Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

On the quest of fragments as inhibitors for SARS-CoV-2 M^{pro} using fdMD

Guillem Cortiada Rovira[1]

Scientific director: Jaime Rubio Martínez

[1] Modelling of Biological Systems and Drug Design lab, Department of Physical Chemistry, Address Faculty of Chemistry, University of Barcelona, Martí i Franqués 1, 08028 Barcelona, Spain.

Abstract

As we are in the middle of a pandemic, the need arises to prevent the spread of SARS-CoV-2 which causes the disease of Covid-19. The main Protease (M^{pro}) of the novel SARS-CoV-2 is responsible for cutting at different sites and producing polyproteins that aids the production of viral proteins. These viral proteins allow the replication of SARS-CoV-2 and therefore, an increase of their spread. Furthermore, the Main Protease does not have a high similarity to human proteases and that means it is a promising therapeutic target. Recently, many studies have used docking strategies to screen huge databases for testing the potential inhibitors of SARS-CoV-2 M^{pro}. However, docking does not consider the full flexibility of the interacting partners and therefore, it is expected to get better results with the use of molecular dynamics. So, our aim was to apply molecular dynamics using the recently published technique of Fragment Dissolved Molecular Dynamics, an efficient method to locate binding sites of each fragment to decipher the binding site of some selected fragments and to develop scripts to efficiently analyse the results to find the Solvent Accessible Surface Area, the Gibbs Binding Energy, Residence time, Linear Interaction Energy, Interacting residues, etc. Our results showed binding to a previously unknown binding site that can be a possible allosteric site that regulates the catalytic site conformation making an intermediate complex that cannot interact with the activator in the catalytic site and hence, inhibit the SARS-CoV-2 Main Protease. Other fragments tested give us the possibility to link between them when they bind to the SARS-CoV-2 M^{pro} that can induce the formation of a bigger fragment and have a greater inhibition on the Main Protease. The study is performed to provide new insight for investigating the capacity of some fragments such as ebselen to design new drugs against SARS-CoV-2.

Motivation

Nowadays, we are in the middle of a pandemic and it raises the need of developing new drugs against Covid-19. The usage of Molecular Dynamics Simulations has increased to understand molecular interactions between ligand and protein and plays an important role in drug development.

Results

The method is able to reproduce the binding site of the X-Ray studied structures and appears as a promising tool to describe the binding site of new fragments.

Supplementary information: Supplementary data are available at GitLab link: https://github.com/guillem99/FinalDegreeProject.git

1 Introduction

At the end of 2019, a new coronavirus infection caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) spread all over the world. As the coronavirus impacted worldwide, the World Health Organization (WHO) declared it as a global pandemic. Nowadays, the coronavirus disease has caused 176.693.988 global confirmed cases and 3.830.304 global deaths until June 17th, 2021 [1].

Coronaviruses (CoVs) belong to the order *Nidovirales*, family *Coronaviridae.* The *Orthocorovirinae* subfamily contains 4 genera: *Alphacoronavirus, Betacoronavirus, Gammacoronavirus and Deltacoronavirus*. SARS-CoV-2 belongs to the *Betacoronavirus* genus which are enveloped viruses with a positive single-stranded RNA genome. The coronavirus (CoV) genome encodes four major structural proteins: the spike (S) protein, nucleocapsid (N) protein, membrane (M) protein and the envelope (E) protein. These proteins are required for the formation of a structurally complete viral particle and infect host cells. [2]

SARS-CoV-2 is transmitted through body secretions (sweat, stool, urine) mainly by droplet infection of respiratory secretions or with close contact person-to-person (when an infected person coughs, sneezes or speaks). Coronavirus infection affects the respiratory tract and its symptoms have a wide range of features among different people. The main symptoms are a common cold, pneumonia, bronchiolitis, sinusitis, rhinitis, pharyngitis, fever, headache, watery and diarrhea. [3]

When the virus enters the body, it uses a certain human cell receptor called Angiotensin Converting Enzyme 2 (ACE-2) to enter into the human cells. The ACE-2 Receptor is expressed on cells throughout the human body but it's more abundantly expressed in the epithelia of the lung and small intestine, so the main symptoms of coronavirus affect the lungs, nose, blood vessels and intestines. [3, 4] It also produces loss of taste and smell but that is not caused by neurons due that these do not have the ACE-2 Receptor. However, the virus infects cells that have the ACE-2 Receptor that help the neurons to process smells and tastes. [5]

The virus uses the Spike protein to interact with the ACE-2 receptor of the human host cell. When these proteins interact, the virus genome is released into the cytoplasm of the host cells, where it can infect the cell and replicate. The viral genome contains two Open Reading Frames genes (ORF1a and ORF1b), which produce polyproteins (pp1a and pp1b) that are able to take command over the host ribosomes for controlling his translation. These polyproteins can produce one replication transcription complex along with the Main protease of SARS-CoV-2 to be able to replicate. So, the Main Protease, also known as 3C-like proteinase (nsp5) is involved in the replication and transcription process of the SARS-CoV-2. It is able to cut at different sites of the polyproteins together with papain-like protease (nsp3) and produces non-structural proteins[1] (NSP) which in terms aids the production of viral structural proteins. With these viral structural proteins, the SARS-CoV-2 can replicate itself and increase its spread through all human body cells. [4] Therefore, Main Protease becomes an essential viral protein for the viral life cycle and could be a promising therapeutic target.

Until now, three main proteins have been detected that are able to reduce the propagation of SARS-CoV-2. These three proteins are the protein Spike, Main protease and ACE-2 receptor of humans.

Our study is based on the Main protease due that is an essential protein for the virus to replicate itself, it has shown a remarkable degree of conservation of the substrate-binding sites and it has been detected that Human proteases have not high similarity to Main protease. [6]

It is also known the three-dimension (3D) structure of the coronavirus Main protease in an X-ray crystal structure. A study showed that no human proteases have similar cleavage specificity are known [7], so this means that inhibitors of the Main protease are unlikely to be toxic or they have less side effects due that they cannot affect human proteases. [8] The Main protease of SARS-CoV has high similarity to SARS-CoV-2 Main protease. [9]

The structure of Main protease has 2 forms: monomeric or inactive form and homodimeric or active form. When it is in the homodimeric form, it has two equal monomers that each one contains three domains: Domain I (8-99 residues), domain II

[1] **Non-structural proteins:** Are proteins encoded by a virus but that is not part of the viral particle. Typically include various enzymes and Transcription Factors that the virus uses to replicate itself.

(100-183 residues) also called N-terminal domain together with domain I, and domain III (200-306 residues) also called C-terminal domain or extra domain, it contains five α-helices. Between domain II and domain III the Main protease has a long loop (184-199 residues) that does not belong to any domain. The active site of Main protease is located in a cleft in the N-terminal domain between domain I and domain II. It has a Cysteine-Histidine catalytic dyad[2] (Cys145 and His41) that is involved in the nucleophile reaction between Main protease and the active substrate. These residues are very conserved among SARS-CoV and SARS-CoV-2. The first residues of the N-terminal domain produce the N-fingers (1-7 residues) of each monomer. These N-fingers of each monomer are criss-crossed and create interactions between N-terminal and C-terminal of each other monomer. These intramolecular interactions are important for the Main protease to be stable in the active form or in the catalysis conformation.

The substrate binding site is different from the catalytic dyad, it only interacts with the substrate residues but it does not produce any reaction. So, the substrate pocket is given by the side chains of His163, Phe140 and the main chains of Glu166, Asn142, Gly143 and His172. It is located next to the catalytic dyad. There are some residues that produce hydrogen bonds to stabilize the substrate binding site. So, the Main protease in the active form is asymmetric. However, when the dimer is formed, it is symmetric until it adopts the catalytic conformation. [10, 11, 12]

As said before, the Main protease is asymmetric in the catalytic conformation and only one protomer is active at a time. Nevertheless, using the monomeric or dimeric structure when doing MD simulations does not affect the binding affinity of each pocket. [12]

Also, it is known that having a mutation at the catalytic residues such as His41 and Cys145 abolishes almost the complete biological function while the enzyme is a dimer. When instead of studying the catalytic dyad, we study a mutation in the catalytic site (His163), we do not see any difference in the substrate-binding affinity. However, the mutation of Ser139 and Phe140 causes the abolishment of the dimer and therefore, inducing the inactivation. [3, 11] Last but not least, when the interaction between C-terminal and N-terminal is not at enough close proximity, the dimerization can cause defective dimerization that lowers the catalytic efficiency by reducing the substrate-binding affinity.

As said in the article “Activation and maturation of SARS-CoV Main protease” [10], the dimerization is essential for the catalysis due that the loop residues Gly138, Ser139, Phe140 and Leu141 as well as Val125 and Tyr126 of the β-hairpin, all these make direct contact with the other monomer and stabilize the structure leaving free the substrate binding site. Therefore, the dimerization is required to adopt the right catalysis conformation of the active site.

The aim of this study is the assessment of different fragments to design possible drugs to fight against Coronavirus infection. However, as fdMD is a recently developed methodology, we decided to test its efficiency in a particular case of the M^{pro}. Thus, four fragments were extracted from the article “Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease” [13] for our study. Once done with these fragments and looking that our approach was working fine, our decision was to apply the method to another fragment that had been described in the article “Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease” [14]. The authors show that ebselen binds to the M^{pro} in an unknown binding site but suggesting an allosteric action mechanism. It could be a good potential inhibitor.

Finally, as our results suggest an allosteric binding site for ebselen in agreement with the article, we studied pd1 and pd2 fragments without experimental information but they can be synthesized by prof. M. Dolors Pujol from the Faculty of Pharmacy (UB).

In all cases, the same methodology was applied. The methodology applied in this study is the Fragment Dissolved Molecular Dynamics technique developed by our research group. This technique is already published. [15] It uses four molecular dynamics for each ligand studied of 200 ns or 400 ns length. The published method reveals that four dynamics of this length is enough to identify the correct binding sites irrespective of the initial structure when no important reorganization is needed for the binding event.

Another parallel objective of this work is to develop pipeline which includes the execution of some scripts to run the molecular dynamic and then, to run

[2] **Catalytic dyad:** is a set of two amino acids that can be found in the active site of some enzymes. Most commonly found in hydrolases (e.g. proteases) and transferases. Catalytic residues participate in a chemical bond (nucleophilic) that is involved in the donation of an electron pair to produce a reaction.

the programs to analyse the results, for instance, discard the non reactive trajectories, make plots of Gibbs binding energy or distances of the ligand to the center of the molecule and to calculate the descriptors to find which binding site is the best and where it is located.
As the fragments are very small molecules, another usual approach to find compounds with better activity would be to find binding sites close enough to link or grow the fragments following Organic Chemistry rules.
Our decision of using Molecular Dynamics instead of Docking technique was to confirm the poses found instead of only predicting possible binding sites. Molecular Dynamics allows establishing if a ligand will remain enough time interacting with the protein in order to get more reliable results.

2 Objectives

The main objective of this research is to assess the potential of some fragments as possible drugs against Covid-19 infection and identify which fragment could be the best drug candidate using the fragment dissolved molecular dynamics technique (fdMD).

3 Methods

Fragment based drug discovery (FBDD) is a known methodology currently used to accelerate the drug discovery process. However, our research group has designed one method that is able to identify hotspots or binding sites using fragments without any previous knowledge of their binding affinity and without any descriptor of the protein we want to inhibit. [16] This method has been designed to be used in early stages of Fragment Based Drug Discovery. [15] In this study we are going to use fragment dissolved Molecular Dynamics technique to figure out where the fragments bind and then, be able to determine if these fragments can cause the inhibition of the Main protease of SARS-CoV-2. This method, referred to as fragment dissolved Molecular Dynamics (fdMD) consists in performing four 200 ns or 400 ns of Molecular Dynamics simulations of the target molecule in our case Main protease surrounded by many copies of the fragment considered as a possible inhibitor. Finally, to find the best binding site, the method uses a set of descriptors obtained from the analysis of the Molecular Dynamics in each run. These descriptors allow us to identify where the fragment will bind (correct binding site), irrespective of the initial structure used.
The fdMD was applied first to four fragments called frag2, frag3, frag5 and frag6 and then to three more fragments called ebselen, pd1 and pd2.

The steps of fdMD are summarized in: ligand box preparation, protein preparation, molecular dynamics run and finally, analysing each dynamic alone. The fdMD procedure flowchart is shown in Fig. 1.
The first step is to prepare the ligand box and the protein to be able to not get errors during the molecular dynamic simulation. These steps are detailed in section 3.1. Ligand box preparation and Protein preparation section 3.2. Preparing the ligand box means to soak the fragment of interest in an equilibrated box with water molecules using the Leap module of amber18. Once the energy minimization is done and the ligand box and the protein are prepared, the protein is solvated with many ligand boxes to obtain the protein surrounded by many copies of the ligand in water media. Then, four molecular dynamics are done as described in section 3.3. The last step of the fdMD technique is to analyse where each fragment goes in each molecular dynamic and find those that can be considered as the best binding sites. That step is explained in section 3.4. Analyse each Molecular Dynamic.
Finally, one step more was added and it is explained in section 3.5. where with the best considered binding sites, we try to find possible links between fragments to create a bigger molecule and have a greater inhibition of the Main protease.

3.1. Ligand box preparation

Each ligand was prepared with MOE using the Protonate 3D option and optimized with semiempirical Austin Model 1 (AM1). Parameters for the ligands were then obtained using Antechamber with generalized Amber force field (GAFF2) parameters. Charges were obtained with the RESP method. One of the central carbon atom names, sulphur in the ebselen, pd1 and pd2 systems, was changed to C99/S99 for a subsequent application of a repulsion term, needed to avoid ligand aggregation. These parameters were loaded

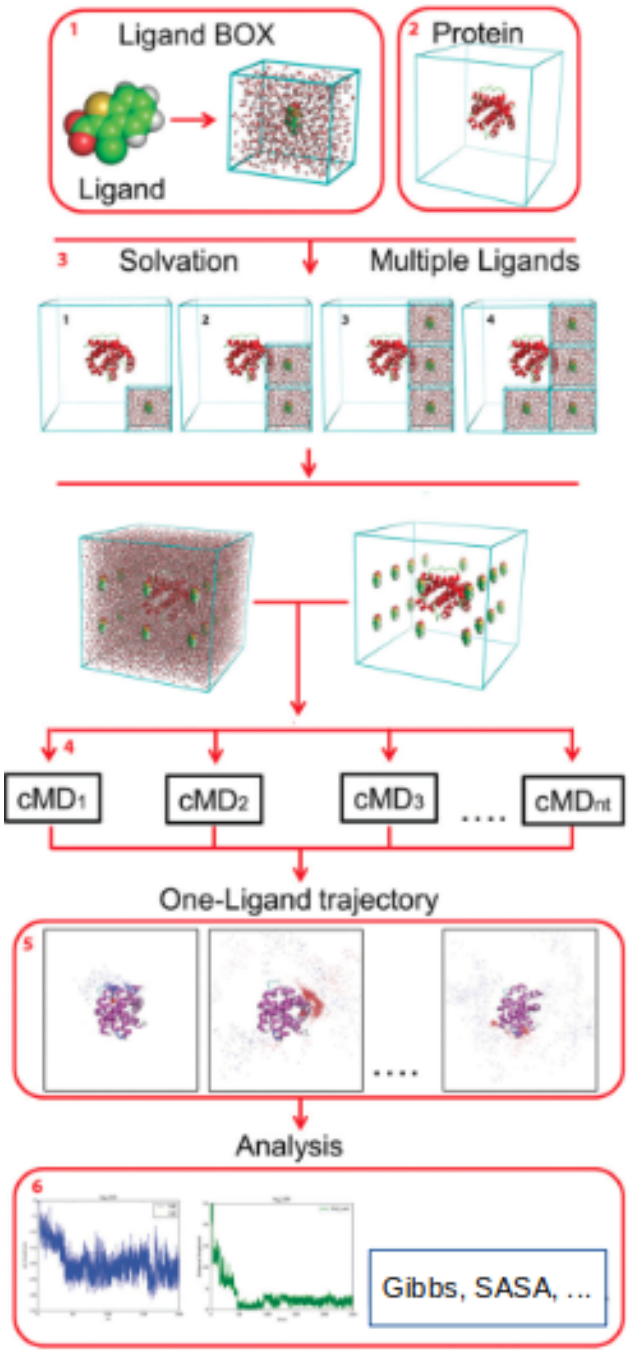

Fig. 1. fdMD flowchart.

into Leap, where counterions were added, if needed, and each ligand was solvated using water molecules. Each box was then subjected to 10.000 steps of Steepest Descent Minimization process before the Molecular Dynamics simulations. This step is done to minimize the Energy in each ligand and to generate a heated and equilibrated box. The last snapshot was the one obtained in the minimization step. It is used as the solvation box.

3.2. Protein preparation

The crystallographic three-dimensional structure of SARS-CoV-2 Main Protease was extracted from Protein Data Bank (PDB code: 6Y84) as a dimeric entity. Subsequently, all residues and organic molecules that do not belong to the Main protease were removed and then, the missing residues were added to the structure. Also, hydrogens were added to all protein residues at their corresponding protonation states at pH 7.0 and side chains orientations were established using Protonate 3D method embedded in MOE. Next step was to place the protein in a cubic box filled with water molecules, setting a minimum distance of 15 Å between the solute and the box walls. Water molecules closer than 1.2 Å to any protein atom were removed. Then, Na^+ and Cl^- ions were added to neutralize the system using the Leap module of Amber18 [17]. All calculations were done using the ff19SB force field [18] with a cutoff of 10 Å for noncovalent interactions, and using the PME method for the treatment of the electrostatic interactions.
Before starting the Molecular Dynamic simulation, the structure has to be firstly relaxed to eliminate possible steric clashes in a multistep energy minimization procedure of 5.000 steps each, using steepest descent method. First, only water molecules and ions are allowed to relax by keeping fixed all the atoms of the protein applying harmonic restrictions of 5kcal/mol*Å^{-2}. In a second step, only the main atoms of the protein were kept fixed. Finally, in a third step all the atoms were allowed to move. This final structure was used for the subsequent calculations.

3.3. Running the Molecular Dynamics

Once prepared the Protein and the ligand, the minimized structure of the Main Protease was used as a starting point for the fdMD simulation. Minimized Main protease was solvated with the corresponding ligand solvation box and equilibrated. When adding the ligand solvation box, any molecule in that box closer than 1.0Å to any protein atom was removed. The number of ligands present in the system depends on the size and shape of the Main protease and each ligand. Therefore, in our study, 12,

20, 20, 34, 34, 34 and 34 ligands for ebselen, pd1, pd2, frag2, frag3, frag4, frag5 and frag6 systems were added respectively.
Furthermore, ParmEd was used to modify the topology in order to use hydrogen mass repartitioning[3] (HMR)[19] and to add a Leonard-Jones repulsion potential, setting to zero the attractive ($B_{C99/S99\text{-}C99/S99} = 0$) part between C99/S99 atoms of the ligands, with e = 0.01 kcal*mol^{-1} and a cut-off distance of 14.0 Å. That Leonard-Jones repulsion term is able to repel if two ligands are close together (less than 14 Å), to avoid collisions/aggregation between fragments and ending all in the same binding site.
After energy minimization and solvation the Main protease with multiple ligands, the system was heated to 300 K stepwisely at a rate of 15 K every 10 ps using the Langevin thermostat algorithm with a collision frequency of 3 ps^{-1} under the canonical ensemble (NVT ensemble). Subsequently, 400 ps simulation was performed at a constant pressure (NPT ensemble) keeping fixed the main atoms of the protein with the same positional restrictions for density equilibration. As each system uses a canonical ensemble (NVT conditions), the system is fixed throughout the simulation: the absolute temperature, the number of atoms and the volume. The system has also a cut-off for non-bonded interactions at 14.0 Å. Finally, a conventional molecular dynamics (cMD) simulation of 400 ns length was carried out for each fragment. Each conventional molecular dynamic (cMD) is divided in 2 steps (1-2), in each one we have and integration time of 0.004ps and 50.000.000 steps:

$$0.004\ ps\ x\ 50.000.000\ steps\ =\ 200.000\ ps\ =\ 200\ ns$$

As we observe, running 200 ns in each step means to have 400 ns of MD simulation.
Once all the protocol was followed in detail, the execution of four independent molecular dynamic simulations for each fragment was done using different initial velocities in each one. Then, the next step was to select the reactive trajectories.

[3] **Hydrogen mass repartitioning:** is a potentially useful tool for accelerating MD simulations. By repartitioning the mass of the heavy atoms into the bonded hydrogen atoms, it is possible to slow the highest frequency motions of the macromolecule under study, thus allowing the time step of the simulation to be increased by up to a factor of 2.

3.4. Analysing each Molecular Dynamic

Once executed the simulations, the next step is to process each cMD with each ligand alone to find these reactive ligands and the good binding sites. The steps followed are summarized in calculating the descriptors for each trajectory (section 3.4.1.), discard the non-reactive trajectories and calculate the MMGBSA Energy of binding for the reactive trajectories (section 3.4.2.) and finally, create some tables with these reactive ligands and get the most probable binding to the Main protease (section 3.4.3.).

3.4.1. Calculate the descriptors for each trajectory

This step involves the calculation of LIE Energy (Linear Interaction Energy) that is an estimation of the binding free energy obtained from the change in Electrostatic Energy and Van der Waals interaction energy between the ligand and the formed complex, the RMSD[4] (Root-Mean-Square-Deviation) for each ligand alone, the percentage of Solvent-Accessible Surface Area[5] along the last 10% of the MD simulation, the number of hydrogen bonds formed during the last 10% of the trajectory, the distance between the center of mass of the M^{pro} and the ligand and finally, the RMSD of the ligand against the pocket it binds to with the pocket superposed.

3.4.2. Discard the Non-Reactive Trajectories

With the descriptors calculated in the previous step, it allow us to find the reactive trajectories[6].
<u>These reactive trajectories need to follow:</u>

- In the last snapshot, the ligand has at least one atom at less than 5 Å of the M^{pro}.
- The distance between a ligand and the M^{pro} is at less than 5 Å during the last 10% of the Molecular Dynamic simulation.

If some point is not fulfilled, the trajectory is considered as non reactive.

[4] **Rmsd** is a measurement of the average distance between atoms (usually the backbone atoms) of superimposed structures.
[5] **Solvent-Accessible Surface Area:** is the surface area of a molecule that is accessible to a solvent.
[6] **Reactive trajectory:** is called to one ligand trajectory when its track during the MD simulation ends up interacting with the protein of interest (e.g. Main protease).

These reactive trajectories are submitted to calculate the MMGBSA Energy. Amber18 [17] was used to calculate the MMGBSA Energy that is an approach that employs molecular mechanics (MM), the generalized Born model (GB) and solvent accessible surface area (SASA) method to estimate the relative binding affinity for a list of ligands. It is reported in kcal/mol and is an approximation of the free energies of binding in which a more negative value indicates a stronger binding.
Finally, some plots are produced with all the data generated until now.

3.4.3. Get the most probable Binding Site

Once all the descriptors are calculated for the reactive ligands, the next step is to produce one table for each of the four MD runs in each fragment. These tables are created to classify if one binding site can be considered equal as another one among different cMD of one fragment. The protein residues that are involved in the interaction between ligand and M^{pro} are used to classify to which binding site it belongs. For each cMD one table is produced and these tables includes the following information: Simulation time, Fragment name, Molecule, Ligand number in the MD, average of LIE Energy, average of Binding Energy, Residence time, average of Solvent-Accessible Surface Area, number of hydrogen bonds, the protein residues involved in the interaction of the ligand-protein and the Binding site classification among cMDs of the same fragment.
Finally, an average table was done for each fragment. This average table is calculated using the data from the four cMD tables of each fragment. Mainly, it is taking care of the number of times a reactive binding site is found along the four cMD of each fragment. Therefore, the final average table includes all the binding sites detected for each fragment, the number of times the binding site appears in the four MD and the average and best case of all the descriptors of that binding site. For instance, binding site 1 can be found in the four run cMD of pd1 and binding site 2 can be found only in one cMD of pd1. So, the best MMGBSA of binding site 1 of pd1 means to compare the binding site 1 in the four cMD of pd1 and find the best MMGBSA and the average MMGBSA of pd1 means to sum all the MMGBSA trajectories where the binding site 1 is found divided by four that is the total number of trajectories. That process is repeated for all the used descriptors.

Once the average table is constructed you can observe which binding site has more reactive trajectories and which has the best value of each descriptor. In conclusion, when we use this average table, we are able to obtain which is the best binding site for each fragment.
The selection of the most probable Binding site follows these steps:

- Select the best descriptor in each case.
- The Binding site that has more selected descriptors is considered the best Binding site.
- If there is a match in the number of selected descriptors, our algorithm is not able to select which is the best binding.

3.5. Analysis of Binding sites detected

Once decided which is the most probable binding site, the next step is to detect possible links between fragments. The aim of this step is to detect binding sites that are close enough to each other (among 6-10Å) to be able to link two fragments and create a bigger molecule that could have a greater inhibition for the M^{pro}. Using some visualization tools such as Chimera [20], allow us to superpose the different structures and compare their ligands.
Our analysis consists of superimposing the last snapshots of the more probable binding sites and measuring the distance between the two fragments in each case.

4 Results and Discussion

As the article called “The SARS-CoV-2 Main protease as a drug target” says, the M^{pro} does not have a great similarity to human proteases. [8] To check that, we have extracted the sequence of amino acids of the M^{pro} and run a blastp [21] to only human sequences and the results show that any human amino acid sequence is similar to the M^{pro} of SARS-CoV-2.
On the other hand, the results obtained from the molecular dynamics simulations include one file for each step of 200 ns. These files include one snapshots every 20 ps along the trajectory.
Once all the pipeline is executed, two tables are obtained to analyse each fragment alone.
First of all, we have executed our approach with the experimental structures and then, when knowing how the descriptors are working with that

experimental data, we have applied the methodology with other fragments.

As we detected that the methodology was working well, we have used the same methodology but using the ligands with unknown experimental data.

As observed in Table 1, where we observe the data from the experimental frag5, the binding site 4 seems to be the best one due to the fact that it has a huge residence time in all molecular dynamics and a good binding energy. However, it does not have hydrogen bonds during at least 50% of the last 80 ns in all the reactive ligands. Indeed, binding site 4 is the one located in the catalytic site and therefore, the experimental one.

Also, in Table 2 of the experimental frag5 is shown that the best selected binding site across all the MD should be Binding site 4 or the experimental one because it has better descriptors or more bold descriptors. It shows a low Solvent-Accessible Surface Area, so it's probably more attached to the M^{pro} than other ligands.

Binding Site 1 has the best LIE Energy and seems to have lower values of binding energy and a good residence time. But it does not have any other best descriptor. Therefore, it needs to be ignored when selecting the best pocket. The same happens with Binding site 2, 5 and 7 with the hydrogen bonds.

Binding site 3 seems to be the worst one because it is only found in one molecular dynamic and it does not show a high affinity to bind to the M^{pro}. Also, the residence time seems to be very short and the SASA value is very high, so that means it is not very attached to M^{pro}.

The same analysis was applied in those fragments whose experimental structure was obtained from the article “Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease”. [13]

Our analysis showed that frag2 and frag3 bind to the experimental site in the second best selected pocket and the frag5 and frag6 bind to the experimental site in the first best selected pocket.

Instead of using all the descriptors, we have used the ones that seem to point out the experimental results. These descriptors are the number of reactive trajectories, best and average binding energy and best and average residence time for each pocket.

Table 3 shows the summary table for frag2 where it is shown that the first pocket is the best selected one. However, the first pocket is not the same as the experimental one. Therefore, if instead of selecting the best descriptor in each case, we select the two best descriptors in each case, the binding site 4 or the experimental one is selected as one of the two best selected pockets.

Table 1. All Descriptors of each Molecular Dynamic and each Reactive ligand of fragment frag5 in complex with M^{pro} in a simulation time of 800 ns.

MD	Binding Site	LIE (kcal/mol)	Delta G (kcal/mol)	RT (ns)	% SASA ligand	Hydrogen Bonds
1	BS1	-61.69	-26.39	90	36.62	1
	BS2	-36.56	-16.86	270	57.58	1
	BS3	-31.16	-14.92	160	50.54	0
	BS4	-61.6	-23.78	700	48.31	1
	BS4	-56.21	-28.91	580	33.09	0
	BS5	-33.87	-19.2	120	57.01	2
2	BS6	-30.38	-19.93	330	47.36	1
	BS2	-36.2	-16.86	440	56.65	1
	BS2	-40.27	-20.39	370	43.18	1
	BS7	-56.06	-19.07	220	61.92	2
3	BS5	-33.46	-18.86	240	60.42	1
	BS2	-35.78	-16.33	250	56.53	1
	BS8	-42.94	-20.46	310	34.53	0
	BS9	-57.96	-27.61	390	22.37	1
	BS4	-35.49	-17.3	90	55.26	0
	BS10	-43.81	-20.47	130	38.41	1
	BS11	-34.96	-16.93	90	50.7	0
	BS2	-43.21	-20.91	520	27.91	0
	BS12	-41.64	-16.72	210	52.48	0
	BS4	-44.79	-22.86	290	40.71	0
	BS1	-59.75	-22.79	420	40.6	1
4	BS4	-42.65	-22.39	490	23.41	0
	BS12	-48.58	-21.45	400	36.2	1
	BS10	-38.76	-20.01	340	45.94	1
	BS12	-44.63	-22.39	490	45.19	1
	BS1	-49.64	-21.94	80	42.36	1
	BS13	-32.39	-15.21	770	51.59	0
	BS14	-35.53	-21.17	720	24.5	0

This table shows the descriptors of each reactive ligand of the fragment frag5. Definition of Table 1: LIE = Average of Linear Interaction Energy, DeltaG= Average of the Binding Energy between Ligand and Receptor, % SASA= Difference between SASA Complex and SASA Receptor divided by the SASA of the Ligand, Hydrogen Bonds = Number of Hydrogen Bonds formed during the last 10% of the trajectory.

So, we can rely on our method to find which is the most probable binding site for the other three fragments that have not been crystallized yet.

As seen in these fragments, frag2, frag3 and frag5 could be used as potential catalytic site inhibitors for SARS-CoV-2 as they show a high affinity to bind to the active site of the M^{pro}.

Furthermore, following the same criteria, we applied this method to the fragments with unknown experimental structure. These fragments are called pd1, pd2 and ebselen. Our results show where each fragment is more likely to interact with the M^{pro}. The obtained results show that these ligands will not bind to the active site nor the catalytic pocket. So, we cannot report that they will inhibit the M^{pro}. Some more studies are needed to show that they could be an allosteric inhibitor.

Table 2. Average Table of the binding sites for fragment frag5 in a simulation time of 800 ns.

Binding Site	Reactive Trajectories	Best Delta G	Avg. Delta G	Best % SASA	Best RT	Avg. RT	Best LIE	Avg. LIE	Best HB	Avg. HB
BS1	3	-26.39	-17.78	36.62	420	147.5	**-62.69**	-43.02	1	0.75
BS2	5	-20.91	-22.84	27.91	520	462.5	-43.21	-48.01	1	**1**
BS3	1	-14.92	-3.73	50.54	160	40	-31.16	-7.79	0	0
BS4/ Exp.	**5**	**-28.91**	**-28.81**	23.41	700	**537.5**	-61.6	**-60.185**	1	0.25
BS5	2	-19.2	-9.52	57.71	240	90	-33.87	-16.83	**2**	0.75
BS6	1	-19.93	-4.98	47.36	330	82.5	-30.38	-7.60	1	0.25
BS7	1	-19.07	-4.76	61.92	220	55	-56.06	-14.02	**2**	0.5
BS8	1	-20.46	-5.12	34.53	310	77.5	-42.94	-10.74	0	0
BS9	1	-27.61	-6.91	**22.37**	310	77.5	-42.94	-10.74	0	0
BS10	2	-20.47	-10.12	38.41	340	117.5	-43.81	-20.64	1	0.5
BS11	1	-16.93	-4.23	50.7	90	22.5	-34.96	-8.74	0	0
BS12	3	-22.39	-15.14	36.2	490	275	-48.58	-33.72	1	0.5
BS13	1	-15.21	-3.80	51.59	**770**	192.5	-32.39	-8.10	0	0
BS14	1	-21.17	-5.29	24.5	720	180	-35.53	-8.89	0	0

This table shows the average table of each binding site for the fragment frag5. Definition of Table 2: Best values are the best value of each descriptor in each BS. Average values are the sum of all the values of each BS divided by 4. Best and Average is explained in the methodology section 3.4.3. (bold descriptors are the best ones).

For instance, a Quantitative Structure-Activity Relationship (QSAR) needs to be done before saying that these fragments could have an allosteric inhibitory response.

Table 3. Average Table of the binding sites for fragment frag2 in a simulation time of 800 ns.

Binding Site	Reactive Trajectories	Best DeltaG	Avg. DeltaG	Best RT	Avg. RT
BS1	**2**	**-18.02**	**-8.65**	230	**107.5**
BS2	**2**	-17.70	**-8.14**	**310**	102.5
BS3	1	-12.97	-3.25	300	65
BS4/ Exp	**2**	**-17.73**	-7.50	**340**	**105**
BS5	**2**	-16.03	-7.31	160	67.5
BS6	1	-12.93	-3.24	80	20
BS7	1	-13.03	-3.36	120	30

This table shows the average table of each binding site for the fragment frag2. Definition of Table 3: Best values are the best value of each descriptor in each BS. Average values are the sum of all the values of each BS divided by 4. Best and Average is explained in the methodology section 3.4.3. (bold descriptors are the best ones).

When observing the pd1 summary table in Supplementary Table 1 and the pd1 average table in Table 4 we see that the Binding site 1 and 3 seem to be the best ones due to the fact that they have low values of binding energy and a large residence time. There is one case that remains almost all the time bound to the third pocket. We do not observe any hydrogen bond in those pockets but as we have discarded this descriptor for detecting the most probable binding site, it is not remarkable.

Selecting the best pocket for the pd1 fragment was done with only the descriptors pointing out the experimental data.

These descriptors are the number of reactive trajectories, best and average binding energy and residence time. So, the selection of the two best values for each descriptor give us the overall selection of two pockets. These pockets are the binding site 1 and binding site 3.

Binding Site 3 of pd1 has the following protein residues involved in the interaction between the ligand and M^{pro}: Q107, P108, G109, Q110, T111, I200, V202, N203, D245, H246, I249, T292, P293 and F294. The ligand binds to three α-helix and one loop. This pocket is the same as the frag6 that goes to an allosteric site. An important point is that the catalytic site cleft seems to not have a huge conformational change, so probably the substrate will be able to perform the catalytic reaction.

As we can observe in the Supplementary Material, the Supplementary Table 2 shows that the pd2 fragment could bind to the BS9 and BS10 and the Supplementary Table 3 shows that the ebselen fragment could bind to BS1 and BS6. So, if we superpose these ligands we observe that the BS3 of pd1, BS9 of pd2 and BS1 of ebselen are going to the same pocket, the same as frag6. This suggests to us that it can be a possible allosteric site.

Table 4. Average Table of the binding sites for fragment pd1 in a simulation time of 1000 ns.

Binding Site	Reactive Trajectories	Best DeltaG	Avg. DeltaG	Best RT	Avg. RT
BS1	**4**	-20.83	**-20.1675**	**820**	**702.5**
BS2	1	-15.41	-3.86	120	30
BS3	**4**	-21.6	**-20.70**	**990**	**720**
BS4	**2**	**-23.44**	-10.42	430	177.5
BS5	1	-11.73	-2.94	340	58
BS6	1	**-23.14**	-5.79	110	27.5
BS7	1	-22.85	-5.72	700	175
BS8	1	-19.14	-4.79	260	65
BS9	1	-13.04	-3.26	150	37.5

This site is located in the first monomer and we observe in Fig. 2D the possible allosteric site where pd1, pd2 and ebselen fragments tend to bind. Other pockets selected for pd1, pd2 and ebselen are also shown in Fig. 2. Pockets in Fig. 2 A, B and C are bound to different sites, but each one has a different binding energy.

We have tried to link some fragments to create a bigger one and have a greater inhibition. When we tried to create possible links between fragments to create a bigger molecule and have a greater

inhibition, we obtained that the binding site 10 and the binding site 4 of pd2 could bind among them.

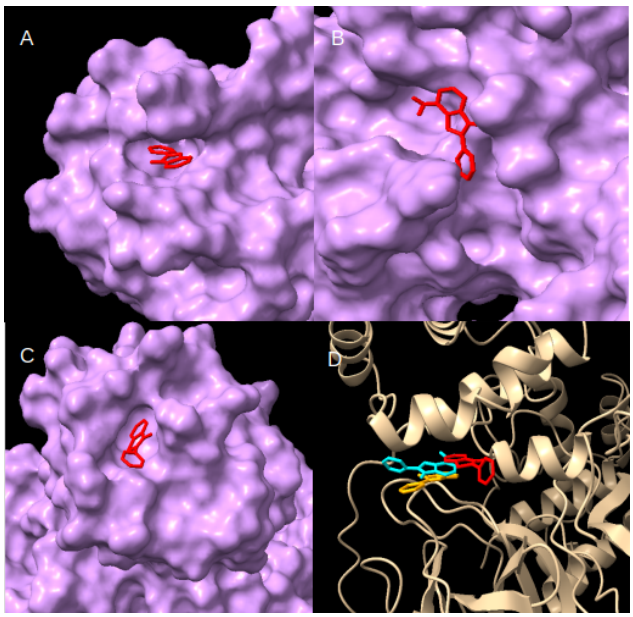

Fig. 2. Pockets where pd1, pd2 and ebselen tend to bind. A. This figure shows the BS1 of pd1 that is the second best pocket for pd1. **B.** This figure shows the BS10 of the first selected pocket for pd2. **C.** This figure shows the BS6 that is the second selected for ebselen. **D.** This figure shows one of the two selected pockets for pd1, pd2 and ebselen. The red ligand is the ebselen, cyan one is the pd2 and the orange one is pd1.

That is shown in Fig. 3. We can observe that it is a bigger fragment and probably will remain bound to M^{pro} more time due to the fact that the structure needs to be more open to let out the fragment. So, as it is more difficult for the fragment to leave the pocket, it would probably have a greater residence time and a greater allosteric inhibition. Also, as BS10 of pd2 is the same as frag6, we can consider that it probably has an allosteric inhibitory response.

As we have done all the analysis in 400, 600 and 800 ns, we have seen that the ligands seem to not remain bound all the time. So, we have created a table with the reactive ligands in 400, 600 and 80 ns. This table is shown in Table 5. From this table we can conclude that there are some fragments that do not remain in the experimental site or in the same binding site for a long time. For instance, Experimental site in the second MD only remains during the 400 and 600 ns. Between 600 and 800 ns, this ligand has left the pocket and therefore, we decide to check if the ligands always come and go or they remain in the same pocket. That was checked running one MD of 1000 ns with the experimental structure. Our results show that the ligand does not remain in the experimental site along all the trajectory. Finally, we observe the ligand bound to another pocket. So, our conclusion is that these ligands do not bind with enough energy or they are not creating enough hydrogen bonds or intermolecular interactions to remain bound to the experimental site.

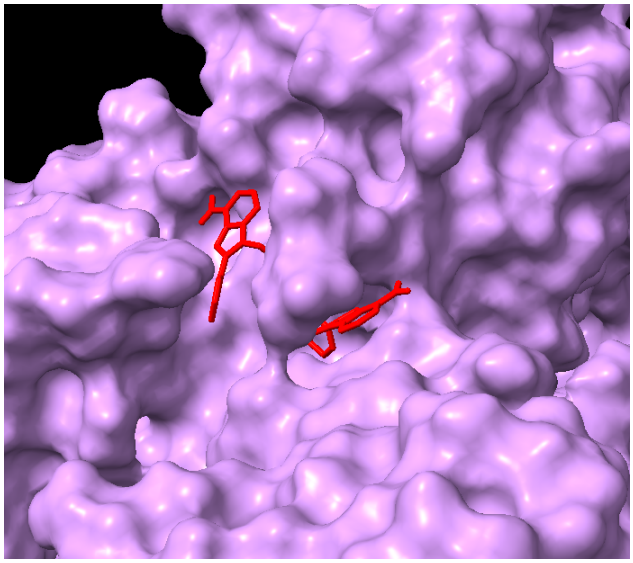

Fig. 3. Link among BS4 and BS10 of pd2. This figure shows in red the two ligands located at their respective sites (BS4 and BS10) and one link created between them. Binding site 10 is the one on the left and binding site 4 is the one on the right

.

Table 5. Reactive ligands in 400, 600 and 800 ns of frag2.

400 ns											
Dyn_1	**Delta G**	**RT**	**Dyn_2**	**Delta G**	**RT**	**Dyn_3**	**Delta G**	**RT**	**Dyn_4**	**Delta G**	**RT**
Exp	-14.33	40	Exp	-21.12	110	BS6	-10.51	50	BS7	-15.31	80
BS8	-15.06	380	BS3	-11.86	70				BS3	-11.22	50
			BS4	-14.66	120				Exp	-14.41	120
			BS5	-12.46	70						
600 ns											
Dyn_1	**Delta G**	**RT**	**Dyn_2**	**Delta G**	**RT**	**Dyn_3**	**Delta G**	**RT**	**Dyn_4**	**Delta G**	**RT**
BS9	-14.93	150	Exp	-21.14	310	BS9	-14.1	160	BS10	-11.31	60
			BS4	-13.1	320	BS8	-13.22	100			
						Exp	-12.73	140			
800 ns											
Dyn_1	**Delta G**	**RT**	**Dyn_2**	**Delta G**	**RT**	**Dyn_3**	**Delta G**	**RT**	**Dyn_4**	**Delta G**	**RT**
BS2	-18.02	230	BS2	-16.57	200	BS8	-12.97	300	BS11	-13.22	110
BS9	-17.70	100	BS9	-14.83	310	Exp	-12.26	340	BS11	-16.03	160
						Exp	-17.73	80	BS12	-12.93	80
									BS3	-13.03	120

5 Conclusion

The first step was to test our methodology with some M^{pro} X-ray structures. Then, when looking for potential inhibitors, we have seen that pd1, pd2 and ebselen are going to bind to an allosteric site but we need a Quantitative Structure-Activity Relationship (QSAR) to show that the ligands could have an allosteric effect.

Also, we have observed that the Fragment Dissolved Molecular Dynamics technique is a very useful technique to detect the best pocket in each fragment studied. Instead of using 200 ns of MD simulation, we have used 1000 ns. Therefore, we expected to get more precise and reliable results. In fact, we were expecting that some fragments would bind to the active site of the Main protease and they could be considered as potential inhibitors. Although we cannot show that, we can conclude that pd1, pd2 and ebselen will not bind to the active site and they can only be an allosteric inhibitor of the selected pockets.

Supplementary Material

Supplementary Material is available at https://drive.google.com/file/d/18XLDsN6rQizSoPfjLxB2AWn8RY4PU2RO/view?usp=sharing .

References

[1] WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/ . Accessed on June 17th, 2021.

[2] Schoeman, D., & Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. *Virology journal*, *16*(1), 1-22.

[3] Prajapat, M., Sarma, P., Shekhar, N., Avti, P., Sinha, S., Kaur, H., ... & Medhi, B. (2020). Drug targets for corona virus: A systematic review. *Indian journal of pharmacology*, *52*(1), 56.

[4] Hamming, I., Timens, W., Bulthuis, M. L. C., Lely, A. T., Navis, G. V., & van Goor, H. (2004). Tissue distribution of ACE2 protein, the functional receptor for SARS coronavirus. A first step in understanding SARS pathogenesis. *The Journal of Pathology: A Journal of the*

Pathological Society of Great Britain and Ireland, *203*(2), 631-637.

[5] Brann, D. H., Tsukahara, T., Weinreb, C., Lipovsek, M., Van den Berge, K., Gong, B., ... & Datta, S. R. (2020). Non-neuronal expression of SARS-CoV-2 entry genes in the olfactory system suggests mechanisms underlying COVID-19-associated anosmia. *Science advances*, *6*(31), eabc5801.

[6] Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., & Hilgenfeld, R. (2003). Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science*, *300*(5626), 1763-1767.

[7] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., ... & Hilgenfeld, R. (2020). Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, *368*(6489), 409-412.

[8] Ullrich, S., & Nitsche, C. (2020). The SARS-CoV-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, 127377.

[9] Amin, S. A., Banerjee, S., Singh, S., Qureshi, I. A., Gayen, S., & Jha, T. (2021). First structure–activity relationship analysis of SARS-CoV-2 virus main protease (Mpro) inhibitors: an endeavor on COVID-19 drug discovery. *Molecular diversity*, 1-12.

[10] Xia, B., & Kang, X. (2011). Activation and maturation of SARS-CoV main protease. *Protein & cell*, *2*(4), 282-290.

[11] Chang, H. P., Chou, C. Y., & Chang, G. G. (2007). Reversible unfolding of the severe acute respiratory syndrome coronavirus main protease in guanidinium chloride. *Biophysical journal*, *92*(4), 1374-1383.

[12] Tam, N. M., Nam, P. C., Quang, D. T., Tung, N. T., Vu, V. V., & Ngo, S. T. (2021). Binding of inhibitors to the monomeric and dimeric SARS-CoV-2 Mpro. *RSC Advances*, *11*(5), 2926-2934.

[13] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., ... & Walsh, M. A. (2020). Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature communications*, *11*(1), 1-11.

[14] Menéndez, C. A., Byléhn, F., Perez-Lemus, G. R., Alvarado, W., & de Pablo, J. J. (2020). Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease. *Science advances*, *6*(37), eabd0345.

[15] Privat, C., Granadino-Roldán, J. M., Bonet, J., Tomas, M. S., Perez, J. J., & Rubio-Martinez, J. (2021). Fragment dissolved molecular dynamics: a systematic and efficient method to locate binding sites. *Physical Chemistry Chemical Physics*, *23*(4), 3123-3134.

[16] Erlanson, D. A., Fesik, S. W., Hubbard, R. E., Jahnke, W., & Jhoti, H. (2016). Twenty years on: the impact of fragments on drug discovery. *Nature reviews Drug discovery*, *15*(9), 605.

[17] Case, D. A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham III, T. E., Cruzeiro, V. W. D., ... & Kollman, K. A. (2018). AMBER 2018; 2018. *University of California, San Francisco*.

[18] Tian, C., Kasavajhala, K., Belfon, K. A., Raguette, L., Huang, H., Migues, A. N., ... & Simmerling, C. (2019). ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation*, *16*(1), 528-552.

[19] Hopkins, C. W., Le Grand, S., Walker, R. C., & Roitberg, A. E. (2015). Long-time-step molecular dynamics through hydrogen mass repartitioning. *Journal of chemical theory and computation*, *11*(4), 1864-1874.

[20] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*(13), 1605-1612.

[21] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, *25*(17), 3389-3402.