

UNIVERSITÀ DEGLI STUDI DI TRENTO

DIPARTIMENTO DI FISICA

SARS-CoV-2 Main Protease (M^{pro}) S144L Mutant
in Complex with Inhibitor GC376



Authors

Francesco MUSSO

Gianluca FIORIO

Matteo SCANDOLA

December 2022

Contents

1	Introduction	2
2	Simulation of M^{Pro} 8DD9 mutant and 6WTT wild type	3
2.1	Simulation Schedule	3
3	Data Analysis	4
3.1	Temperature and Pressure	4
3.2	Protein Analysis	5
3.3	Active site reliability	8
3.4	Ligands and Binding Sites Analysis	9
3.5	Wild type comparison	16
4	Conclusions	20
	Bibliography	21

1 Introduction

M^{pro} is a class of highly conserved cysteine hydrolases from coronaviruses that plays a key role in polyprotein processing: in the process of virus multiplication, M^{pro} is capable of cleaving polyproteins at multiple sites to yield multiple functional proteins. It is composed of 306 amino acids comprising three domains. Domains I (residues 8–101) and II (residues 102–184) together exhibit a chymotrypsin-like fold, and domain III (residues 201–306) comprises a cluster of five alpha-helices connected to domain II by a long loop (residues 185–200). M^{pro} is active in a dimeric form (Figure 1). Each sub-unit harbors an active site consisting of the catalytic dyad H41-C145; despite the monomer form adopting a native-like tertiary fold, monomeric variants of M^{pro} are reported to exhibit very low or no catalytic activity. [1]

The aim of this project is the study of the naturally occurring drug resistant M^{pro} mutant S144L in complex with inhibitors (GC376) (PDB: [8DD9](#)). This mutation concern the exchange of the residue 144, which is a Serine (Ser) in the wild type chain, with a Leucine (Leu). Although they have a similar molecular weight ($M_{Ser} = 105u$ and $M_{Leu} = 131u$) this amino acids has a fundamental difference: the Serine is polar, whereas the Leucine is apolar. We expect that the shielding effect due to hydrophobic and hydrophilic residues will induce a change in the conformation of the active site.

We have simulated the dimeric form of M^{pro} mutant in the HOLO form. After the simulation, we have monitored the protein fluctuations, and in particular the conformations of the binding sites.

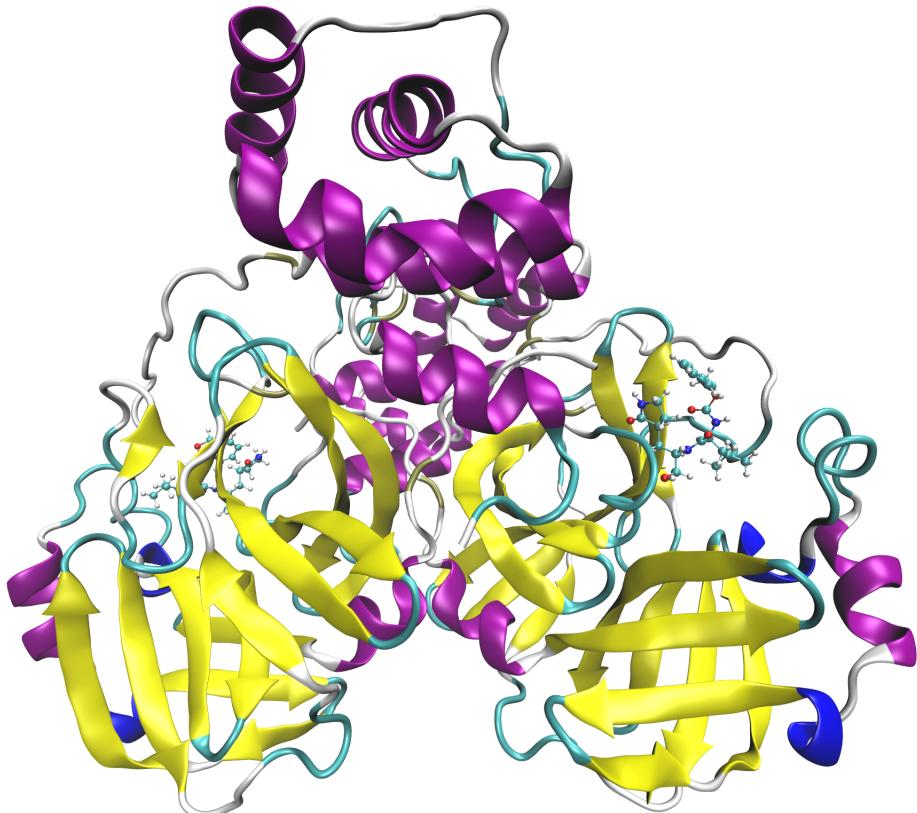


Figure 1: Dimeric form of M^{pro} in complex with GC376 inhibitors.

2 Simulation of M^{pro} 8DD9 mutant and 6WTT wild type

The systems were created using the **CHARMM-GUI** tool, starting from the 8DD9 PDB. Other than allowing to setup the periodic boundary conditions, the solvent, the desired force field (i.e. CHARMM36 in this case) and outputting the system configuration files in the specific format for the most common MD engines, the CHARMM-GUI solution builder tool permits also to mutate the protein.

In particular we started from the mutated 8DD9 input files and we mutated back the protein to the wild type in order to simulate the **6WTT** in the same conditions as the ones given to us for the 8DD9 mutation.

The protein and the ligands are confined in a PBC box of 10.3 nm, solvated in water containing NaCl ions with 0.15 M concentration.

2.1 Simulation Schedule

We performed the following steps in the simulation:

1. **Energy Minimization** - Using gradient descent we minimized the potential energy of the system in order to start the simulation near the equilibrium. The algorithm stops after the biggest force between two components in the system gets below a certain threshold.
2. **NVT Equilibration** - We equilibrated the system to the desired temperature, keeping fixed the protein and the ligands, for 500 ps.
3. **NPT Equilibration** - We equilibrated the system to the desired pressure and temperature, solving possible pathological effects (voids) resulting from the previous NVT equilibration, for 1 ns.
4. **Production Run** A full production run of 500 ns in the NPT ensemble with
 - a timestep $\Delta t = 2$ fs, sampling configurations every 50 ps
 - temperature and pressure are kept constant at 310 K and at atmospheric pressure respectively, to reproduce physiological conditions in human cells.

We checked that the system actually equilibrated after point 2 and 3, by inspecting temperature and pressure values. After some picoseconds in both cases the quantities converged to the required value, fluctuating around it.

The simulations were performed on the university cluster HPC2, utilizing 4 nodes and 64 CPU cores per node, obtaining an average performance of 80 ns per day.

3 Data Analysis

This section is devoted to the analysis of the data gathered from our simulations. The first part will regard the time evolution of the temperature and pressure of the system during the production run. After that some standard metrics computed on the MD simulation are displayed, subsequently a time-windowed analysis will be performed to understand the behaviour of system over time. Finally a comparison to the wild type structure is performed.

Since in the first run we've observed the ligand diffusion we decided to perform another MD simulation of the mutation to confirm that it wasn't due to an error or an artifact: from now on results for both run will be presented.

3.1 Temperature and Pressure

Let's start with the time evolution of temperature and pressure that are computed using Gromacs, a moving average of 5 ns window is performed obtaining:

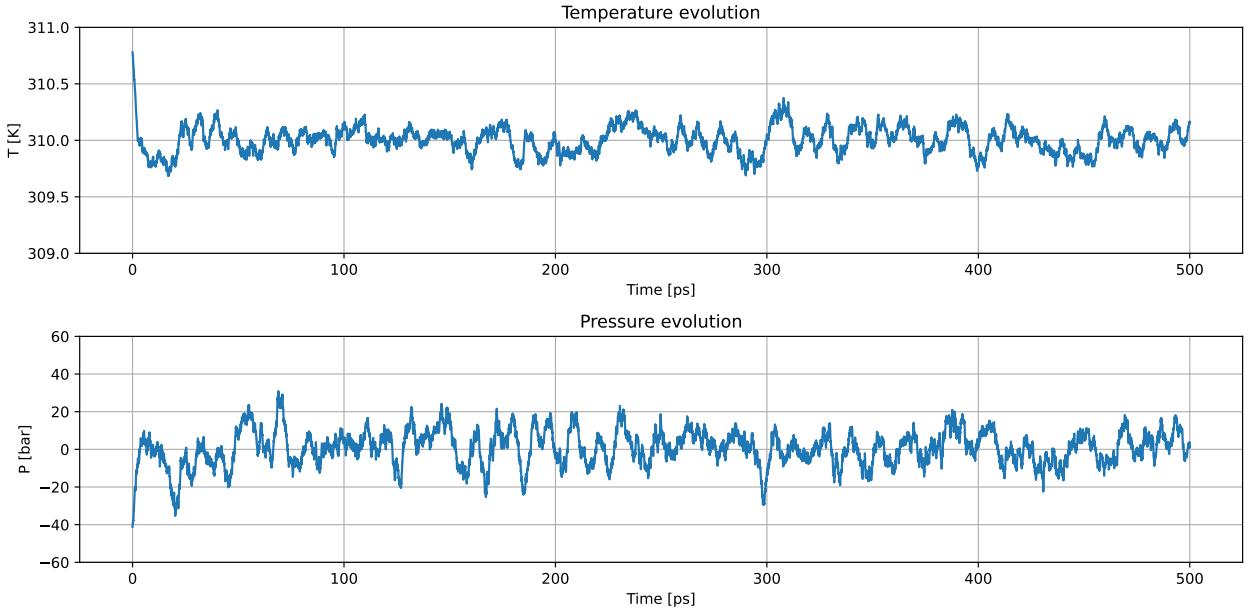


Figure 2: Temperature and pressure time evolution, both observables are time-averaged over 5 ns

Since the difference between runs is negligible, only the results regarding run 2 are considered here. Both temperature and pressure fluctuate around their respective stationary value:

Temperature	Pressure
$(310.0 \pm 0.1) \text{ ns}$	$(0.8 \pm 9.2) \text{ bar}$

We note that the mean pressure value is a little off from the desired value, however due to the large variance of it we don't think that's a problem.

This is a pretty trivial result since our system at this point has already gone under Energy Minimization, NPT and NVT equilibrations, but it's still an important check to verify the correctness of the MD simulation.

3.2 Protein Analysis

In this section some of the standard metrics will be computed using the MDAnalysis library of Python and some proprietary code that can be found at [repository](#).

These are high-resolution metrics, which give insights regarding macroscopic events and help the identification of conformational changes and convergence to equilibrium.

RMSD & RMSD - map

The Root Mean Squared Deviation (RMSD) is the average distance between the atoms of superimposed proteins.

$$\text{RMSD}(t) = \sqrt{\frac{1}{N} \sum_{i=0}^N \left(\mathbf{x}_i(t) - \mathbf{x}_i^{\text{ref}} \right)^2} \quad (1)$$

In the RMSD calculations we consider the reference structure to be the first frame of our production run simulation and we perform these calculations on the C_α atom selection. We also compute the RMSD-map: a matrix representing the RMSD at each frame in respect to the other frames of the trajectory.

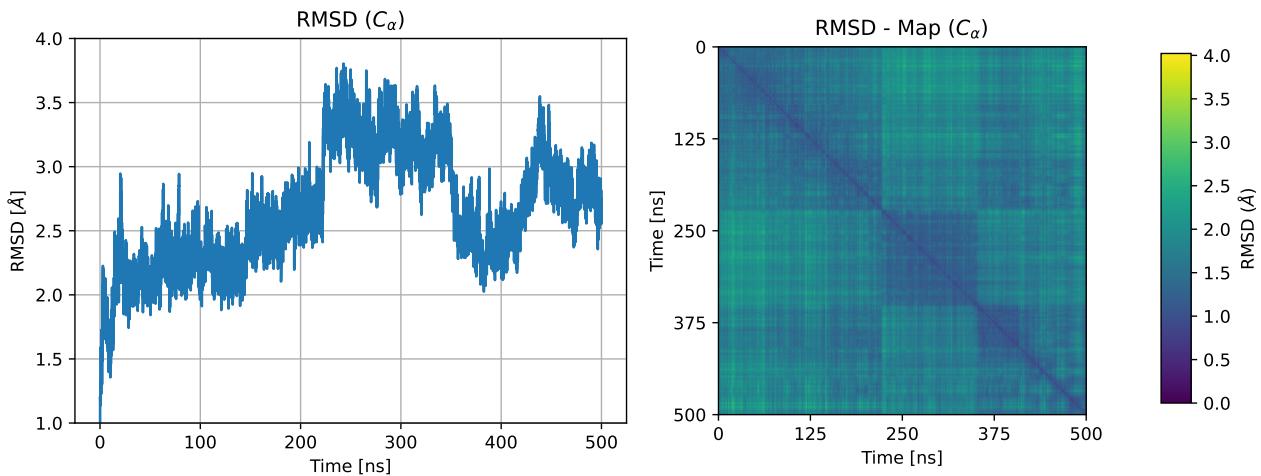


Figure 3: Run 1: C_α Root Mean Squared Displacement

In the first run, we can see that apart from the first transient phase the protein performs a big conformational change at around 220 ns. The structure seems pretty much stable up until around 350 ns, where the protein seems to undergo rapid and frequent conformational changes.

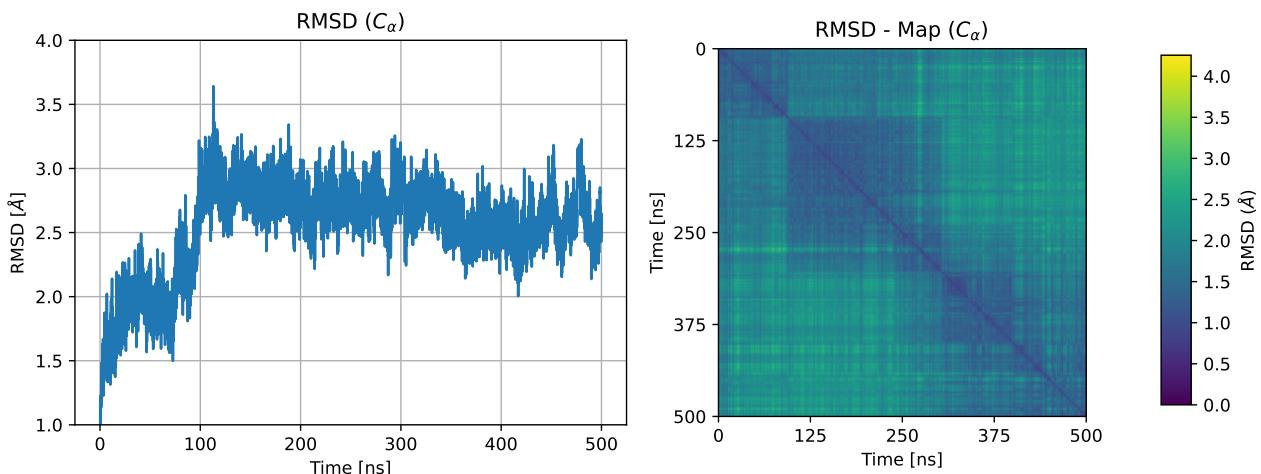


Figure 4: Run 2: C_α Root Mean Squared Displacement

In the second run, we can see that around 100 ns, the protein performs a conformational change and stabilize up to around 300 ns and again start to jump between different macrostates.

Root Mean Squared Fluctuations

The Root Mean Square Fluctuations (RMSF) measures the flexibility of the individual residues of the protein chain:

$$\text{RMSF}_i = \sqrt{\langle (\mathbf{x}_i - \langle \mathbf{x}_i \rangle)^2 \rangle} \quad (2)$$

In the following figure we can see the RMSF of the residues computed on the whole trajectory for each chain:

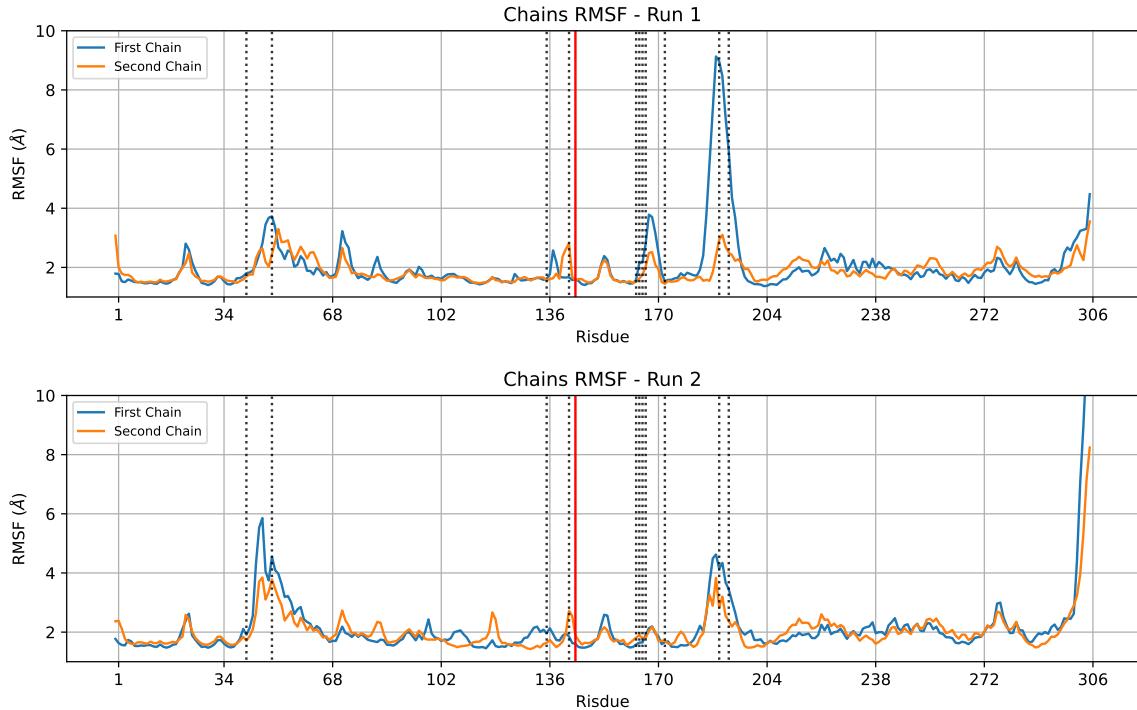


Figure 5: RMSF for each chain and run. Black dotted lines represent the binding sites residues and the solid red line represents the mutated binding site residue.

One can see that in correspondence of the active site residues of the chains there are higher values of RMSF, indicating a higher mobility throughout the trajectory. In particular in the first run, as we'll discuss in the next section, the first binding site seems to "opens up" early in the simulation letting the attached ligand diffuse away.

Since our trajectory presents the diffusion of the ligands we computed the RMSF of the two chains with a time window of 25 ns and a stride of 1 ns in order to see if there's a substantial difference of this metric over time.

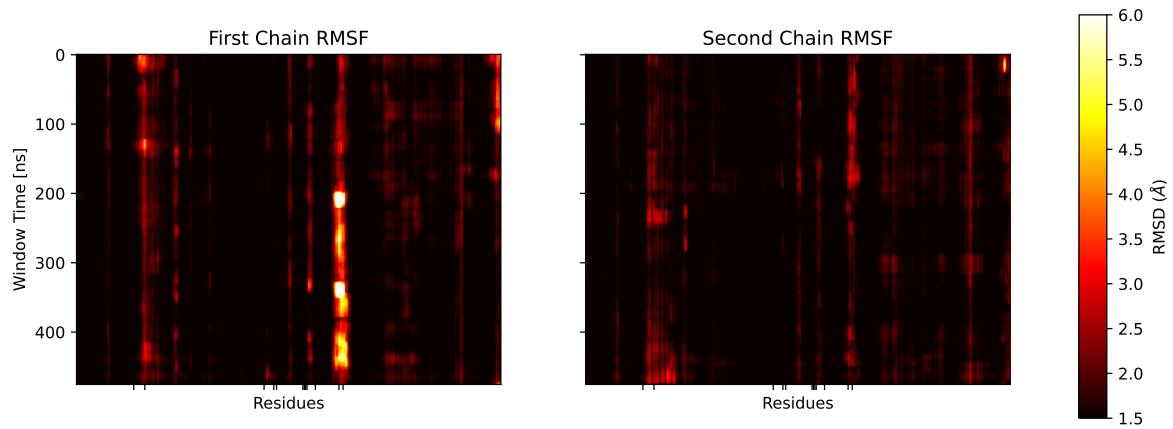


Figure 6: Run 1: RMSF windowed in time, residues ticks are the binding site ones

In this run we can see that the most mobile residues are the binding site ones, in fact around 200 ns the RMSF of two them (Q189 & Q192) increase dramatically, reaching 10.5 \AA suggesting that the binding site is very mobile.

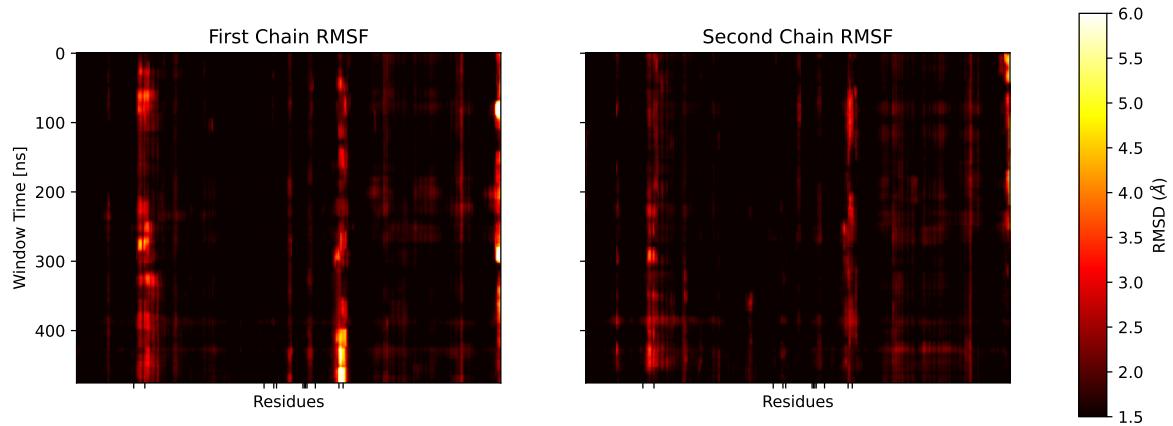


Figure 7: Run 2: RMSF windowed in time, residues ticks are the binding site ones

In the second run we see again that the most mobile residues are the binding site ones, the difference here is that the RMSF increase is more subtle and happens much later in the trajectory, around 400 ns.

3.3 Active site reliability

In the reference paper [1] the active site residues (H41, M49, T135, N142, L144, H163, H164, M165, E166, H172, Q189, Q192) were chosen by looking at the residues that are located within 6 Å of the nirmatrelvir inhibitor binding site. The structure used for this calculation was the SARS-CoV-2 M^{Pro} in complex with Nirmatrelvir (PDB: [7SI9](#)).

We want to check if this active site is a good choice even in our case, that is the S144L mutation of the M^{Pro} in complex with the GC376 inhibitor. For each binding site we consider the first part of the trajectory in which the ligand is docked, and count how many times the ligand is at a smaller distance than 6 Å from different residues, obtaining the following figures:

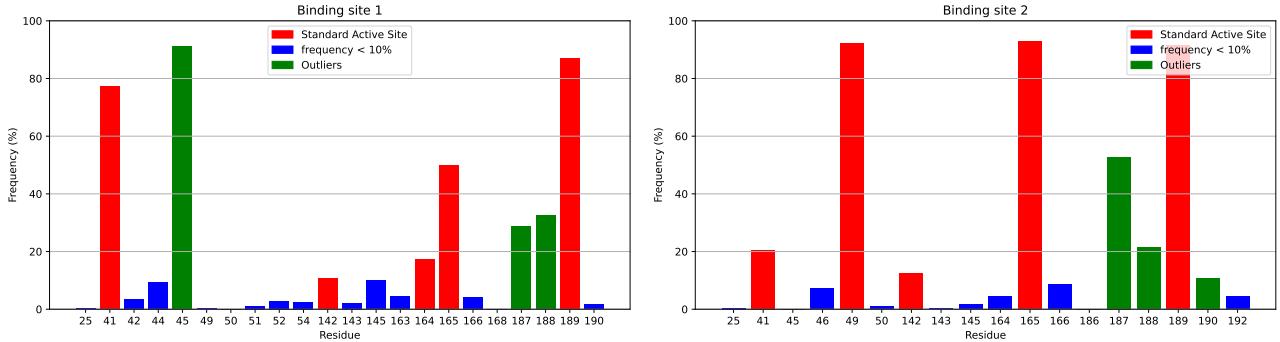


Figure 8: Run 1 - Frequency of residues-ligand contact throughout the simulation

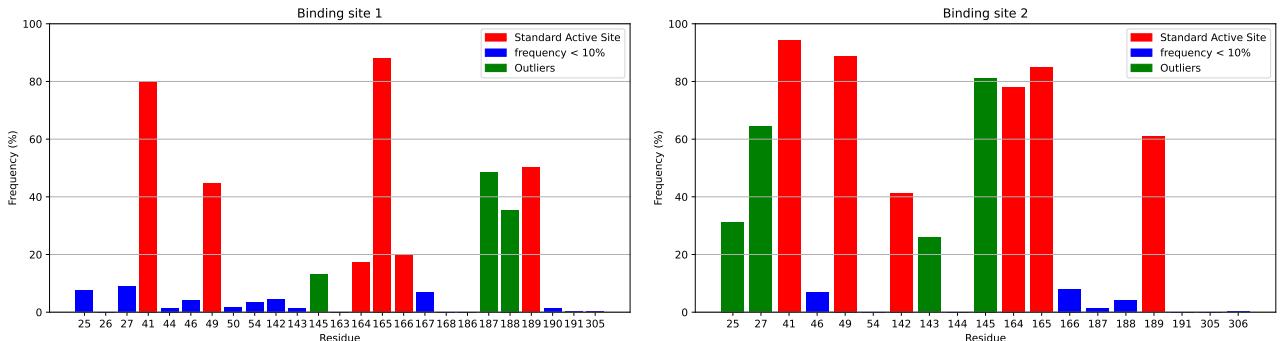


Figure 9: Run 2 - Frequency of residues-ligand contact throughout the simulation

Among the residues that are more frequently in contact with the ligand, there are both residues that are in the standard active site (red bars), but also the so-called "outliers" (green bars), which are those residues that we didn't take into account in the previous analysis.

From the figures we can see that there is a good agreement between residues that are in contact with the ligand in different protomers and different runs, maybe with the exception of the first pocket of the first run. As we'll see later in Figure 13, we think this is due to the fact that the first run has a very short section with the ligand in the site. In particular, we find that standard active site residues H41, M49, N142, H164, M165, E166 and Q189 are frequently in contact in every protomer and run, while most of the residues that we called outliers, such as T25, L27, T45, S46, G143 and C145 are frequently in contact depending on the simulation and on the binding site. The only outliers with a high frequency in every protomer and run are the D187 and R188, and this is probably due to the fact that they are near to the Q189 binding site residue. We then consider the reference binding site as valid.

Lastly, we observe that the mutated residue, L144, is never in contact with the inhibitor. This could be due to Leucine, an apolar amino acid, suffering from hydrophobic shielding and being pushed towards the center of the structure, underneath a hydrophilic rind.

3.4 Ligands and Binding Sites Analysis

With reference to the previous section, the binding site is defined as the one considered in the reference article [1].

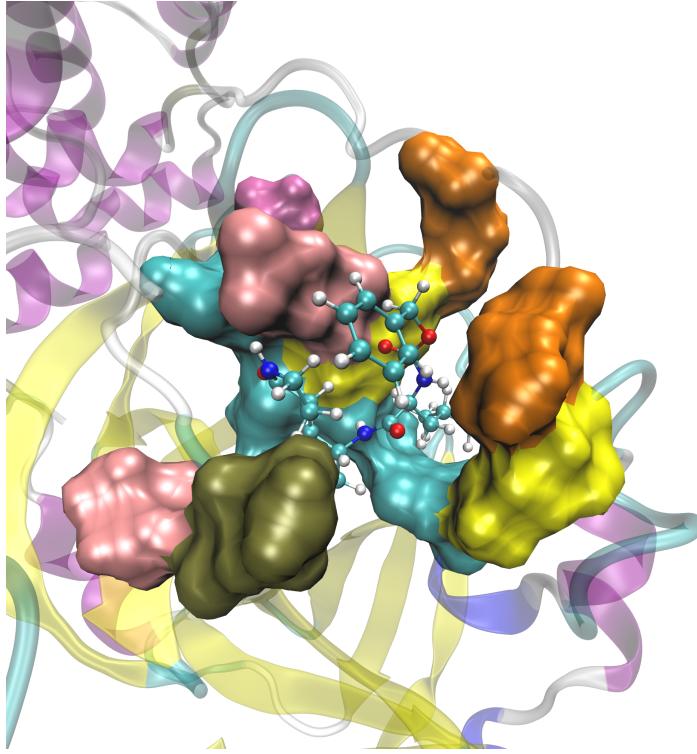


Figure 10: Binding site with ligand.

From the experimental results we know that the mutation under study presents a higher drug resistance than the wild type. We'll study the interaction of the ligand with the site and we'll see that in both cases, in the mutation and in the wild type, the ligands leave the pocket. For this reason we analysed separately the active site conformation while the ligand is in the active site and while it is not.

Ligand position

The first step in this direction is to define when the ligand is actually docked and when it is not. To obtain a cutoff value for the distance between ligand and binding site, we considered multiple factors: the radius of gyration (R_{gyr}) of the active site (11) and the hydrogen bonds between the ligand and the pocket over time (12).

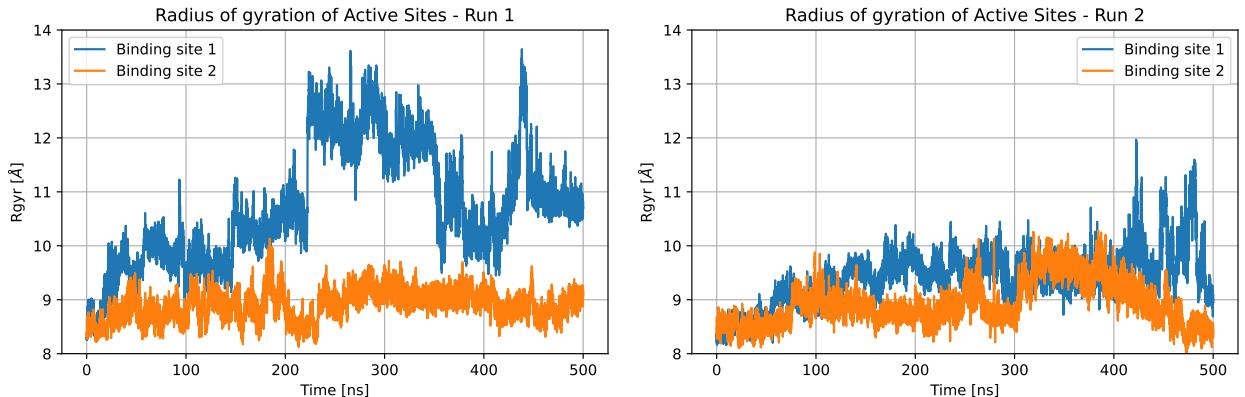


Figure 11: Run 1 (left) and 2 (right). Radii of gyration of the active sites.

The radius of gyration of the active site fluctuate between 8 \AA and 14 \AA . Therefore, with the help of the hydrogen

bond analysis, we expect to find a value for the cutoff in this range.

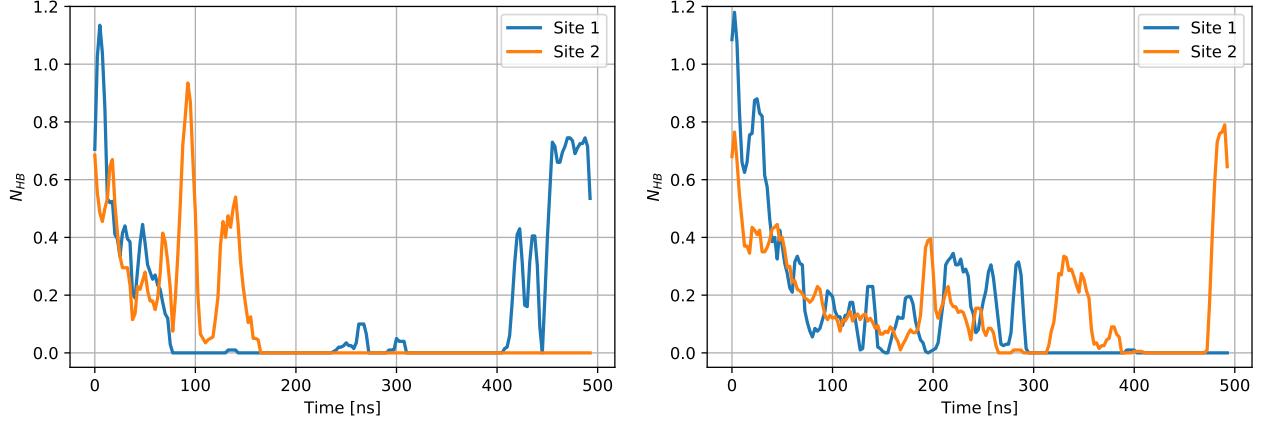


Figure 12: Run 1 and 2 - Moving average of 10 ns window of the number of hydrogen bonds between the ligand and the binding site over time

We established the cutoff at $d_{\text{cut}} = 13 \text{ \AA}$ since it is larger than the R_{gyr} of the site (the size of the pocket), in the time intervals where hydrogen bonds are present. Finally we computed the distance between the center of mass of the ligand and the one of its active site and considered the ligand bound to the pocket when the distance is less than d_{cut} . Then we can divide the trajectory into the two aforementioned active sites macro-configurations, with and without the inhibitor, as in Figure 13.

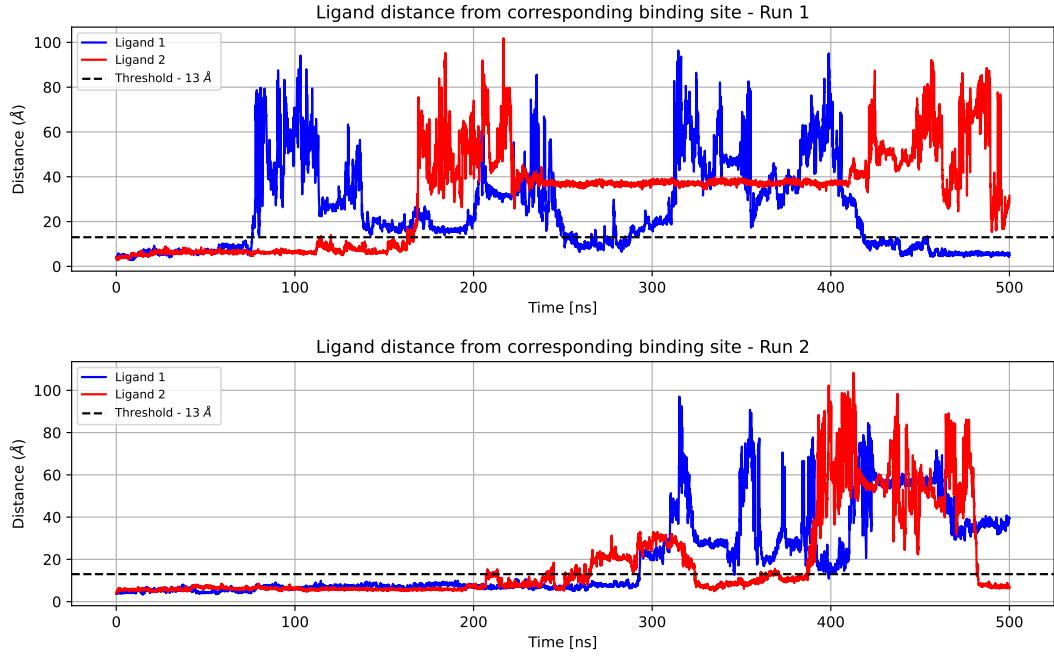


Figure 13: Ligand distance from respective binding site. The horizontal line represents cutoff distance of 13 \AA

We can see that in both runs the ligands depart from the relative binding site. In run 1 this happens early in the trajectory, around 80 ns for the first ligand and around 170 ns for the second one. Ligand 1 seems to re-attach to the binding site for very brief intervals of time, while ligand 2 seems to attach more stably to another point in the protein.

In run 2 the ligand departures happen much later in time than in run 1, around 290 ns for ligand 1 and around 200 ns for ligand 2. We see again that one of the ligand re-attach itself to the binding site for some time.

In the next section, we explore the different configurations of the binding sites separating the configurations with or without the ligand.

RMSD & Radius of Gyration

Firstly, we focus on the configurations of the binding site with the ligand docked. For the time evolution (Figure 14) of the pocket we made a scatter plot just for the second run, since it has much more configurations.

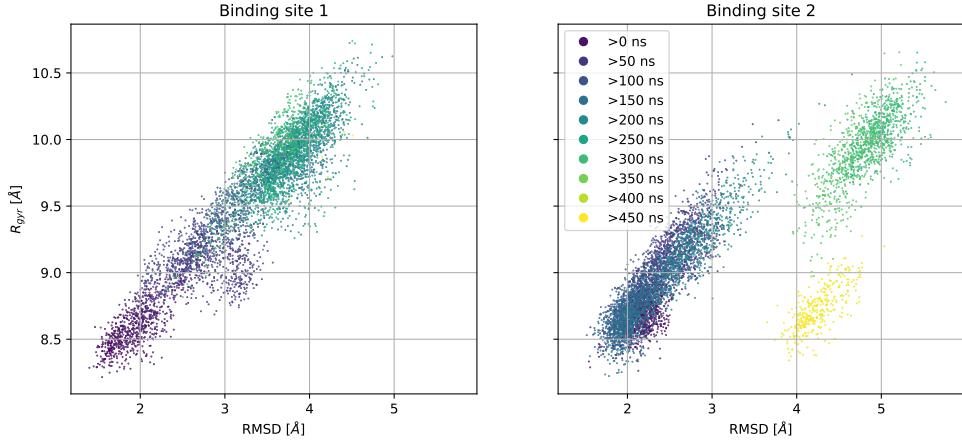


Figure 14: Run 2: Time evolution of the configuration RMSD/ R_{gyr} of the pocket. The different colors has to be intended as an arrow of time.

One can easily see that:

- The configurations of the first active site emphasizes that the binding site tends to open up during the simulation until the ligand leave the pocket.
- The second site, which binds for three times with the inhibitor, changes every time the configuration. In the very last part of the trajectory (yellow dots) the configurations the pocket has almost the same size but it is arranged in a quite different fashion.

Moreover, we can also try to characterize and divide these clouds in clusters. Since it is clear that the binding site is very flexible the number of clusters is arbitrary and based on the RMSD-map in Figure 20. With agglomerative clustering, we end up with Figure 15, which nevertheless reveals a good division in clusters looking at the distribution bellies.

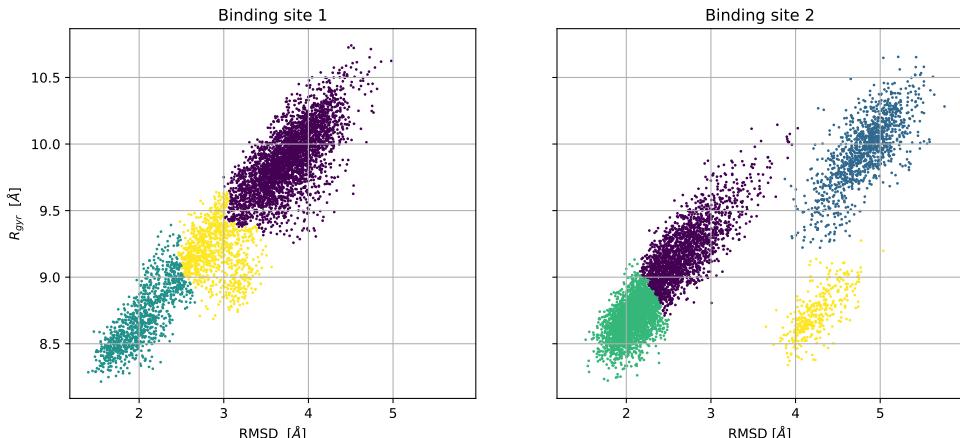


Figure 15: Run 2: Clustering of configurations with an agglomerative clustering algorithm

In order to study the evolution of pocket with or without the ligand we report series of scatter plots that are aimed to show that the presence of the ligand actually forces the active site to a different configuration. Figure 16) reveals that the pocket has a preferential dimension, which is an opening and closing movement, but also that this can happen with the ligand in the pocket or without the ligand in the site. In both runs the first a pocket presents wider opening movement, while the other one explore very different and distinct conformations.

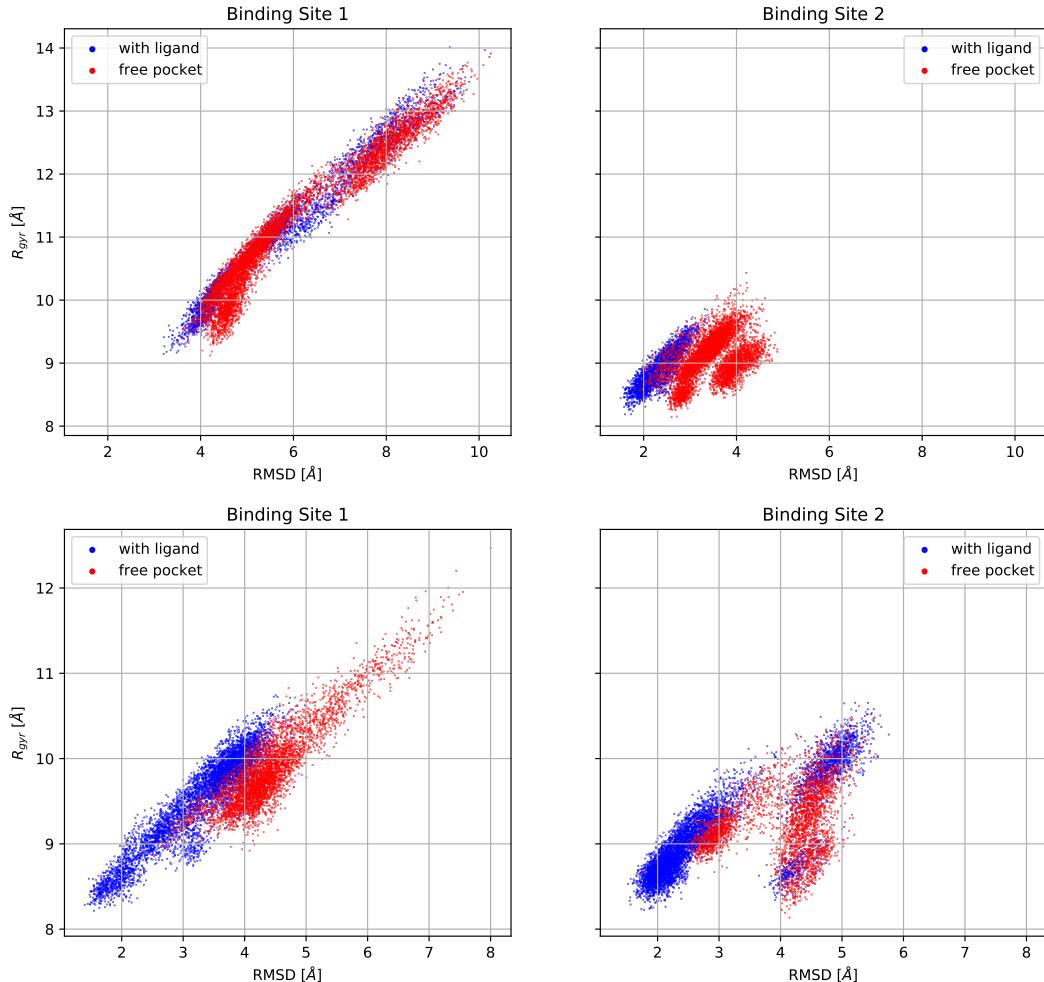


Figure 16: Run 1 and 2 - scatter plots of binding sites configurations with (blue) and without (red) ligands

In all these plots, with the exception of the first one, it is clear that the site without the ligand explores different conformations regions than the site with the ligand does.

RMSD & RMSD-map

In this section we further analyze the conformational changes of both ligands and binding sites. We start by observing the RMSD and the RMSD-map of the ligands.

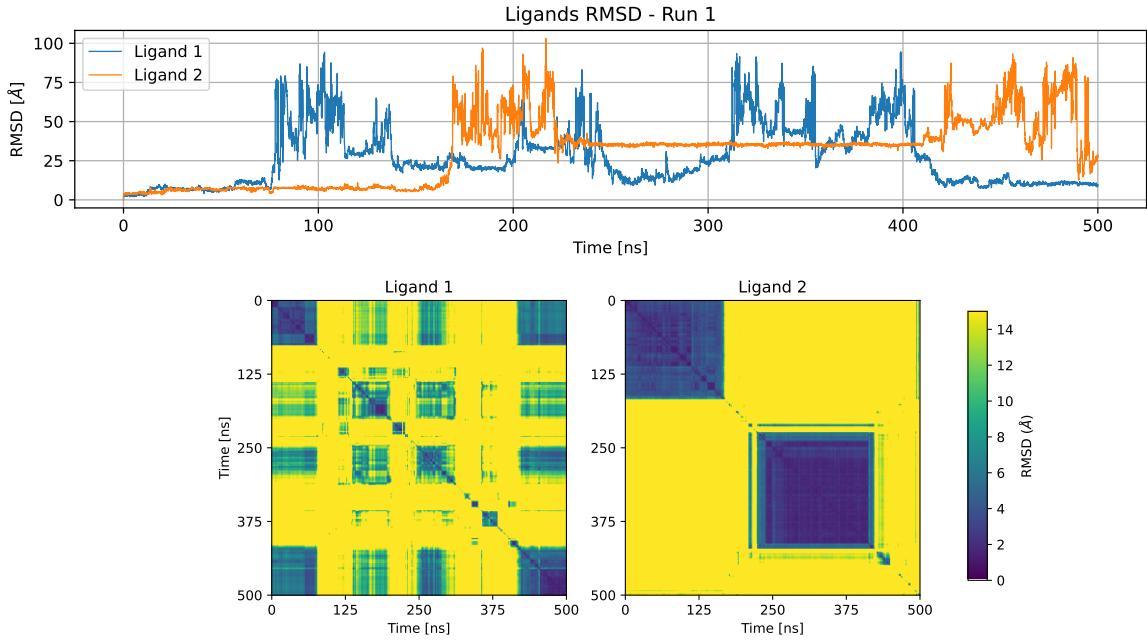


Figure 17: Run 1 - RMSD and RMSD-map of the ligands

With the help of the ligand distance metric (13), we can confirm that in first run the second ligand, after diffusing in the solvent, attaches to the protein and remain stable for around 150 ns. We note that also the first ligand tends to re-attach to the protein, in this case as we'll see in the ligand Distance Heatmap section 3.4 it happens near the original binding site.

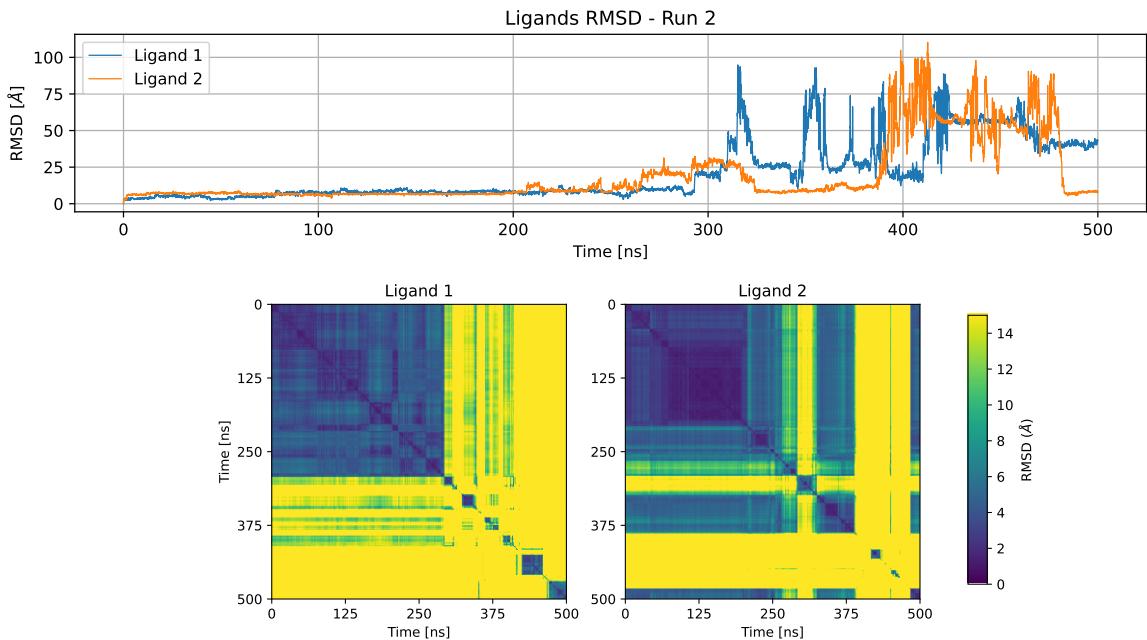


Figure 18: Run 2 - RMSD and RMSD-map of the ligands

In the second run we see a much longer stable initial phase, which when broken indicates the ligands diffusion around the protein. Again, as in the first run, we can see that the ligands tend to stabilize for short time intervals.

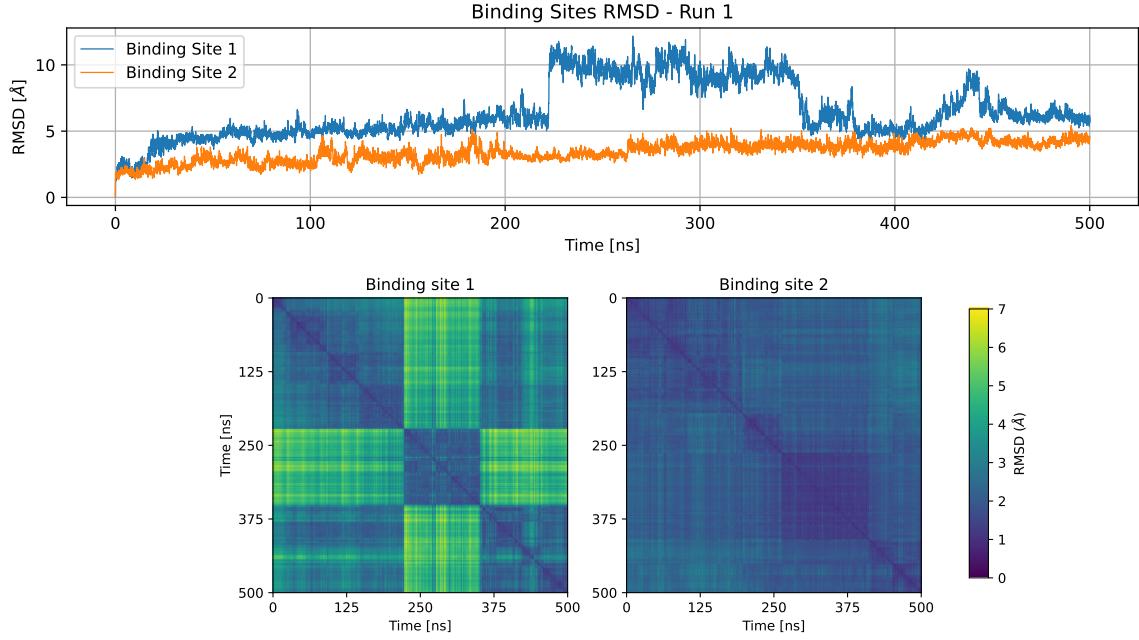


Figure 19: Run 1 - RMSD and RMSD-map of the binding sites

Using the R_{gyr} of Figure 11 and the RMSD/RMSD-map of Figure 19, we see that the most evident conformational change regards the first binding site around 230 ns, which seems to coincide with an opening of the pocket. The opening of the site happens much later than the ligand departure: this behaviour could be attributed to the action of the ligand, but a significant conformational change right after the ligand detaching (which could have lead to the opening in the following) doesn't occur and it was not replicated in the other pocket and neither in the second run. At around 350 ns the site seems to close again and change with the approach of ligand in the last 50 ns.

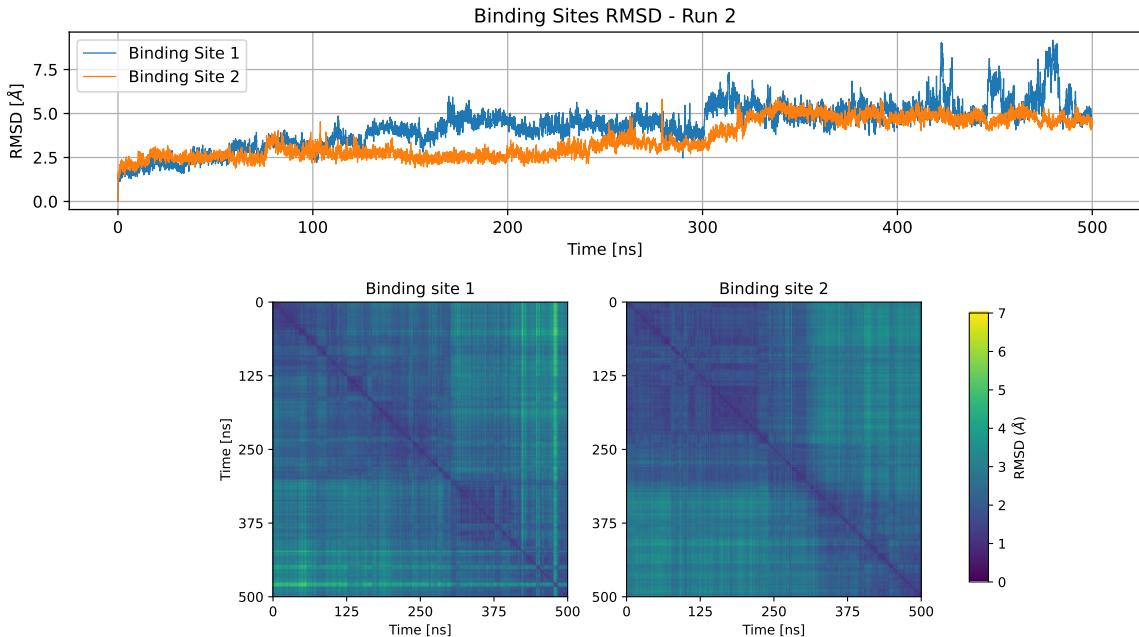


Figure 20: Run 2 - RMSD and RMSD-map of the binding sites

In the second run we see that after around 300 ns both sites undergo a minor conformational change, right after the departure of the ligand. Both sites seems to be much more unstable after the departure of the ligand.

Ligand Distance Heatmap

In this subsection, we try to get some insight in regard to where the ligands are attached to the protein by computing the distance between its center of mass and every C_α atom of the protein.

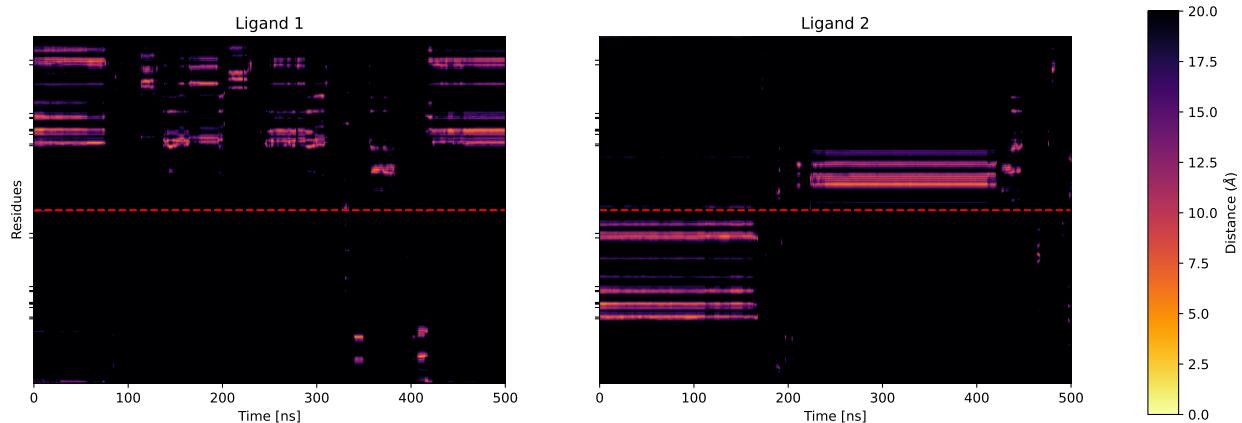


Figure 21: Run 1 - Ligand Distance from C_α , ticks represent the two binding sites. Red middle h-line represent the division between the two protomers.

In the first run we get more information than the previous computed metrics, in particular we can now say that, after the diffusion, the first ligand seems to re-attach to the corresponding binding site while the second one re-attaches to another set of residues of the first chain of the protein, but more stably.

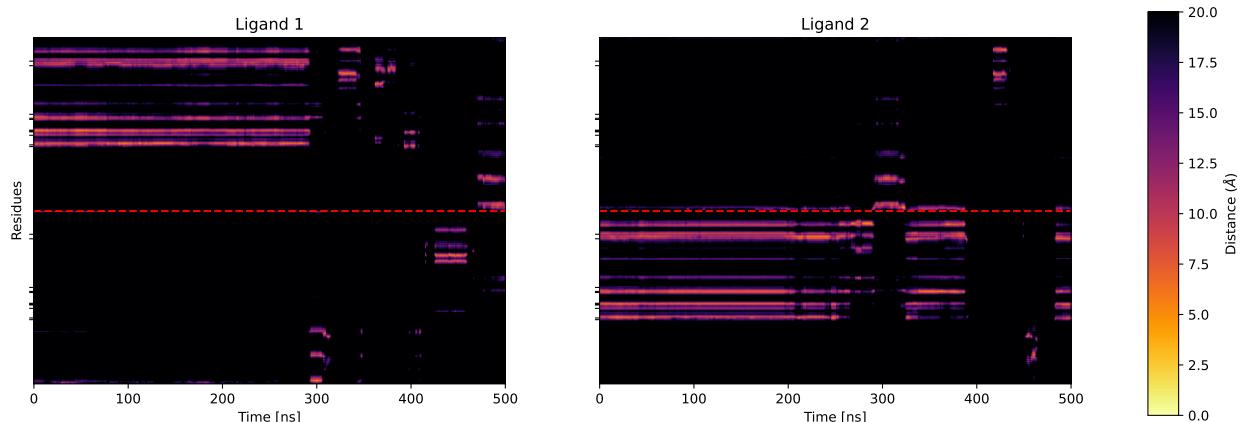


Figure 22: Run 2 - Ligand Distance from C_α , ticks represent the two binding sites. Red middle h-line represent the division between the two protomers.

Again in the second run we can see that the first ligand, after the diffusion, tends to gravitate around the protein in different spots, while the second ligand tends to re-attach to the binding site.

This analysis considers only the distance between the ligands center of mass and the C_α atoms of the protein, for a more rigorous analysis one should consider actually the various bonds between the ligands and the protein.

3.5 Wild type comparison

In this section we will compare some of the results presented above and others to the one obtained for the 6WTT, i.e. the wild type protein corresponding to our mutation 8DD9. We know from the reference article that the mutation should be more resistant to the inhibitor and we look for some confirmation of it in our trajectory.

Ligand and Binding sites

We used the same cutoff distance to consider the ligand detached from the site since it is compatible with the same reasoning and data used to establish the cutoff for the mutation.

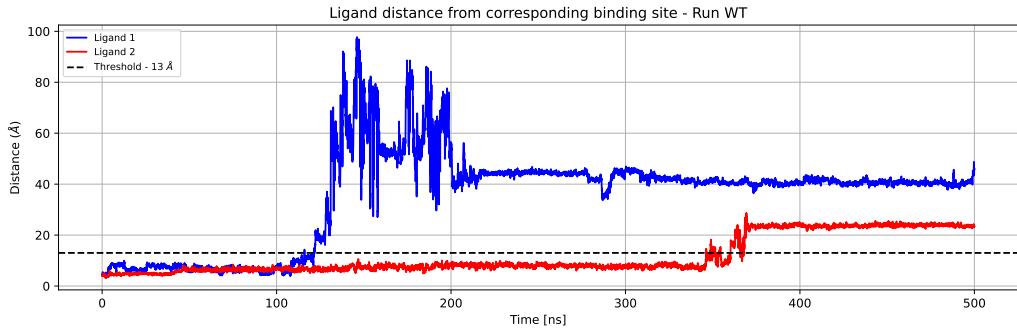


Figure 23: Ligand distance to binding site throughout the simulation of the wild type

Having determined the time of departure of the ligands from their binding site, we can study the effects of the pockets dynamics by observing its radius of gyration and RMSD/RMSD-map:

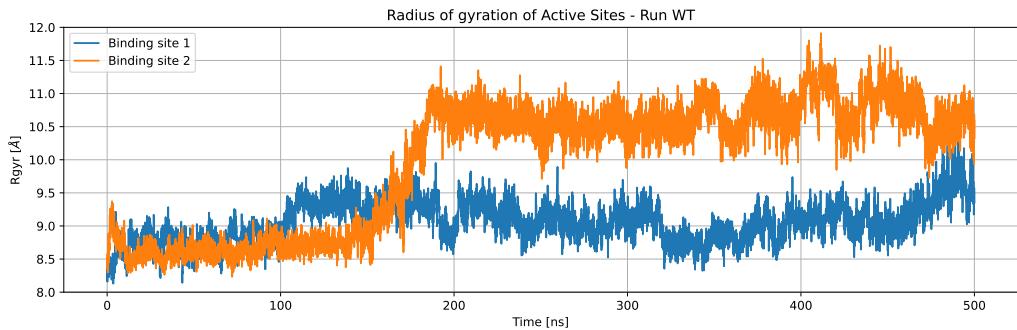


Figure 24: Radius of gyration of the binding sites of the wild type

From the R_{gyr} and the RMSD/RMSD-map in Figure 24 and Figure 25 we can see that in the first binding site the most evident conformational change happens right before the ligand departure and probably coincide with an opening of the pocket, while in second binding site the pocket opening is much more pronounced but it doesn't coincide with the exit of the inhibitor. The second ligand stays stably in the pocket after the opening for around 150 ns, before leaving the site at around 350 ns.

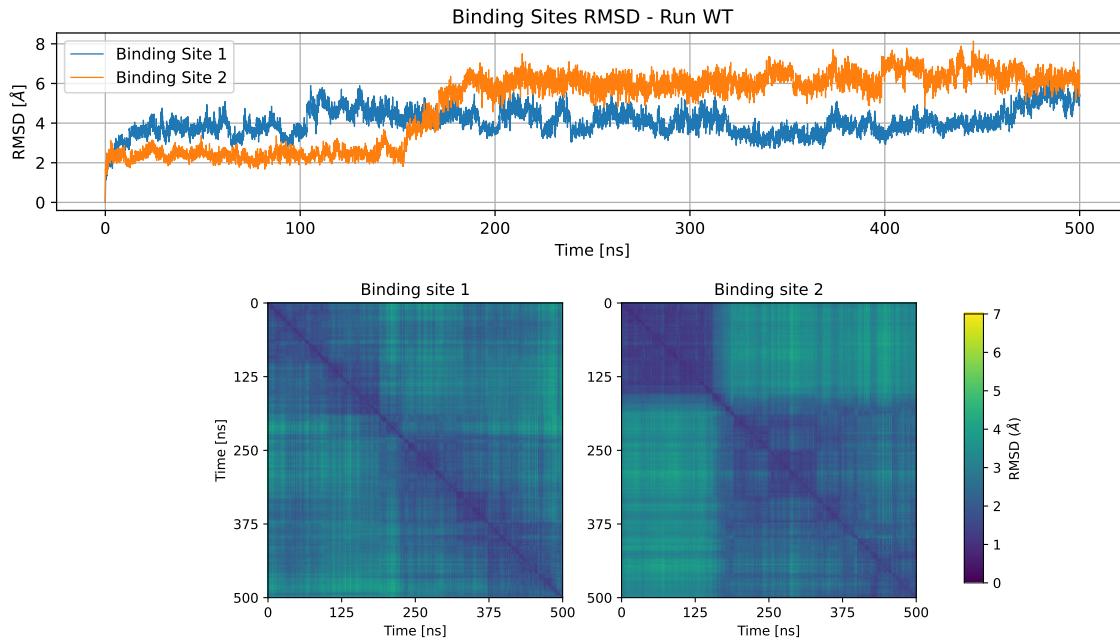


Figure 25: Run 2 - RMSD and RMSD-map of the binding sites

The most apparent difference with the mutation seems to be the effect the ligand has on the binding sites: in the case of the mutation the biggest conformational changes of the sites happened after the ligand exit, while in this case the ligand seems to have a bigger impact in the dynamics of the site. Another important difference is the fact that the ligands seems to reattach to the protein in a different spot and it seems to do it more stably: as can be seen in Figure 23, both ligands do it and the first one stays in the same place, with respect to the protein motion, for almost 300 ns.

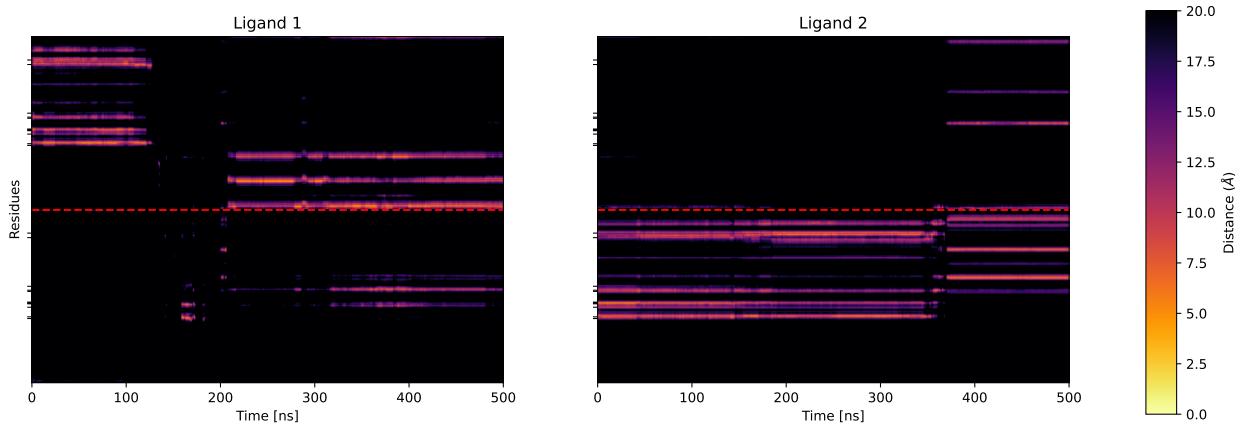


Figure 26: Run wt - Ligand Distance from C_α , ticks represent the two binding sites. Red middle h-line represent the division between the two protomers.

A clearer evidence of this behaviour can be seen in Figure 26 where we report the distance of the ligand to the C_α of the protein over time. In particular we see that the first ligand attaches itself to other residues in the first chain after diffusing for some time in the solvent.

Hydrogen bond analysis

We also compared the hydrogen bonds between the binding sites and the ligands of the mutation and the wild type. We first report the number of hydrogen bonds over the time the ligand is in the binding site. In order to obtain a more intelligible graph we calculated a 10 ns moving average with a stride of 2.5 ns. The data for the

mutation are reported in Figure 12, while the wild type results are reported in Figure 27.

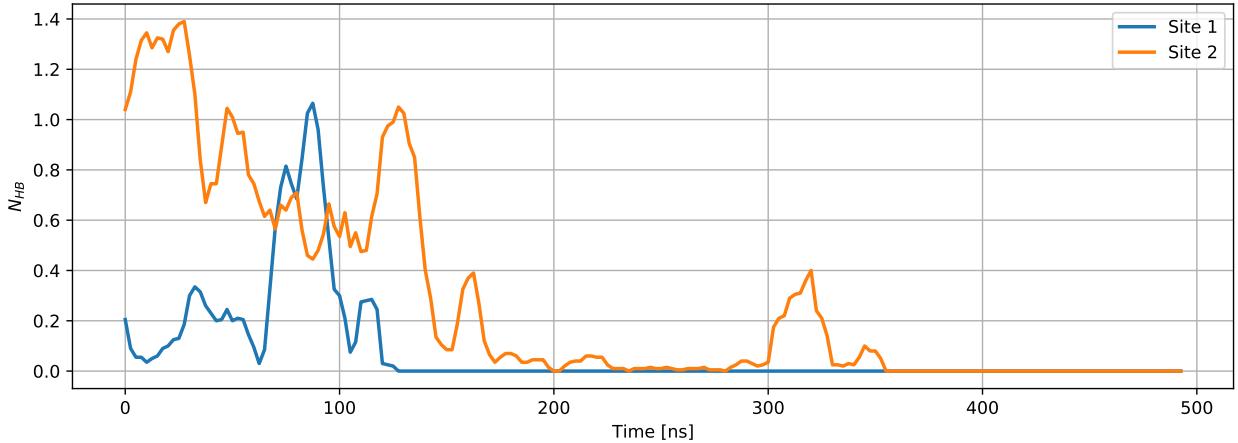


Figure 27: Run wild type - Number of Hydrogen Bond Analysis

As can be seen from the graphs, the hydrogen bonds, that are present in the beginning due to the minimization of the potential energy, are less stable and tend to break more quickly in the mutation than in the wild type. The ligand in the site 1 of the wild type start in a less potentially favorable configuration, where basically no hydrogen bonds are present: as a result of this the first ligand diffuse away from much before the ligand 2 does.

In order to understand which of the residues take part in building the hydrogen bonds we report the persistence of the bonds over the time which the ligand spends in the binding pocket.

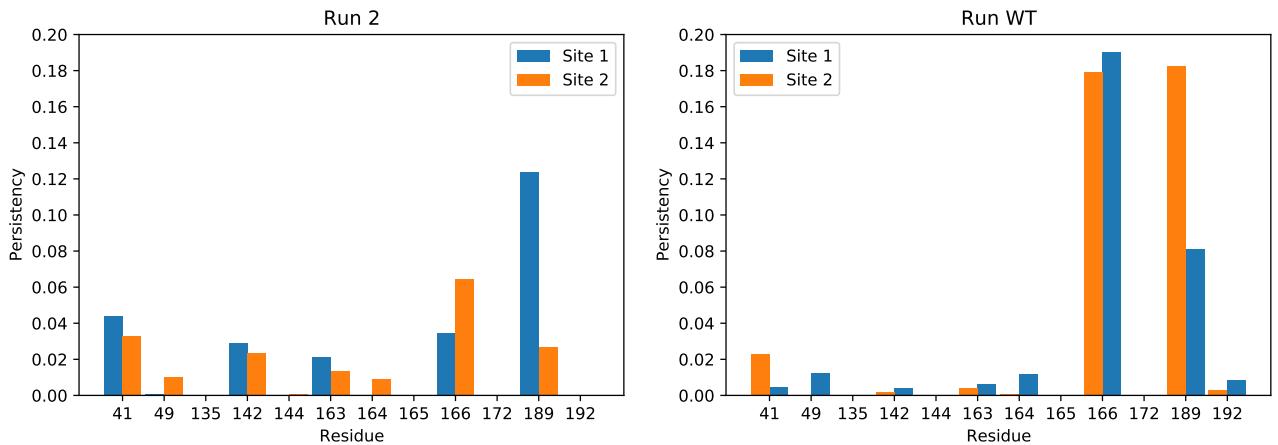


Figure 28: Run 2 and Wild type hydrogen bond persistance

Here it's even clearer that the wild type creates more hydrogen bonds than the mutation. The most active residues are E166 and Q189 in both proteins, but most of the residues don't reach 5% persistence. The main difference between the two is the distribution of the bonds: in the mutation the bonds are more evenly distributed between all the residues in the pocket, while in the wild type the distribution is more peaked in the most active residues.

From this observations we can say hydrogen bonds play a role in the ligand-pocket binding, but not a crucial one. These can be easily seen in the final part of the graph of the number of hydrogen bonds of the second binding site of the wild type in Figure 27: the ligand keeps staying in the binding site without forming approximately any hydrogen bond at all for a considerable period of time.

Binding energy

In order to make a quantitative comparison with the experimental values, we tried to estimate the binding energy of the ligand for both the mutation and the wild type.

We used the linear interaction energy (LIE) method [2]: an approximate method which is based on the difference of potential energies of the ligand in the bound state (in the binding site) and the ligand in the free state (in the solvent). The LIE is obtained by

$$\Delta G_{\text{bind}} \approx \alpha(\langle U^{\text{vdw}} \rangle_b - \langle U^{\text{vdw}} \rangle_f) + \beta(\langle U^{\text{ele}} \rangle_b - \langle U^{\text{ele}} \rangle_f) \quad (3)$$

where U^{vdw} represents the Van der Waals contribution to the potential energy between the ligand and the environment, U^{ele} the electrostatic one and $\langle X \rangle_b$ and $\langle X \rangle_f$ are respectively the averages over the ensemble of configurations with the ligand in the binding site and with the ligand roaming in the solvent. α and β are parameters to balance the contributions to the binding energy: LIE is an end point method, i.e. the binding energy are obtained by considering only averages of the free state and the bound state without considering the docking process. Over the years different values α and β for different kinds of ligands have been found to fit best. The most general values reported in literature, suiting our ligand, are $\alpha = 0.181$ and $\beta = 0.37$.

Unfortunately, due to some bug in GROMACS implementation, we could only use the data from one of the two ligands in both the mutation and the wild type simulation. As a result the estimates, which are rough ones due to a big statistical error on the averages, are plagued by an even bigger error.

Another big source of error in this estimates is due to the lack of clarity in the distinguishing the free state from the bound one. In most of the literature found about LIE, the average over the free state ensemble are obtained by simulating the ligand in a solvated box without the protein. We simulated the ligand in such a condition, but using these values gave unreasonable physical values for ΔG . So instead we used the averages over the configurations of the ligand after it detached from the binding site of the protein and start the float around in the box.

The experimental values where determined from the original article starting from the inhibition constant K_i converted to the binding energy via

$$K_i = \exp\left(\frac{\Delta G}{RT}\right) \quad (4)$$

where $R = 1.985 \times 10^{-3} \frac{\text{kcal}}{\text{mol K}}$ is the universal gas constant and $T = 310 \text{ K}$ is the temperature.

After propagating the errors our estimates and the experimental values are

	Experiment	MD
8DD9	$(-8.12 \pm 0.02) \text{ kcal/mol}$	$(-5.2 \pm 1.9) \text{ kcal/mol}$
6WTT	$(-11.07 \pm 0.05) \text{ kcal/mol}$	$(-10.1 \pm 3.7) \text{ kcal/mol}$

Obviously due to the large uncertainty in our estimates, the value are compatible with the much more precise experimental ones. The huge error one the averages statistical errors and the non-universality of the α and β factor plays the main role in the roughness of the estimation, but nonetheless the LIE methods proves its usefulness in determining which of the two protein will be more likely to bound with the ligand.

4 Conclusions

We conclude with some notes:

- From the protein analysis section we now know that our system is prone to conformational changes, due in particular to the behaviour of the binding sites, which in section (3.3) confirmed to be pretty similar to the ones defined in the reference paper [1].
- In the Ligands and Binding Sites Analysis section (3.4) we saw how the radius of gyration of the binding sites and the number of hydrogen bonds can be used to define a cutoff distance between it and the corresponding ligand. With this cutoff we obtained the subdivision of the trajectory in parts with the ligands in the site and parts outside of the pockets.

With this subdivision we searched for distinct features in conformations of the pockets observing a high mobility of the binding site either in the presence or in the absence of the ligand. This can be considered in accordance with the experimental evidence that the protein shows a low enzymatic activity.

We continued this analysis by observing the ligands RMSD maps, which gave us confidence in affirming that the ligand can actually de-attach and re-attach to the protein, in the binding site and somewhere else, as it was shown in 3.4.

- Finally we confronted our mutation (8DD9) with the wild type (6WTT) with the use of some standard metrics such as the ligand distance to the corresponding binding site, radius of gyration, RMSD of the binding sites and ligand distance to the C_α atoms of the protein. We observed a substantial difference in the mobility of the pockets which are less stable in the mutation and less influenced by the ligand.

We observed a different contribution of the hydrogen bonds to the ligand binding: in both cases hydrogen bonds do not seem to play a crucial role, but the bonds are definitely more frequent in the wild type.

In the end we computed a rough estimation of the binding energy via the linear interaction method: we proved the effectiveness of this simple method, even in the case of very big uncertainties, to establish which one is more likely to have an higher binding affinity.

Finally we list some analysis that one could take further in order to get more refined results:

- A more detailed bond analysis
- A more complex clustering analysis to differentiate more ligand macrostates other than free and bound
- A deeper analysis of the ligands attachment to the other parts of the protein

References

- [1] Yanmei Hu, Eric M. Lewandowski, Haozhou Tan, Xiaoming Zhang, Ryan T. Morgan, Xiujun Zhang, Lian M. C. Jacobs, Shane G. Butler, Maura V. Gongora, John Choy, Xufang Deng, Yu Chen, and Jun Wang. Naturally occurring mutations of sars-cov-2 main protease confer drug resistance to nirmatrelvir. *bioRxiv*, 2022.
- [2] Eko Aditya Rifai, Marc van Dijk, and Daan P. Geerke. Recent developments in linear interaction energy based binding free energy calculations. *Frontiers in Molecular Biosciences*, 7, 2020.