Analysis of the H164N mutant M^{pro} protein in interaction with the GCP-376 ligand

Giuseppe Gambini, Maurizio Gilioli

UNIVERSITÀ
DI TRENTO

Contents

1 ABSTRACT

The M^{pro} is a pivotal protein that has to be produced by the SARS-CoV2, and allows its replication. Because of its importance, the M^{pro} became quickly the target for numerous drugs, such as for example Paxlovid, developed by Pfizer. Another drug that was tested at the start of the pandemic, and that is famous for its activity against FIPV, is GC-376. As reported in the article written by Hu and colleagues[1], particular consideration has to be paid for the naturally occurring mutations, which appear fastly due to the high replicating speed of the virus, as they can be responsible for the comparison of new drug resistances. Among the mutations which are listed in the previously cited article, **H164N** was assigned to us for analysis and testing, **in interaction with the GC-376 ligand**. We performed simulations of both the mutant form of the protein (MUT) and of the wild-type (WT), with the intention of confronting the two proteins and perform basic analysis.

At first we performed a general structure analysis using tools such as RMSD, RMSF or R_g. Then, we focused on the pockets of the MUT protein and found some of the residues most involved in the interaction with the ligand. We also compared the motions of the two pockets along the entire trajectory. Secondly, we made pulling simulations for both proteins.

Despite the numerous attempts for obtaining a good simulation of the WT, unfortunately we did not manage to obtain a comparable one. In fact, both the K36 proteins were strangely dissociating from the protein. To try to understand the mechanisms behind the dissociation of the ligand, we developed an addition for the already present *PyInteraph* program capable of classifying interactions based on the frames of their occurrences. Then, we analyzed the most relevant interactions in both of the two long simulations and the pulling trajectories, and commented some of them.

Our project is associated to a Supplementary Material file, which contains most of the plots we produced. The bar plots are present also in the Appendix of this file.

2 INTRODUCTION

2.1 The Coronaviruses and the Mpro protein

Coronaviruses are single-strand RNA viruses, capable of infecting a series of species of mammals, comprising humans and felines[1,2]. The coronaviruses are divided into different families: α, β, γ, and δ. In particular, the SARS-CoV-2 virus pertains to the β lineage[3] , and shares a similarity of approximately 96% with the SARS-CoV M^{pro}[4].

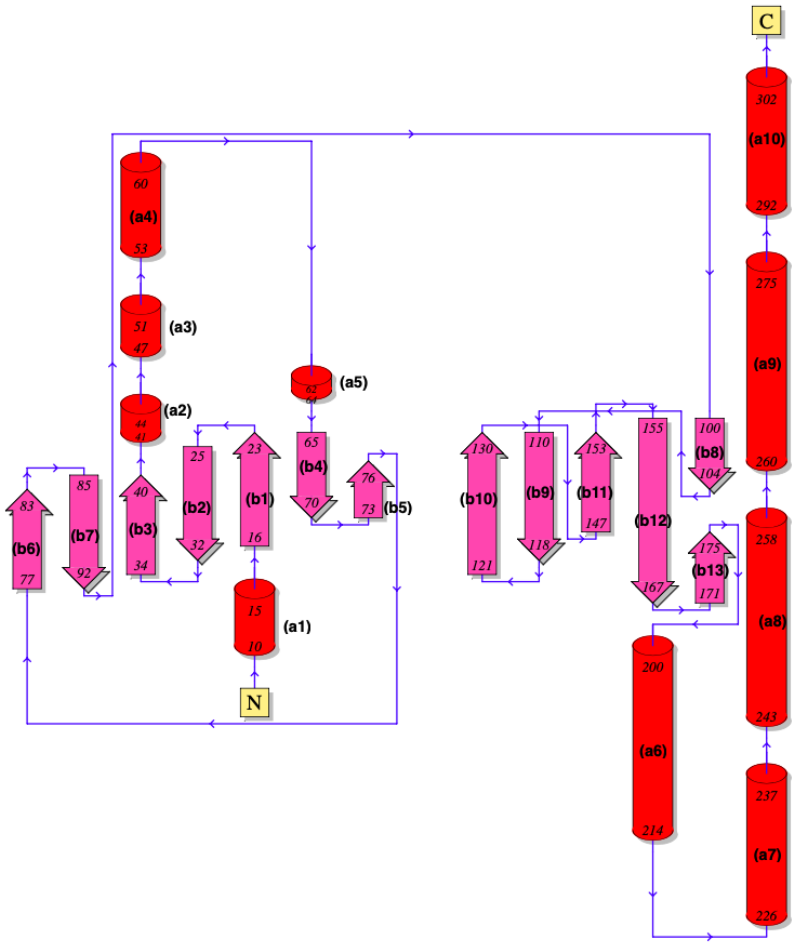


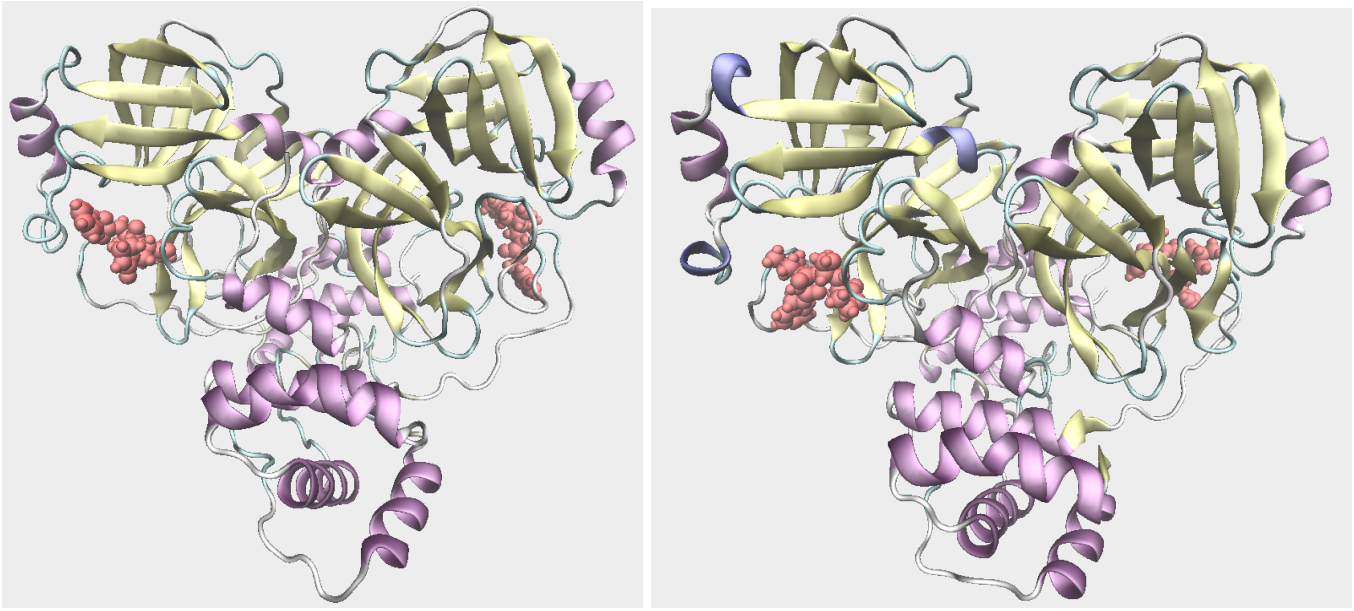
Figure 1. Image taken from the PDB sum site[5] for the *8dd1* .pdb entry. This is an indicative representation: in several cases we prefered to directly view the structure of our proteins on VMD[6].

The SARS-CoV-2 Main Protease (M^{pro}) is a 3-chymotrypsin-like protease[7] and it represents one of the most important proteins needed by the virus to correctly replicate and diffuse. Its simplified structure can be viewed in figure 1. It is pivotal in the maturation step of the virus[4,8,9]: the protease cleaves at 11 conserved sites[8] a polyprotein, formed by pp1a and pp1ab[2], into non-structural proteins (nsps)[7]. Among the produced proteins, for example, there is the RNA dependent RNA polymerase, which is obviously important for the replication of the virus.[10]. Because of its importance, M^{pro} was fastly considered an ultra-potent drug target against the SARS-CoV-2 virus[7,8].

The protein is active in its dimeric form; it has been in fact demonstrated that not only do the monomers have very low catalytic activity, but also the affinity to the proteins to be cleaved is reduced[8,11]. A representation of the dimeric form for the mutant protein and the wild type protein are shown in figures 2.

Each monomer has in total 306 residues and can be subdivided into three domains: the I (8-101), the II (102-184), and the III (201-306)[2,7,11]. The third one, Domain III, regulates the dimerization of the protein, it is formed by 5 α-helices, which are connected to domain II by a long loop (185-200)[11] . The active site, which is responsible for the hydrolytic activity of the protein, in the Wild-type protein is composed of two residues: the CYS145 and the HIS41, which are positioned near the dimer interface[8]. Better visualization of the structure of the protein can be viewed in

figure 1, which has been obtained from the PDB sum website[5] . Instead, the complete reaction of the enzyme could be consulted on the EMBL-EBI website.



(a) MUT structure **(b)** WT structure

Figure 2. Renders of the MUT and WT protein

Importantly, the N-terminal residues of protomer A (1-7) fit between domains II and III of protomer A and interact with elements of the second domain of protomer B[2,11]. Those residues have been considered important in their role in favoring the dimerization of the protein, which is essential for its functionality, as already said.

2.2 The GC-376 ligand and its interaction with the protein

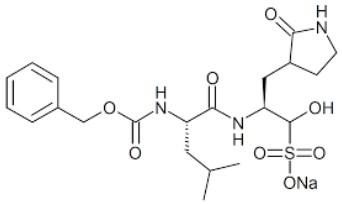


Figure 3. The GC-376 Ligand

GC-376 is a dipeptidyl aldehyde bisulfite adduct[23]. Other than its backbone, it presents in its structure a γ-lactam ring, and a benzyl group, that help the ligand in stably interacting with the protein. The pharmaceutical in question has been used in veterinary clinics to treat feline infectious peritonitis (FIP) in cats, as it was shown to inhibit the FIPV M^{pro} protein, with very low values of IC_{50} (sub-micromolar)[2].

Regarding the SARS-Cov-2 M^{pro}, an IC_{50} value of $0.030 \pm 0.008 \mu M$ was attested[3].

The inhibiting capability of this drug derives from its ability in generating covalent bonds inside the active site of M^{pro}, which consequently causes its inactivation. In fact, if released in an aqueous environment, the GC-376 (figure 3) prodrug releases the bisulfite group to become GC-373, which covalently binds to the catalytic Cys145 aminoacid of M^{pro}[2] , as shown in figure 4. Because of the efficacy of the drug against the FIPV M^{pro} protein, and the similarity between the FIPV M^{pro} and CoV-2 M^{pro}, since the very first moments GC-376 was tested against the latter.[12]. In fact, some researchers verified that the two proteins share a similarity higher than the 60%, a percentage which becomes even higher when considering the conserved sites[12] .

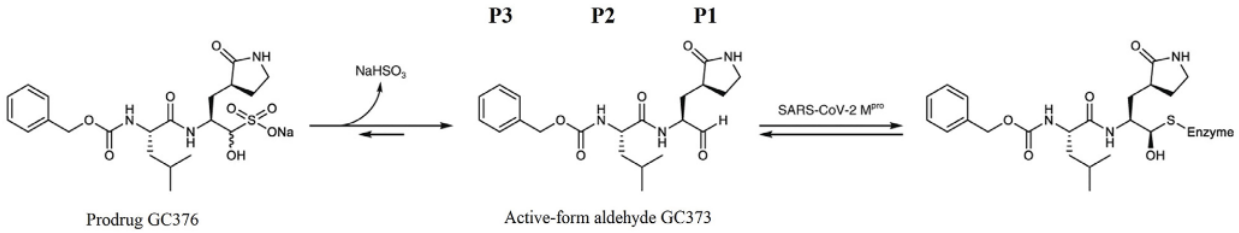


Figure 4. GC-376 transformation and binding to the M^{pro} protein

Regarding the interactions involving the ligand and the protein, several of them are conserved and necessary to allow the suppression of the catalytic activity of the enzyme.[3,7] Some amino acids, like CYS44, THR45, SER46, LEU141, ASN142, GLY143, GLU166, LEU167, PRO168, ALA191, GLN192, and ALA193 play a role in the ligand dissociation process, while instead THR26, SER46, ASN142, GLY143, HIS164, GLU166 and GLN189 were found to be fundamental for the binding[7]. Inside the binding pocket of GC-373, it is possible to observe the presence of five Sites (S1, S2, S3, S4, S5) which have to conserve some of their aminoacids to allow the ligand to bind and stably associate itself to the pocket through hydrogen bond interactions, salt bridges and hydrophobic bonds (see figure 2 in the article published by Kneller et al.[2,13]). Some interactions are conserved in the binding site and are reported in the literature; for example, the M^{pro} protein has a unique substrate preference for glutamine at its P1 binding site[3]. In our project, we tried to consult the present literature, and understand which are the most important connections for the ligand.

2.3 Mutations of the Mpro

Mutations of pivotal proteins are always important to be considered and occur naturally during the spreading of the viruses. Obviously, this is the case also for the M^{pro} protein. A study produced by Hu and colleagues and published in 2022[1] identified 100 mutations of the protein located in proximity to the bonding site, of which 20 were particularly active and resistant to the drugs. They further restricted the plethora of mutations by considering the alterations involving the residues at a distance lower than 6 Åfrom the ligand, which are H41, M49, T135, N142, S144, H163, H164, M165, E166, H172, Q189, and Q192. Among the mutations listed, the professors assigned us the **H164N** mutated protein (MUT). Its alteration, in particular, appears with a very high frequency, keeps the protein sensitive to all three inhibitors ($<$ 4.1-fold change in k_i with respect to the WT protein), and maintains the catalytic activity of the protein (4.2-fold lower in kcat/km)[1] , which means that it is not deleterious for the virus.

2.4 *PyInteraph* introduction and settings

PyInteraph[14] is a "software suite designed to identify intramolecular interactions from protein ensembles and join them in a graph representation, which can be used to identify pathways of structural communication" (from the Github description webpage). At the end of the analysis, the program outputs a matrix of all the interactions, listed with an associated percentage of persistance along the trajectory given in input.

PyInteraph calculates the hydrogen bonds (H-bonds), the hydrophobic interactions, and the salt bridges which arise during the simulation and affect the collective behavior of the protein. Their calculation are described below:

- **H-bonds**: they exist whenever the distance between the acceptor and the donor atoms is smaller than 3.5 Å. Also, a cut-off default value of the angle formed by the donor, its hydrogen and the acceptor of 150° is taken (The mentioned parameters were taken as default by MDAnalysis, version 2.0.0.[15]). To calculate the H-bonds, *PyInteraph* employes the built-in function *hbond_analysis.HydrogenBondAnalysis* of MDAnalysis. The selections for the acceptors, donors, and hydrogens are included in the following list:
  - **Acceptors** = OH, OG, OD1, OD2, OG1, O, ND1, NE2, OE2, OW, SG, OE1, OH2, SD
  - **Donors** = OH, OG, NE2, OG1, NE, N, ND1, NZ, NH1, NH2, OW, ND2, SG, OH2, NE1
  - **Hydrogens** = H*
- **Salt bridges**: Salt bridges are considered to be occurring between oppositely charged ionizable groups, most commonly between the negatively charged side chains of aspartate, glutamate, and the positive charges in arginine, lysine or histidine side chains. In particular, *PyInteraph* uses the following selections for the mentioned amino acids:
  - **Aspartic acid**: CG, OD1,OD2, !HD1

  - **Arginine** = CD, NE, CZ, NH1, HH11, HH12, NH2, HH21, HH22
  - **Glutammic acid** = CD, OE1, OE2, !HE1
  - **Hystidine** = CG, ND1, HD1, CD2, CE1, NE2, HE2
  - **Lysine** = CE, NZ, HZ1, HZ2, HZ3

- **Hydrophobic contacts**: The distance of the center of mass of the sidechains belonging to hydrophobic amino acids (**ALA, VAL, LEU, GLY, PRO, ILE, PHE, MET, TRP**) is monitored. If the distance between the COMs is smaller than 5 Å, then *PyInteraph* considers a contact between the two residues. To calculate hydrophobic bonds, the atoms of the backbone, i.e. CA, C, O, N, H, H1, H2, H3, O1, O2, OXT, OT1, OT2, are excluded from the calculation.

2.5 Umbrella sampling

The umbrella sampling[7] approach is widely employed for the calculation of free energy differences in many biological systems such as ions passing through channels or peptide dissociation. As argued in the article by Tun Ngo[16], umbrella sampling is well suited for solvent-exposed ligand binding sites. In general, the umbrella sampling method is one of the most used methods to characterize the unbinding energy of a ligand with the substrate. It is generally valid to say that the higher is the energy required for the dissociation of the ligand the higher is the affinity of the binding element to the substrate. A study performed by Tam and colleagues in 2022[7] did the calculation of the free energies for a series of pharmaceuticals against the M^{pro} protein were attested in a range of energy values between -8.63 kcal mol^{-1} and -1.80 kcal mol^{-1}, with the mean value of $\Delta G_{\mathrm{US}} - 4.59 \pm 0.41$ kcal mol^{-1}.

In our project, we performed pulling simulations in 5.5 and discussed ideas about a possible umbrella sampling for the calculation of the unbinding free energy in the appendix 8. We did not perform the multiple simulations required in the umbrella sampling approach because the convergence of the PMF is an expensive process from the computational point of view.

3 Aim of the project

This project was made as a requirement for the course of Computational Biophysics 2022/2023, held by professors Gianluca Lattanzi, Luca Tubiana and Thomas Tarenzi at the University of Trento. Our aim was first to gather information about the WT and the mutant protein H164N within the same environment (simulation conditions) and interacting with the ligand GC-376. Secondly, we performed pulling simulations, detaching the ligand from both the proteins. During our study, we built some programs to evaluate the interactions regarding "important residues", and some "important interactions", given their involvement in some important processes of the protein. The meaning of these two terms is explained in figure 20, section 4.9, and in the glossary 7.

4 METHODS AND RESULTS

The conda packages we used are listed in the supplementary file.

4.1 Algorithms and specified options

A short recap of the algorithms employed in the *.mdp* files provided by the professors. The same protocols and algorithms were applied for both the MUT and WT simulations.

	type	**parameters**
minimization algorithm	steepest descent with threshold force ε	$h = 0.01$ nm, $\varepsilon = 1000$ kJ mol^{-1} nm^{-1}
position restraints	on sidechains (k_{sd}) and backbone atoms (k_{bc})	k_{sd} =40 kJ mol^{-1} nm^{-2} k_{bc} =400 kJ mol^{-1} nm^{-2}
thermostat in NVT	V-rescale	$\tau_{\mathrm{T}} = 0.1$ ps
barostat in NPT	Parrinello-Rahman	$\tau_{\mathrm{P}} = 2$ ps $\mathrm{P}_{\mathrm{ref}} = 1$ bar
temperature coupling	Protein and ligand K36	$\mathrm{T}_{\mathrm{ref}}$=310 K
long-range electrostatics	fast smooth particle mesh Ewald (PME)	Fourier spacing = 0.16 nm
cut-off scheme for non-bonded interactions	Verlet neighbor list	$r_{\mathrm{Coulomb}} = 1$ nm $r_{\mathrm{vdw}} = 1$ nm
integrator	leap-frog integration	Δt =0.002 ps
constraints	LINCS	over all the bonds that involve hydrogen bonds
compressibility		$\kappa_{\mathrm{T}} = 4.5 \times 10^{-5}$ bar^{-1}

Table 1. Table with the parameters used during the simulations. The position restraints with k_{bc} have been implemented also on the ligand carbons, nytrogens and oxygens during the NVT and NPT equilibrations

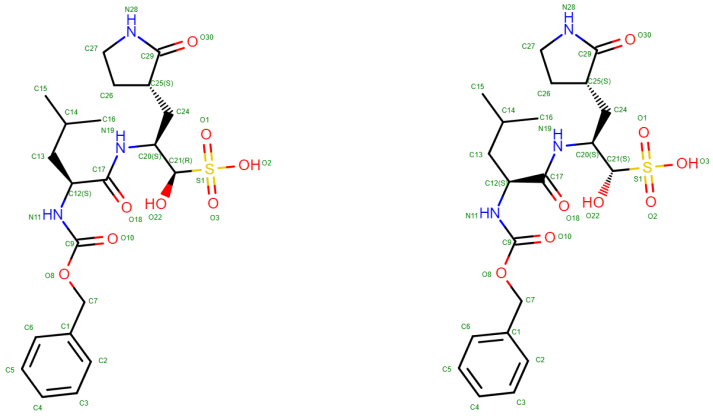
4.2 Mutant (MUT) and wild-type (WT) proteins simulations

4.2.1 Initial data retrieval

- **MUT**: We produced the simulation of the mutant (MUT) (chapter 5.1) protein at first, and secondly we proceeded to analyze the Wild type (WT) (chapter 5.2). The input *.gro* file for the MUT delivered by the professors was employed to start the simulations of the mutant protein. The topology and the force field files were provided as well. Hydrogens, ligands, ions and water were already included in the starting *.gro* file.

- **WT**: The sequence of the WT protein was obtained at the *6wtt* entry of the PDB Protein Data Bank[3,17]. X-ray diffraction experiments were used to determine the structure. We extracted the dimeric configuration with two ligands from the *.pdb* file associated with the *biological assembly 2* listed in the site. The solvation of the protein-ligand complex was performed using the *solution builder* tool of Charmm-gui[18,19] web server. According to the FASTA file contained in the *.pdb* entry, the unresolved residues VAL303 and THR304 have been added to the two monomers. Moreover, we noticed that also the last two residues PHE305 and GLN306 were completely missing both in the *.pdb* structure file and in the FASTA file associated. For an automated homology-modeling process, one could use the Swiss Model[20] web server.
  One of the ligands listed in the pdb structure is a stereoisomer of K36 called B1S. Their stoichiometric formula $C_{21}H_{31}N_3O_8$ and the connectivity of the atoms are exactly the same in the two cases[3]. An overview of the two ligands can be seen in figure 5. The two stereoisomers are considered to be the same GC-376 ligand, for this reason, we kept B1S. The force field for both ligands has been obtained by means of the CSML search tool implemented on the Charmm-Gui server: if the ligand is not included in the database (and this was our case), the server generates a force field based on the *.pdb* coordinates.

  The hystidine in position 164 was, accordingly to the *.pdb* file, in the δ-state (figure 6). For a more specific query on the protonation states, web servers such as PropKa[21] are advisable.



(a) B1S ligand **(b)** K36 ligand

Figure 5. Representations of the structures of K36 and B1S

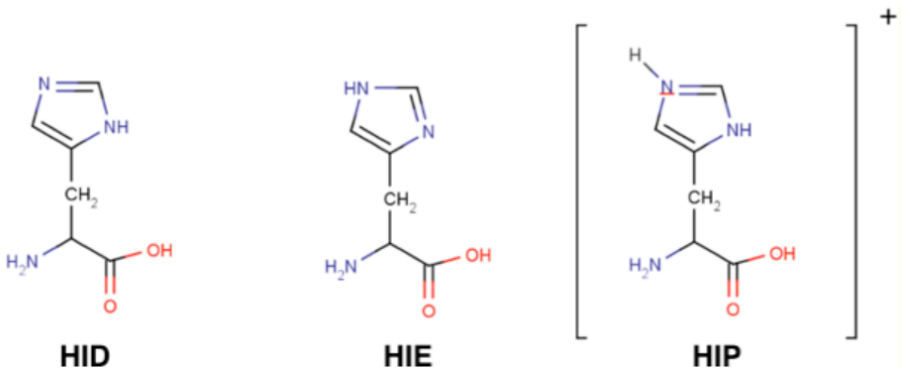


Figure 6. Possible protonation states of a histidine. The HID state is the one with a protonation in δ, while instead, the HIE represents the ϵ state. The fully protonated configuration is labeled as HIP.

4.2.2 General procedure

The protein complex is not elongated in a specific direction, hence, a cubic box represented the natural choice. In the WT case we opted for the optimal choice of the box performed by Charmm-Gui. This choice should prevent contacts with the periodic image of the protein. The cubic box of the mutant had a size of 10.7 nm, while the side of the box for WT was 10.4 nm. In the case of the WT, the size was given and suggested directly by Charmm-Gui. In both cases, the boxes were solvated with an ion concentration of 0.15 M. For the WT simulation we decided to make more NPT equilibration steps. In fact, in this case, the pdb structure is not totally trusted and we presumed it needed some more time to reach a good configuration in the phase space. We did three extra equilibrations of 5 ns in which we decreased the position restraints to k_{sd} =30 kJ mol^{-1} nm^{-2} and k_{bc} =300 kJ mol^{-1} nm^{-2}, then k_{sd} =20 kJ mol^{-1} nm^{-2} and k_{bc} =200 kJ mol^{-1} nm^{-2} and so on.
In both MUT and WT cases, we performed an unrestrained molecular dynamics of duration 300 ns.

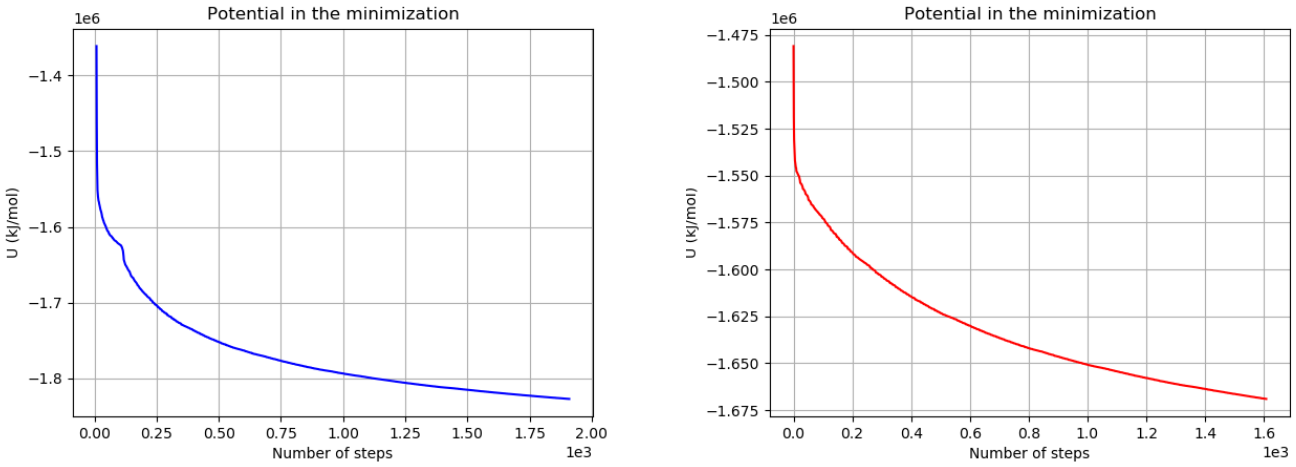
4.2.3 Minimization and equilibration

1. The **minimization**: The *steepest descent* algorithm stops whenever the maximum force between two atoms of the system is smaller than 1000 kJ mol $^{-1}$ nm $^{-1}$. To evaluate the results obtained throughout minimization, we observed the potential energy of the system at each step, which is depicted in figure 7. The mutant protein reached convergence at the 1900$^{\text{th}}$, while the WT protein at around the 1600$^{\text{th}}$ step.

2. An **NVT equilibration** of 0.5 ns was performed in order to reach the desired temperature of 310 K. The *heat bath* is composed of the solvent and ions and it is coupled to the ligand-protein complex. The temperature plots are shown in figure 8.

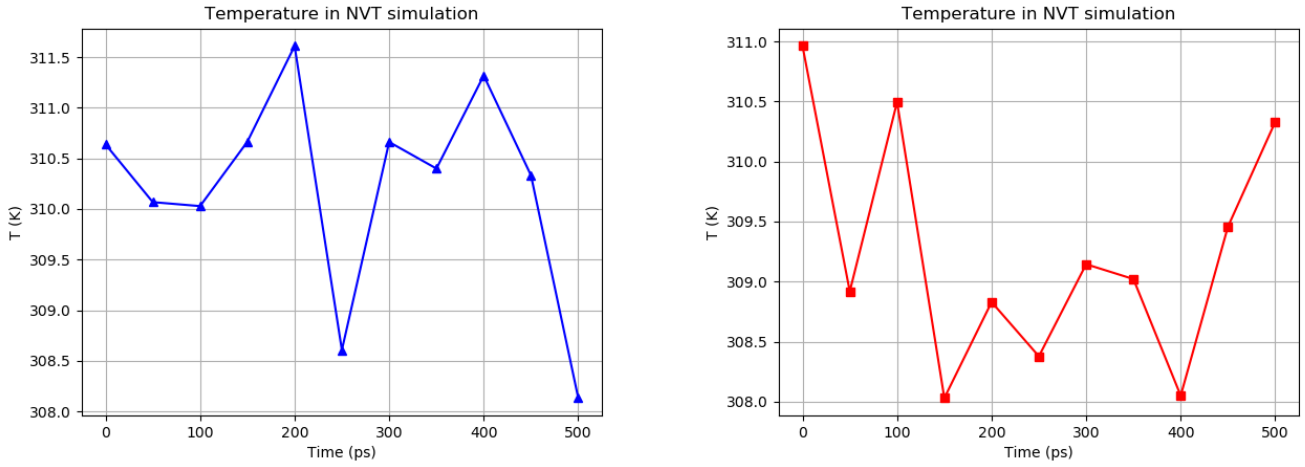
3. An **NPT equilibration** was lastly performed with a duration of 1 ns. This step is used to achieve the right T and P before the unrestrained molecular dynamics starts. We evaluated density (11) as a good quantity for the assessment of the NPT equilibration process. In fact, the volume adjusts so that pressure and temperature are conserved. The changes in volume depend on pressure and temperature through the compressibility showed in table 2.

$$\text{Compressibility } \kappa_{\mathrm{T}} = -\frac{1}{\mathrm{V}}\left(\frac{\partial \mathrm{V}}{\partial \mathrm{P}}\right)_{\mathrm{T}} \tag{1}$$



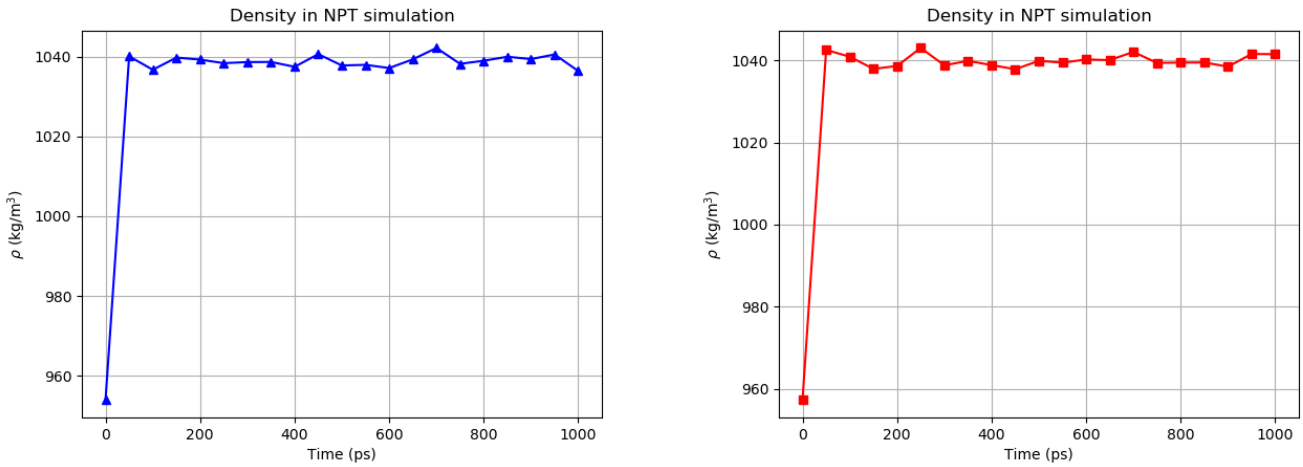
(a) MUT **(b)** WT

Figure 7. Potential energy calculation for the MUT and the WT. The mutant protein reached convergence at the 1900^{th}, while the WT protein at around the 1600^{th} step.



(a) MUT **(b)** WT

Figure 8. Temperature plot of the mutant complex. The system was simulated at a temperature of 310 K. The temperature span a range of 1.5 K around the temperature of 310 K.



(a) MUT **(b)** WT

Figure 9. Density plots obtained after NPT equilibration of the WT protein (left) and of the MUT protein (right). The initial jump is due to the sudden introduction of the barostat.

4.2.4 Processing the simulations

Unwanted movements of the protein outside the simulation box were removed by means of the *trjconv* tool provided by Gromacs[22] by exploiting the functionality *-pbc nojump*. The same tool allowed the alignment of the trajectory using a least squares fit with respect to the α-carbons of the structure. To perform this type of modification, we used the *-fit rot+trans* option.

4.3 General analysis over the trajectory

The following analysis refers to the MUT protein, in this section we do not present results for WT.

4.3.1 Analysis of the RMSD and Rg of the MUT protein

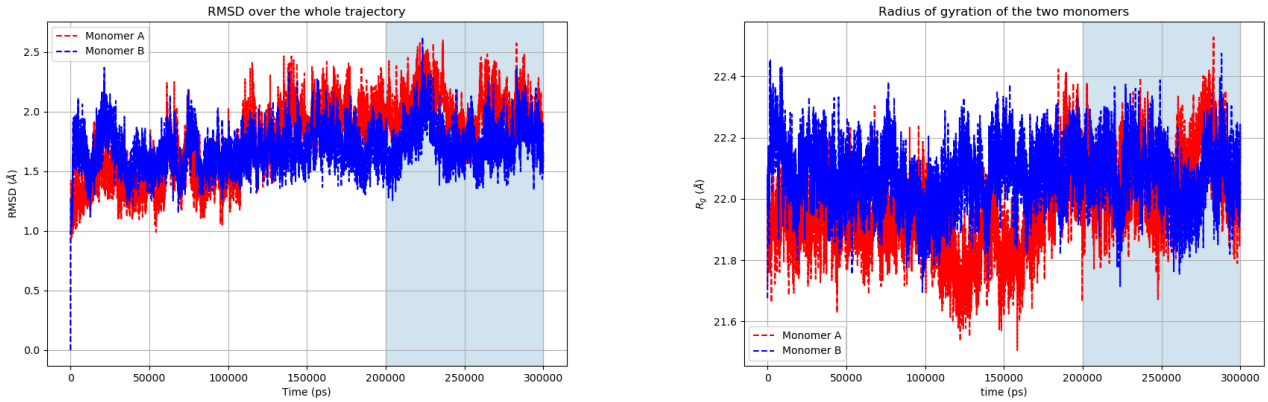
The RMSD can be computed over an arbitrary selection of atoms and the weights are usually chosen to be the masses of the atoms. If all the selected N elements have the same mass, then

$$\text{Root Mean Squared Deviation RMSD}(t) = \sqrt{\frac{\sum_i^N m_i(\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}})^2}{\sum_i^N m_i}} = \sqrt{\frac{1}{N}\sum_i^N (\mathbf{r}_i(t) - \mathbf{r}_i^{\text{ref}})^2} \tag{2}$$

$$\text{Radius of giration } R_g(t) = \sqrt{\frac{\sum_i^N m_i(\mathbf{r}_i - \bar{\mathbf{r}}_i)^2}{\sum_i^N m_i}} = \sqrt{\frac{1}{N}\sum_i^N (\mathbf{r}_i(t) - \bar{\mathbf{r}}_i)^2} \tag{3}$$

To see the definitions of RMSD and of the radius of gyration, see the Glossary in chapter 7. The notation $\bar{\mathbf{r}}_i$ indicates an average over the positions of the atoms in the selection in the current frame. The quantities in (2) and (2) can be used to select the equilibrium part of the trajectory.

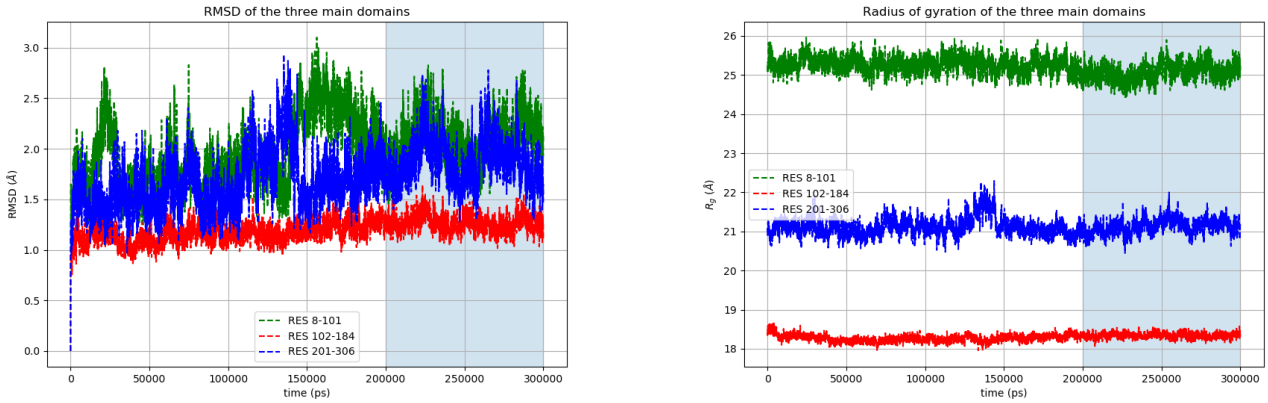
We first analyzed the RMSD and R_g for a selection consisting of all the α-C of the two monomers. The RMSD and the R_g of the two chains do not have to be equal, since each monomer can reach its own equilibrium separately, i.e., its preferential basin in the phase space. Due to the high symmetry of the system, it is highly probable that a longer simulation would eventually lead to more similar results for both RMSD and R_g. The configuration reached at the end of the simulation could be interpreted as a *"metastable"* equilibrium, i.e., an intermediate state represented by a local minimum in the phase space.



(a) **(b)**

Figure 10. Figures representing the values of the RMSD for monomers A and B of the mutant protein. The light blue region starts after 200000 ps frame, where we identified the start of the "*metastable*" portion of the trajectory.

Importantly, the plot of the RMSD does not give information about the fluctuations of specific regions of the protein. More specific selections can highlight different behaviors. The RMSD can, for instance, be computed for the α-C of the three main domains of the protein. The selection does not include the N-terminal loop going from residue 1 to 8 and the loop going from residue 185 to 200 (see figure 1), in order to get a more representative mean square deviation of the biologically important structures. Domain III from residue 201 to 306 has no β-sheets in it, while domain II has no α-helices. Domain I contains both α-helices and β-sheets.



(a) **(b)**

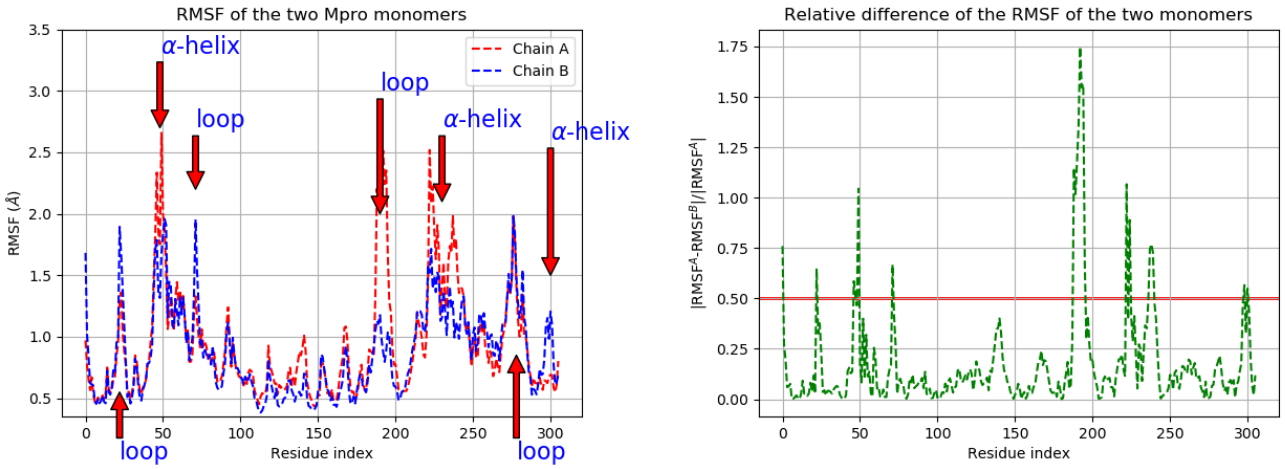
Figure 11. The second domain, with a higher number of β-sheets displays a more stable RMSD compared to the other two regions. From the plot of the R_g on the right, we can conclude that the three regions do not expand or shrink significantly during the trajectory.

4.3.2 RMSF analysis

The knowledge of RMSF (definition in chapter 7) allows us to obtain some preliminary information about the structural details of the protein. Due to the evident symmetry of the molecule, a comparison between the RMSFs of the two monomers is appropriate. The calculation of the RMSF was performed on the equilibrium part of the trajectory from 200 ns to 300 ns, in order to avoid non-representative fluctuations. We define the vector $\mathbf{r}_i$ as the position of the α-C of the i^{th}-residue and $T = N\Delta t$.

$$\mathrm{RMSF}_i = \sqrt{\frac{1}{T}\sum_{j}^{N}(\mathbf{r}_i(j\Delta t) - \langle \mathbf{r}_i \rangle)^2} \qquad \text{Relative difference} = \frac{|\mathrm{RMSF}_i^A - \mathrm{RMSF}_i^B|}{\overline{RMSF}} \tag{4}$$

where the average over all the values of the RMSF, i.e., $\overline{RMSF}$ is the average of all the RMSF values and it represents just a constant factor.



(a) (b)

Figure 12. On the left, the RMSFs of the two monomers. On the right, the relative difference. The sharp peaks identify the residues of the two chains that move differently along the trajectory.

We noticed that the stability of domain II (from 102 to 184 on the x-axis) in RMSD and R_g is reflected in the RMSF: not only the absolute values of RMSF are small compared to the other domains, but also both monomers undergo the same small fluctuations in that region. A low relative difference in RMSF is in fact observed in that region (figure 12b). In order to roughly identify relevant displacements, we can choose 0.5 as a fixed threshold[1] for the absolute value of the relative difference of RMSFs. The majority of the identified residues belong to loop structures, as expected. However, there are some exceptions represented by α-helices, for instance, the residues 297, 298, 299, and 300 belong to one of the α-helices, which are stable constituents of the protein (figure 12b).

The method provided for the determination of most fluctuating residues has to be taken with caution: it could be employed for a rough identification of the loops, but it becomes less precise when dealing with slow-moving residues. In fact, since the RMSF is a scalar quantity that depends on vector quantities, some information about the directionality of the fluctuation is lost.

4.4 Evaluation of the point of equilibrium

The RMSD and R_g are computed along the equilibrium trajectory for the α-C of the whole protein and the error is estimated by means of a *block standard error*, i.e., the standard error on the average is corrected with the factor $\sqrt{\tau}$, where τ is the correlation time. The following table presents the averages of some significant quantities. The temperatures, the potential energy and the volume of the box have low-correlated fluctuations, and therefore, the standard error represents a good estimate of the real error.

[1] This threshold is completely arbitrary and could be adjusted to refine the procedure.

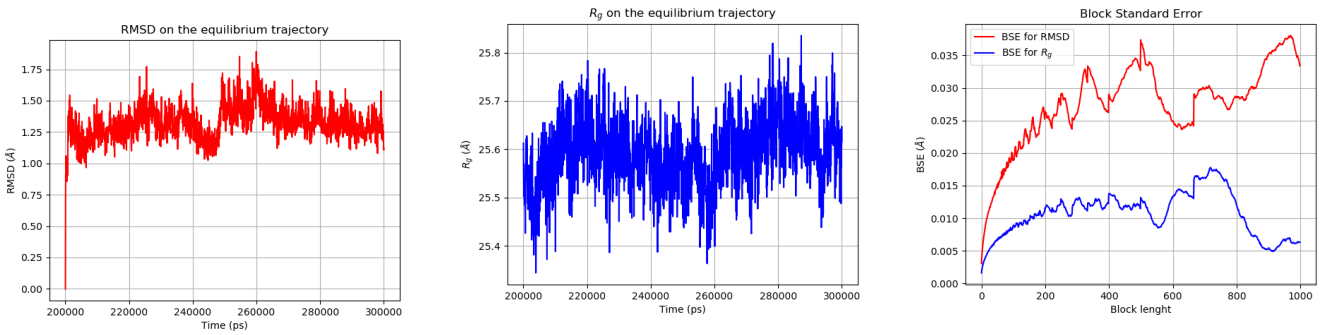


Figure 13

Property	**Value**
Temperature of protein and K36 complex (K)	310.04 ± 0.06
Temperature of solvent (K)	309.96 ± 0.02
Potential energy (kJ/mol)	(-1.54423 ± 0.00003)×10^6
Volume (nm^3)	1125.54 ± 0.04
RMSD (Å)	1.32 ± 0.07
R_g (Å)	25.59 ± 0.03

Table 2. Averages of the main quantities.

Although RMSD, R_g and RMSF are valuable tools to obtain information about overall fluctuations of the protein structure, they do not give information about the directionality and the type of movement happening in the protein.

4.5 Analysis of the behaviour of the binding pockets

We were interested in studying the pockets where the ligands reside. We formulated a first hypothesis about their behaviour. First of all, we analyzed the number of hydrogen bonds between the atoms of the ligand and the protein over the whole trajectory.

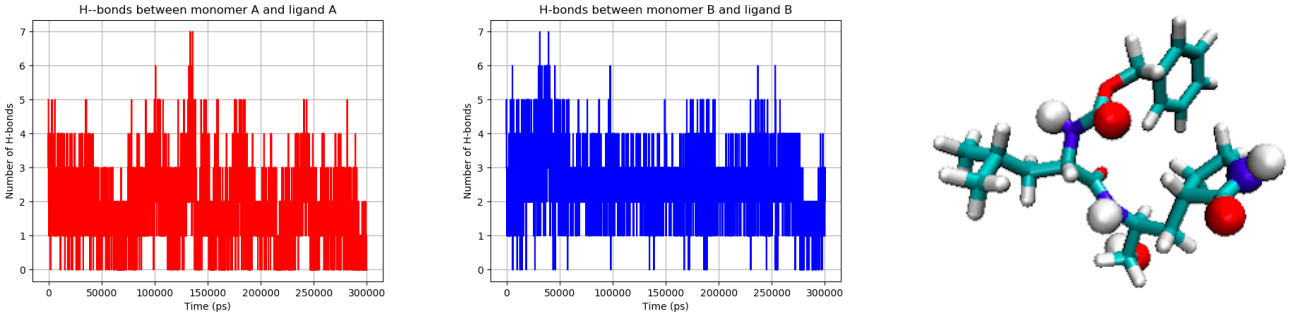


Figure 14. Comparison between the number of hydrogen bonds for monomer A and B with the ligand, respectively on the left and on the center. On the right, a representation of GC-376 in which the hydrogens, oxygens and nitrogens involved in hydrogen bonds are highlighted.

In the case of monomer A, the number of frames with no hydrogen bonds is higher, as reported in figure 14. There could be multiple possible explanations to this: for instance, the ligands might move differently inside the pockets, or some residues might act differently near the ligand. There could be even long-range effects involving a large number of residues.

We want to understand the origin of the different behaviour of the two pockets, the first step is the identification of the residues taking part in the pocket. A visualization of the protein and the ligands is useful to get a first intuition in

this perspective. By directly inspecting the protein region near the ligand, we selected four large groups of residues: (40-50), (140-150), (160-170) and (184-194)[2].

We improved this first selection with the following method for the identification of the residues of the pocket: for each amino acid, the distance between the α-C and each carbon atom of the ligand is computed[3]. A contact is added to the count every time that a given distance is below a cutoff of 6 Å. In the end, all the counts are divided by the number of frames in the trajectory, so that a percentage is obtained.

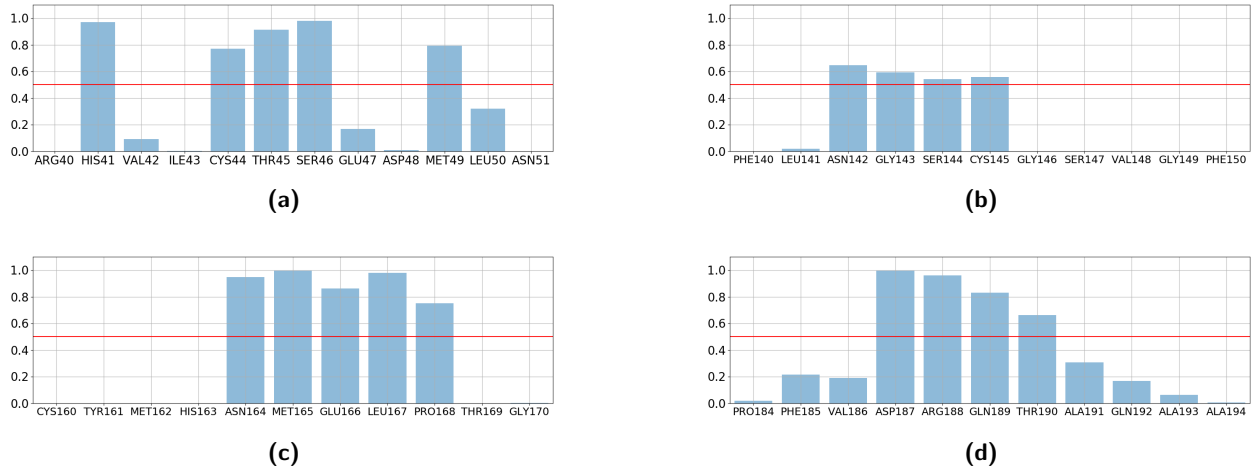


Figure 15. The contacts were evaluated on the portion of trajectory considered at equilibrium. Each column represents the contacts of a particular amino acid with the ligand. The red line indicates a threshold corresponding to 50%.

We considered all the nearest residues with percentages higher than 50% as part of the pocket.
Using the described method on our data, we noticed that some of the residues have no contacts with the carbons of K36, especially in the sequence going from residue 40 to 50. Nonetheless, those amino acids can be considered as a part of the pocket, as they are inserted in a sequence with high percentage of contacts. The total residues selected with this procedure are the following (41-49), (142-145), (164-168) and (187-190). The procedure described above is only an approximation and it can be refined with an analysis of the connections between the residues, i.e., the network of the non-covalent interactions. *PyInteraph*[14] is a valuable tool that allows this type of analysis.

4.6 Networks involving the residues of the pockets

After the calculation of hydrogen bonds, hydrophobic contacts, and salt bridges, *PyInteraph* provides a visualization of the degree of connectivity of the residues. The tool *filter_graph* is used to get a filtered graph of the interactions while *graph_analysis* labels each residue with the number of connections. With reference to figure 16, the residues are colored on the base of their connectivity.

[2]The (i-j) notation includes all the residues in the sequence from i to j.
[3]The distance between the Center of Mass (CM) of the ligand and the α-C of each amino acid is not a significant quantity: different arrangements of the ligand may provide the same CM coordinate. This is largely due to the fast-moving benzene.

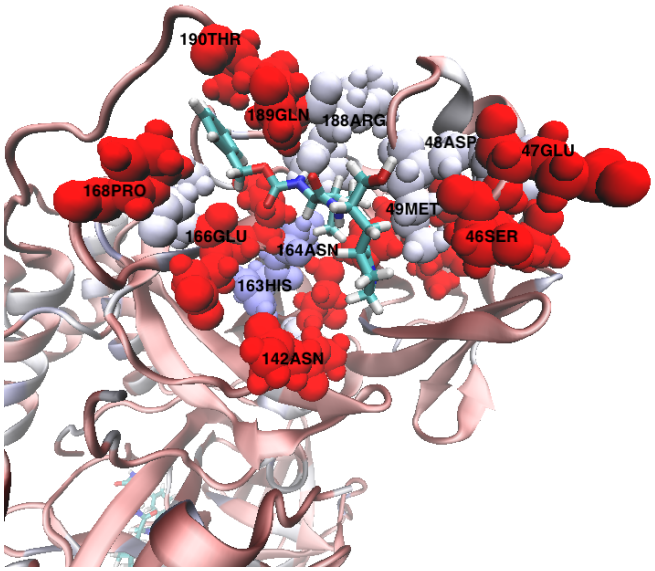


Figure 16. The residues of the pocket are highlighted with a VDW color drawing method. The residues colored in red are involved in less than four connections, the white colored produce exactly four interactions, while the violet residues generate five of them.

By observing the results obtained with *PyInteraph*, we believe that the residue 163HSD should be included in the list of pocket residues, since it is involved in five connections (high connectivity).

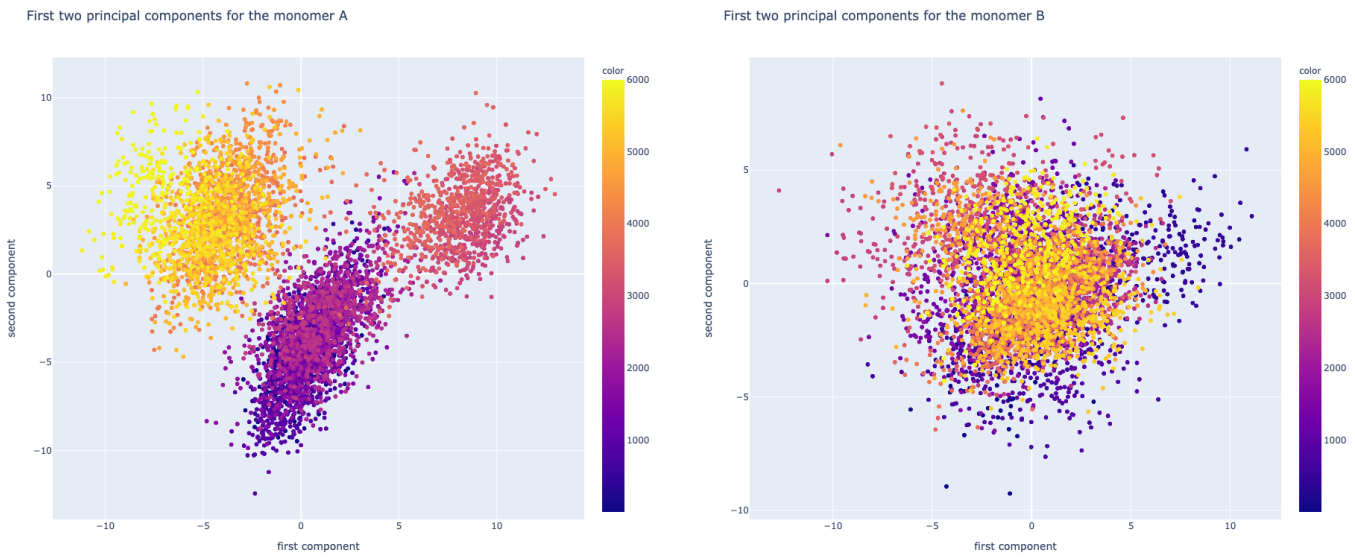
4.7 PC analysis on the pocket residues

Then, starting from the plots of the hydrogen bonds in (14), we evaluated the possibility of a cooperative movement of those pocket residues. We wanted to support this hypothesis with a Principal Component Analysis (PCA). We considered the projections of the principal components[4] $\mathbf{u}_k$ over the displacements $\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle$, where $\langle \mathbf{r}_i \rangle$ is a time averaged position. We selected the α-C of the pocket residues. We did not assign any weight to the positions since all the atoms considered have the same mass.

$$C_{ij} = \langle (\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle)(\mathbf{r}_j(t) - \langle \mathbf{r}_j \rangle) \rangle \qquad PC_k(t) = \sum_{i=1}^{N} (\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle) \cdot \mathbf{u}_k$$

where N is the total number of α-C. The following two plots show the quantities $PC_1(t)$ vs $PC_2(t)$ and the dots are colored according to the time expressed in number of frames.

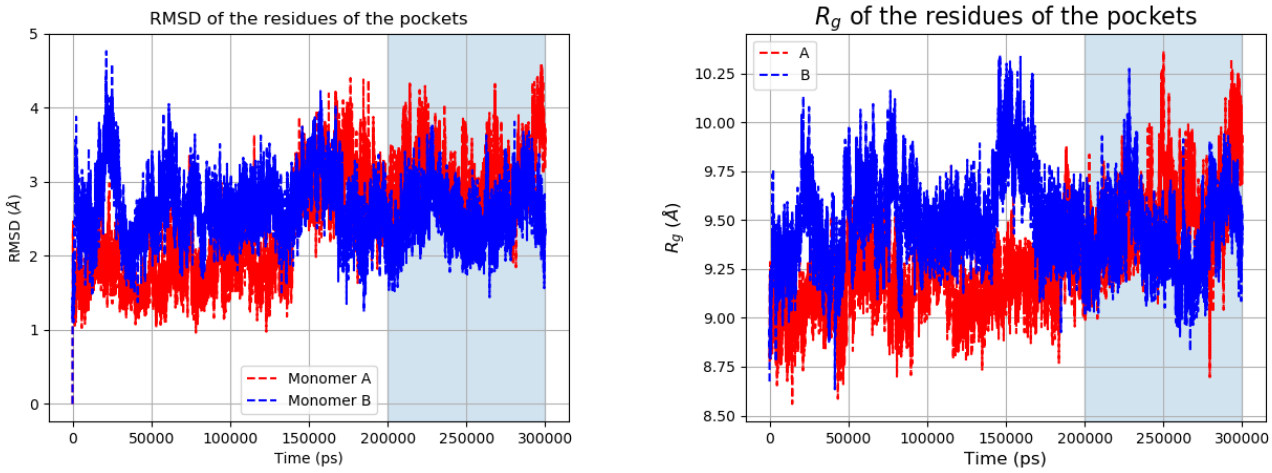
[4] the eigenvectors of the covariance matrix C_{ij}.



(a) **(b)**

Figure 17. First two principal components of a selection composed of the α-C of the pocket residues over the whole trajectory. As it is possible to see in the image on the left, three main clusters could be identified.

The presence of a cluster with points between 4000 and 6000 in the color scale supports our choice to consider the trajectory between 200 ns and 300 ns as equilibrated. The variance captured by the two components is about 63 % for the A case, while it is about 44 % for B. From figure (14) it was not possible to tell wheter the differences in the number of hydrogen bonds were due to a reorganization of the pocket or to a different movement of the ligand. Now, the hypothesis of a reorganization of the residues seems the most probable. Lastly, in order to confirm the hypothesis of a reorganization of the pocket A, we show the RMSD and the R_g of the α-C of the pocket residues.



(a) **(b)**

Figure 18. RMSD and R_g for the pocket residues in both monomers. Both quantities, as expected, seem more stable in the B pocket case.

4.8 Steered molecular dynamics (SMD) setup

The procedure described below has been followed for both SMD of MUT and WT proteins in complex with GC-376. As in the article by Tam et al.[7], we isolated the monomeric form of M^{pro} in complex with a single K36 ligand. Here we

highlight the main differences between the two setups:

- **MUT**: we took the coordinates written in the initial *input.gro* file and isolated one monomer and one ligand. Next, we included the force field contained in the february 2021 version of Charmm36 with the command *gmx pdb2gmx*.
- **WT**: we used the biological assembly 2 listed in the RCSB site for the complex 6wtt. Then, we generated the topology and the force-field files with the Charmm36[19,23] version of February 2021 taken from the MacKerell site. Once the *.gro* file with the protein was obtained, we added the coordinates of the ligand and updated the *topol.top* file with the force-field parameters of the K36 molecule obtained via CGenff[19].

Since the rest of the two setups is the same, we summarize it in a few points:

1. With the Gromacs command *editconf*, we increased the length of the z side of the cubic box from 10.7 nm to 15 nm. This distance prevents the ligand from coming close to the periodic image of the protein.
2. The system has been solvated using the "Simple Point Charge water" *spc216.gro*, i.e., the three-site water model used as default option in Gromacs. After the solvation, we added ions to reach a physiological concentration of 0.15 M.
3. This time the position restraints were implemented only on the protein, with the parameters listed in table 1.
4. As in the other two simulations, a minimization step followed by 0.5 ns of NVT and 1 ns NPT equilibrations with the same couplings and parameters as described in table 1 including the cut-off schemes, compressibility, integrator and constraint algorithm.
5. We employed a pulling velocity of $v = 0.005$ nm/s and a k =600 kJ mol^{-1}nm^{-2}, as suggested in Tam et al. paper[7].
6. We have set a simulation time equal to 0.7 ns for the pulling trajectory, capturing 700 configurations.
7. We performed a constant velocity pulling of the COM of the ligand along the $\hat{z}$ direction. The reaction coordinate is the distance between the COM of the protein and the ligand.

4.9 Description of the *interactions vs time* algorithm

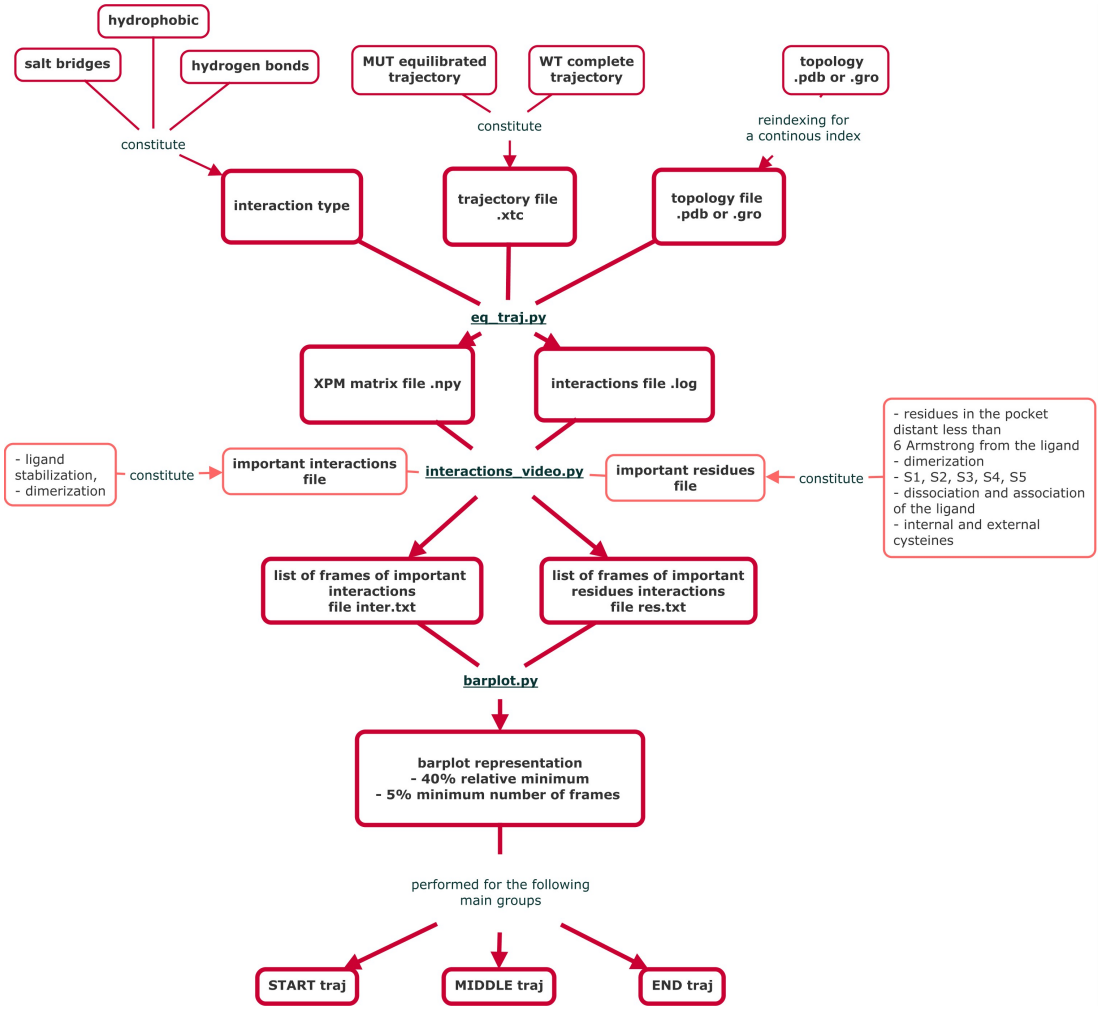


Figure 19. Structure of the *interactions vs time* algorithm, which is an addition to the scheme of *PyInteraph*.

Our aim was to evaluate the presence or absence of each interaction at every time step. To do that, we knew that *PyInteraph* was suitable to analyze each kind of contact. However, *PyInteraph* unfortunately cannot be used to solve the problem of the evaluation of the interactions at each time step.

The steps of the algorithm are reported in the graph 19. The names of the programs we built and used are displayed in the interconnections.

Data were generated for each macro type of interaction following the settings of *PyInteraph* described in 2.4: hydrogen bonds, hydrophobic bonds, and salt bridges. Before doing that, we assigned the masses of the atoms through the Charmm27 all-atom force field[19].

Using the custom-made *log_xpm.py* function, reported in the graph 19, we produced *.xpm* and *.log* files. The latter files contained all the possible residue interactions calculated over the trajectory. Amino acids were written in the form RESIDUEID_CHAIN (e.g. MET49_B means the methionine residue in position 49 on chain B). The *.xpm* files were instead matrices containing, in each row, information about the presence or absence of the corresponding interaction (True or False).

To start it, we gave in input each time a specific trajectory file and a topology, in the *.pdb* or *.gro* format. We performed the calculation of the hydrogen bonds by considering each time a limited number of important residues (glossary 7), to be evaluated with the total protein.

To perform our analysis on the MUT protein, we used the equilibrium trajectory, consisting of the last 2000 frames as stated in 4.3. On the other hand, the WT protein analysis was done over the total trajectory of 6000 frames, because of the bad results we obtained during the simulation. In fact, we were not able to define an equilibrium with confidence.

The topology and trajectory files obtained after pulling (chapter 4.8) were considered too.

We then proceeded by compiling a series of lists of "important residues" and "important interactions" (glossary 7), following the consulted literature[1,2,10,11,24]. We defined as "important residues" the amino acids generating only interesting interactions to be studied. On the other hand, we labeled as "important interactions" all the single connections we were interested to track singularly, with or without the presence of an important residue. Figure 20 explains visually the definitions given above.

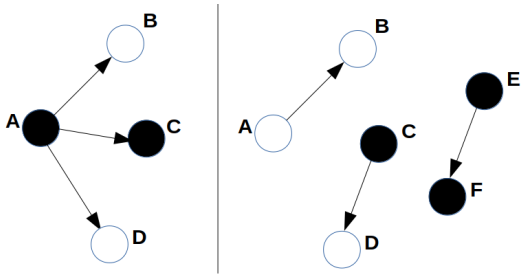


Figure 20. Each one of the black dots represents an "important residue", while each arrow represents an interaction worth consideration. On the left, A is an important residue, and because of that, all its interactions are taken into account. On the right, contacts CD and EF are considered, as in both cases at least an important residue forms the couple. Even though neither A nor B is an important residue, AB is an "important interaction" and must be taken into consideration.

Those were differently associated with the categories listed in tab 3. Importantly, the programs we made can work with asymmetric selections.

Selection	Category	Residues
Ligand (a)	Active site	41HIS, 145CYS, 164HIS/ASN, 187ASP
	Pocket 6Å	41HIS-49MET, 142ASN-145CYS, 163HIS-168PRO, 187ASP-190THR
	Ligand stabilization	1SER, 166GLU, 139SER, 140PHE, 189GLN
	Ligand association	26THR, 46SER, 142ASN, 143GLY, 164HIS/ASN, 166GLU, 189GLN
	Ligand dissociation	44CYS, 45THR, 46SER, 141LEU, 142ASN, 143GLY 166GLU, 167LEU, 168PRO, 191ALA, 192GLN, 193ALA
	Pocket S1	140PHE, 142ASN, 143GLY, 144SER, 163HIS, 166GLU, 172HIS
	Pocket S2	25THR, 41HIS, 49MET, 54TYR, 145CYS, 165MET, 187ASP
	Pocket S3	41HIS, 49MET, 165MET
	Pocket S4	165MET, 166GLU
	Pocket S5	165MET, 166GLU, 189GLN
Dimerization (b)	External cysteines	85CYS, 145CYS, 156CYS
	Internal cysteines	22CYS, 44CYS, 300CYS
	Dimerization	1SER, 2GLY, 3PHE, 4ARG, 5LYS, 6MET 11GLY, 139SER, 142ASN, 166GLU, GLU290, 298ARG
a1 helix (c)	**a1**,**b1** and **a1**/**b1** residues	GLY11, LYS12, VAL13, GLU14, GLY15, CYS16, MET17, VAL18 GLN19, VAL20, THR21, CYS22

Table 3. All the amino acids are numbered starting from 1 to the total length of a monomer (304 in the Wild-type case, 306 in the mutant case). We identified with "pocket 6 Å" the group of residues obtained in chapter 4.5.

Function	Interactions
Dimerization	1SER-166GLU*, 6MET-126TYR, 142ASN-166GLU 4ARG-290GLU*, 4ARG-298ARG**, 6MET-298ARG
Ligand stabilization	1SER-166GLU*, 1SER-140PHE*

Table 4. The interactions with no asterisk are only intra-chain connections. Those with an * are only inter-chain. Finally, the ** are intra- and inter-chain connections.

Then, we consider some of the most important interactions (definition in glossary 7).

Once the *.log* and *.xpm* files were obtained for each type of interaction and each protein (MUT and WT), we filtered the information by using a list of important residues and important interactions. The names of the important amino acids and of the important connections were simply stored in columns in *.txt* files, as shown in figure 21a and 21b. The resulting output gives, for each important interaction and connection involving at least an important residue, a list of frames of occurrences, as reported in the subfigure 21c.

```
164_A
166_A
189_A
26_B
46_B
142_B
```

(a) Important residues file format

```
1_B-166_A
1_B-140_A
126_A-6_A
6_A-126_A
142_A-166_A
```

(b) Important interactions file format

```
LEU57_A-LYS61_A
704
1,6,8,10-11,16-17,
```

(c) *inter.txt* and *res.txt* format. In this case, the frames of interaction between the 57^{th} and the 61^{th} amino acids are reported, which are in total 704

Once these data have been produced, bar plots in the form shown in figures 24 were generated. In blue it is represented the number of frames belonging to the START interval, in orange the number of those in the MIDDLE, and in green the amount of frames in the END. The figures were generated following four types of sorting methods: their presence in the "START", "MIDDLE", and "END" portions of the trajectory (whose ranges are reported in the following dot list), and the total number of frames of appearance. In the first three cases, each horizontal bar was associated with a label indicating the number of frames in the part of interest over the amount of the total frames of occurrence. Most of the graphs we produced are shown in the appendix file (appendix 9).

- **WT main simulation**, sections: START 1-2000, MIDDLE 2001-4000, END 4001-6000.
- **MUT main simulation**, sections: START 1-666, MIDDLE 667-1333, END 1334-2000.
- **WT and MUT pulling simulation**, sections START 1-165, MIDDLE 166-400, END 401-700

The percentages associated with each bar represent the ratio between the number of frames pertaining to a particular part and the total number of frames of appearance of that particular interaction. To consider an interaction especially present in a determined portion, we estimated a lower cutoff percentage value of 40%. Those percentages are highlighted in bold in the tables of the "supplementary material" file.

A different rule was used to evaluate the pulling trajectories. The details are reported in section 5.6.

4.9.1 Production of videos and detailed maps of the interactions

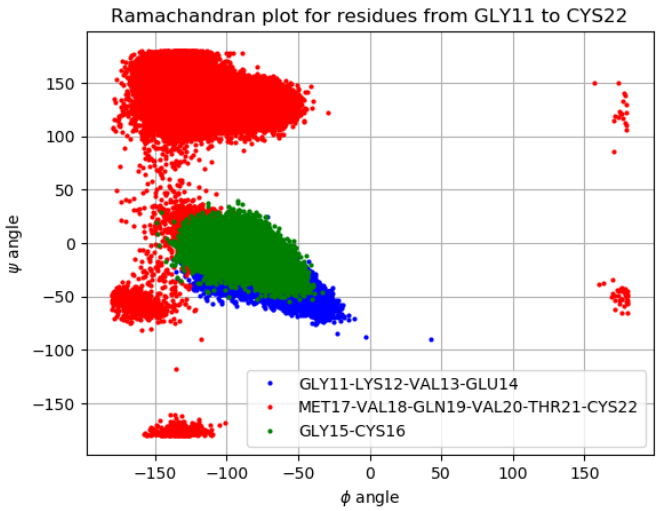
We produced videos, plotting at each time all the active interactions as dots on a grid. The program we built takes into input a color setting, a file of important residues, and a file of important interactions. The dots indicating the presence of an important interaction or of a connection involving at least an important residue are enlarged. Important interactions are labeled by a text reporting the names of the two participating amino acids.

Videos will be shown during the presentation of the project.

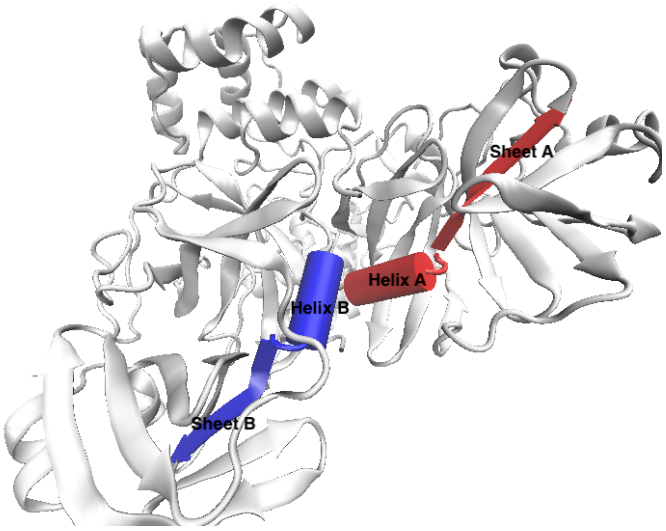
4.10 Targeted uses of the *interactions vs time* procedure

4.10.1 a1 helix analysis - dihedrals and reciprocal orientation of the monomers

The reciprocal orientation of the two monomers depends on many factors. The definition of an angle between the two monomers is clearly not univocal. By means of a visualization tool such as VMD[6], it is possible to see that two regions with anti-parallel β sheets, connected with two α-helices, confer structural rigidity to the protein. The orientations of the two monomers depend on their movements. Specifically, we suppose that it is largely conditioned by the behavior of the symmetry axis of the two helices over the trajectory.



(a) The regions in the (ϕ, ψ) space highlight the presence of a β-sheet (red) linked to an α-helix (blue) via a small loop (green).



(b) Representation of the MUT. In red and blue is shown the selection **a1** and **b1** (see topology in figure 1) for the two monomers.

Figure 22

We considered the two alpha helices composed of the GLY11-LYS12-VAL13-GLU14 amino acids. We obtained the two axes of the helices $(X_A\hat{x}+Y_A\hat{y}+Z_A\hat{z})$ and $(X_B\hat{x}+Y_B\hat{y}+Z_B\hat{z})$ by using the Gromacs[25,26] command *"gmx helixorient"* . Then, we calculated the absolute value of the difference between each component. These differences give information about the bending of the angle. We highlight the overall behavior of the noisy plot of the differences by taking averages every 50 points (red line). In figure 23, we evaluate whether the collected data are stationary or not.

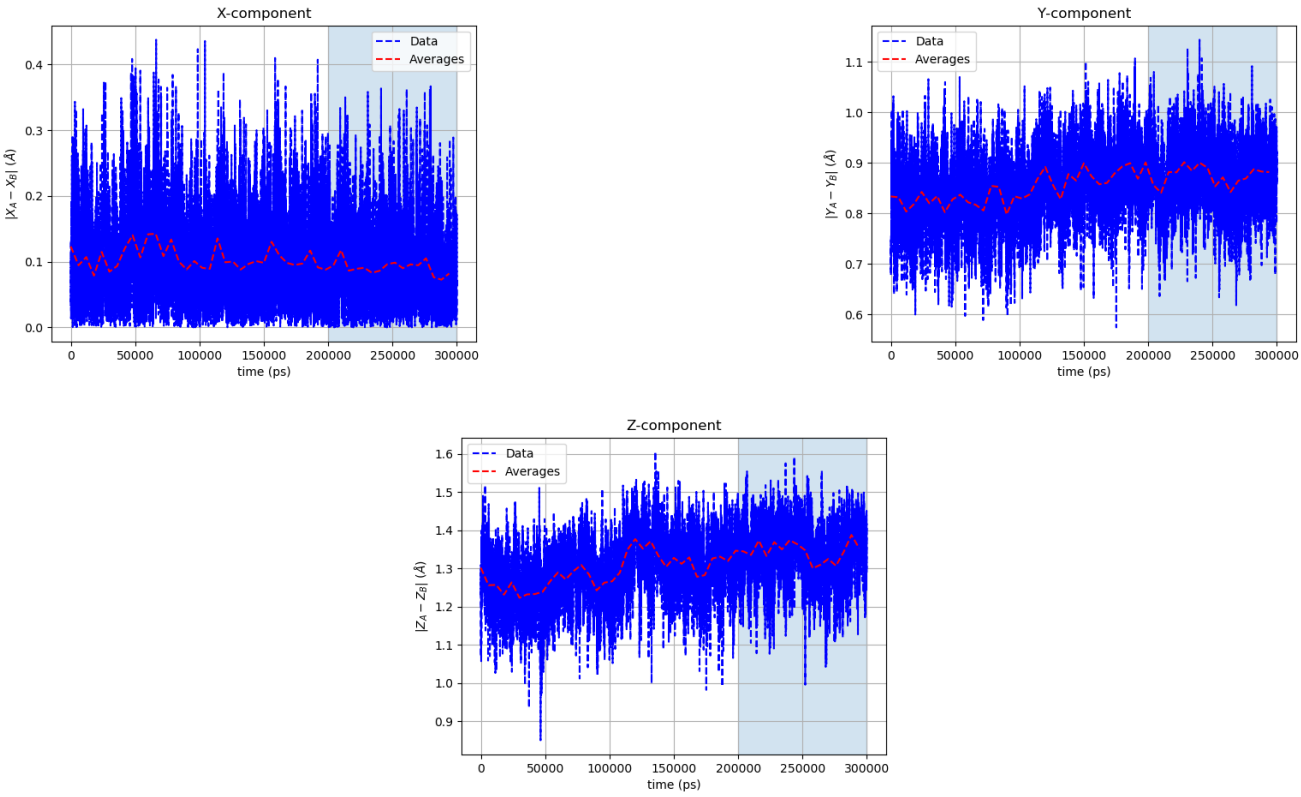


Figure 23. Plots representing the absolute value of the differences of the components of the helices axes. The means over chunks of 50 consecutive points are reported with the red line.

By using the *interactions vs time* addition on the equilibrium part, we produced the bar plots represented in figure 24. By viewing the images, we can conclude that the vast majority of the interactions are stably generated along the trajectory. Therefore, based on these plots and those in the figure 23, we can conclude that there are no significant variations of the angle that we have decided to study.

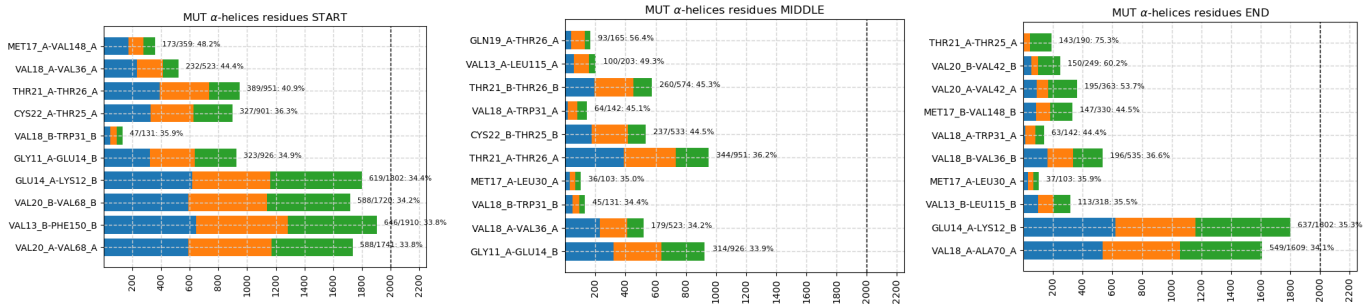


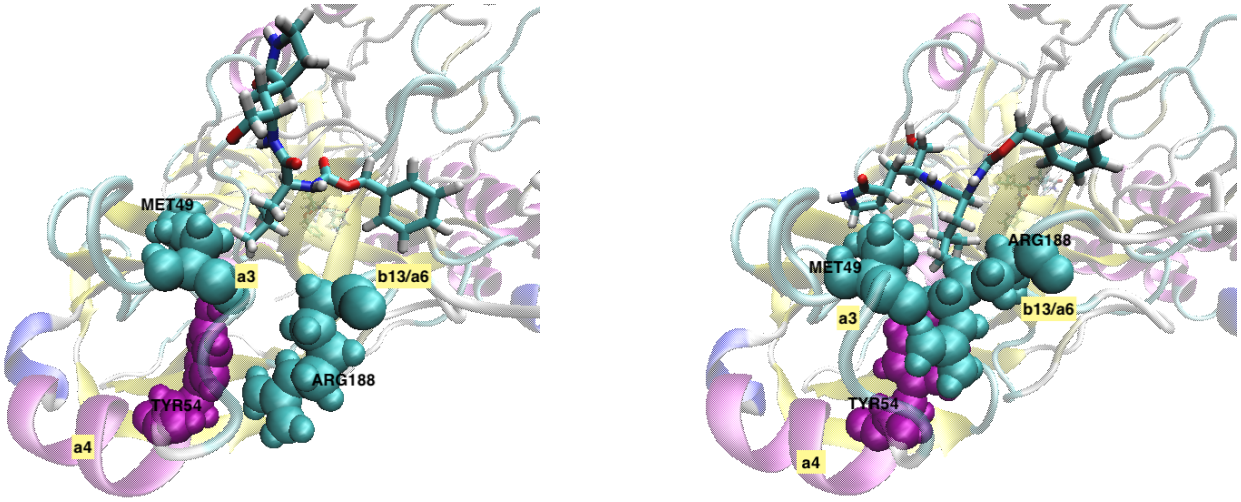
Figure 24. Plots representing the persistence of the interactions in the equilibrium chunk of the trajectory. This last chunk of 100 ns is divided into 3 parts of equal length called "START", "MIDDLE" and "END".

5 DISCUSSION

5.1 Discussion for the MUT simulation

Thanks to the *interactions vs time* modification, we were able to reconstruct some of the specific features of the MUT protein pocket along the equilibrium trajectory from 200 ns to 300 ns. In what follows, the nomenclature used in the diagram 1 will be used. The residues shown in figure 25 are colored according to their position in the secondary structure. First of all, as suggested by figure 25, the residues MET49 and ARG188, situated respectively in **a3** and in

b13/a6 (glossary in chapter 7), have a high degree of connectivity, our discussion will start from there. For both chains A and B, the interaction MET49-ARG188 allows the two mentioned loops to stay together and remain near the ligand. From the bar plots in 9.1, we understand that the residue MET49_A interacts initially with TYR54_A (figure 31, 72.3 % in the START section), then, in the MIDDLE and END sections, it moves away and starts interacting with ARG188_A (42.5% in MIDDLE and 57.4% in END).

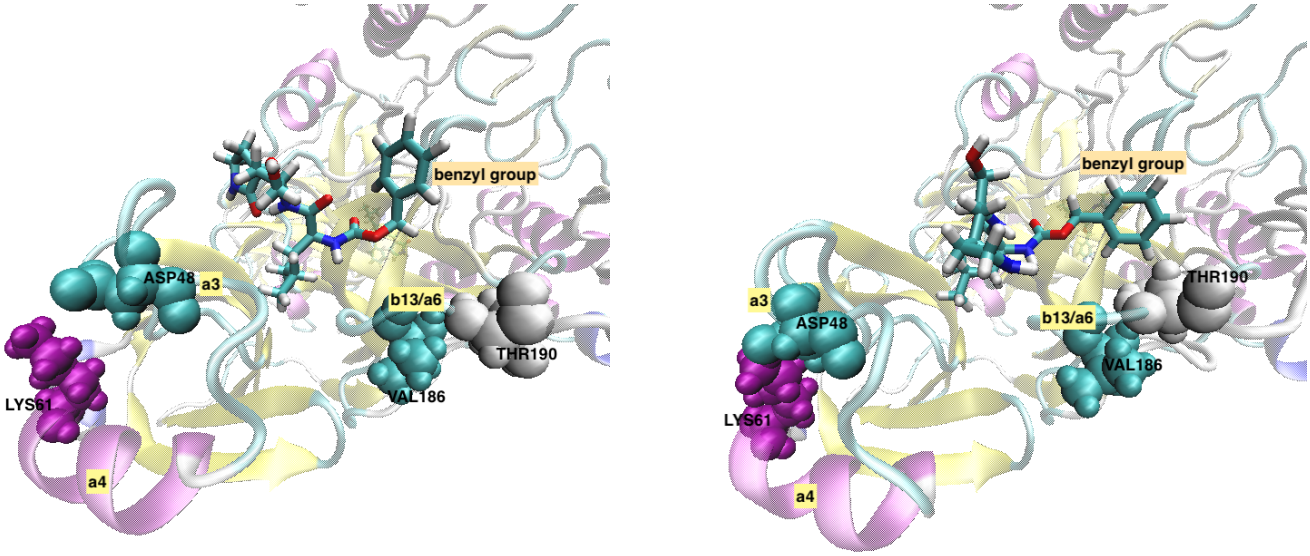


(a) START, pocket A

(b) END, pocket A

Figure 25. In pocket A, the loop **b13/a6** moves towards the **a3**. The helix **a3**, which is present in 1, is not represented by VMD.

ASP48 is another residue with a high degree of connectivity, located in **a3**. The plot bar of the section END in 9.1 tells that ASP48_A interacts with LYS61_A (END 85%), which is part of **a4**. Consequently, the **a3** helix, comes closer to **a4**, expanding the pocket size (figure 26b). On the other hand, the interaction between the two residues 186VAL_A-190THR_A located in **b13/a6** is present towards the end of the trajectory and stabilizes the benzyl group of the ligand (see the structure of the ligand in 2.2). In the last part of the trajectory, the interaction between **a3** and **b13/a6** prevails over the interaction between **a3** and **a4**, stabilizing the ligand.



(a) START, pocket A

(b) END, pocket A

Figure 26. The connection between **a3** and **a4** strengthens in the last part of the trajectory, while **b13/a6** stabilizes the benzyl group of the ligand.

As it can be seen from the bar plots in 31, pocket A has not yet reached a sufficient equilibrium, unlike pocket B, which appears more stable. In fact, interactions involving residues in pocket B tend to be present in all parts of the trajectory. On the other hand, in many cases interactions involving residues of the A pocket are exclusive of some segments. This observation and the fact that not all the interactions between residues are symmetric lead us to affirm that the two pockets show distinct behaviors, as already shown in figure 17 (chapter 4.7).

In addition, we notice that the interactions involving residue 163HIS, which has high connectivity, are stable in time. The same is valid also for its neighboring residue ASN164, which is the mutated residue of our protein. The most present interaction is the hydrogen bond ASP187-ARG40 in both chains A and B. This interaction stabilizes the loops **b13/a6** and **b3/a2**.

The *interactions vs time* addition also identifies the connections with a role in the dimerization process, for example, SER1_A-PHE140_B (87.9 %) in the START or MET6_A-GLN299_A and others (figure 32).

5.2 Discussion of the WT simulation

5.2.1 The problems we found with the WT protein

The unrestrained molecular dynamics of the WT protein did not return the expected results: we observed a strange detachment of the two ligands from their binding site at different time steps. The detachment in chain B is very sudden and happens after approximately 150 ns, while the one regarding chain A is slower and starts at around 200 ns. In the second case, the ligand disengages from the pocket and gets very close to the α-helices of domain III. From the article by Hu and colleagues,[1], we report the values of the IC_{50} and k_i (refer to glossary 7) constants, as reported in table 5.

	MUT	WT
IC_{50} (nM)	126.3 ± 6.53	40.25 ± 1.61
k_i (nM)	37.95 ± 2.58	15.50 ± 1.28

Table 5. IC_{50} and k_i for the MUT and the WT proteins

The values of k_i and IC_{50} are comparable between MUT and WT, meaning that they should have a similar behavior when interacting with the GC-376 ligand. Besides, the ligand should be more bounded to the protein in the WT case.

The ligand detachment could be due to statistical fluctuations: in fact, k_i and IC_{50} are statistical indexes, so the detachment event is not completely excluded.

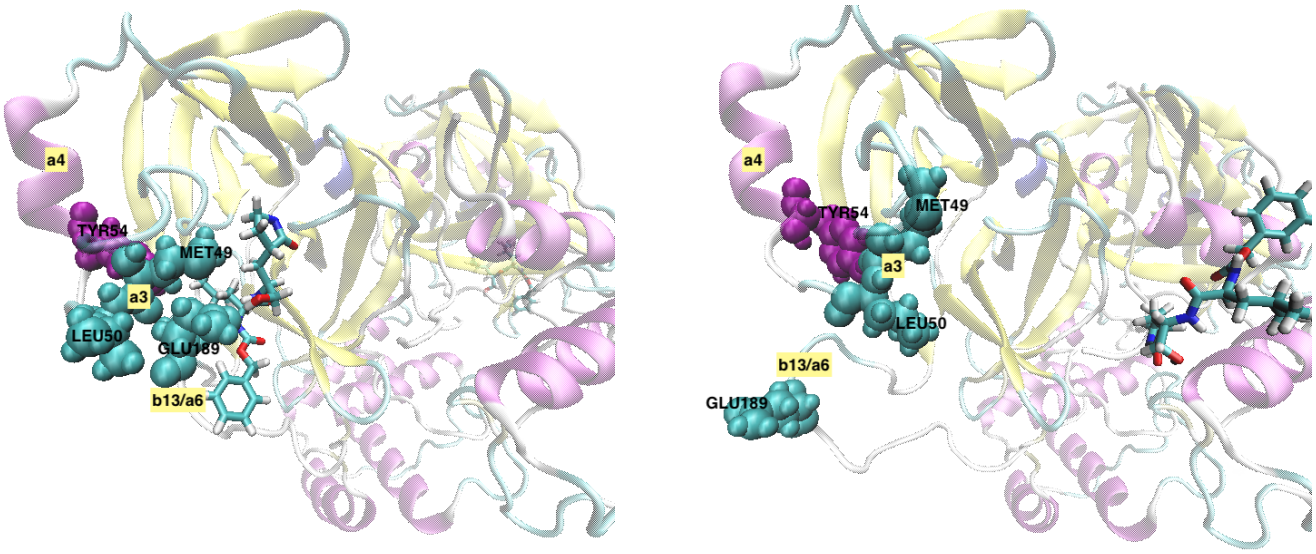
We are aware of the possible problem due to the protonation state of the histidines, which is not known *a priori*. Then, a proper study of the protein should be performed on different simulations with different states of the histidines: HSD/HSE/HSP (figure 6). Moreover, the protonation states of histidine might depend on the charged neighbouring residues. In particular, as suggested in the article by Pavlova et al.[27], the protonation state of the catalytic residues CYS145, HIS41, and HIS164 might be highly correlated. Due to limited computational resources, we decided to trust the given configuration reported in the *.pdb* file on the RCSB site.

In what follows, we use the *interactions vs time* addition to detect the interactions (or network of interactions), more involved in the detachment event. This constitutes a sort of *"debugging"* of the simulation, in which we try to find the root of the problem by looking at the specific interactions involved in the ligand stabilization.

5.3 Discussion about the interactions found in the WT protein

Based on the output of *interactions vs time*, we give an interpretation of the results using graphical representations.

The interaction MET49_A-GLU189_A takes place only in the first portion of the trajectory. This involves two segments **a3** and **b13/a6** which are important for the stabilization of the ligand in the pocket. The absence of this attraction in the last part of the trajectory widens the pocket, allowing the ligand in A to escape towards the domain III. As a consequence of the release of the ligand, the interacting couples MET49_A-TYR54_A and MET49_A-LEU50_A only appear in the END section.

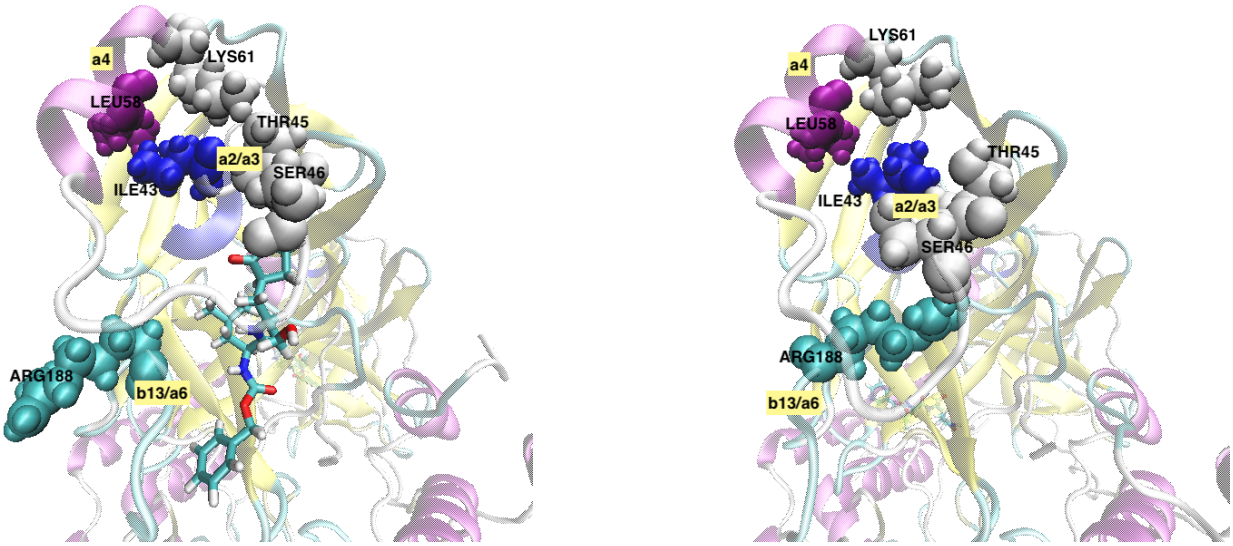


(a) WT case, pocket A, START section.

(b) WT case, pocket A, END section.

Figure 27

The first bars of MIDDLE section plot (figure 9.5) highlight the presence of interactions involving **a2/a3** and **a4**, for instance THR45_B-LYS61_B or ILE43_B-LEU58_B. When these interactions intensify, the loop **a2/a3** moves towards the helix and the ligand comes out of the pocket. A consequence of the release of the ligand is the presence of many interactions between the regions **a2/a3** and **b13/a6** in the last part of the trajectory. From the bar plot sorted by the number of frames in 9.5, it must be noted that the important hydrogen bond 187ASP_B-40ARG_B is not responsible for the leaving of the ligand because it is mostly maintained for the whole trajectory.



(a) WT case, pocket B, MIDDLE section.

(b) WT case, pocket B, END section.

Figure 28

Another possible cause of the detachment is the progressive approach of the 142ASN_A, situated in **b10**/**b11**, to the 300CYS_B in **a10** is a possible cause of the detachment of the ligand of the pocket A: the loop **b10**/**b11**, which is close to the γ-lactamate part of the ligand, is attracted to the α-helix **a10** and leaves space to the ligand, enlarging the pocket size.

5.4 Advantages and disadvantages of the algorithm

To obtain information about different time segments with *PyInteraph*, it would be necessary to produce numerous trajectories from the original (with the "*trjconv*" command of Gromacs[25,26]), and then, to analyze the produced chunks separately. However, there is no automated tool that joins *PyInteraph* with *trjconv*. Not only the trajectory has to be subdivided manually, but also a *.pdb* file of the starting frame of every sub-trajectory has to be retrieved in order to execute the *PyInteraph* code. Some differences between the original algorithm and the modified one could be pointed out:

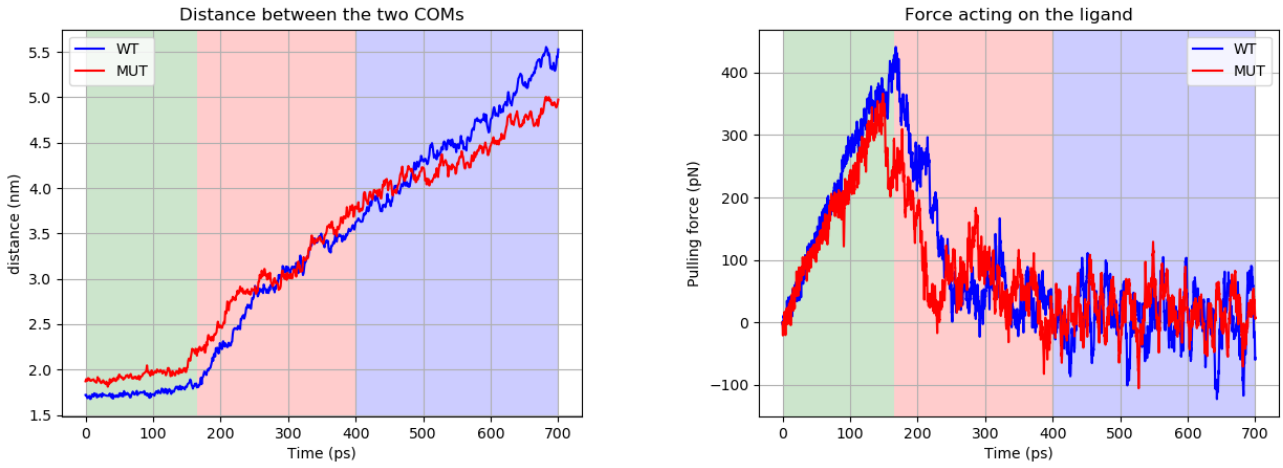
- With our addition it is possible to keep track of specific groups of residues for all types of calculations. Additionally, if some residues are cited in the literature for their biological relevance, it is possible to easily add them to a new *residues* file (see figure 21a) and rerun the program.
- Our addition can be used for making movies of the interactions frame by frame. Moreover, from the files in 21c, it is possible to obtain the exact frame in which a specific interaction ceases to exist or begins to show up. In this optic, the tool could be employed, for instance, in the analysis of steered molecular dynamics or in the evaluation of other *rare events*.
- In order to study the statistics of relevant interactions, we believe that the addition could be employed in a comparison between simulations with the same initial conditions. Additionally, we believe that it could be possible to confront different mutants of the same protein, and characterize the similarities and the differences.

For what concerns the problems related to our addition:

- It does not provide a graphical visualization, which would definitely help in evaluating the relevant interactions.
- The calculation of H-bonds is relatively slow for large selections and has to be performed on a computer cluster.
- The calculation of the salt bridges is overly simplified: for lack of time, we decided to consider the distance between the COMs of the charged amino acids side chains.
- The *interactions vs time* modification does not evaluate connections between the protein and the ligands directly, but it provides interactions between the residues mostly involved in interactions with the ligand.

5.5 Attempts for obtaining a good pulling trajectory

According to the description in 4.8, we performed pulling simulations of MUT and WT monomeric forms.



(a) **(b)**

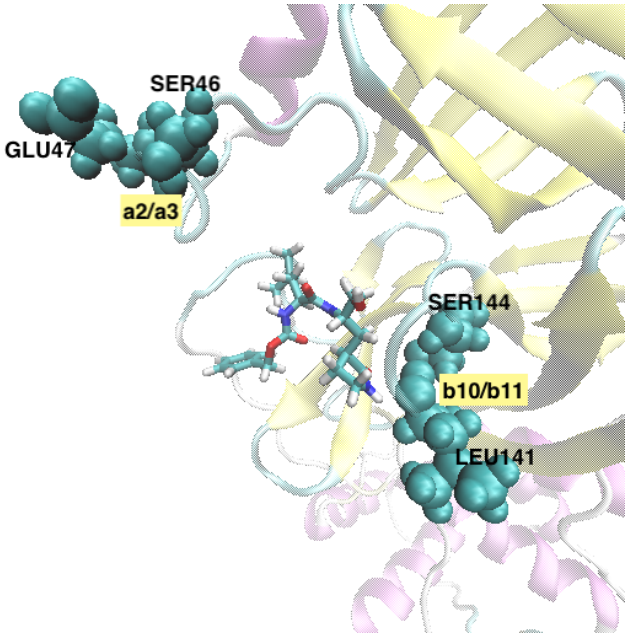
Figure 29. Until approximately 200 ps the ligand feels the attractive forces exerted by the protein, consequently, it is located far from the center of the biasing potential. As the distance increases, the interactions with the protein become progressively less important and the ligand oscillates near the center of the biasing harmonic potential. This oscillation is reflected in the fluctuation of the force in the last section.

5.6 Interactions over time, pulling case

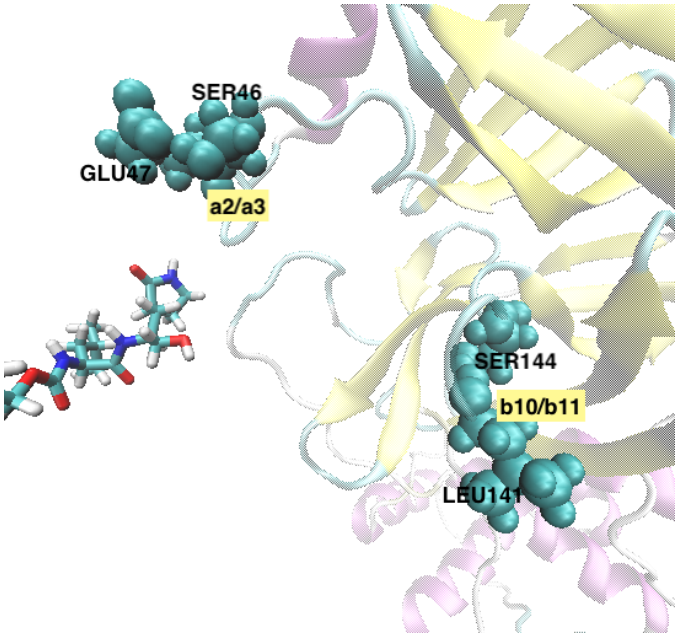
Again, using the *interactions vs time* addition employed in the discussion of the MUT and WT simulation, we can check the persistence of certain interactions over time. We divide the trajectory in three parts according to the coloring used in figure 29, the intervals elencated in 4.9. For each part, we report the percentage of persistence[5] of some significant interactions.

This time, we begin our discussion with the WT case: first of all, as it can be seen in 9.7 in the MIDDLE section plot, SER46-GLU47 and LEU141-SER144 begin to show up after the ligand has been released. These two interactions involve respectively the regions **a2/a3** and **b10/b11**. Regarding the interactions in the MUT case, we do not notice any significant variations over time of the interactions.

[5]The number of occurrences of an interaction divided by the number of frames of the considered chunk of trajectory.



(a) WT pulling simulation, START section.



(b) WT pulling simulation, END section.

6 Conclusions

We produced the simulations for the WT and MUT proteins (section 4.2), we performed basic analysis (section 4.3) and set up pulling simulations of K36 for both proteins (as reported in chapter 4.8). Additionally, the WT protein showed unexpected detachment events of the ligands.

The principal purpose of our project became to find a way to study the emergent interactions along the MUT and WT trajectory and to develop a tool to understand the problems involving the WT simulation. We are aware of the existence of other type of analysis provided by *PyInteraph*, such as the possibility of getting the largest size of the *connected graph*, or even getting a *weighted graph* of the interactions. We think these tools are able to bring important information about the equilibrium regime, but in the studied cases, and especially in the WT case, a true equilibrium cannot be identified.

Using *PyInteraph* we managed to produce a list of residues involved in the pocket, and evaluate their grade of connectivity (4.5). Once obtained the list of residues, we proceeded by using our addition for further analysis.

We believe that *interactions vs time* can be useful in the analysis of simulations characterized by different "moments" or "events". For instance, the three sections identified in the pulling simulation discussion (5.6), or the detachment of the ligand in the WT simulation.

Some changes to the program could help in obtaining an easier algorithmic process, and would definitely help in making it more efficient (refer to advantages/disadvantages in 5.4). For a question of space and time, we limited our analysis to a few set of residues, showing what type of information could be obtained, but of course the study could be continued and deepened.

Unfortunately, we had not the opportunity to have a clear reference for the MUT protein, because of the strange detachment of the ligands from the WT protein (see 5.2). In fact, we think that our addition could be particularly useful to compare Mutant and Wild-type proteins, or protein with different mutations, in specific events or frames. As a future perspective, we think it would be interesting to gather the simulation data of other mutant proteins from other groups, with the aim of finding singular features of the different alterations.

7 Glossary

WT	Wild-type protein.
MUT	Mutant protein.
MPRO	SARS-CoV-2 Main Protease.
important residue	Residues of which we are interested to study each interaction.
important interaction	All the interactions that we consider relevant to track. The residues involved are not necessarily important residues .
COM	Center Of Mass.
AA	Amino acid.
ai, **bj**, **i**=1...10, **j**=1...13	Notation for α-helices an β-sheets.
ai/bj	Notation for the loop located between **ai** and **bj**. The notation is also extended to the case **ai**/**aj** and **bi**/**bj** etc...
Interactions vs time algorithm	Algorithm to study interactions frame by frame.
RMSD	The root-mean-square deviation is the measure of the average distance between the atoms (usually the backbone atoms) with respect to a reference.
R_g	The radius of gyration is a measure of the globularity of the protein.
RMSF	It is a measure of the fluctuation of the position of an atom over the trajectory.
H-bonds	hydrogen bonds
$\mathbf{IC_{50}}$	Ligand concentration that inhibits 50% of the target enzyme.
$\mathbf{k_i}$	Dissociation constant that measures the binding affinity of a ligand.
PMF	Potential of Mean Force.

8 Appendix - Ideas for a possible umbrella sampling

Based on the article by Tam et al[7], we set up a discussion about the complications of an umbrella sampling procedure applied to the MUT and WT proteins in complex with GC376. The article proposes a way of calculating the free energy of the ligand dissociation using the monomeric form of the Mpro, although not enzymatically active. We warn that the conclusions of this section have mostly a qualitative character: due to limited computational resources, we only provide some hypothesis based on our rough attempt.

8.1 Problem of finding the best unbinding pathway

As proposed by Tam et al.[7] we decided to use the distance between the COMs of the ligand and the protein as a good reaction coordinate. The determination of the best reaction coordinate is a non-trivial task because of the large number of atoms in the ligand.

Let ξ be the mentioned reaction coordinate and let $\vec{\xi}/\xi$ be its direction in the three-dimensional space. During the steered simulation, the ligand might move also along directions perpendicular to $\vec{\xi}/\xi$ because of attraction forces coming from different regions of the protein. Although some deviations from a straight path could be admitted in the actual optimal unbinding path, the constant pulling velocity approach needs these fluctuations to be reduced as much as possible. In fact, if the unbinding direction is not chosen carefully, the ligand could be attracted to other parts of the protein once it has been released by the pocket and this could lead to a set of very unrealistic initial configurations for the actual umbrella sampling.

In other words, let $\phi(t)$ be the angle between the direction $\vec{\xi}/\xi$ and the direction of the vector connecting the two COM $\vec{d}(t)$. Then the best steered trajectory minimizes the maximum variation of the angle ϕ. We hypothesize that it could be possible to determine the best direction with a PCA: we could consider different unbinding directions $\vec{\xi}/\xi$, select the atoms of the ligand and compute the principal components of their motion. If the direction is sufficiently good, the first principal component should correspond to $\vec{\xi}/\xi$ with a very high percentage of captured variance.

8.2 Choice of the parameters of the simulation

We hypothesize that using restraints solely on the α-C rather than on sidechains and backbone atoms, would allow a more "natural" movement of the other components under the action of the external force. In a free energy calculation, the choice of the cut-offs of the interactions directly affects the *Potential of Mean Force* PMF 7. The best choice in this regard would be to set very high cut-off distances, but in reality a compromise must be found between the computational cost and the accuracy.

Among the *mdp parameters* listed in (1) the cut-off distance for VDW and Coulomb interactions is the same. Since the VDW potential scales as $1/r^6$ at large distances, the cut-off can remain robust even at $r_{\text{VDW}} = 0.7 - 0.8$ nm. This reduction in the computational cost may be followed by an increase in the Coulomb cut-off, for instance $r_{\text{Coulomb}} = 1.2 - 1.3$ nm.

When the ligand detaches from the protein, as it can be seen in figure 29, the center of the biasing potential is located far from the ligand COM, therefore the pulling force drags the ligand away abruptly and some intermediate configurations might be skipped. This suggests that the velocity of the pulling, as well as an optimal k for the biasing potential should be chosen carefully. We believe that a smaller v and k than the ones reported in 4.8 can be employed to get a good set of initial configurations for an umbrella sampling.

9 Appendix - Bar plots for the WT and MUT simulations

9.1 MUT protein main simulation data - ligand

9.1.1 Bar plots for the ligand residues with four sorting methods

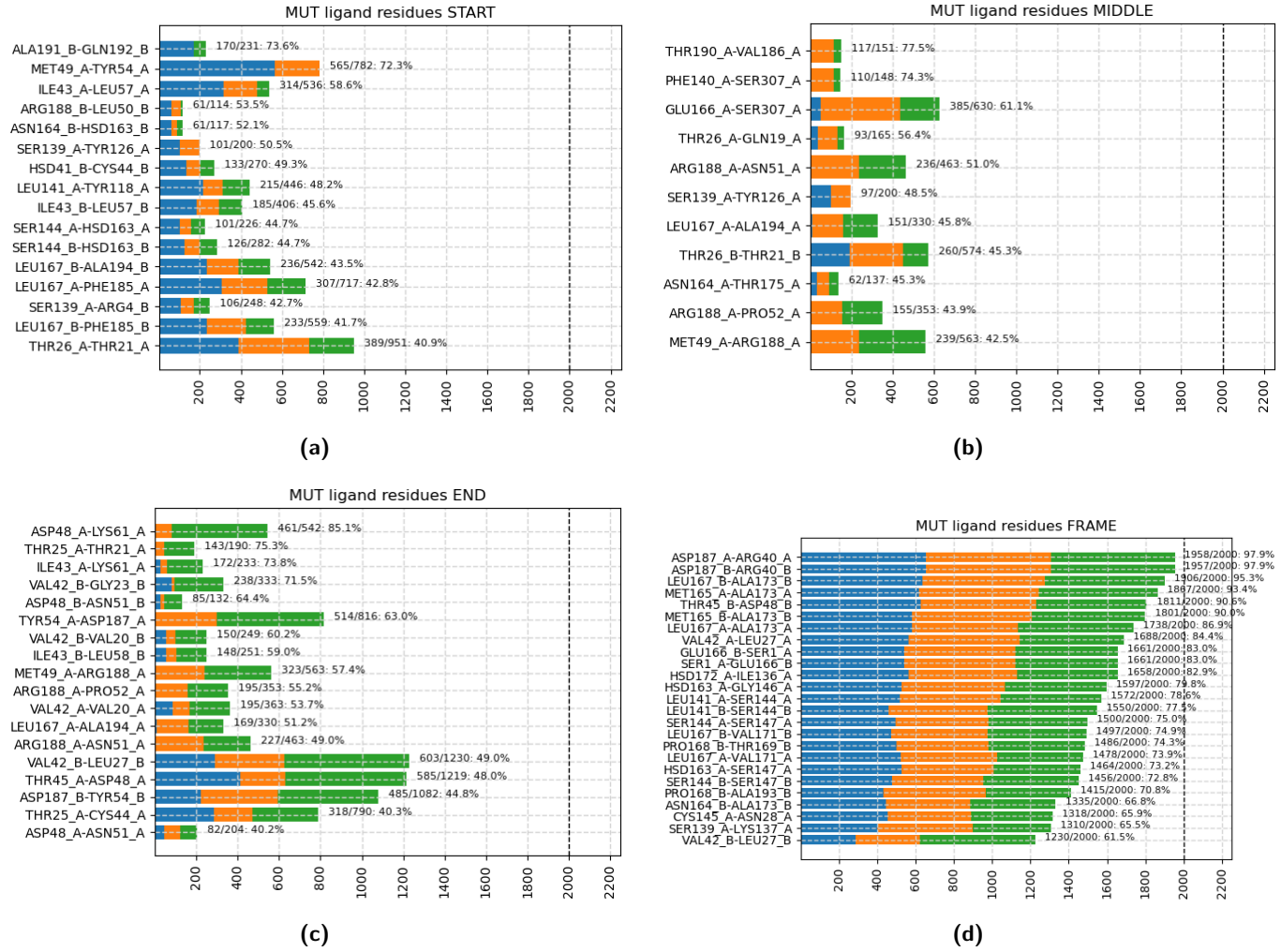


Figure 31. MUT protein, bar plots of the interactions involving residues interacting with the ligand. Each graph lists the interactions with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). We estimated a cut-off of 40% to consider an interaction particularly present in a particular portion of the trajectory

9.2 MUT protein main simulation data - dimerization

9.2.1 Bar plots for the dimerization residues with four sorting methods

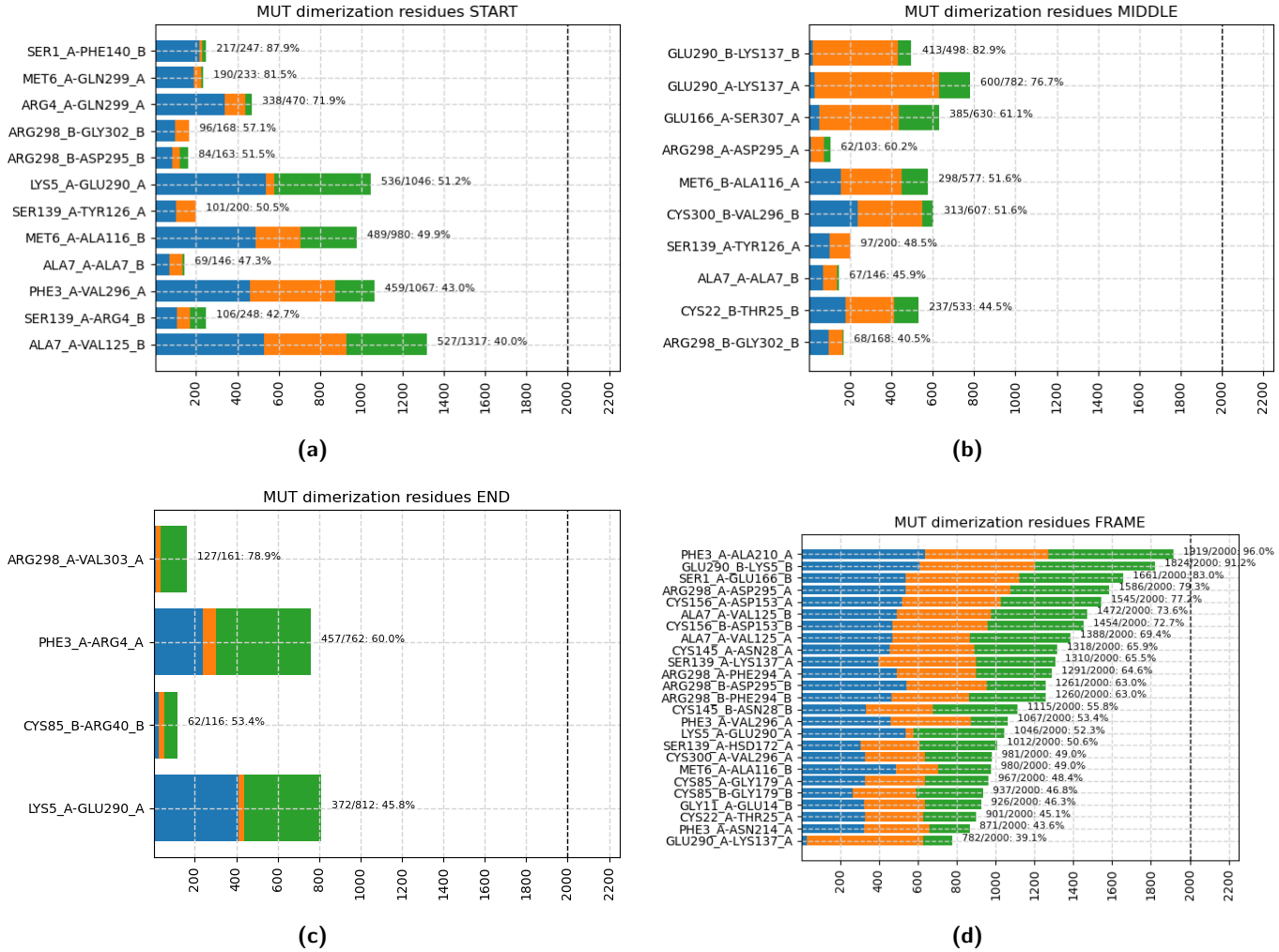


Figure 32. MUT protein, bar plots of the interactions involving residues for the dimerization. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). We considered a cut-off of 40%.

9.3 MUT protein pulling simulation - ligand

9.3.1 MUT: Bar plots for the ligand residues with four sorting methods

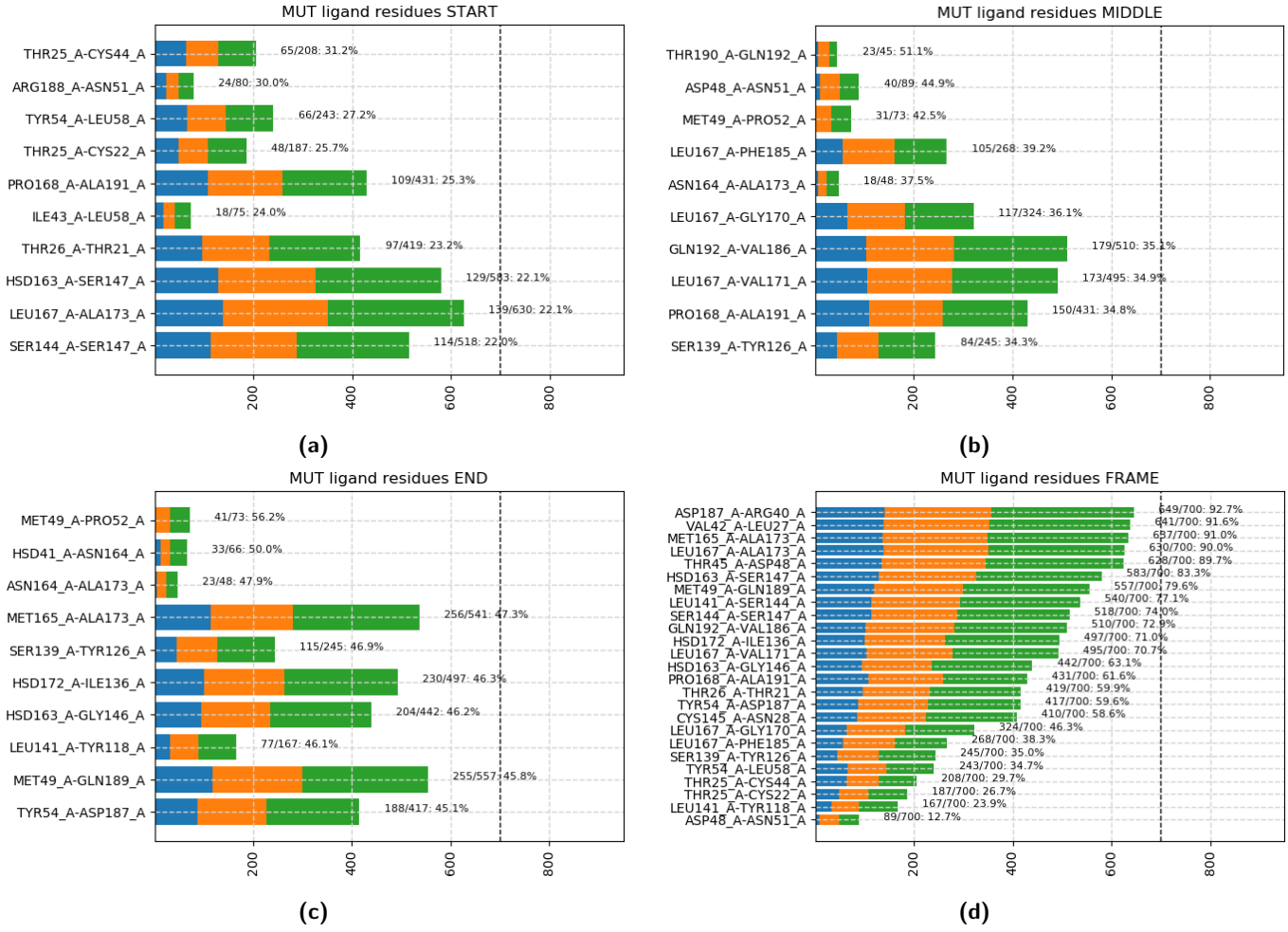


Figure 33. MUT protein, bar plots of the interactions involving residues of the **Ligand** residues. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered cut-offs of 30% for the "START" section, 40% for the "MIDDLE" section and 50% for the "END" section.

9.4 MUT protein pulling simulation - dimerization

9.4.1 MUT protein pulling. Bar plots for the dimerization residues with four sorting methods

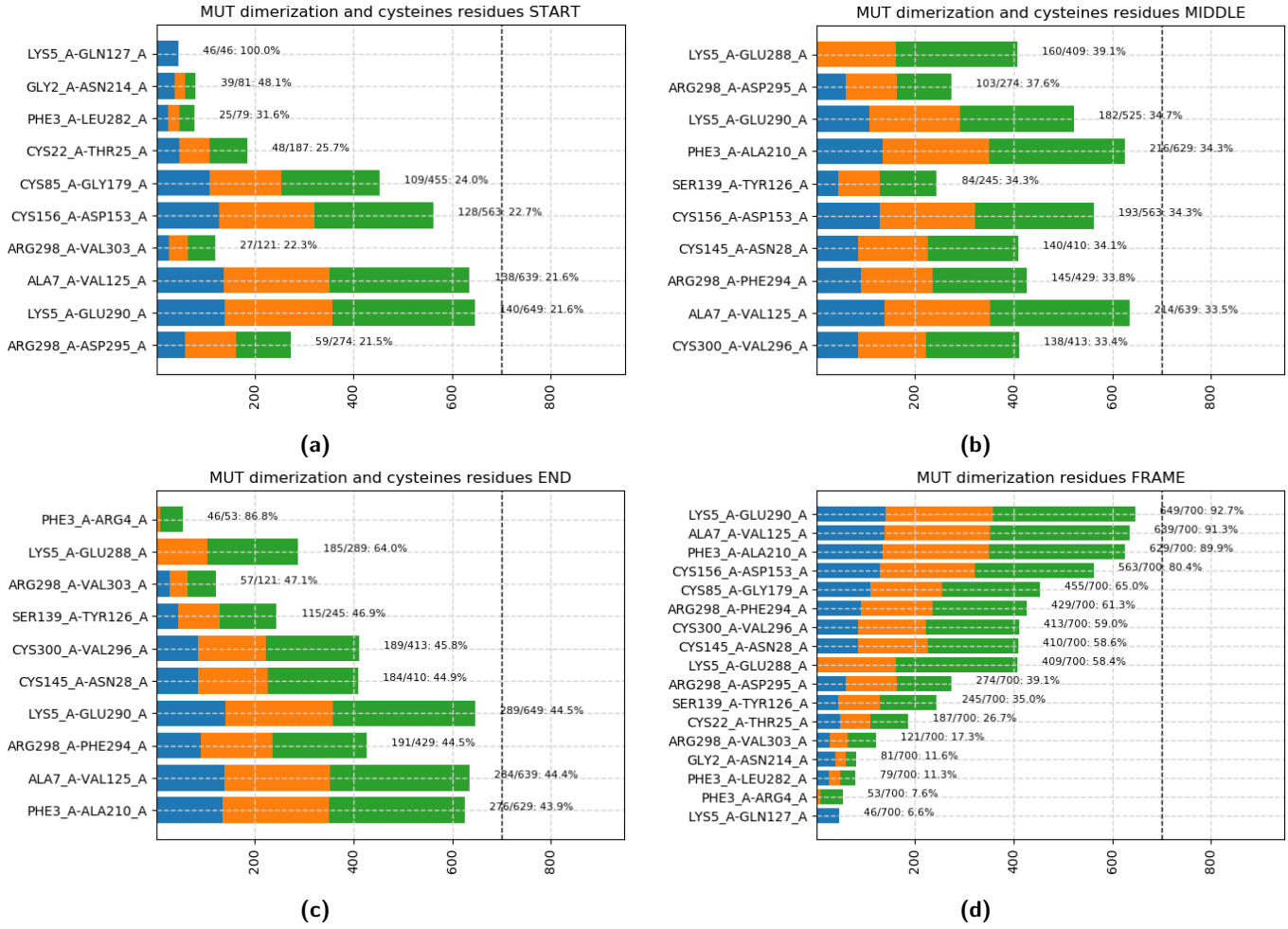


Figure 34. MUT protein, bar plots of the interactions involving residues of the **Dimerization** selection. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered cut-offs of 30% for the "START" section, 40% for the "MIDDLE" section and 50% for the "END" section.

9.5 WT protein main simulation data - ligand

9.5.1 WT: Bar plots for the ligand residues with four sorting methods

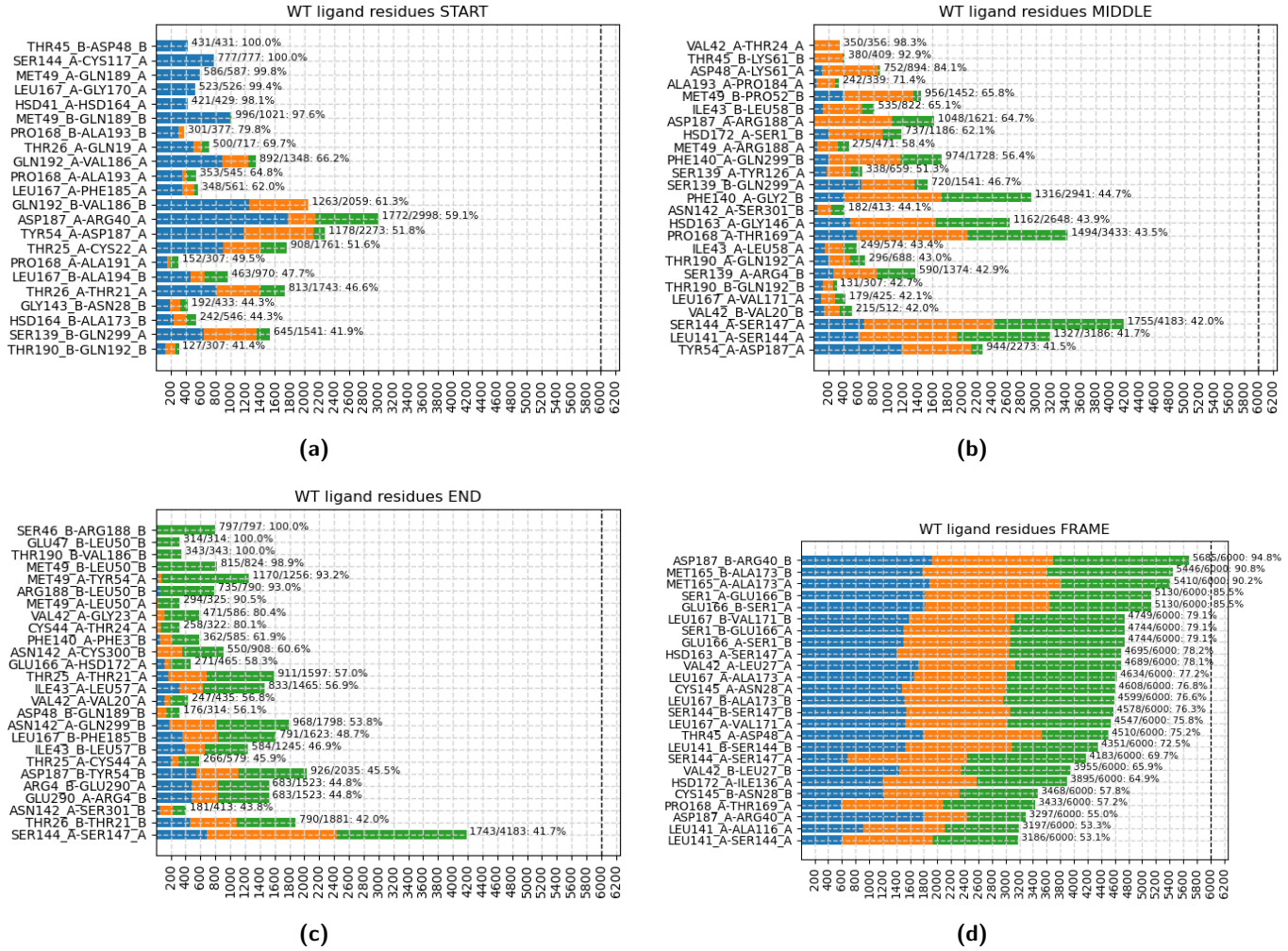


Figure 35. WT protein, bar plots of the interactions involving residues in connection with the ligand. Each graph lists the bonds with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered a cut-off of 40% to retain significant an interaction.

9.6 WT protein main simulation data - dimerization

9.6.1 WT protein: bar plots for the dimerization residues with four sorting methods

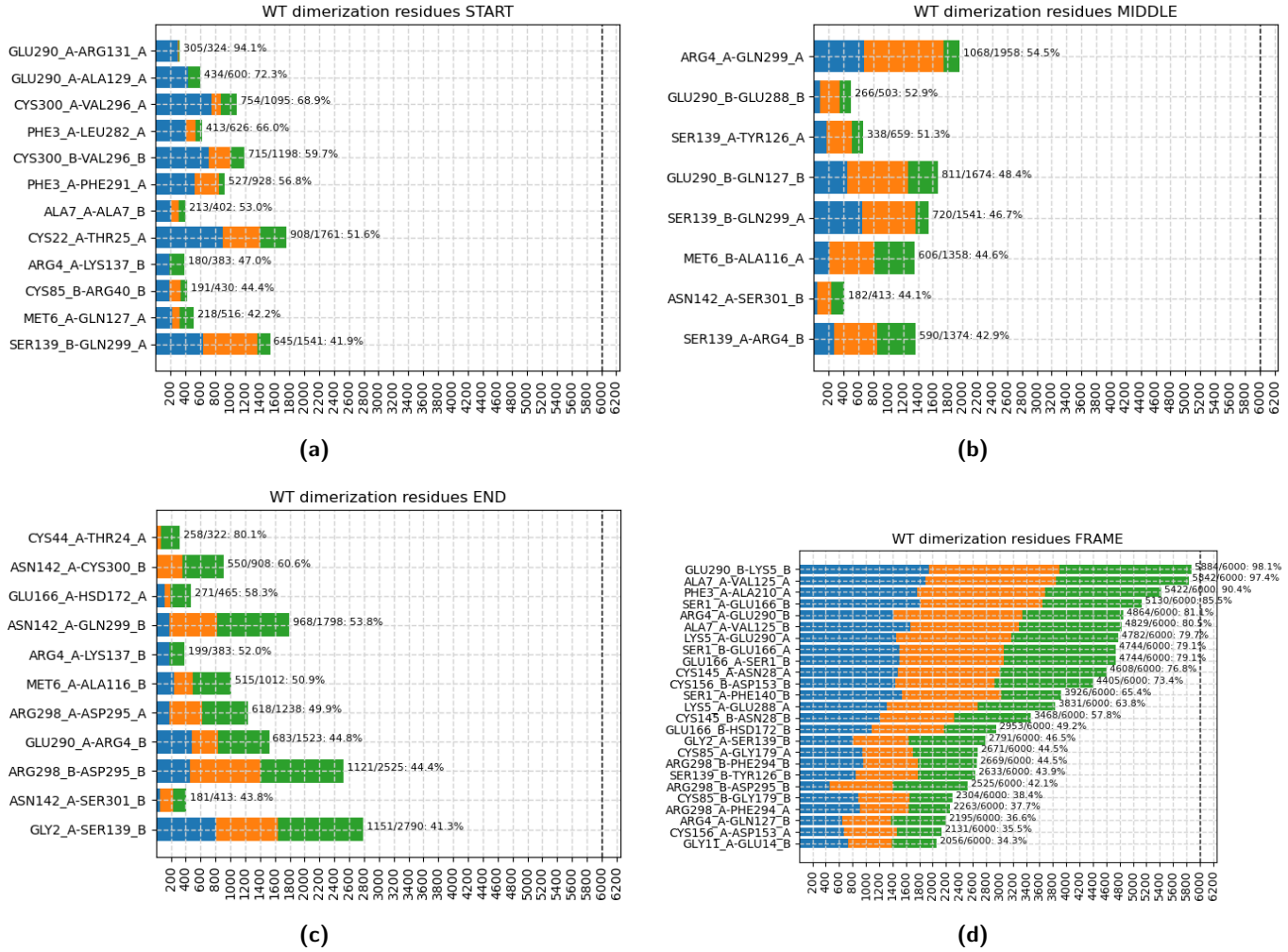


Figure 36. WT protein, bar plots of the interactions involving residues for dimerization. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered a cut-off of 40%.

9.7 WT pulling simulation - ligand

9.7.1 WT protein pulling. Bar plots for the ligand residues with four sorting methods

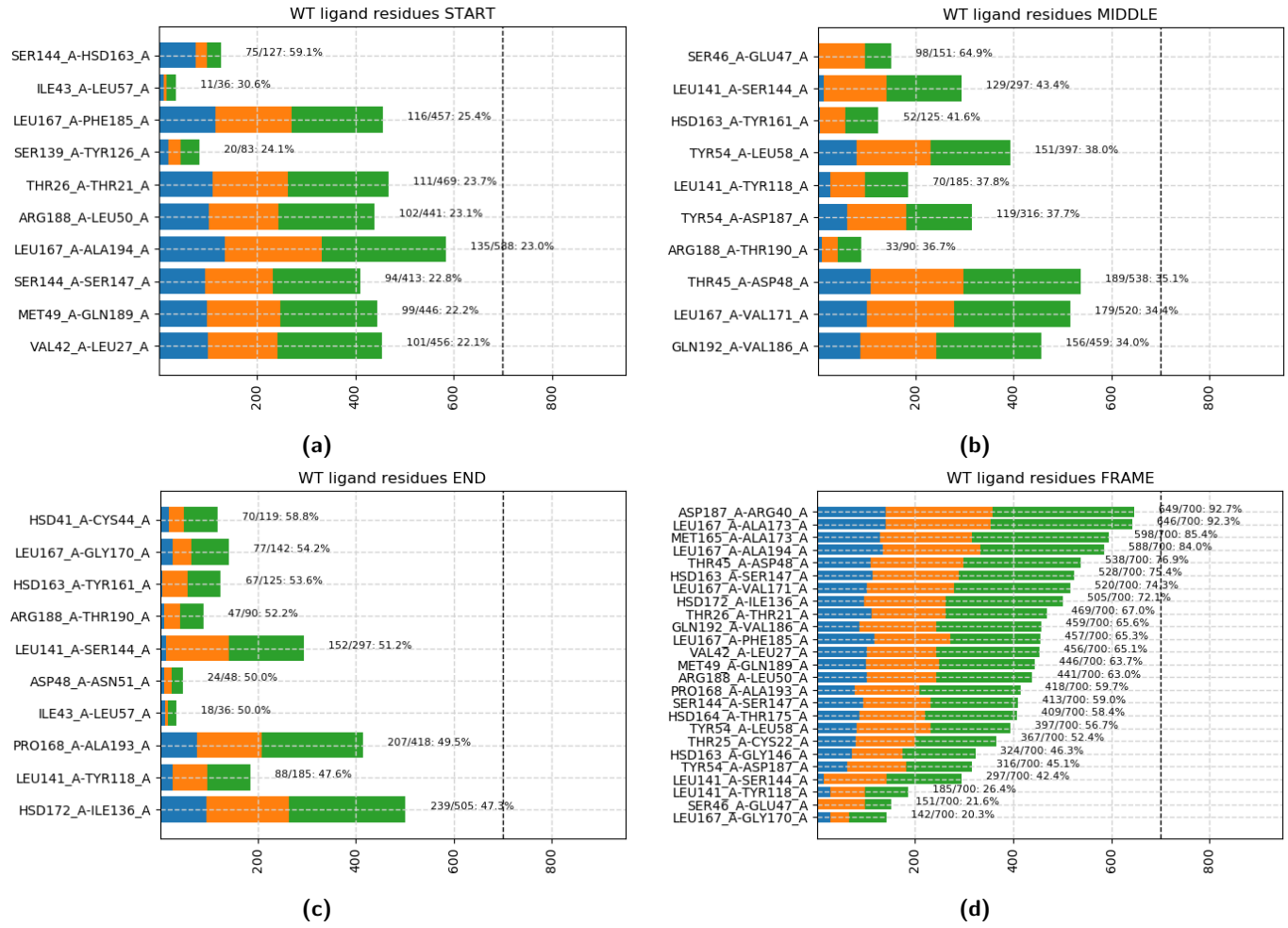


Figure 37. WT protein, bar plots of the interactions involving residues of the **Dimerization** selection. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered cut-offs of 30% for the "START" section, 40% for the "MIDDLE" section and 50% for the "END" section.

9.8 WT pulling simulation - dimerization

9.8.1 WT protein pulling. Bar plots for the dimerization residues with four sorting methods

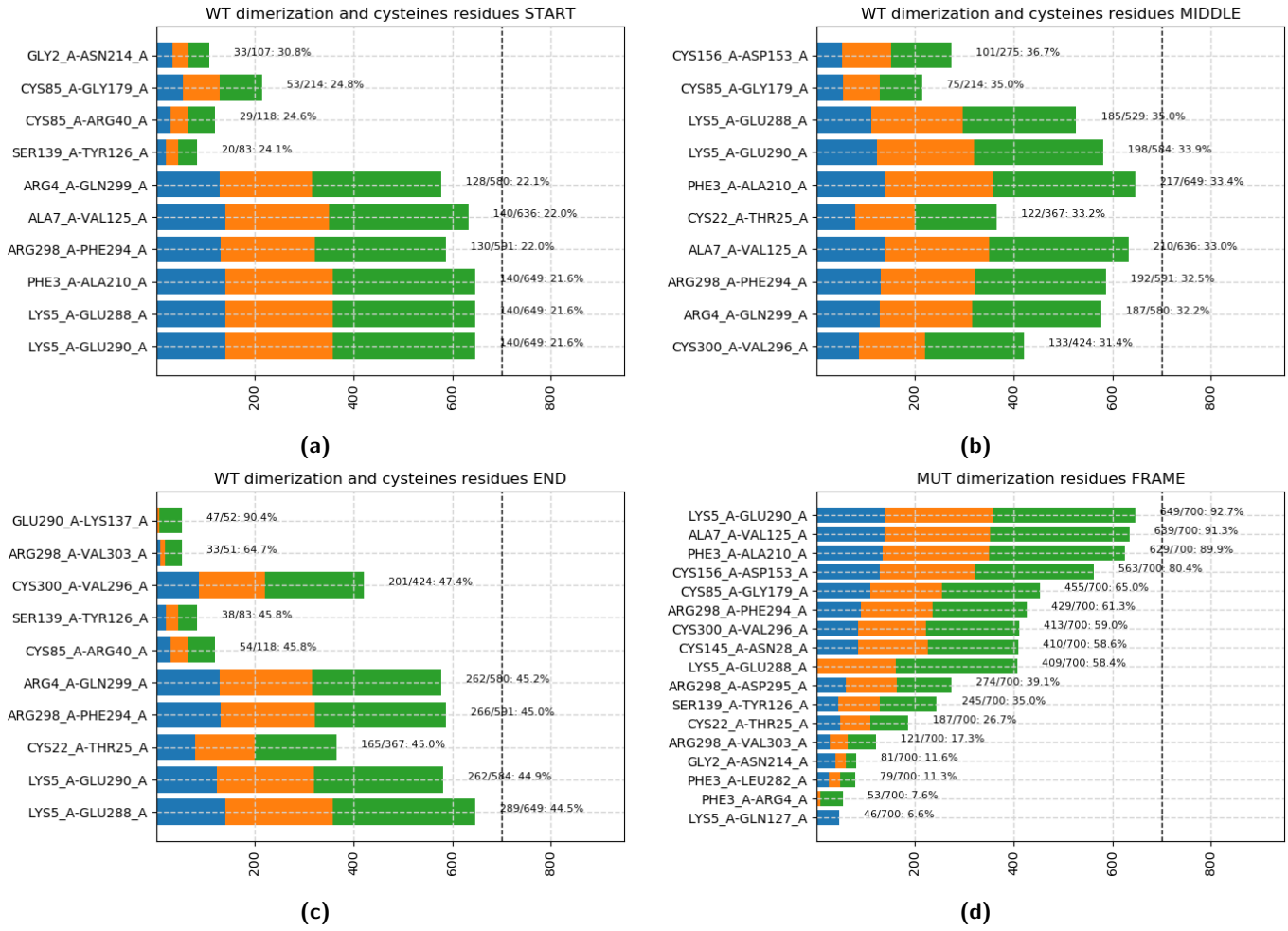


Figure 38. WT protein, bar plots of the interactions involving residues of the **Dimerization** selection. Each graph lists with the highest appearance rate in a specific fragment ("START", "MIDDLE", "END"). Finally, the figure in the bottom right corner represents the most present interactions. We considered cut-offs of 30% for the "START" section, 40% for the "MIDDLE" section and 50% for the "END" section.

References

1. Hu, Y. *et al.* Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir. Preprint, Pharmacology and Toxicology (2022). DOI: 10.1101/2022.06.28.497978.
2. Lu, J. *et al.* Crystallization of Feline Coronavirus Mpro With GC376 Reveals Mechanism of Inhibition. *Front. Chem.* **10** (2022).
3. Ma, C. *et al.* Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res.* **30**, 678–692, DOI: 10.1038/s41422-020-0356-z (2020).
4. Pham, M. Q. *et al.* Rapid prediction of possible inhibitors for SARS-CoV-2 main protease using docking and FPL simulations. *RSC Adv.* **10**, 31991–31996, DOI: 10.1039/D0RA06212J (2020).
5. PDBsum home page. http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/.
6. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38, 27–28, DOI: 10.1016/0263-7855(96)00018-5 (1996).
7. Tam, N. M., Nguyen, T. H., Ngan, V. T., Tung, N. T. & Ngo, S. T. Unbinding ligands from SARS-CoV-2 Mpro via umbrella sampling simulations. *Royal Soc. Open Sci.* **9**, 211480, DOI: 10.1098/rsos.211480 (2022).
8. Chan, H. T. H. *et al.* Discovery of SARS-CoV-2 M ^{pro} peptide inhibitors from modelling substrate and ligand binding. *Chem. Sci.* **12**, 13686–13703, DOI: 10.1039/D1SC03628A (2021).
9. Hu, Q. *et al.* The SARS-CoV-2 main protease (Mpro): Structure, function, and emerging therapies for COVID-19. *MedComm* **3**, e151, DOI: 10.1002/mco2.151 (2022).

10. Arutyunova, E. *et al.* N-Terminal Finger Stabilizes the S1 Pocket for the Reversible Feline Drug GC376 in the SARS-CoV-2 Mpro Dimer. *J. Mol. Biol.* **433**, 167003, DOI: 10.1016/j.jmb.2021.167003 (2021).

11. Nashed, N. T., Aniana, A., Ghirlando, R., Chiliveri, S. C. & Louis, J. M. Modulation of the monomer-dimer equilibrium and catalytic activity of SARS-CoV-2 main protease by a transition-state analog inhibitor. *Commun. Biol.* **5**, 1–9, DOI: 10.1038/s42003-022-03084-7 (2022).

12. Vuong, W. *et al.* Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication. *Nat. Commun.* **11**, 4282, DOI: 10.1038/s41467-020-18096-2 (2020).

13. Kneller, D. W. *et al.* Unusual zwitterionic catalytic site of SARS-CoV-2 main protease revealed by neutron crystallography. *The J. Biol. Chem.* **295**, 17365–17373, DOI: 10.1074/jbc.AC120.016154 (2020).

14. Tiberti, M. *et al.* PyInteraph: A Framework for the Analysis of Interaction Networks in Structural Ensembles of Proteins. *J. Chem. Inf. Model.* **54**, 1537–1551, DOI: 10.1021/ci400639r (2014).

15. Gowers, R. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In *Python in Science Conference*, 98–105, DOI: 10.25080/Majora-629e541a-00e (Austin, Texas, 2016).

16. Ngo, S. T., Vu, K. B., Bui, L. M. & Vu, V. V. Effective Estimation of Ligand-Binding Affinity Using Biased Sampling Method. *ACS Omega* **4**, 3887–3893, DOI: 10.1021/acsomega.8b03258 (2019).

17. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980, DOI: 10.1038/nsb1203-980 (2003).

18. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865, DOI: 10.1002/jcc.20945 (2008).

19. Vanommeslaeghe, K. *et al.* CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. computational chemistry* **31**, 671–690, DOI: 10.1002/jcc.21367 (2010).

20. Waterhouse, A. *et al.* SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303, DOI: 10.1093/nar/gky427 (2018).

21. Rostkowski, M., Olsson, M. H., Søndergaard, C. R. & Jensen, J. H. Graphical analysis of pH-dependent properties of proteins predicted using PROPKA. *BMC Struct. Biol.* **11**, 6, DOI: 10.1186/1472-6807-11-6 (2011).

22. Molecular Dynamics — GROMACS 2019-rc1 documentation. https://manual.gromacs.org/documentation/2019-rc1/reference-manual/algorithms/molecular-dynamics.html.

23. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865, DOI: 10.1002/jcc.20945 (2008).

24. Jin, Z. *et al.* Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582**, 289–293, DOI: 10.1038/s41586-020-2223-y (2020).

25. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25, DOI: 10.1016/j.softx.2015.06.001 (2015).

26. Szilárd, P., Abraham, M. J., Kutzner, C., Hess, B. & Lindahl, E. Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. In *International Conference on Exascale Applications and Software*, vol. 8759, 3–27, DOI: 10.1007/978-3-319-15976-8_1 (Springer International Publishing, 2015). 1506.00716.

27. Pavlova, A. *et al.* Inhibitor binding influences the protonation states of histidines in SARS-CoV-2 main protease. *Chem. Sci.* **12**, 1513–1527, DOI: 10.1039/D0SC04942E (2021).