ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

SCUOLA DI SCIENZE

Corso di Laurea in Applied Physics

Aggregation of the C-terminal fragment of the TAR DNA-binding Protein 43 in relation to the Amyotrophic lateral sclerosis

Advisors:
Prof. Armando Bazzani[1]
Prof. Giancarlo Ruocco[23]

Co-advisor:
Dr. Edoardo Milanetti[23]

Candidate:
Greta Grassmann

Sessione Autunnale
Anno Accademico 2020/21

[1] Department of Physics and Astronomy, University of Bologna, Viale Carlo Berti Pichat 6/2, 40127 Bologna, Italy

[2] Center for Life Nano-science & Neuro-Science, Fondazione Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Rome, Italy

[3] Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185 Rome, Italy

Aggregation of the C-terminal fragment of the TAR DNA-binding Protein 43 in relation to the Amyotrophic lateral sclerosis

Master's thesis. Bologna - University of Bologna

Author's email: gretagrassmann0@gmail.com

Abstract

Many neurodegenerative diseases -including the Amyotrophic Lateral Sclerosis (ALS)- are associated with the presence of protein aggregates. Over 97% of ALS cases feature pathological inclusions that are mainly composed by the human TAR DNA-binding Protein 43 (TDP-43) and affect the cortical and spinal neurons. This thesis studies the aggregation process of the two types of TDP-43 C-terminal fragments (CTFs) -corresponding to different cleavages of the full protein- that can be found in the former. This is interesting not only for the possible implications on the ALS disease, but also because TDP-43 fragments are a useful system model for protein aggregation. The interaction model proposed in this work starts from a cross-β spine model, which hypothesizes that the CTFs' RRM2 fragment is at the core of the aggregation, thanks to the exposition of the aggregation prone β-strands following the proteolysis. By designing specific interfering aptamers which can bind to the RRM2 binding regions, we should be able to prevent another CTF to bind to that site. Still, the structures of these fragments have not been deeply studied yet and their conformations are not available, since their high aggregation propensity makes it difficult to perform an experimental investigation. To propose some possible binding regions, we analyze the RRM2 fragments trajectories resulting from Molecular Dynamics (MD) simulations. By applying a cluster analysis on the two-principal components projections of these trajectories, we find the fragments' equilibrium conformations. Next, we verify the shape complementarity between the 3D molecular surfaces of these equilibrium configurations by means of the **2D *Zernike* polynomial expansion**. Among these ***Zernike*** selected binding regions, we select the ones that would be able to bind an aptamer, i.e. the ones with a positive surface charge. Proposing these binding regions is the final step of this thesis, but not of our work. Starting from these results, we plan to perform additional studies. For example, we will verify our conclusions with Brillouin microscopy: if the suggested binding regions are really at the core of the aggregation, following the insertion of expressively designed aptamers in CTFs-expressing cells, the number and dimension of aggregates will decrease.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many of the molecular mechanisms underlying the pathological aggregation of proteins observed in neurodegenerative diseases are still not fully understood. Among the diseases associated with protein aggregates, the Amyotrophic Lateral Sclerosis (ALS) is of relevant importance. ALS, introduced in the biological overview of Section 2, is a neurodegenerative disease specifically affecting cortical and spinal motor neurons. Although understanding the primary causes of the disease is still an open challenge, its relationship with protein aggregation is widely known. The human TAR DNA-binding Protein 43 (TDP-43), a RNA/DNA binding protein involved in RNA-related metabolism, is a major component of these pathological inclusions [1, 2].

While the deposition of the phosphorylated full-length TDP-43 in spinal-cord cells has been widely studied, it has been shown that the brain cortex presents accumulation of phosphorylated C-terminal fragments (CTFs) [3–7]. The two kind of CTFs that we study correspond to a portion of the full protein including only the last 195 or 206 residues respectively. In this thesis, we want to investigate the CTFs aggregation process. Even if it is debated whether CTFs represent a primary cause of ALS, they are a hallmark of TDP-43 related neurodegeneration in the brain [8]. The analysis of the CTFs could have important implications not only because of their biological role in the cell, but also in relation to the study of the interaction between proteins: a fragment of TDP-43 is a useful system model for protein aggregation, since its small dimension allows us explore the conformational space with high efficiency and non prohibitive computational times.

Here we provide a possible computational model for the molecular interactions, based on extensive Molecular Dynamics (MD) simulations performed with GROMACS [9] to explore the conformational space, and on the evaluation of shape complementarity between the exposed regions of the different sampled conformations. We start our project from what is known in literature to date: CTFs are composed by the disordered C-terminal domain (CTD) and a fragment of RRM2, a folded domain of known structure. The latter could be of fundamental importance for the protein's aggregation, since after the TDP-43 proteolysis it partially misfolds and exposes the aggregation prone β-strands. These β-strands could be at the core of the aggregation, since they are able to give rise to amyloid structures [5, 10]. Since the RRM2 domain is ordered and structured, it is possible to investigate the shape of its molecular surface, and consequently the complementarity between the shapes of different fragments' surfaces.

Aim of this work is to suggest some possible binding regions on the RRM2 fragments that in the future could be taken as starting point for designing specific interfering molecules. This is achieved in three steps:

1. **Molecular Dynamic (MD) simulations** for the two RRM2 fragments (corresponding to a cleavage at two different sites) that can be found in CTFs, with the aim of exploring their equilibrium conformations. To perform a complete study of these regions of TDP-43, we employ MD simulations to study the evolution of the whole RRM2 as well. All simulations are carried on for 10 μs. A theoretical introduction to these topics is given in Sections 3 and 4, whereas in Section 5 we discuss the application of these methods in our case. In Section 7 we present the results of the MD simulations we performed, and in Appendix A we analyse in more details these simulations (in particular their minimization and equilibration phases). To obtain the fragments' equilibrium conformations, we firstly apply a Principal Component Analysis (PCA) on the trajectory resulting from each MD simulation. In this way, we get an essential representation of the dynamics. Then, we implement a cluster analysis on the projection of each trajectory on its first two principal components. Our aim is to find the most representative conformations for each one of the possible conformations that the fragment can take at equilibrium. We are assuming that each cluster's center (or centroid) is a good representative

of that cluster: the structures corresponding to these centroids are the equilibrium conformations. In Appendix B we discuss in more details how the Principal Component (PC) and K-means clustering analyses are implemented.

2. We then sample the set of exposed portions of the 3D molecular surface of these equilibrium conformations to find complementary regions between the molecular surfaces. With this aim, we adopt a newly developed approach based on ***Zernike*** polynomials and presented in Section 6. The **2D *Zernike* polynomial expansion** is a new method [11] (developed at the *Center for Life Nano-science & Neuro-Science, Fondazione Istituto Italiano di Tecnologia*[1] in 2020) for assessing whether and where two proteins can interact with each other to form a complex. In our case we are going to apply it, in Section 7, to the 3D structures obtained with the MD simulations of the two fragments of RRM2.

3. Among these ***Zernike***-selected regions, we identify the ones that could be at the core of the CTFs aggregation, and propose them as candidate **binding regions**. We propose as well a set of binding regions that could be able to bind to specifically designed aptamers. Aptamers are short oligonucleotide or peptide molecules selected via in vitro evolution to bind, with high affinity and selectivity, to a target molecule of interest, including proteins, peptides, and carbohydrates. They can be potentially used in diagnostic and therapeutic applications or as molecular sensors.

Aptamer-binding regions are proposed since in the future we will test our results by inserting these **aptamers** in cells expressing CTFs aggregates: if our predictions are correct, after the aptamers insertion the number and dimension of the aggregates will diminish.

In the future, we plan to test this variation with experimental measurements on cells in vitro, employing **Brillouin microscopy** [12]. Brillouin microscopy can probe the viscoelastic properties of biological samples: since the aggregates are characterised by a more solid consistency compared to the surrounding cytoplasm, they are clearly visible with such a tool. In Section 8 we discuss the design of these aptamers and the Brillouin measurements, together with all the other studies that we would like to implement in

[1] Viale Regina Elena 291, 00161 Rome, Italy.

the future. As an additional future study, we will deepen the analyses of the MD simulations trajectories with our newly developed computational method for the minimal representation of a surface [13], that will allow us to analyze a higher number of MD simulations' frames.

Chapter 2

Biological overview

2.1 TDP-43 in relation to the ALS

The Amyotrophic Lateral Sclerosis (ALS) is a fatal neurodegenerative disorder that is typically adult-onset, and is characterized by progressive loss of upper motor neurons in the motor cortex and corticospinal tract, and lower motor neurons in the spinal cord [14]. This leads to denervation and rapid atrophy of specific muscle groups, which usually eventuates in death by respiratory failure [15].

Over 97% of ALS cases, both sporadic and familial, feature TDP-43-positive inclusions in the cytoplasm of affected neurons [6, 7, 16].

TAR DNA-binding protein 43 (TDP-43) is a nuclear factor that regulates transcription, pre-mRNA splicing and processing, regulation of translation, and RNA stability [3]. It consists of 414 amino acid residues, divided in 4 domains:

- The **N-terminal domain (NTD)**. It has a well-defined fold, and has been shown to form dimer or oligomer in physiological conditions. It contains the Nuclear Localization Signal[1] (NLS).
- The **RNA recognition domain 1 (RRM1)**, spanning residues 106-176. It is a folded RNA recognition motif.
- The **RNA recognition domain 2 (RRM2)**. It stabilizes RRM1 and works

[1] Amino acid sequence that tags a protein for import into the cell nucleus by nuclear transport.

together with it as a RNA recognition motif [17, 18]. Its structure is stable and comprises two α-helices and five β strands assembled in a β sheet, according to a β_1-α_1-β_2-β_3-α_2-β_4-β_5 topology [18].
The RRM2 contains the Nuclear Export Signal[2] (NES).

- The **C-terminal domain (CTD)**. It is unstructured and contains a glycine-rich region. The CTD is involved in protein-protein interactions, is aggregation prone and harbors most of the mutations associated with familial ALS. It also possesses high propensity to phase separate [4].

The human TDP-43 is localized in healthy cells mainly in the nucleus [3, 7], where it forms dimer or oligomer via its NTD [3, 5]. These head-to-tail TDP-43 oligomers represent the functional form of the protein in vivo, and their destabilization results in loss of alternative splicing regulation of known neuronal RNA targets [4].
During neurodegenerative disease, TDP-43 undergoes a vast array of post-translational modifications, including phosphorylation, acetylation, and cleavage [5, 7]. The deposition of the phosphorylated full-length TDP-43 is primarily located in the spinal-cord cells [7].
Nevertheless, the inclusions can be formed not only by the full-length TDP-43, but by the C-terminal fragments (CTFs) that results from its cleavage as well [2–7].
The CTFs aggregates can be mainly found in the brain cortex [6, 7] and are rarely observed in the spinal cord, even if ALS involves dramatic degeneration of spinal motor neurons. Moreover, despite forming disease reminiscent inclusions, TDP-43 CTFs typically do not confer a toxic gain of function, leaving markers of cytotoxicity and apoptosis unaltered. Therefore, they are described as a neuropathological signature of these diseases [8]. However, there is some evidence that they disrupt RNA splicing by TDP-43, because of the loss of the NTD [3, 4, 8].
The aggregation of the CTFs would seem to start from the disruption of the physiological oligomerization of TDP-43 [5]. The NTD-driven head-to-tail oligomerization indeed spatially separates the highly aggregation prone CTDs of consecutive TDP-43 monomers, antagonizing cytoplasmic aggregation [3, 4, 7]. But if a preoteolytic cleavage releases the

[2] Amino acid sequence that tags a protein for export from the cell nucleus to the cytoplasm by nuclear transport.

CTD, together with a truncated RRM2 fragment, these free portions of the protein are free to aggregate [3, 5].
In addition to this, the removal of the NTD increases the cytoplasmic localization, since it deprives the resulting CTF of the NLS.

2.2 C-terminal fragments

The CTFs of TDP-43 correspond to only the last 194 or 206 residues [5] of the full protein and can be obtained from two different cleavages, at site 219 or 208 respectively [19].
In normal conditions, TDP-43 may be cleaved into smaller fragments before being enzymatically degraded to maintain physiological levels [20, 21]. TDP-43 is processed by a range of cysteine proteases, including caspases and calpains. To explain the generation of CTFs in TDP-43 proteinopathies, it has been hypothesised that disease-related factors such as cell stress and genetic mutations may modulate the activity of these enzymes [7, 8]. A second hypothesis is that CTFs may also arise from alterations at the transcriptional level [8].

2.3 CTFs aggregation model

While it is already known that the CTD is aggregation-prone, the RRM2 fragment of the CTFs could be of fundamental importance for the aggregation [22, 23] as well. The truncated RRM2 fragments are prone to aggregation because of the absence of the RRM1 domain, to which RRM2 is connected in the full TDP-43 protein. This absence causes the loss of the stabilizing interaction between the two, together with the exposition of the RRM2 normally buried β-strands [5, 17, 22] after the TDP-43 proteolysis and the consequent RRM2 partial unfolding.
These β-strands have been found to form fibrils in vitro [6]. This means that they could be at the core of the aggregation, because they are able to form steric zippers between different CTFs that then, following a typical atomic model for amyloid fibril structure

[10] formation, give rise to amyloid structures [5].

Amyloid fibrils consist of packed β-sheets that run parallel to the fibril axes. Each β-sheet adheres to its neighboring sheet through the side chains that project roughly perpendicular to the fibril axis, toward the neighboring sheet. This interdigitation between the side chains of mating sheets is the so-called steric zipper.

In support of this hypothesised aggregation model, it has been confirmed that some regions of RRM2 can form different classes of steric zipper structures [7, 19].

This model has already been introduced in [5], but we are going to deepen the investigation of its structure via computational tools. Figure 2.1 shows a schematic representation of the proposed process at the base of the TDP-43 CTFs aggregation. It depicts the cleavage that disrupts the physiological TDP-43 oligomerization and the subsequent aggregation of the resulting fragments, composed by the CTD and a RRM2 fragment. Figure 2.1 shows the β strands within RRM2 prone to fibril formation forming two-dimensional sheet-like fibrils. But the truncated RRM2 are packed into long three-dimensional fibril bundles, indicating that not only the aggregation-prone segments are important, but

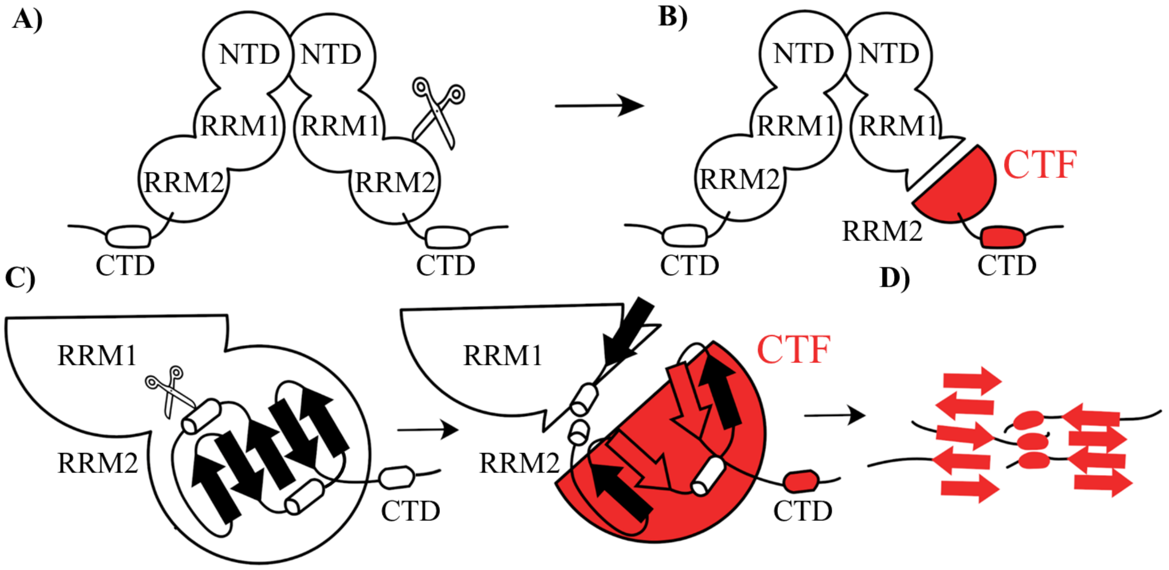

Figure 2.1: **Hypothesised model for the TDP-43 CTFs aggregation.**
A) TDP-43 in physiological conditions forms dimers. **B)** After the cleavage the CTF is split from the whole protein. **C)** The RRM2 fragment resulting from the cleavage exposes its β-strands. **D)** The β-strands from different CTFs allow the formation of aggregates to happen.

also the overall three-dimensional structure of RRM2 may be critical for the formation of large filaments [5].
The importance of the 3D structure plays a key role both in the MD simulation approach and in the ***Zernike*** polynomials based method for analysing the shape complementarity (for more details see Sections 7.1 and 7.2 respectively).

2.3.1 The RRM2 fragment's role in aggregation

In physiological conditions RRM2 is a really stable domain, thanks to a cluster of twelve connected hydrophobic residues in its core [6]. It plays a role in aggregation after its cleavage and separation from the RRM1 domain, which result in the misfolding of RRM2 and, as a second step, in its aggregation.
Indeed, in a study of the RRM2 unfolding model [17], it has been found that the mutually stabilizing interaction between RRM1 and RRM2 reduces the population of an intermediate state of RRM2 linked with pathological misfolding. This intermediate state may enhance the access to the NES contained within its sequence and serve as a molecular hazard linking physiological folding with pathological misfolding and aggregation. Consequently, isolating or fragmenting the RRM2 removes this stabilizing contribution from RRM1 and allows this region of TDP-43 to sample a potentially pathogenic folded state that increases the transport to the cytoplasm and exposes the hydrophobic residues and aggregation prone peptides of RRM2.
This is in accordance with the *conformational selection model* [24]: a protein is a collection of coexisting conformation with different population distributions. Each one of these conformations can selectively bind the most suitable partners. However, according to this model the bound conformations are sampled by the protein even when it is not bounded to a partner. In other words, the conformational change of a protein can occur before a binding event, rather than being induced by the event itself [25]. The assumption of the validity of this model is at the core of our study of the CTFs aggregation, since we are indeed looking at the conformations sampled by each single and free fragment.
This model also suggest that the right partner might act as a "molecular chaperon" by stabilizing a non-pathological state: among the conformations of the dynamically fluctuating protein, this partner selects the one compatible with binding, and shifts the

conformational ensemble towards this state [26].

2.4 Unveiling the self-assembly of the C-terminal fragments

The aggregation of TDP-43 is strongly influenced by the interaction with DNA and RNA: RNA aptamers are able to interfere with the aggregation kinetic, as a function of their nucleotides composition, binding affinity and length [24].

Assuming the validity of the *cross-β spine* model for the CTFs aggregation, by binding an aptamer to the RRM2 site that forms the "spine" of the fibril, we should be able to prevent another CTF to bind to that site. This is indeed an almost mandatory choice since the RRM2 fragment is the only part of the CTFs that we can control (with both MD simulations and the ***Zernike*** method) because of the CTD disordered structure: such a disordered structure does not have an equilibrium conformation that can be selected as the most representative one for the study of the binding.

The designing of this interfering molecule should obviously consider the binding compatibility to the misfolded conformation of the RRM2 fragments, which is not available yet in literature.

Because of this, we use MD simulations to study the conformations of the fragments after the cut.

Chapter 3

General Molecular Dynamics approach

Molecular Dynamics (MD) simulations are a technique for computing the equilibrium and transport properties, that can be applied to classical many-body system to perform measurements of their physical observables. For a basic implementation of MD simulations, we have to select a model system consisting of N particles, set the initial conditions at a time t_0 and then we integrate the Newton's equations of motion for this system. These equations are give by

$$m_i \frac{d^2 \mathbf{r}_i(t)}{dt^2} = \mathbf{F}_i, \quad i = 1, ..., N, \tag{3.1}$$

where m_i and $\mathbf{r}_i$ are the ith particle mass and position respectively. $\mathbf{F}_i$ is the force experienced by the ith particle, and its expression is fixed by the assumption of a force field $V_i(\mathbf{r}_1, ..., \mathbf{r}_N)$ acting on the ith particle and calculated by its classical expression:

$$\mathbf{F}_i = -\frac{\partial V_i(\mathbf{r}_1, ..., \mathbf{r}_N)}{\partial \mathbf{r}_i} \tag{3.2}$$

MD simulations solve these equations simultaneously in small time steps: the system is followed in this way for some time, and the coordinates are written to an output file at regular intervals, so that we are able to know dynamical variables such as the position and the velocity coordinates for each particle at each step of the integration. These variables are necessary to measure an observable, which to be calculated must first be expresses as a function of them. As an example we can look at the operative definition of temperature. Assuming that the equipartition theorem for the average kinetic energy per degree of freedom for a system in equilibrium at temperature T holds, we can write

$$< \frac{1}{2} m_i v_i^2 >= \frac{1}{2} k_B T, \tag{3.3}$$

from which

$$T(t) = \sum_{i=1}^{N} \frac{< m_i v_i^2(t) >}{k_B N}. \tag{3.4}$$

The brackets <> indicate a statistical ensemble average of quantities.
Despite the complexity of solving 6N non linear differential equations, thanks to this technique we can achieve a precision when determining positions and velocities that is not accessible in real experiments.

Despite being a successful and commonly used technique, the MD simulation may fail if a starting conformation is very far from the equilibrium state in typical experimental conditions, since in this case the forces may be excessively large. In such a situation, a robust energy minimization (for more details see Section 4.1.2) is required. Another reason to perform an energy minimization is the removal of all the kinetic energy from the system: if several snapshots from dynamic simulations must be compared, energy minimization reduces the thermal noise in the structures and potential energies so that they can be compared better.

In addition to the energy minimization, there is a set of necessary conditions for the implementation of this method:

- The initial conditions on positions and velocities must be given.
- The expression of the Hamiltonian H describing the system whose potential is used to calculate forces must be known as well.
- The particles positions and momenta at each time step have to be integrated and updated.

To choose the best algorithm to integrate Newton's equation of motion, we must take into account some considerations:

- Its speed is not fundamental, because the fraction of time spent on integrating the equations of motion is small compared to the computation of the interactions.

- We must give a bigger importance to the large time step accuracy, because the longer the time step that we can use, the fewer evaluations of the forces are needed. Algorithms that allow the use of a large time step are based on the storing of information on increasingly higher-order derivatives of the particles coordinates.

- Another important criterion is energy conservation, that we can divide in two kind: short time and long time energy conservation. The higher-order algorithms allow bigger time steps, while on the other hand tend to have a good energy conservation for short time but overall energy drifts for long times.
  On the contrary, Verlet-style algorithms tend to have moderate short-term energy conservation but little long-term drift.

- None of these algorithms can predict accurately particles' trajectories for both long and short times. This is because, usually, the systems studied with MD simulations are in a regime whose trajectory thorough the phase space depends strongly on the initial conditions: two trajectories that are initially close will diverge exponentially as time progresses. The integration error of the algorithm causes the initial small difference between the true trajectory and the one generated by the simulation, and as a consequence an exponential divergence between them. This is the so-called Lyapunov instability.
  Nevertheless, these inaccurate trajectories can be used because considerable numerical evidence [27] suggests the existence of the so-called shadow-orbits. A shadow orbit is a true trajectory of a many-bodies system that closely follows the numerical trajectory for a time that is long compared to the time it takes the Lyapunov instability to develop. In other words, the results of the simulation are representative of a true trajectory in the phase space, even though we cannot tell a priori which.

- Another requirement we must check for when choosing an integration algorithm, is time reversibility: since Newton's equations of motion are time reversible, so should be the algorithm. However, even when considering a time-reversible algorithm, the numerical implementation will not be truly time-reversible, because of the computer's finite machine precision.

- Many numerical schemes, especially the ones that are not time reversible, differ in another crucial aspect from Hamilton's equation of motion: the area-preserving property, in that they define a dynamic that changes the magnitude of any volume element in the phase space. The expansion of the system in the phase space is not compatible with energy conservation: non-reversible algorithms will have long-term energy drift problems.

With respect to these considerations, the most simple and best performing algorithms used to integrate the equation of motions are the Verlet-like ones. Verlet-like algorithms are a good choice for most MD applications, because higher-order schemes require more storage and are often neither reversible nor area preserving.
In particular the Leap-Frog algorithm is the default integration algorithm for GROMACS [9] MD simulations (as discussed in Section 4.1).

Verlet

The Verlet algorithm is fast and requires little memory but, since it is not particularly accurate for long time steps, needs to compute frequently the forces. On the other hand, it has a fair short-term energy conservation and a little long-term energy drift: this is related to the fact that the Verlet algorithm is time-reversible and area preserving. It does not conserve the precise total energy of the system, but it does conserve a pseudo-Hamiltonian approaching the true Hamiltonian, in the limit of infinitely short steps. It does not generate really accurate trajectories , but no algorithm is good enough to keep the trajectories close to the true ones for a time comparable to the duration of a typical MD simulation: a better algorithm would at best postpone the unavoidable exponential growth of the trajectories' errors by a few hundred time steps.

The derivation of the Verlet algorithm start from a Taylor expansion of the coordinate of a particle $r(t)$ around time t, for a subsequent and a preceding interval:

$$\begin{aligned} r(t+\Delta t) &= r(t) + v(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 + \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4) \\ r(t-\Delta t) &= r(t) - v(t)\Delta t + \frac{F(t)}{2m}\Delta t^2 - \frac{\Delta t^3}{3!}\dddot{r} + O(\Delta t^4). \end{aligned} \tag{3.5}$$

The sum of these two equations yields:

$$r(t+\Delta t) + r(t-\Delta t) = 2r(t) + \frac{F(t)}{m}\Delta t^2 + O(\Delta t^4). \tag{3.6}$$

This Equation leads to the Verlet position integrator:

$$r(t+\Delta t) = 2r(t) - r(t-\Delta t) + \frac{F(t)}{m}\Delta t^2 + O(\Delta t^4). \tag{3.7}$$

We can see how the Verlet algorithm does not use the velocity to compute new positions. Nevertheless, their knowledge is essential in order to perform measurements on macroscopic quantities.

The velocity can be derived from the knowledge of the trajectory, by performing the same Taylor expansion for $r(t+\Delta t)$ and $r(t-\Delta t)$, only up to the second order, and subtracting them:

$$r(t+\Delta t) - r(t-\Delta t) = 2v(t)\Delta t + O(\Delta t^3), \tag{3.8}$$

so that

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\Delta t} + O(\Delta t^2). \tag{3.9}$$

Equation 3.9 is accurate only to the order of Δt^2; in our simulation we will use the Leap-Frog integrator, which is an extension of the Verlet algorithm that performs more accurate estimates.

Leapfrog

Several algorithms are equivalent to the Verlet scheme, and the Leap-Frog algorithm is the simplest one.
To obtain the Leapfrog velocity integrator we rewrite Equation 3.7 as

$$r(t+\Delta t) - r(t) = r(t) - r(t-\Delta t) + \frac{F(t)}{m}\Delta t^2 + O(\Delta t^4). \tag{3.10}$$

Then we divide by Δt so as to obtain the Leapfrog velocity integrator

$$v(t+\frac{\Delta t}{2}) = v(t-\frac{\Delta t}{2}) + \Delta t\frac{F(t)}{m} + O(\Delta t^3), \tag{3.11}$$

where we have defined $v(t+\frac{\Delta t}{2}) = \frac{r(t+\Delta t)-r(t)}{\Delta t}$ and $v(t-\frac{\Delta t}{2}) = \frac{r(t)-r(t-\Delta t)}{\Delta t}$ to represent the half time step velocities.
It is now clear how the integration of the half time step is accurate to the order $O(\Delta t^3)$, whereas the full time step velocities integration of Verlet results in a worse precision (that goes as $O(\Delta t^2)$).

To obtain the Leapfrog position integrator we rewrite Equation 3.10 as

$$\begin{aligned} r(t+\Delta t) &= r(t) + \big[r(t) - r(t-\Delta t)\big] + \frac{F(t)}{m}\Delta t^2 + O(\Delta t^4) \\ &= r(t) + \big[v(t-\frac{\Delta t}{2}) + \frac{F(t)}{m}\Delta t\big]\Delta t + O(\Delta t^4). \end{aligned} \tag{3.12}$$

Using Equation 3.11 we finally obtain the Leapfrog position integrator

$$r(t+\Delta t) = r(t) + \Delta t v(t+\frac{\Delta t}{2}) + O(\Delta t^4), \tag{3.13}$$

which has the same precision as the Verlet position integrator from which it derives.

3.1 Molecular dynamics in the canonical ensemble

The MD simulation technique discussed up to know, is a scheme for studying the natural time evolution of a classical system of N particles in a volume V, where the total energy E is a constant of motion. If we assume the validity of the ergodic hypothesis, the averages obtained from a conventional MD simulation correspond to ensemble averages in the microcanonical NVE ensemble. In conventional MD the microcanonical ensemble NVE is indeed generated due to the conservation laws of Hamilton's equations. However, this ensemble is not the best choice in our case. Andersen was the first, in 1980 [28], to suggest that ensembles other than the microcanonical one could be generated in a MD run in order to better mimic some experimental conditions.
The first reason is that we want to simulate biological macromolecules (in particular proteins) in the cellular environment. This is reproduced more accurately by a system with

constant number of particle, volume and temperature (NVT ensemble), or constant number of particles, pressure and temperature (NPT ensemble), than by a NVE ensemble. The second reason is that the microcanonical ensemble does not allow thermodynamics fluctuations of quantities. Thus the solutions of Newton's equations of motions can not be used to study a dissipative non-equilibrium system, which means they can not be used to obtain transport properties. In addition to this, the fluctuations of some quantities are necessary to keep some other constant. For example, fluctuations in temperature are needed to maintain a constant pressure constant; moreover, they give a more likely representation of a real physical system.
Both this remarks lead us to conclude that the canonical ensemble is a better choice over the microcanonical one.

In particular, due to its greater likelihood with the physical cellular system, we will perform all simulations in the canonical isothermal-isobaric ensemble NPT. The Boltzmann distribution for the canonical ensemble is given by

$$< ... >= \frac{1}{Z_N} \int (...) e^{-\beta H} d\mathbf{r}^N d\mathbf{p}^N, \tag{3.14}$$

where Z_N is the partition function of the system:

$$Z_N = \int e^{-\beta H} d\mathbf{r}^N d\mathbf{p}^N. \tag{3.15}$$

From now on, all the ensemble averages will be defined with the density of phase space described by Equation 3.14.

3.1.1 Temperature coupling

From a statistical point of view, we can impose a specific temperature on a system by bringing it into thermal contact with a large heath bath at the desired temperature T_0. In the standard MD simulations to calculate the instantaneous temperature we can measure the mean kinetic energy, as showed by Equation 3.4.

The condition of constant T is not equivalent to the condition that the kinetic energy per particle is constant: indeed in a system that is in thermal equilibrium with a bath the relative variance in the kinetic energy of each particle is related to the second and fourth moments of the Maxwell-Boltzmann distribution.

Consequently, by using the kinetic energy per particle as a measure of the instantaneous temperature, we can see how the instantaneous kinetic temperature T in a canonical ensemble fluctuates.

This is the reason why the so-called isokinetic MD schemes [29] or the velocity-scaling schemes, which keep the average kinetic energy per particle constant and do not allow fluctuations of T, do not correctly simulate the true constant-temperature ensemble.

These schemes give incorrect results especially in the case where the measured equilibrium averages are sensitive to fluctuations, since they do not allow them. Moreover, they are not time reversible.

Fortunately, there are several techniques usually implemented with success in MD to realize an ensemble with constant (in the just discussed sense) temperature. The most common ones are:

- The Andersen thermostat.
- The Berendsen thermostat.
- The modified Berendsen thermostat, or velocity rescaling temperature coupling.

The Andersen thermostat

This method, introduced by Andersen [30], employs an NVE integrator and periodically re-selects each component α of the velocities of each particle i from a Maxwell-

Boltzmann distribution at the desired temperature:

$$P(v_{\alpha,i}) = \left(\frac{m_i}{2\pi k_B T_0}\right)^{\frac{1}{2}} e^{-\frac{m_i v_{\alpha,i}^2}{2k_B T_0}}. \tag{3.16}$$

As the system evolves, the distribution of the velocities will depart from this distribution: in order to control the temperature, we can ”refresh” the velocities so as to go back to the Maxwell-Boltzmann distribution at the desires temperature.
This is intended to mimic collisions with the particles in a heath bath at a specified T_0. The strength of the coupling to the heath bath is specified by a collision frequency ν. The stochastic collisions can be considered as Monte Carlo moves that transport the system from one constant-energy shell to another, accordingly to their Boltzmann weight; between them the system evolves according to the normal Newtonian laws.
Thus this MD scheme is turned into a Markov process [31]. Because of its stochastic nature this method does not yield good result for dynamic properties: the stochastic collisions disturb the dynamic in an unphysical way and lead to sudden random decorrelation of the particles' velocities. Moreover, by randomizing correlated motions it slows down the kinetics of system.

Berendsen thermostat

The Berendsen thermostat (or proportional thermostat) was introduced in 1984 [32] and reproduces a weak coupling to an external bath using the principle of least local perturbation. It is based on supplementing the Hamilton's equations by a first-order equation for the kinetic energy, whose driving force is the difference between the instantaneous kinetic energy and its target value [33]. In this way it allows the temperature fluctuations that are present in the canonical ensemble.
This thermostat tries to correct the deviations of the actual (or instantaneous) temperature $T(t)$ from the prescribed one T_0 by multiplying the velocities by a certain factor λ, defined as

$$\lambda = 1 + \gamma\Delta t\left(\frac{T_0}{T} - 1\right), \tag{3.17}$$

where γ is a dumping constant related to the strength of the coupling to the bath. In practice, the velocities are scaled at each time step so that the rate of temperature change is proportional to the difference in temperature. This operation is usually performed at a predetermined frequency during equilibration, or when the kinetic energy exceeds the limits of an interval centered around the target value.

This method of coupling has the advantage that the strength of the coupling can be varied and adapted to the user requirement. However, this method suffers from the same problem as the velocity rescaling scheme, in that the energy fluctuations are not captured correctly and a correct canonical ensemble is not generated.

As a consequence, for small systems or when the observables of interest are dependent on the fluctuations rather than on the averages, this method cannot be used.

Because of this, the modified Berendsen velocity-rescaling thermostat is introduced.

Modified Berendsen thermostat

In the modified Berendsen thermostat, the rescaling factor λ is calculated so as to enforce a canonical distribution for the kinetic energy thanks to an additional stochastic term. Instead of forcing the kinetic energy to be equal to a chosen value, we select its target value K_0 with a stochastic procedure aimed at obtaining the desired ensemble (in our case the canonical one); this means that K_0 is drawn from the canonical equilibrium distribution for the kinetic energy.

Nevertheless, this procedure disturbs the particles' velocities: each time the rescaling is applied, the modulus of the velocities will exhibit a fast fluctuation. To obtain a more smoother result, we can distribute among a number of time steps the rescaling procedure instead of extracting a K_0 at each time step [34]. This can be done because our only requirement is that the random changes in the kinetic energy leave a canonical distribution unaltered. Moreover, we can base the choice of K_0 on its previous value, so as to obtain a smoother evolution.

To obtain this result this method follows these steps:

1. It evolves the system for a single time step according to the Hamilton's equations, using a time-reversible area-preserving integrator.

2. Then it calculates the kinetic energy and evolves it for a time corresponding to a single time step using an auxiliary continuous stochastic dynamics (that must

preserve the canonical distribution).

3. Finally it rescales the velocities so as to enforce this new value of the kinetic energy.

This leads to a new equation in which the addition of a stochastic term ensures a correct kinetic energy distribution, resulting in a correct canonical ensemble.

3.1.2 Pressure coupling

In the same spirit as temperature coupling, we can couple the system to a pressure bath to achieve a constant pressure. There are several techniques to realize a coupling, the most common ones including:

- The Berendsen pressure weak coupling scheme.
- The Parrinello-Rahman pressure coupling.

Berendsen pressure coupling

The Berendsen representation for a pressure bath follows the same idea of the Berendsen thermostat introduced in Section 3.1.1: it is based on the weakly coupling with a large system at constant pressure.
The Berendsen algorithm rescales the coordinates and box[1] vectors by adding an extra term to the equation of motion that has the effect of a first-order kinetic relaxation of the pressure towards a given reference pressure P_0. Since the equations of motion are modified by pressure coupling, the conserved energy quantity also needs to be modified. For first order pressure coupling, the work the barostat applies to the system every step needs to be subtracted from the total energy to obtain the conserved energy quantity.
As for the Berendsen temperature coupling, this approximation does not yield to the NPT ensemble that we need, despite producing a simulation with the correct average pressure. This is a problem especially in those cases in which we are interested in calculating the fluctuations in pressure or volume (for example to calculate thermodynamic properties).

[1] This box refers to what is usually done in MD simulations: a cellular environment is performed by defining a box and by filling it with a solvent.

Because of this, we chose to perform all the simulations using the Parrinello-Rahman thermostat, which has a true correspondence to the NPT canonical ensemble.

Parrinello-Rahman pressure coupling

The representation of constant pressure by the Parrinello-Rahman pressure coupling is obtained by introducing friction terms in the equation of motions that are linked with a changing of the box coordinates. The box is defined by a matrix $\hat{b}$ with three vectors, and deforms accordingly to

$$\frac{d\hat{b}^2}{dt^2} = V\hat{W}^{-1}\hat{b'}^{-1}(\hat{P} - \hat{P}_{ref)}, \tag{3.18}$$

where V is the volume of the box, $\hat{W}$ is a matrix determining the strength of the coupling, $\hat{P}$ is the current pressure and $\hat{P}_{ref}$ is the reference pressure.
The equation of motion is then given by

$$\ddot{\mathbf{r}}_i(t) = \frac{\mathbf{F}_i}{m_i} - \hat{M}\dot{\mathbf{r}}_i(t), \tag{3.19}$$

where

$$\hat{M} = \hat{b}^{-1}\left[\hat{b}\frac{d\hat{b'}}{dt} + \frac{d\hat{b}}{dt}\hat{b'}\right]\hat{b'}^{-1}. \tag{3.20}$$

In the GROMACS molecular dynamics package implementation of this algorithm, the coupling strength is given by

$$W_{ij}^{-1} = \frac{a\pi^2\chi_{ij}|_T}{3\tau_P^2 L}, \tag{3.21}$$

where L is the largest box element, $\chi|_T$ is a tensor corresponding to the isothermal compressibility of the system and τ_P is the time constant of coupling between the system and the barostat.
If the pressure is very far from equilibrium, the Parrinello-Rahman coupling may result in very large box oscillations that could even crash the run. In that case we would have to increase τ_P, or use the Berendsen pressure coupling scheme to reach the target pressure, and then switch to Parrinello-Rahman coupling once the system is in equilibrium.

As discussed in Section 5.2, we choose to implement the modified Berendsen thermostat and the Parrinello-Rahman pressure coupling for the simulations on which our study is based.

Chapter 4

Implementation and analysis of MD simulations

All kind of MD program follow the same draft:

1. **Initial conditions:** the starting point of the MD is the initial conformation of the system, which must be provided by the user and includes:

   - The initial structure, usually in the form of a file with the coordinates of all the atoms in the system.
   - The initial velocities of all the atoms. These values can be generated from the Maxwell-Boltzmann distribution of velocities for a canonical ensemble, given by:

     $$P_v(\mathbf{v}_i, T) = \left(\frac{m_i}{2\pi k_B T} \right)^{\frac{3}{2}} e^{-\frac{m_i \mathbf{v}_i^2}{2k_B T}}. \tag{4.1}$$

     Since the resulting total energy will not correspond exactly to the required temperature T, we have to apply a correction. As a first step, the center-of-mass motion is removed (as it should remain constantly zero, since there are no external force acting on the system), and then all velocities are scaled so that the total energy corresponds exactly to T.
   - The definition of the potential that appears in the Hamiltonian.
   - The box size, which is determined by three vectors $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ that represent the three basis vectors of the periodic box.

The system topology, including the description of the force field, is a static information, in the sense that it will never be modified during the run.

2. **Forces computation:** the forces have to be calculated starting from the given potential, according to Equation 3.2. These forces can be divided in

   - Forces acting between non bonded pairs.
   - Forces due to bonded interactions. These can depend from up to four atoms.
   - Restraining and external forces.

   After the forces have been calculated, we can also compute the potential and kinetic energies, as well as the pressure tensor.

3. **Update of the conformation:** this is done by integrating the equations of motion, after having taken into account the pressure and temperature coupling. In practice, updating the conformation can be divided in three main passages:

   (a) Computing the velocities and box coordinates scaling factors due to the temperature and pressure coupling.

   (b) Integrating the scaled equations of motion.

   (c) Scaling the velocities and box coordinates values.

4. **Output:** in this final step the program writes down the positions, the velocities and the requested thermodynamic quantities (like energy, pressure, temperature, ecc). This values are recorded according to a saving step which is much larger than the integration step, in order to have a not-too-big final file.

The last three steps are executed for each step of the total time of integration.

4.1 The GROMACS engine

These four steps are at the core of ***GROningen MAchine for Chemical Simulations*** (GROMACS) [9]. GROMACS is an engine to perform molecular dynamics

simulations and energy minimization, originally developed in 1991 in the Biophysical Chemistry department of University of Groningen.
The GROMACS procedure can be divided into this four parts:

- Topology generation and solvatation (Section 4.1.1).
- Energy minimization (Section 4.1.2).
- Equilibration (Section 4.1.3).
- Dynamic simulations (Section 4.1.4).

4.1.1 Topology generation and solvatation

This initial part is divided in three steps.

- To start, the user has to give in input a structure file (of the type of the Protein Data Bank PDB structure files [35]) with the initial conformations of all the atoms of the macromolecule.
  A force field has to be selected as well taking into account all the several contributions to the potential. GROMACS already provides the implementation of different force fields.
- Once the structure file with the potential constants of interaction between atoms is generated, the system is inserted in a box defined by the minimum distance d that all the atoms must have from the box surface. Typically $d \sim 1\ nm$. The box is filled with pure water molecules. This system is simulated with periodic boundary conditions and the distance to the box surface has to be chosen in order to avoid that a certain part of the solute reaches its counterpart on the other side of the box.
- Some water solvent molecules are replaced with ions, in order to obtain an electrically neutral system and avoid divergences during the force calculations. Na^+ and Cl^- are the ions used to add a positive and a negative charge respectively.

At the end of these steps, we obtain the initial topology file.

4.1.2 Energy minimization

The minimization of the potential energy is necessary, not only because of the already discussed divergences during the integration of the equation of motion, but also because of the how the positions of the constituent atoms of a macromolecular structure are usually determined. Larger structures are obtained through X-ray crystallography, while for the smaller molecules through nuclear magnetic resonance (NMR). In both cases, the structure may not be in a relaxed state (in terms of the force field) and the constituent atoms may be distorted from their natural positions. Consequently, bond lengths and bond angles may be distorted and steric clashes in between atoms may occur. Distances between atoms little shorter than the equilibrium positions give rise to high energetic contributions in terms of Van der Waals interactions: the minimization of the potential energy of the macromolecular structure brings them back to the equilibrium values. The resulting structure is more similar to the one observable in physiological conditions, in the typical solvated form.

Generally speaking, the potential energy function of a molecular system is a very complex hypersurface; it has one deepest point, the global minimum, and a very large number of local minima.

Knowledge of all minima and of all saddle points would enable us to describe the relevant structures and conformations and their free energies, as well as the dynamics of structural transitions. Unfortunately, the dimensionality of the conformational space and the number of local minima is so high that it is impossible to sample the space at a sufficient number of points to obtain a complete survey. In particular, no minimization method exists that guarantees the determination of the global minimum in any practical amount of time. However, given a starting conformation, it is possible to find the nearest local minimum moving down the steepest local gradient of the potential energy function.

Aim of GROMACS is indeed to find this nearest minimum with this energy minimization steepest descent method based on the derivative information: GROMACS simply takes a step in the direction of the negative gradient (hence in the direction of the force), without any consideration of the history built up in previous steps. The step size is adjusted such that the search is fast, but the motion is always downhill.

Steepest descent method: Despite not being the most efficient algorithm, the steepest descent method is robust and easy to implement.

We define $\mathbf{r}$ as the vector of all $3N$ coordinates. As a first step the forces $\mathbf{F}$ end the potential energy V are calculated. Next, the new positions are calculated as

$$\mathbf{r}_{n+1} = \mathbf{r}_n + \frac{\mathbf{F}_n}{max(|\mathbf{F}_n|)} h_n, \tag{4.2}$$

where $max(|\mathbf{F}_n|)$ is the largest scalar force on any atom and h_n is the length of the step n. An initial h_0 must be given.

The forces and energy are again computed for the new positions:

If $(V_{n+1} < V_n)$ the new positions are accepted and $h_{n+1} = 1.2h_n$.

If $(V_{n+1} \leq V_n)$ the new positions are rejected and $h_{n+1} = 0.2h_n$.

The algorithm stops when either a user-specified number of force evaluations has been performed, or when the maximum of the absolute values of the force components is smaller than a specified value. Since force truncation produces some noise in the energy evaluation, the stopping criterion should not be made too tight to avoid endless iterations. A reasonable value for can be estimated from the root mean square force a harmonic oscillator would exhibit at a temperature T. This value is

$$f = 2\pi\nu\sqrt{2mkT}, \tag{4.3}$$

where ν is the oscillator frequency, m the mass, and k the Boltzmannâs constant.

4.1.3 Equilibration

Energy minimization ensures that we have a reasonable starting structure in terms of geometry and solvent orientation. To begin real dynamics, we must equilibrate the solvent and ions around the protein. If we were to attempt unrestrained dynamics at this point, the system may collapse because the solvent is optimized within itself and not necessarily with the solute. Equilibration is conducted in two phases.

1. **Thermalization:** the first phase is conducted under an NVT canonical ensemble, thus not turning on the pressure coupling. The system needs to be brought to the temperature we wish to simulate.

2. **Pressurization:** after the correct temperature is obtained, we can apply the equilibration of pressure. In this case we are going to work under a canonical NPT ensemble.

At the end of these two procedures the system is stable and the simulation can be run.

4.1.4 Dynamic simulations

At this point, the system is ready to run the simulation. One of the most important action performed by GROMACS is neighbours searching, which is fundamental for the computation of the forces.
Internal forces are either generated from fixed (static) lists, or from dynamic lists. The latter consist of non-bonded interactions between any pair of particles.
The non-bonded pair forces need to be calculated for those pairs i, j for which the distance r_{ij} between i and the nearest image of j is less than a given cut-off radius R_c (beyond which particle interactions are considered close enough to zero to be ignored).
GROMACS employ the Verlet list to efficiently maintain a list of all particles within a given cut-off distance of each other.
For each particle, it constructs a Verlet list that lists all other particles within R_c, plus some extra distance so that the list needs to be updated only every *nstlist* integration steps: these results in the buffered Verlet lists.
This searching, usually called neighbor search (NS) or pair search, involves periodic boundary conditions and determining the image.

Periodic boundary conditions: Periodic boundary conditions are used to minimize edge effects in a finite system. The atoms of the system to be simulated are put into a space-filling box, which is surrounded by translated copies of itself; in this way the artifact caused by unwanted boundaries in an isolated cluster is replaced by the artifact of periodic conditions. The periodicity still causes errors in non-periodic systems, but these errors are less severe than the ones resulting from an unnatural boundary with vacuum.
GROMACS uses periodic boundary conditions combined with the minimum image con-

vention: only the nearest image of each particle is considered for short-range non-bonded interaction terms.

4.2 Classical MD analysis

After the protein has been simulated, several analyses can be performed on the resulting trajectory. For this study, we compute in particular the Root Mean Square Deviation (described in Section 4.2.1) and the radius of gyration (described in Section 4.2.2).

4.2.1 Root Mean Square Deviation:

In MD the Root Mean Square Deviation (RMSD) is the measure of the mass weighted average distance between certain atoms of a molecule with respect to a reference structure, and is defined as:

$$RMSD(t_{ref}, t) = \left[\frac{1}{M} \sum_{i=1}^{N} ||\mathbf{r}_i(t_{ref}) - \mathbf{r}_i(t)||^2 \right]^{\frac{1}{2}}. \tag{4.4}$$

Where N is the number of considered atoms (usually the backbone of the protein), $M = \sum_{i=1}^{N} m_i$ and $\mathbf{r}_i(t)$ is the position of atom i at time t.
The molecule is fitted to the reference structure in order to not take into account the translational motion.

4.2.2 Radius of gyration:

The radius of gyration of a body about the axis of rotation is defined as the radial distance to a point which would have a moment of inertia the same as the body's actual distribution of mass, if the total mass of the body were concentrated there.
It is defined as the root mean square distance of the object's parts from either its center of mass or a given axis:

$$R_g(t) = \left(\frac{\sum_i ||\mathbf{r}_i(t)||^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}}. \tag{4.5}$$

The R_g of a protein is a measure of its compactness. If a protein is stably folded, it will likely maintain a relatively steady value of R_g, whereas if it unfolds, its R_g will change over time. As a consequence, the R_g value is related to the shape of a protein: a change of the latter can be identified by a change on R_g. For example, a globular protein can be characterized by a certain R_g value, but if it "opens-up" and takes a more elongated structure the R_g will increase.

4.3 Post-simulation analysis

Once the MD simulations have been performed, we have to analyze the resulting trajectories in order to find which are the structures chosen at equilibrium by the RRM2 fragments.
With this objective, we perform the following procedure:

1. As a first step, we apply on the trajectory of each fragment a Principal Component Analysis (PCA), described in Section 4.3.1, to study its essential motion.

2. As a second step, we apply a K-means clustering algorithm (described in Section 4.3.2) on the trajectories resulting from this dimensionality reduction technique, in order to identify the classes of possible structures at equilibrium.

3. Finally, we take each centroid as the structure representative of the corresponding class. In this way, we obtain for each fragment a certain number of possible equilibrium conformations.

These are the conformations that we will study with a recently developed method based on the ***Zernike*** formalism [11] (see Section 6). This method allows us to describe compactly the shape of molecular surfaces' portions.

4.3.1 Principal component analysis

PCA is a multivariate statistics technique that reduces the high number of degrees of freedom in a dataset. It transforms the input data by projecting them into a lower number of dimensions, called components.
Collective variable descriptions are particularly adapt for describing internal protein dynamics because of the way proteins are constructed [48]: rigid secondary and super-secondary structures and compact domains are often connected together by flexible loops that allow them to move as quasi-rigid bodies. Consequently, to characterize the large-scale motion of such molecules, one needs only to determine the variables that describe the relative coordinates of these quasi-rigid elements. Indeed studies of molecular dynamics simulations focusing on the motions of individual atoms have found evidence for

collective motions [48].

This gives a considerable reduction in the number of degree of freedom in comparison to that required to describe the dynamics in atomic detail. It is this reduction that leads to the prediction of a low-dimensional subspace in which essential protein motion is expected to take place.

This reduction is achieved starting from a transformation of the basis vectors describing the data (in our case the atoms positions) into an the orthogonal basis composed by the eigenvectors of the covariance matrix $\hat{C}$ for the set of observables. The components of such a basis, called Principal Components (PCs), are linearly uncorrelated in spite of the possible correlation present in the "natural" basis of data. The PCs are then sorted in order of decreasing values of the corresponding eigenvalues. In other words, they are ordered according to how much information about the variability of the data they contain, so that the first d ones describes most of the positional deviations (where d is small compared to $3N$). The reduction of degrees of freedom is then obtained by projecting the coordinates into a subset of this basis defined by its first d PCs and generating the so-called d-dimensional essential space. In this way we can reduce the information loss, where by saying "information" we are referring to the eigenvalues of the covariance matrix.

To implement PCA on the dynamics of a molecular structure in equilibrium in a given environment, we start by eliminating the overall transitional and rotational motion (since we are interested in the internal motion).

For each t_k time frame of a simulation we can then define the coordinate vector $\mathbf{X}$ as

$$\mathbf{X}(t_k) = \Big(x_1(t_k)\ y_1(t_k)\ z_i(t_k)\ ...\ x_N(t_k)\ y_N(t_k)\ z_N(t_k) \Big). \tag{4.6}$$

The we define the $3N \cdot M$-dimensional matrix $\hat{X}$ whose rows are the vectors $\mathbf{X}$ at each time frame (M is the number of time frames):

$$\hat{X} = \begin{pmatrix} x_1(t_1) & y_1(t_1) & z_1(t_1) & ... & x_N(t_1) & y_N(t_1) & z_N(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1(t_M) & y_1(t_k) & z_1(t_M) & ... & x_N(t_M) & y_N(t_M) & z_N(t_M) \end{pmatrix} \tag{4.7}$$

We can now define the time-covariance matrix $\hat{C}$, whose elements are defined as:

$$\begin{aligned} C_{ij} &= \sigma^2_{\tilde{x}_i,\tilde{x}_j} \\ &= \frac{1}{M-1}\sum_{k=1}^{M}\big(\tilde{x}_i(t_k) - < \tilde{x}_i > \big)\big(\tilde{x}_j(t_k) - < \tilde{x}_j > \big), \end{aligned} \tag{4.8}$$

where $\tilde{x}_i$ and $\tilde{x}_j$ could be any cartesian components of the coordinates vectors of the data set objects i and j respectively. $< >$ denotes the time average. By definition $\hat{C}$ is a squared symmetric matrix composed by real values: therefore it is hermitian and according to the spectral theorem, diagonalizable.
The diagonalized matrix $\hat{C}_{diag}$ has $3N$ eigenvectors; to obtain a sorted progress of information through these basis components we sort them by decreasing eigenvalues and define in this way the matrix $\hat{C}_{PCA}$.
By taking the first d columns of $\hat{C}_{PCA}$ we finally obtain the projection in the d-dimensional essential space for each time step.

A quantitative estimate of the information collected by each infect corresponding to an eigenvalue λ_i, is given by the Explained Variance Ratio (EVR):

$$EVR(\lambda_i) = \frac{\lambda_i}{\sum_j^{3N}\lambda_j}. \tag{4.9}$$

4.3.2 Cluster analysis

Cluster analysis refers to several machine learning algorithms that group similar objects into groups called clusters. There are several clustering algorithms, including the the K-means clustering, which is a partitional technique based on unsupervised machine learning. It can be summarized as follows:

1. The user has to specify a required number K of clusters.

2. The algorithm starts with initial estimates for the K centroids, which can be either randomly generated or randomly selected from the data set.

3. Then, the cluster centroids (or means) are computed, and objects are allocated to the cluster corresponding to the closest (according to the Euclidean distance)

centroid.
Each centroid is a vector with length defined by the number of variables (in our case, the number d of principal components) containing the means of all variables for the observation in that cluster.

4. Each cluster centroid is updated by calculating the new mean values of all the data points in the cluster.

5. These last two steps are iteratively repeated to minimize the total within Sum of Squared Error (SSE) (i.e., the sum of squared Euclidean distances between items and the corresponding centroid), until the cluster assignment stop changing significantly or the maximum number of steps is reached.

To evaluate the appropriate number of clusters (i.e. the value of K) we want to maximize the Silhouette Coefficient (SC), a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on two factors: how close the data point is to other points in its cluster, and how far away the data point is from points in other clusters. The first quantity is called similarity, and for a point x_i belonging to a cluster C_i is defined as:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{\substack{j \in C_i \\ i \neq j}} d(i,j), \tag{4.10}$$

where $d(i,j)$ is the distance between data points x_i and x_j in the cluster C_i. For clusters with size$= 1$ we set $a(i) = 0$.
The second quantity is called dissimilarity and is defined as:

$$b(i) = \min_{k \neq i} \sum_{j \in C_k} d(i,j). \tag{4.11}$$

From them we can define the silhouette value for a data point x_i as

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{max\left(a(i),b(i)\right)}, & \text{if } |C_i| > 1 \\ 0, & \text{if } |C_i| = 1 \end{cases} \tag{4.12}$$

The silhouette coefficient ranges between -1 and 1; larger numbers indicate that x_i is closer to its clusters than to other clusters, and consequently it has been clustered appropriately. Values near 0 denote overlapping clusters.

The mean $s(i)$ over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. Thus the mean $s(i)$ over all data of the entire dataset is a measure of how appropriately the data have been clustered. From the thickness of the silhouette plot the cluster size can be visualized: if there are too many or too few clusters the silhouette plots of the clusters will have very different widths.

The maximum value of the mean $s(i)$ over all data of the entire dataset is called silhouette coefficient:

$$SC = \underset{k}{max}\ \tilde{s}(k), \tag{4.13}$$

where $\tilde{s}(k)$ represents the mean $s(i)$ over all data of the entire dataset for a specific number of clusters k.

Chapter 5

CTFs simulations

5.1 Starting structures

The study of CTFs is a complicated matter in terms of MD simulations. The reason for this is the presence of the disordered CTD, which can not be well controlled with neither MD simulations nor the ***Zernike*** method. Unlike folded proteins, disordered proteins have native states that lack a well-defined tertiary structure: because of this, it has been unclear whether the physical models (i.e. the force fields) used in simulations are sufficiently accurate and if the MD simulation results (that are strongly dependent on the accuracy of the physical model used) is of good quality [49]. For what concerns the ***Zernike*** method, disordered proteins do not have an equilibrium conformation that can be selected as the most representative one for the application of this method.
To overcome this problems we consider only the RRM2 contribution to the CTF, the evolution of which can be followed with standard MD simulations. This approximation is justified by our hypothesis that the RRM2 structure is independent from the rest of the protein, so that during the MD simulation we can neglect the CTD effect on this ordered domain. To better understand the RRM2 role, we start by simulating its whole isolated domain, and as a second step the two possible fragments resulting from the cleavages at residue 208 and 219.
To obtain the respective initial starting structures, we begin from the PDB file of the Nuclear Magnetic Resonance (NMR) structure of the TDP-43 tandem RRMs in complex

with UG-rich RNA (PDB id: 4BS2) and perform three cuts:

- With the first one we isolate the whole RRM2, corresponding to the residues 192-269 of TDP-43.
- With the second one we select from this domain only the residues 209-269; this is what we call Fragment A.
- With the last cut we delete all the residues except the range 220-269; this is what we identify as Fragment B.

Figure 5.1 shows the three resulting structures: However, the resulting dynamic is probably much different from the one of the RRM2 fragments in the CTFs in vivo, for two reasons:

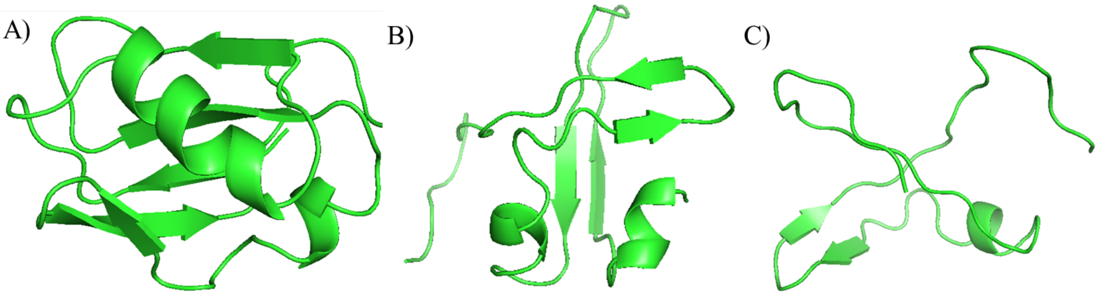

Figure 5.1: **Starting structures for the MD simulations.**
A) Starting structure of the whole RRM2. **B)** Starting structure of Fragment A. **C)** Starting structure of Fragment B.

1. We are neglecting the presence of the CTD, which constitute the bigger portion of the CTFs.
2. The misfolded conformations of the RRM2 fragments (that is, the conformation that are chosen by the fragments after the two possible cleavages) are not available jet in literature. Our simulations simply started from the structure obtained after a cut of a structure formed by RRM1 and RRM2 alone.

> In order to observe the evolution of the protein towards these misfolded conformations (as well as to find the equilibrium conformations with a lower uncertainty) we perform extensive simulations (each of 10 μs).

To verify if our simulations' results are correct, as a continuation of this thesis work, we will carry out experimental measures on cells in vitro, as described in Section 8.

5.2 Settings of the MD simulations

Generation of the system topology and definition of box and solvate

To simulate the evolution of each of these fragments, we generate the system topology using the CHARMM-27 force field [50], the standard force field for proteins, and the Verlet cutoff-scheme.
Each fragment is placed in a rhombic dodecahedron simulative box, with periodic boundary conditions, filled with TIP3P water molecules [51]. The system of the whole RRM2 includes 5269 water molecules, Fragment A 4607 and Fragment B 4658. The rhombic dodecahedron box is built so that each atom of each fragment is at least at a distance of 11 Å from the box borders. Its volume is 71% of the one of a cubic box of the same periodic distance: fewer water molecules have to be added to solvate the protein. For a protein to have the correct behavior there need to be at least two or three layers of water around it: with 11 Å there is space for approximately five layers.

Minimization and equilibration

After the topologies of the systems are built , the final system of the whole RRM2, consisting of 17038 atoms, is first minimized with 371 steps of steepest descent. In the same way, the system of Fragment A, consisting of 14777 atoms, is minimized with 102 steps, whereas the system of Fragment B, consisting of 14759 atoms, is minimized with 346 steps. Each step has a size of 0.01, while the force limit value is set to $max(|\mathbf{F}_n|) < 10^3\ kJmol^{-1}nm^{-2}$.
The thermalization and pressurization of the systems in NVT and NPT environments are run each for 0.1 ns at 2 fs time-step (for a total duration of 100 ps each), with

a saving step for coordinates, velocities and energies of 1 ps. The temperature is kept constant at 300 K with a Modified Berendsen thermostat and the final pressure is fixed at 1 bar with the Parrinello-Rahman algorithm [52] (with a time constant of coupling between the system and the barostat of $\tau_P = 2\ ps$), which guarantees a water density close to the experimental value of the SPC/E model of water[1] of 1008 kg/m^3.
The LINCS algorithm [53] is used to constraint h-bonds.

For all the three simulations, we can perform a first initial control on the correctness of the minimization and equilibration phases by looking at the evolution in time of the related quantities:

1. During the energy minimization, the average potential energy should be minimized by the steepest descent method.

2. During the temperature equilibration, the temperature of the system should reach quickly the desired plateau (in our case, 300 K) and then remain stable during the rest of the equilibration.

3. During the pressure equilibration, we impose a stabilization of the system pressure around a final value of 1 bar, which guarantees a water density close to the experimental value of the SPC/E model of water of 1008 kg/m^3.
   Actually, pressure is a quantity that fluctuates widely over the course of a MD simulation: what we have to check is that the average value computed for each simulation is not, statistically speaking, distinguishable from this reference value.

4. To have an ulterior control on the stabilization of the pressure, we can look at the evolution of the density, which should be indeed similar to the one of water. In addition to this, we expect the density values to be very stable over time, indicating that the system is well-equilibrated.

All these requests are respected in our simulations, as discussed in Appendix A.

[1] The Extended Simple Point Charge model (SPC/E) is a slight reparameterisation of the Simple Point Charge (SPC) model of water.

MD production

Once the minimization and equilibration processes have been performed and checked, the simulation can begin.
The systems are simulated with a 2 fs time-step for 10 μs in periodic boundary conditions, using a Verlet cutoff-scheme of 12 Å for the evaluation of short-range non-bonded interactions and the Particle Mesh Ewald method [54] for the long-range electrostatic interactions. For the MD production the saving step for coordinates, velocities and energies is 10 ps.
For all these steps the Leap-Frog integrator and the Verlet cut-off scheme are used.
The resulting trajectory needs to be corrected for periodicity: since during the simulation the protein will diffuse through the unit cell, we need to recenter it to avoid "jumps" across to the other side of the box.

5.3 Parameters choice

In Section 7 we present the results of the CTFs simulations and their analysis. Here, in Table 5.1 we summarize the values of the parameters (already presented in the preceding Sections) chosen for these simulations.

Topology generation and definition of the box and solvate	
Force field	CHARMM 27
Cutoff-scheme	Verlet
Unit cell	Rhombic dodecahedron box
Solvent	Water (TIP3P geometry)
Energy minimization	
Force limit value	$max(\lvert\mathbf{F}_n\rvert) < 10^3\ kJmol^{-1}nm^{-2}$
Minimization step size	0.01
Maximum number of minimization steps	$5 \cdot 10^4$
Thermalization	
Integrator	Leap-Frog
Maximum number of steps	$5 \cdot 10^4$
Integration step	$2\ fs$
Temperature coupling	Modified Berendsen thermostat
Reference temperature	$T = 300\ K$
Saving step (for coordinates, velocities and energies)	$1\ ps$
Pressurization	
Integrator	Leap-Frog
Maximum number of steps	$5 \cdot 10^4$
Integration step	$2\ fs$
Pressure coupling	Parrinello-Rahman
Time constant of coupling between the system and the barostat	$\tau_P = 2\ ps$
Reference pressure	$p = 1\ bar$
Saving step (for coordinates, velocities and energies)	$1\ ps$
MD production	
Integrator	Leap-Frog
Maximum number of steps	$5 \cdot 10^9$
Integration step	$2\ fs$
Saving step (for coordinates, velocities and energies)	$10\ ps$
Cutoff-scheme	Verlet

Table 5.1: **Parameters values chosen for the CTFs MD simulations.**

Chapter 6

Zernike polynomial expansion

Recently, in the year 2020, a new method based on the ***Zernike*** 2D polynomial expansion has been developed [11], with the aim of evaluating whether and where two proteins can efficiently interact with each other to form a complex. This new method, that we are going to call 2D ***Zernike***, is an unsupervised computational approach that looks at the shape complementarity between molecular surfaces. Indeed a key aspect for the evaluation of interactions is the identification of the binding interfaces (or hot-spots) [36–40].

Even if much of the information about the interaction is encoded in the chemical and geometric features of the structures (interactions between proteins sum up a very complex interplay between electrostatic, hydrophobic, and geometrical requirements), the set of possible contact patches and of their relative orientations are too large to be computationally affordable in a reasonable time, thus preventing the compilation of a reliable interactome.

Fortunately, the shape of local surface regions has a key role as well in predicting protein ability to bind its molecular partner [41]. This is because at shorter distances, the shape complementarity between the interacting portions dictates the stabilizing role exerted by van der Waals interactions. Biological complexes typically exhibit intermolecular interfaces of high shape complementarity.

Indeed, by expanding the well-exposed molecular surface patches in term of 2D ***Zernike*** polynomials, the ***Zernike*** method is able to rapidly and quantitatively measure the geometrical complementarity between interacting proteins by comparing their molecular surfaces.

Compared to 3D ***Zernike*** [42–47], this method is much faster. Both evaluate the shape complementarity of protein-protein interfaces with the ***Zernike*** expansion, which associates each portion of molecular surfaces with an ordered set of numerical descriptors. These descriptors are invariant under rotation, allowing easy metric comparison between the shape of different protein regions without the considerable computational cost that would be required to consider all possible relative angles between the surfaces. But 2D ***Zernike***, while preserving all the salient traits of the 3D description, decreases the computational cost, since an expansion at the same order has far less coefficient in 2D. The gained velocity allows for the exploration of a very high number of protein regions, which is an important advantage for the application of the method to MD simulation data.

6.1 Computational protocol

The first step of this algorithm is to select from the molecule a patch Σ, defined as the set of surface points constituting the region of interest. Σ was chosen to be defined from a spherical region having radius $R_{zernike}$ and centered in one point of the surface. The points contained in this sphere are divided, with a clustering from a random point that includes only the points closer than a distance D_p, in points belonging to the surface and points not directly connected to it (for example coming from a protuberance included in the sphere). Only the former will constitute the patch.

Once the patch has been selected, a plane passing through Σ is built, and the coordinates are oriented so that the z-axis has the same direction as the mean of the normal vectors of Σ. Thus, given a point C on the z-axis, the angle θ is defined as the largest angle between the z-axis and a secant connecting C to any point of the surface Σ. C is then set so that $\theta = 45°$ and each surface point is labeled with its distance r from C. As a next step, a square grid that associates each pixel with the mean r value calculated on the points inside it is built. This grid will present some pixels where no point of the surface has been projected, as well as some discontinuities on the border corresponding to the regions where the surface has a deep saddle point that can not be captured by the plane of projection. Since in this situation the ***Zernike*** method would specialize in distinguishing only where there is or there is not the surface and would not be able to describe it, we have to fill this gaps by using the average value of the surrounding pixels. Using too many pixels for this grid would results in too many of this empty pixels; on the other hand, the ***Zernike*** polynomials become increasingly complex, so that the higher order one would not be distinguishable with too few pixels. Because of this, each pixel in the grid is divided in many pixels with that same value.

Such a 2D function can now be expanded on the basis of the ***Zernike*** polynomials. Indeed, each function of two variables $f(r, \psi)$ defined in polar coordinates inside the region of the unitary circle ($r < 1$) can be decomposed in the ***Zernike*** basis as

$$f(r, \psi) = \sum_{n'=0}^{\infty} \sum_{m=0}^{n'} c_{n'm} Z_{n'm}(r, \psi), \tag{6.1}$$

with

$$c_{n'm} = \frac{n'+1}{\pi} \int_0^1 dr\ r \int_0^{2\pi} d\psi Z^*_{n'm}(r, \psi) f(r, \psi) \tag{6.2}$$

and

$$Z_{n'm} = R_{n'm}(r) e^{im\psi}. \tag{6.3}$$

$c_{n'm}$ are the expansion coefficients, while the complex functions $Z_{n'm}(r, \psi)$ are the ***Zernike*** polynomials. The radial part $R_{n'm}$ is given by

$$R_{n'm}(r) = \sum_{k=0}^{\frac{n'-m}{2}} \frac{(-1)^k (n'-k)!}{k! \left(\frac{n'+m}{2} - k\right)! \left(\frac{n'-m}{2} - k\right)!}. \tag{6.4}$$

Since for each couple of polynomials it is true that

$$Z_{n'm} | Z_{n''m'} = \frac{\pi}{n'+1} \delta_{n'n''} \delta_{mm'}, \tag{6.5}$$

the complete sets of polynomials forms a basis, and knowing the set of complex coefficients $c_{n'm}$ allows for a univocal reconstruction of the original patch.

The norm of each coefficient $z_{n'm} = |c_{n'm}|$ constitutes one of the ***Zernike*** invariant descriptors.

Figure 6.1 shows a schematic representation of the steps implemented for the *Zernike* method. Once a patch is represented in terms of its ***Zernike*** descriptors, the shape

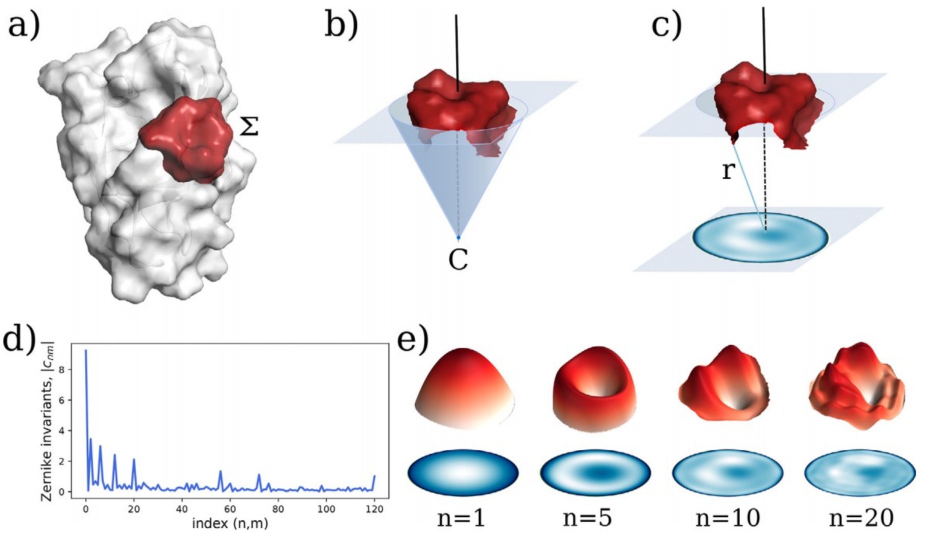

Figure 6.1: **Schematic representation of the steps of the *Zernike* method.**
a) Molecular representation of a protein surface. The red region highlights a possible patch. **b)** Each patch is firstly oriented along the z-axis, then a cone is build so that all surface points are contained inside of it. **c)** 2D projection of the patch. The origin of the cone is used to assign the color in the plane, as the distance between the origin and each point of the surface. **d)** ***Zernike*** invariant associated to the selected patch. Each invariant is defined as the modulus of the coefficients obtained projecting the patch against the ***Zernike*** basis. **e)** Surface reconstruction at different maximum expansion orders. *Figure taken from [11].*

relation between that patch and another one can be simply measured as the Euclidean distance between the invariant vectors. The relative orientation of the patches before the projection in the unitary circle must be considered. In fact, if we search for similar regions we must compare patches that have the same orientation once projected in the 2D plane, i.e. the solvent-exposed part of the surface must be oriented in the same direction for both patches, for example as the positive z-axis. If instead, we want to assess the complementarity between two patches, we must orient the patches contrariwise, i.e. one patch with the solvent-exposed part toward the positive z-axis ('up') and the other toward the negative z-axis ('down'). Figure 6.2 shows an example of what we mean by 'up' or 'down' orientation. In the former (depicted in Figure 6.2 B), the cone is built inside the molecular surface, whereas in the latter (depicted in Figure 6.2 C) the cone is built outside of it.

What is in practice done to understand if two surfaces have some complementary patches is described in the following:

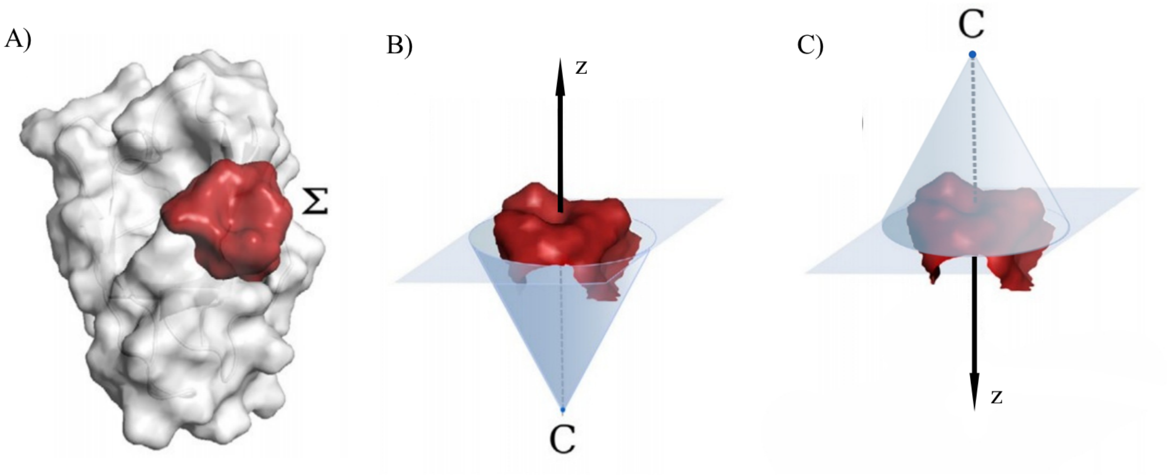

Figure 6.2: **Different orientation of the cone employed by the *Zernike* algorithm with respect to the considered patch.**
A) Molecular representation of a protein surface. The red region highlights a possible patch. **B)** The patch can be oriented towards the positive z-axis and then a cone is built so that all the patch points are contained in it. **C)** The same patch can be oriented towards the positive z-axis, depending on the situation, and then again a cone is build so that all the patch points are contained in it.

1. For both surfaces we compute the ***Zernike*** descriptors of the patches centered in all the points of the two surfaces up to the selected maximum expansion order n.

2. For each point i of the surface 1, we compute the distance between the ***Zernike*** descriptors of its patch and all the patches built on the points of the surface 2. The minimum of these values is selected, and after all the points have been studied these minimum values are mapped in $[0, 1]$ and inverted. In this way, low values of the distance are associated to high values of complementarity: at the end of the process, the points whose corresponding patches have a high complementarity with the other surface are associated to a value near one.

3. After all surface points are associated with these binding propensities, we perform a smoothing process.
   In this process each point is associated with a novel binding propensity (BP) computed as the mean value of the points in its neighborhood, defined as all the points having a spatial distance from it smaller than 6 Å.

The interacting regions should be made up mostly of elements with high complementarity and therefore a high average value of BP values.

6.2 Classification efficacy

The precision that this method can achieve in determining if the so-found value of complementary can be associated to a stable binding is related to the radius $R_{zernike}$ of the sphere which defines the patch and the ***Zernike*** maximum expansion order n. Increasingly higher n can capture more and more details. Nevertheless, an excessively accurate level of description of the molecular surface, corresponding to a too large order of expansion, would model molecular details unnecessary for the study of binding. Moreover, with a too large order of expansion we would not be able to represent the higher order ***Zernike*** polynomials with the necessary precision: with increasing values of n, the description of the noise increases as well. Here, with the term ”noise” we mean the peculiarity of the regions of interaction that have not evolved to maximize the complementarity with the partner. Figure 6.3 shows as example of how when we increase the order too much (*high order* in the Figure), the description departs more from the real situation than the case with a medium value of n (*medium order* in the Figure). This is indeed because of the increased noise description.

For what concerns $R_{zernike}$ instead, when too small patches are considered the details necessary to distinguish compatibility between interacting regions are lacking, whereas too large patches will include non-interacting zones that will have per se a low complementarity.
It can be seen from Figure 6.4 that an optimum in the complementarity is obtained when considering patches of $6 - 8$ Å of radius: at this distance the interaction regions have a specific, more than random complementarity. For what concerns the value of n, an equilibrium must be found between a not too low nor too high value, so as to well capture the overall shape without too much noise.

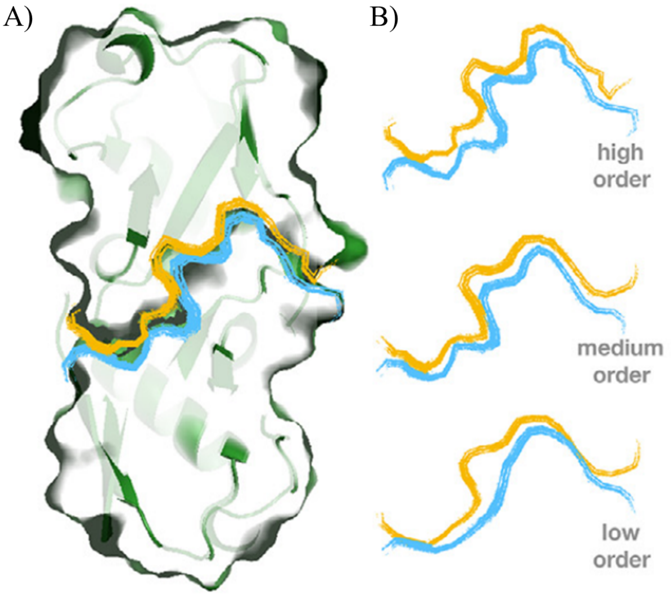

Figure 6.3: **Sketch of three possible representation of the binding region obtained by the *Zernike* expansion, depending on the expansion order n.**
A) Sketch of a binding regions between two surfaces (one in blue and one in orange). **B)** Sketch of the three possible representation of the binding region in A obtained by the Zernike expansion with different expansion order n. *Figure taken from [11].*

6.3 Parameters choice

In Section 7.2 we discuss the application of the 2D ***Zernike*** method on the 3D molecular surfaces of the CTFs equilibrium conformations resulting from the MD simulations. For that case study, we choose the parameters values, for the implementation of the ***Zernike*** expansion, shown in Table 6.1.

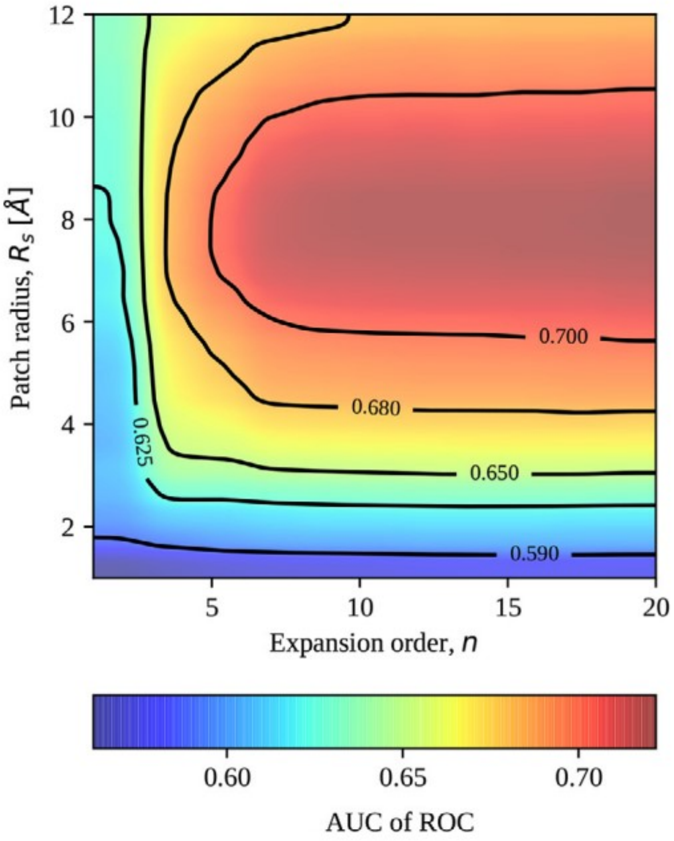

Figure 6.4: **Classification efficacy of the *Zernike* method.**
Performance, measured by the AUC of the ROC curve, in discriminating the real binding region against a set of random patches from the Protein Dataset, upon varying the patch radius $R_{zernike}$ (R_s in the Figure) and the expansion order n of the ***Zernike*** basis. *Figure taken from [11].*

Radius of the sphere	R_s =6 Å
Maximum distance between neighbouring points	D_p =1 Å
Zernike maximum order of expansion	n =20
Square grid (initial dimension)	25 × 25 *pixels*
Square grid (increased dimension)	300 × 300 *pixels*

Table 6.1: **Parameters selected for the implementation of the Zernike method.**

Chapter 7

Results

7.1 CTFs simulations and equilibrium configurations

7.1.1 Analyses of the trajectories

Finally, we can analyze these corrected trajectories to study how each fragments evolves in time and how it reaches its equilibrium conformations. We can derive a first impression of how the conformation of each fragment changes in time by looking at the evolution of the RMSD and the radius of gyration R_g.
Then, to identify the equilibrium conformations we firstly apply a PCA on the trajectory resulting from the MD simulation, after the subtraction of the rotational and translational motions. In this way we obtain an essential representation of the molecule's motion.
Then, we implement a clustering analysis to find the more representative conformations for each one of the possible conformations that the molecule can take at equilibrium: we are assuming that each cluster's center is a good representative of that cluster.
The choice of the number of PCs and clusters is justified in Appendix B.

7.1.2 Equilibrium conformations of the whole isolated RRM2 domain

Once the MD simulation for the whole RRM2 has been performed and corrected, the RMSD and the R_g defined in Equations 4.4 and 4.5 respectively have been calculated. The results are shown in Figure 7.1, together with the distribution of the former with respect to its starting minimized and equilibrated structure ($t_{ref} = t^*$): Figure 7.1 A shows the RMSD of the corrected trajectory respect to both the structure present in the

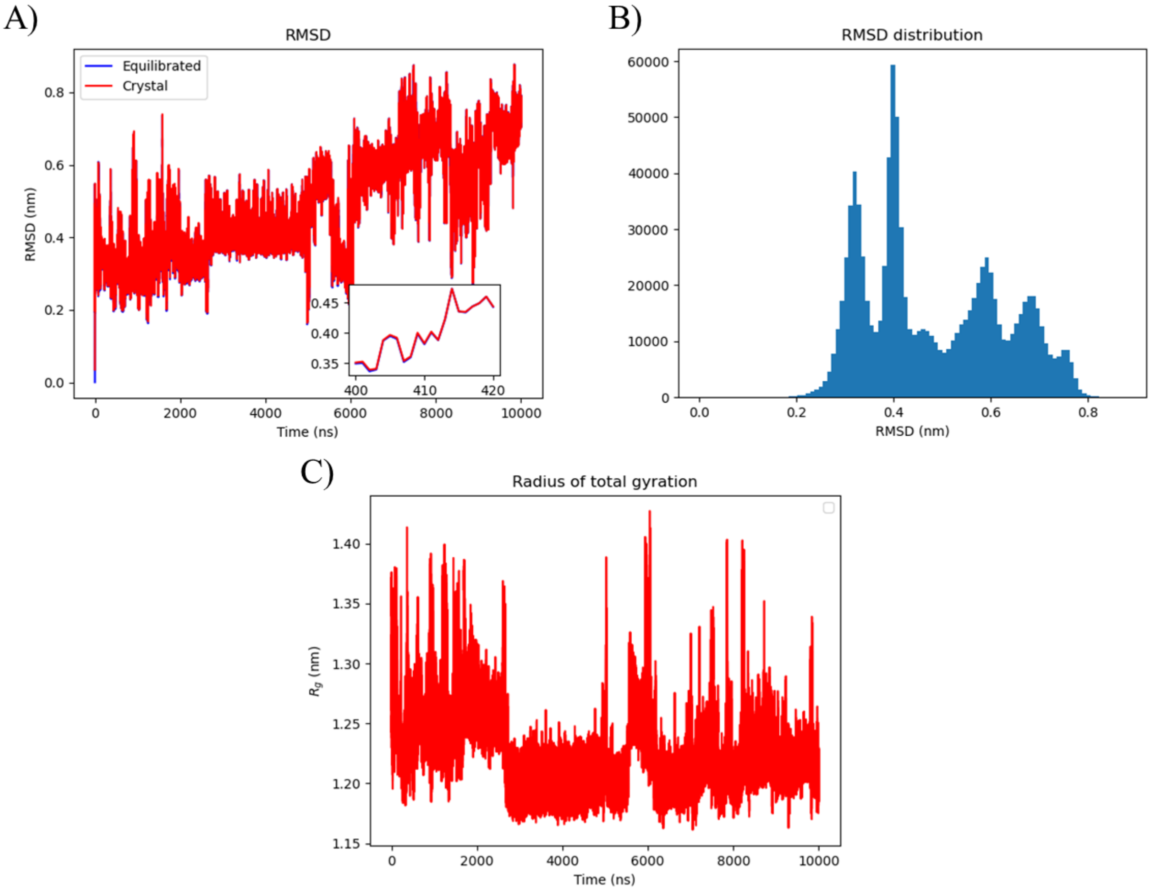

Figure 7.1: **Evolution of RMSD and R_g during the whole RRM2 MD simulation.** **A)** Time evolution of the RMSD of the evolving system respect to the equilibrated system (blue line) and the crystal system (red line). A segment of the two functions is zoomed in so as to give a better visualization of their small separation. **B)** Distribution of the RMSD of RRM2 with respect to its starting equilibrated structure. **C)** Evolution of the radius of gyration for the whole RRM2.

minimized, equilibrated system and the crystal structure (i.e., before the minimization, at $t_{ref} = 0$). Subtle differences between the two lines in this plot indicate that the equilibrated structure is slightly different from the crystal structure. This is to be expected, since it has been energy-minimized. The RMSD and R_g mean values are respectively 0.481 ± 0.141 nm and 1.225 ± 0.030 nm, but what really is interesting is their behaviour: Figure 7.1 suggests that the system during its evolution stops in certain conformations, each one characterized by a range of RMSD and R_g values.

To find these conformations, we study the essential motion of the system: at first we look at the projection of the trajectory on its first two Principal Components (PCs), and as a second step we identify the centroids of this projection, as shown by Figure 7.2. The four so-found equilibrium conformations are shown in Figure 7.2 C.

7.1.3 Equilibrium conformations of Fragment A

The same analysis can be done for the corrected trajectory of Fragment A, and Figure 7.3 shows the results. The RMSD and R_g mean values are respectively 0.598 ± 0.065 nm and 1.209 ± 0.053 nm, but again we are interested in their trend, which suggests that the system during its evolution stops in certain conformations, each one characterized by a certain range of RMSD and R_g values. To find them we look again at the centroids of the two-dimensional projection of the trajectory, as shown by Figure 7.4. The find five equilibrium conformations, shown in Figure 7.4 C.

7.1.4 Equilibrium conformations of Fragment B

Finally, Figure 7.5 shows the results for Fragment B. The RMSD and R_g mean values are respectively 0.759 ± 0.260 nm and 1.135 ± 0.068 nm and their trend suggests that the system during its evolution stops in certain conformations, each one characterized by a certain range of RMSD and R_g values.

To find them we look again at the centroids of the two-dimensional projection of the trajectory, as shown by Figure 7.6. In this case we find two equilibrium conformations, reported in Figure 7.6 C.

Going back to Figure 7.5, it is interesting to note that in this case the RMSD has a peak

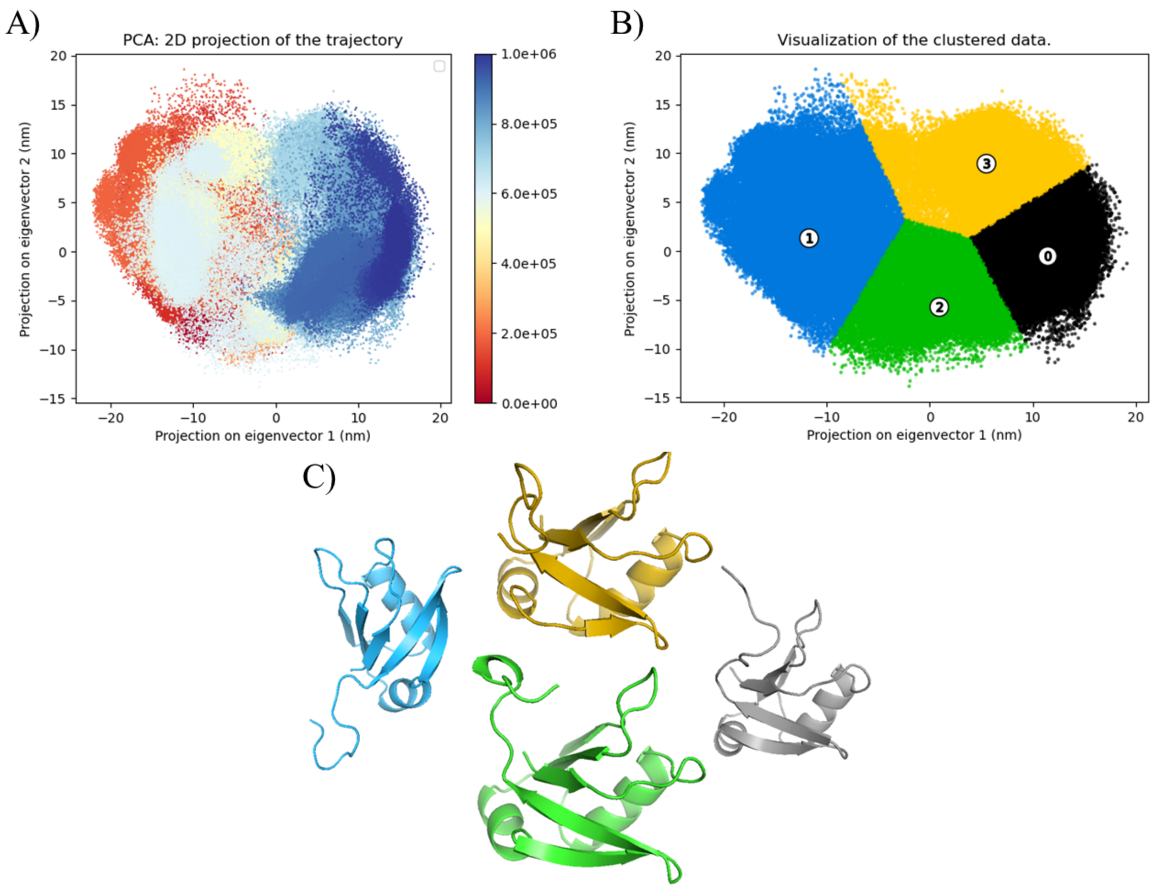

Figure 7.2: **Representative equilibrium conformations of the whole RRM2.**
A) Two-dimensional projection of the sampled conformations in the subspace spanned by the first two PCs during the simulation. Each point corresponds to the whole RRM2 domain conformation after a number of steps indicated by the color-bar; each step corresponds to an increase of 10 *ps*. **B)** Clustering of the scatter plot of the two-dimensional projection of the sampled conformations. Four clusters and their centers (labeled by the numbered white circle) are depicted. **C)** Visualization of the found four equilibrium conformations of the whole RRM2.

between $5.5 \cdot 10^2$ ns and $6.5 \cdot 10^2$ ns: there is a clear change between two equilibrium conformations. This transition goes through a completely unfolded structure, as shown in Figure 7.7, which depicts the conformation of fragment B at ~ 6201 ns, corresponding to the maximum RMSD value. This conformation has a fast dynamic (≈ 500 ns), but could be a recurrent conformation adopted by the fragment in a longer period of time. Because of this, in Section 7.1.4 we will analyse separately this part of the trajectory, going from 5500 to 6500 ns.

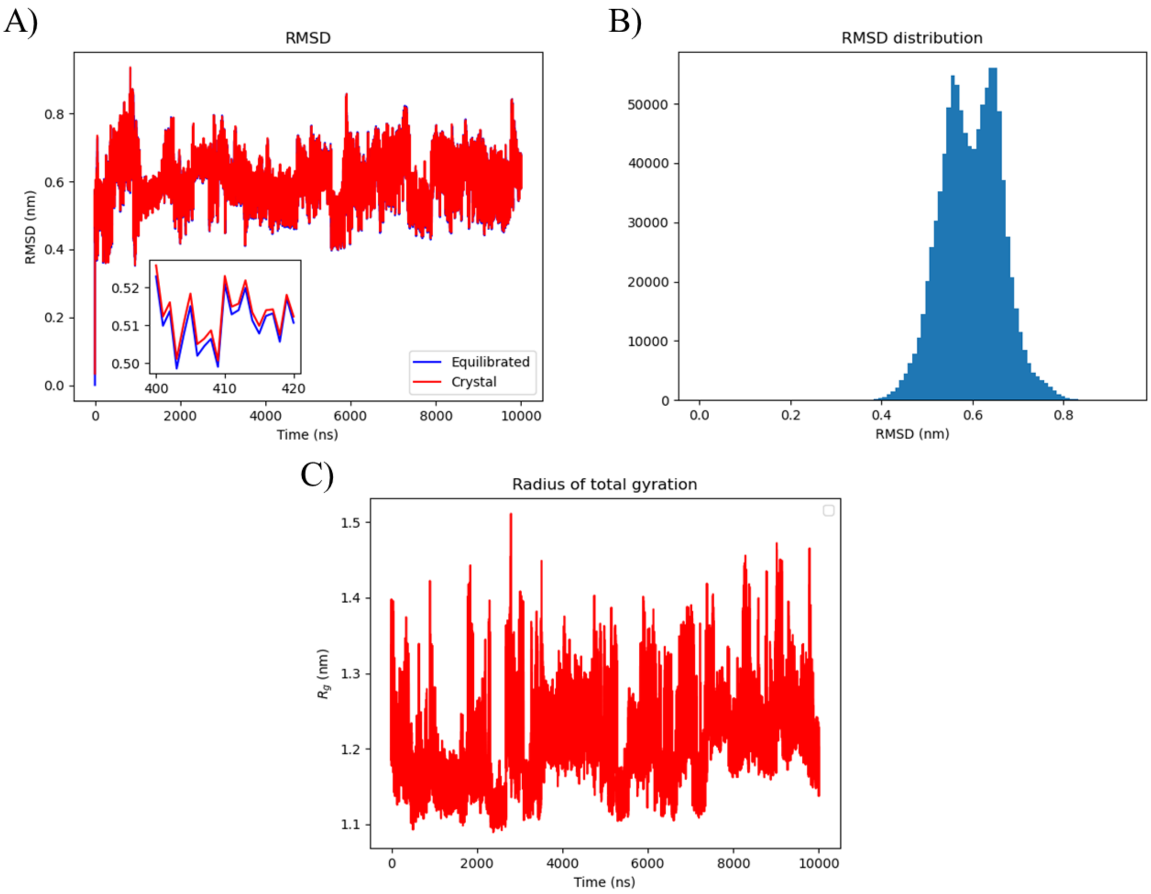

Figure 7.3: **Evolution of RMSD and R_g during the Fragment A MD simulation.**
A) Time evolution of the RMSD of the evolving system respect to the equilibrated system (blue line) and the crystal system (red line). A segment of the two functions is zoomed in so as to give a better visualization of their small separation. **B)** Distribution of the RMSD with respect to the starting equilibrated structure. **C)** Evolution of the radius of gyration.

For the same reason, it could be interesting in the future to further lengthen the MD simulation of this fragment. With a longer dynamics, we will for example be able to see it the conformational change between folded and unfolded is repeated in time: this observation will allow us to elaborate a new series of considerations on the stability of this structure. As another example, we will be able to verify if there is a particular unfolded state that the fragment assumes for a longer time.

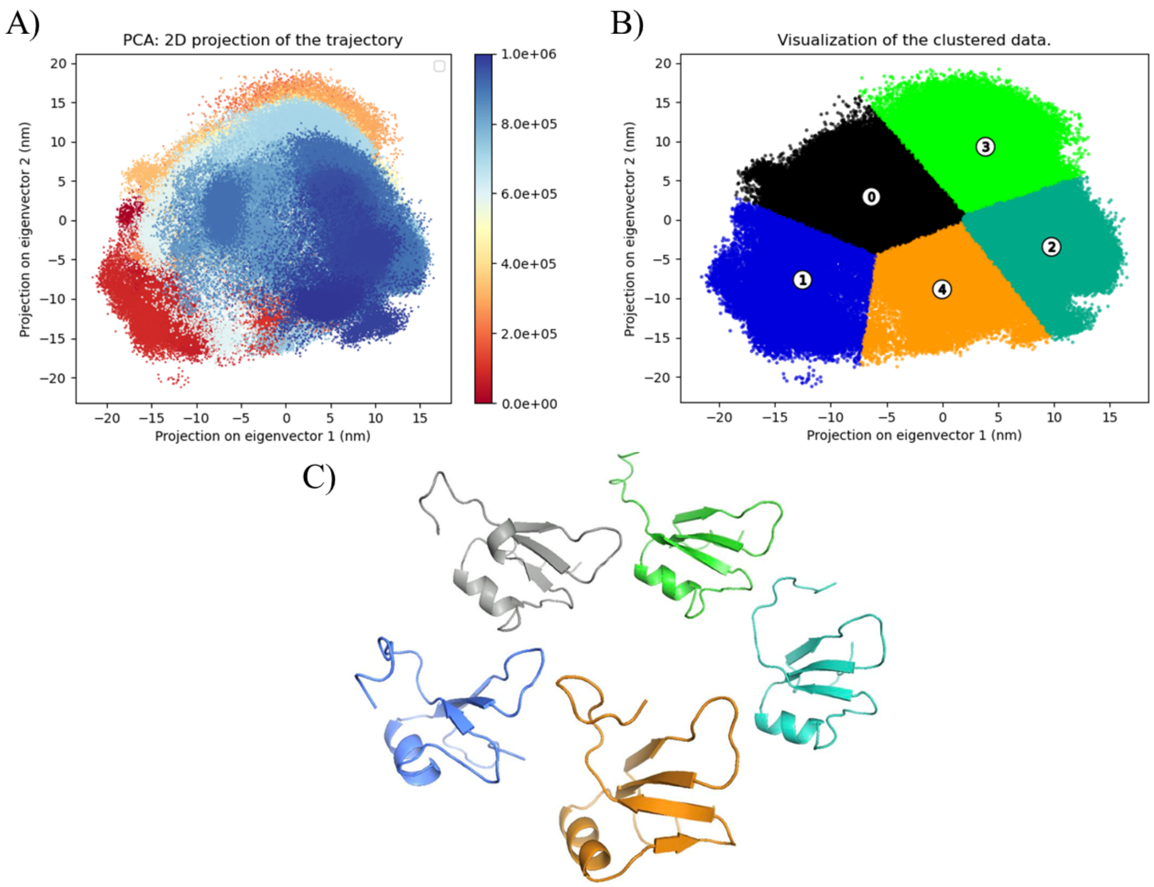

Figure 7.4: **Representative equilibrium conformations of Fragment A.**
A) Two-dimensional projection of the sampled conformations in the subspace spanned by the first two PCs during the simulation. Each point corresponds to Fragment A conformation after a number of steps indicated by the color-bar; each step corresponds to an increase of 10 *ps*. **B)** Clustering of the scatter plot of the two-dimensional projection of the sampled conformations. Five clusters and their centers (labeled by the numbered white circle) are depicted. **C)** Visualization of the five equilibrium conformations.

Unfolding of Fragment B

In the following, we are going to take a closer look to the portion of the trajectory corresponding to the unfolding of Fragment B and shown in Figure 7.8. To find the representative conformations of this unfolding we follow the same steps as before, as shown in Figure 7.9. Figure 7.9 C shows the three conformations found for the unfolding of Fragment B.

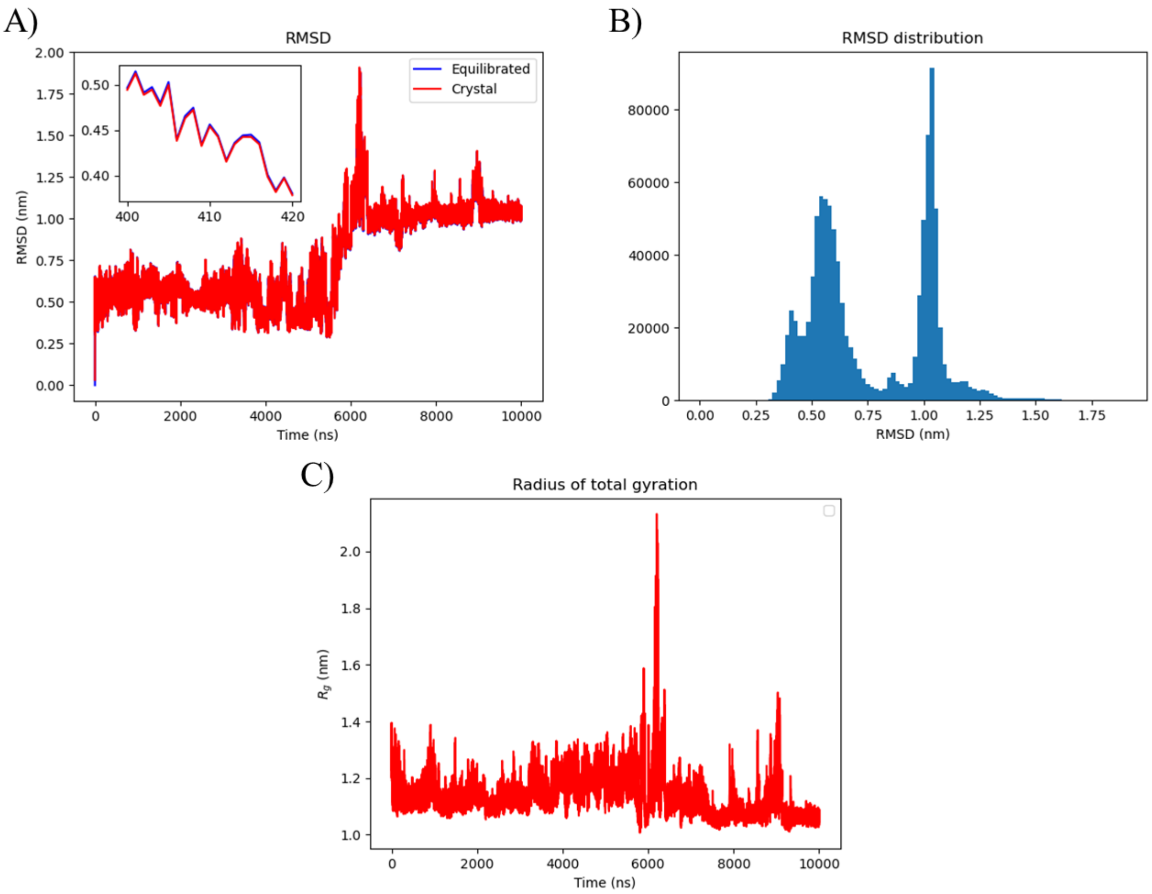

Figure 7.5: **Evolution of RMSD and R_g during the Fragment B MD simulation.**
A) Time evolution of the RMSD of the evolving system respect to the equilibrated system (blue line) and the crystal system (red line). A segment of the two functions is zoomed in so as to give a better visualization of their small separation. **B)** Distribution of the RMSD with respect to the starting equilibrated structure. **C)** Evolution of the radius of gyration.

7.2 *Zernike* method to identify the candidate binding regions

7.2.1 Analysis of the representative conformations

We compute for each one of the conformations found for Fragment A and B the corresponding 3D molecular surface. In particular, all the presented surfaces are obtained starting from the PDB files describing these conformations. To compute for this structures the solvent-accessible surface, we use DMS [55], with a density of 5 points per Å^2

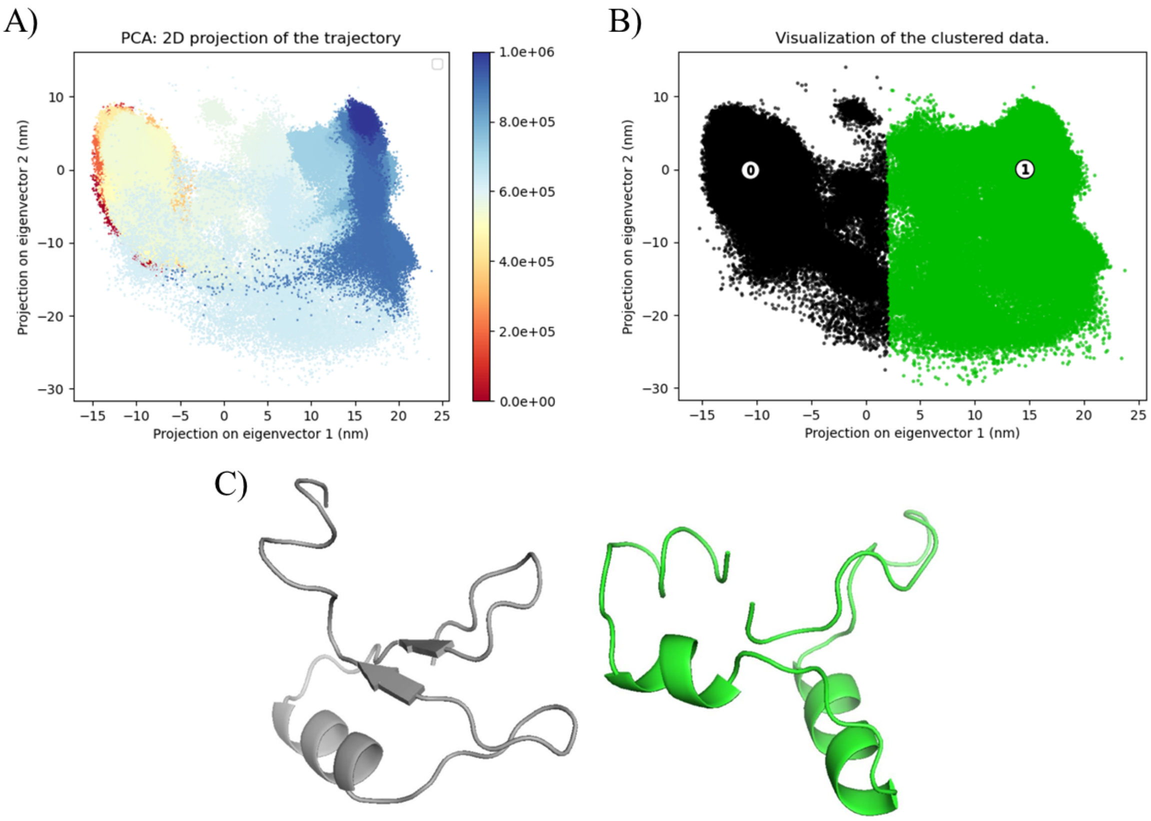

Figure 7.6: **Representative equilibrium conformations of Fragment B.**
A) Two-dimensional projection of the sampled conformations in the subspace spanned by the first two PCs during the simulation. Each point corresponds to Fragment B conformation after a number of steps indicated by the color-bar; each step corresponds to an increase of 10 *ps*. **B)** Clustering of the scatter plot of the two-dimensional projection of the sampled conformations. Two clusters and their centers (labeled by the numbered white circle) are depicted. **C)** Visualization of the two equilibrium conformations.

and a water probe radius of 1.4 Å. For each surface point, we also calculate the unit normal vector with the flag $-n$.

The shape complementarity between these molecular surfaces can now be studied with the ***Zernike*** polynomial expansion: we apply it to all the possible pairs between the 3D surfaces of the two CTFs RRM2 fragments to find the binding regions on each surface. We consider the binding both between two fragments of the same kind and between two different CTFs: since both can be present inside a cell, there is no reason to rule out a

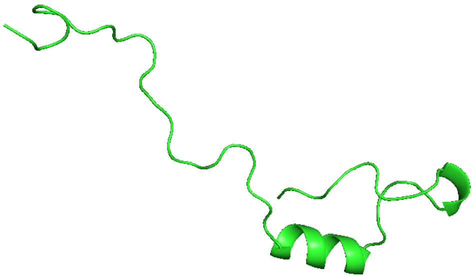

Figure 7.7: **Maximum unfolded conformation of Fragment B.**
Conformation of fragment B when the RMSD has the maximum value (at 6201540 *ps*).

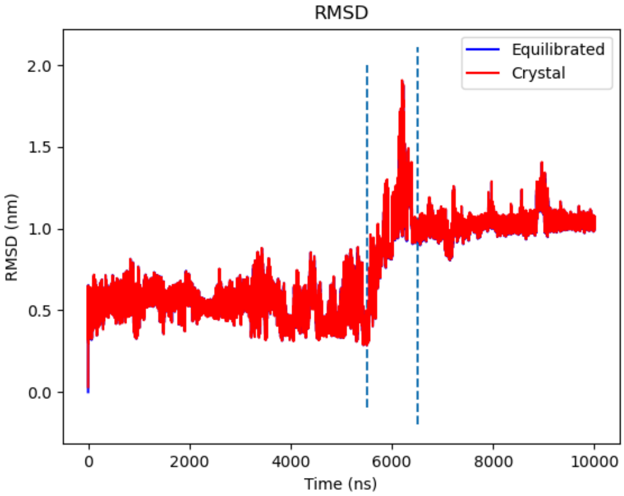

Figure 7.8: **Peak in the RMSD evolution of Fragment B and portion of its trajectories corresponding to an unfolding.**
The time interval of the unfolding of the fragment is delimited by the vertical bars.

priori the interaction between the fragments resulting from the two different cuts. We verify the shape complementarity, with the procedure described in Section 6, between all the possible pairs of conformations. The five conformations of Fragment A are called A1, A2, A3, A4 and A5. The two conformations of Fragment B found from the analysis of the whole trajectories are called B1 and B2, whereas the three identified by looking at the unfolding are called B3, B4 and B5. For more details see Appendix C.

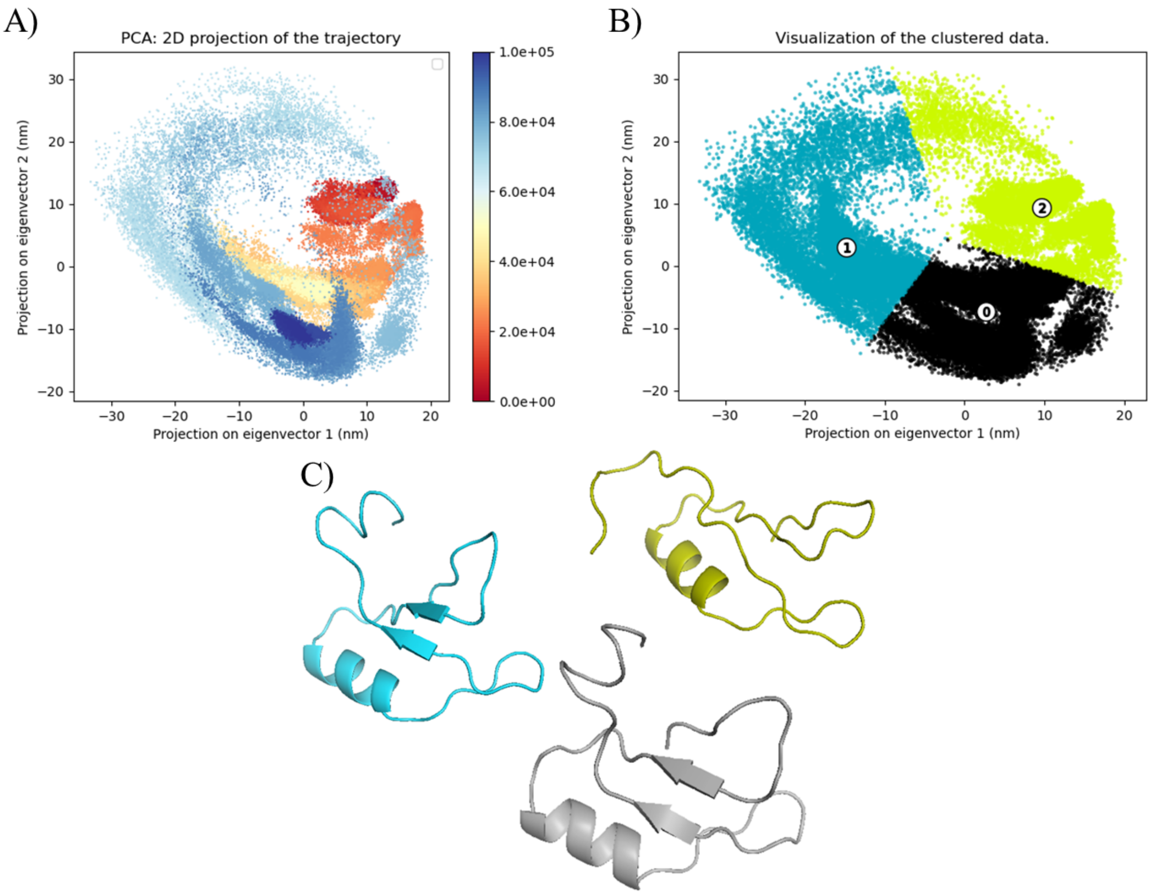

Figure 7.9: **Representative equilibrium conformations of the unfolding of Fragment B.**
A) Two-dimensional projection of the sampled conformations in the subspace spanned by the first two PCs during the unfolding portion of the simulation. Each point corresponds to Fragment B conformation after a number of steps indicated by the color-bar; each step corresponds to an increase of 10 *ps*, in the interval from 5500 to 6500 *ns*. **B)** Clustering of the scatter plot of the two-dimensional projection of the sampled conformations. Two clusters and their centers (labeled by the numbered white circle) are depicted. **C)** Visualization of the three equilibrium conformations.

7.2.2 Identification of the β-strands residues prone to aggregation

To identify the most promising regions of interaction between the two fragments (i.e., the regions that we expect to be at the core of the CTFs aggregation), we select, for each conformation, the pairing, for each fragment, that results in the highest mean BP of the residues corresponding to β-strands in the conformation sequence. In this way, we are selecting the pairings that are more prone to bind through β-strands.

Figure 7.10 shows, as an example, the result of this procedure for the first conformation of Fragment A (that we are going to call A1). After computing the binding propensity between each residue of A1 and all the equilibrium conformations' surfaces of Fragment A, we obtain for each pairing a binding propensity profile (as the one reported in Figure 7.10 in blue) and the corresponding mean BP of the β-strand residues. We select between these pairings the one that maximizes this value (in this specific case, the fifth conformation of Fragment A, that we are going to identify as A5). We apply the same procedure between A1 and all the equilibrium conformations of Fragment B. This process is repeated for each one of the five equilibrium conformations of both Fragment A and Fragment B.

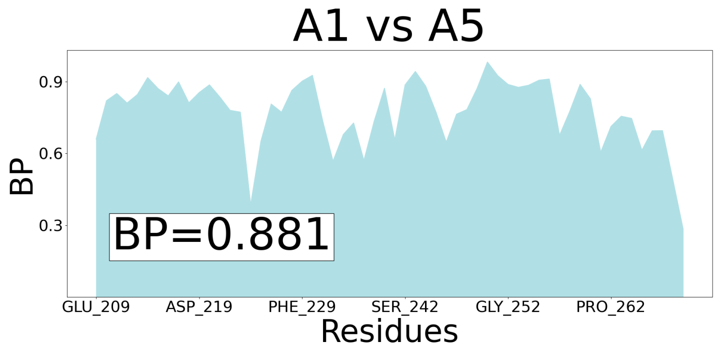

Figure 7.10: **Binding propensity profile of the first equilibrium conformation of Fragment A (that we call A1) with the conformation of Fragment A itself that corresponds to the highest mean BP of the residues associated to β-strands.**
On the y-axis, the binding propensities scores for the first conformation of Fragment A when compared the equilibrium conformation of the same fragment that results in the highest mean BP of the β-strand residues (that is, the fifth conformation, that we call A5). On the x-axis the residues composing Fragment A (one residue every ten is labeled for clarity). In the legend, the mean value of the binding propensity profile of the residues associated to β-strands is reported.

Figure 7.11 shows the results for each of these ten conformations. The pairings found in the this way should be the more interesting ones for the CTFs aggregation process.

In agreement with previous studies, new RNA aptamers could be in the future proposed as candidates for the interruption of the molecular interaction between the CTFs of the TDP43 protein: to propose some binding regions suited for testing with aptamers we follow a second approach.

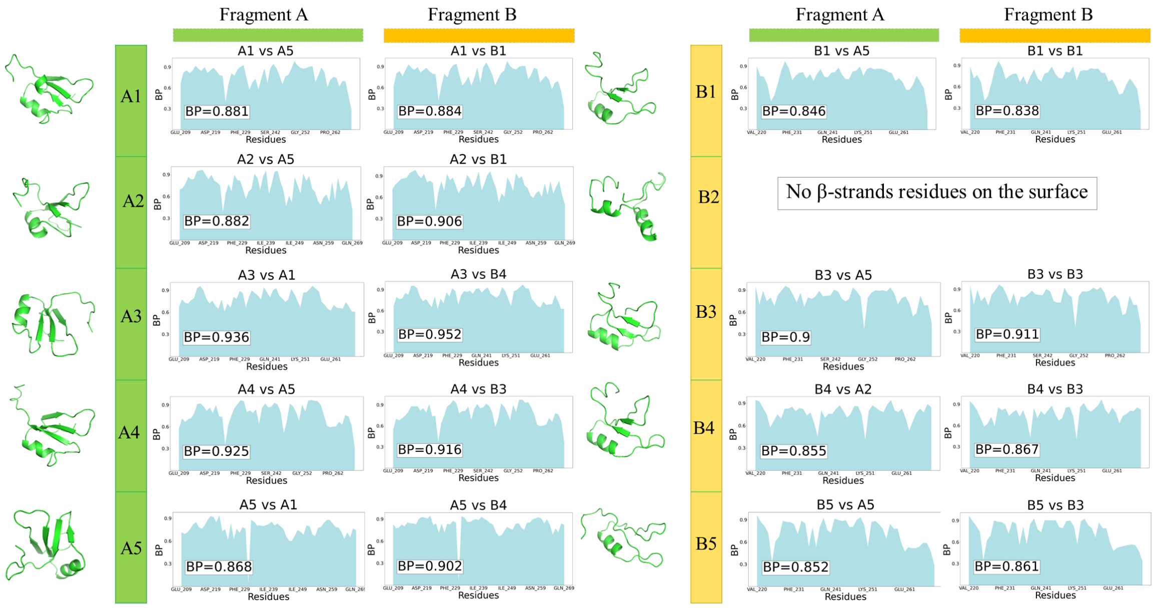

Figure 7.11: **Binding propensity profiles for the highest mean BP of the residues corresponding to β-strands.** Binding propensities scores for the conformations where the mean BP of the β-strand residues in the starting conformation has the highest value. Each row corresponds to one of the ten conformations: for each row i, the plot on the left corresponds to the BP of the residues of conformation i respect to the conformation of Fragment B that results in the highest mean BP for the β-strands residues. The plot on the right instead shows the same results, but with the best pairing with a conformation of Fragment B. The row corresponding to B2 is empty because this conformation has no β-strand residues on its surface.

7.2.3 Proposal of new binding regions for the insertion of aptamers

In this second approach, for each conformation we sum the BPs obtained for all of its possible pairings with the other conformation. In this way, we obtain a clear representation of the residues in each conformation that are in general more involved in the interaction with other surfaces.

As a second step, we have to consider the fact that the ***Zernike*** method looks only at the shape complementarity between surfaces, which is a necessary but not sufficient condition for the interaction to take place. We also have to remember that our objective is to determine the structures of the specific aptamers able to prevent the aggregation between fragments, by interacting only with the binding regions and preventing their interactions with other fragments.

Consequently, to select among the residues found with ***Zernike*** the ones corresponding to the binding regions that can be bind by an aptamer, we apply a chemical-based constraint.

In particular, we consider the Coulombic interaction: since aptamers are characterized by a negative charge, we select among the ***Zernike***-selected residues the ones associated with a positive charge.

These are the only ***Zernike***-selected residues on which an aptamer should be able to bind.

To select these regions we use ***UCSF Chimera*** [56] to visualize the Coulombic surface colouring of each conformation: negative regions are red-coloured, positive ones are blue-colored. Then we analyze the Coulomb colouring of the surface regions corresponding to our ***Zernike***-selected residues, and select only the residues corresponding to non-negative regions.

Figure 7.12 shows an example of how a binding region is selected: That said, we are interested in studying the aggregation of these fragments as hypothesised by a model according to which their interaction is mediated by the β-strands. To better understand the importance of these β-strands we select, among the ***Zernike***-found residues, the ones corresponding to β-strand fragments as well.

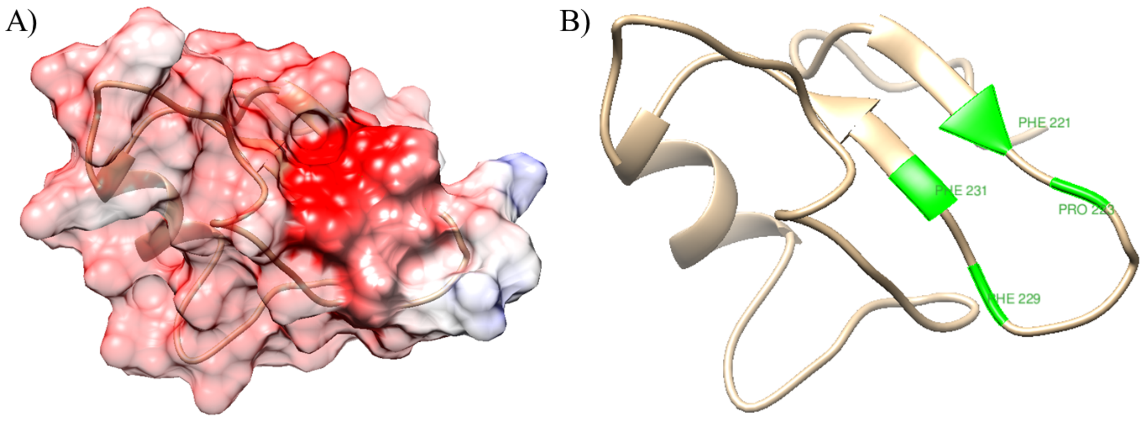

Figure 7.12: **Example of binding region selection.**
A) Coulomb surface colouring of the binding region of the first binding region of the first equilibrium conformation of Fragment A. **B)** Selection of the binding region sequence.

The results of these two selections for Fragment A are shown in Table 7.1. Table 7.2 shows instead the results for Fragment B.

Conformation A1	
I binding region	PHE221, PRO223, PHE229, PHE231
II binding region	GLN213, TYR214, ILE250, LYS251
I β-strand region	PHE221
II β-strand region	PHE231
Conformation A2	
I binding region	SER212, GLN213, TYR214
I β-strand region	GLN213, TYR214, GLY215, ASP216, VAL217, MET218, ASP219, VAL220, ILE222
Conformation A3	
I binding region	PRO223, ARG227, PHE229, PHE231
I β-strand region	SER254, VAL255, IHS256
Conformation A4	
I binding region	PHE221, PHE226, ARG227, ALA228, PHE229, PHE231
II binding region	VAL255, HIS256, ILE257, SER258
I β-strand region	PHE221, ALA228, PHE229, PHE231, THR233
II β-strand region	SER254, VAL255, HIS256, ILE257, SER258
Conformation A5	
I binding region	PHE221, ILE222
I β-strand region	VAL217, MET218, ASP219, VAL220, PHE221, ILE222
II β-strand region	THR233, PHE234
III β-strand region	ILE257

Table 7.1: **Binding and β-strands regions found for each conformation of Fragment A.**

Conformation B1	
I binding region	VAL220, ILE222, LEU248, ILE250
II binding region	ARG227, ALA228, PHE229
I β-strand region	PHE229
II β-strand region	IHS256, ILE257, SER258
Conformation B2	
I binding region	LYS224, PRO225, PHE226, ARG227, ALA228, PHE231, VAL232, THR233, PHE234, ALA235, ILE239
II binding region	HIS256, ILE257
Conformation B3	
I binding region	VAL220, ILE222, LYS224, LEU248, ILE250
II binding region	ARG227, PHE229
III binding region	ILE253
I β-strand region	PHE229
II β-strand region	HIS256, ILE257
Conformation B4	
I binding region	VAL220, PHE221, ILE222, PHE229, PHE231, ILE249, ILE250, GLY252, ILE253, HIS256, ASN267, ARG268
I β-strand region	PHE229, PHE231
II β-strand region	HIS256, ILE257, SER258
Conformation B5	
I binding region	VAL220, PHE221, ILE222, ARG227, PHE229, PHE231, THR233, PHE234, ILE250, GLY252
I β-strand region	PHE229, PHE231

Table 7.2: **Binding and β-strands regions found for each conformation of Fragment B.**

Chapter 8

Future developments

Finding the set of residues corresponding to possible regions of interaction, underlined in Tables 7.1 and 7.2, is the end point of this thesis. That said, there is still much work to do to understand the mechanisms underlying the TDP-43 CTFs aggregation: in the future we are going to further develop the here presented study.
Our future work will be articulated in three main steps:

1. We will examine more in depth the fragments' trajectories and the corresponding equilibrium conformations. In preparation of this more extensive analysis we developed a new computational strategy for defining the minimal protein molecular surface representation. Thanks to this novel method, briefly introduced in Section 8.1, we will be able to reduce the computational time needed to study the shape complementarity or similarity between surfaces. This will allow us to apply the ***Zernike*** method to a high number of fragments' conformations and verify, for example, how much the surface of a fragment changes during a simulation.

2. As a next step, we will test if our proposed binding-regions are indeed at the core of the CTFs aggregation, by verifying if after their obstruction from an aptamer the aggregation is hindered.
   We already sent our proposed binding regions to the *Department of Neuroscience and Brain Technologies, Istituto Italiano di Tecnologia*[1], and the designing of

[1] Via Morego 30, 16163 Genoa, Italy

region-specific aptamers, introduced in Section 8.2, is already under way.

3. Finally, by means of the Brillouin microscopy, briefly described in Section 8.3, we will verify if after the insertion of the aptamers in CTFs expressing cells, the number and dimension of aggregates is reduced.

8.1 New minimal molecular surface representation

Predictive methods, like the complementarity search we perform with the ***Zernike*** formalism, often rely on extensive samplings of molecular patches with the aim to identify hot spots on the surface. Intuitively, the more the surface is sampled the more the reaching for hot-spot is accurate. Similarly, the higher the number of different points used to represent the surface the higher the level of detail of the molecular shape. However, time and computational costs limit both the resolution of the surface and the number of patches that can be sampled, especially for large protein complexes and/or in analyses that involve a big set of surfaces, like, as in what will be our case, many molecular dynamics frames.

Thus we want to find an optimal way to reduce the number of points to be sampled maintaining the biological information carried by the molecular surface: we define a new theoretical and computational algorithm [13] with the aim of defining a set of molecular surfaces composed of points not uniformly distributed in space, in such a way as to maximize the information of the overall shape of the molecule by minimizing the number of total points.

The basic idea of the proposed new algorithm is the selection of molecular surface points according to the local roughness (that is, the degree of complexity of shape of each surface region): increasing the sampling in high roughness and decreasing the sampling in the more flat regions. In particular, we define a sampling probability that depends on the local roughness of the surface.

To begin with, we numerically represent the molecular surface with a set of N points in the 3D space (the discretization of the surface). For each point i, we evaluate the exiting normal vector, $\bar{v}_i$, to the surface, originating from i. Next, we evaluate the local roughness of the molecular surface by looking at the relative orientation of the normal vectors with respect to each point i. To do so, starting from each point i, we define a patch including all the surface points within with a sphere of radius R_{patch} centered on the point i. We calculate the roughness of each patch as the mean of the cosines of the angles formed by the normal vectors associated to each of the n_p points of the patch and the average normal vector:

$$\mathcal{R}_i = \frac{1}{n_p} \sum_{j=1}^{n_p} \cos(\theta_{ij}) \tag{8.1}$$

with $\cos(\theta_{ij}) = \frac{\bar{v}_i \cdot \bar{v}_j}{|\bar{v}_i||\bar{v}_j|}$ and $\bar{v}_i = \frac{1}{n_p} \sum_j \bar{v}_j$. Figure 8.1 A shows for example the molecular surface for conformation A1 colored according to the local roughness. Being a mean of cosines, the roughness ranges from zero to one (see Figure 8.1 B). When the considered patch is plane, the mean value of the cosine between the normal vectors of each point i of the surface and the mean normal vector of that patch, $\mathcal{R}_i$, is close to one, while lower values of $\mathcal{R}_i$ indicate rougher patches. Then, we associate to each point j in the patch

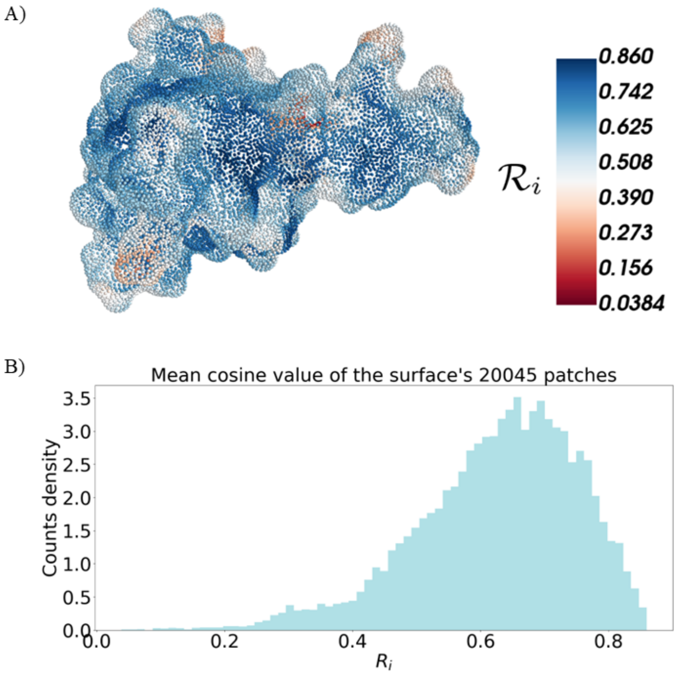

Figure 8.1: **Local roughness of the patches.**

A) Discretized representation of a molecular surface of conformation A1. Each point of the surface is coloured according to the local roughness value, $\mathcal{R}_i$.

B) Distribution of the roughness $\mathcal{R}_i$ found for each point i of the considered surface.

centered on a point i, the probability to be accepted for the sampling, defined as:

$$p(j) = \alpha(1-\mathcal{R}_i)^{\beta}\left(\frac{r_{i,j}}{R_{patch}}\right)^{\gamma(1+\mathcal{R}_i)+\delta}, \tag{8.2}$$

where $r_{i,j}$ is the distance of the point j from the center i, and α, β, γ and δ are parameters that can be optimized to yield different sampling scenarios. For the specific case of the CTFs aggregation, we are interested in selecting the absolute best combination of parameters, but we designed this method with the aim of it being as user-friendly as possible: if desired, the parameters optimization can start from a pre-determined maximum number of selected surface points or with constrictions on the shape of $p(j)$.

In general, when a patch i has a high roughness, more points are needed to describe it. On the other hand when it is more plane we need fewer points, and indeed $(1-\mathcal{R}_i)$ becomes smaller. Finally, the center of a patch is always selected, but then to capture the surface's irregularities we can use as centers for the *Zernike* patches the points further away from it, i.e. the ones with a high value of $r_{i,j}$. By elevating this term to the $(1+\mathcal{R}_i)$ we are changing the distribution of sampled points in each patch as a function of that patch roughness.

By means of ***Zernike***, we verified if the patches centered around the sampled points are indeed the most representative of the surface: in order to evaluate the ability of the reduced molecular surface to capture the information of the complete surface, we defined a descriptor based on the local characterization of the molecular surface patch shape. To evaluate the resulting representation of the surface, we compared the shape similarity between a portion of the surface obtained from the complete surface of the protein and the same portion of the surface obtained via our algorithm. In order to study the gain of the proposed algorithm in terms of information preserved in the reduced version of the molecular surface, we compare the description of the complete surface also with random sampling, which represents the approach of trivial reduction of each molecular surface by decreasing, without criteria, the density of the number of points in space. The parameters optimization can be so summarized:

1. For each combination of the four parameters, we sample from the original total surface a number n_S of points and we define a new surface determined by these n_S

points. Then, we extract from the original total surface again n_S points, but this time with a uniform distribution (or random extraction).

2. We select from the total surface n_{test} points, and define around each of them a region with radius $R = 6$ Å.

3. Next, we associate to each one of these points, j, three vectors: $z_{tot}(j)$, $z_S(j)$ and $z_R(j)$. $z_{tot}(j)$ contains the *Zernike* descriptors that describe that patch as defined by all the total points included in it, $z_S(j)$ describes the patch as defined by the sampled points included in it and $z_R(j)$ describes the patch as defined by the included randomly extracted points.

4. For each of the n_{test} patches we compute the distances $Z_{t-S}(j) = z_{tot}(j) - z_S(j)$ and $Z_{t-R}(j) = z_{tot}(j) - z_R(J)$. We average all the obtained $Z_{t-S}(j)$ and $Z_{t-R}(j)$, and obtain respectively the values Z_{t-S} and Z_{t-R}. Since we are considering the description given by all the original points as our "ideal", for a good sampling we expect the value of Z_{t-S} to be small, and in particular smaller than Z_{t-R}.

5. Finally, we compute the difference $d = Z_{t-R} - Z_{t-S}$. The best sampling for a surface should result in the maximization of d.

When there is no restriction on the number of sampled points or on the parameters' values, we can fix $\alpha = 1$, since it is a multiplicative parameter that causes no variation of the distribution of sampled points between patches with different roughness values. Consequently, we are interested in finding the combination of β, γ, and δ that results in the highest d. While it is true that a good sampling should result in a high d combined with a low n_S, the weights that these two components should have in an optimization function will change according to the application and cannot be generalized.

We showed that our proposed sampling reduces the number of considered points, minimizing the loss of information about the protein surface shape. Figure 8.2 shows an example of how the surface of conformation A1 is described when all its points are considered versus when only some subsets -including increasing number of points- are selected, with the sampling or randomly. When a small subset of points is used to reconstruct the surface, the difference between the sampling or a random extraction of the same number of points is clearly distinguishable. The more points are considered, the more the two selections become similar.

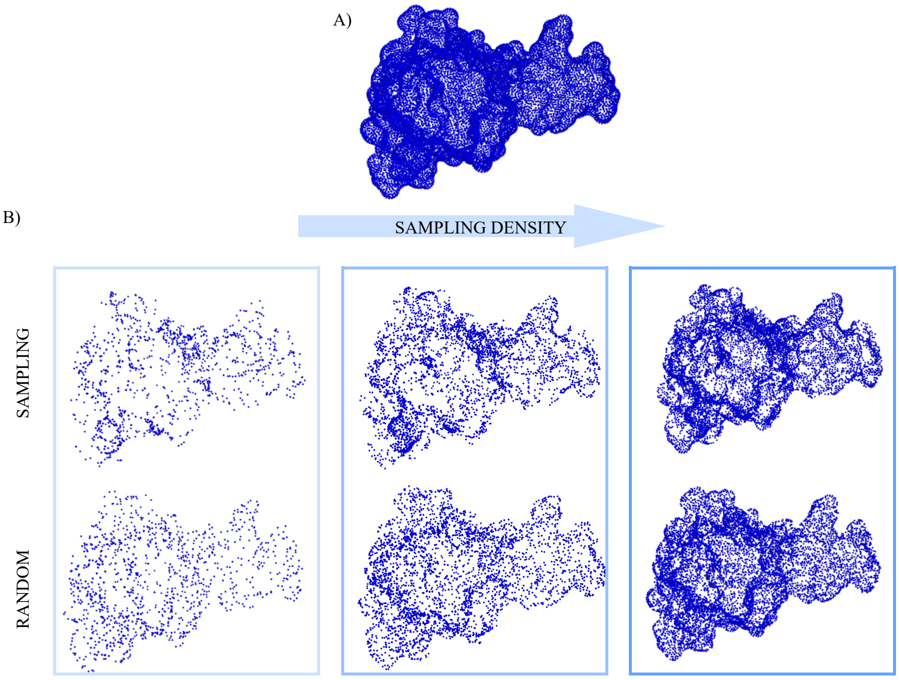

Figure 8.2: **Visualization of the 3D surfaces reconstruction of conformation A1.**
A) 3D reconstruction of the A1 surface from all its surface points.
B) The three columns depict the reconstruction of the same surface, with an increasing sampling density. In each column, the first row shows the reconstruction with a subset of the original points selected with the sampling, whereas the second row shows the reconstruction with a subset that counts the same number of points selected with the sampling, but in this case randomly extracted.

8.2 Aptamer design

The binding regions found in Section 7.2.3 are relevant to describe the CTFs aggregation process.

Artificial molecules, such as RNA aptamers or peptides, could be in the future proposed as candidates for the interruption of the molecular interaction between the CTFs of the TDP43 protein: these artificial molecules should be able to obstruct these CTF binding regions and prevent the binding of other ones. For the aptamer identification we will employ ***catRAPID*** [57] and ***STRIDE*** [58].

catRAPID is an algorithm to estimate the binding propensity of protein-RNA pairs. By combining secondary structure, hydrogen bonding and van der Waals contributions, this software predicts protein-RNA associations with great accuracy.

STRIDE is a software tool for secondary structure assignment from atomic resolution protein structures. It extracts information about the secondary structure and the accessible surface of a protein from its PDB file, in our case the PDB file of the ten equilibrium conformations found with the MD simulations. To be precise, α-helical, β-strand and turn contributions are extracted directly from the PDB. Polarity and hydrophobicity are derived from the accessible surface area by normalizing the values in the range $[0, 1]$.

Getting the PDB description of these conformations is indeed the first important result of the here presented work, since usually these kind of information have to be predicted and are not known like in our case.

For the testing of aptamers, we will use as input for ***catRAPID*** some conformers composed by ten nucleotides identified in a preceding experiments (with experimental measures) as the most prone to bind the whole TDP-43. Starting from these sequences we will modify one nucleotide at a time to identify the sequences most likely to bind each of the ten proposed conformations of the two RRM2 fragments.

As a next step, we need to identify the specific regions in which the binding between a RRM2 fragment and an aptamer happens: we expect some of these regions to match to the ones proposed in this thesis. To do this, we will cancel the contribution of each amino acid, one at a time, and evaluate the difference in the ***catRAPID*** computed binding propensity.

8.3 Brillouin microscopy

Once the aptamers able to bind to our proposed binding regions will be identified, we will test the relation of our binding regions with the CTFs aggregation with experimental measures. These measures will employ Brillouin microscopy to verify if after the insertion of the aptamers in CTFs expressing cells, the number and dimension of aggregates is reduced.

Brillouin microscopy is a type of optical elastography that has recently [12] emerged as a non-destructive, label- and contact-free method that can probe the viscoelastic properties of biological samples with diffraction-limited resolution in 3D. Obtaining an image whose resolution is limited by the unit diffraction spot, rather than by scattered light or lens aberrations, is what is meant by the term diffraction limited. Brillouin microscopy is based on Brillouin scattering. Analysis of the Brillouin spectrum can provide, for a known material density and refractive index, a unique characterization of the materialâs mechanical properties, because the sound wave properties (such as their velocity or attenuation) exhibit an intrinsic dependence on the viscoelastic properties of the material.

Since aggregates have an higher viscosity compared to the rest of the cell, this method will allow us to verify the effect of the aptamers insertion in the cells.

8.3.1 Brillouin scattering

When photons hit a sample, a small fraction of them ($\sim 10^{-12}$) interacts with the medium by exchanging (either releasing or absorbing) energy and momentum, and we can observe Stokes or anti-Stokes frequency shift. The former corresponds to a scattering at lower frequencies ($\omega_0 - \Omega$) corresponding to phonons annihilation, the latter to a scattering at higher frequencies ($\omega_0 + \Omega$) corresponding to phonons generation.

If the photons are exchanging energy and being scattered with acoustic phonons, the process is called Brillouin scattering [59].

Acoustic photons can be seen as a population of microscopic acoustic waves (with wavelength Λ and period T, related by $\Lambda = VT$, where V is the medium's sound velocity) that describe spontaneous, thermally induced density fluctuations.

Since phonons can be interpreted as density (acoustic) waves, their interaction with photons can be interpreted as an effective grating that diffracts the light; as the grating is travelling with velocity V, the scattered light experiences a frequency shift due to the Doppler effect [60]. The gain or loss of energy of the scattering field depends on the propagation direction of phonons with respect to the incident photons. This is a second interpretation that we can give to the rise to the two peaks in the scattered light spectrum, that we already introduced as Stokes and Anti-Stokes Brillouin peaks.

The Brillouin frequency shift is given by:

$$\nu_B = \frac{2n}{\lambda_i} V \sin\frac{\theta}{2}, \tag{8.3}$$

where ν_B is the frequency that characterize the acoustic photons (typically on the order of 1-20 GHz), λ_i is the wavelength of the incident light and θ is the scattering angle. Since the medium's acoustic velocity is given by $V = \sqrt{\frac{M'}{\rho}}$, where ρ is the density, from Equation 8.3 we can see that the Brillouin frequency shift is related to the the stiffness of a material. Indeed M' is the real part of the longitudinal modulus M, which recapitulates the viscoelastic properties of a material. M', also called storage modulus, provides information about the elastic properties of a material. Its relation to the frequency shift highlights the fact that, during the fast timescale of the materialâs deformation, some of the slower molecular relaxation processes cannot follow the perturbation. Because of this they behave like an effective ”stiffer” material [60].
Equation 8.3 implies that the greater the frequency shift of the Brillouin peak, the stiffer the material.

It is now clear why the combination of Brillouin scattering with scanning confocal microscopy gives a clear access to the mechanical properties of cells and tissues, which are of fundamental importance since they play intricate roles in determining biological function.

Conclusions

The investigation of the molecular mechanisms that lead to the accumulations of aggregated proteins is crucial for understanding the pathophysiology of many neurodegenerative diseases. The accumulation of aggregates containing TDP-43 in the central nervous system is a common feature in diseases such as ALS. However, the mechanisms of aggregation are not yet fully understood and various aggregation models have been proposed. In this scenario, the fundamental role of the C-terminal fragments of TDP-43 in the formation of aggregates has already been widely confirmed. Main objective of this work was indeed to propose some regions on the TDP-43 CTFs (in particular on their RRM2 fragment) as candidate cores of their aggregation.

The structures of these fragments have not been deeply studied yet and their conformations are not yet available, since their high aggregation propensity makes them difficult to be investigated experimentally. Within this framework, we began our project by studying the time evolution of the two possible RRM2 fragments constituting the CTFs, i.e. Fragment A and B, with MD simulations of 10 μs; we studied the whole RRM2 as well.

From the analysis of the trajectories, we found four equilibrium conformations for the whole RRM2 (shown in Figure 7.2 C), five for Fragment A (shown in Figure 7.4 C)), and two for Fragment B (shown in Figure 7.6 C). Since the plot of the RMSD evolution presented a clear peak for the trajectory of Fragment B, which should correspond to the fragment unfolding, we implemented the same analysis specifically for that time interval. We found as the most representative conformations for the unfolding of Fragment B three conformations, as depicted in Figure 7.9 C. These are the possible conformations that should be observable in a cell, as well as the ones that the fragments assume while interacting with each other. The definition of these equilibrium conformations is our first result.

As a next step, we searched on the surfaces of these conformations all the possible regions of interaction, by verifying their shape complementarity by means of a ***Zernike*** polynomials based characterization. Bringing further this research will include the verification in vitro of our results: following the insertion of expressively designed aptamers (starting from our suggested binding regions' residues) in CTFs-expressing cells, if our results are correct, the aggregates number and dimension should decrease. With this aim, we identified among the ***Zernike*** selected binding regions, the ones that would be able to bind an aptamer, i.e. the ones with a positive surface charge. Complementarity and positivity are necessary but not sufficient conditions for interaction. Even though having narrowed the possible region of interaction to a limited set of residues (listed in Tables 7.1 and 7.2) is already an interesting result, we would like to further develop this study in the future, for example by applying additional constraints on the region selection and a more extensive use of the ***Zernike*** method, which we will apply on all the MD simulations' frames.

Since the model that we have chosen as a starting point (shown in Figure 2.1) states that the β-strands are at the core of the CTFs aggregation, we selected among these proposed binding residues the ones located on β-strands. These residues, collected in Table 8.1, should correspond, according to our model, to the regions where the interaction between different fragments happens.

Conformation	Proposed β-strands binding residues
A1	PHE221, PHE231
A2	GLN213, TYR214
A4	PHE221, ALA228, PHE229, PHE231, VAL255, HIS256, ILE257, SER258
A5	PHE221, ILE222
B1	PHE229
B3	PHE229
B4	PHE229, PHE231, HIS256
B5	PHE229, PHE231

Table 8.1: **Proposed binding β-strands residues found for each conformation of Fragment A and B.**

Appendix A

Minimization and equilibration phases

A.1 Whole RRM2

As shown by the plots depicted in Figure A.1, the simulation of the whole RRM2 starts from a regular minimization and equilibration, which respect the behaviours described in Section 5.1. In particular, we can point out that Fragment B quickly reaches the target value of $T = 300\ K$ and after this has a stable temperature with an average value $300.0 \pm 2.5\ K$. The average value of the pressure is $-2.1 \pm 222.5\ bar$, while the reference pressure was set to 1 bar: as anticipated, statistically speaking, one cannot distinguish them.
Moreover, the average value of the density is $1012 \pm 4\ kg/m^3$, which is compatible with the expected value of 1008 kg/m^3.

A.2 Fragment A

The simulation of Fragment A starts as well from a regular (as defined in Section 5.1) minimization and equilibration, as shown by Figure A.2. Fragment A quickly reaches the target value of $T = 300\ K$ and is characterized by a stable temperature for the rest of the equilibration, with an average value $299.7 \pm 2.7\ K$. Importantly, the average value

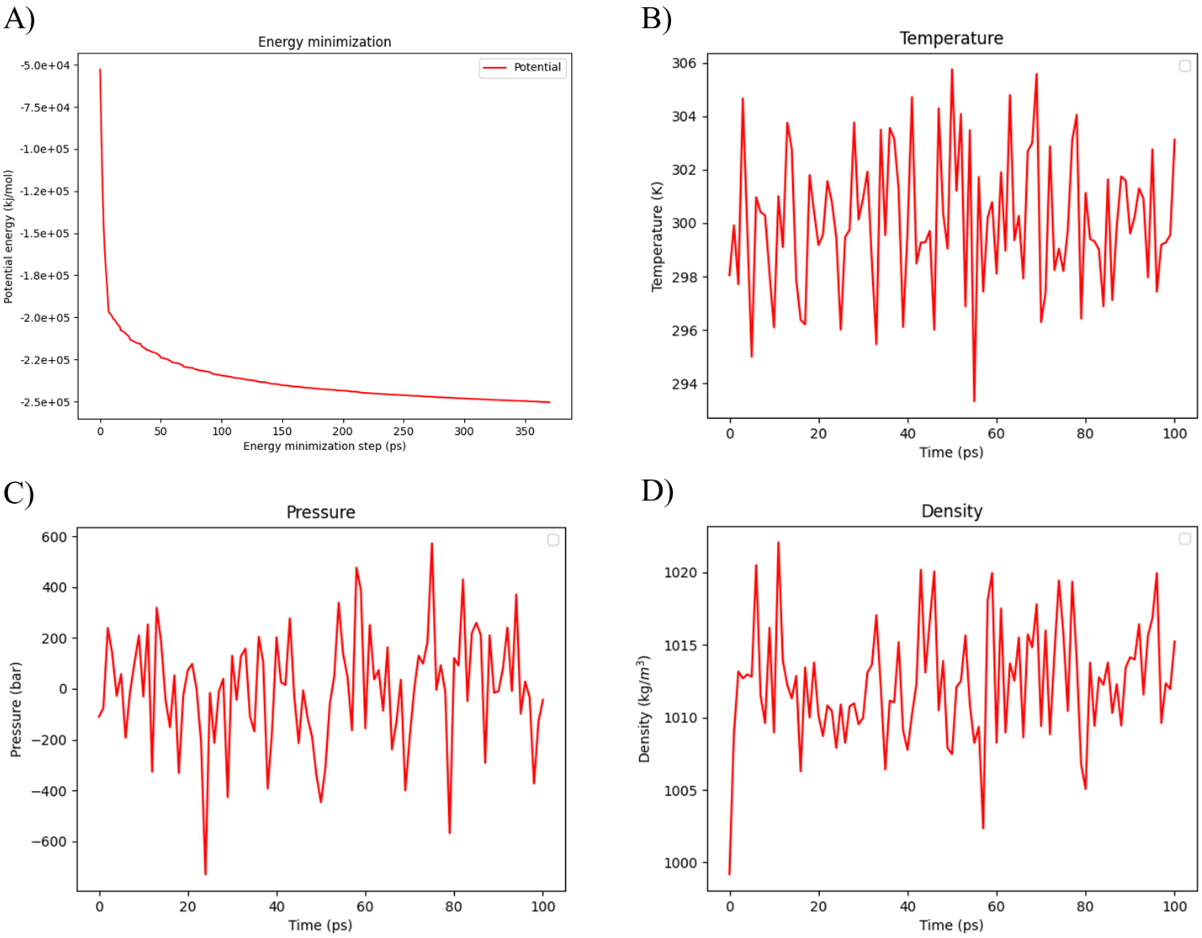

Figure A.1: **Evolution in time of the whole RRM2 system variables during the minimization and equilibration phases.**
A) Time evolution of the potential energy during the energy minimization. **B)** Time evolution of the temperature during the thermalization (NVT ensemble). **C)** Time evolution of the pressure during the pressurization (NPT ensemble). **D)** Time evolution of the density during the pressurization.

of the pressure is -59.5 ± 205.2 bar, which is again statistically indistinguishable from the reference value. Moreover, the average value of the density is 1006 ± 5 kg/m^3, which is compatible with the expected value of 1008 kg/m^3.

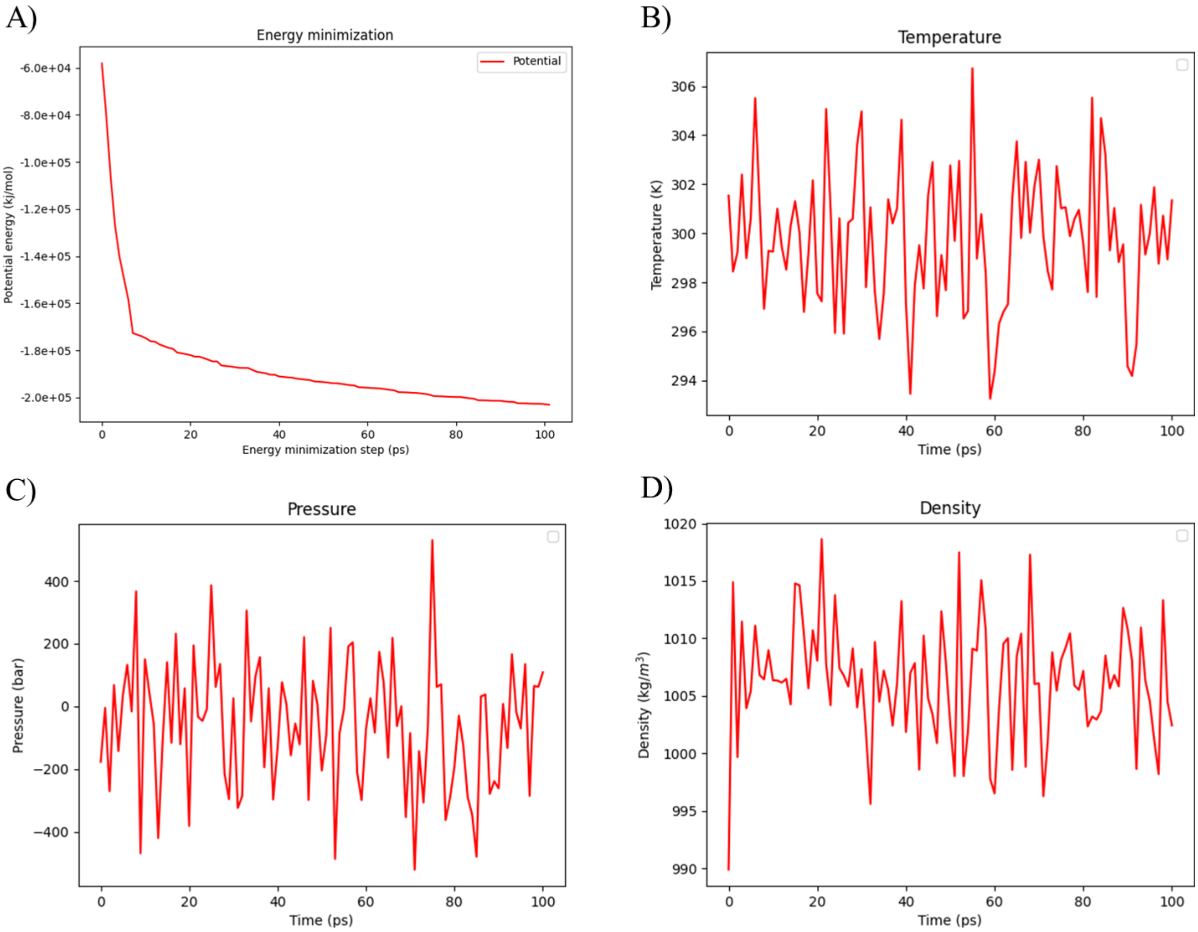

Figure A.2: **Evolution in time of the Fragment A system variables during the minimization and equilibration phases.**

A) Time evolution of the potential energy during the energy minimization. **B)** Time evolution of the temperature during the thermalization (NVT ensemble). **C)** Time evolution of the pressure during the pressurization (NPT ensemble). **D)** Time evolution of the density during the pressurization.

A.3 Fragment B

Figure A.3 leads to the same conclusions for what concerns Fragment B. Fragment B quickly reaches the target value of $T = 300\ K$ and is characterized by a stable temperature for the rest of the equilibration, with an average value $299.8 \pm 2.4\ K$. The average value of the pressure is $13.7 \pm 223.7\ bar$ and the one of the water is $1002 \pm 4\ kg/m^3$, which are again statistically indistinguishable from the respective reference values.

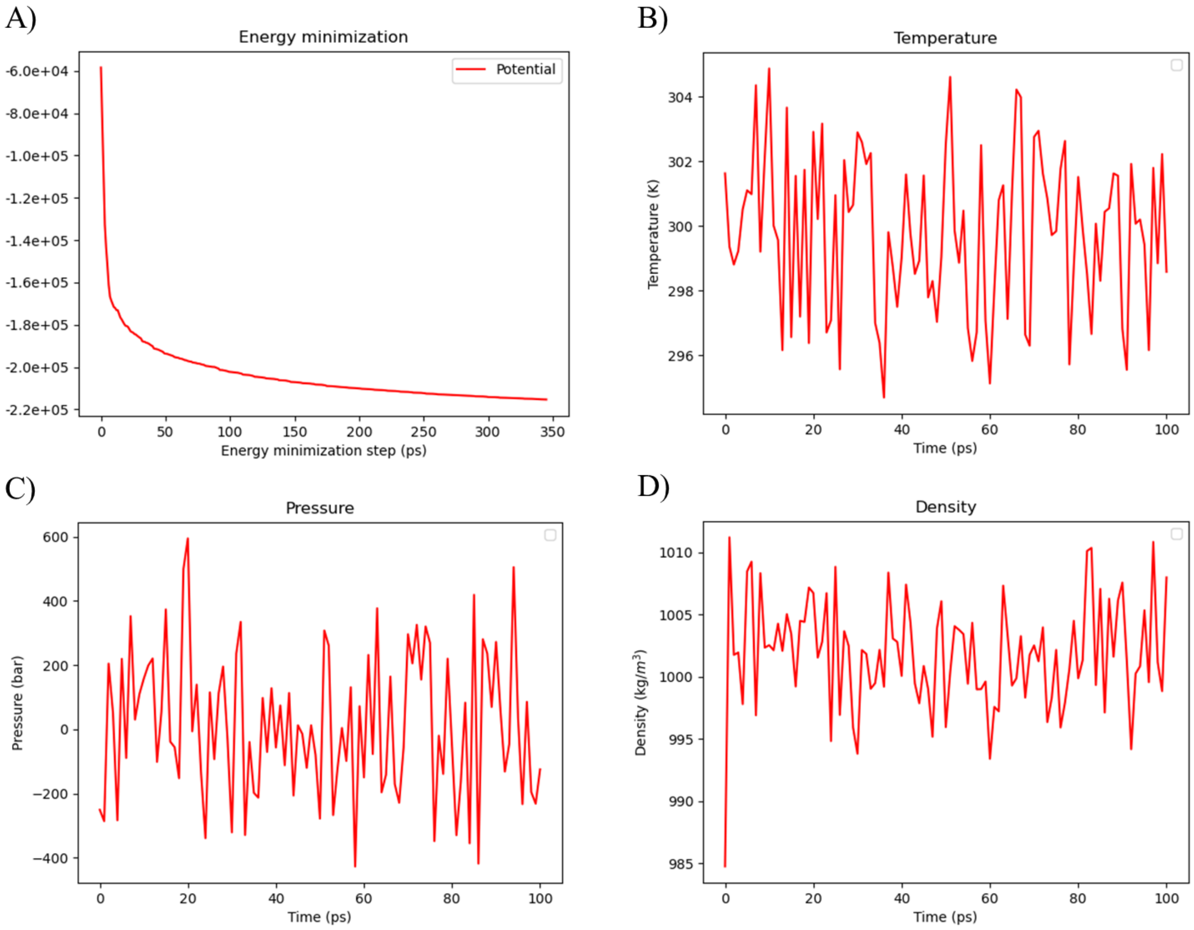

Figure A.3: **Evolution in time of the Fragment B system variables during the minimization and equilibration phases.**

A) Time evolution of the potential energy during the energy minimization. **B)** Time evolution of the temperature during the thermalization (NVT ensemble). **C)** Time evolution of the pressure during the pressurization (NPT ensemble). **D)** Time evolution of the density during the pressurization.

Appendix B

Choice of the number of PCs and clusters

B.1 Principal Components

Figure B.1 shows why only the first two PCs are considered for the projection of all the considered trajectories: their EVRs, as defined in Equation 4.9, are much higher compared to the ones of the other eigenvectors.

B.2 K-means clustering analysis

Table B.1 shows the average silhouette coefficient $\tilde{s}(k)$ for different k number of clusters, for each of the considered trajectories. In each case, we divide the trajectory's points in the number of clusters that maximizes $\tilde{s}(k)$. Figure B.2 shows the silhouette plots for each of these selected number of clusters. The size of each cluster can be visualized from the thickness of each plot.

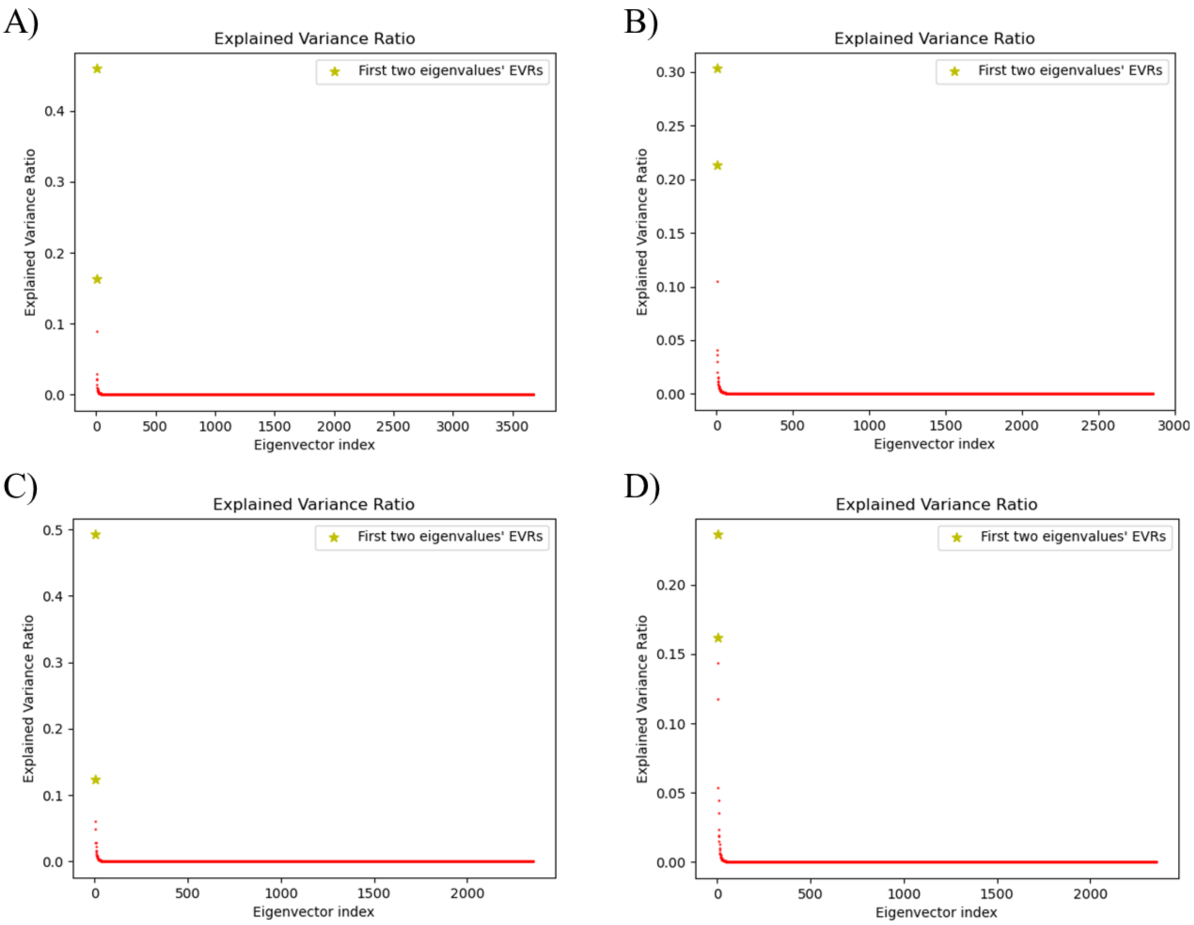

Figure B.1: **EVRs for the eigenvectors describing each of the considered trajectories.** **A)** EVRs of the eigenvectors of the whole RRM2 trajectory. **B)** EVRs of the eigenvectors of Fragment A trajectory. **C)** EVRs of the eigenvectors of Fragment B trajectory. **D)** EVRs of the eigenvectors of the Fragment B trajectory corresponding to its unfolding.

	Whole RRM2	Fragment A	Fragment B	Fragment B unfolding
$\tilde{s}(k=2)$	0.5410	0.4489	0.7095	0.4566
$\tilde{s}(k=3)$	0.5626	0.4568	0.7066	0.5355
$\tilde{s}(k=4)$	0.5996	0.5023	0.6795	0.5241
$\tilde{s}(k=5)$	0.5447	0.5148	0.6393	0.4768
$\tilde{s}(k=6)$	0.5245	0.4833	0.6735	0.4660

Table B.1: $\tilde{s}(k)$ **for different** k **number of clusters, for each of the considered trajectories.** For each trajectory in red the highest $\tilde{s}(k)$, that corresponds to the best number of clusters.

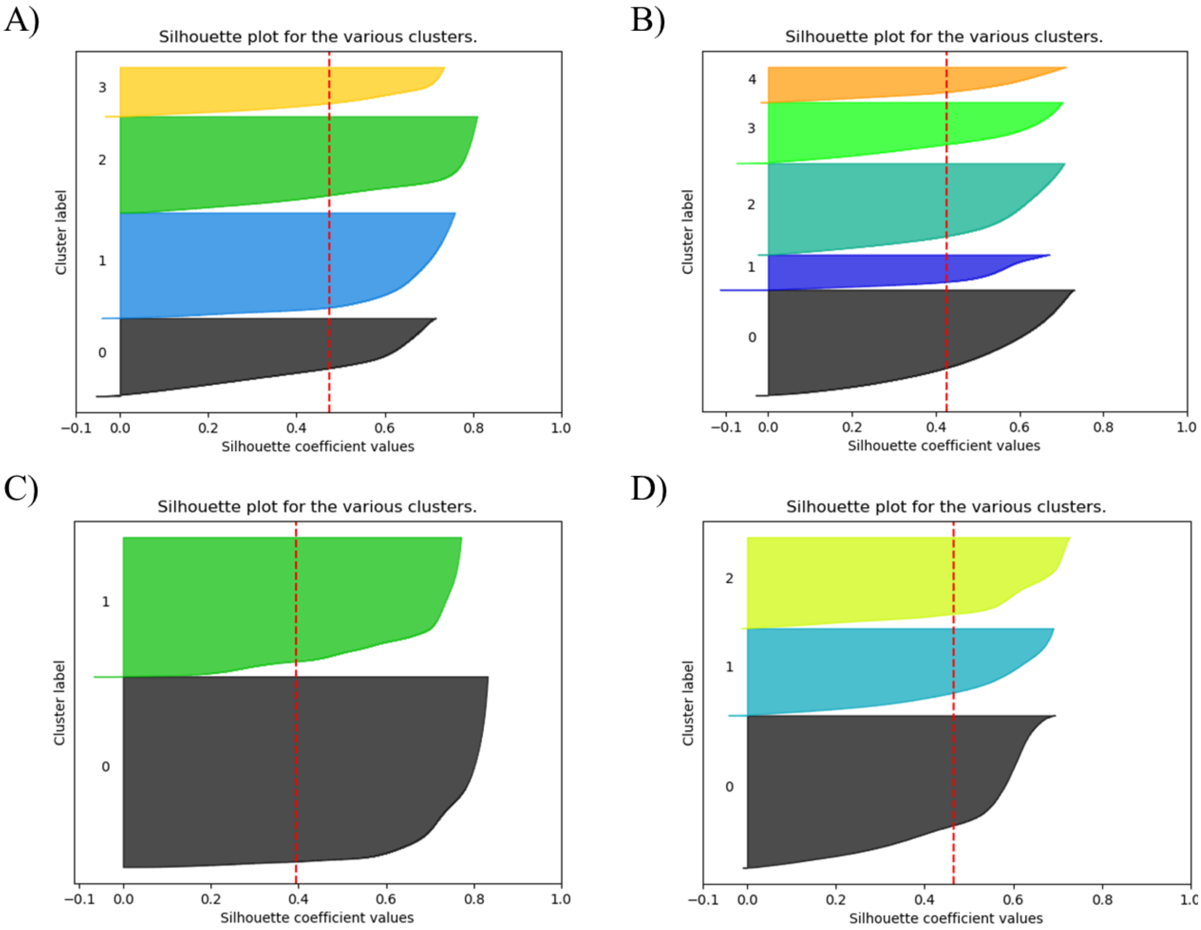

Figure B.2: **Silhouette plots for the selected number of clusters for each trajectory.** **A)** Silhouette plot for the four clusters of the whole RRM2 trajectory. **B)** Silhouette plot for the five clusters of Fragment A trajectory. **C)** Silhouette plot for the two clusters of Fragment B trajectory. **D)** Silhouette plot for the three clusters of the Fragment B trajectory corresponding to its unfolding.

Appendix C

Identification of regions able to bind aptamers

Figure C.1 shows the ten conformations found for Fragment A and B. Then we compute the molecular surface of each of these ten conformations and use ***Zernike*** to compute the shape complementarity between each one of its residue and all the other surfaces, one at a time. For each residue of each conformation, we sum the BPs obtained for all the pairings with the other surfaces. The results are shown in Figure C.2. The

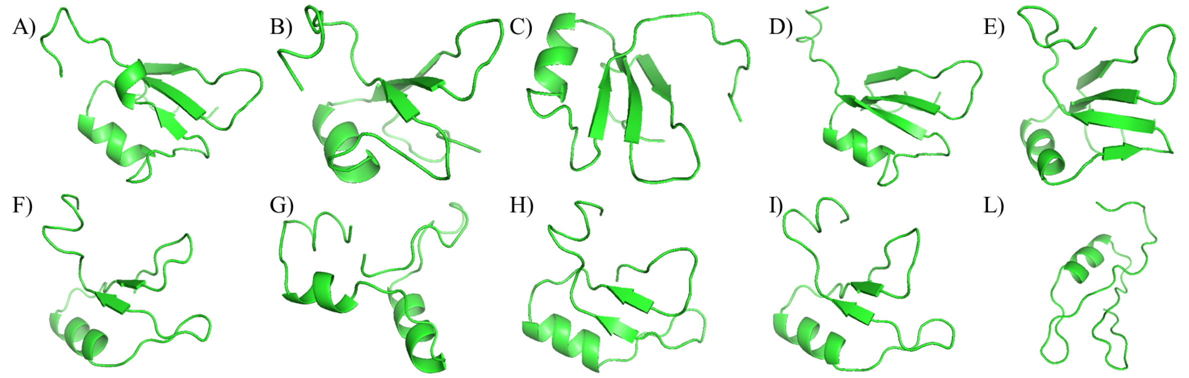

Figure C.1: **Conformations of RMM2 fragments proposed to be in CTFs aggregates.**
Plots from **A)** to **E)**: conformations A1, A2, A3, A4 and A5, representative of the clusters from the one labeled as ⓪ to the one labeled as ④ in Figure 7.4. **F)** and **G)**: conformations B1 and B2, representative of the clusters labeled as ⓪ and ① respectively in Figure 7.6. Plots from **H)** to **L)**: conformations B3, B4 and B5, representative of the clusters from the one labeled as ⓪ to the one labeled as ② in Figure 7.9.

residues associated to a non-null value in these plots are the ones that we searched for on each molecular surfaces with ***Chimera***.

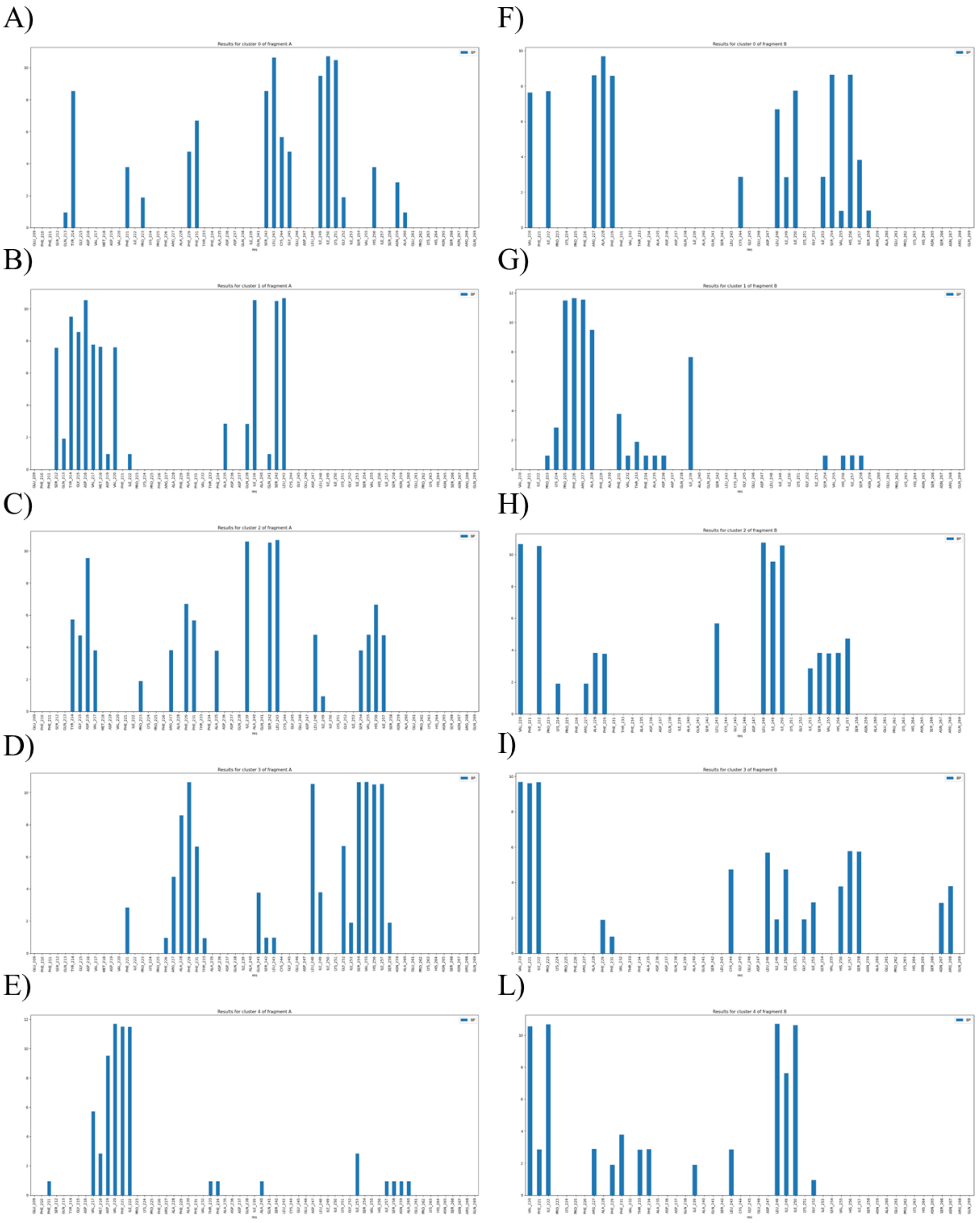

Figure C.2: **Sum of the BPs for the residues for each of the RRM2 equilibrium conformations.** Plots from **A)** to **E)**: Sum of the BPs calculated from the comparison with all the other surfaces of each residue of conformations from A1 to A5 respectively. Plots from **F)** to **L)**: Sum of the BPs calculated from the comparison with all the other surfaces of each residue of conformations from B1 to B5 respectively.

Bibliography

[1] X. Zuo, J. Zhou, Y. Li, K. Wu, Z. Chen, Z. Luo, X. Zhang, Y. Liang, M. A. Esteban, Y. Zhou, et al.*Tdp-43 aggregation induced by oxidative stress causesglobal mitochondrial imbalance in als.* Nature Structural & Molecular Biology 28, 132 (2021).

[2] M. Jo, S. Lee, Y. M. Jeon, S. Kim, Y. Kwon, and H.-J.Kim. *The role of tdp-43 propagation in neurodegenera-tive diseases: integrating insights from clinical and experimental studies.* Experimental & Molecular Medicine 52, 1652 (2020).

[3] Jiang, LL., Xue, W., Hong, JY. et al. *The N-terminal dimerization is required for TDP-43 splicing activity.* Sci Rep 7, 6196 (2017).

[4] Afroz, T., Hock, EM., Ernst, P. et al. *Functional and dynamic polymerization of the ALS-linked protein TDP-43 antagonizes its pathologic aggregation.* Nat Commun 8, 45 (2017). https://doi.org/10.1038/s41467-017-00062-0.

[5] Wang YT, Kuo PH, Chiang CH, et al. *The truncated C-terminal RNA recognition motif of TDP-43 protein plays a key role in forming proteinaceous aggregates.* The Journal of Biological Chemistry. 2013 Mar;288(13):9049-9057. DOI: 10.1074/jbc.m112.438564.

[6] Davide Tavella, Jill A. Zitzewitz, Francesca Massi. *Characterization of TDP-43 RRM2 Partially Folded States and Their Significance to ALS Pathogenesis.* Biophysical Journal, Volume 115, Issue 9, 2018, Pages 1673-1680, ISSN 0006-3495, https://doi.org/10.1016/j.bpj.2018.09.011.

[7] Prasad Archana, Bharathi Vidhya, Sivalingam Vishwanath, Girdhar Amandeep, Patel Basant K. *Molecular Mechanisms of TDP-43 Misfolding and Pathology in Amyotrophic Lateral Sclerosis.* Frontiers in Molecular Neuroscience 12 (2019) . 10.3389/fnmol.2019.00025.

[8] Berning Britt A., Walker Adam K. *The Pathobiology of TDP-43 C-Terminal Fragments in ALS and FTLD.* Frontiers in Neuroscience Volume 13 (2019). https://www.frontiersin.org/article/10.3389/fnins.2019.00335.

[9] Lindahl, Abraham, Hess, Spoel VD. *GROMACS 2020 Source code.* 2020. Available from: https://zenodo.org/record/3562495.YQBZFLris2w

[10] Nelson, R. and D. Eisenberg. *Recent atomic models of amyloid fibril structure.* Current opinion in structural biology 16 2 (2006): 260-5.

[11] Milanetti E, Miotto M, Di Rienzo L, Monti M, Gosti G, Ruocco G. *2D Zernike polynomial expansion: Finding the protein-protein binding regions.* Comput Struct Biotechnol J. 2020 Dec 4;19:29-36. doi: 10.1016/j.csbj.2020.11.051. PMID: 33363707; PMCID: PMC7750141.

[12] Prevedel, Robert and Diz-Munoz, Alba and Ruocco, Giancarlo and Antonacci, Giuseppe. (2019). *Brillouin microscopy: an emerging tool for mechanobiology.* Nature Methods. 16. 10.1038/s41592-019-0543-3.

[13] G. Grassmann, M. Miotto, L. Di Rienzo, F. Salaris, B. Silvestri, E. Zacco, A. Rosa, G. G. Tartaglia, G. Ruocco, E. Milanetti. *A computational approach to investigate TDP-43 C-terminal fragments aggregation in relation to Amyotrophic Lateral Sclerosis*, arXiv:2107.08931.

[14] Robberecht, W., Philips, T. *The changing scene of amyotrophic lateral sclerosis.* Nat Rev Neurosci 14, 248-264 (2013). https://doi.org/10.1038/nrn3430.

[15] Neudert, C., Oliver, D., Wasner, M. et al. *The course of the terminal phase in patients with amyotrophic lateral sclerosis.* J Neurol 248, 612-616 (2001). https://doi.org/10.1007/s004150170140.

[16] Shuo-Chien Ling, Magdalini Polymenidou, DonÂ W. Cleveland. *Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis.* Neuron 79, Issue 3, 416-438 (2013). https://doi.org/10.1016/j.neuron.2013.07.033.

[17] Brian C. Mackness, Meme T. Tran, Shannan P. McClain, C. Robert Matthews, Jill A. Zitzewitz. *Folding of the RNA Recognition Motif (RRM) Domains of the Amyotrophic Lateral Sclerosis (ALS)-linked Protein TDP-43 Reveals an Intermediate State.* Journal of Biological Chemistry, Volume 289, Issue 12, 2014, Pages 8264-8276, ISSN 0021-9258, https://doi.org/10.1074/jbc.M113.542779.

[18] E. Zacco, S.R. Martin, R. Thorogate, A. Pastore. *The RNA recognition motifs of TAR DNA-binding protein 43 may play a role in the aberrant self-assembly of the protein.* Front. Mol. Neurosci. 11 (2018) 32.

[19] Guenther EL, Ge P, Trinh H, et al. *Atomic-level evidence for packing and positional amyloid polymorphism by segment from TDP-43 RRM2.* Nature Structural Molecular Biology. 2018 Apr;25(4):311-319. DOI: 10.1038/s41594-018-0045-5.

[20] Ling, S. C., Albuquerque, C. P., Han, J. S., Lagier-Tourenne, C., Tokunaga, S., Zhou, H., et al. (2010). *ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS.* Proc. Natl. Acad. Sci. U.S.A. 107, 13318-13323. doi: 10.1073/pnas.1008227107.

[21] Huang, C. C., Bose, J. K., Majumder, P., Lee, K. H., Huang, J. T., Huang, J. K., et al. (2014). *Metabolism and mis-metabolism of the neuropathological signature protein TDP-43.* J. Cell Sci. 127(Pt 14), 3024-3038. doi:10.1242/jcs.136150.

[22] Vijay Kumar, Wahiduzzaman, Amresh Prakash, Anil Kumar Tomar, Ankit Srivastava, Bishwajit Kundu, Andrew M. Lynn, Md. Imtaiyaz Hassan. *Exploring the aggregation-prone regions from structural domains of human TDP-43.* Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, Volume 1867, Issue 3, 2019. Pages 286-296, ISSN 1570-9639. https://doi.org/10.1016/j.bbapap.2018.10.008.

[23] Brian S. Johnson, J. Michael McCaffery, Susan Lindquist, Aaron D. Gitler. *A yeast TDP-43 proteinopathy model: Exploring the molecular determinants of TDP-43 ag-*

gregation and cellular toxicity. Proceedings of the National Academy of Sciences Apr 2008, 105 (17) 6439-6444; DOI: 10.1073/pnas.0802082105.

[24] Zacco E, Grana-Montes R, Martin SR, de Groot NS, Alfano C, Tartaglia GG, Pastore A. *RNA as a key factor in driving or preventing self-assembly of the TAR DNA-binding protein 43.* J Mol Biol. 2019 Apr 5;431(8):1671-1688. doi:10.1016/j.jmb.2019.01.028. Epub 2019 Feb 8. PMID: 30742796; PMCID: PMC6461199.

[25] Paul F, Weikl TR (2016). *How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates.* PLoS Comput Biol 12(9): e1005067. https://doi.org/10.1371/journal.pcbi.1005067

[26] Csermely P, Palotai R, Nussinov R. *Induced fit, conformational selection and independent dynamic segments: an extended view of binding events.* Trends Biochem Sci. 2010 Oct;35(10):539-46. doi: 10.1016/j.tibs.2010.04.009. Epub 2010 Jun 11. PMID: 20541943; PMCID: PMC3018770.

[27] G.D. Quinlan, S. Tremaine. *On the reliability of gravitational n-body integrations.* Mon. Not. R. Astron. Soc., 259:505-518. 1992.

[28] H.C. Andersen, J. Chem. Phys. 72, 2384. 1980.

[29] D.J. Evans, G.P. Morris. *Statistical mechanics of non-equilibrium liquids.* Academic Press, London, 1990.

[30] H.C. Andersen. *Molecular dynamics at constant pressure and/or temperature.* J. Chem. Phys, 72: 2384-2939, 1980

[31] N.G. van Kampen. *Stochastic processes in physics and chemistry.* North-Holland, Amsterdam, 1981.

[32] Herman JC Berendsen et al. *Molecular dynamics with coupling to an external bath.* In: The Journal of chemical physics 81.8 (1984), pp. 3684-3690.

[33] Bussi G, Donadio D, Parrinello M. *Canonical sampling through velocity rescaling.* J Chem Phys. 2007 Jan 7;126(1):014101. doi: 10.1063/1.2408420. PMID: 17212484.

[34] Giovanni Bussi, Davide Donadio, Michele Parrinello. *Canonical sampling through velocity rescaling.* In: The Journal of chemical physics 126.1 (2007), p. 014101.

[35] Frances C Bernstein et al.*The Protein Data Bank: A computer-based archival file for macromolecular structures.* In: European journal of biochemistry 80.2 (1977), pp. 319-324.

[36] Moreira IS. *The role of water occlusion for the definition of a protein binding hot-spot.* Current topics Med Chem 2015; 15(20):2068-79

[37] Xue LC, Dobbs D, Bonvin AM, Honvar V. *Computational prediction of protein interfaces; a review of data driven methods.* FEBS Lett 2015;589(23):3516-26.

[38] Vakser IA. *Protein-protein docking: from interaction to interactome.* Biophys J 2014;107(8):1785-93.

[39] De Vries SJ, Bonvin AM. *How proteins get in touch: interface prediction in the study of biomolecular complexes.* Current protein peptide Sci 2008;9(4):394-406.

[40] Brender JR, Zhang Y. *Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles.* PLos Comput Biol 11(10).

[41] Teyra J., Hawkins J., Zhu H., Pisabarro MT. *Studies on the inference of protein binding regions across fold space based on structural similarities.* Proteins: Struct, Funct, Bioinformatics 2011;79(2):499-508.

[42] Daberdaku S, Ferrari C. *Antibody interface prediction with 3d Zernike descriptors and svm.* Bioinformatics 2019;35(11):1870-6.

[43] Kihara D, Sael L, Chikhi R, Esquivel-Rodriguez J. *Molecular surface representation using 3d Zernike descriptors for protein shape comparison and docking.* Current Protein Peptide Sci 2011;12(6):520-30.

[44] Zhu X, Xiong Y, Kihara D. *Large-scale binding ligand prediction by improved patch-based method patch-surfer2. 0.* Bioinformatics 2015;31(5):707-13.

[45] Venkatraman V, Yang YD, Sael L, Kihara D. *Protein-protein docking usingregion-based 3d Zernike descriptors.* BMC Bioinformatics 2009;10(1):407.

[46] Di Rienzo L, Milanetti E, Lepore R, Olimpieri PP, Tramontano A. *Superposition-free comparison and clustering of antibody binding sites: implications for the prediction of the nature of their antigen.* Sci Rep 2017;7(1):1-10.

[47] Di Rienzo L, Milanetti E, Alba J, DâAbramo M. *Quantitative characterization of binding pockets and binding complementarity by means of Zernike descriptors.* J Chem Inform Model 2020;60(3):1390-8.

[48] S. Hayward and N. Go. *Collective variable description of native protein dynamics.* Annu. Rev. Phys. Chem., 46 223-250 (1995).

[49] P. Robustelli, S. Piana, David E. Shaw. *Developing a molecular dynamics force field for both folded and disordered protein states.* Proceedings of the National Academy of Sciences May 2018, 115 (21) E4758-E4766; DOI: 10.1073/pnas.1800690115

[50] B.R.Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. *CHARMM: Thebiomolecular simulation program.* Journal of Computational Chemistry 30, 1545 (2009)

[51] Jorgensen, William & Chandrasekhar, Jayaraman & Madura, Jeffry & Impey, Roger & Klein, Michael. (1983). *Comparison of Simple Potential Functions for Simulating Liquid Water.* J. Chem. Phys. 79. 926-935. 10.1063/1.445869.

[52] M. Parrinello and A. Rahman. *Crystal structure and pairpotentials: A molecular-dynamics study.* Physical Re-view Letters45, 1196 (1980)

[53] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M.Fraaije. *LINCS: A linear constraint solver for molecular simulations.* Journal of Computational Chemistry 18, 1463 (1997)

[54] T. E. I. Cheatham, J. L. Miller, T. Fox, T. A. Darden,and P. A. Kollman. *Molecular dynamics simulations onsolvated biomolecular systems: the particle mesh ewald method leads to stable trajectories of DNA, RNA, and proteins.* Journal of the American Chemical Society117,4193 (1995)

[55] Richards FM. *Areas, volumes, packing and protein structure.* Annu Rev Biophys Bioeng. 1977;6:151-76. doi: 10.1146/annurev.bb.06.060177.001055. PMID: 326146.

[56] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004). *UCSF Chimera- A visualization system for exploratory research and analysis.* J. Comput. Chem., 25: 1605-1612. https://doi.org/10.1002/jcc.20084

[57] Bellucci M, Agostini F, Masin M, Tartaglia GG. *Predicting protein associations with long noncoding RNAs.* Nat Methods. 2011 Jun;8(6):444-5. doi: 10.1038/nmeth.1611. PMID: 21623348.

[58] Heinig M, Frishman D. *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.* Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W500-2. doi: 10.1093/nar/gkh429. PMID: 15215436; PMCID: PMC441567.

[59] L. Brillouin. *Diffusion de la lumiere et des rayons X par un corps tranparent homogene-Influence de l'agitation thermique.* In: 9.17 (1922), pp. 88- 122.

[60] R. Prevedel et al. *Brillouin microscopy: an emerging tool for mechanobiology.* In: Nature methods 16.10 (2019), pp. 969-977.

Acknowledgments

My grateful thanks to Professor Armando Bazzani, who followed and encouraged me during this period, and to all the IIT collaborators in Rome and Genoa, whose availability and interest in this project have been incredibly effective for its development and for the growth of my passion for this field.

In particular my gratitude goes to Professor Giancarlo Ruocco, for introducing me to this project and for his incredible kindness and helpfulness: he has always been available whenever I had a doubt or curiosity. Not only he opened me a window on the world I want to be part of in the future, but he also managed to make me already feel a part of it during my internship.

Special thanks to Dr. Edoardo Milanetti and Dr. Mattia Miotto: they were always available to spend their time discussing with me, and taught me that when you are studying something you are passionate about, sleeping is not that important anymore. I hope one day to be able to work with a group as great and stimulating as theirs.

Thanks to Professor Gian Gaetano Tartaglia, Dr. Claudia Testa, Dr. Beatrice Silvestri and Dr. Alessandro Rosa, whose kindness and availability to answer all my questions made easier my first step into new topics.

Thanks to Professor Nico Lanconelli, who trusted my work and encouraged me to publish my first paper, and to the Camplus managers Dr. Guidetti and Dr. Bonafede, who helped me a lot and made my life much easier in Bologna and Rome.

Finally, thanks to my family for the continuous support: I will do my best to reciprocate all you have done for me. Thanks to Antonio, for the happiness he brought me even in the worst times. And of course thanks to all my friends, both the ones still around and those far away (as the wise says, *such is life*). From the stairway in Trieste, to the porch

in Bologna and the flying-rock in Rome, thank you for helping me to build such great memories.