Academic Year 2022-2023
Computational Biophysics [145546] - Proff. G. Lattanzi, L. Tubiana

Structural and functional analysis of Sars-Cov-2 main protease, wild type and H172Y mutant

Alessia Guadagnin Pattaro[1], Teresa Dalle Nogare[2]

Abstract
This work focuses on the study of the H172Y mutation of Sars-Cov-2 main protease (Mpro) to investigate differences from two wild-type structures, namely 6XHU and 7VH8. Both structural and statistical analyses were performed on trajectories obtained from MD simulations.
Concerning the structural part of the analysis, configurational distances such as the RMSD, radius of gyration, and RMSF were computed to monitor the system's behaviour. Further insights were gained by combining these qualitative results with the ones obtained from the second part of the analysis. The degree of correlation in each trajectory was established by both fitting the autocorrelation function and exploiting a block analysis, obtaining different autocorrelation times for the mutant and WTs. From the spectral analysis, it emerged that the fluctuating part of the RMSD resembled pink noise. Clustering and PCA were ultimately implemented to highlight the presence of any emerging pattern from the data, related to structural features. Very few quantitative conclusions could be obtained mostly because of the poor sampling that could be achieved with the available resources.
The starting PDB structures, the code and the Jupyter Notebooks used for this project are available in a dedicated Github repository: https://github.com/alessiagp/8D4JProject

[1] *Master's degree in Quantitative and Computational Biology, University of Trento*
[2] *Master's degree in Physics, University of Trento*

Introduction

1.1 The main protease of Sars-Cov-2

Severe acute respiratory syndrome coronavirus-2 (Sars-Cov-2) is a single-stranded RNA virus responsible of the 2020 global pandemic. Its viral pathogenesis results in symptoms such as fever, acute pneumonia, and respiratory failure. The virus was a threat for global health for its high death rate and its long time needed for recovery, effects enhanced if a patient already had debilitating conditions, was in advanced age, or in particular life stages (i.e., pregnancy). The Sars-Cov-2 genome, from a phylogenetic point of view, has many similarities with the Sars-Cov genome and this was very helpful to try and immediately contrast the diffusion of the virus: the inhibition targets that were found to be effective against Sars-Cov during the 2003 epidemics could be effective, by similarity, also against Sars-Cov-2 [1].

One of the primary pharmacological targets that were investigated against the two Sars-Cov viruses was their main protease (Mpro), needed for the synthesis of functional polypeptides useful for viral replication [2] by hydrolisation of the peptide bond between the aminoacids.

Sars-Cov-2 Mpro is a cysteine hydrolase of the 3CL type composed of two protomers, A and B. The protein has the highest hydrolytic activity in a dimeric form. Three subdomains can be identified in the monomers:

- domain I (residues 8–101)
- domain II (residues 102–184), characterised by an antiparallel β-barrel structure
- domain III (residues 201–303), with a five-fold antiparallel α-helix structure

The rendering in Fig. 1 shows that domain II and domain III are connected via a loop region (residues 185–200). Aminoacids between 140–146 are called ”oxyanion loop” and are useful to stabilise the transition state of the enzymatic reaction (highlighted in red in 1). The catalytic site is located at the centre of the cleft region between domains I and II and it is due to the dyad His41-Cys145, where the sulphur of the Cystein acts as a nucleophile towards substrates and the imidazole moiety of the Histidine acts as a general acid-base catalyst. The native state of the active site is the thiolate-imidazolium ion pair. The active region can be divided into five subpockets (S1, S2, S3, S4, and S5) [3–5].

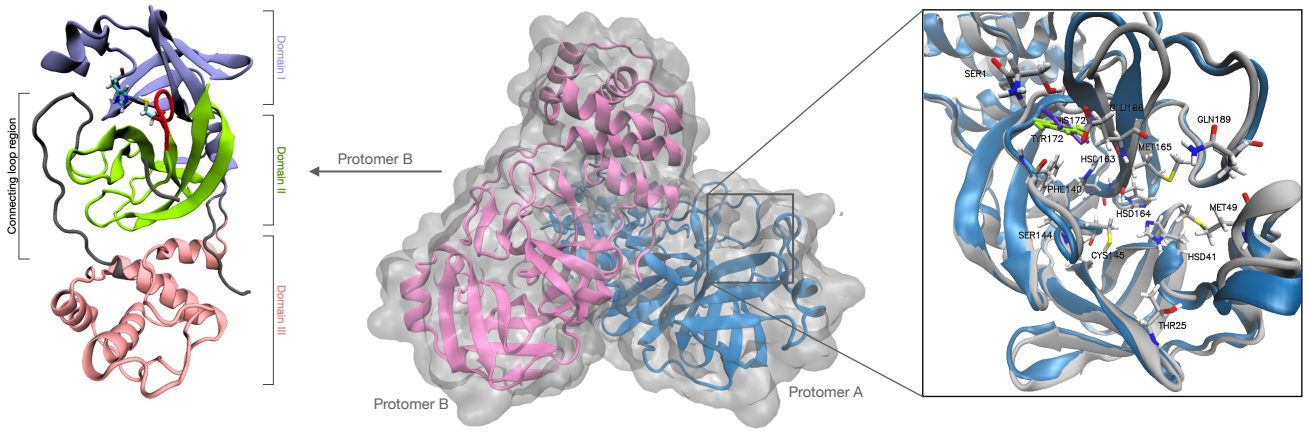


Figure 1. Rendering of Mpro (PDB: 6XHU) showing: (left) protomer B of 6XHU with the three domains. The catalytic dyad and the oxyanion loop (in red) are also reported; (center) full protein 6XHU highlighting chain 1 in blue and chain 2 in pikn; (right) magnification of the binding site in chain 1. His172 is colored in purple while Tyr172 is colored in green.

The S1 pocket of one monomer A is stabilised by the interaction between Glu-A166 and His-A172, and between Glu-A166 and Ser-B1, which is the N-terminus of protomer B (also called N finger). The N finger of B protomer also interacts with Phe-A140 that is stabilised by His-A163 via aromatic ring-stacking [6]. The reasons for this structural arrangement are to be found in the formation of the quaternary structure of the protease, released after the autoproteolysis reaction with its substrates: the reaction occurs in a *trans* conformation and it allows the positioning of the N-termini near the active sites of the opposite monomer [3].
It is important to highlight that the two monomers of the protease have different hydrolytic activity within the catalytic site [2]: in the active site of protomer B at pH 6, the aromatic ring-stacking between Phe-B140 and His-B163 is not present, since His-B163 now interacts with Glu-B166, preventing the noncovalent bonds with His-B172, and the N finger of protomer A. This results in a rearrangement of the tertiary structure of domains I and II of protomer B, causing a collapse of its S1 active site. The activity of the protease depends on pH (pH-dependent activation switch), it has been shown that at pH 7.3-8.5 the protease has its maximum activity and both the active sites can act as catalysts. This means that Sars-Cov Mpro can regulate its proteolytic activity in the different microenvironments present in the cell [6].

1.2 The H172Y mutation of Sars-Cov-2 Mpro

Mutations in viruses occur as RNA replication happens. Most of the time mutations do not affect the pathogenesis of the virus and go undetected, or lead to internal instabilities and, ultimately, to the virus destruction, or they can increase the virus replication and, for this reason, proliferate. Mutations are also relevant because they allow the virus to go undetected by the host’s immune system once its RNA has been changed.

Several structures of novel mutations of Sars-Cov-2 main protease have been published in September 2022 by the research group of Hu *et al.*. These mutations were naturally occurring and were found to increase the drug resistance against several Mpro inhibitors, along with increased protease activity. There were 100 mutations in total, 20 of which were located in residues near the active site. This report focuses on mutation H172Y, which alters the interactions between the N-terminus of the opposite protomer and the active site.

H172Y is a mutation involving the histidine at position 172 being substituted by tyrosine in both protomers. As shown in Fig. 2, the two amino acids are different, His has an imidazole residue, which is a basic heterocyclic aromatic compound. Tyr has an aromatic phenol residue, whose -OH group gives the amino acid a mild acidity and allows a larger delocalisation of the negative charge of the aromatic ring. Tyr has also a larger steric hindrance than His. This mutation is not very frequent but it has been reported that it reduces the enzymatic activity of Mpro

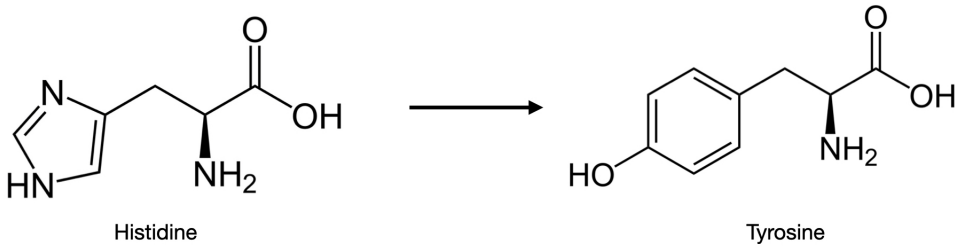


Figure 2. Graphical representation of the aminoacids that vary in the H172Y mutation of Sars-Cov-2 Mpro

(a 13.9-fold lower K_{cat}/K_m with respect to WT is observed) and it increases the pharmacoresistance against drug Nirmatrelvir (233-fold increase in K_i with respect to WT). This has been confirmed experimentally by plaque assay and viral growth kinetics. [7] H172Y mutation is close to the S1 subpocket of the active site and interacts with the aromatic ring-stacking and with the N-finger of the opposite protomer, disrupting its interaction with the active site. The aromatic ring-stacking is an important constituent of the S1 pocket conformation. The research group of De Oliveira [8] has explored this mutation in detail employing molecular dynamics simulations, they confirmed the disrupting of the aromatic ring-stacking and the partial collapse of the oxyanion loop, and the loss of the interaction between Phe140 and the N-finger. These result in a change of interactions between the protease and its substrate, and consequently to an increased pharmacoresistance. The group also reported the unfolding temperatures of the wild-type protease (T_m=57.9°C) and the mutated protein (T_m=53.7°C): a lower unfolding temperature is related to less energy needed to disrupt the structure, hence the structure is less stable. These results are consistent with the simulations. De Oliveira also corroborated experimentally the reduction of the binding affinity of Nirmatrelvir for the mutated protease. Overall, the H172Y mutation is not thought to become dominant since the protein activity is so altered despite having high resistance against drugs.

1.3 Aim of the project

This report presents the complete analysis of Sars-Cov-2 Mpro for two wild-type structures and the assigned mutant structure. This analysis aims to observe any differences in the structure that could give rise to different behaviours between the two protomers, and between the wild-type and the mutant protease. The report is composed of two parts: a structural analysis, focusing on the chemical interactions between residues with a focus on the active site, and a functional analysis, which includes a statistical evaluation of the observables involved in the first part. Special attention was also given to the spectral analysis of the time series, which gave some insights into some intrinsic properties of the system. Ultimately, some unsupervised methods-based analyses were performed to retrieve some information about the principal modes of the protease.

Computational methods

2.1 Structures

Three different structures were used in this project:

- 8D4J: H172Y mutant structure, investigation target, from Hu Yanmei paper [7]
- 6XHU: wild type structure, from Hu Qing paper [5]
- 7VH8: wild type structure, from De Oliveira paper [8]

Two different wild-type structures were chosen since both of them were used in papers taken as references. None of them was discarded after having performed the simulation since they both showed meaningful behaviours. While 7VH8 exhibited a stable structure with no significant motions, 6XHU seemed to agree with the pH-dependent activation switch reported in the Introduction.

Table 1. List of residues composing the active site of Sars-Cov-2 Mpro, valid for both WT and mutant structures. (*) Tyr172 for 8D4J

Subpocket	Residues
S1	Phe140, Gly143, Ser144, Cys145, His163, Glu166, His172(*)
S2	Thr25, His41, Cys145
S3	His41, Met49, Met165
S4	Met165, Glu166
S5	Glu166, Met165, Gln189
N-finger	Ser1 (opposite protomer)

Each protease is divided into two protomers (CH1 and CH2). CH1 refers to the first 306 residues appearing in the `.pdb` file, from Ser1 to Gln306. 7VH8 is considered as a whole, so CH1 refers to residues Ser1 to Gln306, CH2 refers to Ser307 to Gln612. For all the structures, the full protein structure will be referred to as FP.

The spotlight of the project is on the residue H172Y, which is located close to the protease active site. The residues composing the active site have been listed by [5] and are reported in Table 1

7VH8 PDB structure had to be edited with CHIMERA software and the bash shell because the initial file involved one chain only, and a ligand. The resulting structure correctly resembled the protease Mpro.

2.2 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulations were performed with GROMACS 2021.5 on the University HPC Cluster [9]. Before launching the simulations, a benchmark was performed for each step. Benchmarks are useful because they allow the optimisation of the number of cores used in the process. Short runs of 10 minutes each were performed for each simulation step, and the number of cores was plotted against the velocity of the simulation calculated in ns/day. As it is shown in Figure (3) the number of cores chosen for the NVT and NPT run was 60 cores, for the production run 80 cores were selected. The reason for this choice was in the cluster infrastructure - and especially in the queue that was employed: no simulation with a number of cores greater than 80 started.

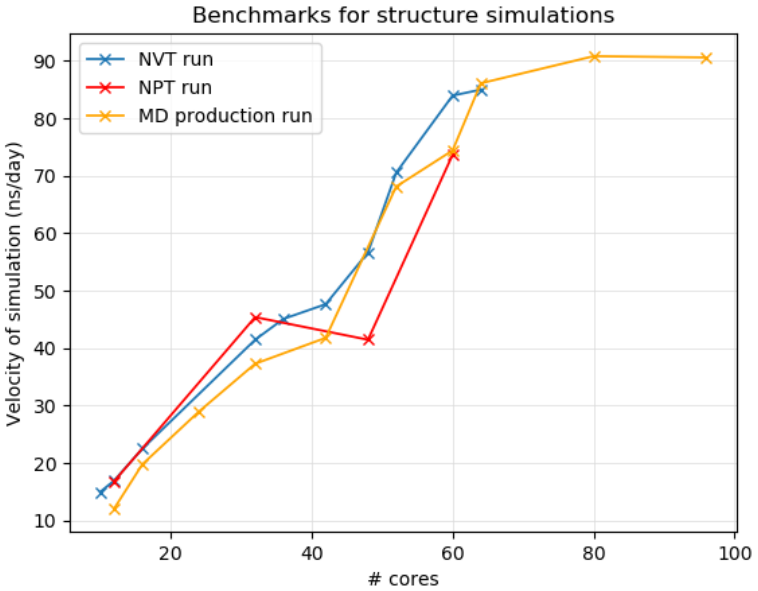


Figure 3. Benchmark of the NVT, NPT and production run used to determine the number of cores to run simulations

8D4J and 6XHU structures setups were generated with the web server CHARMM-GUI, developed by the research team of Dr. Im at Lehigh University, Bethlehem [10] while the 7VH8 system, for issues in the starting PDB file, was built locally using GROMACS as it was not recognized correctly by CHARMM-GUI. Both 6XHU and 7VH8 were solvated in a box with dimensions as similar as possible to the provided 8D4J input file. The simulations were all performed with a fixed box length of 10.9 nm, an ion concentration of NaCl 0.15 M, at a temperature of 310 K. The

force field used was CHARMM36FF.

The steepest descend algorithm was used to perform the energy minimisation and for all simulations the maximum force was lower than 1000 $\mathrm{kJ\,mol^{-1}\,nm^{-1}}$. The temperature equilibration step was operated with the V-rescale algorithm for 0.5 ns, this step also allowed conformational sampling of the velocities from the Maxwell distribution at 310 K with `gen-vel=yes` parameter. The pressure equilibration step used the V-rescale thermostat at 310 K, and the Parrinello-Rahman barostat at 1 atm for 1 ns. The production run used the same parameters as the pressure equilibration step, all production runs were performed for 600 ns. Positions and velocities were written every 50 ps and the integration step was every 2 fs. In total, one simulation for each WT structure, and three simulations for 8D4J were performed.

Since the three 8D4J simulations are analysed separately, they will be referred to, from now on, as 8D4J-1, 8D4J-2, and 8D4J-3.

Base parameters By using GROMACS, one can have some insights into the so-called base parameters of the system, such as potential energy, temperature, and density.

Potential energy (Figure 4a) is referred to the energy minimisation step and highlights the number of steps needed to reach the minimum potential energy. All the 8D4J simulations started from the same initial `.pdb` structure, hence they all took 160 steps to converge, while WT 7VH8 and WT 6XHU took a slightly longer time. For all, the short convergence time may highlight that the structures were obtained in a quite energetically favourable conformation, that is coherent with the experimental technique used to obtain the structures (X-ray crystallography).

Temperature (Figure 4b), referred to as the thermalisation step of the simulation procedure, oscillates around the selected value of 310 K for all the structures. Only one trajectory was taken as an example for 8D4J.

Density (Figure 4c) is related to the third step of the simulation process. It measures the density of the system as a whole, considering both the solvent and the protein structure. All the structures have a value of density around 1035 $\mathrm{kg\,m^{-3}}$ that confirms the pressure equilibration of the system.

2.3 Data analysis

Molecular graphics and visual inspections were performed with the Visual Molecular Dynamics (VMD) software [11, 12] while structural and statistical analyses were performed with the Python library MDAnalysis [13, 14].

Results and discussion

3.1 Structural analysis

To investigate similarities and differences among the wild-type (WT) Sars-Cov-2 Mpro and the H172Y mutant structure, and among the two monomers composing each protein, an introductory structural analysis was performed. The analysis aimed firstly to highlight possible differences between the WT and 8D4J, and ultimately to give some new insights about this biological system.

3.1.1 Visual inspection of the structures with VMD

For the structural characterisation of the protease, the relevant residues belonging to the active site are considered. The preliminary analyses involve the catalytic dyad His41 - Cys145, the aromatic ring-stacking Phe140-His163, its interaction with His172/Tyr172, and Ser1 of the opposite protomer. The `RMSF_visualisation.tcl` script on VMD was useful to represent fluctuations over the protein structure. Of all structures, both the mutant and the wild type(s), high fluctuations were observed for the C-termini, for loops in domain III, and for the α-helix structure in domain I. Since fluctuations in the C-termini were very high compared to the rest of the protein, residues 303-306 were excluded from the RMSF calculations to highlight other relevant changes in the protease, except for WT 6XHU. Table 2 reports the most fluctuating residues for each chain of each structure to have a general overview of the system.

Wild types For **WT 7VH8** structure (Figure 5a), the highest fluctuating parts are in subdomains I and III. The residues that fluctuate the most are not close to the active site except for residues between Thr45-Asn65 for CH1, and Thr45 - Pro52 for CH2, which are partly located in the S3 subpocket, as reported in Table 2. On average, the residues with high fluctuations are located in the same areas of the two chains.

Figure 5b reports **WT 6XHU** structure. As for WT 7VH8, the most fluctuating residues are in subdomains I and III. One difference that can be found compared to the other wild type is that the C-terminus is not as fluctuating, thus residues 303-306 are accounted for in the calculation. Both the two chains fluctuate the most in residues

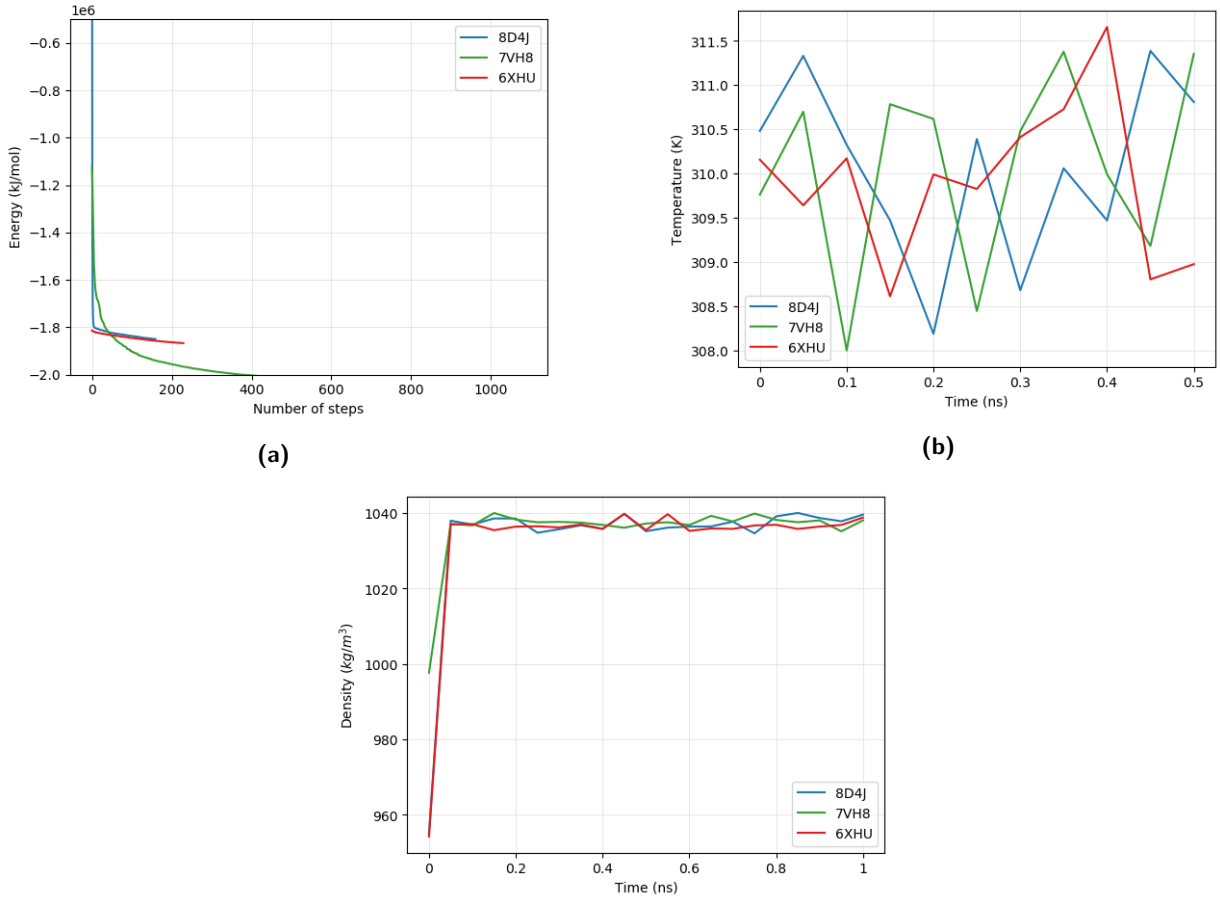
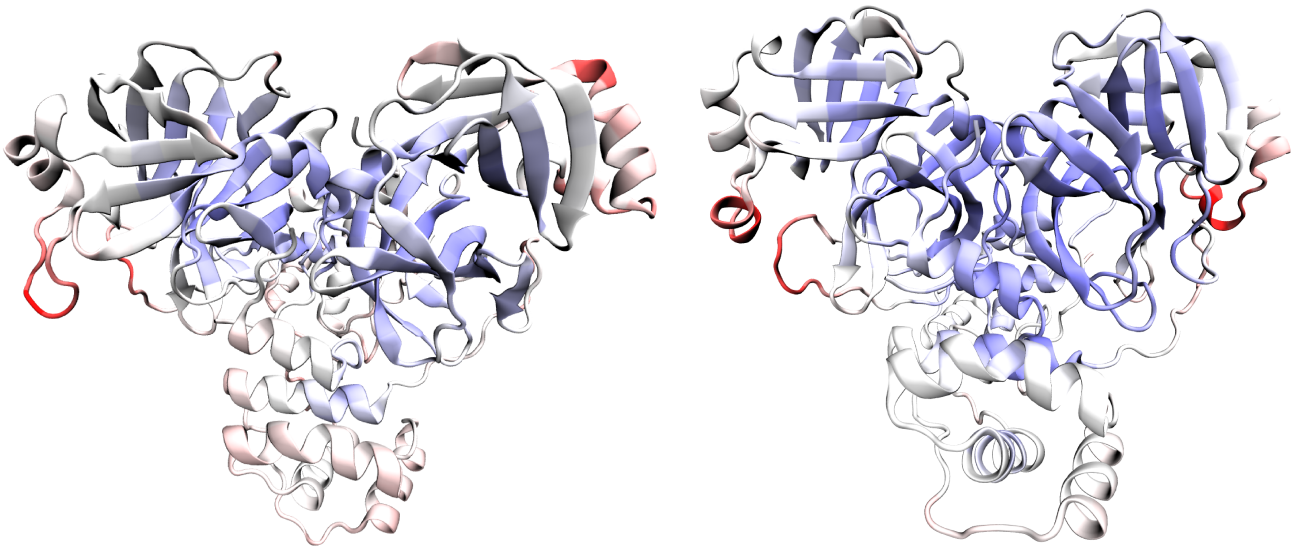


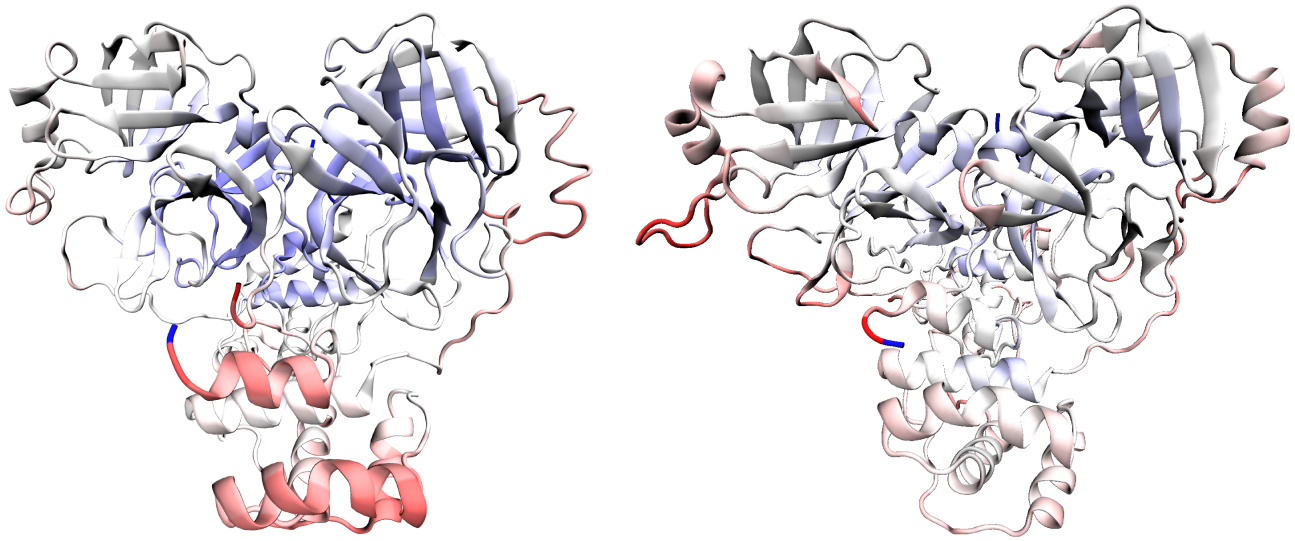
Figure 4. Base parameters calculated during the simulation of both WT, 7VH8 and 6XHU, and mutant 8D4J : **(a)** energy minimization, **(b)** temperature, and **(c)** density.



(a) 7VH8 **(b)** 6XHU

Figure 5. Renderings of wild type structures 7VH8 and 6XHU, coloured according to the most fluctuating residues.

Thr45-Pro52 and Phe185-Ala194; CH1 with a lower magnitude than CH2. From the visualisation of the trajectory in VMD, it is shown that residues Leu50-Asn51 interact repulsively with Gln189 and the backbone of Arg188 around 450 ns that cause the protein segment Thr45-Pro52 to ”flap”. This behaviour is predominant in CH2. A particular thing happening in this simulation is the breakage of the ring-stacking interaction that occurs more frequently in CH2. This is probably caused by a change of interactions of the N-terminus from Phe140 and His172 to Glu166. The ”flapping” movement also prevents the functionality of the catalytic dyad since they move away from each other. Interestingly, whenever this happens in one chain, in the other chain they are close enough to interact.



(a) 8D4J-1 **(b)** 8D4J-3

Figure 6. Renderings of 8D4J for simulations 1 and 3, coloured according to the most fluctuating residues.

H172Y Fluctuations in **8D4J** were consistent in the three performed simulations. In particular, the N-terminus of CH1 moves significantly more than the other termini (especially in the first 200 ns) because of the interactions with the S1 subpocket of CH2. The motion is also due to the residues 189-196, located in the connecting loop, and mostly due to the residues in subdomain III. In all simulations of the mutant, the disruption of the interaction between the N-finger Ser1* and the ring-stacking Phe140-His163 by the presence of the mutation Tyr172 is observed. The second chain of 8D4J-2, in the final part of the trajectory, shows large interactions between residues Ser46-Asn53 and the C-terminus, similarly to 6XHU.

The particular active site behaviour (explicit separation of the ring-stacking and catalytic dyad breakage) that was seen for WT 6XHU has not been observed in any of the mutant simulations.

3.1.2 Analysis based on configurational distance measures

A first analysis of the simulated trajectory was performed both for the wild type and mutant structure to gain insights on the relevant time scales that characterise both systems. Thus, observables such as the radius of gyration and RMSD were inspected to acknowledge the presence of transition states or equilibrated configurations. It is worth mentioning that a first semi-quantitative analysis based on the evaluation of configurational distance measures can not assess the presence of an equilibrated state for the system under investigation.

After aligning the trajectory to the first frame, both the RMSD and radius of gyration were computed. A transient behaviour representing the relaxation of the initial configuration towards more representative ones has been discarded through a visual inspection of these parameters. The equilibration time, namely the temporal coordinate that separates the equilibration from the production data, is chosen to be short enough to avoid the discard of too much data but, at the same time, long enough to remove biased initial configurations. The first frame of the production data was then used as a reference for the alignment of each structure.

Because of the limited duration of simulations, it doesn’t make complete sense to refer to the production data as the equilibrated part of the trajectory. Indeed, according to Grossfield *et al*[15], the notion of separating a trajectory into its equilibration and production parts is meaningful only if the system has reached important configurations in the equilibrium ensemble. However, since it is impossible to guarantee whether this has occurred, the only way to

Table 2. Table with the most fluctuating residues of both wild type and mutant simulations with the exclusion of the C-termini.

Structure	CH1	CH2
6XHU	Thr45-Leu58 Phe185-Ala194	Thr45-Pro52 Phe185-Ala194
7VH8	Thr45-Asn65 Asp187-Ala193	Thr45-Pro52 Val186-Gln192
8D4J-1	Thr45-His64 Gln189-Thr196	Thr45-Asn63 Arg188-Ala193
8D4J-2	Cys44-Asn53 Leu167-Gly170 Gln189-Ala193	Ser46-Asn53 Ile59-His64 Arg188-Ala193
8D4J-3	Thr45-Asn53 Leu167-Val171 Val186-Ala194	Cys44-Asn63 Leu167-Val171 Arg188-Ala194

assess it is via a series of *configuration-space distances*, namely scalar functions that quantify the similarity between two molecular configurations.

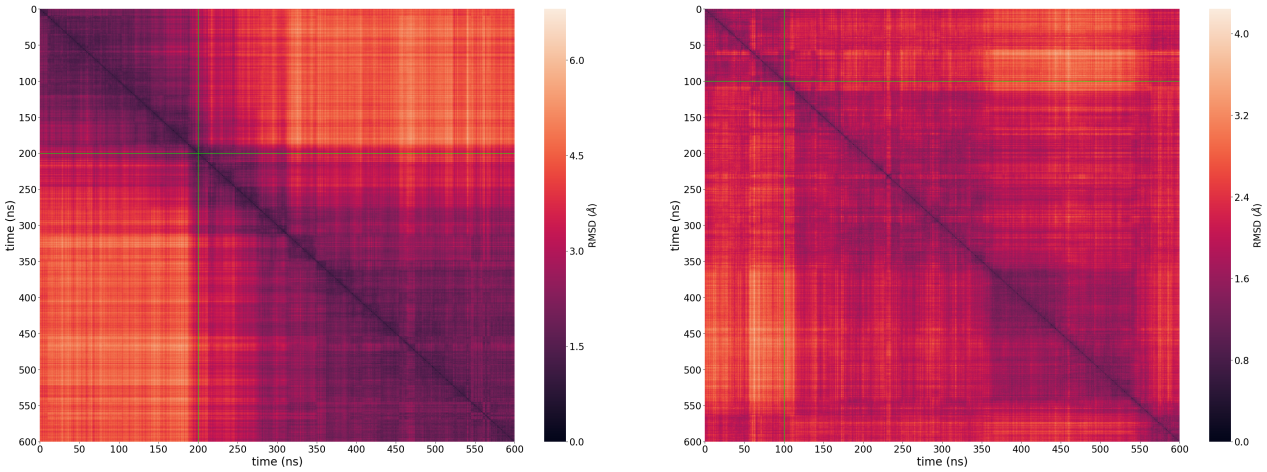
RMSD and pairwise RMSD of production data The root-mean-square deviation (RMSD) is a configuration distance commonly employed to quantify differences between conformations in MD simulations. In the framework of molecular dynamics, RMSD computed as a function of time expresses the deviation of each atomic position from the one in the reference structure.
RMSD was measured for a selection of α-carbon atoms both for the mutant protein and wild-type to filter higher-order fluctuations. However, before computing this quantity, a structural alignment to the first frame of the production data was performed. For each trajectory, the reference frame was chosen through a visual inspection of both the pairwise RMSD and the 1D RMSD values measured considering the overall initial frame as a reference.
The pairwise RMSD is useful to identify similar structures throughout the trajectory since it represents the RMSD of each snapshot of the trajectory computed with respect to all the others. In each map (Fig. 7) the green line indicates the separation between equilibration data, which are discarded from further analysis, and the production ones. The origin of time is then shifted to match the first frame of the trimmed trajectory
The matrix in Fig. 7a, which reports the first simulation of the mutant 8D4J, shows a transition occurring at about 200 ns between two metastable states where the protein has low RMSD values. According to Fig. 8, on average the RMSD fluctuates around 3.5 Å, although the rising trend observed in the first 100 ns suggests that the aforementioned transition has not been completely discarded. Moreover, in the range 100-400 ns, the average RMSD value of 8D4J-1 (3.5Å) is higher compared to the average RMSD of the other two simulations (2.1 Å and 2.3 Å) and of the WT 7VH8 (1.8 Å), suggesting a possible conformational change in the former structure.

For what concerns **8D4J-2** (Figure **S3**), the structure is initially in a metastable state approximately 480 ns long where different smaller configurations are constantly revisited by the protein. The final block in the RMSD matrix (480 ns - 600 ns) hints at a shift to a new conformation that does not reach an equilibrium state in the simulated time since the average RMSD value has a steep increase from 2 Å to approximately 2.5 Å with high fluctuations. This behaviour raised doubts about cutting the simulation both at the beginning and at the end to keep the part of the trajectory in the bigger state but this decision was discarded [1] and only the first 100 ns of the trajectory were trimmed.

For **8D4J-3**, reported in Figure 7b, the reason for the cut came naturally after looking at the RMSD plot and heatmap. The equilibration part of the run is comprised of up to 100 ns. From 100 ns to 550 ns, the protein enters a metastable state that, as it was for 8D4J-2, is composed of smaller substates that are revisited. These substates can be identified by slight variations in RMSD:

1. 100-350 ns, with average RMSD of 2.35 Å
2. 375-450 ns, with average RMSD of 2.75 Å

[1] Teresa commented that it would have been like "hiding behind a hat"



(a) 8D4J - first simulation **(b)** 8D4J - third simulation

Figure 7. Pairwise RMSD map of the **8D4J-1** simulation (C_α selection). The green line in both figures indicates the equilibration data that were discarded in the rest of the analysis.

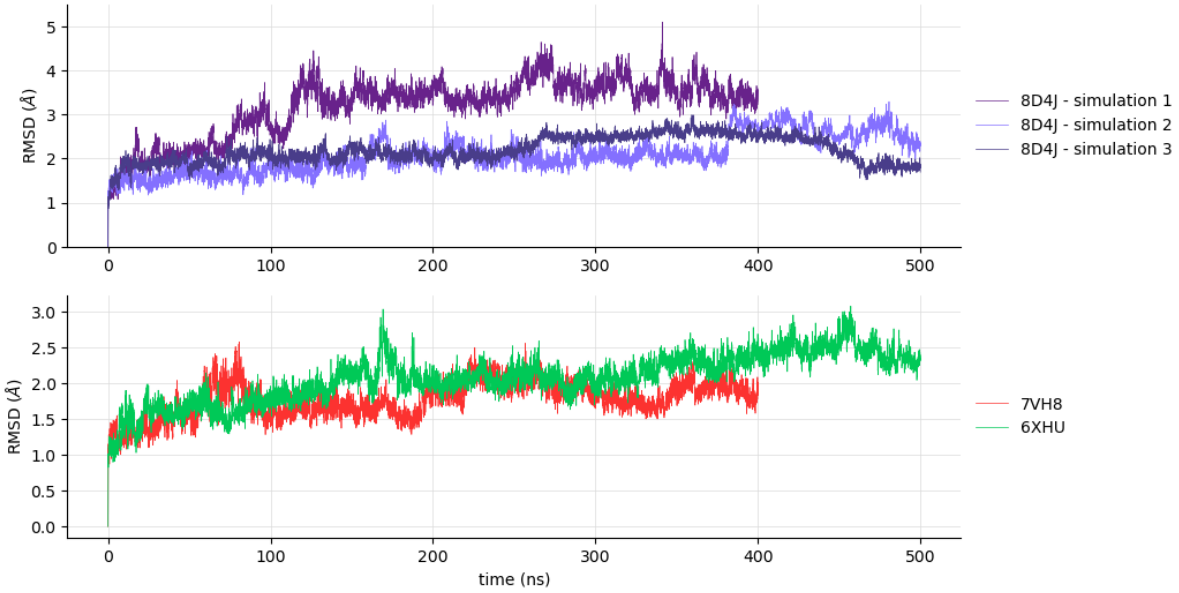


Figure 8. Plot showing RMSD for the two wild types and all three simulations of the mutant for C_α selection.

3. 450-550 ns, RMSD has a descending trend from about 3.0 Å to about 2.0 Å

These states are visited interchangeably, with the lowest values of RMSD - probably indicating a particularly favourable conformation - found in state 2. The transition at the end of the RMSD plot can be seen in the map, confirming that the protein has not reached equilibrium yet.

Regarding the **WT 7VH8** (Figure **S2**), a rising trend is observed in the first 75 ns of the production run. This hints that the trajectory is not equilibrated yet at 100 ns. RMSD pairwise matrix does not show any metastable states of relevance. The ”darkest spot on the map” is at about 175-250 ns and, by comparing it with the RMSD plot, it can be regarded as a transition state. Also in this case, equilibration cannot be easily assessed since an oscillating behaviour is observed in the trajectory.

WT 6XHU (Figure **S1**) has a continuously rising trend in RMSD with a very first equilibration step (first 50 ns), a steady trend of RMSD at 1.8 Å until 240 ns, and a quite oscillating trajectory until the end, with a first rise to 2.0 Å and peaks up to 3.0 Å. The pairwise RMSD map does not highlight any particular metastable states, at about 260 ns a highly unstable cross can be observed, meaning that the protein is changing abruptly its conformation. This can

be referred to as the first ”flapping” motion reported in visual inspection. The reason for trimming the trajectory at 100 ns was the right balance between the exclusion of the equilibration step only (first 50 ns) and the complete exclusion of the first quasi-equilibrated part of the trajectory (cut at 240 ns) leaving only the heavily oscillating part.

Radius of gyration of production data The radius of gyration R_g is a parameter used as an indicator of structural compactness and global dimension of a protein. Overall, no significant configurational changes are noticed in any simulation since neither abrupt increases nor significant drops are observed. However, a comparison between structures can be inferred from the time series of both the radius of gyration (Fig. 9) and RMSD (Fig. 8).

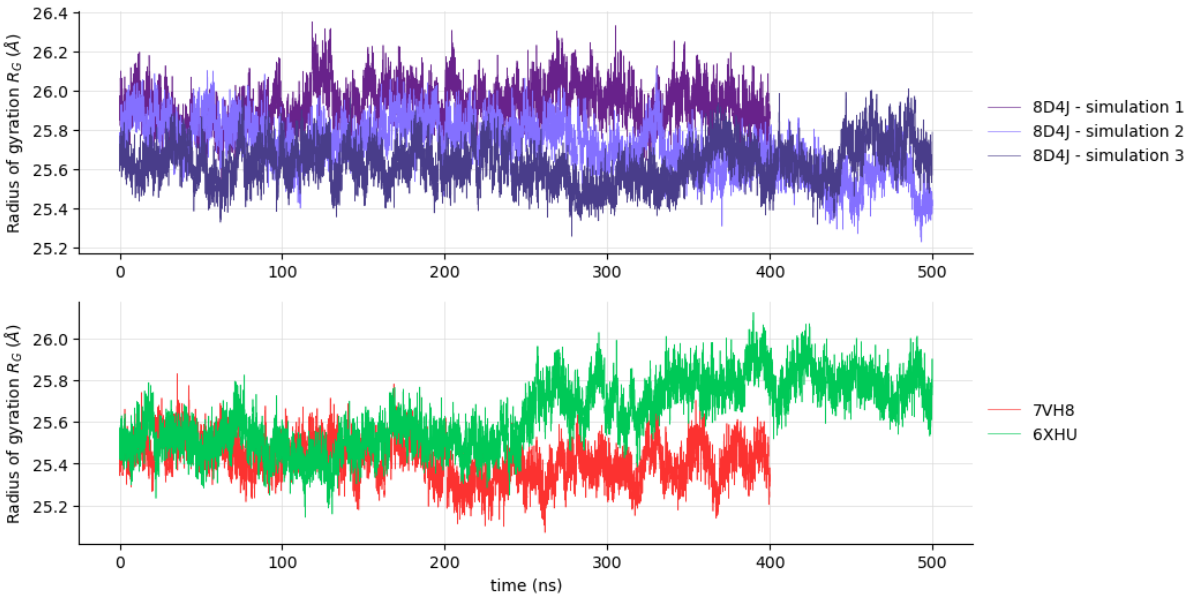


Figure 9. Plot showing the radius of gyration for the two wild types and all three simulations of the mutant for C_α selection.

As mentioned above, the **8D4J-1** simulation requires further investigation since the values of RMSD and the average R_g values are higher compared to those obtained from the other simulations. Here, R_g increases steadily by about 0.2 Å from 0 to 400 ns, meaning that the structure is compact during the transition and lacks visible structural rearrangements. The radius of gyration of **8D4J-2** behaves differently than the other simulation, it decreases by about 0.4 Å overall. The radius of gyration does not variate significantly for **8D4J-3** as well since it fluctuates around an average value of 25.6 Å. When the protein enters the third part of the trajectory, a R_g magnitude increase from 25.4 Å up to 25.9 Å is shown, eventually settling to 25.8 Å in the final state ($>$ 450 ns).

For **WT 7VH8**, R_g is not coherent with the variations seen in the RMSD. The radius of gyration has a smaller average value (25.4 Å) compared to the mutant ones (25.9 Å, 25.7 Å, and 25.6 Å respectively), suggesting a more compact and more stable structure, thus confirming observations carried out by De Oliveira [8].

The radius of gyration of **WT 6XHU** paired with RMSD has a behaviour that is similar to 8D4J-3. Both the observables have a constantly rising trend, hinting that the structure visibly changes during the simulation. The radius of gyration of WT 6XHU oscillates around an average value of 25.5 Å from the beginning to about 250 ns, after this time there’s a sudden increase to an average of 25.8 Å due to the reported ”flapping movement”.

RMSD and radius of gyration of single chains For **8D4J-1**, as reported in Fig.10, both the magnitude and trend of the RMSD of CH1 are comparable to the RMSD calculated on the C_α selection for the FP, suggesting that CH1 influences the protein structure more than CH2. The latter appears to be more compact, with a radius of gyration fluctuating around 22.0 Å and is characterized by a more steady RMSD, showing fluctuations of smaller magnitude.

The separated chains of **8D4J-2** (Figure **S4**) have a particular behaviour: the two chains oscillate at a similar value of RMSD (between 1 and 2 Å) until 150 ns. Between 150 ns and 250 ns CH1 has two spikes due to a sweeping

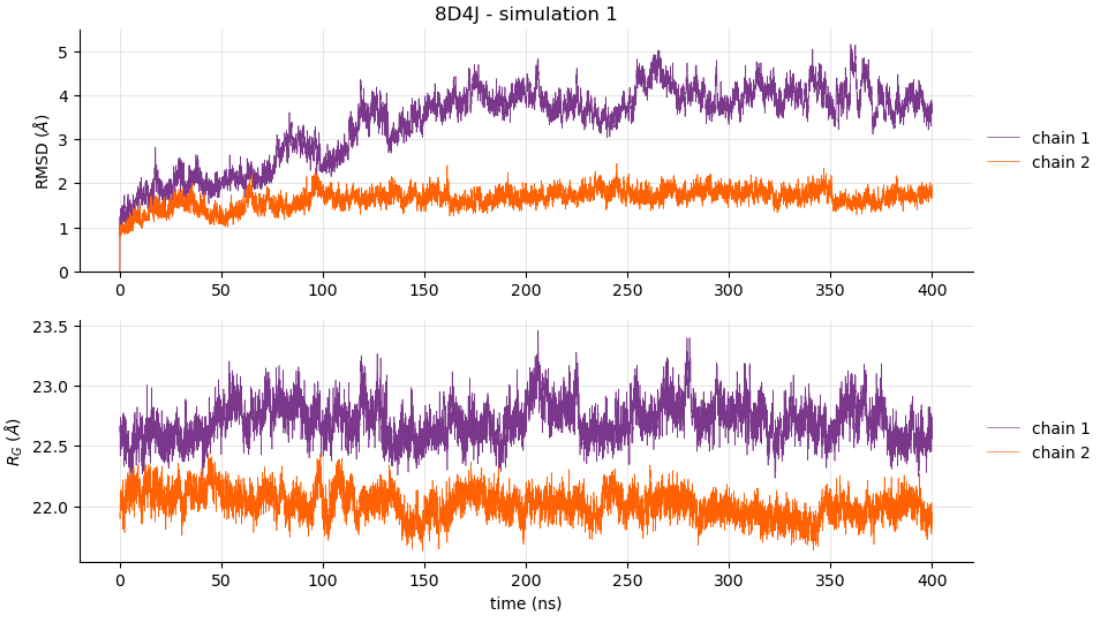


Figure 10. RMSD (top) and Radius of gyration (bottom) of the separated chains of 8D4J-1 simulation.

movement of the C-terminus, after which the chain has a RMSD value higher than CH2. After 150 ns, CH2 is on average 0.5 Å lower than CH1, its RMSD rises steadily from 1.5 Å to 2.0 Å until 380 ns then its value has a steep increase of 2.0 Å and keeps fluctuating between 2.0-3.5 Å until the end. This simulation is remarkable because the motions are present in the second chain, not in the first, and refer to the interactions reported in the visual inspection. This suggests the passage to a new conformational state that is less stable than the previous and the RMSD heatmap highlights this instability as well. Surprisingly, RMSD variations are not found in R_g data, that is fluctuating for both CH1 and CH2 between 21 and 23 Å. At about 370 ns, CH1 R_g diminishes and CH2 R_g increases, probably due to the "transition" to the new disordered state in the RMSD pairwise matrix.

About **8D4J-3** (Figure **S5**), CH1 has a behaviour similar to the FP. After the rapid initial equilibration rise left from the cut, the chain has an average value of RMSD of about 2.5 Å that increases to above 3.0 Å after 250 ns. This state stays constant until about 430 ns, then RMSD decreases and settles to an average of 2.0 Å until the end. CH2 does not show such behaviour, its RMSD is always lower than CH1, it increases slightly from the beginning to 150 ns from an average of 1.25 Å to 2.0 Å, then it fluctuates around this value until the end of the simulation.

The separated chains of **WT 7VH8** (Figure **S6**) have different behaviours, as expected, but fluctuate around an average value in both RMSD (1.6 Å) and radius of gyration (22 Å). The FP behaviour of RMSD is comprised of both chains, showing in particular the CH1 spike at 50 ns and an increase between 170 and 270 ns due to both chains, CH1 first, followed by CH2. Both are due to random fluctuations.

For all the structures so far, the discrepancies in the RMSD and R_g behaviour (constant R_g and variating RMSD) could mean that the movements of the residues do not disrupt the protein structure enough to cause an increment of R_g. This is corroborated by the fact that

1. the residues that fluctuate the most are in the external part of the protein, so they mostly interact with the solvent. R_g would significantly increase if repulsive or disrupting interactions were located in the internal part of the protein (i.e., in subdomain III or in the contact residues between the two chains);

2. the maximum RMSD magnitude is always at least three-fold lower than the maximum R_g magnitude and the simulations where the ratio is max (8D4J-1, Figures 8, 9) show an increase of R_g coherent with an increase of RMSD.

WT 6XHU (Figure **S7**), differently from all the other structures, has a behaviour that is similar to FP for both CH1 and CH2. Moreover, the FP behaviour seems to be mostly due to CH2. A spike at about 150 ns is shown in the RMSD plot in Figure 8. This spike is also reported for CH2 in both RMSD and R_g plots and is supposedly coherent with the "instability cross" found in the RMSD pairwise map. This is the first instance of the already-cited "flapping

movement", with the plots indicating an alternating activity between the active sites. The RMSD and R_g increase of the FP structure after 300 ns is due to the flaps that occur much more frequently from this timestep. Interestingly, something similar to an alternation between CH1 and CH2 is shown in RMSD, the trend is not so evident for R_g. The latter part of the trajectory of the FP structure is hence due to both chains, in an alternate fashion.

RMSF The root-mean-square fluctuation, RMSF, is a configurational metric that quantifies the individual residue flexibility during a simulation, often exploited to assess the stability of the various parts of a protein. It is related to the visual inspection that has been operated with VMD and helps individuate the residues that contribute the most to visual motion in the structure. The same results that were observed during the visual inspection should be expected here as well with greater detail. An example of what a RMSF plot looks like is reported in Figure 11,

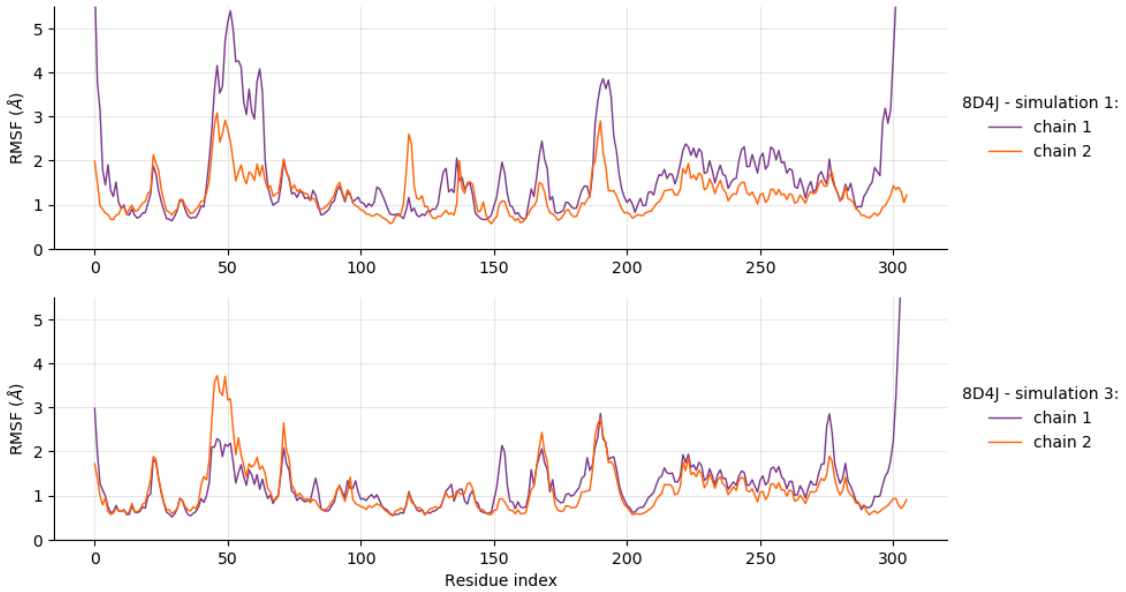


Figure 11. RMSF comparison of simulations 8D4J-1 (top) and 8D4J-3 (bottom), overlaid C_α selection for CH1 and CH2

referring to simulations 8D4J-1 and 8D4J-3.

Concerning **8D4J-1** (Figure 11), both RMSF and visual inspection assess that CH1 has higher fluctuations than CH2, especially

- residues 46-64, corresponding roughly to pocket S2 (see Table 1);
- residues 185-195;
- residues 220-270, with lower magnitude than the previous two groups but higher magnitude than CH2, corresponds to the lower part of subdomain III;
- terminal residues

In this simulation, the N-terminus of CH1 moves significantly more than the other termini - especially in the first 200 ns - because of the interactions with the S1 subpocket of CH2. The motion is also due to the connecting loops 189-196, and mostly due to the residues in subdomain III but since this latter subdomain has only structural significance, it is not relevant to protease activity.

For **8D4J-2** (Figure **S9**) there are no relevant discrepancies in RMSF except for residues 45-56, whose high fluctuations have been also observed for 7VH8 and 8D4J-1. The situation is similar for **8D4J-3** (Figure 11), visual inspection suggests that, as before, the most fluctuating residues are external and do not directly involve anyone of the residues of the active site. To better examine the reason for the very high RMSD in 8D4J-2, the trajectory was further cut at 375 ns and the last 125 ns were used to thoroughly analyse the RMSF. RMSF is at its highest for the C-terminus of CH2 (Gln306) that, according to visual inspection of the trajectory, interacts with the β-sheet

His41-Pro52 and with the loop Arg188-Ala193.

RMSF plots of separated chains for **WT 7VH8** (Figure **S11**) show different behaviours. CH2 has very high fluctuations around residue 47, that is coherent with visual inspection with VMD.
WT 6XHU (Figure **S12**) RMSF plot of separated chains confirms the considerations done by visually inspecting the interactions between the active sites. According to the plot, the most fluctuating residues are 44-60, 186-192, and 221-223. The difference lies in the magnitude of the second group of residues (186-192), where the maximum value of RMSF is 2.80 Å for CH1 and 5.75 Å for CH2. Both values refer to residue Gln189 that interacts repulsively with Met49 and with His41 and show how CH2 has the most intense ”flapping movement”.

RMSF analysis of the active site A focused RMSF analysis on the active site was performed to understand whether the two active sites differ between the two chains and between WT and mutant structures. RMSF plot of the active site residues was done by selecting all the residues reported in Table 1 by `atomnum` in MDAnalysis and plotting them sorted by residue number. The AS residues location in the chain are highlighted in Figure 12.

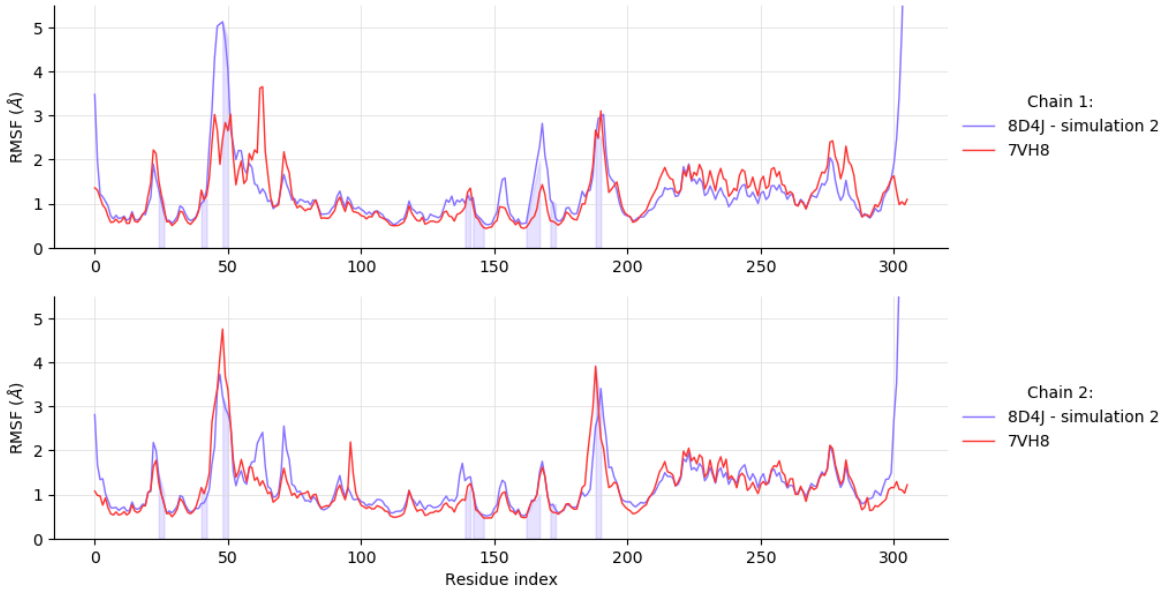
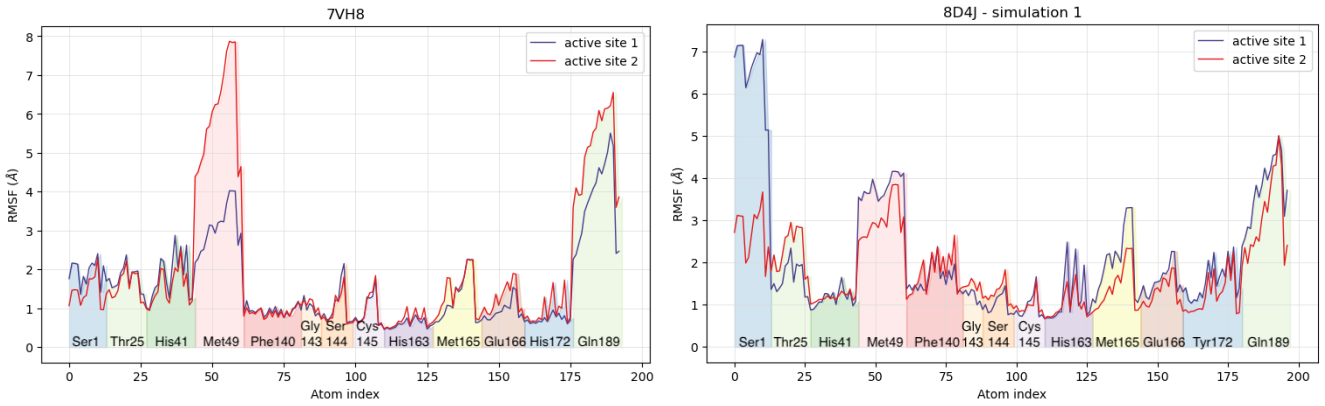


Figure 12. RMSF plot showing the positions of the active site residues with comparison between the two chains of simulations WT 7VH8 and 8D4J-2



(a) 7VH8 active site residues **(b)** 8D4J-1 active site residues

Figure 13. Plots of RMSF of residues of the active site of WT 7VH8 and 8D4J-1. The colours only help distinguish the different residues and have no particular meaning.

The colours of the plot have no meaning besides helping distinguish the different residues, that are highlighted by their name.

For all the structures, there are differences in the fluctuations of the two chains that involve particularly residues Met49 and Gln189. By visual inspection in VMD, these two residues interact during the simulation and can account for the high fluctuations observed in the overall RMSF of the chains. More specifically, the interactions are between

- Gln189 amide group - Met49 backbone carbonyl group
- Gln189 central methylene group - Met49 sulphide group

The fluctuation magnitudes reflect those of the chains.

In all three **8D4J** simulations (Figures 13b, **S13**, **S14**), higher fluctuations for the N-termini are observed due to the interaction of Ser1 with the ring-stacking. The Ser1-His172 interaction is prevented by the mutation of the histidine to a tyrosine, as explained before. Another highly fluctuating residue is Met165 in the mutant that interacts, through Asp187, with the aromatic ring of His41. This fluctuation is alike for both chains for the WT but is always higher for CH1 in all 8D4J simulations. It is interesting to note that 8D4J-3 has the least fluctuating active site among the mutants. Other relevant differences that can be observed are the increased fluctuations of Phe140 and His163 which are doubled with respect to 7VH8. They represent the aromatic ring stacking interaction that is disrupted by the presence of Tyr172.

For **WT 7VH8** (Figure 13a) the fluctuations of all the other AS residues are not higher than 3 Å, for both active sites the N-terminus has the same fluctuation magnitude. This suggests a stable bond with its target residues.

WT 6XHU (Figure **S15**) has the highest magnitude of fluctuations for residues Met49 and Gln189 of all the structures (up to 9 Å for CH2 Met49). This is a further confirmation of the presence of the flapping movement that does not involve any other residues of the active site. The AS behaviour overall is similar to WT 7VH8, except for Ser1 of both chains which has fluctuations that are similar in magnitude to 8D4J-2, and 8D4J-3.

3.1.3 Summary and further considerations

The structural analysis of the two WT and the three simulations of 8D4J gave great insights into the behaviour of the protease in both forms. This phase of the project was useful to isolate a possible production part of the simulation and to identify particularly active residues. The activity of the chains and the active site was observed thoroughly both by visual inspection and by analysing configurational observables such as RMSD, R_g, and RMSF.

In general, the two WT structures behave very differently: 7VH8 stays compact and does not show any particular conformational changes during the simulation, while 6XHU shows strong and repulsive interactions between Met49-Gln189 and the breaking of the ring-stacking Phe140-His163-His172.

The mutation in 8D4J leads to a change of interactions in the active site. As explained, Tyr172 has a greater steric hindrance and acidity compared to WT His172. The most important consequence, which was also reported in [7], is the change of interactions between the N-finger and the active site, disrupting the catalytic activity of the protease. This behaviour was very present in the first 8D4J simulation, which shows also a visible conformational transition from one state to another in the RMSD heatmap of Figure 7a. 8D4J-2 and 8D4J-3 have similar N-finger interactions with the corresponding AS of smaller influence on the overall trajectory, as suggested by the RMSF plot of the AS. The two simulations do not show any particular transitions and seem to explore different areas of conformational space with respect to 8D4J-1 but as of now it is not possible to discriminate between the phase space exploration of 8D4J-2 and 8D4J-3.

The experimental information in RCSB Protein Data Bank for 6XHU reports that the structure was obtained by X-ray diffraction at pH 6. Besides the crystallization of the structures being performed at different pH when the system is prepared for MD simulation, in the ionization step the electron neutrality is reached. This does not ensure that we are at neutral pH since there is no information about the pH of the system. Hence, the breakage of the ring-stacking interaction and the flapping movements observed in the simulation could be related to the pH-dependent activity of the protease, as reported in the Introduction.

Indeed, WT 7VH8 was crystallised at pH 5.6 and the simulation does not show any notable difference in the two chains or between the active sites. Mutant structure 8D4J, as described in [7], was crystallised at neutral pH of 7-7.5. Since no information about the pH-dependent activity was reported for mutant structures, and since no particular differences were observed for the two active sites, according to information reported in the Introduction, it could be inferred that the mutant structure has a coherent activity with the pH at which it was obtained.

This part of the work has been performed by considering the pH of the simulated molecular system to be the same pH of the initial crystallographic structure. In a MD simulation, the pH of a solution cannot be assessed because the water autoprotolysis reaction ($H_3O^+ + OH^- \rightleftharpoons 2\,H_2O$) is not explicit and should be treated with quantum mechanics. This assumption is also validated by the fact that the salt used for the charge neutralisation is NaCl, that generates neutral solutions and should not interfere with the system.

This though does not exclude the possibility that the structures present this activity due to the sampling of the conformational space.

3.2 Functional analysis

To further characterise the protein behaviour and quantify the sampling quality of the conformational protein space, a statistical analysis was computed. The following study is not pretentious in giving quantitative results but still attempts to highlight differences between the WT and mutant structures by exploiting a statistical-based approach besides the restricted amount of available data.

It is worth mentioning that some of the tools exploited to perform this functional analysis, such as the expression for the autocorrelation function and spectral analysis, are true for, at least, *wide-sense stationary* signals [2]. Even though equilibrium can not be assured, these techniques are still used because it is assumed that the major part of transient conformations is discarded with the equilibration part of the trajectory. However, we are aware of the fact that it is not accurate to apply these statistical tools to the available data.

In this work, the observable of interest on which statistical analysis is performed is the RMSD, which is calculated on α-carbon selection for the full protein (FP[3]), chain 1 (CH1), and chain 2 (CH2).

Correlation graph A first exploratory analysis to assess correlation is carried out examining *lag plots*. The shape of a lag plot is often used to evaluate whether values in a time series are completely random or if they are affected by any hidden pattern. In this type of analysis, the X-axis represents the original RMSD time series, and the same time

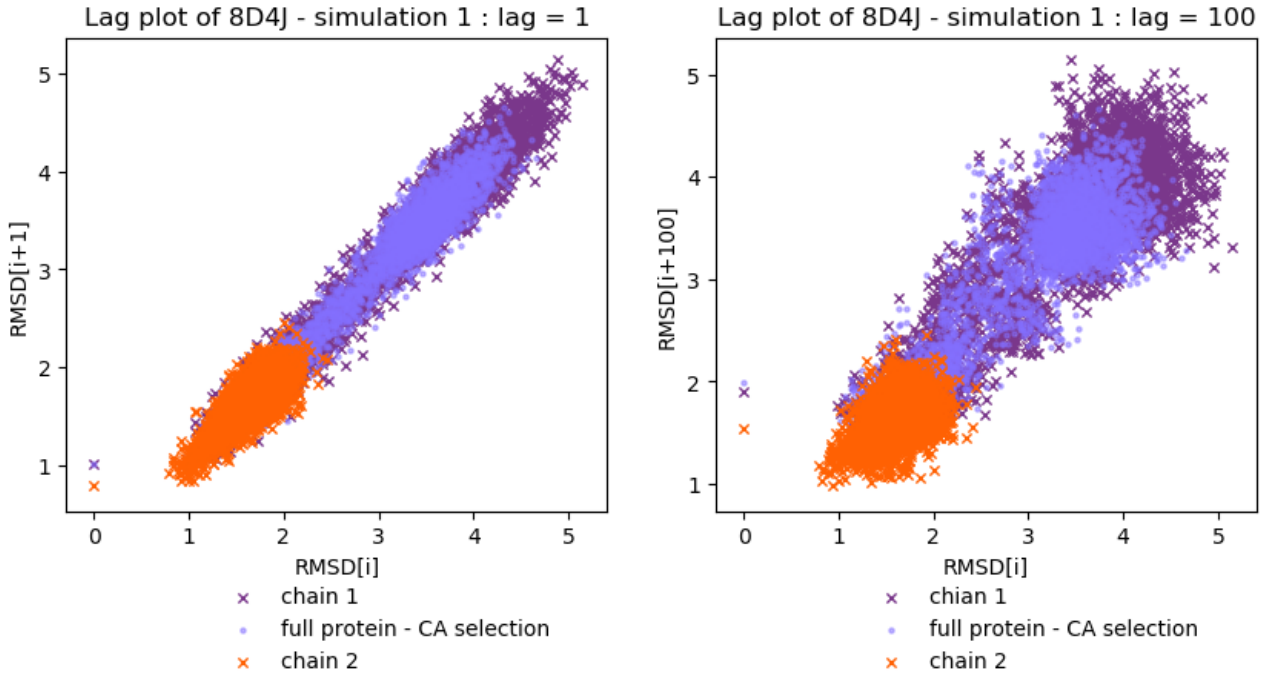


Figure 14. Lag plot for 8D4J-1 with a lag of 1 time step (left) and 100 time step of delay, corresponding to a delay of 5 ns of the time series (right).

series shifted by a chosen temporal lag k is displayed on the Y-axis.

The lag plot for simulation **8D4J-1** is reported in Fig. 14 while the other results can be found in **SI**. A linear pattern is observed for all simulations with the RMSD plotted against the 1^{st} order lag spreading out for higher-order delays.

[2] WSS processes require that the mean value and autocovariance are time-invariant and the second moments are finite.

[3] can also refer to *Faccioli Pietro*

This trend is expected since a higher degree of correlation is typical for time series plotted against smaller lags. Concerning the 1^{st} order lag plot, points of **WT 7VH8** (Fig. **S18**) are less tightly bound along the diagonal compared to data points of the 8D4J simulations. A more dispersed cloud of points in the lag plot obtained for a 5ns-shift of the WT 7VH8 RMSD time series suggests a lower autocorrelation time. On the contrary, the lag plot for **WT 6XHU** (Fig. **S19**) is more similar to the ones of 8D4J, suggesting a characteristic time scale alike the one of the mutant, hence greater correlation.

Autocorrelation analysis One of the main weaknesses in data simulation is the presence of correlation in the generated time series, an unavoidable consequence due to the nature of the algorithms exploited to produce data. Autocorrelation analysis performed on a time series allows the measurement of the similarity between observations of a random variable as a function of the temporal delay between them. Methods such as the computation of the autocorrelation function or Block averaging are often used to understand the relevant time scales typical of a biological system.

Autocorrelation function The *autocorrelation function* (ACF) can be used to establish how the correlation between any two values of a time series changes with the temporal lag. In the case of a wide-sense stationary process, the ACF calculated for any observable $f(t)$ is derived from the formula:

$$C_f(t') = \frac{1}{\sigma_f^2}\frac{1}{N}\sum_{i=1}^{N-t'/\Delta t}(f(i\Delta t) - \langle f\rangle)(f(i\Delta t + t') - \langle f\rangle) \tag{1}$$

where N is the number of frames, Δt is the time step, σ_f^2 is the sample variance, and $\langle f\rangle$ is the sample mean estimated from data. It is worth noticing that Eq. 1 is an approximation of ACF since the macromolecule under study may not have equilibrated yet but it could be in a local minimum of the potential instead.
Concerning 8D4J, the ACF evaluated on the RMSD time series of the full protein C_αs displays a similar pattern characterised by a first slow decrease, followed by a soft rise towards zero. Even though it seems that comparable time scales characterise the behaviour of 8D4J FP in all three simulations, a deeper analysis highlights that the ACF of the two chains contributes differently to the overall trend.

In both **8D4J-1** and **8D4J-3** (Fig.15a + **S21**), the ACF for CH2 decreases more steeply than the one calculated for CH1, hinting that the former reaches equilibrium before the latter. The FP behaviour is more influenced by CH1 than CH2, as it is suggested by the similar trend of FP and CH1. For both simulations, this consideration is also observed in the configurational distances measured for both chains in the structural analysis (Fig. 10 + **S5**).
In Fig. **S20**, the ACF of **8D4J-2** exhibits an opposite behavior compared to the other two simulations. Specifically, CH1 reaches equilibrium earlier than CH2 due to the steeper initial decrease of the ACF. The plot also reveals that CH2 has the most significant influence on FP behavior. The structural analysis of RMSD of 8D4J-2 (see Fig. **S4**), it reveals high fluctuations in the last 120 ns of simulation that may account for the observed behaviour.

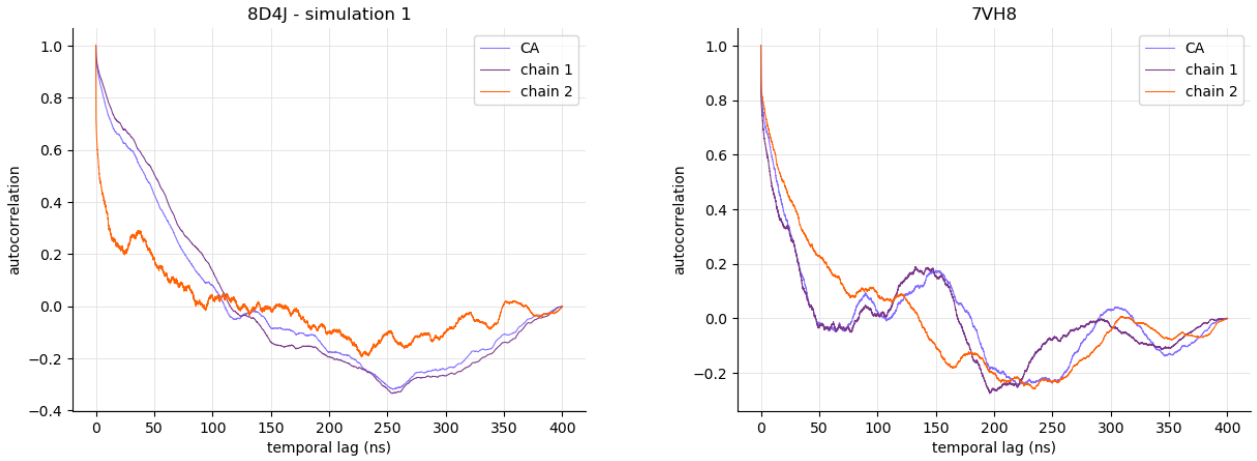
The ACF of **WT 7VH8** (Fig. 15b) displays more oscillations around 0 than the one calculated for the mutant. An overall steeper drop for shorter lags is also a symptom that memory of the WT is lost faster compared to the mutant. In particular, CH1 has a sharper fall towards zero compared to CH2. The two chains are characterised by different autocorrelation times, as seen for the mutant structures.

Finally, the ACF calculated for **WT 6XHU** (Fig. **S22**) displays a trend that is more similar to the ones of 8D4J; hence its typical time scales resemble the mutants more than the other WT.

Autocorrelation time The ACF is useful to calculate the *autocorrelation time* τ_f of the observable $f(t)$ in order to extract the number of independent conformations from the simulation. Once the number of independent configurations is known, time-invariant statistical properties can be exploited to describe the system at equilibrium. The autocorrelation time is defined as

$$\tau_f = \int_0^\infty C_f(t)\, dt \tag{2}$$

and quantifies the period that must elapse before the process loses memory of the past values of the observable.



(a) 8D4J - first simulation **(b)** 7VH8

Figure 15. Autocorrelation function for the RMSD of the first simulation of 8D4J (left) and the WT 7VH8 (right). The ACF is computed for the C_α selection of the full protein and single chains.

A first estimate for τ_f was obtained from its definition (Eq. 2), thus by performing a numerical integration of the ACF over the simulation period. However, this method yielded negative values of autocorrelation time for all simulations. An explanation of this result lies both in the shape of the ACF and the limited length of the simulation. According to the definition, the integration of the ACF should have been performed over an infinitely long simulation period while only about 400 ns were simulated. As the time delay increases, fluctuations around zero of the ACF are expected to diminish. However, the simulated time is not enough for fluctuations to drop to zero. On the contrary, the results very close to zero of the autocorrelation time suggest that the positive contribution almost balances the negative one. According to [16], the integration to infinity should be performed with care due to noise in the long-time tail of the correlation function.

Given the difficulties of obtaining a reliable estimate of the autocorrelation time, a non-linear least square fit to the ACF obtained from the RMSD data was performed using the fitting function:

$$y(t) = \exp\left(-t/\tau_f\right) \cdot \left(\cos\left(\omega t\right) + A\sin(\omega t)\right) \tag{3}$$

to account for both the exponentially decaying trend and the fluctuations at longer lags. Only the first harmonic is considered in the fit since the aim of the analysis is to obtain a ”reliable” estimate for the autocorrelation time and not a precise result. As an example, the plot with the ACF calculated from the RMSD of the FP 8D4J-1 with selection C_α together with its fitting function is reported in Fig. (16 + **S23**).

Since different sources of error enter this discussion, in the first place because of the limitation in the available data, a more robust result could be obtained by performing repeated simulations to carry out a statistical analysis. This strategy is highly resource-demanding since equilibration must be reached in all the simulations and the poor statistics achieved in this work do not consent to making any relevant estimate. Nevertheless, a qualitative comparison between the autocorrelation time obtained from the fitting procedure (Table 3) performed on different protein selections was attempted.

Concerning **8D4J-1**, the autocorrelation time for CH2 is much smaller compared to the one for FP and CH1, confirming the hypothesis that CH2 is more stable than CH1 which, however, influences the behaviour of the full protein the most. For all simulations of the mutant, a number between 2 and 3 independent samples were obtained through this procedure, suggesting that the macromolecule takes time to reach equilibrium. Ultimately, no relevant quantitative conclusions can be drawn from the data.

From Table 3, a discrepancy between the autocorrelation time values for **WT 7VH8** can be highlighted. The FP and CH1 have much shorter autocorrelation times than the other structures, suggesting that they could reach an equilibrium configuration in less time and could have a greater number of independent samples. However, since the structural analysis did not confirmed equilibration of the protease, these results could strongly underestimate

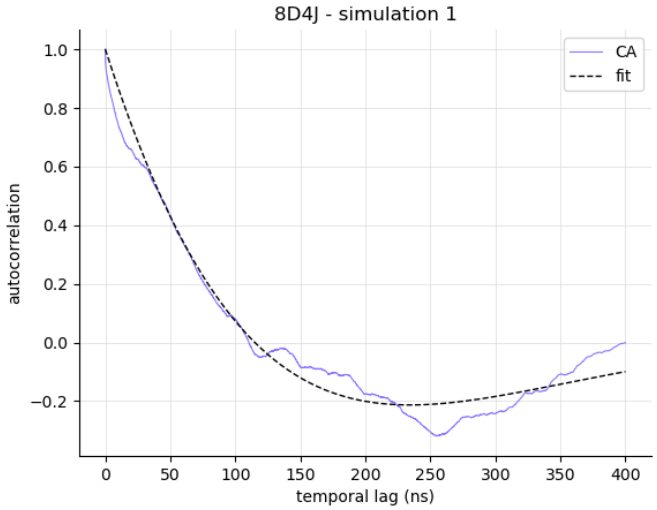


Figure 16. Example of the ACF calculated from the RMSD of the full protein with C_α selection plotted together with its fitting function.

the autocorrelation time due to the choice of the fitting function. Figure **S23** illustrates how the chosen fit is too loose for 7VH8 - but it is not entirely appropriate either for the other plots - and does not represent evenly the oscillating behaviour of the ACF.

6XHU is expected to take longer to reach a steady state and only 2 independent samples are obtained from this procedure.

Table 3. Table with estimates of the autocorrelation time and number of independent blocks calculated for three selections (α-carbon): full protein, chain 1 and chain 2. Both results obtained fitting the autocorrelation function and from a block average analysis are reported.

	8D4J - 1			8D4J - 2			8D4J - 3		
	C_α	chain 1	chain 2	C_α	chain 1	chain 2	C_α	chain 1	chain 2
Autocorrelation function									
τ_f [ns]	149	179	15	182	113	167	151	146	63
N_{ind}	2	2	26	2	4	3	3	3	7
Block averaging									
τ_f [ns]	99	100	50	125	125	83	125	125	125
N_{ind}	4	4	7	4	4	6	4	4	4
	7VH8			6XHU					
	C_α	chain 1	chain 2	C_α	chain 1	chain 2			
Autocorrelation function									
τ_f [ns]	19	16	98	191	225	173			
N_{ind}	21	25	4	2	2	2			
Block averaging									
τ_f [ns]	100	100	100	125	125	125			
N_{ind}	4	4	4	4	4	3			

Block average Block average analysis is another method employed to deal with correlation in time series. It is based on the idea of generating uncorrelated data from the time series by creating independent samples from the original data set. The implemented algorithm follows the guideline presented in [16] according to which a running

estimate of the standard error for the observable $f(t)$ - in this case RMSD - can be calculated as:

$$BSE(f,n) = \frac{\sigma_n}{\sqrt{M}} \tag{4}$$

M is the number of blocks of variable length n into which a given trajectory can be divided, and σ_n is the standard deviation of the mean. For small n, consecutive blocks are highly correlated; thus, BSE largely underestimates the correct statistical error since Eq.4 is true only if the M blocks are independent.
In principle, from the trend of the function $BSE(f,n)$ it is possible to see when the error estimate has converged and thus identify the length $n = \tilde{n}$ for which blocks become independent. When blocks are statistically independent, the BSE should asymptotically reach the value of the correct estimator for the true standard error SE:

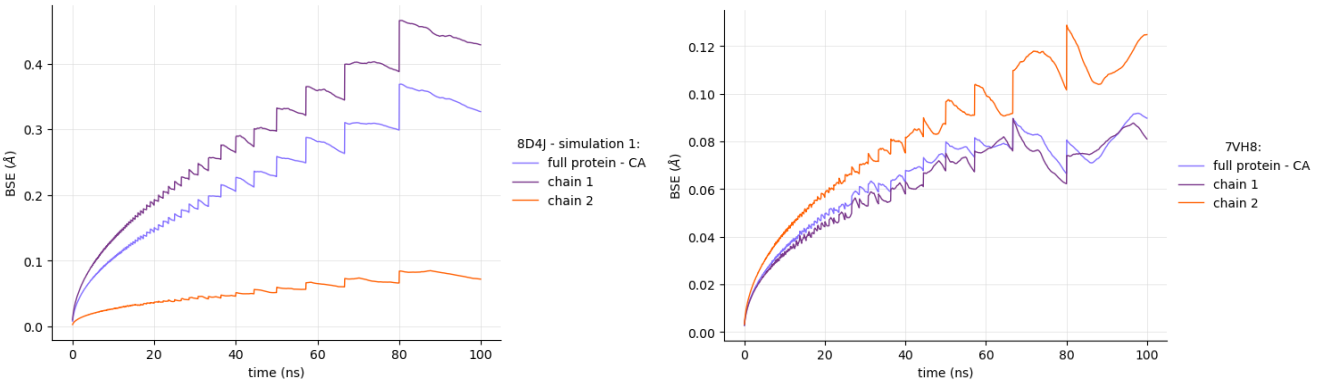
$$BSE(f,\tilde{n}) \approx SE(f) = \frac{\sigma_n}{\sqrt{N_{ind}}} \tag{5}$$

Recalling that $N_{ind} \approx \frac{t_{sim}}{\tau_f}$, the autocorrelation time solving for τ_f can be estimated as:

$$\tau_f = \frac{BSE(f,\tilde{n})^2}{\sigma_n^2} \cdot t_{sim} \tag{6}$$

Note that this autocorrelation time could underestimate the correlation time by a factor of 2 [16]. Results are reported in Tab. 3 together with the estimate of the number of independent blocks.
It is worth recalling that only qualitative conclusions can be drawn from these results since they are affected by various sources of error. The major issue was encountered when establishing $\tilde{n}$ from the plot of BSE (Fig.17 + **S23**, **S24**, **S25**). Indeed, even though a plateau value was not reached in most of the simulations - for instance, Figure 17 - the choice of approximately 100 ns (50 ns for 8D4J-1) was made to have a limited amplitude of the fluctuations due to poor sampling of the equilibrium configurations.

Concerning **8D4J**, lower outcomes for the autocorrelation time were obtained with the BSE method compared to the ones derived from the fitting procedure, probably due to an underestimation of the number of frames $\tilde{n}$ within independent blocks. For **7VH8**, the estimate of τ_f obtained from the fitting procedure is way smaller compared to the outcomes from the Block analysis. In this case, both the inaccuracy of the fit and the choice of $\tilde{n}$ may have influenced the estimate. An accurate distinction between the fit and the block analysis estimate of the standard error is not possible: the amount of data available is insufficient, and the algorithm should undergo further testing.



(a) 8D4J-1 **(b)** 7VH8

Figure 17. BSE plots for the first simulation of 8D4J (left) and the WT 7VH8 (right) computed for the C_α selection of the full protein and single chains.

Spectral analysis of the RMSD time series The role of correlation in biological matter was further investigated by turning into the frequency domain [17]. Spectral analysis is a powerful tool to analyse random signals by decomposing them into simpler signals that can be represented as a sum of many individual frequency components.

In this study, the Fourier analysis was performed on the RMSD time series, which is composed of N elements and is labeled as f_n, $n = 1, ...N$. By computing the *power spectral density* (PSD) $S_k(\nu)$ of f_n, $n = 1, ...N$, the distribution

of power into frequency components characterising that signal can be seen.

Many techniques have been developed to perform a spectral density estimation of a signal's power starting from a sequence of time samples of the signal. The *periodogram*, for example, estimates the power spectrum by taking the square modulus of the DFT components calculated from the original sequence.

Despite the drawbacks of the periodogram [4], no sophisticated spectral analyses will be performed. Thus, the periodogram of the RMSD time series is considered to be a sufficient tool to investigate correlation properties. Periodograms for 8D4J-1 and 7VH8 are reported in Fig. 18.

The fluctuating component of the RMSD, is obtained by subtracting the mean value from the original time series. Then, the Discrete Fourier Transform F_k of the 0-mean time series was computed, and the periodogram $\tilde{S}_k(\nu)$ derived as :

$$\tilde{S}_k(\nu) = \frac{\Delta t}{N}\left|F_k\right|^2 = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} f_n \exp\big(-2\pi i \frac{nk}{N}\big)\right|^2 \tag{7}$$

where N comes from the fact that a normalized square window is considered and Δt is the sampling period.

Subsequently, the type of noise characterising the RMSD was studied. In general, the power-law noise spectrum is related to frequency ν by the relation:

$$PSD(\nu) = \frac{a}{\nu^\beta} \tag{8}$$

where the color of the noise depends on the value of the parameter β. For example, $\beta = 0$ corresponds to white noise since its PSD is independent of frequency while $\beta = 1$ is related to pink noise, also known as "1/f noise" that decays by 3.01 dB per octave.

The value of β characteristic for the simulated data was obtained through a fitting procedure, accounting for the fact that:

$$10 \cdot \log(PSD(\nu)) = 10 \cdot \log\left(\frac{a}{\nu^\beta}\right) \tag{9}$$

that for the property of logarithms:

$$10 \cdot \log(PSD(\nu)) = 10 \cdot \log(a) - 10\beta \cdot \log(\nu) \tag{10}$$

Thus, data can be fitted with a line : $y = A + B \cdot x$, where:

- $y = 10 \cdot \log(PSD(\nu))$
- $x = \log(\nu)$
- $A = 10 \cdot \log(a)$
- $B = -10\beta$

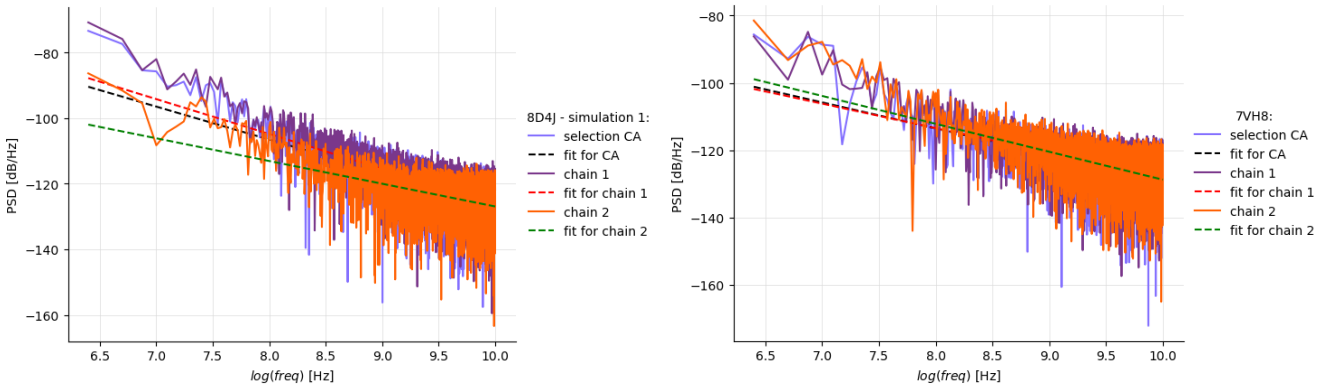
From the fit, the parameter β can thus be determined from the slope of the line as:

$$\beta = -\frac{B}{10} \tag{11}$$

It is worth recalling that values of β close to 1 indicate pink noise while the closer it goes to 0, the more features of white noise are present in the data. Results in Tab. 4 show that, overall, β obtained for 7VH8 are smaller compared to the ones obtained for simulations of 8D4J, suggesting that less correlation is present in the WT structure compared to the mutant. Moreover, results obtained for 8D4J-1 validate the hypothesis that CH2 is less correlated than CH1 - once again - since the noisy part of the signal resembles white noise the most.

The typical 1/f trend obtained from the 0-mean RMSD time series suggests that:

[4] Periodogram estimator of the PSD is affected by a bias that often results in higher variance and, consequently, higher signal power.



(a) 8D4J-1 **(b)** 7VH8

Figure 18. Periodogram plot for different protein selections of both the first simulation of 8D4J and 7VH8. A linear fit for each selection, from which the value of the parameter β is determined, is also displayed.

Table 4. Table with values of the parameter β obtained fitting the periodogram obtained from simulated data with the function $y = A + B \cdot x$. Different selections of the protein were considered -α-carbon atoms -: full protein, chain 1 and chain 2.

Structure	Full protein	Chain 1	Chain 2
8D4J - 1	1.01	1.06	0.69
8D4J - 2	0.91	0.95	0.97
8D4J - 3	0.82	0.92	0.72
7VH8	0.77	0.73	0.83
6XHU	0.87	0.91	0.87

1. Both a long time scale and short time scale characterize the evolution of the system. The high contribution observed at low frequency can be attributed by 'macroscopic' oscillations observed in the RMSD data, not excluded by the discard of the mean value. Higher frequency contributions may be due to both thermal fluctuations and other sources of noise instead.

2. The presence of a stochastic contribution in the original RMSD time series introduces correlation and this is observed by the presence of a 1/f-decaying trend in the PSD. If no correlation was present, a spectrum resembling white noise would have been expected.

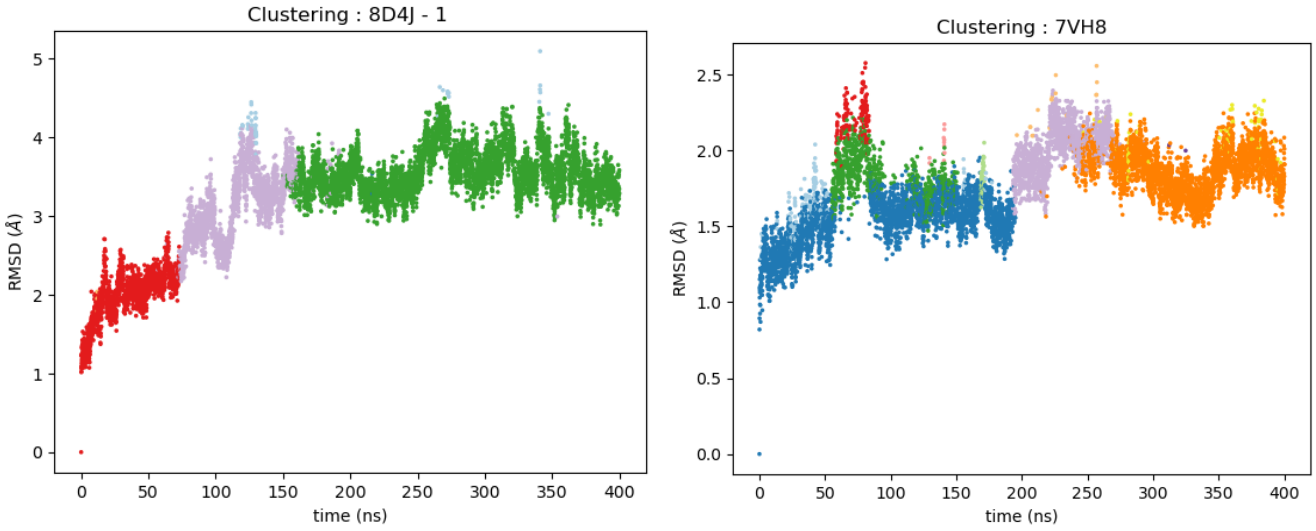
The fact that a pink-colored noise is obtained from the noisy part of the RMSD time series is in agreement with studies carried out by Grove *et al* [18] concerning the role of noise color in biological [5] systems. Indeed, empirical studies applied to environmental and ecological time series showed that lower-frequency sources of variation had higher amplitude than higher-frequency noise. Thus, time series with pink - or red - PSD are considered to be more realistic as proxies for real environments.

3.3 Dimensionality reduction analysis

Clustering In the framework of MD simulations, *cluster analysis* is a useful technique to find patterns within data by grouping similar conformations visited by a biological system. Since clustering identifies a general task to be solved and not a specific classification procedure, different algorithms have been developed [19]. The choice of the algorithm depends on the features of the selected dataset. Moreover, the clustering procedure is intrinsically subjective, mainly due to the freedom of choosing the type of linkage and method used to label different clusters. Hence, different results can be obtained starting from the same input data.
In this study a hierarchical bottom-up clustering approach is applied to equilibrated pairwise RMSD data, which are given in input as a distance matrix to the Python function `scipy.cluster.hierarchy.linkage`. The *average-linkage* method is exploited to compute the inter-cluster distance, which is established by calculating the average pairwise

[5] Grove refers to *evolutionary* systems, based upon much longer time scales (up to hundreds of years). We have found no related information about molecular systems, but it would be an interesting consideration, along with other theoretical frameworks

distance of all pairs of objects from different clusters.
Subsequently, dendrograms are plotted as a tree representation of the arrangement of clusters obtained from the previous linkage choice (plots are reported in **SI**). The Python function `scipy.cluster.hierarchy.fcluster` is then used to label clusters starting from the previously calculated linkage matrix exploiting the *distance* criterion according to which flat clusters are generated so that the original observations in each group are within a threshold cophenetic distance t. This value is chosen by visually inspecting each dendrogram, and the resulting number of clusters is reported in Tab. 5. RMSD time series is plotted as subdivided into groups obtained from the cluster analysis (Fig. 19) that can be physically interpreted as different protein configurations characterised by stable RMSD values.



(a) 8D4J - first simulation **(b)** 7VH8

Figure 19. RMSD time series obtained from production data with highlighted groups corresponding to different protein configurations obtained form the cluster analysis

In **8D4J-1** (Fig. 19a) three principal regions are identified by different clusters. The first group represents the part of the transition after the trimming procedure, hence the transition towards the second metastable state in Figure 7a. A further transition period of about 100 ns is represented by a second group (lilac), after which a third cluster (green) corresponds to the region in which the protein reaches a stable conformation

Simulations 8D4J-2 and 8D4J-3 were found to have between 4 to 6 optimal clusters. In particular, the clustering of **8D4J-2** (Fig.**S31**) is separated into two main regions. The lilac cluster refers to the map area up to 480 ns, with a smaller green group associated with the small transition at 250 ns. The final cluster (light orange) represents the last part of the trajectory where CH2 influences the protein motion, moving the FP towards a new conformation.
Concerning **8D4J-3** (Fig. **S32**), the previously discussed interchangeability of states visited by the protein - observed in the pairwise RMSD - is highlighted clearly. Interestingly, in this simulation, the protein visits a new state (red) and goes back to the previous one (lilac) before moving towards a new conformation (green).

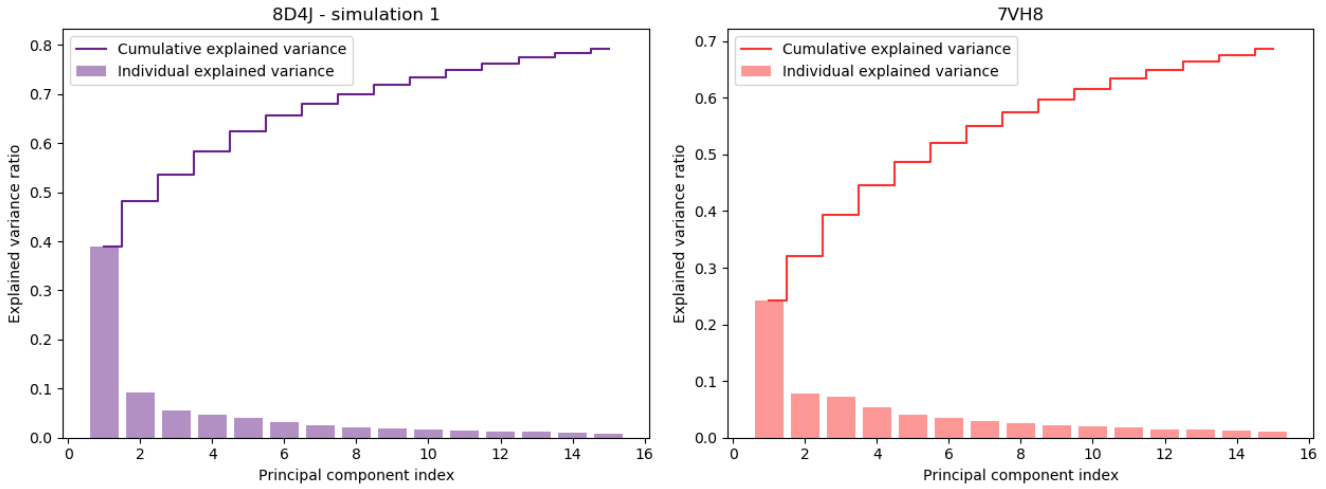
More clusters were obtained for the **WT 7VH8** (Fig. 19b), validating the observations drawn from the RMSD matrix about the lack of equilibration of the protein which visits different states, either transition states (lilac, green, red) or conformations (blue, orange). Clustering analysis corroborates the unphysical nature of the low autocorrelation time value obtained for the FP with the fitting procedure.

Despite the constant rising trend of the RMSD, three major conformations emerge from the clustering analysis applied to **WT 6XHU** (Fig. **S33**). The green region corresponds to the low initial RMSD values in the map. The orange spike at about 150 ns corresponds to the first flapping motion observed through the visual inspection of the protein, associated with the unstable cross in the RMSD map in Figure **S1**. The red and lilac regions have no direct correspondence with the map. Probably, a different clustering algorithm could have given some more insights about this WT structure.

In this study, clustering analysis is applied only to the FP: the different role played by single chains in influencing the global behaviour of the protein is already assessed by the structural analysis. More insights into the motion of specific parts of the protein could be gained by applying this technique to definite protein selections (e.g. single chains, single domains).

Principal component analysis Principal component analysis (*PCA*) is a technique used to analyse large amount of data to reduce the dimensionality of the original data set. PCA can be interpreted as a linear change of basis that highlights directions of maximum variance and projects data in those directions. In this study, PCA was applied to C_α selection of both the mutant and WT proteins.
Explained variance is the quantity that expresses how much variation in the data set can be attributed to each of the principal components. By telling how much of the total variance is 'explained' by each component, explained variance allows one to rank components in order of importance: the larger the variance accounted for by the component, the more that component is important.
Percentages of both the cumulative and individual variance of the first PCs were plotted (Fig. 20 + **S34**, **S35**, **S36**) to understand the relevance of the first modes with respect to all the others. In Tab. 5 the variance obtained for the low values of the first PC suggests that the collective motions of each structure is mostly due to the local portion of the protein characterised by particularly high RMSF values. Moreover, the low values obtained for the cumulative variance of the first 3 PCs suggest that a great part of the information is lost when projecting data onto the reduced dimensional space, in particular for WT 6XHU whose first component accounts for the 18% of variance only.

In MD simulation framework, one application of PCA is to detect *essential dynamics* in biological systems since it allows the identification of the most meaningful basis to reduce the number of degrees of freedom and thus the relevant modes to describe the motion of the system.



(a) 8D4J - first simulation **(b)** 7VH8

Figure 20. Individual and cumulative explained variance percentages plotted for the first 15 PC.

By diagonalising the covariance matrix for the atom selection, a set of eigenvectors and eigenvalues is obtained. Eigenvectors represent the principal components while the eigenvalues represent the associated variances.
The dynamics in the low-dimensional space is often referred to as 'essential dynamics' to reflect the notion that the principal modes of motion are those that are essential to describe the protein's function. In this work, PCA is used to try to highlight if any functional difference between the mutant and WTs can be detected by considering the principal mode of motion.

Firstly, the time series for the projections of the first 3 modes is plotted for each protein. It is worth mentioning that a proper physical interpretation of PC is difficult to assess; according to [16], the first three principal components can be plotted as "reduced" spatial coordinates, but they are lacking in any physical meaning of relevance. The PCA plot, consequently, can only give some hints about the motions of the structures.

Figure 21 shows the general motions of 8D4J-1 in the reduced coordinate space, where colour range represents time. A comparison between conformational spaces of different proteins (Figures **S37**, **S38**, **S39**, and **S40**) highlights

Table 5. Table with the number of clusters obtained from the clustering analysis and variance explained by the first PC and cumulative variance of the first three principal components evaluated for C_α selection of both the WT and mutant structures. The cosine content of the first PC is also reported.

Structure	# of clusters	Var (PC1)	Cumulative variance	Cosine content PC1
8D4J - 1	6	0.39	0.54	0.67
8D4J - 2	4	0.3	0.44	0.73
8D4J - 3	6	0.22	0.39	0.47
7VH8	11	0.24	0.39	0.67
6XHU	6	0.18	0.37	0.76

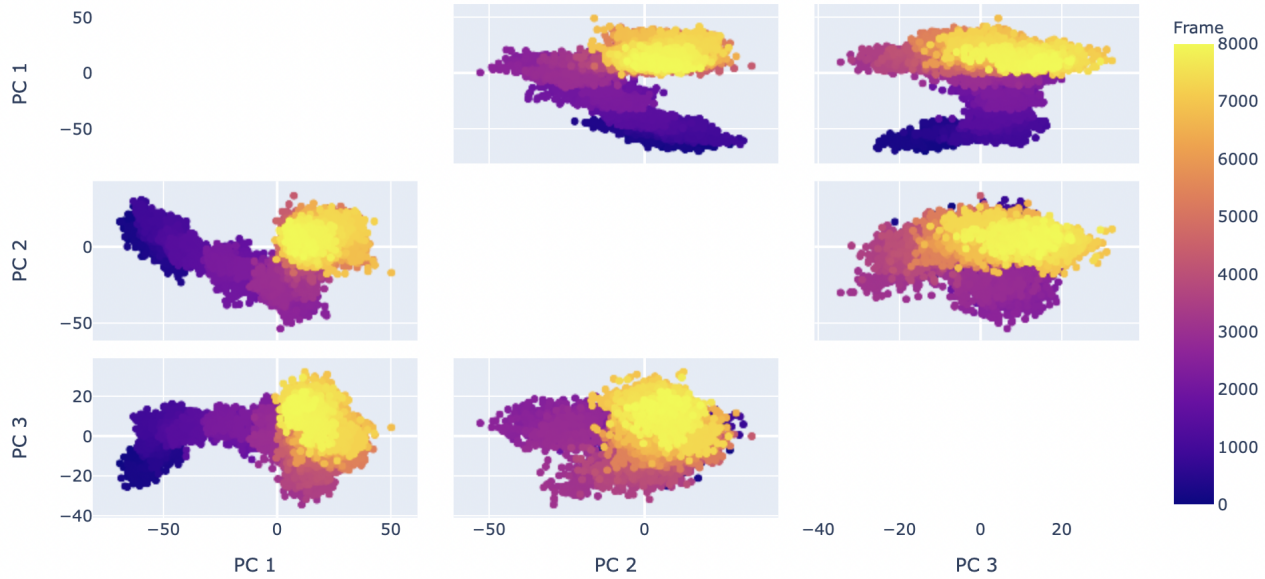


Figure 21. Trajectory of simulation 8D4J-1 projected onto the first three PCs.

that the cloud of points corresponding to the projections onto the first 3 PCs is more clustered in the WTs than the mutant. The proteins can be seen as ”moving” in the reduced space; in particular, 8D4J-2 shows a jump towards a new basin but the motion can not be related to any previously mentioned structural evaluation.

Another way of assessing the stability of PCA results is by calculating the cosine content of the PCA projections. Values calculated for PC1 (Tab. 5), that are not close to 1, suggest that all simulations have ’converged’. No further conclusions can be drawn from the obtained results since each protein visits different conformational basins. Sadly, there is no real physical convergence in the simulations.

In an attempt to find a connection between the original physical trajectory and the one in reduced space, the RMSF of the trajectory projected onto the first mode is computed using GROMACS. The covariance matrix is calculated and diagonalised through the command `gmx covar`; the function `gmx anaeig` is then used to analyse the obtained eigenvectors, and the command `-rmsf` is added to compute the RMS fluctuation per residue of the first PC. Although a meaningful contribution to the RMSF may also be provided by higher modes, especially for 6XHU, only the first principal component is considered, despite its low variance values.

The trend of the RMSF reported in Fig. 22 and Fig. 23 appears to be similar to the one obtained from the structural analysis (Fig.11 + **SI**), although its magnitude is 10 times smaller. This result is expected since PC1 is the direction of maximum variation of the system, which thus captures the most relevant collective motions.

Some considerations between the WT and mutant structures, and between the chains of each structure can be made, with the exception of 6XHU where the first PC accounts only for 18% of the total variance. Globally, CH1 of

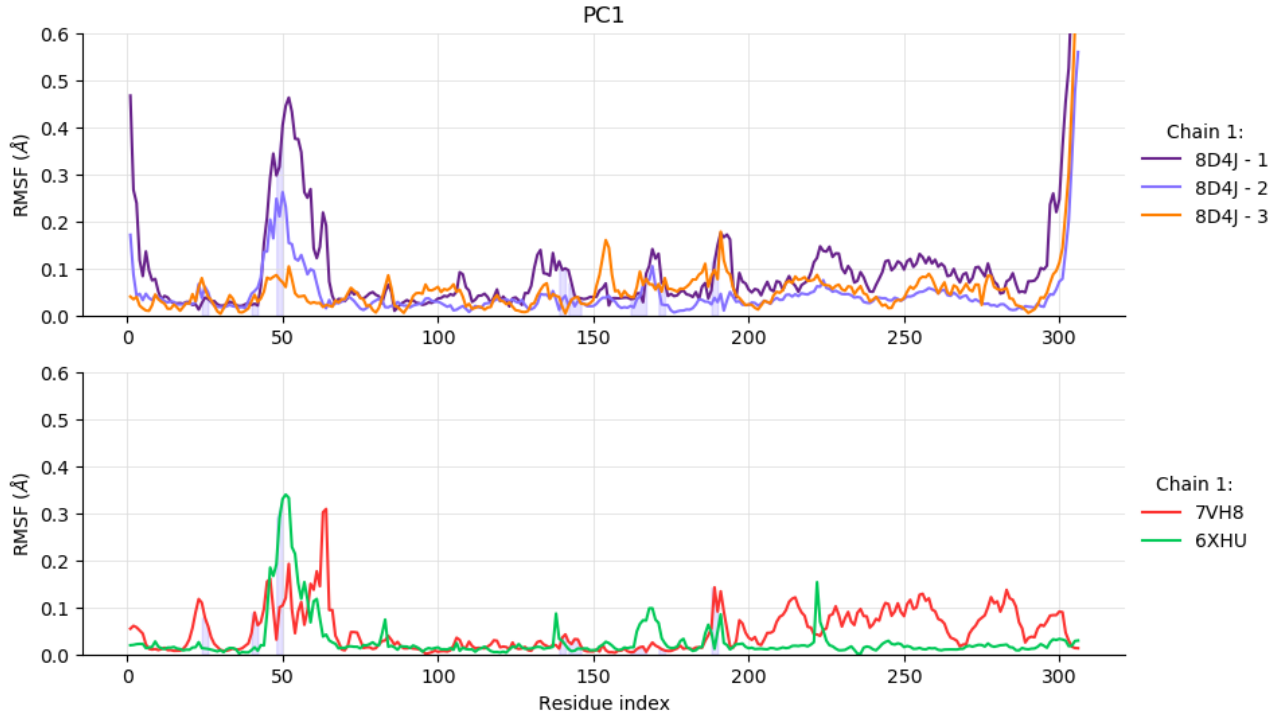


Figure 22. RMSF per residue of the first principal component evaluated for CH1 for all mutant simulations and both wild types. Residues belonging to the active site are highlighted with a purple shaded area.

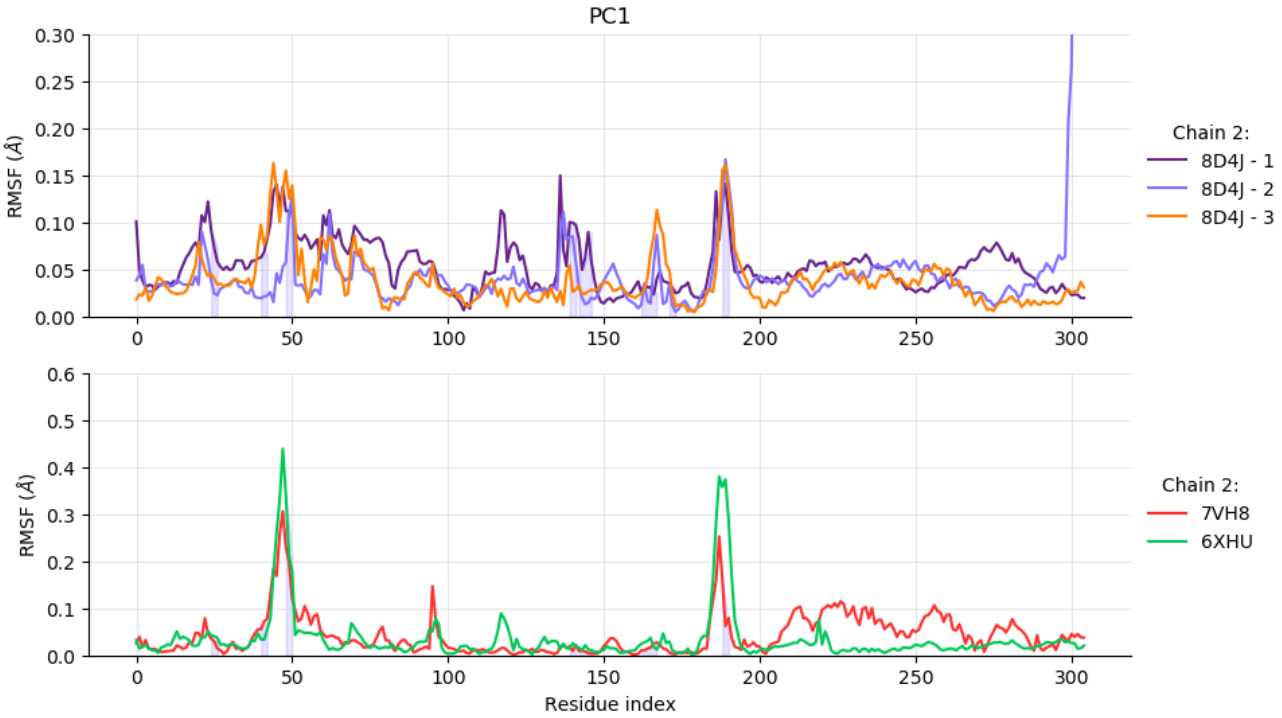


Figure 23. RMSF per residue of the first principal component evaluated for CH2 for all mutant simulations and both wild types. Residues belonging to the active site are highlighted with a purple shaded area.

8D4J has higher fluctuations in all residues compared to the WTs while for CH2 the fluctuations are comparable. PC1 captures main motions that are distinct for the two chains, corroborating the initial hypothesis of the different

chain activity.

3.3.1 Summary

Functional analysis started with an exploratory study based on lag plots that were used to qualitatively understand the role of correlation in data. Not only various time scales were hypothesised between structures, but also diverse time scales were supposed to be found in some structures between different chains.
Autocorrelation analysis and block average analysis allowed to quantify the autocorrelation time typical for each protein and for their chains pair. Even though contrasting results were obtained when exploiting these two methods, they were mostly justified by both the poor sampling quality and inaccuracies in the methods.
Spectral analysis of the fluctuating part of the RMSD was then used to assess deeper insights in the role of correlation by turning the prospective from time to frequency domain. Exploiting the periodogram estimate for the PSD, a pink-noise nature of fluctuations was highlighted. The presence of correlation is thus attributed to the 1/f-decaying trend of the PSD: the more the PSD resembles a flat distribution of power - white noise - the less correlation is present in data.
Finally, a dimensionality reduction analysis was performed with the aim of both highlighting the presence of significant clusters and of filtering out the main modes of collective motions from local fluctuations.

Conclusions

The combination of the results from the structural and functional analyses outlines a diverse behaviour between the mutant and WTs 6XHU and 7VH8.

Radius of gyration and RMSD suggest that the mutant is less stable and less compact compared to the WTs, thus confirming De Oliveira conclusions.

Concerning separated chains, RMSD plots from the structural analysis, autocorrelation time and the block from the functional analysis highlight differences between CH1 and CH2 in both the mutant and WT 6XHU. While CH1 influences the behaviour of the FP 8D4J the most, CH2 has the greatest impact on the FP 6XHU. On the contrary, both CH1 and CH2 characterise the behaviour of the WT 7VH8 structure. In this case variations in the RMSD of one chain are always counterbalanced by the other chain, resulting in a small oscillating activity and possibly corroborating the pH-activation switch of the AS in WT proteins.

The H172Y mutation has also an impact on the stability of the 2 chains. Indeed, from the RMSF plots of PCA it emerges that CH1 has fluctuations of higher magnitude compared to CH2, suggesting that CH1 could be less stable than CH2. In particular, since higher fluctuations (captured as motions projected onto PC1) in CH1 interest the active site as well, these may suggest that a ligand binding AS1 is released sooner than the one that binds the AS2. This hypothesis is also supported by the higher fluctuations observed in the physical RMSF for the mutant (Fig. 13b). A biological *guesstimate* for the different chain activity could lie in both the presence of the Tyr172 and the pH-dependent switch that, though, was not reported for the mutant. This could be the reason why the H172Y mutant is more pharmacoresistant, while also having a diminished enzymatic activity, as reported in [7].

The considerations of this project are based on insufficient sampling and an approximated analysis, hence they should be regarded as only qualitative. For a quantitative validation of the results, more simulations of longer duration should be performed to ensure the trajectory equilibration and to allow a more robust statistical evaluation.

Possible tasks to have a complete picture of the system could involve an analysis of the free energy landscape of the system by using the PCA (inspired by the work of Elena Papaleo [6]). Of course, enhanced sampling methods (e.g., replica exchange Monte Carlo, metadynamics or umbrella sampling) could be employed to see whether a different biological activity is present between the two chains.

[6] We tried to contact her but she never replied to our e-mail

References

1. Yoshimoto, F. K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. *The Protein Journal* **39,** 198–216 (June 1, 2020).
2. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature* **582,** 289–293 (7811 June 2020).
3. Anand, K. *et al.* Structure of Coronavirus Main Proteinase Reveals Combination of a Chymotrypsin Fold with an Extra -Helical Domain. *The EMBO Journal* **21,** 3213–3224 (July 2002).
4. Calligari, P., Bobone, S., Ricci, G. & Bocedi, A. Molecular Investigation of SARS–CoV-2 Proteins and Their Interactions with Antiviral Drugs. *Viruses* **12,** 445 (Apr. 2020).
5. Hu, Q. *et al.* The SARS-CoV-2 Main Protease (Mpro): Structure, Function, and Emerging Therapies for COVID-19. *MedComm* **3,** e151 (2022).
6. Yang, H. *et al.* The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences* **100,** 13190–13195 (2003).
7. Hu, Y. *et al.* Naturally Occurring Mutations of SARS-CoV-2 Main Protease Confer Drug Resistance to Nirmatrelvir (Sept. 6, 2022).
8. De Oliveira, V. M., Ibrahim, M. F., Sun, X., Hilgenfeld, R. & Shen, J. H172Y Mutation Perturbs the S1 Pocket and Nirmatrelvir Binding of SARS-CoV-2 Main Protease through a Nonnative Hydrogen Bond. *bioRxiv,* 2022.07.31.502215 (Aug. 1, 2022).
9. *HPC@Unitrento* https://sites.google.com/unitn.it/hpc/home?authuser=1. Accessed: March, 7, 2023.
10. *CHARMM-GUI* https://www.charmm-gui.org. Accessed: August, 24, 2022.
11. Humphrey, W., Dalke, A. & Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14,** 33–38 (1996).
12. *VMD* http://www.ks.uiuc.edu/Research/vmd/. Accessed: March, 6, 2023.
13. Gowers, R. J. *et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations* in *Proceedings of the 15th Python in Science Conference* (eds Benthall, S. & Rostrup, S.) (2016), 98–105.
14. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32,** 2319–2327 (2011).
15. Grossfield, A. *et al.* Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations. *Living J Comput Mol Sci.* **1** (2018).
16. Grossfield, A. & Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annu Rep Comput Chem.* **5,** 23–48 (2009).
17. Koopmans, L. H. in *The Spectral Analysis of Time Series* (ed Koopmans, L. H.) (Academic Press, San Diego, 1995).
18. Grove, M., Timbrell, L., Jolley, B., Polack, F. & Borg, J. The Importance of Noise Colour in Simulations of Evolutionary Systems. *Artif Life.,* 1–19 (2022).
19. Shao, J., Tanner, S., Thompson, N. & Cheatham, T. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J Chem Theory Comput.* **6,** 2312–34 (2007).