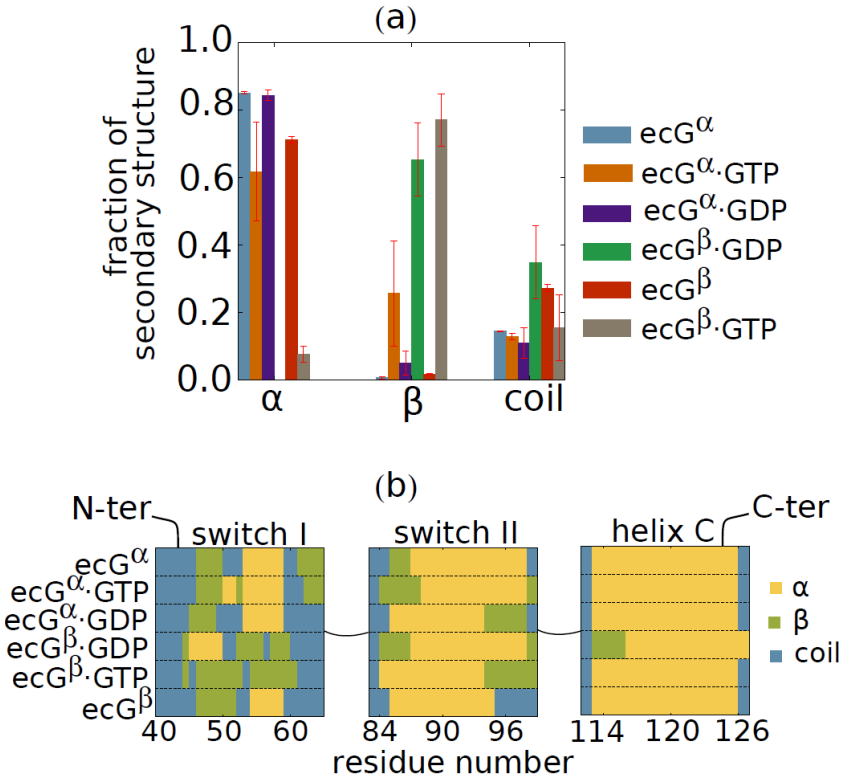

Figure 3.3: Percentage of secondary structure motifs for a part of the switch I region, residues P53-G59, that is reported to undergo a secondary structure change in the catalytic cycle. The bottom part of the figure shows the most occupied secondary structure per residue for three key regions in the protein, shown for different representative states of the protein during the catalytic cycle.

G59 portion of switch I is weakly populated in the ecG$^{\beta}$·GTP simulation, see Figure 3.3 (a). However, it is interesting to note that the switch I fragment acquires α-helix when ecG$^{\beta}$ is simulated in the apo state. This fact suggests that β to α conformational change during the enzymatic turnover may occur via the ligand-free state. The average secondary structure of two other key regions for protein activity is shown in Figure 3.3 (b): the switch II and helix C (residues P113-V125). The switch II is characterized by an extended helix (10 residues) and is found to rigidly shift upon GTP hydrolysis, one turn unfolds at the C-ter side and one refolds at the N-ter side [163, 180, 174]. This shift is for instance observed in the ecG$^{\alpha}$·GDP, but depending on the initial state, we remark that the extension of the helix is fluctuating across the simulations. The helix C represents the anchoring for the P-loop involved in nucleotide binding [181]. The helical structure is preserved in all the simulations meaning that any conformational change associated to substrate locking are caused by a rigid body motion of this region.

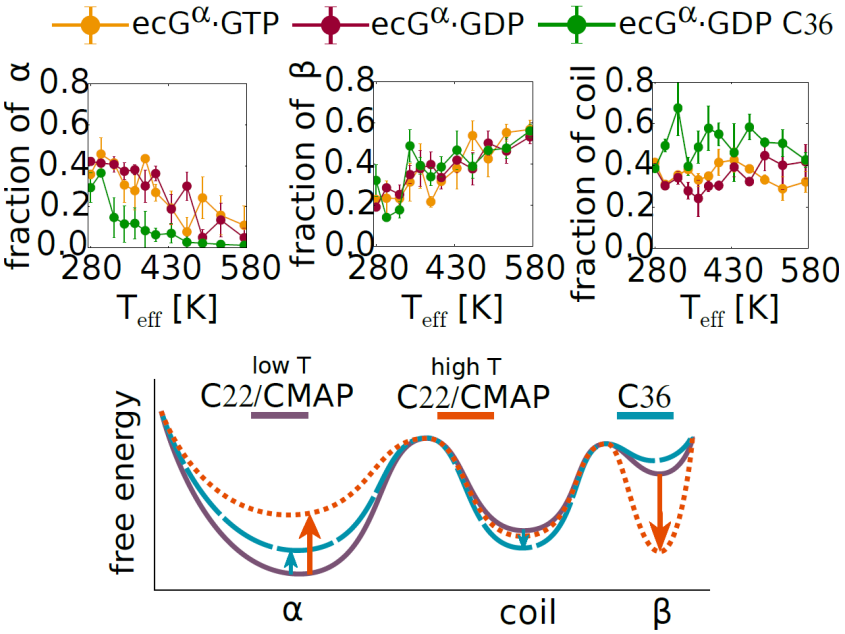

Figure 3.4: Enhanced sampling of the switch I region in the REST2 simulations. In the top panel, we report the fraction of secondary structures (α, β, coil) in the fragment as a function of the effective temperature exciting the switch I. Lower panel schematically compares data obtained from the simulations of the holo state ecG$^{\alpha}$·GTP(GDP) at different temperatures, based on CHARMM22/CMAP and CHARMM36 force field.

The lack of $\alpha \rightarrow \beta$ transition in the brute force MD of the holo ecG$^{\alpha}$·GDP state could depend on several factors: i) the presence of a high kinetic barrier separating the two states that would confine the sampling in the initial state only, ii) an intrinsic bias of the force field used, for instance it is widely reported that CHARMM22/CMAP favors helical states [96, 107], iii) the lack of inter-domain interactions in our model, in fact the crystallographic evidence of the transition was based on the resolution of the structures of the whole EF-Tu protein, iv) a temperature effect, since the quench into the two separate states could be caused by the low temperature at which X-ray experiments are performed. In the following we will address some of these issues.

We have performed enhanced sampling Hamiltonian-Replica Exchange simulations [123, 124] designed to "thermally" excite the switch I region of the protein. The simulations were performed on both holo states of the ecG$^{\alpha}$ conformers using 12 replicas for each simulation, see Figure 3.4 (for ecG$^{\beta}$ see Figure 3.5). According to a mean-field rescaling scheme [124], the sampling allowed to scan the effective temperature (T$_{eff}$) window 280-580 K. In Figure 3.4 we report the fraction of secondary structure accessed by the switch I as a function of the effective temperature T_{eff}. We observe that for

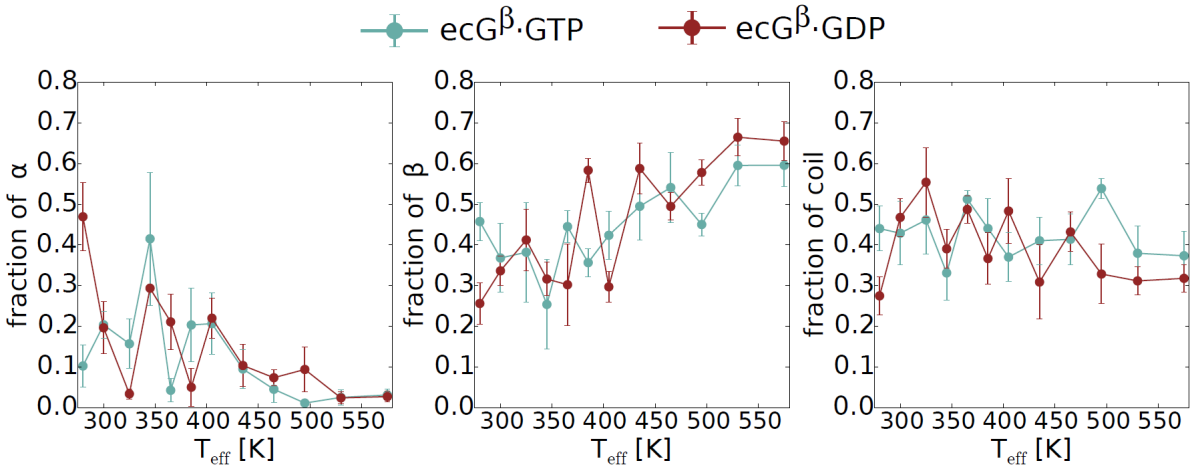

Figure 3.5: Enhanced sampling of the switch I region in the REST2 simulations. We report the fraction of secondary structures (α, β, coil) in the fragment as a function of the effective temperature exciting the switch I. Data refer to the holo state built from the conformer ecG$^{\beta}$.

both the GTP and GDP ligands, the α conformation is the most populated at ambient condition, and its occupation decays with temperature. The most important result is the meaningful fraction (20%) of β-like state - mainly turn - that can be accessed in both the holo states, and starting from both conformers (see Figure 3.5). This finding indicates that the α to β transition can occur, although the details of the associated kinetics would require ad-hoc calculations and will be reserved for further work. The population of the β-like structure increases as a function of temperature and compensates the thermal instability of the helical structure. The high, and almost temperature independent, fraction of unstructured coil state (~40%) further confirms the intrinsic flexibility of the fragment.

In order to account for the force-field dependence of the secondary structure propensities and their relative temperature changes, we have performed simulations of the ecG$^{\alpha}$·GDP state using the CHARMM36 force field, which was designed to give a better helix-coil balance [105]. While the helix state is much less populated, the population of β-like conformation is unchanged when compared to CHARMM22/CMAP. Overall, CHARMM36 renders the fragment highly unstructured, with the coil population being as high as ~50%.

A better sampling of the Hamiltonian-Replica Exchange could be achieved by extending the simulation time per replica, however despite the force field effects, the obtained results clearly show that the α to β conversion in the switch I region is possible. The

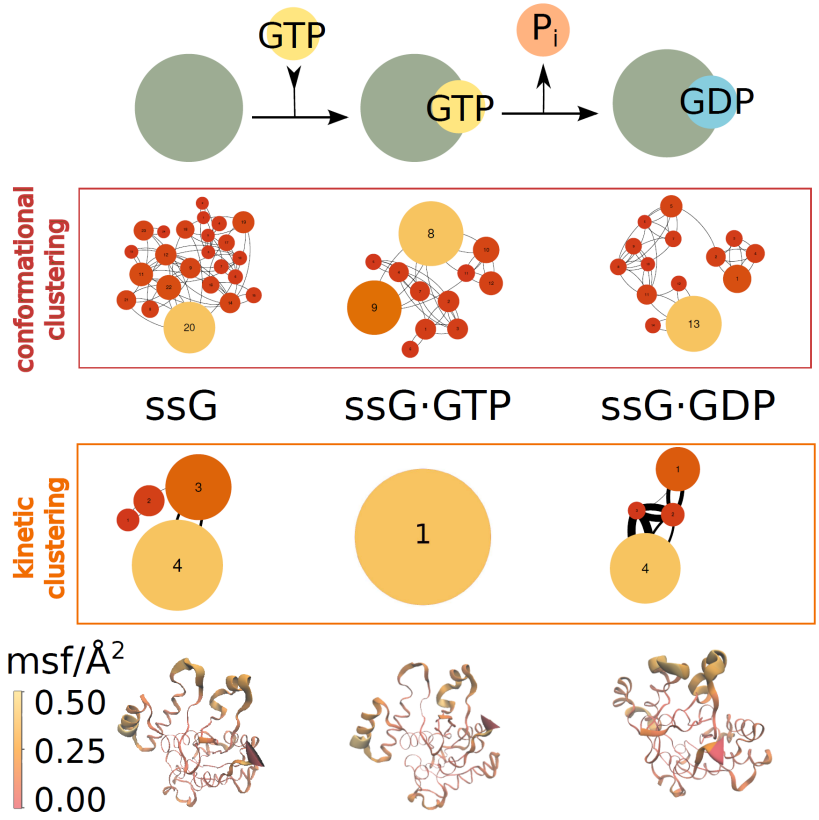

Figure 3.6: Binding the GTP or GDP to the EF-1α changes the number of representative conformational substates as shown by both conformational and kinetic clustering. The lowest panel shows the amplitude of the mean squared fluctuations of the protein backbone, coded in thickness of the backbone representation and color. Data refer to the MD simulations that were performed at T=300 K.

temperature effect on this transition can not be rigorously estimated because of present force field inaccuracies.

3.3.3 The Holo States of the Hyperthermophilic G-domain.

In a previous work [47, 45, 182] it was observed that the hyperthermophilic apo state ssG spans a comparable and even larger conformational space than the mesophilic variant ecG, and shows well-defined substates separated by high kinetic barriers. The enhanced flexibility is due to the rigid body motion of the highly structured switch I region. Similarly to the case of the homologous mesophilic ecG, the binding of the reactant GTP molecule to the hyperthermophilic domain quenches the protein flexibility. The number of conformational states explored on the same time-scale reduces by a factor

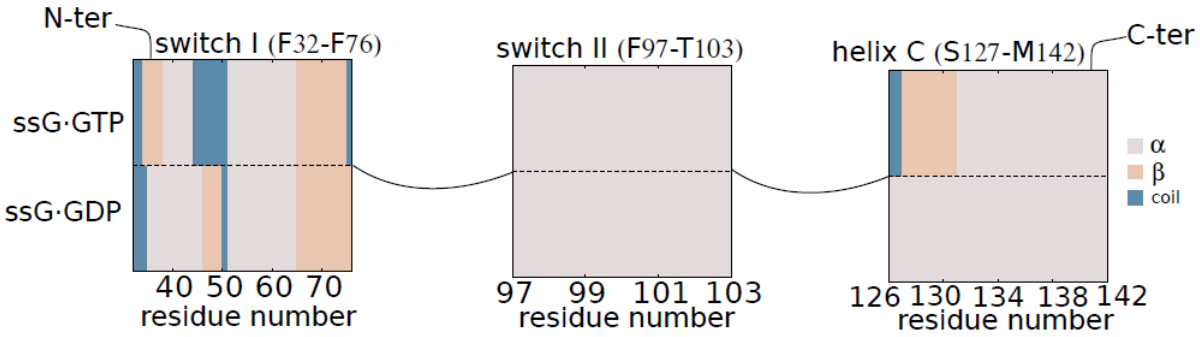

Figure 3.7: Most occupied secondary structure per residue for three key regions in the protein. Data refer to the simulations of the two holo states, ssG·GTP and ssG·GDP.

of 2 with respect to the apo state (Figure 3.6 and Table 3.1). In the product holo state ssG·GDP, the protein gets slightly more excited, and it partially recovers its intrinsic flexibility that allows sampling a larger number of kinetically relevant states. This conformational flexibility caused by the GTP hydrolysis localizes at the level of the switch I and II regions, see the bottom part of Figure 3.6, and is much less pronounced at ambient temperature than in the mesophilic variant. The secondary structure patterns of these regions are quite insensitive to the ligand hydrolysis, see Figure 3.7, although the switch I of the GTP-bound state cages the ligand more extensively by linking the triphosphate tail.

The increase of conformational entropy in the protein matrix upon catalysis is a signature of functional efficiency since it relates to both substrate unbinding and long-range communication. Therefore, the weak excitation observed at ambient temperature following the virtual GTP hydrolysis could correlate to the known low activity of the hyperthermophilic domain at ambient condition. Actually we see that at high temperature (T= 380 K), slightly exceeding the optimal growth temperature of the *S. solfataricus* archeon, the flexibility of the ssG·GDP is enhanced when compared to ssG·GTP, an increase by a factor of 5, similar to what found for the mesophilic domain when comparing the reactant ecG$^{\alpha}$·GTP and the product ecG$^{\beta}$·GDP states at ambient condition, see Table 3.1. The high temperature release of excitation in ssG·GDP concentrates again mainly in switch I and switch II regions.

3.4 What Specializes the Thermophilic G-domain

In this final section, we present a comparative discussion of the molecular factors that cooperate during the enzyme activity of the G-domain in particular, and of the EF

protein in general. The focus is firstly placed on the behavior of the switch I region considered an essential element of the protein matrix to regulate both the ligand binding kinetics and long-range communication upon catalysis [183]. This region has also been pointed out as the weak spot of the mesophilic ecG domain, where the early steps of thermal unfolding take place [45]. In the hyperthermophilic variant ssG, the same region is structurally stabilized by the insertion of two extra small α-helices, α' and α''. By performing enhanced sampling on the two homologues in their holo states via REST2, the stability of the region upon thermal stress has been assessed. In Figure 3.8 we compare the stability curve obtained for three secondary structure states populated by the fragment in the reactive states of the mesophilic and hyperthermophilic domains, the ecG$^{\alpha}$·GTP and ss·GTP. The mesophilic fragment is not only systematically more flexible at all temperatures, as seen from higher content of coil, but more importantly, its helical component is shown to be significantly less stable than that of its hyperthermophilic counterpart. The switch I region loses half of its initial fraction of helical content at about T_{eff}=400 K in the mesophilic ecG$^{\alpha}$·GTP, while for the hyperthermophilic domain, the helical disruption occurs at much higher value of the effective temperature exciting the fragment. These data confirm that the stability and function of the ssG domain are granted by the more robust structure of its switch I region.

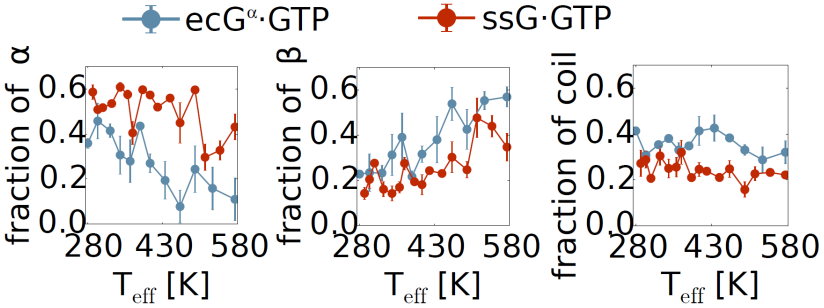

Figure 3.8: Fraction of secondary structure in the switch I region as a function of temperature for the holo state of the mesophilic and hyperthermophilic domains when bound to the reactive substrate GTP, ecG$^{\alpha}$·GTP and ssG·GTP, respectively.

A second motif, structurally conserved across EFs G-domains, and more broadly in NTPases [184], acting as molecular gate for substrate binding and unbinding is the P-loop [181]. The changes in flexibility caused by the substrates is obtained by clustering the conformations explored by this short fragment in the apo and holo states. As it was observed for the global behavior of both proteins, the number of states visited by the P-loop is strongly reduced when the GTP is bound, see Figure 3.9. The rigidification of

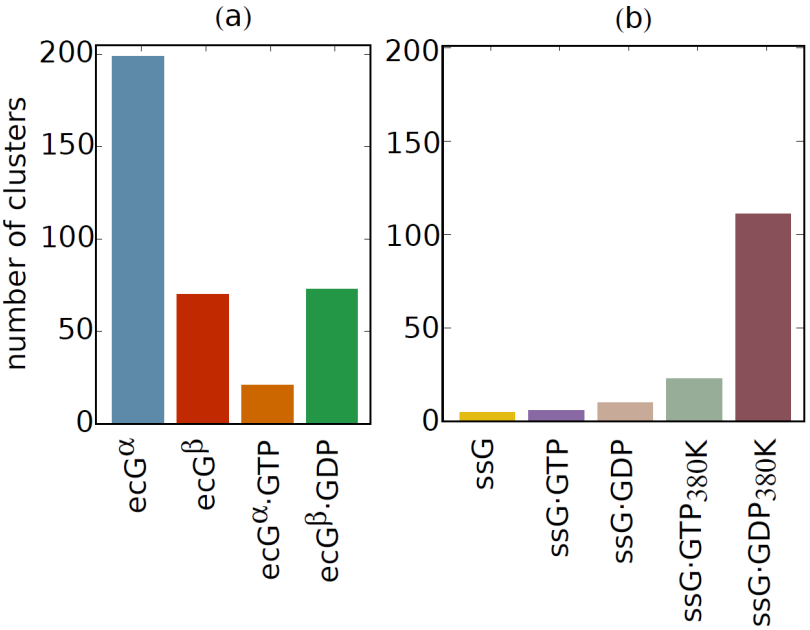

Figure 3.9: Number of conformational states of the P-loop obtained by cluster analysis of the fragment. Panel (a) refers to the mesophilic domain, panel (b) to the hyperthermophilic domain. RMSD is used as the collective variable in the clustering, with the cut off of 0.5 Å.

the P-loop is the consequence of an extended network of interactions formed with the GTP substrate. This connectivity is graphically represented in Figure 3.10, for both ecG and ssG.

In the mesophilic variant, the rigidification of the P-loop is alleviated when the substrate is changed in GDP and when the product β conformer is considered. This mimics the effect of GTP hydrolysis. The recovered flexibility of the loop is the result of a cooperative effect involving the cleavage of the final phosphate bond of the triphosphate tail, and the associated conformational change of the switch I region, $\alpha \rightarrow \beta$ [181]. The excitation of the loop flexibility upon GTP toward GDP conversion is also found in the hyperthermophilic variant ssG, although no secondary structure change occurs at level of the switch I. For ssG, the flexibility gap of the switch I between the GTP and GDP bound states further increases when considering the simulations at high temperature (T=380 K). It is important to stress that only at high temperature the P-loop in ssG·GTP(GDP) exhibits a flexibility comparable to that of the mesophilic domain at ambient condition. This is an indication of the validity of the corresponding state principle for the involved degrees of freedom. However, a precise connection of the observed variabilities of the P-loop flexibility with the dissociation kinetics of the product molecule GDP is beyond

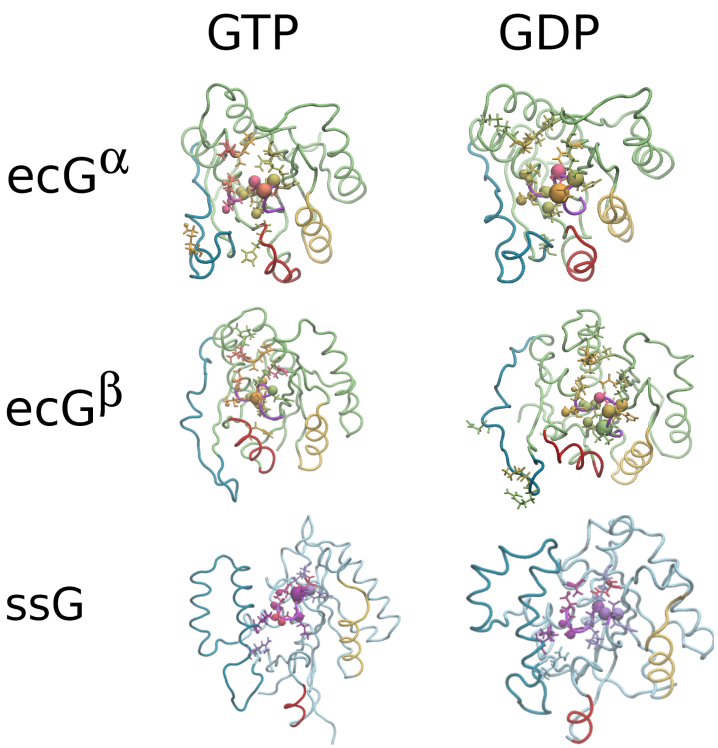

Figure 3.10: The occupancy of hydrogen bonds formed with the ligand in the protein binding pocket. The sidechains that form hydrogen bonds with the ligand are represented in ball-and-stick style, and the radius and color of the ball is equal to the proportion of hydrogen bond existence in the total simulation time. Different backbone elements are also emphasized by color coding (blue - switch I, red - switch II/helix B, yellow - helix C, purple - P-loop). We used a geometrical definition to identify the hydrogen bonds, the donor-acceptor distance cut off is set to 3.5 Å, and the hydrogen bond angle limit to 120°.

the scope of the present work since it requires a specialized approach. For example, it is worth mentioning that when considering the dissociation kinetics of the mant-GTP molecule from EF-Tu, it was surprisingly found that a mutation supposed to increase the local flexibility of the P-loop actually slows down the dissociation kinetics as compared to the wild-type protein [181]. This slowing down is the result of a delicate entropy/enthalpy compensation. The *in silico* estimate of the enthalpic and entropic contributions to the free energy barrier controlling the substrate dissociation kinetics is challenging because of the difficulties to individuate correct reaction coordinates for the process of interest, and to perform correct sampling.

We conclude by inspecting the correlation between the orientation of His84, a universally conserved residue in translation GTPases [185, 161, 156], and the dynamics of the so-called hydrophobic-gate (Val20 and Ile61) [160, 161]. The residue His84 is considered

as being a key residue for GTP hydrolysis and several point mutations of this residue in EF-Tu from *E. coli* showed anti-catalytic effects [156, 185]. It is however not clear if the role for catalysis is direct, i.e. by activating a water molecule for a nucleophilic attack on the γ-phosphate [180], or indirect, i.e. by helping the conformational rearrangement of the catalytic site upon ribosome-binding [156]. In a cryo-electron microscopic map of the aa-tRNA·EF-Tu·GDP·kirromycin bound to ribosome, and reconstructed by the help of atomistic modelling [160], the position of the His84 toward the GTP substrate was correlated to the opening of the hydrophobic gate, see PDB 4V69. Although in our simulations we lack the effect of ribosome binding, we explored the dynamics of the hydrophobic gate and of the orientation of His84. The former was monitored by the distance between the sidechain centre of mass of the two residues, and the latter by measuring the distance between the His84 sidechain centre of mass and the P_β of GTP and GDP molecules. The analysis is extended to the hyperthermophilic domain where upon structural superimposition, we identified analogous residues [165, 186].

In Figure 3.11 (a) and Figure 3.11 (b) we report the two dimensional probability distributions of both distances in GTP and GDP bound states for the two homologous domains ecG and ssG, respectively. For each protein and each holo state, in the top panels we report the data from MD trajectories and in the bottom panels that from Replica Exchange simulations. In the mesophilic domain, we find that although the hydrophobic gate is always in the open state, the His84 is oriented quite far from the GTP molecule. In the cryo-electron microscopy derived structure (PDB code 4V69), the gate distance is about 12 Å, and His84 approaches P_β up to 5 Å, see the symbol in Figure 3.11. Our results show that in the isolated mesophilic G-domain, the orientation of His84 and hydrophobic gate are uncorrelated. Interestingly, when moving to the hyperthermophilic ssG, we find that at ambient temperature the analogue of the mesophilic His84, His94 (see sequence alignment in [165]), is localized far from the catalytic site, at a distance preventing any direct contribution to the GTP hydrolysis. Only at high temperature does the distance decrease to values $\sim$ 6 Å, supporting a possible contribution to the catalysis. Putative analogue of the mesophilic hydrophobic gate in ssG [186] is always found in open state in our simulations, shown with a dashed line in the figure.

3.5 Conclusions

In this Chapter, we have investigated the effect of substrate binding on the conformational flexibility of two homologous GTPase domains of different stability content. In

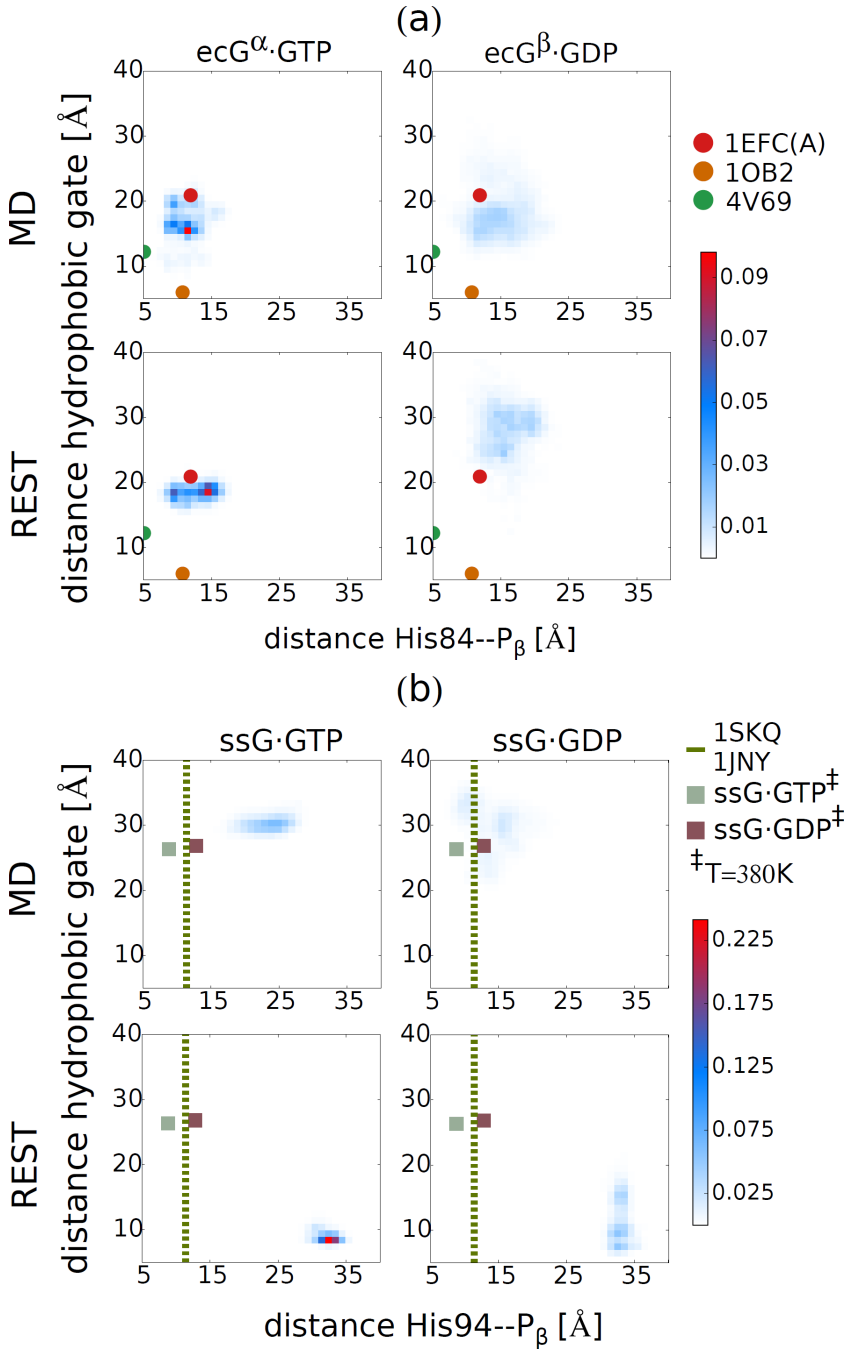

Figure 3.11: 2D probability distribution of the hydrophobic gate and His84(94)-P$_{\beta}$ distances. The top chart refers to the mesophilic domain while the bottom chart to the hyperthermophilic domain. For each domain, data are reported for the GTP bound state (left panels) and for the GDP bound state (right panels). For each system we compare results from MD (top panels of a and b) and REST2 (bottom panels of a and b). Symbols refer to the distances measured in the crystallographic structures indicated in the legend by their PDB codes. For the hyperthermophilic domain, the average value of the hydrophobic gate distance extracted from the simulation at T=380 K is also reported. The dashed line represents the average His94-P$_{\beta}$ distances in both 1SKQ and 1JNY crystal structures which are the same.

both homologues, the flexibility of the apo protein is readily quenched when the reactant GTP molecule binds to the protein, but conversely recovered when the virtual hydrolysis is mimicked by considering the GDP bound state. The flexibility changes are localized at the level of structural key motifs, the switch I, switch II, and the P-loop. The magnitude of the entropy released in the protein matrix upon the reaction differs between the mesophilic and the hyperthermophilic enzymes. For the latter, a flexibility of switch I and of P-loop comparable to the mesophilic variant is only attained at high temperature. This finding confirms the validity of the corresponding state principle for these modes, despite the fact that the apo hyperthermophilic domain shows comparable, if not enhanced, flexibility with respect to the apo mesophilic domain. As a final remark, we point out that the stability/function trade-off in the two species relates to the different structure of the switch I region. In the mesophilic domain, the high flexibility of the fragment allows for a secondary structure rearrangement along the functional process, but at the same time renders the fragment highly unstable in temperature. On the other hand, in the hyperthermophilic species, the insertion of extra secondary structure motifs renders the fragment more resistant to temperature, reflecting once and again, the evolutionary pluralism in optimizing the function in different temperature regimes.

APPENDIX A

APPENDIX OF CHAPTER 3

System	N(t_{sim})	N$_\infty$	τ (ns)
ecG$^\alpha$	46	45.2	61.4
ecG$^\alpha$·GTP	6	6.6	195.6
ecG$^\alpha$·GDP(*)	10	9.3	89.3
ecG$^\alpha$·GDP	11	10.6	62.3
ecG$^\beta$	896	1289.0	540.6
ecG$^\beta$·GTP	13	13.9	168.2
ecG$^\beta$·GDP	16	16.9	144.4
ssG	4	4.1	44.9
ssG·GTP	3	5.7	815.4
ssG·GDP	7	7.9	229.5
ssG·GTP (T=380 K)	19	18.1	45.3
ssG·GDP (T=380 K)	47	44.6	133.2

Table A.1: Fraction of native contacts clustering of the MD simulations. The conformational clustering was based on the collective variable $Q(t)$ formally defined as $Q(t) = \frac{1}{N_{C_\alpha}} \sum_{i=1}^{N_{C_\alpha}} \frac{l_i(t)}{l'_i}$, where N_{C_α} is the number of carbon alphas, l'_i is the number of native contacts given by the number of carbon alphas within 8 Å cut off from the C^i_α in the reference state, and $l_i(t)$ is the number of native contacts calculated for the configuration at time t. Clustering cut off used was 0.35. The total number of clusters obtained is indicated in the first column, $N(t_{sim})$. In the third and fourth columns, we report the parameters of a simple exponential growth model fitting the data, $N(t) = N_\infty \cdot (1 - exp(t/\tau))$.

System	N(t_{sim})	N$_\infty$	τ (ns)
ecG$^\alpha$	34	33.5	250.4
ecG$^\alpha$·GTP	9	10.9	377.7
ecG$^\alpha$·GDP(*)	36	52.1	609.1
ecG$^\alpha$·GDP	25	26.4	217.1
ecG$^\beta$	174	332.7	804.7
ecG$^\beta$·GTP	21	84.9	1909.3
ecG$^\beta$·GDP	34	45.8	470.4
ssG	21	19.8	136.5
ssG·GTP	8	7.3	113.7
ssG·GDP	11	32.9	1423.3
ssG·GTP (T=380 K)	77	76.6	212.6
ssG·GDP (T=380 K)	149	180.1	373.2

Table A.2: Fraction of native torsion angles clustering of the MD simulations. The conformational clustering was based on the collective variable that can be expressed as $n_t(t) = \frac{1}{N_\theta}\sum_{i=1}^{N_\theta} \exp[-\frac{(\theta_i(t)-\theta_i')^2}{\sigma^2}]$, where $|\theta_i(t)-\theta_i'| < 180°$, $\sigma = 60°$, N_θ is the number of torsion angles θ, θ_i' are the torsion angle values in the reference state, and $\theta_i(t)$ are the number of torsion angles in the configuration at time t. Both ϕ and ψ dihedrals were used in the calculations. Clustering cuto ff used was 0.20. The total number of clusters obtained is indicated in the first column, $N(t_{sim})$. In the third and fourth columns, we report the parameters of a simple exponential growth model fitting the data, $N(t) = N_\infty \cdot (1 - exp(t/\tau))$.

CHAPTER 4

THERMAL RESPONSE OF MESOPHILIC AND THERMOPHILIC DEHYDROGENASE HOMOLOGUE

In this chapter, we inspect the thermal activation of protein soft modes by combining Neutron Spin-Echo scattering experiments and Molecular Dynamics simulations. The ultimate goal is comparing the thermal response of functional modes in a pair of homologous tetrameric Lactate/Malate Dehydrogenase proteins. The Neutron Spin-Echo scattering can probe motions at nano length- and timescales that are relevant in large protein conformational reorganization, while the Molecular Dynamics simulations support the microscopic interpretation of the experimental spectra. The results obtained for the mesophilic species, the eukaryotic Lactate Dehydrogenase from Rabbit muscle 5 (LDH M5), will be extensively presented. The analysis of the experiments and simulations for the thermophilic homologue are still under way. For the Lactate Dehydrogenase, we probed the thermal activation of functional modes spanning the lengthscales of interdomain separations, matching the allosteric reorganization previously probed for bacterial LDHs. A manuscript presenting the results for the LDH protein is in preparation.

4.1 Introduction

Protein dynamics and functionality are intimately related. Despite general consensus, the fine details on how protein conformational changes modulate and regulate activity is constantly debated [73, 55, 187, 188]. It is now accepted that protein dynamics is

characterized by a hierarchy of timescales, from picoseconds to microseconds, reflecting a rough manifold conformational landscape [189, 190, 191]. Great effort has been devoted to link these dynamics to functionality, i.e. substrate binding/unbinding kinetics [70], allosteric relaxation [192, 141], and catalysis [60, 178]. While so far the Nuclear Magnetic Resonance [191] and single molecule spectroscopy represented the privileged means of investigating these phenomena [178, 152], recent development of the Neutron Scattering spectroscopy paved the way for further exciting applications in exploration of protein soft modes at the nanometer and nanosecond scales [150, 151, 193, 194]. Specifically, the Neutron Spin Echo (NSE) spectroscopy has been applied to systems exhibiting long-range signaling modes via domain displacement, as in the case of the NHERF1 [193], Taq polymerase [150], and Phosphoglycerate Kinase [194], as well as to a more compact multimeric protein Alcohol Dehydrogenase [151]. The studies conducted so far have combined experiments and molecular modelling, e.g. normal mode (NM) analysis, and molecular dynamics simulations (MD), to determine the contribution of specific functional modes along with the changes associated to substrate binding. While quasielastic NS is routinely used in monitoring the thermal variation of the protein atomistic fluctuations [129, 66, 195], this is the first time NSE is used in probing the temperature response of a protein system in the nanoscale regime. In our investigation, we have combined NSE spectroscopy and MD simulations in order to investigate the thermal activation of the soft modes in two tetrameric homologous proteins in the apo state, the Lactate Dehydrogenase from rabbit muscle 5 (M5) (optimum growth temperature range 37.7-39.4 ^{o}C) and the Malate Dehydrogenase from *Methanocaldococcus jannaschii* (optimum growth temperature range 48-94 ^{o}C [196]).

Seminal study on LDH M5 [197] showed the activation of the enzyme in the presence of pyruvate, while enhanced activity is measured upon high-temperature incubation. Conformational changes accompanying the enzymatic turnover are expected to occur, as in the case of the majority of the bacterial LDHs that are allosteric in a proper sense [198], the latter is extensively verified through the fructose 1, 6-bisphosphate allosteric regulation. By comparing the crystallographic structures of the bacterial LDHs in apo and holo states, the reorganization of key parts of the protein matrix upon substrate binding was detected [199, 200, 201, 202]. The structural reorganization of the tetramer includes movements of various amplitudes, e.g. the closure of the active site loop, the rearrangements of several mobile regions (MR), and the favorable positioning of catalytic residues. Moreover, the active site loop gating was found to be the rate-limiting step for the catalysis of the wild type bacterial LDH from *B. Stearothermophilus* [203]. The

paramount importance of the loop region in the catalysis of MDHs is reflected by its high conservation across the protein family members [204]. Importantly, as pointed out by experiments and theoretical calculations, the kinetic heterogeneity of the enzymatic activity is related to conformational fluctuations of this loop [205, 206, 207].

The presence of long-range correlated motions potentially relevant for functionality makes the NSE an advantageous/pertinent choice. Moreover, the protein crystallographic structure has been solved, offering a good starting point to carry out MD simulations, spanning the protein motion at the atomistic resolution.

4.2 Methods

4.2.1 Neutron Spin Echo

Neutron Spin Echo (NSE) spectroscopy (see Chapter 2) extends the experimental time resolution of Inelastic Neutron Scattering to hundreds of nanosecond by encoding the velocities of polarized neutrons in their precession motion across a highly homogeneous magnetic field [208]. Unlike conventional Inelastic Neutron Scattering, the NSE technique returns the Intermediate Scattering Function $I(Q,t)/I(Q,0)$ directly in the time domain.

The experiments were performed on the J-NSE spectrometer at the FRM-II reactor in Munich at two wavelengths, 8 and 10 Å, giving a maximum spin echo time of 65 ns in Q-range $0.037 < Q < 0.214$ Å^{-1}. The LDH concentration in the D_2O buffer was 90 mg/ml, a concentration well higher than that corresponding to the dilute regime, necessary to achieve a sufficient Spin Echo signal. Note that the LDH protein is soluble even in this high concentration regime and that an exchange between deuterium and hydrogen happens on the protein surface, but that all atoms contribute to the experimental signal and the exchange process is thus not explicitly treated. Both the protein solution and the buffer alone were measured at the same experimental conditions to allow background subtraction. The samples were investigated at three temperatures: 283 K, 298 K, and 313 K.

4.2.2 Small-Angle X-Ray Scattering

Small-Angle X-ray Scattering (SAXS) experiments were performed at the high-brilliance beamline ID02 at the European Synchrotron Radiation Facility (ESRF) in Grenoble, France. A sample to detector distance of 2 m was chosen to cover a Q-range $0.05 \leq$

Q ≤ 3 nm^{-1}. The incident X-ray wavelength λ was 0.1 nm. The measurements were performed in a Peltier-controlled flow-through capillary of 1.8 mm diameter to minimize beam damage of the samples and to ensure an accurate subtraction of the background (buffer solution). The two-dimensional scattering patterns were recorded using a Rayonix MX-170HS fiber-optic taper coupled charge-coupled device camera. The two-dimensional spectra were normalized to absolute intensity scale after applying the detector corrections for spatial homogeneity and linearity. Normalized SAXS patterns were azimuthally averaged to obtain the one-dimensional scattering profiles [$I(Q)$ vs. Q]. The background corrected protein SAXS curves are displayed in the left panel of Figure 4.1 as a function of Q, after rescaling by the concentration C (mg/ml).

The scattering $I(Q)$ of a solution of N proteins is proportional to the product of the structure factor $S(Q)$, associated to the concentration-dependent interaction between different proteins, and the form factor $F(Q)$, which accounts for the spatial correlations of the atoms within the single protein and is independent of the concentration, $I(Q) \sim NS(Q)F(Q)$. At low concentration, the interaction is negligible ($S(Q) \sim 1$), so that one can estimate $F(Q)$ directly by extrapolating the concentration-dependent scattering to $C = 0$. On the other hand, at higher concentrations, a drop of the scattering intensity at low Q is caused by the interparticle interactions, i.e. the $S(Q)$ is in turn extracted by the ratio $I(Q,C)/(C \cdot I(Q,C = 0))$. As expected, at high Q, the scattering curves are unaffected by the interaction and overlap.

4.2.3 Dynamic Light Scattering

Dynamic Light Scattering (DLS) measurements were carried out using an ALV CGS-3 Compact Goniometer equipped with a HeNe Laser with a wavelength of 632.8 nm, a 22 mW output power, and an ALV LSE-5004 Correlator. Samples were measured at a scattering angle of 90^o, while the sample temperature was controlled via an external waterbath circulator.

4.2.4 Molecular Dynamics Simulations

The protein structure of the LDH M5, as obtained by X-ray scattering (PDB 3H3F, chains E, F, G, H), was embedded in a simulation box containing 34275 water molecules, producing a system of 123,627 particles including counter-ions. We used the NAMD 2.9 software [89] and the CHARMM22/CMAP force-field [94] to perform all-atom simulations in the NPT ensemble using Periodic Boundary Conditions. Water was modelled by the

TIP3/CHARMM model [91]. Four simulations were carried out at P=1 atm and T=[283 K, 298 K, 313 K, 330 K], where the first three temperatures match those in experiments, and the fourth further probes the high-temperature regime. The integration time step was set to 2 fs. The non-bonded interactions and the short range electrostatic interactions were cut off at 9 Å, while the long range electrostatic interactions were treated with the PME algorithm with a grid spacing of 1.3 Å. After initial equilibration, the simulations were run another 0.6 μs, and the latter were used for analysis purposes. The trajectories were recorded with a frequency of 5 ps. The simulations of the thermophilic variant MDH (PDB 1HYG) were conducted using the same protocol, and the analysis is under way at the moment of the redaction of the thesis.

4.2.5 Analysis of Molecular Dynamics Trajectories

4.2.5.1 Conformational and Kinetic Clustering

As a probe in exploring the change in protein flexibility as a function of temperature, conformational and kinetic clustering were employed, see Chapter 2 for details. In this work, we used the root mean square deviation (RMSD) among conformations as an order parameter to distinguish the different substates with a cut off of 1.5 Å. The kinetic networks were generated by setting the granularity parameter to 2. The networks of substates obtained by both clustering strategies are graphically visualized by using a force-based algorithm implemented in GEPHI [137].

4.2.5.2 Principal Component Analysis

To decompose the complex dynamics of the multimeric LDH to independent motions, we used Principal Component Analysis (PCA) [209], see Chapter 2. By using the PCA, the principal modes on which the larger fluctuations of the protein motion is concentrated can be individuated and additionally, a virtual trajectory projected on these modes can be obtained (Eq. 2.53) in order to quantify the contribution of these modes to the Neutron Scattering spectra. The procedure was performed by exploiting the analysis suite available through the Gromacs package [145].

4.2.5.3 Diffusion Coefficient in Harmonic Approximation

In order to characterize modes relevant for functionality, we estimated the diffusivity associated to particular elements of the LDH, namely for the distance between each

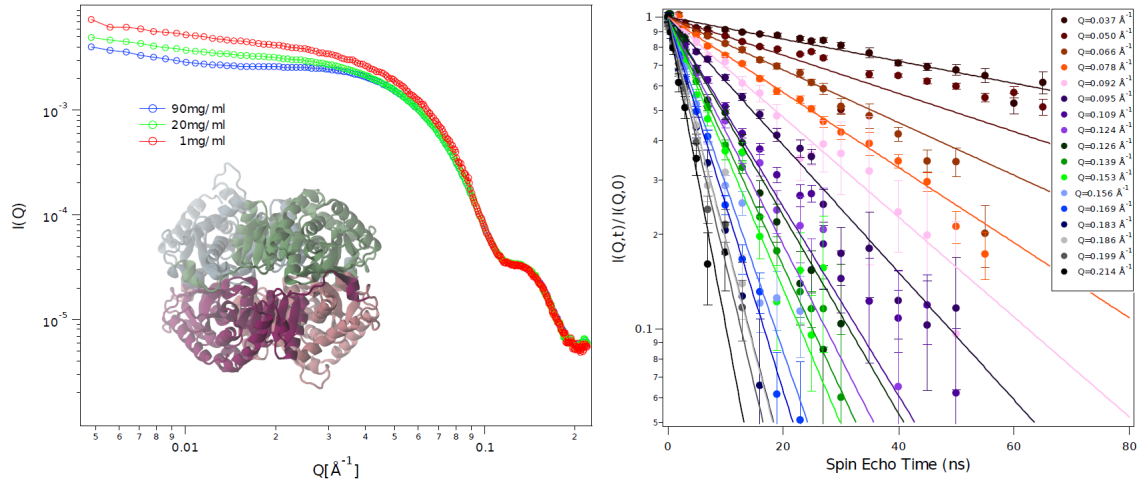

Figure 4.1: Left panel shows the SAXS spectra as a function of Q, rescaled to the protein concentrations, and a graphical representation of the protein structure, where different colors correspond to different subunits. Right panel shows the intermediate scattering function $I(Q,t)/I(Q,0)$ as a function of the spin echo time (circles) measured at T=298 K, shown for different Q along with the exponential fits to the data (lines). The Figure is reported by the courtesy of M. Maccarini (University Grenoble Alpes, Grenoble).

domain pair (α,β) of the catalytic loops (Ala95-Arg105). The collective variable (CV) combining the distances between the C_α atoms in the loops is first designed:

$$d_{\alpha,\beta}(t) = \left(\frac{1}{N_\alpha N_\beta} \sum_i^{N_\alpha} \sum_j^{N_\beta} |r_i - r_j|^2 \right)^{1/2}, \tag{4.1}$$

with the sum running over the C_α atoms in the loops α and β. The diffusion of the relative collective distance between the loops, fluctuating around their equilibrium values as seen in Figure B.1 in the Appendix of this chapter, is estimated in the harmonic approximation. The exponential fit of the time correlation function of $d_{\alpha,\beta}$ is shown in Figure B.2 in the Appendix of the chapter.

4.3 Protein Diffusion

In this section we present the results on the effective protein diffusion as obtained by Spin-Echo spectroscopy. The short time decay of the $I(Q,t)/I(Q,0)$ can be approximated by a cumulant expansion given in Eq 2.63, from which the diffusion coefficient can be extracted by using Eq 2.64. The Q-dependent diffusion of the protein can therefore be

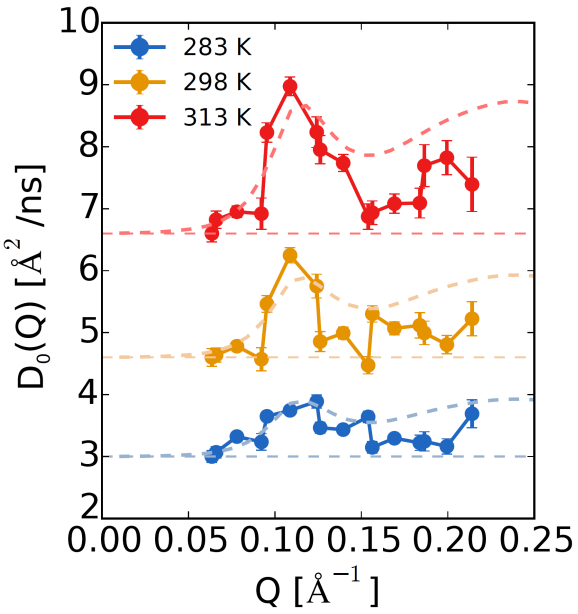

Figure 4.2: Diffusion spectra of LDH at three different temperatures. Circles indicate the experimental points, the horizontal dashed lines indicate the value of the translational diffusion evaluated by DLS measurements. The dashed line curves indicate the Q-dependent diffusion constant calculated for a rigid-body (X-ray structure) and using the mobility tensors D^R and D^T calculated by the HYDROPRO program.

extracted by fitting the normalized intermediate scattering function $I(Q,t)/I(Q,0)$, as shown in Figure 4.1. The lines correspond to the exponential fit performed on the short time decay, 0.1-10 ns, of the $I(Q,t)/I(Q,0)$ set of functions obtained at T=298 K. The exact time window of the fit depends on the Q value considered. Since we performed measurements on protein solutions at concentrations where protein-protein interactions might be significant, we extracted the single-protein diffusion coefficient in the infinite dilution limit $D_0(Q)$ from the effective diffusion $D_{eff}(Q)$ of Eq 2.64, by considering concentration-dependent interprotein interactions and solvent-mediated interactions [210]:

$$D_0(Q) = \frac{D_{eff}(Q) \cdot S(Q)}{H(Q)}, \tag{4.2}$$

where $H(Q)$ is the hydrodynamic function representing Q-dependent hydrodynamic interparticle interactions mediated by the solvent, while the structure factor $S(Q)$ describes the direct interactions.

To account for the hydrodynamic contribution, we followed the procedure described in Ref [151]. Consequently, we have scaled the term $D_{eff}(Q) \cdot S(Q)$ to match the Q-

independent value of the translational diffusion measured by DLS. Theoretical calculations on model solutions of spherical charged proteins have shown that $H(Q)$ has a similar trend as $S(Q)$ and, quite importantly, does not manifest significant oscillatory behavior in the Q-range above 0.06 Å^{-1} [211]. Thus, any modulation of the $D_0(Q)$ spectrum, reported in Figure 4.2, can be ascribed to the single tetramer dynamics.

At all temperatures, the spectra show a well defined first peak at Q=0.11 Å^{-1}, whose intensity grows with temperature. This peak, at the characteristic length $\lambda = \frac{2\pi}{Q} \sim 57$Å, relates to the correlated motion of regions of the tetramer at the external surface. A second peak is visible at shorter length scales, namely for Q approaching the value 0.2 Å^{-1}. The resolution of the second peak region is poor, and the rise in the protein diffusion is evident only at the highest temperature 313K in this Q-range, thus we concentrate our efforts in discussing the first peak only.

In Figure 4.2 we also report the theoretical curves for the rigid body roto-translational diffusion calculated according to Eq 2.65 in Chapter 2 [150]. The translational and rotational tensors needed for the calculation were obtained by hydrodynamic calculations performed using the HYDROPRO software [153]. The values for the translation diffusion coefficients obtained by DLS (horizontal baselines in Figure 4.2) and those obtained by HYDROPRO are close at all considered temperatures, they differ by a factor 1.2 (see Table B.1 in the chapter Appendix). This factor was used to rescale the elements of the translational and rotational diffusion tensors in Eq 2.65 for a correct comparison with the experimental spectra. The plotted curves immediately reveal that the rigid body rotational diffusion (dashed curves) accounts for the main features of the spectra, apart from the peak zone around $Q \simeq 0.11$Å^{-1}, which contains an additional contribution at 298 K and 313 K. The latter stems from the internal motion of the protein that, along with the rotation and the translation, contributes to the the overall protein diffusivity: $D_0(Q) = D_0^{Tra} + D_0^{Rot}(Q) + D_0^{Int}(Q)$. The internal contribution is negligible at the lowest temperature T=283 K but, at higher temperatures, it represents 10-15 % of the total diffusion of the protein when translation is removed.

The contribution from the internal Q-dependent motion can be dissected by using NM analysis [212] (see Chapter 2) and focusing on the low frequency modes. These have been extracted for the crystallographic configuration, Figures B.4 and B.3 in the Appendix of the chapter. Because of the highly symmetric nature of the LDH protein, at variance with previously investigated systems where large domains displacements occur at low frequency [150, 193], it is difficult to single out a dominant mode. Therefore, in order to gather microscopic insights on the protein internal dynamics, we have performed MD

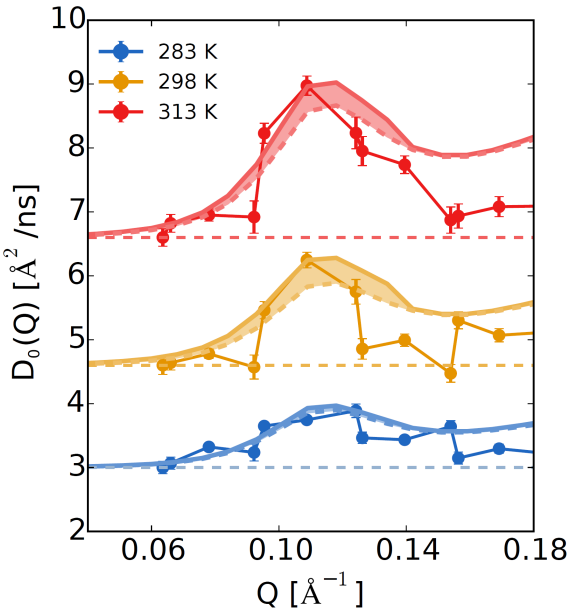

Figure 4.3: Experimental diffusion spectra at different temperatures compared to the theoretically reconstructed spectrum (solid line) obtained by adding the rigid-body contribution (dashed line) to the internal-dynamics contribution derived from long MD simulations (shown in the shaded area).

simulations and estimated the contribution of internal motion to the $D_0(Q)$ spectra.

4.4 Protein Internal Motion

Following the work flow proposed by Smolin *et al* [154], the diffusion spectrum $D_0(Q)$ extracted from the MD simulations can be decomposed in its translational, rotational, and internal contributions by adequate post-processing of the trajectories. For instance, by removing the translation of the protein's center of mass from the original MD trajectory and fitting on it a rigid reference structure, a virtual trajectory is generated where only the rigid body rotation is present. In the same spirit, if the MD trajectory is fitted, frame by frame, on a reference structure, thus removing roto-translation, only the internal modes will be maintained. The processed trajectories are used in the calculations of the intermediate scattering function $I(Q,t)$ in order to extract respectively the rotational diffusion $D_0^{Rot}(Q)$ and the internal diffusion $D_0^{Int}(Q)$ by fitting the initial decay of the obtained $I(Q,t)/I(Q,0)$. However, when employing this strategy for a direct comparison with experimental data, *ad hoc* numerical manipulations are needed. In fact, molecular force-fields routinely used in MD simulations of protein-water solutions generally bias

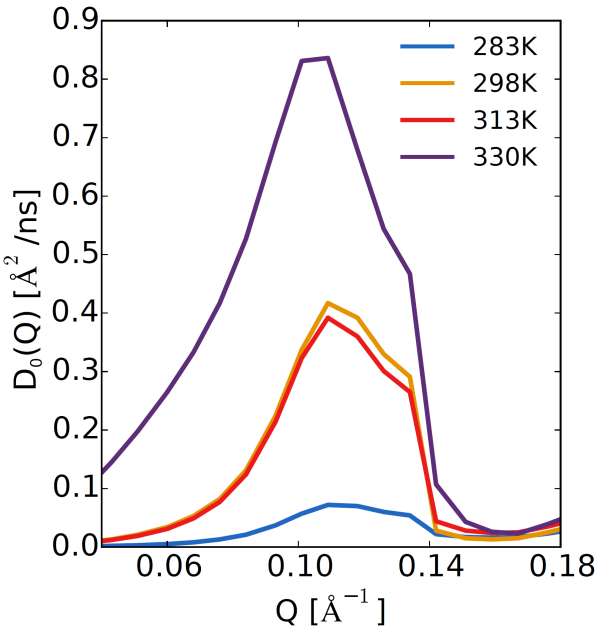

Figure 4.4: The internal contribution to the diffusivity as obtained from MD trajectories, by fitting the calculated I(Q, t) for different temperatures, without rescaling for the translational diffusion.

protein motion and empirical rescaling of roto-translation is necessary [154]. Moreover, care should also be placed on practical issues such as the system-size dependence as well as the accuracy of the fitting procedure to extract the diffusion coefficients [213].

In order to limit the impact of these weaknesses, we used MD-based calculations for the internal motion only. From our MD simulations, we removed roto-translation, and obtained the internal component of the spectra. The internal diffusion has been added to the curve calculated by considering the rigid-body motion as described in the previous section, and compared to the experimental data, see Figure 4.3. Most notably, we observe the activation of the internal motion at T=298 K. Between T=298 K and T=313 K, the internal contribution at the peak is comparable (0.4 Å^2/ns), becoming much larger at the highest simulated temperature, T=330K (~0.8 Å^2/ns), see Figure 4.4. Importantly, the addition of the internal contribution allows to quantitatively reproduce the experimental value of the diffusion coefficient $D_0(Q)$.

To inspect these thermally activated internal protein modes in a length-range corresponding to the first peak of the calculated spectra, and to further understand the temperature effect on these modes, we performed Principal Component Analysis (PCA) [209] on the processed trajectories with removed roto-translations, at T=298 K and T=313 K (see section Methods). By projecting these trajectories on a different number of modes,

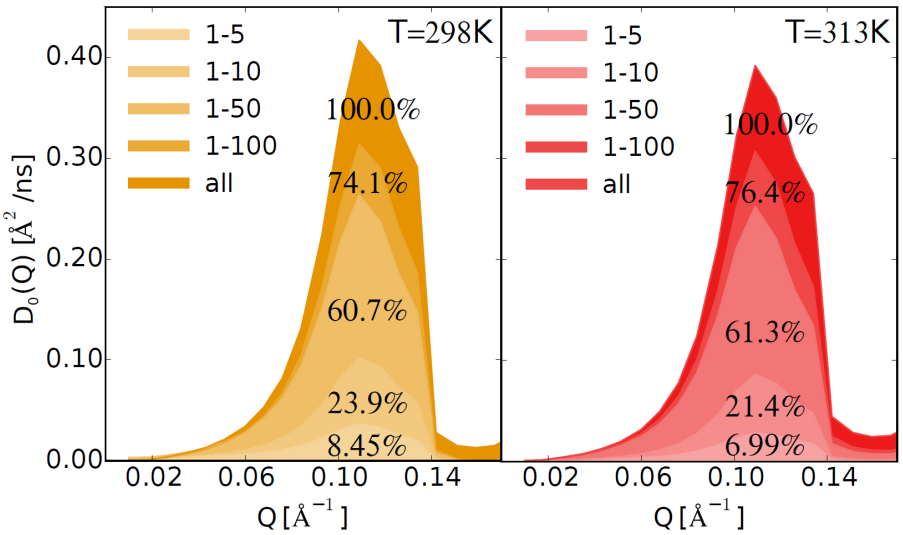

Figure 4.5: Contribution of the principal modes to the diffusion spectra at temperature 298 K and 313 K.

components summed	T=283K [Å^2/ns]	T=298K [Å^2/ns]	T=313K [Å^2/ns]
1-5	0.0035 (4.99%)	0.033 (8.45%)	0.025 (6.99%)
1-10	0.014 (20.2%)	0.094 (23.9%)	0.077 (21.4%)
1-50	0.041 (59.3%)	0.24 (60.7%)	0.22 (61.3%)
1-100	0.050 (74.4%)	0.29 (74.1%)	0.28 (76.4%)

Table 4.1: The dynamics of protein described by a varying number of Principal Components taken into account, and represented by the value of the diffusion coefficient associated to $Q = 0.118$ Å^{-1}, where the experimental data peak is demonstrated. The parentheses contain the percentage of the total internal dynamics at $Q = 0.118$ Å^{-1} described by these components.

we are able to extract the contribution of specific internal motions to $D_0(Q)$. This is done by calculating, and subsequently fitting $I(Q,t)/I(Q,0)$ from the projected trajectories. For the thermally activated state (T>283 K), it is found that the first 100 modes account for ~75% of the diffusive contribution in the first peak, see Figure 4.5 and Table 4.1. The regions of the protein interested by these modes are highlighted in Figure B.5 in the chapter Appendix. The Figure clearly shows that the larger flexibility induced by the modes localizes at the level of the loops on the protein surface, most notably the catalytic site loop, and this distribution of flexibility is the same for the four subunits.

We now provide a complementary view of the protein soft modes underlying the internal motion by performing conformational and kinetic clustering of the MD trajectories, see section Methods. The results are reported in the Figure 4.6. In the upper layer of

the Figure, the network of states visited during the dynamics is represented for the four simulated temperatures. At the lowest temperature, only few conformational states are accessed by the protein at the 0.6 μs time scale, this number increasing exponentially with temperature (see Table B.2 in Appendix). While conformational clustering classifies conformational states only on the basis of their proximity according to the RMSD, a more subtle casting is achieved by merging together the frequently interconverted states, yielding a representation of states that are mutually distinguished by high kinetic barriers, as represented in the middle part of the Figure 4.6 [45, 78]. The thermally activated conformational disorder is in fact due to different orientation of the binding site loops and adjacent peripheral helices, as the reader can appreciate from the last layers of the Figure 4.6, where the protein regions manifesting the highest flexibility are magnified in proportion and accentuated in color. The pictorial representation of the protein is complemented by the sequence profile of the mean square fluctuation for one domain. Among the flexible regions activated in temperature, we individuated two of the principal protein regions involved in the allosteric reorganization of bacterial LDHs [199, 200, 201, 202]; the active site loop (CL) and region MR2 around the E222 residue.

The results of the PCA and the clustering clearly show that the contribution to internal diffusivity stems from an ensemble of modes involving the peripheral regions nearby the catalytic site. This collective reorganization of the protein matrix is the source of the plasticity necessary for functionality, i.e. the conformational shifts due to cofactor and substrate binding, as well as the gating of the binding site loop required for efficient catalysis. Since the binding site loops play a fundamental role in LDH activity [203, 205, 206, 207], we have specifically targeted them in further investigation by considering their correlated motion. The distances between the loops span the range 36-75 Å, their correlated motion therefore falling in the probed region of the peak, 0.08 $Å^{-1}$ <Q<0.14 $Å^{-1}$. By using the harmonic approximation, details being provided in the Methods section, it is found that specific correlated motions account for about 2% of the internal diffusivity in the peak region, see Table B.3 in Appendix. The same is found by processing a trajectory containing the internal dynamics only, and excising the loop region, see Figure 4.7. Interestingly, the thermal response of the correlated loops motions seems to depend on the considered subunit, see Figure B.6 in Appendix.

The obtained results reinforce the notion that a wide set of correlated motion across the four domains contribute to the internal component of the diffusivity peak. It is important to note that the relative distances between flexible amino-acids in the MR2 across all four domains fall in the range of ~36 Å, probably contributing to high Q region

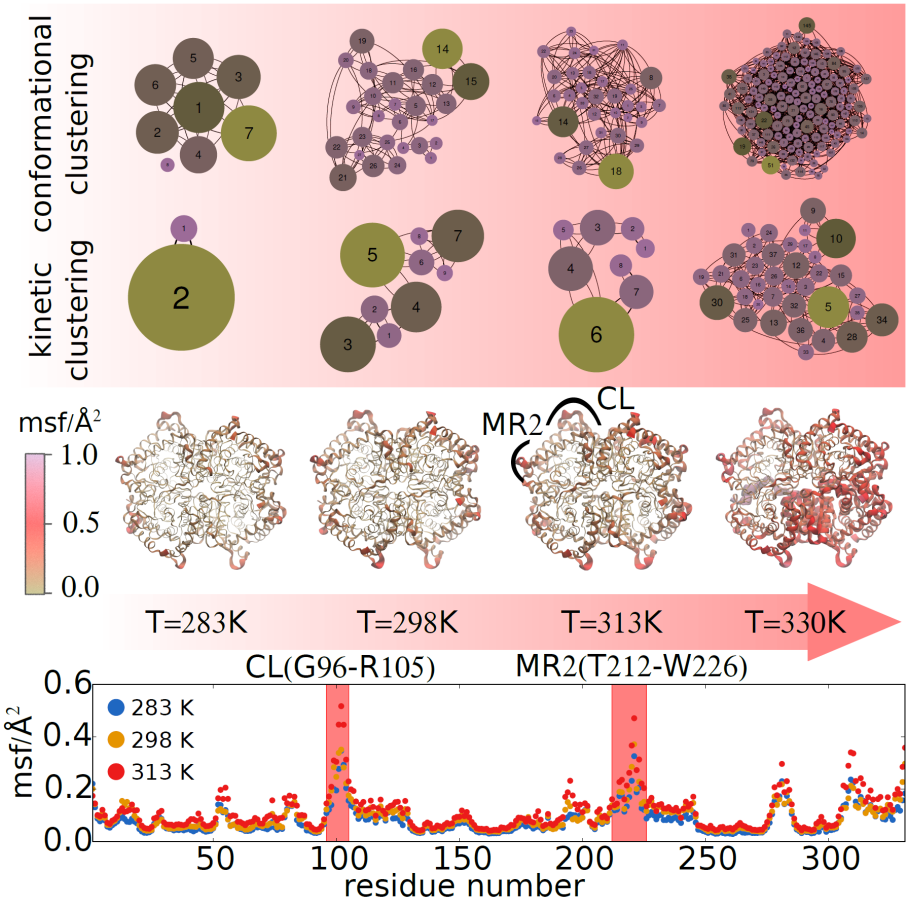

Figure 4.6: Network of conformational states visited by the LDH protein in MD simulations at different temperatures. In the top layers we report the networks obtained by conformational and kinetic clustering, respectively. In the bottom layers, the flexible regions of the protein individuated by the local atomistic fluctuations are highlighted in the protein structure and along the domain sequence. We also emphasize the position of the loop of the catalytic site (CL) and the adjacent helical region (MR2 according to the annotation of Ref [200]) on one of the proteins.

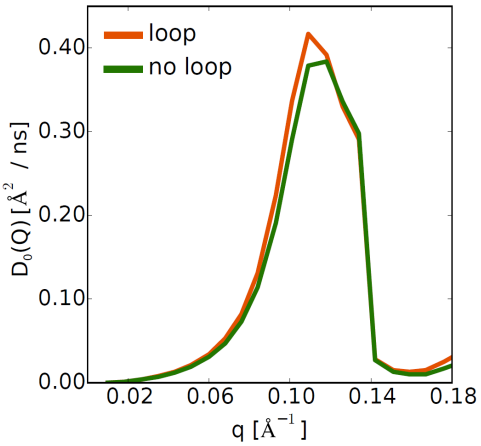

Figure 4.7: The internal dynamics diffusion coefficent extracted from fitting the I(Q, t) from a single trajectory where the loop was removed ('no loop') and kept ('loop'). After calculating and fitting the I(Q, t), the results indicate the loop accounting for 9% of the 0.417 Å^{2}ns^{-1} peak. The loop dynamics is thus characterized with a diffusion coefficient of 0.038 Å^{2}ns^{-1}, agreeing well with calculations using the harmonic approximation. T=298 K.

of the spectra, while the distance across the four domains between MR2 and the active site loop atoms are ~60 Å for adjacent and intra-dimer domains, and 80 Å for diagonal distances, respectively. Thus, the correlated fluctuations of the catalytic loop with the MR2 region are a likely contributor to the observed first peak in the diffusion spectrum.

4.5 Discussion and Conclusions

NS Spin-Echo spectroscopy enabled us to probe the thermal activation of the soft modes in a mesophilic tetrameric LDH enzyme relevant for its activity. At the lowest investigated temperature, T=283 K, the protein motion is substantially dominated by rigid body roto-translation. Only by considering the ambient temperature condition do the internal motions give a significant contribution to the Q-dependent diffusivity. At Q=0.166 Å^{-1} , this contribution, D^{Int}, is about 0.4 Å^2/ns, and is equal at 298 K and 313 K. This finding suggests that in the optimal temperature window for the protein activity, 298 K<T<313 K [66], protein requires and sustains a steady level of internal flexibility. Additional thermal excitation would have a degrading effect on functionality by provoking the distortion of the catalytic site and dissipating conformational changes in non-allosteric paths. LDH maximal activity occurs at ~313 K and is compromised at higher temperatures [66]. The loss of activity anticipates the thermodynamic unfolding detected at 330-340 K [214, 215].

On the basis of MD simulation in this high temperature regime, our findings identify that high flexibility interests a greater portion of the protein and induces a substantial distortion of the structure around the catalytic pocket.

The analyses of the MD trajectories based on PCA and clustering strategies highlight that the functional internal diffusivity cannot be reduced to a single dominant motion, but is rather attributed to a wide range of modes covering the peripheral region of the protein around the catalytic sites, the binding site loops Ala95-Arg105, and the helices Arg105-Ser127 (MR2 region). It is speculated that the activation of some of these modes are fundamental to the functionality of the protein similarly to the allosteric regulation in bacterial LDHs. In fact, by comparing the X-ray structures of apo and holo bacterial LDHs [200, 201], it was possible to single out the regions involved in the allosteric conformational shifts.

In our experiments and simulations, the protein is in the apo state, where we always find the binding site loop sampling the open configuration. The closed loop configuration, that is rate-limiting the catalysis, is most likely only accessible upon co-factor and substrate binding. However it is important to note that in the MD simulations we observe the reorganization of the catalytic site toward a reactive configuration as temperature increases. At variance with bacterial LDHs, in the rabbit muscle 5 LDH, the extension of a supplemental N-ter helix from one domain to another locks the movement of the MR1 region, the latter found to be rigid in our structure. In bacterial LDHs, the displacement of MR1 (Leu164-Gly186) from the apo to the holo state allows the reorganization of the active site to its reactive state, notably inducing the correct positioning of the catalytic residue, equivalent to Arg168 in the rabbit muscle 5 LDH, towards the oxamate. In our system, the positioning of Arg168 toward the reactive configuration is enhanced by the temperature increase. This is shown by monitoring the conformational fluctuations of the residue side chain oscillating between an extended configuration, mimicking the configuration it assumes when pyruvate is bound in the catalytic pocket, to state bound with the Asp165. Between 298 and 313 K, the temperature eases the breaking of the ion pair Asp165-Arg168 and increasingly favors the extension of Arg168 as well the interaction of Asp165 with the catalytic His192, see Figure 4.8. Further increase in temperature enables the Arg168 to start visiting conformations pointing outside the active site, which resembles the inactive state of bacterial LDHs. For a complete overview of the process, see the discussion in Appendix B and Figure B.7. The analysis of the conformational shifts induced by the substrate binding [151, 215, 78] will be the next step of the investigation.

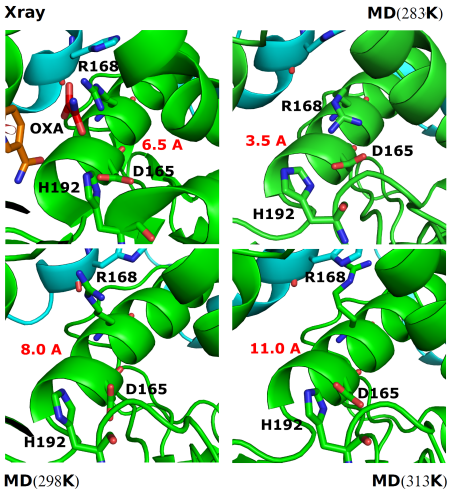

Figure 4.8: Representation of different conformations sampled by the residue Arg168 in the active site during the MD simulations at different temperatures. In the top right panel we also report the organization of the catalytic site in the presence of oxamate and pyruvate molecules as resolved in the X-ray structure of the LDH of rabbit muscle 5 (PDB code 3F3H). For the sake of comparison with figure B.7, we have also reported the instantaneous distance between the Arg168 and Asp165 charged terminals. The Figure is reported by the courtesy of D. Madern (IBS, Grenoble).

In conclusion, we have shown the effectiveness of a combined use of Spin-Echo NS spectroscopy and MD simulations in studying the thermal response of the protein motion relevant to functionality. The experimental and numerical tools presented here allow the characterization of the multiple time- and length scales of protein dynamics with a specific focus on functionally relevant modes. The exploration of thermal response of these modes is essential in comparing proteins with different optimal working temperatures [71] and addressing different evolutionary mechanisms of functional regulation. The NSE experiments of the thermophilic MDH from *Methanocaldococcus jannaschii* have been completed in a large temperature range up to the optimal working temperature of the protein; the maximum scanned temperature was T=343 K. Long MD simulations at the μs time scale have been carried out at the experimental temperatures and the comparative analysis is under way.

APPENDIX B

APPENDIX OF CHAPTER 4

temperature	DLS [Å^2/ns]	HYDROPRO [Å^2/ns]
283K	3.0	3.6
298K	4.6	5.1
313K	6.6	8.2

Table B.1: Comparison between translational diffusion coefficients as measured by the Dynamic Light Scattering and calculated in the HYDROPRO program (see main text for details).

temperature	N(t_{sim})	N$_\infty$	τ (ns)	N$_{kinetic}$
283K	8	23.0	1427.7	2
298K	27	34.4	470.0	9
313K	33	68.0	939.0	8
330K	150	-	-	37

Table B.2: Conformational and kinetic clustering of the MD simulations. The conformational clustering was based on the collective variable RMSD and using a cut off of 1.5 Å. The total number of clusters obtained is indicated in the first column, $N(t_{sim})$. In the second and third column, we report the parameters of a simple exponential growth model fitting the data, $N(t) = N_\infty \cdot (1 - exp(t/\tau))$. In the last column, we report the number of independent kinetic states as obtained by applying Markov state model based clustering algorithm with a threshold of 2.0.

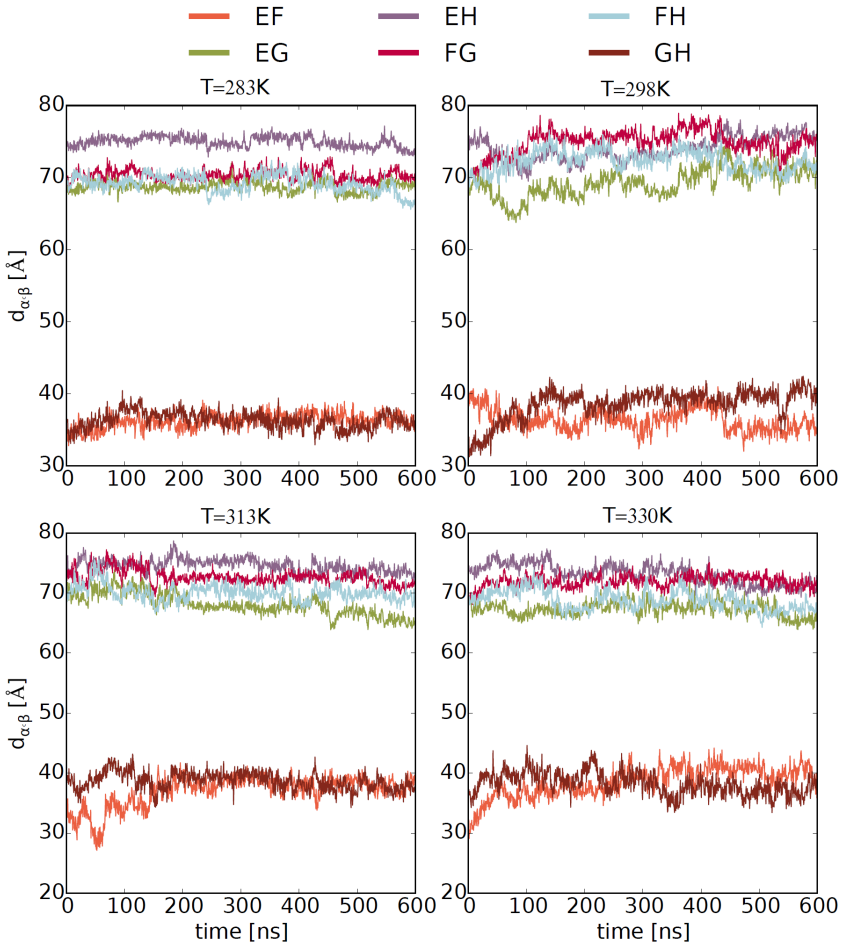

Figure B.1: Timelines of interloop distances between four subunits named E, F, G, H, defined as $d_{\alpha,\beta} = (\frac{1}{N_\alpha}\frac{1}{N_\beta}\sum_i^{N_\alpha}\sum_j^{N_\beta}|r_i - r_j|^2)^{1/2}$.

subunits involved	T=283K [Å$^2/ns$]	T=298K [Å$^2/ns$]	T=313K [Å$^2/ns$]	T=330K [Å$^2/ns$]
intradimer (FH, EG)	0.053	0.13	0.15	0.20
adjacent (EF, GH)	0.046	0.061	0.23	0.30
diagonal (FG EH)	0.019	0.056	0.095	0.13

Table B.3: Diffusion coefficient of the interloop distances obtained as $D_{\alpha,\beta} = \frac{\langle \delta d^2_{\alpha,\beta}\rangle}{\tau}$, where d is the interloop distance defined for all atoms in the loop as $d_{\alpha,\beta} = (\frac{1}{N_\alpha}\frac{1}{N_\beta}\sum_i^{N_\alpha}\sum_j^{N_\beta}|r_i - r_j|^2)^{1/2}$, and τ is the relaxation time of the autocorrelation function of this distance.

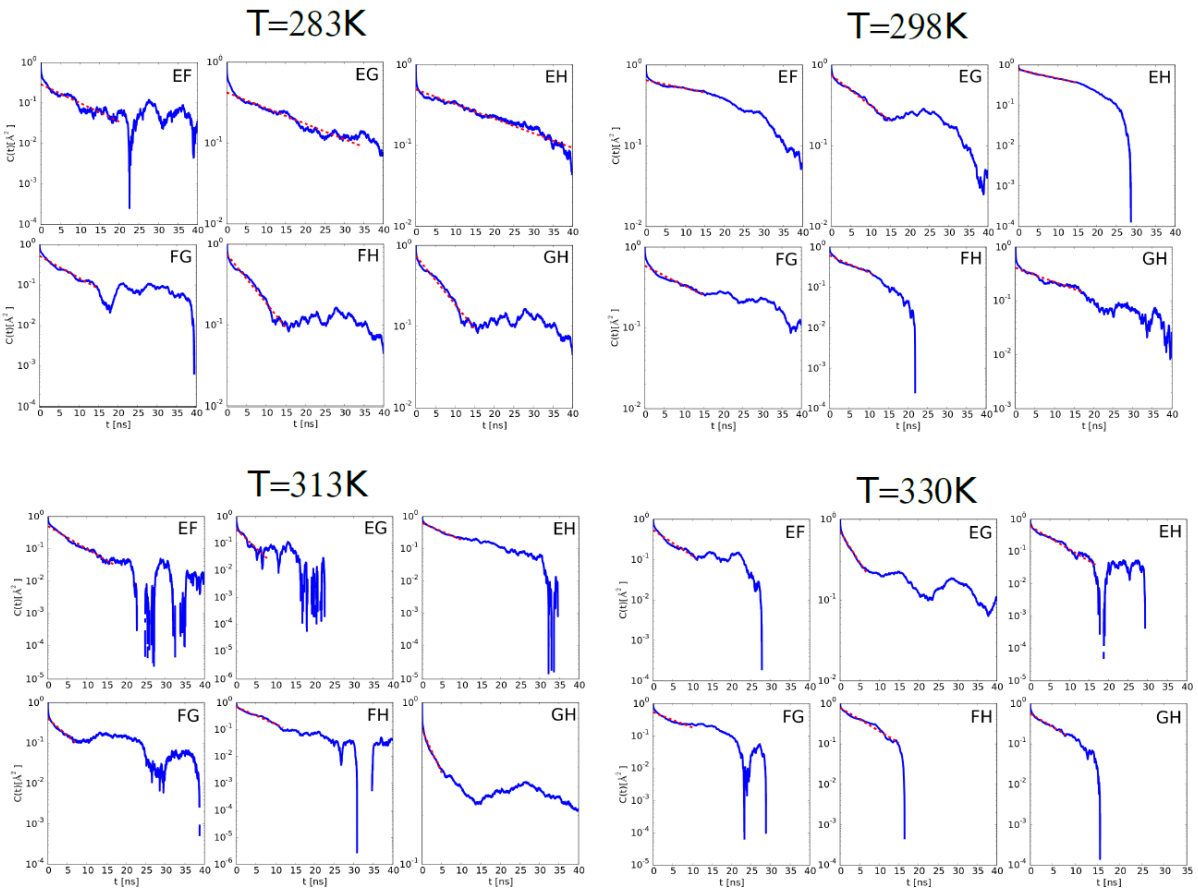

Figure B.2: Fitting the exponential model $e^{-t/\tau}$ to the autocorrelation function, $C(t) =< d_{\alpha,\beta}(t)d_{\alpha,\beta}(0) >$, of the interloop distances as defined by the metric $d_{\alpha,\beta} = (\frac{1}{N_\alpha N_\beta}\sum_i^{N_\alpha}\sum_j^{N_\beta}|r_i - r_j|^2)^{1/2}$ to obtain relaxation time τ, necessary for calculating the diffusion coefficient in harmonic approximation (see main text). The results are shown for all pairs of loops α, β in the four subunit protein, and for all simulated temperatures.

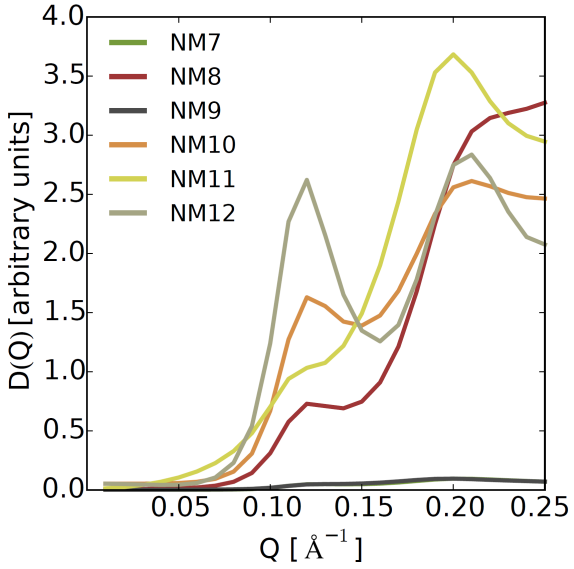

Figure B.3: Diffusion coefficient obtained from projecting the normal modes of the X-ray structure PDB 3H3F on vectors scattered over a sphere so as to mimic a point scatterer, see Chapter 2 and Section on NMA in it for details on the calculation. For each mode α, the diffusion coefficient is calculated as follows; $D^{\alpha}(Q) = \frac{C}{Q^2 F(Q)} \langle \sum_{k,l} b_k b_l \exp[i\mathbf{Q}(\mathbf{r}_k - \mathbf{r}_l)](\mathbf{Q}\cdot\mathbf{e}_k^{\alpha})(\mathbf{Q}\cdot\mathbf{e}_l^{\alpha})\rangle$, where $F(Q) = \sum_{k,l} b_k b_l \exp[i\mathbf{Q}(\mathbf{r}_k - \mathbf{r}_l)]$, and $C = \lambda_{\alpha}\frac{k_B T}{m\omega_{\alpha}^2}$, where ω_{α}^2 is the eigenvalue associated to each mode, and the λ_{α} is the mode-dependent relaxation rate, containing friction coefficients within the molecule and with the surrounding water. As the latter are unknown, we cannot estimate the prefactor C and thus show D in arbitrary units.

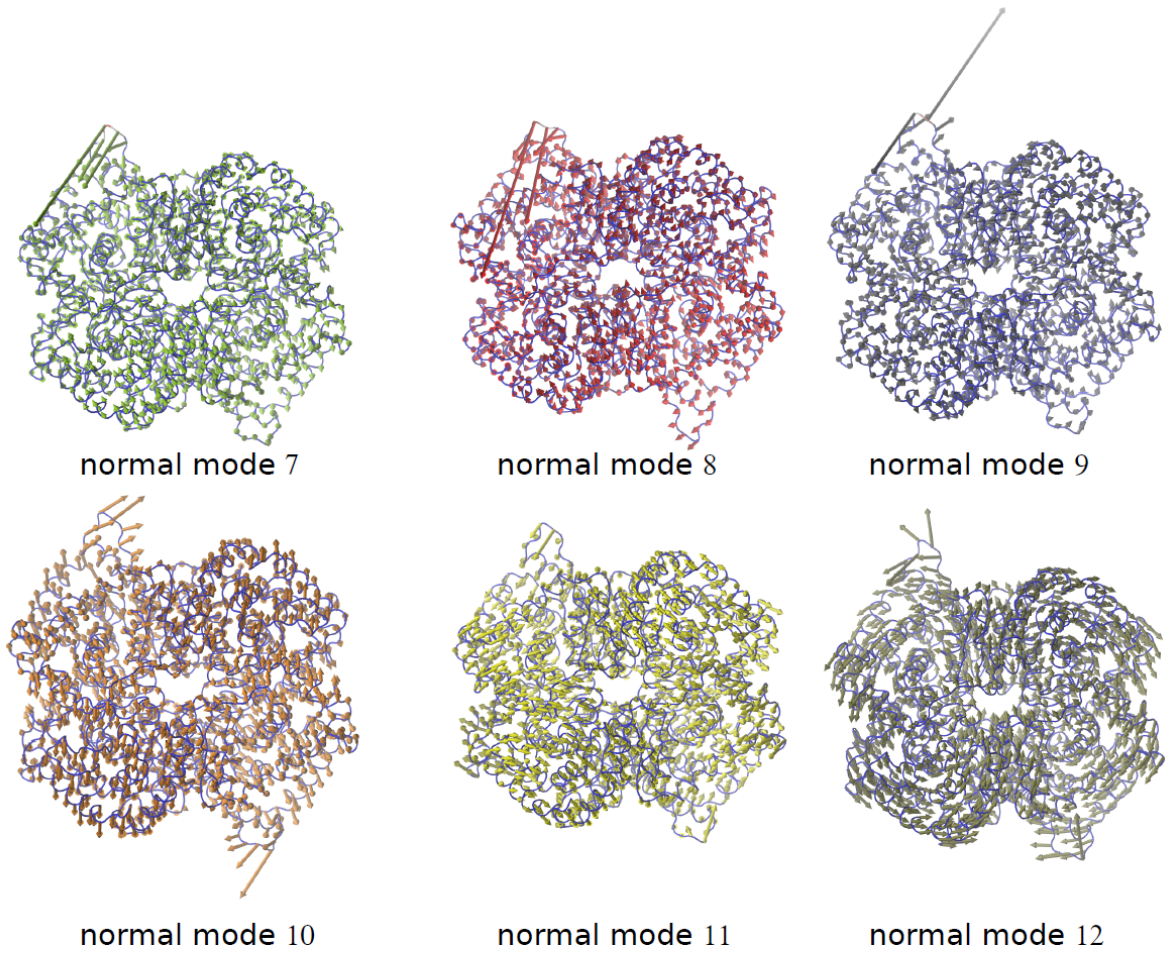

Figure B.4: Representation of the first 6 nontrivial Normal Modes of the X-ray structure PBD 3H3F. The arrows' magnitude and direction reflect the intensity and the direction of movement in the harmonic perturbation.

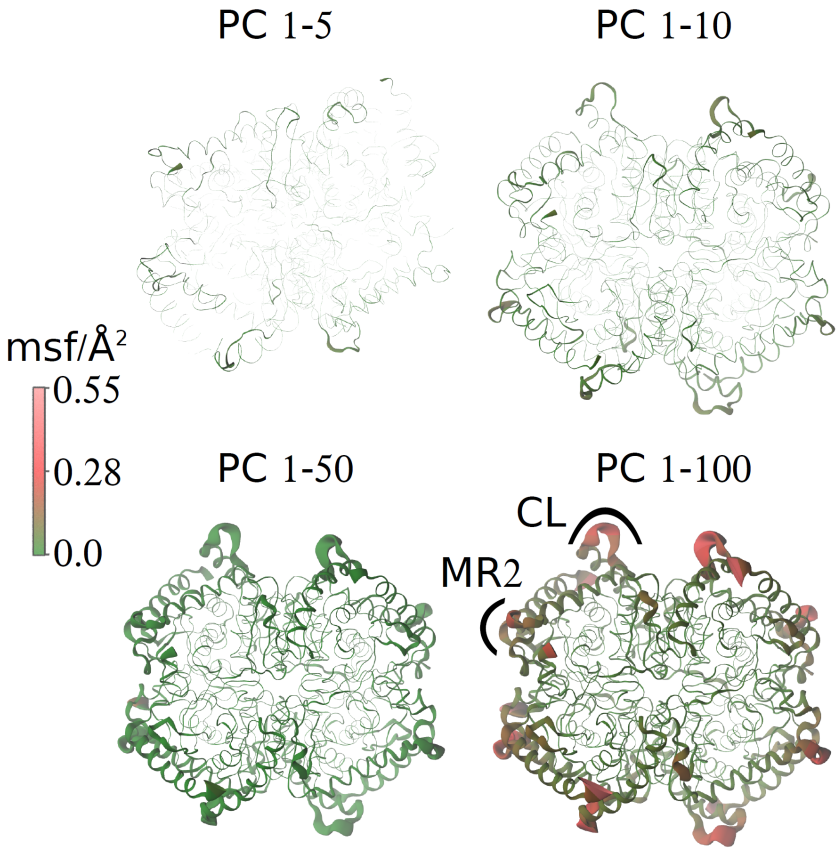

Figure B.5: Displacements of the protein backbone described by summing different numbers of Principal Components, coded in color and thickness. T=298 K.

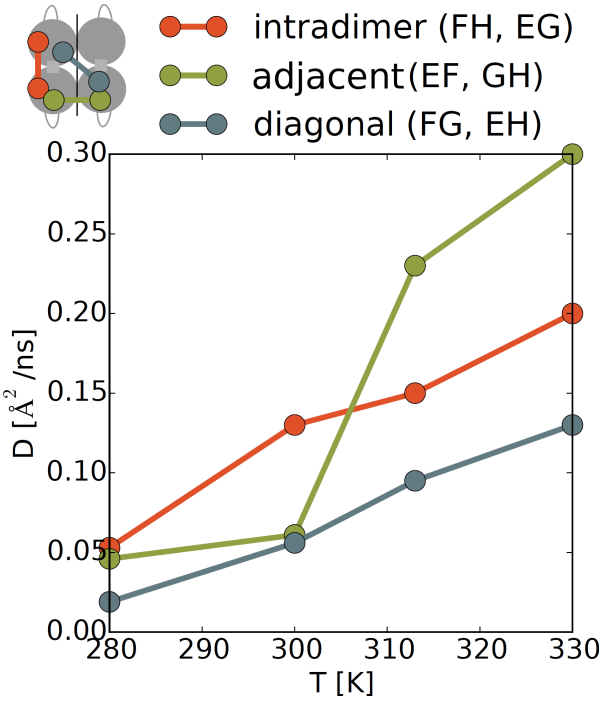

Figure B.6: Table B.3 plotted.

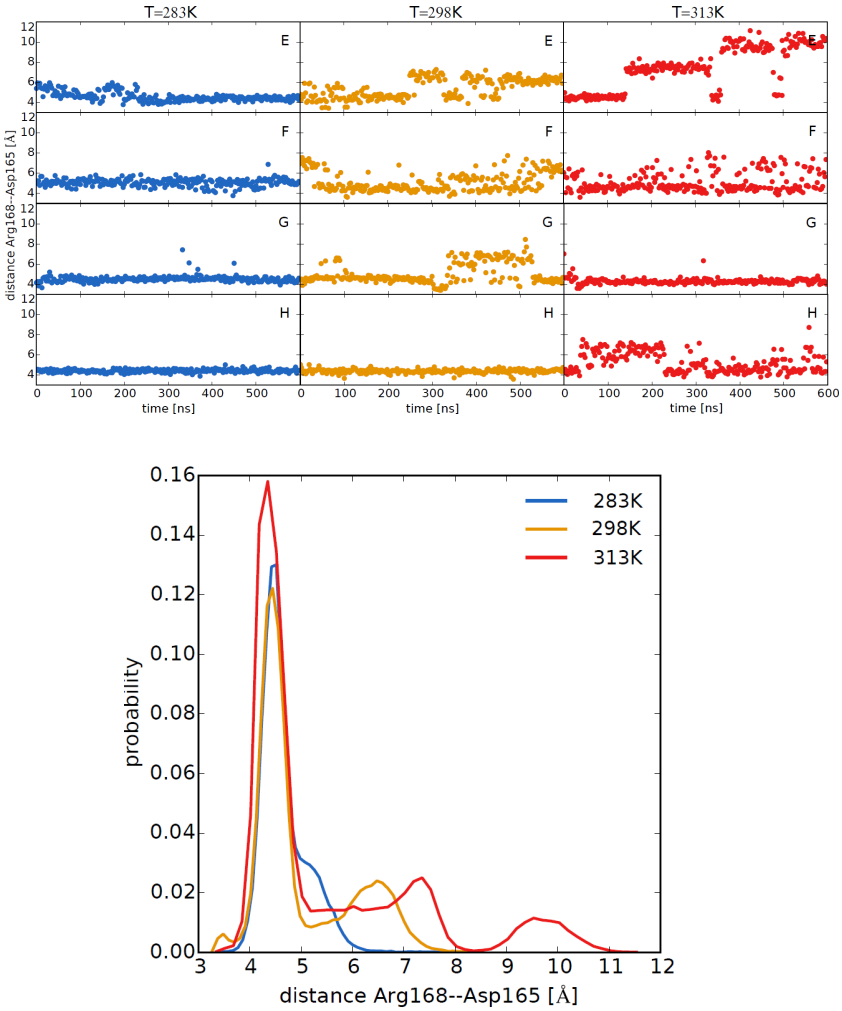

Figure B.7: In the top panel of the figure, we report the distance variations in Å between R168 and D165 amino-acids for each domain of the apo-state protein (labeled E, F, G, and H), and at various temperatures during the 0.6 μs MD simulations. The distance computed is the center of mass of the charged terminals of the sidechain ends. The bottom panel reports the probability distribution of the distances averaged over the four domains.

B.1 Arg168-Asp165 Sidechain Center of Mass Distances

In Figure B.7 we report the distance variations in Å between R168 and D165 amino-acids for each domain of the apo-state protein and at various temperatures during the 0.6 μs MD simulations (labeled E, F, G, and H). The distance of approximately 4 Å corresponds to the R168 lateral chain protruding within the catalytic site and forming a salt bridge with D165, as illustrated by the representative view extracted from the simulation at 283 K in Figure 4.8. A distance higher than 4 Å, typically between 6 and 8 Å, indicates that the R168-D165 salt bridge is broken, as shown in representative conformation extracted from the MD simulation at 298 K, see Figure 4.8. The distance of 10 Å indicates the R168 lateral chain is located outside of the catalytic site as it is illustrated on the snapshot of active site of the domain E taken at the end of the simulation at 313 K, see Figure 4.8. For sake of comparison, a close-up view of the ternary complex crystal structure (PDB code 3F3H) is shown with the same orientation. The substrate analogue (oxamate) is colored in red, while the coenzyme is shown in orange. Domains E and F are in green and blue, respectively. For the sake of clarity, other domains are not shown. Important residues are represented in sticks. It is to be noted that the residue D165 helps in polarizing the catalytic residue H192 during the catalysis. R168 lateral chain is orientated within the catalytic site and interacts with the negatively charged analogue. The R168-D165 distance in the X-ray structure of the holo state is ~6.5 Å. The MD simulation data show that in the apo state, as is expected for an eukaryotic LDH, R168 from rabbit LDH is mainly located within the catalytic site at 283 K. Because there is no substrate analogue, R168 forms a strong salt bridge with D165, and fluctuates around this tight position. At 298 K, R168 is always in the catalytic site, but the salt bridge is more labile as it is indicated by the distance increasing up to 8 Å, especially in domains E, F, and G. This position of R168 in the apo state, with an extended lateral chain, mimics the one observed within the ternary complex. At 313 K, fluctuations become of wider amplitude. It is striking to observe that in domain E, during the last part of the simulation, R168 lateral chain goes outside the catalytic site. This drastic local change was not detected in others domains. In domain E, some other structural reorganizations involving the kinked helix (H1G-H2G) and helix H2F which carries R168 are observed. The mobile loop covering the catalytic site is not show in the bottom panels of Figure 4.8. During the simulations, this loop always samples the open conformation for the higher temperatures T=298 K and T=313 K. We put forth that in

the snapshot at 313 K, H192 and D168 of the apo state display the same conformation as in the crystal structure.

Eukaryotic LDH are considered non-allosteric i.e, both the apo and the ternary complex look like the R-(active) state of bacterial LDH. Our data indicate that in certain conditions, fluctuations of the apo state of an eukaryotic LDH may sample some local conformations which look like those observed with the T-(inactive) state of allosteric bacterial LDH.

CHAPTER 5

TRACKING THE LINDEMANN CRITERION FOR PROTEIN MELTING

The universal scaling of atomic fluctuations has often been used to describe phase transitions in solids. A simple relationship, the Lindemann criterion, predicts the onset of the melting once the thermal fluctuations exceed a threshold value upon which the crystal "shakes itself to pieces". Here we aim to verify whether the concept can be extended to biological inhomogeneous matter by using as a model the protein Lysozyme, and performing Molecular Dynamics and enhanced sampling simulations to meet this goal. To this purpose, we considered the Lysozyme protein embedded in three different environments: in dilute aqueous solution and two powder systems, one solvated with water and another with glycerol. An effect of these conditions is the shift of the Lysozyme melting temperature. The systems were simulated and analyzed with the final goal of elucidating whether protein melting is accompanied by universal scaling of atomic fluctuations. The work has been inspired by recent Elastic Incoherent Neutron Scattering experiments performed by our collaborators.

5.1 Introduction

Protein thermal stability is a key issue in both research and industrial application [216, 217, 54, 218, 219, 220]. Design of thermally stable proteins requires the understanding of means by which proteins lose their stability and of the extent to which this process can be

controlled. To this purpose, one can modify the protein structure and sequence [216, 217, 54, 218, 219] or tune the physico-chemical properties of the environment surrounding the biomolecule [221, 222, 223, 224]. Even though still an object of debate [225], it has been proposed in the past that protein flexibility is inversely correlated with thermal stability [226, 227, 228, 229], identified by the value of its melting temperature T_m, the midpoint of the thermal transition. The intrinsic flexibility of a protein and how it responds to external perturbations can be probed by the magnitude of atomic thermal fluctuations [230] spanning the picosecond time scale, which deserve to be carefully described along the path toward unfolding.

In this final part of the thesis, we investigate the effect of protein crowders and bioprotectants to the change in atomic fluctuations magnitude and their scaling. Crowding usually refers to the effect of the volume excluded by one molecule on the thermodynamics and kinetics of folding, binding, and chemical reactivity of another molecule [231]. The crowding effect is mainly attributed to the reduction of the entropy of the unfolded state under the influence of the crowders due to the excluded volume effect. A seminal work on the Lysozyme shows different crowders having qualitatively similar effects on protein refolding [232], therefore showing that the excluded volume argument is sufficient in explaining the protein behavior under crowding. However, there certainly exists a dependence on the nature of proteins, crowding agents, and their interactions [233], a consequence of electrostatic interactions and the hydrophobic effect [234, 235, 236]. In fact, studies on the Lysozyme embedded in a single or multiple crowder environment point to the presence of non-additive effects, emphasizing further additional contributions to macromolecular crowding [237]. Here we use the Lysozyme protein as a model to study the effect of homogeneous crowding with the additional effect of solvent, i.e. water versus glycerol. By using these well-known bioprotectant media, which make proteins more stable against thermal degradation [222], we progressively increase the Lysozyme melting temperature and study the behavior of its flexibility at different critical thermal conditions.

The exploration of the universal scaling of the atomic fluctuation while approaching melting relies on the Gillvary's modernization [238] of the Lindemann theory [239], where melting is onset once the MSD of atomic thermal vibrations reach a critical fraction of the interatomic separations, while in the original Lindemann model the assumption was made that the neighboring atoms must collide to onset melting. The main criticism of the Lindemann theory is that it only considers the atomic vibrational amplitude, without taking into account the configurational entropy of the system, particularly important

in the molten phase [240]. Translating the argument to protein atomic fluctuation, the question remains whether to consider the folded phase alone or to explicitly take into account the atomic fluctuations of the unfolded state in addition.

The investigation presented here aims to clarify whether atomic fluctuations can be used to predict system large-scale behavior when approaching protein melting. The results indicate that the thermal fluctuations of the Lysozyme in solution, powder, and crowding environment are similar, reminiscent of the Lindemann criterion. The protein melting temperature is upshifted in the crowded powder environment and further shifted when water is replaced by glycerol. This shows the importance of the excluded volume effect and solvent in the protein melting, while further suggesting that melting can be described by universal concepts independent of the thermoprotective mechanism in place.

5.2 Methods

5.2.1 Elastic Incoherent Neutron Scattering Experiments

The Elastic Incoherent Neutron Scattering experiments, described in Chapter 2, were performed by A. Paciaroni (University of Perugia, Italy) on the following samples: Lysozyme in dehydrated powder (mass ratio of protein to D_2O = 0.4) in the temperature range [20-350] K, and Lysozyme embedded in deuterated glycerol (mass ratio of protein to glycerol = 1) in the temperature range [20-410] K. The melting temperatures T_m of Lysozyme in these crowded conditions are 347 K and 398 K, respectively, as measured by Differential Scanning Calorimetry [241]. Both the dialyzed salt-free chicken egg white Lysozyme and solvents have been purchased by SIGMA. Lysozyme was previously dissolved in D_2O to allow the substitution of all the exchangeable hydrogen atoms, which are essentially located at the protein surface. Additionally, deuterated solvents were used to minimize the contribution of the solvent to the overall signal [242]. The measurements were done at the IN13 backscattering spectrometer (Institut Laue-Langevin, Grenoble), with an energy-resolution of Γ=4.5 eV (half-width at half-maximum) in the wide Q range [0.3-4.7] $Å^{-1}$. An amount of about 0.5 g of sample was held in a standard flat aluminium cell with internal spacing of 0.5 mm, placed at an angle of 120^o with respect to the incident beam. The data were corrected to take into account for incident flux, cell scattering, self-shielding, and detector response. Finally, the intensity of each sample has been normalized with respect to the corresponding lowest measured temperature.

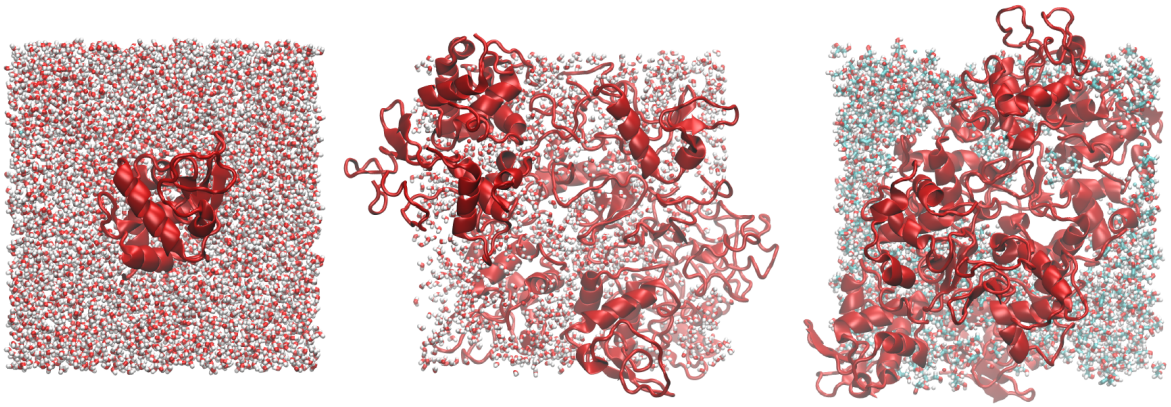

Figure 5.1: Lysozyme in solution, and in powder with water and glycerol, shown from left to right. Note that while the powder systems contain 8 equivalent proteins, only one protein is targeted in simulations and analysis, while the others represent protein crowders.

5.2.2 System Preparation

Three setups of Lysozyme were simulated, Lysozyme in solution, hydrated powder, and in powder with glycerol. The latter two were created to match exactly the experimental samples in the Elastic Incoherent Neutron Scattering Experiments described previously. Lysozyme in solution was created by placing the protein structure resolved by X-ray crystallography (PDB code 2LYM [243]) in a box of TIP3P water [102], and adding counterions, creating a system of 31341 particles. This system was built to contrast the crowded conditions of the protein powders, where 8 proteins are arranged within a cubic box by using symmetry transformations of the 2LYM crystal space group, and adding TIP3P water and counterions to create a dehydrated powder of 22467 atoms in total. This system matches the experimental system where the mass ratio of protein and D_2O was 0.4. Lastly, Lysozyme in the powder containing glycerol as solvent was created in an equivalent manner as the protein powder system, with the difference of adding glycerol as solvent along with the counterions, obtaining 30654 particles in total, corresponding to the experimental system where the mass ratio of protein and glycerol was 1. The three systems will be termed 'solution', 'powder/water', and 'powder/glycerol' in the Results section, and are shown in Figure 5.1.

5.2.3 Replica Exchange with Solute Scaling (REST2)

As our main aim was to determine fluctuations when approaching melting, an enhanced sampling technique is used to achieve the unfolding in a reasonable simulation time. Here we exploit REST2 [123, 124], described in detail in Chapter 2, a Hamiltonian-exchange replica exchange technique that allows us to achieve unfolding within a 200 ns/replica simulation time and estimate the melting temperature *in silico*. In each of the three systems, prepared as described in the previous section, only one protein is subjected to Hamiltonian rescaling and targeted for unfolding, while the other proteins, along with the solvent, are subjected to an unscaled Hamiltonian. The entire simulation setup is run at 300 K, although the mean-field rescaling scheme [124] grants the rescaled protein to explore a potential energy surface available in the temperature range T=[292, 528] K for Lysozyme in solution and hydrated powder, and T=[291, 720] K for Lysozyme embedded in a powder with glycerol. Each system was simulated using 24 replicas allowing simultaneously the protein to unfold and achieving good energy replica overlap, finally yielding a 40 % exchange efficiency with exchanges attempted every 10 ps. The simulations were performed with the CHARMM22/CMAP force-field and are run in the NpT ensemble using algorithms described in Chapter 2. The non-bonded and short-range electrostatic interactions are calculated for atoms within a 9.0 Å cut off, while PME with a grid spacing of 1.2 Å is used for long-range electrostatics. The equations of motion were integrated using the multiple timestep integration scheme, with the shortest integration time of 2 fs. Simulation snapshots were saved every 20 ps.

5.2.4 Molecular Dynamics Simulations

All-atom Molecular Dynamics simulations were performed for the Lysozyme in solution and powder systems, with the goal of calculating mean squared displacement of hydrogen atoms at temperatures approaching melting. After determining the *in silico* melting temperature and defining the criteria of folded and unfolded state, as explained later in the Results section, the Molecular Dynamics simulations were ran with rescaling the Hamiltonian in an equivalent manner as for the REST2 simulations, see previous section and Chapter 2. For each system, four configurations corresponding to the unfolded state and one configuration corresponding to the folded state were chosen, and separate MD runs were performed. Again, the Hamiltonian was rescaled for one protein in each system, and according to the mean-field rescaling scheme [124], the rescaled portion was effectively thermalized over a temperature range T=[300, 393] K for Lysozyme

in solution, T=[300, 469] K for Lysozyme in powder with water, and T=[300, 586] K for Lysozyme in powder with glycerol. The simulation scheme is equivalent to the one described in the previous section for REST2 MD, with the difference that the snapshots were saved every 2 ps and the production ran of 20 ns.

5.2.5 Analysis of Native Contacts

The Native Contacts definition was adopted from Ref [85]. The native state of the Lysozyme in the three systems was defined by first performing conformational clustering (see Chapter 2) on the REST2 MD trajectories at T=300 K, with the RMSD=2.0 Å as clustering criterion. The native contacts that occur in configurations belonging to the most populated cluster more than 80 % of the time were used to define the Lysozyme native contacts, while only taking into account C_α atoms separated by 7 or more residues and by less than 10 Å. The average number of native contacts in a trajectory was then computed by averaging the following metric:

$$Q(t) = \frac{\sum_{i=1}^{N_{res}} \sum_{j=1}^{N_i} \frac{1}{1+e^{10(d_{ij}(t)-(d_{ij}^0+1))}}}{\sum_{i=1}^{N_{res}} N_i}, \tag{5.1}$$

where d_{ij}^0 are native contact distances, $d_{ij}(t)$ are the native contact distances existing in frame t, N_i are the native contacts of a residue i, and N_{res} is the number of residues, i.e. the number of C_α atoms present in the system. According to this definition, a protein folded in all trajectory frames would result in $< Q >= 1$, while the fully unfolded protein would yield $< Q >= 0$.

5.3 Results

5.3.1 The Experiment

As a starting point, we put forth the experimental results of the Elastic Incoherent Neutron Scattering Experiments (see Methods), shown in Figure 5.2. The experimental signal stems mainly from the hydrogen atoms, which have high incoherent scattering cross section. Hydrogen atoms are abundantly distributed throughout the protein and EINS allows sampling atomic protein fluctuations by estimating the atomic mean square displacements (MSD, see Chapter 2) of protein non-exchangeable hydrogen atoms [129, 244] from the dynamic structure factor measured in the experiment. The quantitative information in terms of MSD is extracted by adopting the double-well jump model (see

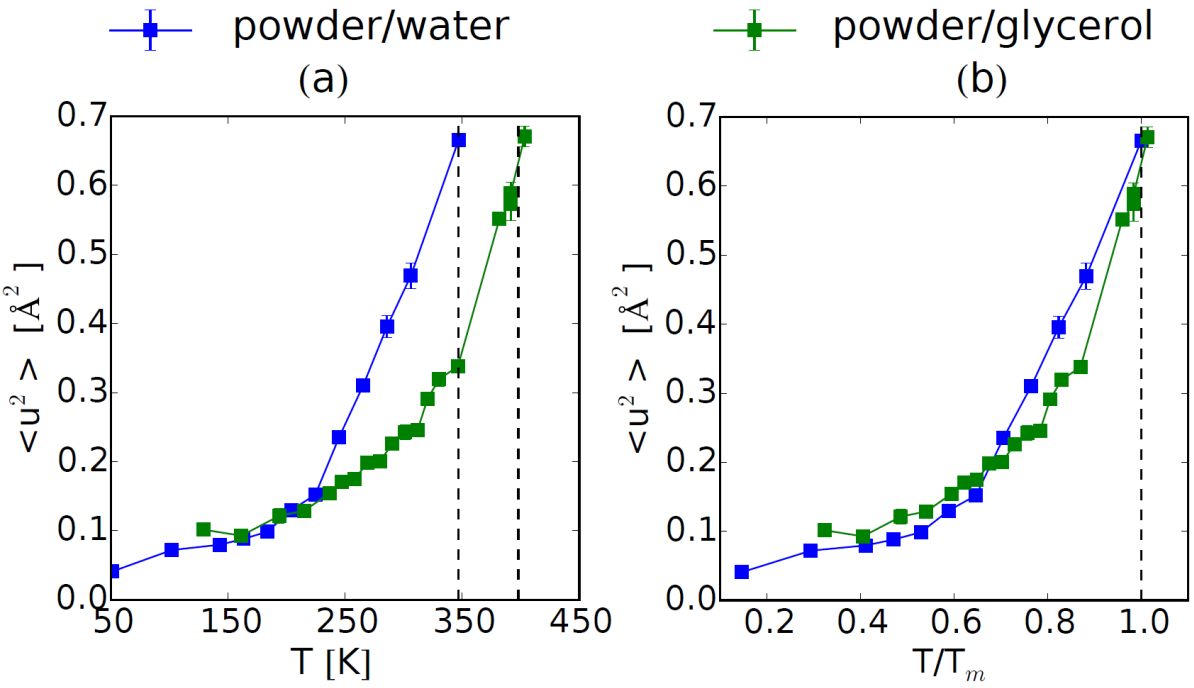

Figure 5.2: Mean squared displacement as a function of temperature as reported by an Elastic Incoherent Neutron Scattering experiment. Panel (a) shows the temperature in Kelvin, while panel (b) shows the normalized temperature scale so that the reader can easily appreciate the convergence of fluctuations when approaching the melting temperature, marked in the figures with a dashed vertical line. This plot is reported with the courtesy of A. Paciaroni.

Chapter 2), which has been successful in describing the dynamics of protein powders [129, 245] and proteins in glassy environments [244, 246, 247] on the same energy resolution and Q-range as in the present experiments. Additionally, the double well model can properly describe the non-Gaussian behavior of the elastic data. The experimental curves are normalized with respect to signal at a temperature of 20 K, thus showing in Figure 5.2 the hydrogen MSD in Lysozyme embedded in powders with water and glycerol, relative to the MSD at 20 K, as a function of temperature.

In Figure 5.2 (a), it can be appreciated that for both samples, the MSD deviate from the initial low-temperature trend at 175 K. This departure has been assigned in the past to the onset of methyl group dynamics [248, 249, 250] which is indeed independent of both the hydration degree and the type of glassy matrix in which the biomolecule is embedded [247, 249]. Further temperature increase drives the protein dynamical transition at T_d, marked by a steep rise of the protein MSD . The experimental data point to the fact that the dynamical route of Lysozyme towards thermal unfolding is

dependent on the surrounding matrix, reinforced by the shift of the protein dynamical transition onset closer to the melting temperature in the thermoprotective glycerol when compared to the dehydrated protein powder.

Most importantly, Figure 5.2 (b) shows similar internal dynamics of both samples at melting, indicated by comparable amplitudes of atomic displacements near the melting temperature $\langle u^2 \rangle$=0.68 Å^2. The reported value is expected to depend on the experimental energy resolution, possibly increasing as the accessible time range is extended to include slower dynamical processes. Using atomic fluctuations to probe melting is a familiar concept in the melting of solids and quantified in the Lindemann criterion, predicting the onset of melting when the atomic thermal fluctuations exceed a threshold value with respect to the equilibrium interatomic distances [239, 251]. To further inspect the experimental results, we make use of Molecular Dynamics simulations of the experimental systems and additionally Lysozyme in solution, aiming to compute the MSD average directly from the atomic-level resolved trajectories.

5.3.2 Protein Melting *in silico*

We begin our computational exploration by devising means to determine the melting curve of a protein *in silico*, a necessary step in order to follow the scaling of atomic fluctuations when approaching melting. The melting of small globular proteins, such as the Lysozyme, can be approximated with a two-state model in which only two states exist, the folded (native) and the unfolded (denatured) states, separated by an energy barrier [252]. It is to be noted that both the folded and unfolded states are in fact represented by an ensemble of protein substates. The model is kinetically justified by assuming the transitions between the unfolded states to be fast [253] and assuming that any intermediate states, i.e. the molten globule, are unstable and transiently populated. Within the framework of the two-state model, protein melting is defined by an equilibrium of folded and unfolded states in equal proportions, and finding the temperature at which this condition is satisfied will yield the *in silico* melting temperature.

To this end, we have used the REST2 method (see section Methods) to overcome the long simulation time necessary to sample the transition of folded to unfolded state in classical MD, as well as to overcome the limitations in sampling the unfolded state at appropriate temperatures due to inherent force fields limitations (see Chapter 2). REST2 scales well with system size and allows the exploration of the thermodynamics of protein stability, while the standard Replica Exchange MD is too expensive for our purposes. REST2 applies the corresponding state principle and instead of raising the

temperature in subsequently arranged protein replicas, the potential energy of a single protein in each system is scaled to allow for extensive sampling of the potential energy surface, including the unfolded state. It is to be noted that the scaling interests the intramolecular protein potential energy and the interaction between the target protein and the rest of the system. To distinguish the folded and the unfolded states, a 'reaction' coordinate or order parameter must be defined. Our choice is the RMSD, measuring the distance between the protein structures in the trajectories with respect to a reference, e.g. the equilibrated structure at T=300 K. The RMSD makes an appropriate choice as it measures the deformation of the protein matrix and is responsive to thermal excitation, as evident from Figure 5.3 (a). The temperature scale shown in all Figures is associated to different values of scaled potential energy in the REST2 method through the mean field rescaling scheme (see Chapter 2 and Ref [124]).

Based on the distribution of the RMSD in the trajectories, we chose the threshold RMSD=4.0 Å as the separating surface to distinguish the folded and unfolded states. The population of the folded state f was determined by applying Fermi function to the RMSD values:

$$f_i = \frac{1}{1+\exp(RMSD-4.0)}, \tag{5.2}$$

and counting the proportion of the folded state. Subsequent determination of the fraction of unfolded state is trivial with the assumption of the two-state model and the free energy is computed from the two fractions. Finally, the free energy is fit [254, 124]:

$$\Delta G = -\Delta C_v[T(ln(\frac{T}{T_m})-1)+T_m]+\Delta H_m(1+\frac{T}{T_m}), \tag{5.3}$$

where ΔC_v is the change in heat capacity between the folded and unfolded state, here considered constant, ΔH_m is the melting enthalpy, and T_m is the melting temperature. The main result reported in the Figure 5.3 (b) is the increase in the melting temperature upon introduction of crowders, and further increase when glycerol replaces water, agreeing well with the experimental trend presented in the previous section. The result is a familiar one, as the hydration of a dehydrated Lysozyme sample has previously found to lower the protein stability [255] and lyophilization is generally used as means of preserving proteins for transport and storing. Furthermore, glycerol has shown to favor the native state of Lysozyme [256], producing the observed cryoprotective effect.

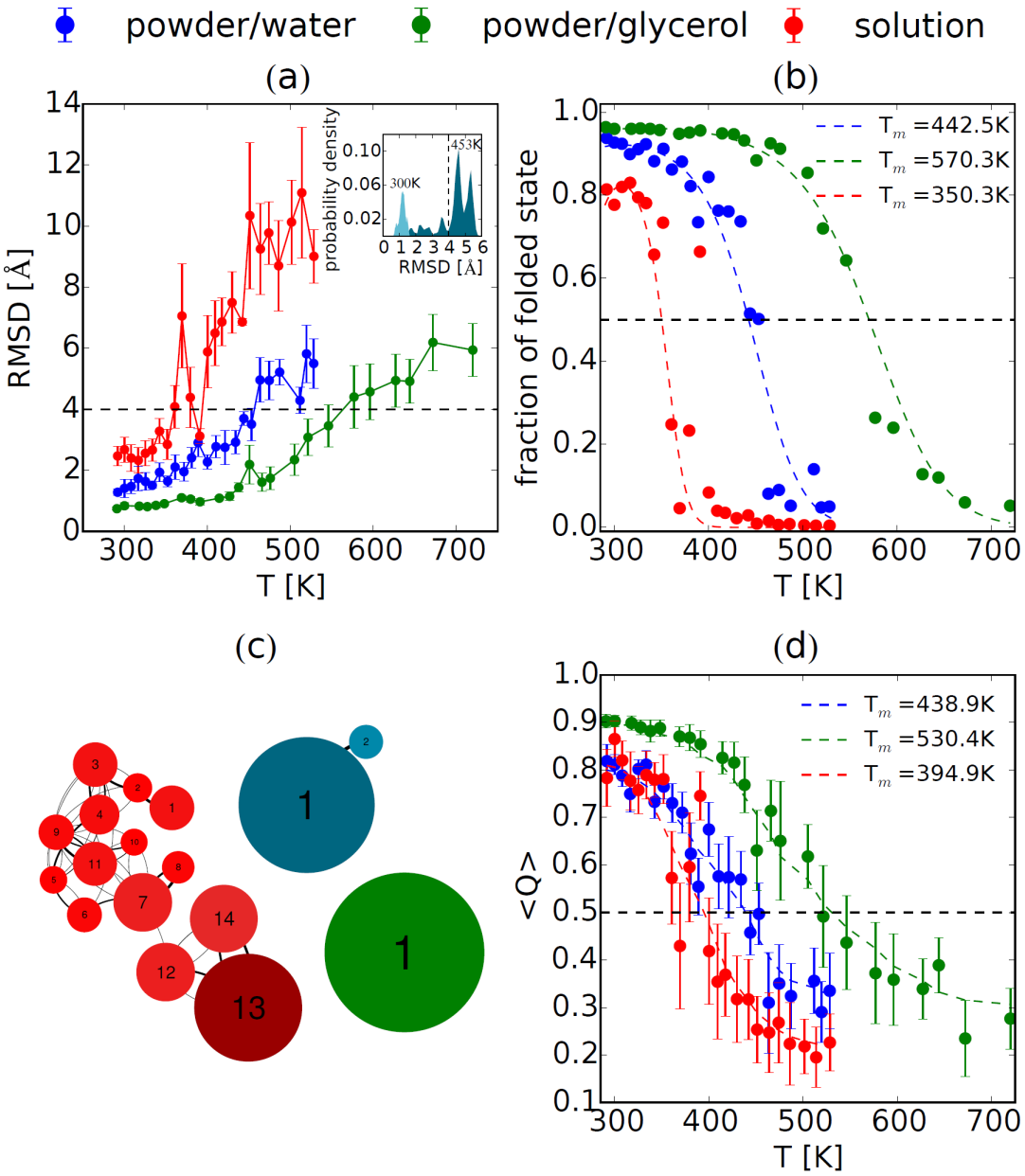

Figure 5.3: The melting curves of the protein Lysozyme shown by choosing different criteria to distinguish the folded and the unfolded substates. The RMSD temperature dependence is shown in (a) and a dividing criterion between the folded and unfolded state is chosen as 4.0 Å , illustrated by the inset figure where the RMSD distributions for the powder/water system are shown for two temperatures, T=300 K and T=453 K, and the dividing surface chosen is marked by a vertical line. The melting curve produced with the RMSD criterion is shown in (b). The native contacts can also be used to generate the melting curve, shown in (d). The native contact analysis is based on choosing the native contacts from the ensemble of most visited clusters. The clustering is shown in (c), where clustering C_α positions with a cut off RMSD=2.0 Å was applied. Note that the color code for the networks of conformational substates is equivalent as in the remainder of the figure.

As the results are entirely dependent on the choice of the order parameter and dividing surface used to separate the folded and unfolded state, we chose yet another parameter, the fraction of native states, to determine the melting curve of the three systems. The native contacts (see Methods and Ref [85]) are calculated form the ensemble of the most populated cluster in protein conformational clustering of trajectories at T=300 K. The networks of conformational substates for the three systems are shown in Figure 5.3 (c) and reveal the constraints on the configurational space of native state Lysozyme in powder, reducing the number of cluster leaders from 14 in solution to 2, and the combined effect of crowding and glycerol, producing protein dynamics described by a single cluster. The computed average fraction of native states at different temperatures represents the melting curve shown in Figure 5.3 (d). The melting temperatures at <Q>=0.5 are extracted after Gaussian smoothing, and they fully agree with previous observations. Interestingly, the new metric represents a strongly conservative definition of the folded and the unfolded state, as the melting curves show no fully folded or unfolded state for any of the three systems. Once with the melting temperatures in hand, we proceed with observing the atomistic fluctuations when approaching melting.

5.3.3 Scaling of the Atomic Fluctuations

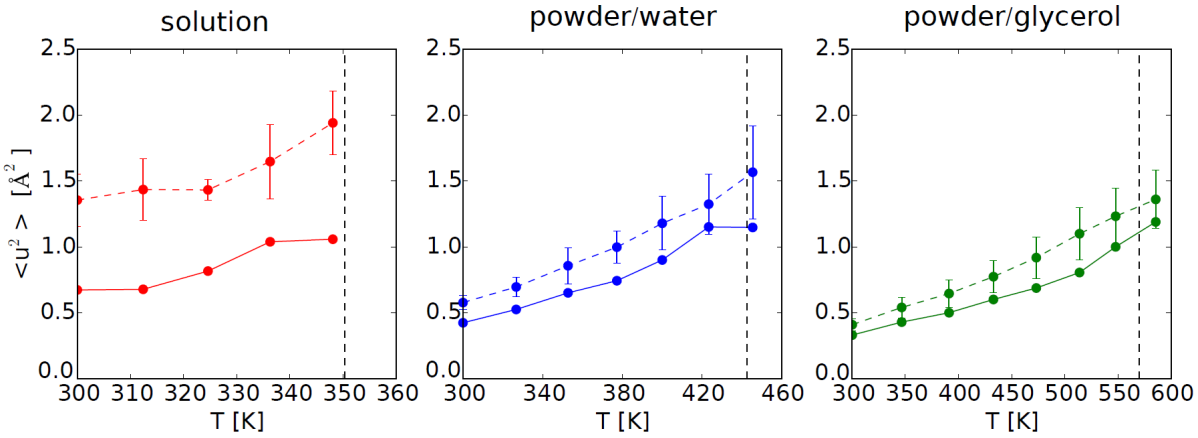

Figure 5.4: Atomistic fluctuations, expressed as MSD and calculated over a time window of 150 ps, of the protein Lysozyme in folded (full line) and unfolded (dashed line) states as a function of temperature. The three panels refer to the three systems, Lysozyme in solution, powder/water, and powder/glycerol. The vertical line marks the melting temperature as determined by using the RMSD as the unfolding order parameter.

Since approaching the melting the contribution from the unfolded state becomes increasingly important, we have carried out MD simulations of configurations belonging

to four different unfolded states and a single folded state, and computed the temperature dependence of their fluctuations for all three systems, shown in Figure 5.4. For the sake of coherence, we use the same Hamiltonian rescaling scheme as in the REST2 simulations, which allows us to maintain the same temperature scale and assure that we are indeed exploring the melting-approaching temperature regime. Atomistic fluctuations were monitored by computing the mean square displacement of hydrogen atoms attached to carbons, and using a time window of 150 ps as suggested by the experimental resolution (see Chapter 2 and section Methods). Upon examining the data in Figure 5.4, it is worth to note that for the Lysozyme in solution, the folded and unfolded systems exhibit markedly different atomistic fluctuations, with the unfolded protein exploring larger motion amplitudes. On the contrary, in the crowded systems, the atomistic fluctuations of the unfolded and folded states are quenched to very similar values. The data thus suggests that macromolecular crowding is the dominant factor in determining not only the large scale configuration changes due to volume exclusion [231, 232], but that it also exerts a local influence at an atomistic level. Further steps in the investigation would include the explicit determination of Voronoi volumes to quantify the extent of the excluded volume effect as well as running more simulations of the folded state protein for all systems.

As a final step in determining the fluctuations when approaching melting, we have weighted the atomistic fluctuations of the folded and unfolded states according to their statistical weights, i.e. the fractions of the folded and unfolded states:

$$< u^2 >= f < u^2 >_f + (1-f) < u^2 >_u, \tag{5.4}$$

where the MSD of the folded and unfolded states are given indices f and u, respectively, and f is the fraction of folded state. Results in Figure 5.5 clearly show the protein flexibility increase with increasing the temperature, as the system may access a larger number of conformational substates. Most notably, on increasing the temperature, the system reaches critical conditions where thermal melting takes place and shows a striking similarity in the magnitude of the atomic fluctuations, specifically in the powder systems. Note that the Lysozyme in solution is specific in that full hydration grants the protein full flexibility and a large number of available conformational substates exist in this condition. Thus, to achieve a proper weighing of fluctuations under full hydration, more configurational averages most likely have to be performed in order to achieve the convergence of the fluctuation magnitudes with crowded systems. Additionally, the high divergence of fluctuations for Lysozyme solution in Figure 5.5 (b) at all temperatures

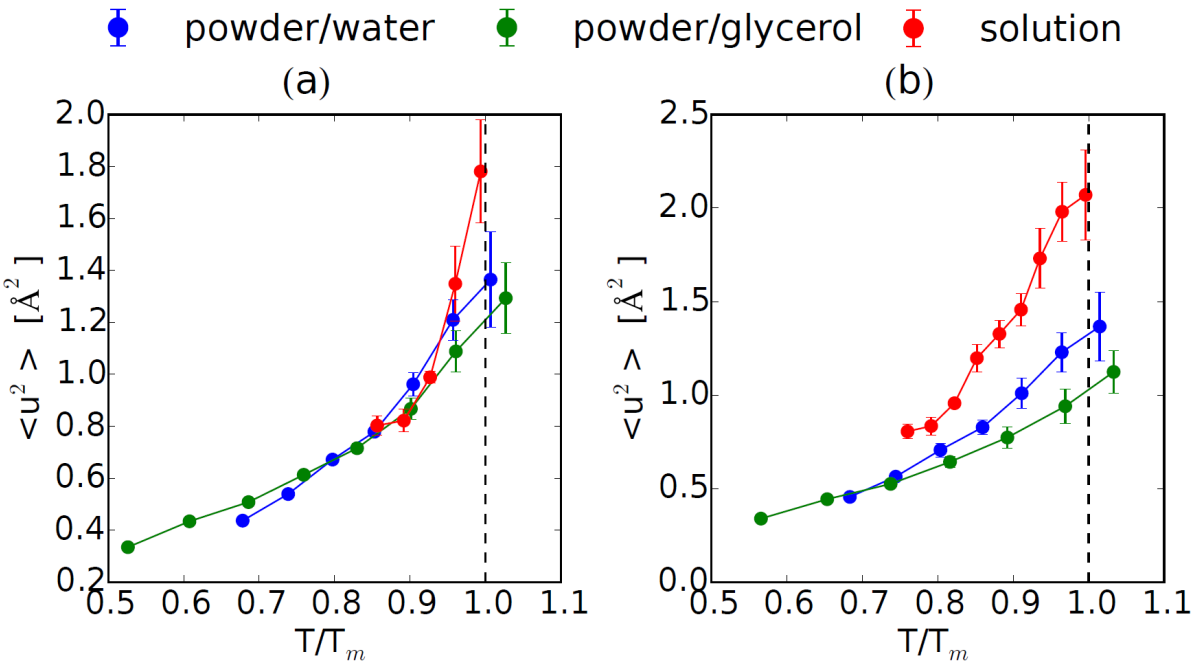

Figure 5.5: Combined total MSD calculated over a time window of 150 ps as a function of temperature. Panel (a) shows the MSD when RMSD is used to distinguish the folded from the unfolded state, while the panel (b) shows the same result when the fraction of native contacts is utilized to distinguish the two states. The temperature scale in panel (a) is normalized with values obtained from Fig 5.3 (b), while the temperature scale in panel (b) is normalized with values obtained from Fig 5.3 (d). The vertical line marks the normalized T_m, while the average is performed over four different unfolded states.

reflects that the fraction of native contacts yields a strong contribution of high amplitude fluctuations due to the lack of unfolding cooperativity the metric produces in the melting curve, and the subsequent high contribution of the fluctuations of the unfolded state when approaching melting. The effect is absent in the crowded conditions due to the similarities in the fluctuation magnitudes of the folded and unfolded states.

The similarity in the magnitudes of atomic fluctuations when approaching melting is a strong model-independent suggestion that the protein structural fluctuations at the melting point are similar, irrespective of the matrix around its surface. It can be argued that the common dynamical behavior corresponds to a condition in which the protein conformational substates are populated in the same critical way. Thus, the matrix surrounding the protein surface would manifest its bioprotectant character by abating the thermal fluctuation amplitude and shifting the critical flexibility to higher temperature conditions needed for protein melting.

5.3.4 Validating the Lindemann Criterion

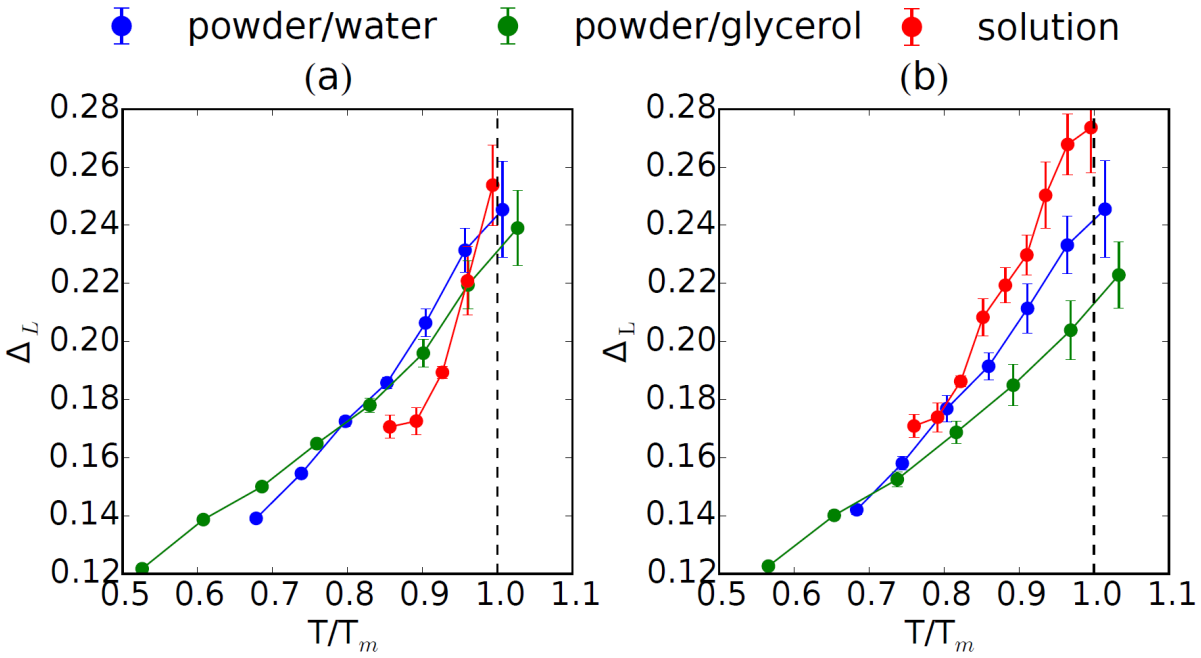

Figure 5.6: The Lindemann parameter, the root mean squared displacement divided by the typical nonbonded atomic distance, as a function of temperature. The typical lengths were determined separately for each system, as explained in text. Panel (a) shows the result when RMSD is used to distinguish the folded from the unfolded state, while the panel (b) shows the same result when the fraction of native contacts is utilized to distinguish the two states. The temperature scale in panel (a) is normalized with values obtained from Fig 5.3 (b), while the temperature scale in panel (b) is normalized with values obtained from Fig 5.3 (d). The vertical line marks the normalized T$_m$, while the average is performed over four different unfolded states.

Inspired by previous computational studies tracking the Lindemann criterion on heavy atom noble gas clusters [251] and proteins [257], we scaled the MSD with respect to the typical interatomic distance of the matrix, an equivalent to lattice space in solids. As the protein is not homogeneous, the characteristic distance reflects the criterion defined in Ref [257], computed as the most probable non-bonded interatomic heavy atom distance between different and non-neighboring residues of the main chain, with no cut off applied in the calculation. The calculations were performed on trajectories at 300 K, and reveal the effect of the crowding condition on the protein configuration and the interatomic separations. In the crowded conditions of the powder and glycerol system, the typical interatomic distance amounts to 4.75 Å, while in the solution conditions it increases to 5.25 Å. Previously reported values for Crambin, Ribonuclease A, Barnase,

and Myoglobin [257, 258] fall between 4.0-5.0 Å, agreeing reasonably well with our result. The data on the distribution of interatomic distances additionally reveals the well known effect of the crowded environment, in which the volume available to the protein is decreased, favoring the compact protein substates and conversely rendering smaller the observed interatomic separations [231].

By dividing the MSD with the typical interatomic separation $\bar{r}$, as previously argued, the Lindemann parameter is obtained $\Delta_L = \frac{\sqrt{<u^2>}}{\bar{r}}$, shown in Figure 5.6 as a function of temperature. Using the RMSD as the order parameter, Δ_L is similar in all Lysozyme systems and clusters at $\Delta_L \sim 0.25$. The native contact Δ_L values converge relatively well as compared to the MSD values in Figure 5.5 (b), all pointing to universal behavior. The Lindemann parameters for proteins have been previously estimated by utilizing MD simulations [257], yielding results in the range 0.12-0.16 at 300 K. The value of the Lindemann parameter predicting the onset of melting in solids is in this same range of values (0.1-0.15) [259, 260, 261], indicating that the values for proteins might be different than those for solids, as protein melting at ambient temperatures is not a likely event.

The temperature dependence of the Lindemann parameter has been studied for proteins Crambin and Ribonuclease A, associating an observed change in slope at Δ_L=0.14 with melting [257]. While falling in the range of the values predicted for solids, the result must be commented in the context of the following: (i) the critical value determined from the slope change corresponds to that determined at physiological temperature in the same type of simulations and should not be equated with melting, (ii) the melting temperature *in silico* has not been determined in the simulated conditions, thus the relationship between the simulated temperature range and the T_m is not familiar and the represented Δ_L values might not correspond to melting nor approaching melting, (iii) the calculations were 200 ps long, while we have shown, based on enhanced sampling 200 ns simulations, that the atomic fluctuations are configuration-dependent, and the relevant configuration changes are most likely not explored on a picosecond time scale. Nonetheless, the same study offers another interesting perspective, where the melting of the surface is onset prior to the melting of the core [257] and a computational analysis can be easily envisaged along similar lines, where computations presented in this chapter are accompanied by a separate treatment of the protein surface and core.

5.4 Conclusion

In conclusion, the present results provide a description of protein thermal fluctuations approaching the melting of the protein in different environments. The amplitude of the protein MSD at the melting point is rather independent of the environment, thus suggesting the existence of a threshold for the dynamical contribution of thermal fluctuations in native Lysozyme. The slight deviations from this universal scaling can be ascribed to a series of approximations made while determining the T_m (order parameter choice, two-state model, sampling in the simulation), calculation of $<u^2>$ (taking into account relevant protein configurations), and finally calculation of the typical interatomic separation. With this in mind, we put forth the result that the values of the Δ_L are similar for all three systems, showing that with the same error in aforementioned approximations, one can use the Δ_L as a hallmark of the melting process. It is particularly interesting to note that the magnitude of the atomistic fluctuations in the representative folded and unfolded states are rather similar in the powder systems, while in the dilute aqueous solution, the unfolded state exhibits systematically larger atomistic fluctuation due to its less compact nature. This implies that the vibrational entropy from the folded and unfolded states combine similarly at melting, but that their relative contribution is different depending on the environmental conditions.

CHAPTER 6

CONCLUSION

This thesis seeks to understand the effect of temperature on protein function and stability by considering three topical study-cases. As the optimal temperature window for protein activity is relatively narrow, we are interested in observing where and how the fine tuning is achieved in order to better understand the relationship between protein structure and activity. For this purpose, thermophilic proteins represent an ideal model as their stability at moderate and high temperatures is accompanied by protein activity in the high temperature regime only, suggesting that the relationship between a stable protein fold and an active protein is not as straightforward as is widely presented. In our investigation, we have systematically compared similar proteins with different optimal working temperatures in order to tackle the stability/function trade-off, as well as its relationship with protein mechanical flexibility.

In performing Molecular Dynamics studies and coupling them to Neutron Scattering experiments, we have investigated the validity of two classic paradigms related to temperature effects - the Somero's corresponding state principle and the Lindemann criterion. The former correlates the emergence of enzymatic activity in thermophilic proteins to the thermal activation of protein flexibility, while the latter defines the critical magnitude of atomic fluctuations to define melting. Both principles witness the central role the temperature plays in modulating evolutionary adaptations.

In the first study presented in the thesis, we investigated the G-domain conformational changes occurring during the ideal enzymatic turnover of a pair of mesophilic and hyperthermophilic homologues. Considering the essential functional modes, we verified

that the two proteins behave similarly when placed at their working temperatures, which is in agreement with the essence of Somero's principle. The scaled relationship among flexibility and function was also investigated for a pair of homologous Lactate/Malate Dehydrogenase proteins. The data we present in the manuscript regard the mesophilic species only, but a preliminary comparison with the results for the thermophilic protein confirms a shifted thermal activation of the allosteric-like, functional modes of the two proteins at their working temperatures. The final system considered, Lysozyme in powder condition, helped us in assessing a complementary aspect of the relationship among mechanical flexibility, stability, and function, with the result of proving *universal* scaling of atomistic flexibility when approaching the unfolding transition.

Although our studies offer substantial data on different effects in protein thermal (in)stability and (in)activity, they also open new doors and make opportunities to further push the frontiers in several directions. Regarding the activity of the G-domain, an interesting prospect is focusing on the chemical step of the enzymatic activity, therefore employing quantum/classical simulations to evaluate the barrier for the GTP hydrolysis in the two proteins. It would be also of interest to estimate the free energy of α to β secondary structure transition with the goal of establishing whether the structurally conserved portions of the protein have evolved as energetic drains in enzymes, and to which extent these drains can be used to influence protein stability. Furthermore, the long-range communication studies on the Lactate/Malate Dehydrogenases will continue in protein holo states as the next logical step. The complexity of the system offers numerous possibilities that can be undertaken to examine the role of its symmetry and multiple oligomeric states, as well as to understand the substrate binding and population in multimeric enzymes. MD explorations will prove to be, once and again, indispensable in these endeavors, as experimentally impossible situations will be easily achievable - populating varying number of domain active sites and observing the effect in long-range communication across the domains is an exciting possibility to be explored. Lastly the Lysozyme remains a favorite 'toy' model to further push in understating protein folding and stability in crowded, cytoplasm-like conditions. Our studies have only scraped the surface of the matter, and many questions are left to answer. Among them the most obvious are the effect of inert crowders, i.e. with no long-range interactions, and examining the model case of reduced Lysozyme, i.e. without disulfide bonds. These studies are under way and will illuminate the atomistic details of the crowded condition.

BIBLIOGRAPHY

[1] O. Holderer and O. Ivanova. “J-NSE: Neutron spin echo spectrometer”. In: *Journal of Large-Scale Research Facilities JLSRF* 1 (2015), p. 11.

[2] G. L. Früh-Green et al. “30,000 Years of Hydrothermal Activity at the Lost City Vent Field”. In: *Science* 301.5632 (2003), pp. 495–498.

[3] K. Ludwig et al. *U/Th Geochronology of Carbonate Chimneys at the Lost City Hydrothermal Field*. Washington: AGU Fall Meeting, 2005.

[4] S. Akanuma et al. “Experimental Evidence for the Thermophilicity of Ancestral Life”. In: *Proceedings of the National Academy of Sciences* 110.27 (2013), pp. 11067–11072.

[5] W. Martin et al. “Hydrothermal Vents and the Origin of Life”. In: *Nature Reviews Microbiology* 6.11 (2008), pp. 805–814.

[6] S. L. Miller and A. Lazcano. “The Origin of Life - Did it Occur at High Temperatures?” In: *Journal of Molecular Evolution* 41.6 (1995), pp. 689–692.

[7] S. C. Cary et al. “On the Rocks: the Microbiology of Antarctic Dry Valley Soils”. In: *Nature Reviews Microbiology* 8.2 (2010), pp. 129–138.

[8] E. Blöchl et al. “Pyrolobus fumarii, gen. and sp. nov., Represents a Novel Group of Archaea, Extending the Upper Temperature Limit for Life to 113^oC”. In: *Extremophiles* 1.1 (1997), pp. 14–21.

[9] K. Takai et al. “Cell Proliferation at 122^oC and Isotopically Heavy CH_4 Production by a Hyperthermophilic Methanogen under High-Pressure Cultivation”. In: *Proceedings of the National Academy of Sciences* 105.31 (2008), pp. 10949–10954.

[10] R. Corkrey et al. “The Biokinetic Spectrum for Temperature”. In: *PLOS ONE* 11.4 (Apr. 2016), pp. 1–29.

[11] A. Ilari and C. Savino. “Protein Structure Determination by X-ray Crystallography”. In: *Bioinformatics: Data, Sequence Analysis and Evolution* (2008), pp. 63–87.

[12] D. Marion. “An Introduction to Biological NMR Spectroscopy”. In: *Molecular & Cellular Proteomics* 12.11 (2013), pp. 3006–3025.

[13] Z. H. Zhou. “Atomic Resolution cryo-Electron Microscopy of Macromolecular Complexes”. In: *Advances in Protein Chemistry and Structural Biology* 82 (2011), p. 1.

[14] E. Callaway. “The Revolution Will not be Crystallized: a New Method Sweeps Through Structural Biology”. In: *Nature* 525.7568 (2015), p. 172.

[15] A. Quintas. “What Drives an Amyloid Protein Precursor from an Amyloidogenic to a Native-Like Aggregation Pathway?” In: *OA Biochemistry* 1 (2013), p. 6.

[16] P. G. Wolynes. “Recent Successes of the Energy Landscape Theory of Protein Folding and Function”. In: *Quarterly Reviews of Biophysics* 38.04 (2005), pp. 405–410.

[17] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes. “Theory of Protein Folding: the Energy Landscape Perspective”. In: *Annual Review of Physical Chemistry* 48.1 (1997), pp. 545–600.

[18] J. N. Onuchic and P. G. Wolynes. “Theory of Protein Folding”. In: *Current Opinion in Structural Biology* 14.1 (2004), pp. 70–75.

[19] H. Nymeyer, A. E. Garcia, and J. N. Onuchic. “Folding Funnels and Frustration in off-Lattice Minimalist Protein Landscapes”. In: *Proceedings of the National Academy of Sciences* 95.11 (1998), pp. 5921–5928.

[20] N. Go. “Theoretical Studies of Protein Folding”. In: *Annual Review of Biophysics and Bioengineering* 12.1 (1983), pp. 183–210.

[21] J. A. Schellman. “Temperature, Stability, and the Hydrophobic Interaction”. In: *Biophysical Journal* 73.6 (1997), p. 2960.

[22] S. Kumar, C. J. Tsai, and R. Nussinov. “Temperature Range of Thermodynamic Stability for the Native State of Reversible Two-State Proteins”. In: *Biochemistry* 42.17 (2003), pp. 4864–4873.

[23] C. N. Pace. “Energetics of Protein Hydrogen Bonds”. In: *Nature Structural and Molecular biology* 16.7 (2009), pp. 681–682.

[24] K. Teilum, J. G. Olsen, and B. B. Kragelund. “Protein Stability, Flexibility and Function”. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1814.8 (2011), pp. 969–976.

[25] D. C. Rees and A. D. Robertson. “Some Thermodynamic Implications for the Thermostability of Proteins”. In: *Protein Science* 10.6 (2001), pp. 1187–1194.

[26] T. Haltia and E. Freire. “Forces and Factors that Contribute to the Structural Stability of Membrane Proteins”. In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1228.1 (1995), pp. 1–27.

[27] J. Backmann and G. Schaefer. “Thermodynamic Analysis of Hyperthermostable Oligomeric Proteins”. In: *Methods in Enzymology* 334 (2001), pp. 328–342.

[28] W. J. Becktel and J. A. Schellman. “Protein Stability Curves”. In: *Biopolymers* 26.11 (1987), pp. 1859–1877.

[29] P. L. Privalov. “Cold Denaturation of Proteins”. In: *Critical Reviews in Biochemistry and Molecular Biology* 25.4 (1990), pp. 281–306.

[30] P. C. Rathi, H. W. Höffken, and H. Gohlke. “Quality Matters: Extension of Clusters of Residues with Good Hydrophobic Contacts Stabilize (Hyper)Thermophilic Proteins”. In: *Journal of Chemical Information and Modeling* 54.2 (2014), pp. 355–361.

[31] F. Sterpone, G. Stirnemann, and D. Laage. “Magnitude and Molecular Origin of Water Slowdown Next to a Protein”. In: *Journal of the American Chemical Society* 134.9 (2012), pp. 4116–4119.

[32] A. Razvi and J. M. Scholtz. “Lessons in Stability from Thermophilic Proteins”. In: *Protein Science* 15.7 (2006), pp. 1569–1578.

[33] G. Feller. “Protein Stability and Enzyme Activity at Extreme Biological Temperatures”. In: *Journal of Physics: Condensed Matter* 22.32 (2010), p. 323101.

[34] A. Karshikoff and R. Ladenstein. “Ion Pairs and the Thermotolerance of Proteins from Hyperthermophiles: a Traffic Rule for Hot Roads”. In: *Trends in Biochemical Sciences* 26.9 (2001), pp. 550–557.

[35] I. N. Berezovsky et al. “Entropic Stabilization of Proteins and its Proteomic Consequences”. In: *PLOS Computational Biology* 1.4 (2005), e47.

[36] L. Xiao and B. Honig. “Electrostatic Contributions to the Stability of Hyperthermophilic Proteins”. In: *Journal of Molecular Biology* 289.5 (1999), pp. 1435–1444.

[37] J. M. Sanchez-Ruiz and G. I. Makhatadze. “To Charge or Not to Charge?” In: *Trends in Biotechnology* 19.4 (2001), pp. 132–5.

[38] S. Basu and S. Sen. “Turning a Mesophilic Protein into a Thermophilic One: a Computational Approach Based on 3D Structural Features”. In: *Journal of Chemical Information and Modeling* 49.7 (2009), pp. 1741–50.

[39] M. J. Thompson and D. Eisenberg. “Transproteomic Evidence of a Loop-Deletion Mechanism for Enhancing Protein Thermostability”. In: *Journal of Molecular Biology* 290.2 (1999), pp. 595–604.

[40] A. Karshikoff and R. Ladenstein. “Proteins from Thermophilic and Mesophilic Organisms Essentially Do Not Differ in Packing”. In: *Protein Engineering* 11.10 (1998), pp. 867–72.

[41] B.Schuler et al. “Role of Entropy in Protein Thermostability: Folding Kinetics of a Hyperthermophilic Cold Shock Protein at High Temperatures Using 19F NMR”. In: *Biochemistry* 41.39 (Oct. 2002), pp. 11670–11680.

[42] G. Hernandez et al. “Millisecond Time Scale Conformational Flexibility in a Hyperthermophile Protein at Ambient Temperature”. In: *Proceedings of the National Academy of Sciences* 97.7 (2000), pp. 3166–3170.

[43] J. Fitter and J. Heberle. “Structural Equilibrium Fluctuations in Mesophilic and Thermophilic α-Amylase”. In: *Biophysical Journal* 79.3 (Sept. 2000), pp. 1629–1636.

[44] L. Meinhold et al. “Protein Dynamics and Stability: The Distribution of Atomic Fluctuations in Thermophilic and Mesophilic Dihydrofolate Reductase Derived Using Elastic Incoherent Neutron Scattering”. In: *Biophysical Journal* 94.12 (June 2008), pp. 4812–4818.

[45] M. Kalimeri et al. “How Conformational Flexibility Stabilizes the Hyperthermophilic Elongation Factor G-domain”. In: *The Journal of Physical Chemistry B* 117.44 (2013), pp. 13775–13785.

[46] S. Cavagnero et al. “Kinetic Role of Electrostatic Interactions in the Unfolding of Hyperthermophilic and Mesophilic Rubredoxins”. In: *Biochemistry* 37.10 (1998), pp. 3369–76.

[47] F. Sterpone et al. “Key Role of Proximal Water in Regulating Thermostable Proteins”. In: *The Journal of Physical Chemistry B* 113.1 (2009), pp. 131–7.

[48] X. Zhang et al. “X-ray Structure Analysis and Crystallographic Refinement of Lumazine Synthase from the Hyperthermophile Aquifex aeolicus at 1.6 Å Resolution: Determinants of Thermostability Revealed from Structural Comparisons”. In: *Journal of Molecular Biology* 306.5 (2001), pp. 1099 –1114.

[49] R. Salari and L. T. Chong. “Desolvation Costs of Salt Bridges across Protein Binding Interfaces: Similarities and Differences between Implicit and Explicit Solvent Models”. In: *The Journal of Physical Chemistry Letters* 1.19 (2010), pp. 2844–2848.

[50] F. Sterpone and S. Melchionna. “Thermophilic Proteins: Insight and Perspective from In Silico Experiments”. In: *Chemical Society Reviews* 41 (5 2012), pp. 1665–1676.

[51] F. Sterpone and S. Melchionna. *Thermostable Proteins: Role of Packing, Hydration, and Fluctuations on Thermostability*. 2012, pp. 21–46.

[52] S. Kumar and R. Nussinov. “How Do Thermophilic Proteins Deal with Heat?” In: *Cellular and Molecular Life Sciences* 58 (9 2001). 10.1007/PL00000935, pp. 1216–1233.

[53] R. Jaenicke and G. Böhm. “The Stability of Proteins in Extreme Environments”. In: *Current Opinion in Structural Biology* 8.6 (Dec. 1998), pp. 738–748.

[54] C. Vieille and G. J. Zeikus. “Hyperthermophilic Enzymes: Sources, Uses, and Molecular Mechanisms for Thermostability”. In: *Microbiology and Molecular Biology Reviews* 65.1 (2001), pp. 1–43.

[55] K. Henzler-Wildman and D. Kern. “Dynamic Personalities of Proteins”. In: *Nature* 450 (2007), pp. 964–972.

[56] G. N. Somero. “Proteins and Temperature”. In: *Annual Review of Physiology* 57 (1995), pp. 43–68.

[57] G. N. Somero. “Temperature Adaptation of Enzymes”. In: *Annual Review of Ecology, Evolution, and Systematics* 9 (1978), pp. 1–29.

[58] M. J. Danson et al. “Enzyme Thermostability and Thermoactivity”. In: *Protein engineering* 9.8 (1996), pp. 629–630.

[59] P. Závodszky et al. “Adjustment of Conformational Flexibility is a Key Event in the Thermal Adaptation of Proteins”. In: *Proceedings of the National Academy of Sciences* 95 (1998), pp. 7406–7411.

[60] A. Kohen and J. P. Klinman. “Protein Flexibility Correlates with Degree of Hydrogen Tunneling in Thermophilic and Mesophilic Alcohol Dehydrogenases”. In: *Journal of the American Chemical Society* 122.43 (2000), pp. 10738–10739.

[61] T. Lazaridis, I. Lee, and M. Karplus. “Dynamics and Unfolding Pathways of a Hyperthermophilic and a Mesophilic Rubredoxin”. In: *Protein Science* 6.12 (Dec. 1997), pp. 2589–2605.

[62] M. Roca et al. “On the Relationship Between Thermal Stability and Catalytic Power of Enzymes”. In: *Biochemistry* 46.51 (2007), pp. 15076–88.

[63] E. D. Merkley, W. W. Parson, and V. Daggett. “Temperature Dependence of the Flexibility of Thermophilic and Mesophilic Flavoenzymes of the Nitroreductase Fold”. In: *Protein Engineering, Design and Selection* 23.5 (2010), pp. 327–36.

[64] S. Radestock and H. Gohlke. “Protein Rigidity and Thermophilic Adaptation”. In: *Proteins* 79.4 (2011), pp. 1089–1108.

[65] P. C. Rathi, K.-E. Jaeger, and H. Gohlke. “Structural Rigidity and Protein Thermostability in Variants of Lipase A from Bacillus subtilis”. In: *PLOS One* 10 (2015), e0130289.

[66] M. Tehei et al. “Neutron Scattering Reveals the Dynamic Basis of Protein Adaptation to Extreme Temperature”. In: *The Journal of Biological Chemistry* 280.49 (2005), pp. 40974–40979.

[67] E. Marcos, P. Mestres, and R. Crehuet. “Crowding Induces Differences in the Diffusion of Thermophilic and Mesophilic Proteins: A New Look at Neutron Scattering Results”. In: *Biophysical Journal* 101.11 (2011), pp. 2782 –2789.

[68] P. A. Calligari et al. “Adaptation of Extremophilic Proteins with Temperature and Pressure: Evidence from Initiation Factor 6”. In: *The Journal of Physical Chemistry B* 119.25 (2015), pp. 7860–7873.

[69] T. Collins et al. “Activity, Stability and Flexibility in Glycosidases Adapted to Extreme Thermal Environments”. In: *Journal of Molecular Biology* 328 (2003), pp. 419–428.

[70] M. Wolf-Watz et al. “Linkage Between Dynamics and Catalysis in a Thermophilic-Mesophilic Enzyme Pair”. In: *Nature Structural and Molecular Biology* 11.10 (2004), pp. 945–949.

[71] M. Kalimeri et al. “Interface Matters: The Stiffness Route to Stability of a Thermophilic Tetrameric Malate Dehydrogenase”. In: *PLOS One* 9 (2014), e113895.

[72] A. Kohen et al. “Enzyme Dynamics and Hydrogen Tunnelling in a Thermophilic Alcohol Dehydrogenase”. In: *Nature* 399.6735 (1999), pp. 496–499.

[73] M. H. M. Olsson, W. W. Parson, and A. Warshel. “Dynamical Contributions to Enzyme Catalysis: Critical Tests of A Popular Hypothesis”. In: *Chemical Reviews* 106.5 (2006), pp. 1737–1756.

[74] H. S. Kim et al. “Structure and Hydride Transfer Mechanism of a Moderate Thermophilic Dihydrofolate Reductase from Bacillus stearothermophilus and Comparison to Its Mesophilic and Hyperthermophilic Homologues”. In: *Biochemistry* 44.34 (2005), pp. 11428–11439.

[75] J. Guo et al. “Effect of Dimerization on Dihydrofolate Reductase Catalysis”. In: *Biochemistry* 52.22 (2013), pp. 3881–3887.

[76] L. Y. P. Luk, E. J. Loveridge, and R. K. Allemann. “Different Dynamical Effects in Mesophilic and Hyperthermophilic Dihydrofolate Reductases”. In: *Journal of the American Chemical Society* 136.19 (2014), pp. 6862–6865.

[77] M. Karplus. *Martin Kaplus Nobel Lecture: Development of Multiscale Models for Complex Chemical Systems From H+H2 to Biomolecules*. 2013.

[78] M. Katava et al. “Stability and Function at High Temperature. What Makes a Thermophilic GTPase Different from its Mesophilic Homologue”. In: *The Journal of Physical Chemistry B* 120 (2016), pp. 2721–2730.

[79] Y. Shan et al. “How Does a Drug Molecule Find Its Target Binding Site?” In: *Journal of the American Chemical Society* 133.24 (2011), pp. 9181–9183.

[80] K. Schulten and M. Tesch. “Coupling of Protein Motion to Electron Transfer: Molecular Dynamics and Stochastic Quantum Mechanics Study of Photosynthetic Reaction Centers”. In: *Chemical Physics* 158.2-3 (1991), pp. 421–446.

[81] M. Karplus and A. J. McCammon. “Molecular Dynamics Simulations of Biomolecules”. In: *Nature Structural and Molecular Biology* 9.9 (2002), pp. 646 –652.

[82] S. A. Adcock and J. A. McCammon. “Molecular Dynamics: A Survey of Methods for Simulating the Activity of Proteins”. In: *Chemical Reviews* 106.5 (2006), pp. 1589–1615.

[83] C. Liu et al. “Cyclophilin A Stabilizes the HIV-1 Capsid Through a Novel Non-Canonical Binding Site”. In: *Nature Communications* 7.10714 (2016).

[84] D. S. D. Larsson, L. Liljas, and D. van der Spoel. “Virus Capsid Dissolution Studied by Microsecond Molecular Dynamics Simulations”. In: *PLOS Comput Biol* 8.5 (May 2012), pp. 1–8.

[85] K. Lindorff-Larsen et al. “How Fast-Folding Proteins Fold”. In: *Science* 334.6055 (2011), pp. 517–520.

[86] M. E. Tuckerman. *Protein-Ligand Interactions*. Oxford Graduate Texts, 2010.

[87] M. P. Allen and D. J. Tildesley. *Computer Simulations of Liquids*. Oxford Science Publications, 1987.

[88] D. Frenkel and B. Smit. *Understanding Molecular Simulations*. Academic Press, 2002.

[89] J. C. Phillips et al. “Scalable Molecular Dynamics with NAMD”. In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1781–1802.

[90] P. Ballone. “Modeling Potential Energy Surfaces: From First-Principle Approaches to Empirical Force Fields”. In: *Entropy* 16.1 (2013), p. 322.

[91] J. A. D. MacKerell et al. “All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins”. In: *The Journal of Physical Chemistry B* 102.18 (1998), pp. 3586–3616.

[92] J. Ponder and D. Case. “Force Fields for Protein Simulation”. In: *Advances in Protein Chemistry* 66 (2003), pp. 27–95.

[93] G. A. Kaminski et al. “Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides”. In: *The Journal of Physical Chemistry B* 105.28 (2001), pp. 6474–6487.

[94] A. D. MacKerell, M. Feig, and C. 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”. In: *Journal of Computational Chemistry* 25 (2004), pp. 1400–1415.

[95] D. I. Freedberg et al. “Discriminating the Helical Forms of Peptides by NMR and Molecular Dynamics Simulation”. In: *Journal of the American Chemical Society* 126.33 (2004), pp. 10478–10484.

[96] K. Lindorff-Larsen et al. “Systematic Validation of Protein Force Fields against Experimental Data”. In: *PLOS ONE* 7.2 (Feb. 2012), e32131.

[97] S. Piana, K. Lindorff-Larsen, and D. E. Shaw. “How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization”. In: *Biophysical Journal* 100.9 (2011), pp. L47 –L49.

[98] M. Buck et al. “Importance of the CMAP Correction to the CHARMM22 Protein Force Field: Dynamics of Hen Lysozyme”. In: *Biophysical Journal* 90.4 (2006), pp. L36 –L38.

[99] R. B. Best et al. “Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain $\chi 1$ and $\chi 2$ Dihedral Angles”. In: *Journal of Chemical Theory and Computation* 8.9 (2012), pp. 3257–3273.

[100] C. M. Baker. “Polarizable Force Fields for Molecular Dynamics Simulations of Biomolecules”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 5.2 (2015), pp. 241–254.

[101] E. Neria, S. Fischer, and M. Karplus. “Simulation of Activation Free Energies in Molecular Systems”. In: *The Journal of Chemical Physics* 105.5 (1996), pp. 1902–1921.

[102] W. L. Jorgensen et al. “Comparison of Simple Potential Functions for Simulating Liquid Water”. In: *The Journal of Chemical Physics* 79.2 (1983), pp. 926–935.

[103] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. “The Missing Term in Effective Pair Potentials”. In: *The Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[104] M. W. Mahoney and W. L. Jorgensen. “A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions”. In: *The Journal of Chemical Physics* 112.20 (2000), pp. 8910–8922.

[105] J. Huang and A. D. MacKerell. “CHARMM36 All-Atom Additive Protein Force Field: Validation Based on Comparison to NMR Data”. In: *Journal of Computational Chemistry* 34.25 (2013), pp. 2135–2145.

[106] R. B. Best et al. “Inclusion of Many-Body Effects in the Additive CHARMM Protein CMAP Potential Results in Enhanced Cooperativity of α-Helix and β-Hairpin Formation”. In: *Biophysical Journal* 103.5 (2012), pp. 1045 –1051.

[107] E. A. Cino, W.-Y. Choy, and M. Karttunen. “Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations”. In: *Journal of Chemical Theory and Computation* 8.8 (2012), pp. 2725–2740.

[108] W. Chen et al. “Conformational Dynamics of Two Natively Unfolded Fragment Peptides: Comparison of the AMBER and CHARMM Force Fields”. In: *The Journal of Physical Chemistry B* 119.25 (2015), 7902,Äì7910.

[109] R. B. Best and J. Mittal. “Free-Energy Landscape of the GB1 Hairpin in All-Atom Explicit Solvent Simulations with Different Force Fields: Similarities and Differences”. In: *Proteins: Structure, Function, and Bioinformatics* 79.4 (2011), pp. 1318–1328.

[110] W. L. Jorgensen and C. Jenson. “Temperature Dependence of TIP3P, SPC, and TIP4P Water from NPT Monte Carlo Simulations: Seeking Temperatures of Maximum Density”. In: *Journal of Computational Chemistry* 19.10 (1998), pp. 1179–1186.

[111] J. Zhang, Y. Shi, and P. Ren. *Polarizable Force Fields for Scoring Protein-Ligand Interactions*. Wiley-VCH Verlag GmbH and Co. KGaA, 2012, pp. 99–120.

[112] B. A. Bauer and S. Patel. “Recent Applications and Developments of Charge Equilibration Force Fields for Modeling Dynamical Charges in Classical Molecular Dynamics Simulations”. In: *Theoretical Chemistry Accounts* 131.3 (2012), pp. 1–15.

[113] G. Lamoureux and B. Roux. “Modeling Induced Polarization with Classical Drude Oscillators: Theory and Molecular Dynamics Simulation Algorithm”. In: *The Journal of Chemical Physics* 119.6 (2003), pp. 3025–3039.

[114] P. Ren and J. W. Ponder. “Polarizable Atomic Multipole Water Model for Molecular Mechanics Simulation”. In: *The Journal of Physical Chemistry B* 107.24 (2003), pp. 5933–5947.

[115] U. Essmann et al. “A Smooth Particle Mesh Ewald Method”. In: *The Journal of Chemical Physics* 103.19 (1995), pp. 8577–8593.

[116] P. P. Ewald. “Die Berechnung Optischer und Elektrostatischer Gitterpotentiale”. In: *Annalen der Physik* 369.3 (1921), pp. 253–287.

[117] D. Quigley and M. I. J. Probert. “Langevin Dynamics in Constant Pressure Extended Systems”. In: *The Journal of Chemical Physics* 120.24 (2004), pp. 11432–11441.

[118] G. J. Martyna, D. J. Tobias, and M. L. Klein. “Constant Pressure Molecular Dynamics Algorithms”. In: *The Journal of Chemical Physics* 101.5 (1994), pp. 4177–4189.

[119] S. E. Feller et al. “Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method”. In: *The Journal of Chemical Physics* 103.11 (1995), pp. 4613–4621.

[120] M. Tuckerman, B. J. Berne, and G. J. Martyna. “Reversible Multiple Time Scale Molecular Dynamics”. In: *The Journal of Chemical Physics* 97.3 (1992), pp. 1990–2001.

[121] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen. “Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes”. In: *Journal of Computational Physics* 23.3 (1977), pp. 327 –341.

[122] D. J. Earl and M. W. Deem. “Parallel Tempering: Theory, Applications, and New Perspectives”. In: *Physical Chemistry Chemical Physics* 7 (2005), pp. 3910–3916.

[123] L. Wang, R. Friesner, and B. Berne. “Replica Exchange with Solute Scaling: a More Efficient Version of Replica Exchange with Solute Tempering (REST2)”. In: *The Journal of Physical Chemistry B* 115 (2011), p. 11305.

[124] G. Stirnemann and F. Sterpone. “Recovering Protein Thermal Stability Using All-Atom Hamiltonian Replica-Exchange Simulations in Explicit Solvent”. In: *Journal of Chemical Theory and Computation* 11 (2015), pp. 5573–5577.

[125] C. Abrams and G. Bussi. “Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration”. In: *Entropy* 16.1 (2014), p. 163.

[126] F. Natali et al. “IN13 Backscattering Spectrometer at ILL: Looking for Motions in Biological Macromolecules and Organisms”. In: *Neutron News* 19.4 (2008), pp. 14 –18.

[127] S. Capaccioli et al. “Evidence of Coexistence of Change of Caged Dynamics at T_g and the Dynamic Transition at T_d in Solvated Proteins”. In: *The Journal of Physical Chemistry B* 116.6 (2012), pp. 1745–1757.

[128] Z. Yi et al. “Derivation of Mean-Square Displacements for Protein Dynamics from Elastic Incoherent Neutron Scattering”. In: *The Journal of Physical Chemistry B* 116.16 (2012), pp. 5028–5036.

[129] W. Doster, S. Cusack, and W. Petry. “Dynamical Transition of Myoglobin Revealed by Inelastic Neutron Scattering”. In: *Nature* 337 (1989), pp. 754–756.

[130] S. Prokudaylo. “Calculations for Neutron Spin Echo: Optimization of the Magnetic Field Geometries and Preparations and Analysis of Experiments on Crystal Lattice Dynamics”. PhD thesis. Technische Universität München, 2004.

[131] F. Mezei, C. Pappas, and T. Gutberlet. *Neutron Spin Echo Spectroscopy: Basics, Trends and Applications*. Springer Science and Business Media, 2003.

[132] J. K. Dhont et al. *Lecture notes of the 42nd IFF Springschool: Macromolecular Systems in Soft and Living Matter*. Vol. 20. Schriften des Forschungszentrum Jülich, 2011.

[133] N. Michaud-Agrawal et al. “MDAnalysis: A Toolkit for the Analysis of Molecular Dynamics Simulations”. In: *Journal of Computational Chemistry* 32.10 (2011), pp. 2319–2327.

[134] B. Lindner and J. C. Smith. “Sassena - X-ray and Neutron Scattering Calculated from Molecular Dynamics Trajectories Using Massively Parallel Computers”. In: *Computer Physics Communications* 183.7 (2012), pp. 1491 –1501.

[135] J. Hartigan. *Clustering Algorithms*. New York:Wiley, 1975.

[136] X. Hu et al. “The Dynamics of Single Protein Molecules is Non-Equilibrium and Self-Similar Over Thirteen Decades in Time”. In: *Nature Physics* 12 (2016), pp. 171–174.

[137] M. Bastian, S. Heymann, and M. Jacomy. “Gephi: An Open Source Software for Exploring and Manipulating Networks”. In: *International AAAI Conference on Weblogs and Social Media* (2009).

[138] S. M. van Dongen. “Graph Clustering by Flow Simulation.” PhD thesis. University of Utrecht, The Netherlands, 2000.

[139] C. Peter et al. “Estimating Entropies from Molecular Dynamics Simulations”. In: *The Journal of Chemical Physics* 120.6 (2004), pp. 2652–2661.

[140] G. Lipari and A. Szabo. “Model-Free Approach to the Interpretation of Nuclear Magnetic Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity”. In: *Journal of the American Chemical Society* 104.17 (1982), pp. 4546–4559.

[141] N. Popovych et al. “Dynamically Driven Protein Allostery”. In: *Nature Structural and Molecular Biology* 13.9 (2006), pp. 831–838.

[142] D. Yang and L. Kay. “Contributions to Conformational Entropy Arising from Bond Vector Fluctuations Measured from NMR-derived Order Parameters: Application to Protein Folding.” In: *Journal of Molecular Biology* 263 (1996), pp. 369–382.

[143] S. Hayward and N. Go. “Collective Variable Description of Native Protein Dynamics”. In: *Annual Review of Physical Chemistry* 46.1 (1995), pp. 223–250.

[144] S. Hayward and B. L. de Groot. *Molecular Modeling of Proteins*. Totowa, NJ: Humana Press, 2008. Chap. Normal Modes and Essential Dynamics, pp. 89–106.

[145] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. “Essential Dynamics of Proteins”. In: *Proteins: Structure, Function, and Bioinformatics* 17.4 (1993), pp. 412–425.

[146] R. Elber and M. Karplus. “Multiple Conformational States of Proteins: a Molecular Dynamics Analysis of Myoglobin”. In: *Science* 235.4786 (1987), pp. 318–321.

[147] A. Atilgan et al. “Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model”. In: *Biophysical Journal* 80.1 (2001), pp. 505 –515.

[148] A. Bakan, L. M. Meireles, and I. Bahar. “ProDy: Protein Dynamics Inferred from Theory and Experiments.” In: *Bioinformatics* 27.11 (2011), pp. 1575–1577.

[149] A. Thomas, B. Roux, and J. C. Smith. “Computer Simulations of the Flexibility of a Series of Synthetic Cyclic Peptide Analogues”. In: *Biopolymers* 33.8 (1993), pp. 1249–1270.

[150] Z. Bu et al. “Coupled Protein Domain Motion in Taq Polymerase Revealed by Neutron Spin-Echo Spectroscopy”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.49 (2005), pp. 17646–17651.

[151] R. Biehl et al. “Direct Observation of Correlated Interdomain Motion in Alcohol Dehydrogenase”. In: *Physical Review Letters* 101 (2008), p. 138102.

[152] G. Luo et al. “Single-Molecule and Ensemble Fluorescence Assays for a Functionally Important Conformational Change in T7 DNA Polymerase”. In: *Proceedings of the National Academy of Sciences* 104 (2007), pp. 12610–12615.

[153] G. D. L. Torre, J. M. L. Huertas, and B. Carrasco. “Calculation of Hydrodynamic Properties of Globular Proteins from their Atomic-Level Structure”. In: *Biophysical Journal* 78 (2000), pp. 719–730.

[154] N. Smolin et al. “Functional Domain Motions in Proteins on the 1–100 ns Timescale: Comparison of Neutron Spin-Echo Spectroscopy of Phosphoglycerate Kinase with Molecular-Dynamics Simulation”. In: *Biophysical Journal* 102 (2012), pp. 1108–1117.

[155] G. Hummer. “Position-Dependent Diffusion Coefficients and Free Energies from Bayesian Analysis of Equilibrium and Replica Molecular Dynamics Simulations”. In: *New Journal of Physics* 7.1 (2005), p. 34.

[156] A. J. Adamczyk and A. Warshel. “Converting Structural Information into an Allosteric-energy-based Picture for Elongation Factor Tu Activation by the Ribosome”. In: *Proceedings of the National Academy of Sciences* 108.24 (2011), pp. 9827–9832.

[157] D. Kavaliauskas, P. Nissen, and C. R. Knudsen. “The Busiest of All Ribosomal Assistants: Elongation Factor Tu”. In: *Biochemistry* 51.13 (2012), pp. 2642–2651.

[158] G. R. Andersen, P. Nissen, and J. Nyborg. “Elongation Factors in Protein Biosynthesis”. In: *Trends in Biochemical Sciences* 28.8 (2003), pp. 434 –441.

[159] T. Pape, W. Wintermeyer, and M. V. Rodnina. “Complete Kinetic Mechanism of Elongation Factor Tu-Dependent Binding of Aminoacyl-tRNA to the A Site of the E.coli Ribosome”. In: *The EMBO Journal* 17.24 (1998), pp. 7490–7497.

[160] E. Villa et al. “Ribosome-Induced Changes in Elongation Factor Tu Conformation Control GTP Hydrolysis”. In: *Proceedings of the National Academy of Sciences* 106.4 (2009), pp. 1063–1068.

[161] R. M. Voorhees et al. “The Mechanism for Activation of GTP Hydrolysis on the Ribosome”. In: *Science* 330.6005 (2010), pp. 835–838.

[162] N. Fischer et al. “Structure of the E. coli Ribosome-EF-Tu Complex at < 3 Å Resolution by Cs-Corrected cryo-EM”. In: *Nature* 520 (2015), pp. 567–570.

[163] H. Song et al. “Crystal Structure of Intact Elongation Factor EF-Tu from Escherichia coli in GDP Conformation at 2.05 Å Resolution”. In: *Journal of Molecular Biology* 285.3 (1999), pp. 1245 –1256.

[164] H. Šanderová et al. “Thermostability of Multidomain Proteins: Elongation Factors EF-Tu from Escherichia coli and Bacillus stearothermophilus and their Chimeric Forms”. In: *Protein Science* 13.1 (2004), pp. 89–99.

[165] V. Luigi et al. “The Crystal Structure of Sulfolobus Solfataricus Elongation Factor 1α in Complex with Magnesium and GDP”. In: *Biochemistry* 43.21 (2004), pp. 6630–6636.

[166] L. Vitagliano et al. “The Crystal Structure of Sulfolobus solfataricus Elongation Factor 1α in Complex with GDP Reveals Novel Features in Nucleotide Binding and Exchange”. In: *The EMBO Journal* 20.19 (2001), pp. 5305–5311.

[167] A. Parmeggiani et al. “Properties of a Genetically Engineered G Domain of Elongation Factor Tu”. In: *Proceedings of the National Academy of Sciences* 84 (1987), pp. 3141–3145.

[168] M. Masullo et al. “Properties of Truncated Forms of the Elongation Factor 1α from the Archaeon Sulfolobus solfataricus”. In: *European Journal of Biochemistry* 243.1-2 (1997), pp. 468–473.

[169] M. Jensen et al. “Structure-Function Relationships of Elongation Factor Tu. Isolation and Activity of the Guanine-Nucleotide-Binding Gomain.” In: *European Journal of Biochemistry* 182 (1989), pp. 247–255.

[170] H. Šanderová et al. “The N-terminal Region is Crucial for the Thermostability of the G-domain of Bacillus”. In: *Biochimica et Biophysica Acta* 1804 (2010), pp. 147–155.

[171] P. Stouten et al. “How Does the Switch II Region of G-domains Work?” In: *FEBS Letters* 320.1 (1993), pp. 1 –6.

[172] C. J. Thomas et al. “Uncoupling Conformational Change from GTP Hydrolysis in a Heterotrimeric G Protein α-Subunit”. In: *Proceedings of the National Academy of Sciences* 101.20 (2004), pp. 7560–7565.

[173] K. Abel et al. “An α to β Conformational Switch in EF-Tu”. In: *Structure* 4 (1996), pp. 1153–1159.

[174] G. Polekhina et al. “Helix Unwinding in the Effector Region of Elongation Factor EF-Tu–GDP”. In: *Structure* 4 (1996), pp. 1141–1151.

[175] T. Darden, D. York, and L. Pedersen. “Particle Mesh Ewald: An $N \cdot log(N)$ Method for Ewald Sums in Large Systems”. In: *The Journal of Chemical Physics* 98 (1993), pp. 10089–10092.

[176] K. Kulczycka, M. Długosz, and J. Trylska. “Molecular Dynamics of Ribosomal Elongation Factors G and Tu”. English. In: *European Biophysics Journal* 40.3 (2011), pp. 289–303.

[177] J. M. Bui et al. “Analysis of Sub-tauc and Supra-tauc Motions in Protein Gbeta1 Using Molecular Dynamics Simulations”. In: *Biophysical Journal* 97.9 (2009), pp. 2513–2520.

[178] B. P. English et al. “Ever-Fluctuating Single Enzyme Molecules: Michaelis-Menten Equation Revisited”. In: *Nature Structural and Molecular Biology* 2 (2006), pp. 87–94.

[179] L. Vogeley et al. “Conformational Change of Elongation Factor Tu (EF-Tu) Induced by Antibiotic Binding”. In: *The Journal of Biological Chemistry* 276 (2001), pp. 17149–17155.

[180] M. Kjeldgaard et al. “The Crystal Structure of Elongation Factor EF-Tu from Thermus aquaticus in the GTP Conformation”. In: *Structure* 1 (1993), pp. 35–50.

[181] E. Mercier, D. Girodat, and H.-J. Wieden. “A Conserved P-loop Anchor Limits the Structural Dynamics that Mediate Nucleotide Dissociation in EF-Tu”. In: *Scientific Reports* 5 (2015), p. 7677.

[182] O. Rahaman et al. “Role of Internal Water on Protein Thermal Stability: The Case of Homologous G Domains”. In: *The Journal of Physical Chemistry B* 119.29 (2015), pp. 8939–8949.

[183] K. Abel and F. Jurnak. “A Complex Profile of Protein Elongation: Translating Chemical Energy into Molecular Movement”. In: *Structure* 4.3 (1996), pp. 229 –238.

[184] A. Wittinghofer and I. R. Vetter. “Structure-Function Relationships of the G Domain, a Canonical Switch Motif”. In: *Annual Review of Biochemistry* 80 (2011), pp. 943–971.

[185] R. H. Cool and A. Parmeggiani. “Substitution of Histidine-84 and the GTPase Mechanism of Elongation Factor Tu”. In: *Biochemistry* 30.2 (1991), pp. 362–366.

[186] K.Kobayashi et al. “Structural Basis for mRNA Surveillance by Archaeal Pelota and GTP-bound EF1-α Complex”. In: *Proceedings of the National Academy of Sciences* 107 (2010), pp. 17575–17579.

[187] Z. D Nagel and J. P. Klinman. “A 21^{st} Century Revisionist’s View at a Turning Point in Enzymology”. In: *Nature Chemical Biology* 5.8 (Aug. 2009), pp. 543–550.

[188] H. N. Motlagh et al. “The Ensemble Nature of Allostery”. In: *Nature* 508.7496 (Apr. 2014), pp. 331–339.

[189] H. Frauenfelder, S. Sligar, and P. Wolynes. “The Energy Landscapes and Motions of Proteins”. In: *Science* 254.5038 (1991), pp. 1598–1603.

[190] L. Milanesi et al. “Measurement of Energy Landscape Roughness of Folded and Unfolded Proteins”. In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 19563–19568.

[191] S. J. Kerns et al. “The Energy Landscape of Adenylate Kinase During Catalysis”. In: *Nature Structural and Molecular Biology* 22.2 (Feb. 2015), pp. 124–131.

[192] N. M. Goodey and S. J. Benkovic. “Allosteric Regulation and Catalysis Emerge via a Common Route”. In: *Nature Chemical Biology* 4 (2008), pp. 474–482.

[193] B. Farago et al. “Activation of Nanoscale Allosteric Protein Domain Motion Revealed by Neutron Spin Echo Spectroscopy”. In: *Biophysical Journal* 99 (2010), pp. 3473–3482.

[194] R. Inoue et al. “Large Domain Fluctuations on 50-ns Timescale Enable Catalytic Activity in Phosphoglycerate Kinase”. In: *Biophysical Journal* 99 (2010), pp. 2309–2317.

[195] G. Schirò et al. “Translational Diffusion of Hydration Water Correlates with Functional Motions in Folded and Intrinsically Disordered Proteins”. In: *Nature Communications* 6 (2015), p. 6490.

[196] W. J. Jones et al. “Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent”. In: *Archives of Microbiology* 136.4 (1983), pp. 254–261.

[197] P. J. Fritz. “Rabbit Muscle Lactate Dehydrogenase 5; A Regulatory Enzyme”. In: *Science* 150 (1965), pp. 364–366.

[198] S. Fushinobu, T. Ohta, and H. Matsuzawa. “Homotropic Activation via the Subunit Interaction and Allosteric Symmetry Revealed on Analysis of Hybrid Enzymes of L-Lactate Dehydrogenase”. In: *The Journal of Biological Chemistry* 273 (1998), pp. 2971–2976.

[199] T. Ohta et al. “Mechanism of Allosteric Transition of Bacterial L-Lactate Dehydrogenase”. In: *Faraday Discussions* 93 (1992), pp. 153–161.

[200] S. Iwata et al. “T and R States in the Crystals of Bacterial L-Lactate Dehydrogenase Reveal the Mechanism for Allosteric Control”. In: *Nature Structural and Molecular Biology* 1 (1994), pp. 176–185.

[201] N. Coquelle et al. “Activity, Stability and Structural Studies of Lactate Dehydrogenases Adapted to Extreme Thermal Environments”. In: *Journal of Molecular Biology* 374.2 (2007), pp. 547–562.

[202] J. Colletier et al. “Sampling the Conformational Energy Landscape of a Hyperthermophilic Protein by Engineering Key Substitutions.” In: *Molecular Biology and Evolution* 29 (2012), pp. 1683–1694.

[203] A. R. Clarke et al. “The Rates of Defined Changes in Protein Structure During the Catalytic Cycle of Lactate Dehydrogenase”. In: *Biochimica et Biophysica Acta* 829 (1985), pp. 397–407.

[204] C. R. Goward and D. J. Nicholls. “Malate Dehydrogenase: A Model for Structure, Evolution, and Catalysis”. In: *Protein Science* 3.10 (1994), pp. 1883–1888.

[205] H.-L. Peng et al. “Energy Landscape of the Michaelis Complex of Lactate Dehydrogenase: Relationship to Catalytic Mechanism”. In: *Biochemistry* 53.11 (2014), pp. 1849–1857.

[206] H.-L. Peng et al. “Mechanism of Thermal Adaptation in the Lactate Dehydrogenases”. In: *The Journal of Physical Chemistry B* 119.49 (2015), pp. 15256–15262.

[207] S. Ferrer et al. “A Theoretical Analysis of Rate Constants and Kinetic Isotope Effects Corresponding to Different Reactant Valleys in Lactate Dehydrogenase”. In: *Journal of the American Chemical Society* 128.51 (2006), pp. 16851–16863.

[208] F. Mezei. *Neutron Spin Echo*. Vol. 128. Berlin, Heidelberg, New York: Springer, 1980.

[209] A. E. García. “Large-Amplitude Nonlinear Motions in Proteins”. In: *Physical Review Letters* 62 (1992), pp. 2696–2699.

[210] G. Nägele. “On the Dynamics and Structure of Charge-Stabilized Suspensions”. In: *Physics Reports* 272 (1996), p. 215.

[211] J. Gapinski et al. “Diffusion and Microstructural Properties of Solutions of Charged Nanosized Proteins: Experiment Versus Theory”. In: *The Journal of Chemical Physics* 123 (2005), p. 054708.

[212] E. Eyal, L.-W. Yang, and I. Bahar. “Anisotropic Network Model: Systematic Evaluation and a New Web Interface”. In: *Bioinformatics* 22 (2006), pp. 2619–2627.

[213] K. Takemura and A. Kitao. “Effect of Water Model and Simulation Box Size on Protein Diffusional Motions”. In: *The Journal of Physical Chemistry B* 111 (2007), pp. 11870–11872.

[214] A. L. Jacobson and H. Braun. “Differential Scanning Calorimetry of the Thermal Denaturation of Lactate Dehydrogenase.” In: *Biochimica et Biophysica Acta* 493 (1977), pp. 142–153.

[215] L. Qiu, M. Gulotta, and R. Callender. “Lactate Dehydrogenase Undergoes a Substantial Structural Change to Bind its Substrate”. In: *Biophysical Journal* 93 (2007), pp. 1677–1686.

[216] L. D. Unsworth, J. van der Oost, and S. Koutsopoulos. “Hyperthermophilic Enzymes - Stability, Activity, and Implementation Strategies for High Temperature Applications”. In: *FEBS Journal* 274.16 (2007), pp. 4044–4056.

[217] P. Turner, G. Mamo, and E. N. Karlsson. “Potential and Utilization of Thermophiles and Thermostable Enzymes in Biorefining”. In: *Microbial Cell Factories* 6.1 (2007), pp. 1–23.

[218] A. Korkegian et al. “Computational Thermostabilization of an Enzyme”. In: *Science* 308.5723 (2005), pp. 857–860.

[219] E. Bae, R. M. Bannen, and G. N. Phillips. “Bioinformatic Method for Protein Thermal Stabilization by Structural Entropy Optimization”. In: *Proceedings of the National Academy of Sciences* 105.28 (2008), pp. 9594–9597.

[220] G. Vogt, S. Woell, and P. Argos. “Protein Thermal Stability, Hydrogen Bonds, and Ion Pairs”. In: *Journal of Molecular Biology* 269.4 (1997), pp. 631–643.

[221] M. J. Pikal et al. “Solid State Chemistry of Proteins: II. The Correlation of Storage Stability of Freeze-Dried Human Growth Hormone (hGH) with Structure and Dynamics in the Glassy Solid”. In: *Journal of Pharmaceutical Sciences* 97.12 (2008), pp. 5106–5121.

[222] J. J. Hill, E. Y. Shalaev, and G. Zografi. “Thermodynamic and Dynamic Factors Involved in the Stability of Native Protein Structure in Amorphous Solids in Relation to Levels of Hydration”. In: *Journal of Pharmaceutical Sciences* 94.8 (2005), pp. 1636–1667.

[223] C. P. Schneider, D. Shukla, and B. L. Trout. “Effects of Solute-Solute Interactions on Protein Stability Studied Using Various Counterions and Dendrimers”. In: *PLOS ONE* 6.11 (2011), e27665.

[224] D. Shukla, C. P. Schneider, and B. L. Trout. “Complex Interactions Between Molecular Ions in Solution and Their Effect on Protein Stability”. In: *Journal of the American Chemical Society* 133.46 (2011), pp. 18713–18718.

[225] J. M. Vinther, S. M. Kristensen, and J. J. Led. “Enhanced Stability of a Protein with Increasing Temperature”. In: *Journal of the American Chemical Society* 133.2 (2010), pp. 271–278.

[226] K. E. Tang and K. A. Dill. “Native Protein Fluctuations: the Conformational-Motion Temperature and the Inverse Correlation of Protein Flexibility with Protein Stability”. In: *Journal of Biomolecular Structure and Dynamics* 16.2 (1998), pp. 397–411.

[227] A. M. Tsai, T. J. Udovic, and D. A. Neumann. “The Inverse Relationship between Protein Dynamics and Thermal Stability”. In: *Biophysical Journal* 81.4 (2001), pp. 2339–2343.

[228] A. De Francesco et al. “Picosecond Internal Dynamics of Lysozyme as Affected by Thermal Unfolding in Nonaqueous Environment”. In: *Biophysical Journal* 86.1 (2004), pp. 480–487.

[229] D. Russo et al. “Dynamic Transition Associated with the Thermal Denaturation of a Small Beta Protein”. In: *Biophysical journal* 83.5 (2002), pp. 2792–2800.

[230] G. Zaccai. “How Soft is a Protein? A Protein Dynamics Force Constant Measured by Neutron Scattering”. In: *Science* 288.5471 (2000), pp. 1604–1607.

[231] H. X. Zhou, G. Rivas, and A. P. Minton. “Macromolecular Crowding and Confinement: Biochemical, Biophysical, and Potential Physiological Consequences”. In: *Annual Review of Biophysics* 37 (2008), pp. 375–97.

[232] B. van den Berg, R. J. Ellis, and C. M. Dobson. “Effects of Macromolecular Crowding on Protein Folding and Aggregation”. In: *The EMBO journal* 18.24 (1999), pp. 6927–6933.

[233] J. Li, S. Zhang, and C.-c. Wang. “Effects of Macromolecular Crowding on the Refolding of Glucose-6-phosphate Dehydrogenase and Protein Disulfide Isomerase”. In: *Journal of Biological Chemistry* 276.37 (2001), pp. 34396–34401.

[234] A. H. Elcock. “Atomic-Level Observation of Macromolecular Crowding Effects: Escape of a Protein from the GroEL Cage”. In: *Proceedings of the National Academy of Sciences* 100.5 (2003), pp. 2340–2344.

[235] A. C. Miklos et al. “Protein Crowding Tunes Protein Stability”. In: *Journal of the American Chemical Society* 133.18 (2011), pp. 7116–7120.

[236] Y. Wang et al. “Macromolecular Crowding and Protein Stability”. In: *Journal of the American Chemical Society* 134.40 (2012), pp. 16614–16618.

[237] B.-R. Zhou et al. “Mixed Macromolecular Crowding Accelerates the Oxidative Refolding of Reduced, Denatured Lysozyme: Implications for Protein Folding in Intracellular Environments”. In: *Journal of Biological Chemistry* 279.53 (2004), pp. 55109–55116.

[238] J. J. Gilvarry. “The Lindemann and Grüneisen Laws”. In: *Physical Review* 102.2 (1956), p. 308.

[239] F. A. Lindemann. “Über die Berechnung Molekularer Eigenfrequenzen”. In: *Physikalische Zeitschrift* 11 (1910), pp. 609–612.

[240] J. Maddox. “Calculating Melting Temperature”. In: *Nature* 323 (1986).

[241] Y. Fujita and Y. Noda. “Effect of Hydration on the Thermal Denaturation of Lysozyme as Measured by Differential Scanning Calorimetry”. In: *Bulletin of the Chemical Society of Japan* 51.5 (1978), pp. 1567–1568.

[242] S. W. Lovesey. *Theory of Neutron Scattering from Condensed Matter*. Oxford University Press, 1988.

[243] C. E. Kundrot and F. M. Richards. “Crystal Structure of Hen Egg-White Lysozyme at a Hydrostatic Pressure of 1000 Atmospheres”. In: *Journal of Molecular Biology* 193.1 (1987), pp. 157–170.

[244] A. Paciaroni, S. Cinelli, and G. Onori. “Effect of the Environment on the Protein Dynamical Transition: a Neutron Scattering Study”. In: *Biophysical Journal* 83.2 (2002), pp. 1157–1164.

[245] A. Paciaroni et al. “Fast Fluctuations in Protein Powders: the Role of Hydration”. In: *Chemical Physics Letters* 410.4 (2005), pp. 400–403.

[246] E. Cornicchi, G. Onori, and A. Paciaroni. “Picosecond-Time-Scale Fluctuations of Proteins in Glassy Matrices: the Role of Viscosity”. In: *Physical Review Letters* 95.15 (2005), p. 158104.

[247] E. Cornicchi et al. “Controlling the Protein Dynamical Transition with Sugar-Based Bioprotectant Matrices: a Neutron Scattering Study”. In: *Biophysical Journal* 91.1 (2006), pp. 289–297.

[248] W. Doster and M. Settles. “Protein–Water Displacement Distributions”. In: *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1749.2 (2005), pp. 173–186.

[249] J. Roh et al. “Onsets of Anharmonicity in Protein Dynamics”. In: *Physical Review Letters* 95.3 (2005), p. 038101.

[250] G. Schiro et al. “Direct Evidence of the Amino Acid Side Chain and Backbone Contributions to Protein Anharmonicity”. In: *Journal of the American Chemical Society* 132.4 (2010), pp. 1371–1376.

[251] F. H. Stillinger and D. K. Stillinger. “Computational Study of Transition Dynamics in 55-Atom Clusters”. In: *The Journal of Chemical Physics* 93.8 (1990), pp. 6013–6024.

[252] S. E. Jackson and A. R. Fersht. “Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition”. In: *Biochemistry* 30.43 (1991), pp. 10428–10435.

[253] R. Zwanzig. “Two-State Models of Protein Folding Kinetics”. In: *Proceedings of the National Academy of Sciences of the United States of America* (1997), pp. 148–150.

[254] S. Hawley. “Reversible Pressure-Temperature Denaturation of Chymotrypsinogen”. In: *Biochemistry* 10.13 (1971), pp. 2436–2442.

[255] L. N. Bell, M. J. Hageman, and L. M. Muraoka. “Thermally Induced Denaturation of Lyophilized Bovine Somatotropin and Lysozyme as Impacted by Moisture and Excipients”. In: *Journal of Pharmaceutical Sciences* 84.6 (1995), pp. 707–712.

[256] T. Knubovets et al. “Structure, Thermostability, and Conformational Flexibility of Hen Egg-White Lysozyme Dissolved in Glycerol”. In: *Proceedings of the National Academy of Sciences* 96.4 (1999), pp. 1262–1267.

[257] Y. Zhou, D. Vitkup, and M. Karplus. “Native Proteins are Surface-Molten Solids: Application of the Lindemann Criterion for the Solid Versus Liquid State”. In: *Journal of Molecular Biology* 285.4 (1999), pp. 1371–1375.

[258] R. Diamond. “Real-Space Refinement of the Structure of Hen Egg-White Lysozyme”. In: *Journal of Molecular Biology* 82.3 (1974), 371IN5375–374IN11391.

[259] J. H. Bilgram. “Dynamics at the Solid-Liquid Transition: Experiments at the Freezing Point”. In: *Physics Reports* 153.1 (1987), pp. 1–89.

[260] H. Löwen. “Melting, Freezing and Colloidal Suspensions”. In: *Physics Reports* 237.5 (1994), pp. 249–324.

[261] F. H. Stillinger. “A Topographic View of Supercooled Liquids and Glass Formation”. In: *Science* 267.5206 (1995), p. 1935.