

# Potential drug candidates for SARS-CoV-2 using computational screening and enhanced sampling methods

Sadanandam Namsani<sup>1§</sup>, Debabrata Pramanik<sup>2,§</sup>, Mohd Aamir Khan<sup>1,2</sup>, Sudip Roy<sup>\*1</sup> and Jayant Kumar Singh<sup>\*1,2</sup>

<sup>1</sup>Prescience Insilico Private Limited  
Old Madras Road, Bangalore 560049, India

<sup>2</sup>Department of Chemical Engineering  
Indian Institute of Technology, Kanpur, India

<sup>§</sup> Same contribution

\*Corresponding authors: Sudip Roy (sudip@prescience.in) and Jayant Kumar Singh (jayantks@iitk.ac.in)

## Abstract

Here, we report new chemical entities that exhibit highly specific binding to the 3-chymotrypsin-like cysteine protease (3CLpro) present in the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Because the viral 3CLpro protein controls coronavirus replication, 3CLpro is identified as a target for drug molecules. We implemented an enhanced sampling method in combination with molecular dynamics and docking to reduce the computational screening search space to four molecules that could be synthesized and tested against SARS-CoV-2. Our computational method is much more robust than any other method available for drug screening (e.g., docking) because of sampling of the free energy surface of the binding site of the protein (including the ligand) and use of explicit solvent. We have considered all possible interactions between all the atoms present in the protein, ligands, and water. Using high-performance computing with graphical processing units, we were able to perform a large number of simulations within a month and converge the results to the four most strongly bound ligands (based on free energy and other scores) from a set of 17 ligands with lower docking scores. Additionally, we have considered N3 and 13b  $\alpha$ -ketoamide inhibitors as controls for which experimental crystal structures are available. Out of the top four ligands, PI-06 was found to have a higher screening score compared to the controls. Based on our results and analysis, we confidently claim that we have identified four potential ligands, out of which one ligand is the best choice based on free energy and the most promising candidate for further synthesis and testing against SARS-CoV-2.

## 1. Introduction

The current situation of the world is extraordinary due to the coronavirus disease 2019 (COVID-19) pandemic. COVID-19 is caused by a new pathogen, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus (Coronaviridae: *Betacoronavirus*)<sup>1,2</sup>. Infection with the new pathogenic SARS-CoV-2 can result in long-term reduction in lung function, arrhythmia, and death. This virus is found to have much stronger binding energy with the host cell than its predecessors and thus spreads more efficiently. Given the novelty of this virus and the global pandemic it has caused, there is an urgent need for drug candidates and vaccines to be developed as soon as possible. The current crisis is mainly due to the lack of any specific antiviral drugs that could function against SARS-CoV-2 as well as a lack of preparedness to find and produce a new vaccine. To mitigate the risks posed by viruses, including SARS-CoV-2, it is imperative that research efforts for the development of new antiviral agents targeting this virus be pursued with renewed energy.

The scientific world is responding to the COVID-19 pandemic via three major innovation paths. The most focused research currently in progress is the development of vaccines and clinical trials of existing FDA-approved drugs for other relevant diseases and developing new chemical entities (NCEs). As the repurposing of drugs may fail, developing NCEs is also essential.

The goal of NCEs research is to find the target proteins, i.e., those proteins that are envisaged as moderators of functions that help the virus propagate in the human body. NCEs are designed to inhibit these proteins either on the viruses themselves or in human cells, in order to stop the biological pathways and control the disease. The *ab initio* design of NCEs begins with mining the knowledge related to the chemical space from large sets of compounds available in chemical databases. This virtually unlimited chemical data is used to develop NCEs, and the developed NCEs are then subjected to high-throughput computational screening<sup>3</sup> to identify those entities that may have a therapeutic effect on the coronavirus. The main protease (Mpro 3CLpro) of coronavirus is an attractive drug target because of its function in processing the polyproteins that are translated from the viral RNA. Mpro is a key CoV enzyme that mediates viral replication and transcription. Mpro has cleavage-site specificity similar to that of picornavirus 3C protease (3C pro). Therefore, it is also known as 3C or 3C-like main protease (3CL Mpro). Jin *et al.* recently reported X-ray structures<sup>4</sup> of the SARS-CoV-2 Mpro and its complex with an N3 inhibitor. The crystal structure of SARS-CoV-2 main protease in complex with an inhibitor N3 was reported<sup>4</sup> at RCSB Protein Data Bank as entry 6LU7. Liu *et al.*, in a subsequent publication, predicted a list of commercial medicines that might work as inhibitors<sup>5</sup> of 2019-nCoV. Zhang *et al.*<sup>6</sup> recently reported an X-ray crystal structure complexed with 13b  $\alpha$ -ketoamide (PDB ID: 6Y2G). The 13-b ketoamide exhibits increased solubility in plasma and reduced binding with plasma proteins compared to other peptidomimetic  $\alpha$ -ketoamides<sup>7</sup>. It has also been shown that the 13-b  $\alpha$ -ketoamide enhanced antiviral activity compared to other ketoamides. Walls *et al.* recently showed that SARS-CoV-2 S uses ACE2, a membrane-associated protein, to enter human cells<sup>8</sup>. The SARS-CoV-2 S and SARS-CoV S (SARS coronavirus identified in 2003) binding domains were found to exhibit similar affinities<sup>6</sup> to bind with the receptor, ACE2 human protein. They found that the SARS-CoV-2 S glycoprotein uses a furin cleavage site at the boundary between the S1/S2 subunits, which is processed during biogenesis, and this sets the virus SARS-CoV-2 apart from SARS-CoV and SARS-related CoVs. They reported a cryo-EM structure of the SARS-CoV-2 S ectodomain trimer, which is another hotspot for designing vaccines and inhibitors.

Bung *et al.* recently published<sup>9</sup> their initial work on *de novo* design of NCEs for SARS-CoV-2, targeting the 3CLpro protein using deep neural-network (DNL)-based generative and predictive methods for in silico design of NCEs. They trained the DNL model using ~1.6 million small molecules available in

the ChEMBL database<sup>10</sup>. Subsequently, based on various physicochemical properties such as drug likeness<sup>11</sup> and synthetic accessibility<sup>12</sup> to filter the designed NCEs. Finally, these filtered small molecules were screened and ranked using the virtual screening scores obtained by docking with 3CLpro using AutoDock Vina<sup>13</sup>. They docked 3,960 molecules and reported a final set of 31 high-potential<sup>9</sup> (to qualify as drug candidates) NCE molecules with virtual screening scores between -8.3 and -7.5. The highest virtual screening score they reported was -9.1 and the highest Tanimoto coefficient<sup>14</sup> with the existing protease inhibitors among the top 15 molecules was 0.90.

While virtual screening tools have become popular, the limitations of methods such as docking with different variations in methodologies are well-established in the literature<sup>15,16</sup>. The target ligand docking often fails to produce or identify the correct ligands (NCEs) with high specificity towards the target. Failure is caused by multiple factors, such as lack of proper sampling of the binding site, degree of flexibility of the protein (change of conformation of protein due to binding of the ligand), nonexistence of solvent (i.e., the solvent-ligand interaction) in the model, and finally, results in the definition of the scoring function with missing entropic contributions. Therefore, docking is often used for qualitative estimation of the chemical space of the target, and, subsequently, medicinal chemists intuitively design scaffolds and synthesize large numbers of molecules to form a chemical library. This library of molecules is then tested in biological assays to further screen for better efficacy.

With the current challenges in mind, we propose a robust methodology to address some of the issues mentioned above and reduce the subset of NCEs for further synthesis and testing. Our methodology is to identify the set of ligands with high specificity using a large-scale all-atom molecular simulation followed by enhanced free-energy methods on the target-ligand complexes. In this method, we used molecular docking to select a set of chemical entities that showed significant interaction (high score) with the protein. These molecules are then subjected to molecular dynamics simulations with water as the explicit solvent. Solvation of molecules in water (or any other solvent) is critical for identifying the ligands that could bind at the binding site of the protein with high stability and not become solvated in water. Molecular dynamics (MD) simulations are therefore used as an additional filter to identify ligands with high stability. The stable structures (i.e., protein-ligand bond state in water) identified from the MD simulations were further used for enhanced free-energy sampling. Since entropy plays an important role in the specificity of binding, quantitative estimation of free energy is essential for better comparative binding specificity among various ligands interacting with proteins. In addition, defining a score associated with the binding of these ligands (chemical entities) is important to choose the best set of NCEs as potential drug candidates.

In this work, we have considered the final molecules reported by Bung *et al.*<sup>7</sup> for enhanced sampling using the methodology described above. We have selected molecules that showed higher binding affinities toward the protein. In addition, we have also considered a few molecules that are very similar to darunavir (Tanimoto similarity of 0.91 and 0.90). Darunavir is currently in clinical trials for COVID-19 (ClinicalTrials.gov Identifier: NCT04252274). As controls, we performed additional simulations for N3 and 13b  $\alpha$ -ketoamide inhibitors with 3CLpro. The experimental crystal structures of both inhibitors with protease are available<sup>4,6</sup>, and these are considered as the starting configurations for MD simulations and subsequently enhanced sampling simulations. These existing crystal structure-based inhibitor data are used as controls to compare the performance of the ligands studied in this work. The details of all ligand structures (in 2D and geometry-optimized 3D) are given in Table 1. The computational details that, to the best of our knowledge, are novel for selecting NCEs for SARS-CoV-2 are described in section 2 of the paper. In section 3, we provide the results along with discussion.

## 2. Computational Method

In this work, we have used a combination of quantum chemicals calculations to optimize the structures of ligands (Table 1), molecular docking at the binding site of the protein, and all-atom molecular dynamics (MD) of protein-ligand complexes in water to determine the stability of the complex, and enhanced free energy sampling for final identification of potential drug molecules. A detailed simulation protocol is described here.

The geometry optimizations of all the ligands (listed in Table 1) were performed using a semi-empirical method at the PM6 level, followed by geometry optimization using density functional theory (DFT) with the M06 functional and 6-311g (d,p) basis set. To account for the bulk solvent effects, the PCM method is used. Further, the partial atomic charges for the ligands are computed by fitting the electrostatic potential using the CHELPG method as implemented in the Gaussian09 code<sup>17</sup>. These charges are computed for the optimized structures using a single point calculation at the DFT with the M06 functional with 6-311g (d,p) basis set and water as the solvent.

All the ligand structures used in this study were taken from the work of Bung *et al.*<sup>9</sup>, in which the authors used parameters such as drug likeliness and toxicity to filter the molecules. They have used deep neural network-based generative and predictive methods for in silico design of NCEs. They used a dataset of ~1.6 million small molecules from the ChEMBL database to train the DNL model. The DNL model and filtration method are explained in detail in their paper. Finally, these filtered small molecules were docked using AutoDock Vina to the energy-minimized 3CLpro structure (PDB ID: 6LU7) and ranked based on their virtual screening scores. They docked 3960 molecules and obtained 1333 small molecules with virtual screening scores below -7.0. In this paper, Bung et al. reported 31 final high potential (to qualify as drug candidates) NCE molecules with virtual screening scores between -8.3 and -7.5. Of these 31, they refer to 16 molecules that are already FDA-approved drugs. The pharmacokinetics and toxicities of the considered ligands are discussed in the article by Bung *et al.*

We selected the list of ligands from that study to perform our enhanced sampling calculations to screen further and select the best candidates. To prepare the docking structures, we used the crystal structure of protease (PDB entry 6LU7, Jin *et al.*<sup>4</sup>) with an N3 inhibitor. We removed the N3 inhibitor from the crystal structure and used the same binding site (residues HIS41 and CYS145) for docking the ligands. A 3D grid of 60×60×60 Å is used around the active binding site of the protease by taking the binding site centroid as a grid center. The grid map was created with a spacing of 0.375 Å. All the ligands were then placed in the protease binding site to perform docking calculations and obtain a bound form of protease and ligands.

The virtual screening scores for binding were generated through docking to determine the affinity of all the ligands, with 3CL protease. In general, docking involves finding the optimal binding between protein and ligand. To obtain this optimal binding score, a ligand conformational search is performed around the binding sites. Here, a genetic algorithm-based conformational search was employed to find the lowest energy conformation of the ligand. The ligand conformational search is carried out by creating a grid around the binding site of the protein. The binding site of 3CLpro is already known and is reported to the HIS-41 and CYS-148 protein amino acid residues cavity<sup>4</sup>. The docking of ligands and proteins was conducted using Autodock4<sup>18</sup> software. The docking was carried out using the Lamarckian genetic algorithm (LGA), and a total of 100 GA-LA hybrid runs were used to perform the conformational search for the ligand. Furthermore, the lowest energy protein-ligand cluster was used to repeat the docking twice, and the consistency of the results were combined to obtain the best score. All docking calculations were performed using the same set of parameters.

The lowest energy docked complexes, the protein-ligand systems obtained from docking, were used to perform MD simulations. The protein was modeled using the CHARMM27 force-field<sup>19</sup> parameters. The CHARMM27 force field was employed for all the ligands, and the force-field parameters were generated using SwissParam<sup>20</sup>. The ligand partial atomic charges were computed by fitting the electrostatic potential using the CHELPG method<sup>21</sup> as implemented in the Gaussian09 code. The protein-ligand systems were solvated in water and equilibrated using MD simulations at room temperature. The systems were first equilibrated using an NVT ensemble at 300 K for 0.5 ns and extended to the NPT ensemble at 300 K and 1 atm for another 1 ns. The temperature and pressure during the simulations were maintained using a velocity rescaling thermostat and Parrinello-Rahman barostat, respectively. A time step of 2 fs was used to integrate the equation of motion, and a non-bonded cutoff of 10 Å was used to perform the MD simulations. These simulations were used to understand the stability of the interaction of the ligand with respect to the protein binding site in explicit water. We quantified the interactions between the amino acids in the binding pocket and the ligand using hydrogen bond analysis. All MD simulations were performed using the GROMACS-5.1.4 simulation package<sup>22,23</sup>. Further, the equilibrated structure obtained from the 1ns MD simulations was used to perform the free energy analysis.

Since protein-ligand systems are complex in nature, exploring various important quantities such as thermodynamics, kinetics, and microscopic description at the all-atom level, remain a challenge owing to the length scales of the systems and the time scales involved in processes such as dissociation or association<sup>24,25,26</sup> of the ligand from/to the protein binding pocket, etc. The available computational resources are generally not sufficient to address these types of complexes where sampling is very important through all-atom descriptions with brute force MD. Therefore, here we have performed enhanced sampling using metadynamics<sup>27</sup> (metaD) and its variant well-tempered metadynamics<sup>26</sup> (wt-metaD) using Plumed 2.3.0<sup>28</sup> patched with MD engine GROMACS 5.1.4. Free energy perturbation and thermodynamic integration methods are theoretically rigorous and computationally expensive. These methods produce very accurate and reliable free energy surfaces preserving entropic and enthalpic contributions for all-atom systems. However, large complex systems, for example, protein-protein binding, are often difficult to converge mainly because of the large number of collective variables and computational resources. On the other hand, the molecular mechanics Poisson-Boltzmann surface area (MMPBSA) uses approximations to calculate enthalpic and entropic contributions using implicit continuum solvent models. Therefore, there is less accuracy in comparison to free energy perturbation methods; however, this method can be used for more complex systems with proper approximations<sup>28</sup>. On the other hand, the enhanced sampling-based method metadynamics developed by Laio *et al.*<sup>27</sup>, is widely applied from simple to complex systems in a variety of fields. In a metadynamics simulation, a time-dependent bias is added to the system along some suitably chosen reaction coordinate(s) such that the deposited bias will eventually push the complex away from its minimum energy state; otherwise, the system would have generally been trapped for a sufficiently long time. This method is insensitive to the choice of reaction coordinates. A crude reaction coordinate can bias the system and help the system escape from its minima within a short time, providing a qualitative free energy surface (FES). In addition, it is not very sensitive to the precise choice of the biasing parameters except in instances where these parameters are chosen to be very high or very low. In this method, a history-dependent bias is added, which prevents the system from revisiting those regions of the phase space that are already visited. Thus, it is efficient and computationally less expensive than other methods. In addition, in comparison to MMPBSA, we performed all-atom simulations in explicit solvent medium including the dynamics of the solvent, solute, and ions.

In our metadynamics simulations, we added a bias  $V(s,t)$  in the form of Gaussians with every 500 steps (1 ps) deposition stride, with a Gaussian hill-height of 2.0 kJ/mol, width of ( $\sigma$ ) 0.1 nm, bias factor of 15, and temperature ( $T$ ) of 300 K. Once the system converges, free energy  $F(s)$  (Eq. 1)<sup>28</sup> can be extracted by adding the deposited hills along the biased reaction coordinates. In a wt-metaD, the amplitude of the bias is tuned such that the system converges smoothly. Here, we used a tempering factor  $\Delta T$  to tune the height of the hills and thus we achieved smooth convergence of the free energy landscape.

The wt-metaD simulations derived from the MD equilibrated structure as starting configurations. Since the association of a ligand from aqueous medium to the binding site in protein is an entropy-driven process and a much slower process in comparison to the dissociation of the ligand from the binding site, we mainly focused on the dissociation of ligands (Figure S2) in enhanced sampling simulations.

$$F(s) = -\frac{T+\Delta T}{T}V(s, t) + C(t) \quad (1)$$

Because we are mainly interested in the dissociation of the ligand from the binding site, we considered the center of mass distance between the heavy atoms in the ligands and the protein backbone in the vicinity of the binding pocket (Figure. S2) as the reaction coordinates. We performed 20 independent simulations for each ligand to obtain better sampling and statistically reliable results.

Because a strong binding pose will be more stable, its root-mean-square deviation (RMSD) will be lower. Therefore, high RMSD values can be used as indicators of poor binding pose, and lower RMSD indicates a stable binding pose. Hence, to determine the top binding ligands according to their binding specificity, we performed wt-metaD simulations using aligned RMSD as the reaction coordinate. We chose the RMSD for the heavy atoms of the ligands and protein backbone, as shown in Figure S2. One important thing to mention here is that in our RMSD metadynamics run, we started from the configuration corresponding to the minimum free energy value of the FES profile along the reaction coordinate of the center of mass distance ( $d$ ). We performed 30 independent wt-metaD simulations with RMSD as the reaction coordinate for each ligand, with each run extending up to 2 ns. As described here, we performed several independent short metadynamics simulations with RMSD as the reaction coordinate. Therefore, this is similar to performing a much longer unbiased MD simulation where the starting structure of the ligand-protein complex could overcome the local barriers and reach a global minimum. Therefore, these independent trajectories were used to evaluate the stability that was translated into scores for the ligand-protein complexes. The analysis (scoring) methods from these trajectories are described along with the results and discussion.

### 3. Results and Discussion

The protein-ligand docking scores obtained from this study are shown in Table 2. The docking trend was found to be in qualitative agreement with the results obtained from Autodock Vina<sup>11</sup>. However, the absolute scores obtained from our simulations are different from those reported using Autodock Vina, owing to the differences in the force fields used in Autodock 4 and Vina. The lowest energy docked complexes were examined to determine the ligand location with respect to the protein binding site. The PI-06 ligand was found to exhibit a high binding affinity with the protein. Among the 17 ligands, PI-04, PI-06, PI-10, and PI-12 ligands were found to exhibit higher binding scores with the protein. We also performed docking for N3 and 13b with 3CLpro and present the scores in Table 2. The binding affinity of the top four ligands was found to be higher than that of the control inhibitors, N3 and 13b. The binding energies are found to be -8.20 kcal/mol and -8.28 kcal/mol for N3 and 13b, respectively.

The rigid docking-based binding energies are -6.74 kcal/mol and -9.07 kcal/mol for N3 and 13b respectively. All ligands considered in this study were found to be in the binding pocket and interact with HIS41 and CYS 148. The best docking poses of the four ligands obtained from docking are shown in Figure 1. The ligand position clearly indicates that the ligand has a tendency to stay at the binding site of the protein.

The docked poses of the 13 other ligands are shown in Figure S1, which shows that the ligands tend to stay in the binding pocket. To further understand the docked complex stability and the interactions of the ligand with protein in the binding pocket, the best-docked complexes were solvated with water, and MD simulations were performed.

In the case of docking, the protein is considered to be rigid, and a conformational search is carried out in the gas phase. It is very difficult to presume the docked complex as stable. Docking is mainly useful to eliminate the ligands that are very improbable. Hence, it is very important to perform all-atom MD simulations to assess the stability of the docked complex. Thus, the docked complexes with all-atom description were simulated in the presence of a solvent. The aqueous solvent environment plays an important role in the stability of the docked complex because of the solvation of ligands and the dynamics of the solvated protein.

The docked complexes were used as initial configurations to perform the MD simulations. These systems are equilibrated in water, and simulations are performed at room temperature and 1 atm pressure. The last 0.5 ns trajectory data obtained from NPT ensemble simulations were used to compute the RMSD for the binding site and ligand to assess the stability (see Table S1) of the protein-ligand complexes and to validate the docking pose. The RMSD values were found to be less than 0.2 nm for all the protein-ligand complexes. This clearly shows that the protein-ligand systems are stable, and the ligands tend to be in the binding pocket.

Furthermore, to elucidate the main interactions of the ligand with the protein amino acids in the binding pocket, we analyzed different hydrogen bonding scenarios. The interacting groups of proteins and ligands through hydrogen bonds are listed in Table 2. Almost all the ligands interact with the -NH<sub>2</sub> groups of the protein. The most common interacting amino acids in the binding pocket are THR26, ASN142, and GLN189. The PI-04, PI-06, PI-08, PI-10, PI-11, PI-13, and PI-17 ligands are found to be interacting with a greater number of residues in the binding pocket than the other ligands. However, this observation is only based on the hydrogen bonding performed on the structure obtained from equilibrium NPT simulations, and it is highly probable that ligands might show other predominant interactions. To determine the contribution from all possible interactions, one needs to explore the complete free energy surface associated with ligand-protein binding.

The entropic contributions associated with the solvent and the conformational changes of the protein-ligand complexes are not accounted for in the docking. In the case of MD simulation, the sampling around the binding site of the protein is also not sufficient, as conformations might get stuck in local minima. Therefore, enhanced sampling of ligand binding and analysis of changes in conformation of ligands is important to ascertain the most stable (bound) protein-ligand complex from the set of 17 complexes reported here. The equilibrium structure obtained from MD simulations was used as the starting configuration in the enhanced wt-metaD simulations. The average free energy of dissociation for all the ligands obtained from the wt-metD simulations is reported in Table 2, and the corresponding free energy profiles are shown in Figure 2 (a). We also performed enhanced sampling simulations for N3 and 13b ketoamide, and the obtained FESs are shown in Figure 2(a). Here, the average free energy

values were obtained from 20 independent dissociation simulations for each ligand to obtain better sampling and statistically reliable results. The free energy values are found to be in the range of -13.40 to -2.83 kcal/mol for all the ligands (see Table 2). The PI-06, PI-08, PI-11, and PI-14 ligands were found to exhibit higher energy barriers in the same order compared to the other ligands. The maximum free energy of binding is -13.40 kcal/mol, which is observed for the PI-06 ligand. These four ligands clearly outperformed all other ligands. However, PI-06 was the best among the four ligands with -4.63 kcal/mol lower free energy than the second-best ligand PI-14. To better understand the free energy behavior, the profiles for these four ligands are shown separately in Figure 2 (b). To compare the binding affinities with the top 4 ligands, we have also shown the FESs for N3 and 13b (control inhibitors) in Figure 2(b). As can be seen from Figure 2(b), our top four ligands show comparable binding energies to control ligand 13b. The free energy surfaces displayed in Figure 2(a) and (b) show a complex and rugged free energy landscape with multiple local minima and one global minimum, that is, at the binding site. This behavior represents multiple interactions between the ligands and residues of the binding site.

As the solvent effects are not included in the docking, the ligand-protein interactions are expected to be different from the wt-metD simulations, in which the protein-ligand system is solvated in water. Thus, after performing wt-metaD simulations, the protein-ligand complex configuration corresponding to the free energy minimum position (Figure 2) was superimposed with the complex obtained from docking. We present in Figure 3 the superimposed structures of the FES minimum configuration and the docked complex for the PI-06 ligand. In wt-metD, the ligand position was found to be in the binding pocket marginally away from residues HIS41 and CYS148. The ligand in the docked pose is shown as red sticks, whereas the FES minimum pose is shown as green sticks. In addition, to assess the binding landscape of the ligand-protein and to validate the binding pose from the FES, the RMSD based free energy was computed.

To understand the FES of the binding poses of the ligands in detail, we looked into the FES as a function of RMSD (reaction coordinate) as described in the computational method section. For poorly bound ligands, the RMSD (with respect to the lowest energy binding structure obtained from FES described in Figure 2) is expected to be higher than that of the strongly bound structures. Therefore, RMSD can be attributed as a measure of the binding between the ligands and proteins. Thus, we took the minimum free energy configuration from Figure 2 as the starting structure for the wt-metaD simulations and RMSD as the reaction coordinates. In Figure 4(a), we present the free energy as a function of the aligned RMSD for all ligands. Here, each FES is averaged over 30 independent runs. From the FES of Figure 4 (a), it is evident that there are stable conformations for all the ligands below 0.2 nm of RMSD. Therefore, there is a global minimum for all the ligands close to the starting conformation, and almost no other local or global minima were observed. However, there some metastable states exist after 0.3 nm RMSD. Therefore, to quantify the binding of ligands to the protein, we computed the probability of the ligand-protein complex within 0.2 nm of RMSD from all trajectories we obtained from FES calculations with RMSD as the reaction coordinate. The trajectories (RMSD as a function of time) for the top four ligand-protein complexes, including N3 and 13b, are shown in Figure 5 to elucidate the stability of the ligands in the binding site (see Figure S3 for all the ligands). In Figure 4 (b), we have depicted the distribution of the probability of RMSD for these six ligands. For ligands PI-06, PI-08, PI-11, N3, and 13b, we observed sharp peaks in the distribution of probability values for RMSD below 0.2 nm, and for PI-14 it is slightly lower. This signifies that PI-06, PI-08, PI-11, and PI-14 ligands are strongly bound at the binding site of the protein, and the binding energies of PI-06, PI-08, and PI-11 are comparable to those of the control ligand 13b. However, the binding energy of PI-14 means that this ligand emerged as the best among all the ligands, including the controls. The probability of the



RMSD value below 0.2 nm could, therefore, be an indicator of binding. Thus, a higher probability would indicate stronger bonding. These values are reported in Table 2 along with the free energy change for all the ligands.

In a similar way, to find the stability of ligand-protein complexes, we calculated the average RMSD ( $\langle s \rangle$ ) from all the independent biased trajectories using the following equation:

$$\langle S \rangle = \frac{\int ds s e^{-F(s)/k_B T}}{\int ds e^{-F(s)/k_B T}} \quad (2)$$

Here,  $F(s)$  is the energy associated with the RMSD. A higher estimate of the average or thermodynamically preferred RMSD can be considered an indication of poor instability of the complex. Thus, the higher the value (score) the lowers the stability and vice-a-versa. All these quantitative estimations of the stability (score) for each ligand using Eq. 2 are reported in Table 2.

We calculated two types of scores (probability of RMSD below 0.2 nm and average RMSD as per Eq. 2) from the biased trajectories obtained from the metadynamics simulations. In Figure 6 we present the correlation of these two types of scores with the FESs for all ligands. It is evident that for the ligands with a lower free energy barrier for dissociation (from the binding site), the average RMSD is lower and the probability of RMSD (below 0.2) is higher. These distinct correlations confirm that our method could well segregate ligands that show higher stability than others. We used docking structures with similar docking scores and separated the four ligands that bind 3CLpro with much higher affinity. These ligands are in the order PI-06 > PI-14 > PI-11 > PI-08 according to the free energy barrier and average RMSD. However, if we consider the probability of RMSD values less than 0.2 nm, then the resulting order is PI-06 > PI-08 > PI-11 > PI-14. From all these scores, it is evident that PI-06 has a much higher probability compared to the other three ligands to bind the protein.

We have shown the FES of dissociation in Figure 2. We observed that the free energy profile for PI-06 has a much higher energy of solvation (at the dissociated state) than the other ligands. Furthermore, in the case of all the ligands, there were local minima present along with one global minimum in the free energy landscape. To understand this feature, we looked into the dissociation trajectory for the PI-06 ligand (see Figure 7). We found full dissociation of the ligand from the binding pocket to the aqueous environment. Initially, the ligand is at the binding pocket and explores various conformations (red wire representation). Due to the applied bias along the center of mass-center of mass distance ( $d$ ), the ligand gradually escapes from the minimum of the potential well and explores other regions of the phase space (gray wire representation). Later, the ligand fully escaped from the binding pocket to the solvent (blue wire). As can be seen from the trajectory, the ligand strongly interacts with the protein backbone near the vicinity of the binding pocket, which gives rise to these local features in the free energy landscape, as shown in Figure 7.

We further analyzed the hydrogen bonds (Hbond) between all ligand-protein systems from unbiased MD simulations starting with the minimum free energy conformation obtained from FES. From these unbiased MD trajectories, we calculated Hbonds as a function of time using a distance cutoff of 0.35 nm between the donor and acceptor, and an angle cutoff of 30 ° for the hydrogen-donor-acceptor. In Figure 8, we show the probability distribution of the Hbonds with the number of Hbonds formed for the top four ligands, including N3 and 13b, as per the free energy barrier. PI-06 has a higher probability of forming two and three Hbonds in comparison with the other three ligands (see Figure S4 for all the ligands). The calculated weighted average of the Hbonds for the top 4 ligands elucidates the same order

of PI-06 > PI-08 > PI-11 > PI-14, as observed from the RMSD probability values mentioned above. We also computed the hydrogen bond lifetime correlation functions, which are depicted in Figure 8(b), for the top four ligands, including N3 and 13b. The correlation function for PI-06 was found to exhibit a very slow decay, that is, a higher Hbond lifetime compared to other ligands. We presented the Hbond lifetime correlation function for all other ligands in Figure S5 of the supporting information. The Hbond analysis clearly shows that in the case of PI-06, the contribution of hydrogen bond interactions is the highest among all other ligands that stabilize the protein-ligand complex.

#### 4. Conclusion

In this study, we performed large-scale all-atom molecular dynamics simulations with enhanced sampling for ligands that bind to the 3CL protease of SARS-CoV-2. These calculations are robust and are modeled similarly to the experimental system by incorporating explicit solvent molecules and considering all-atom molecular models and interactions. We considered a set of 17 ligands with lower virtual screening scores (for 3CLpro of SARS-CoV-2) and high Tanimoto score with respect to known HIV inhibitors, for example, currently FDA-approved drugs darunavir, lopinavir, ritonavir, indinavir, saquinavir, and ASC09. Our method could distinctively isolate these 17 ligands into four possible NCEs and could even identify the best compound with very high confidence. In addition, we validated our method by performing similar calculations for N3 and 13b  $\alpha$ -ketoamide inhibitors as controls. Upon successful synthesis and testing, these four NCEs are expected to have a much higher probability of success in clinical trials.

The method described in this work is scalable for multiple targets (proteins from the same family with similarities) of ligand binding that could result in a much smaller subset of NCEs compared to docking or any other drug screening method. The method demonstrated here is envisaged to significantly reduce the time of drug design and discovery.

#### Acknowledgment

DP thanks IIT Kanpur for its generous financial support. SN would like to thank Dr. Amit Kumawat for helpful discussions. JKS acknowledges the support of the Science and Engineering Research Board, Department of Science and Technology (grant number: SB/S1/Covid-9/2020) for partial funding. This work is also supported by the CSR funding of GST IN and Suraj Logistix Pvt. Ltd.

#### Supplementary Material

In the supplementary we have shown docking pose of all the ligands, RMSD values obtained from NPT MD simulation with respect to the docking pose, highlighted groups used calculation of center of mass collective variable for metadynamics simulation, time evaluation of RMSD for biased trajectories where RMSD of heavy atoms were considered as collective variables, normalized probability distribution of the hydrogen bonds with the number of hydrogen bonds for all the ligands from unbiased MD runs.

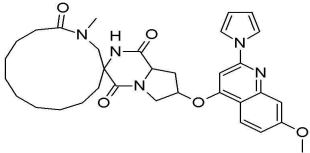
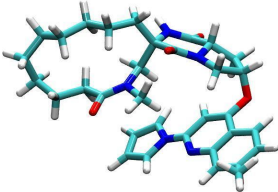
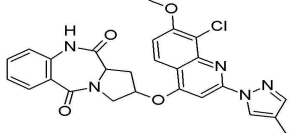
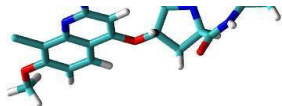
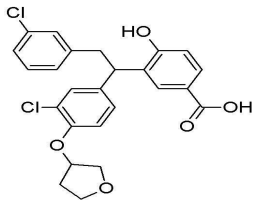
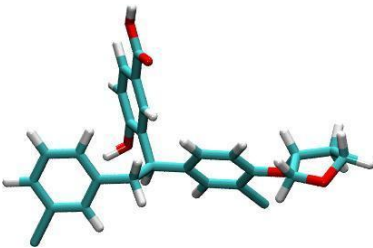
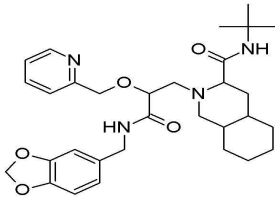
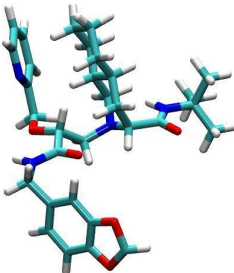
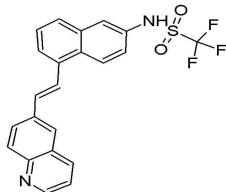
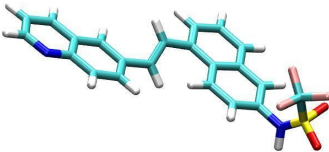
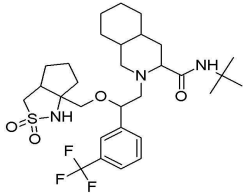
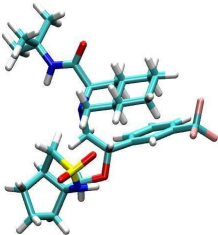
#### References

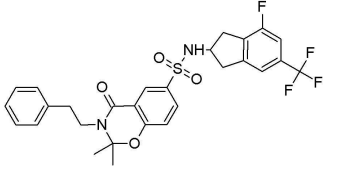
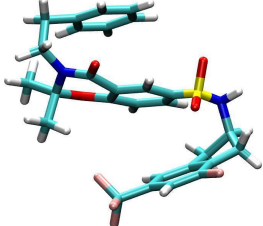
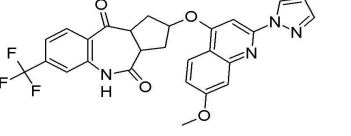
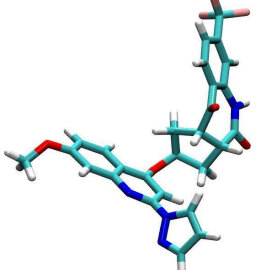
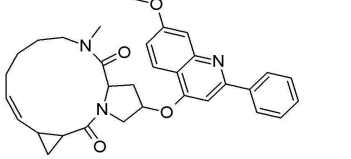
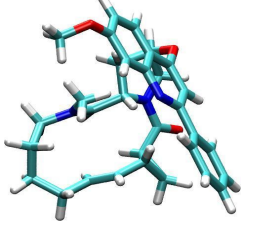
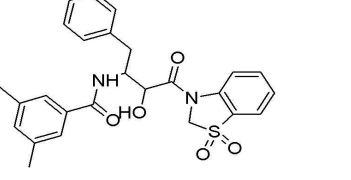
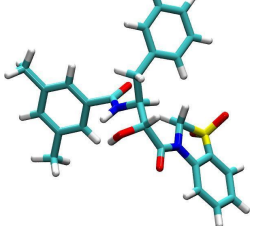
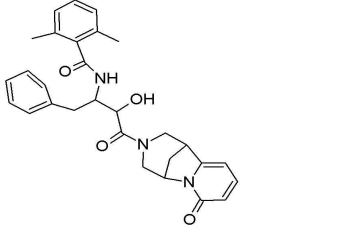
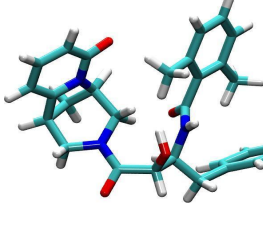
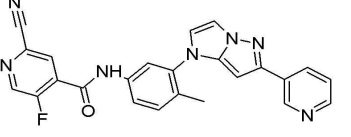
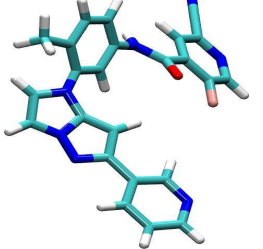
1. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, *Lancet* **395**, 565–574 (2020).
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

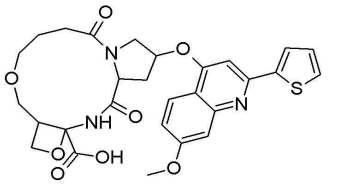
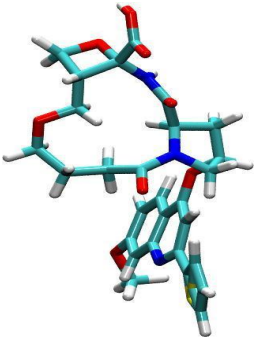
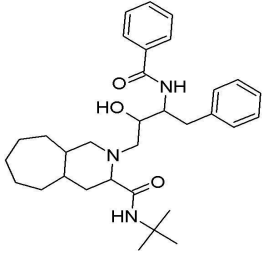
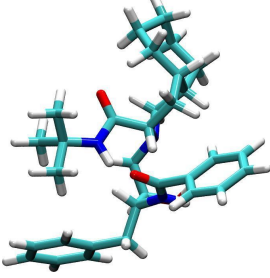
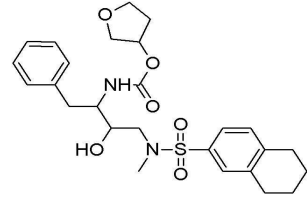
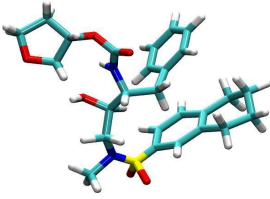
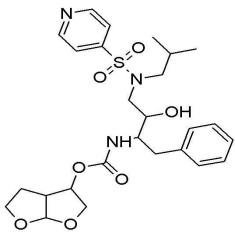
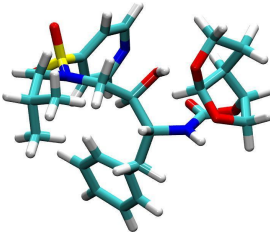
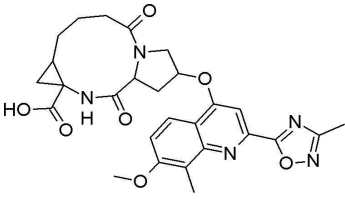
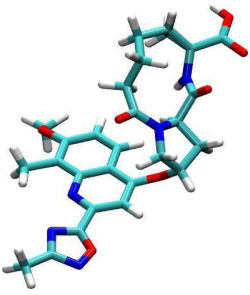
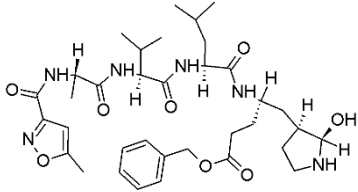
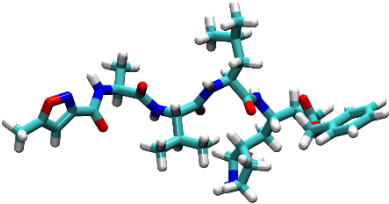
3. Zumla, A., Chan, J. F. W., Azhar, E. I., Hui, D. S. C. & Yuen, K. Y. Coronaviruses-drug discovery and therapeutic options. *Nat. Rev. Drug Discov.* **15**, 327–347 (2016).
4. Jin, Z. *et al.* Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* **582**, 289–312, (2020).
5. Liu, X. & Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *J. Genet. Genomics*, **47**, 119–121, (2020).
6. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved  $\alpha$ -ketoamide inhibitors, *Science*, **368**, 409–412 (2020)
7. Zhang, L. *et al.*  $\alpha$ -Ketoamides as Broad-Spectrum Inhibitors of Coronavirus and Enterovirus Replication: Structure-Based Design, Synthesis, and Activity Assessment, *J. Med. Chem.*, **63**, 9, 4562–4578 (2020)
8. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281–292.e6 (2020).
9. Bung, N., Krishnan, S. R. K., Bulusu, G. & Roy, A. De novo design of new chemical entities (NCEs) for SARS-CoV-2 using artificial intelligence, DOI: 10.26434/chemrxiv.11998347.v2
10. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–7, (2012).
11. Bickerton, G. R. *et al.* Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
12. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, Article number 8, (2009).
13. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461, (2009).
14. Lipkus A H. A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* **26**, 263–265 (1999).
15. Hassan Baig, M. *et al.* Computer Aided Drug Design: Success and Limitations. *Curr. Pharm. Des.* **22**, 572–581 (2016).
16. Pons, C. *et al.* Present and future challenges and limitations in protein-Protein docking. *Proteins Struct. Funct. Bioinforma.* **78**, 95–108 (2010).
17. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ort. Gaussian 09, Revision A.02. (2016).
18. Morris, G. M. *et al.* Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
19. Foloppe, N. & Mackerell, A. D. All-Atom Empirical Force Field for Nucleic Acids: I. Parameter Optimization Based on Small Molecule and Condensed Phase Macromolecular target data I. *J. Comput. Chem.* **21**, 86–104 (2000).
20. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. *J. Comput. Chem.* **32**, 2359–2368 (2011).
21. Breneman, C. M. & Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **11**, 361–373 (1990).
22. Pronk, S. *et al.* GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
23. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
24. Pramanik, D., Smith, Z., Kells, A. & Tiwary, P. Can one trust kinetic and thermodynamic observables from biased metadynamics simulations: detailed quantitative benchmarks on

- millimolar drug fragment dissociation. *J. Phys. Chem. B*, **123**, 3672-3678 (2019).
25. Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding Kinetics of a p38 MAP Kinase Type II Inhibitor from Metadynamics Simulations. *J. Am. Chem. Soc.* **139**, 4780–4788 (2017).
  26. Pan, A. C., Xu H., Palpant T. & Shaw D. E. Quantitative Characterization of the Binding and Unbinding of Millimolar Drug Fragments with Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **13**, 3372-3377, (2017).
  27. Laio, A., Parrinello, M. *Escaping free-energy minima*, PNAS, **99**, 12562-12566 (2002)
  28. Wang C., Greene D., Xiao L., Qi R., & Luo R. Recent Developments and Applications of the MMPBSA Method, *Front. Mol. Biosci.*, **4**, Article 87 (2018)
  29. Barducci, A. et al., M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.*, **100**, 020603, (2008).
  30. Bonomi, M. et al. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180**, 1961-1972, (2009).
  31. Clark, A. J. et al. Prediction of Protein-Ligand Binding Poses via a Combination of Induced Fit Docking and Metadynamics Simulations. *J. Chem. Theory Comput.* **12**, 2990-2998, (2016).

Table 1. Code-name, chemical structures of the ligands and QM Density Functional Theory (DFT) optimized structures. The control inhibitor structures, N3 and 13b, are also included.

Code Name	2D structure	3D DFT optimized structure
PI-01		
PI-02		
PI-03		
PI-04		
PI-05		
PI-06		

PI-07		
PI-08		
PI-09		
PI-10		
PI-11		
PI-12		

PI-13		
PI-14		
PI-15		
PI-16		
PI-17		
N3		

13b  $\alpha$ -ketoamide

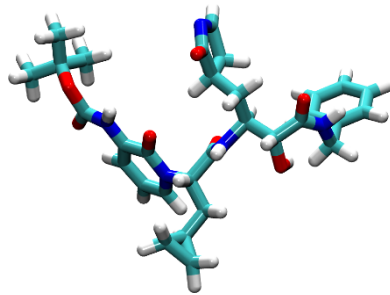
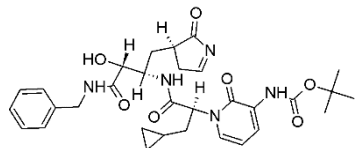


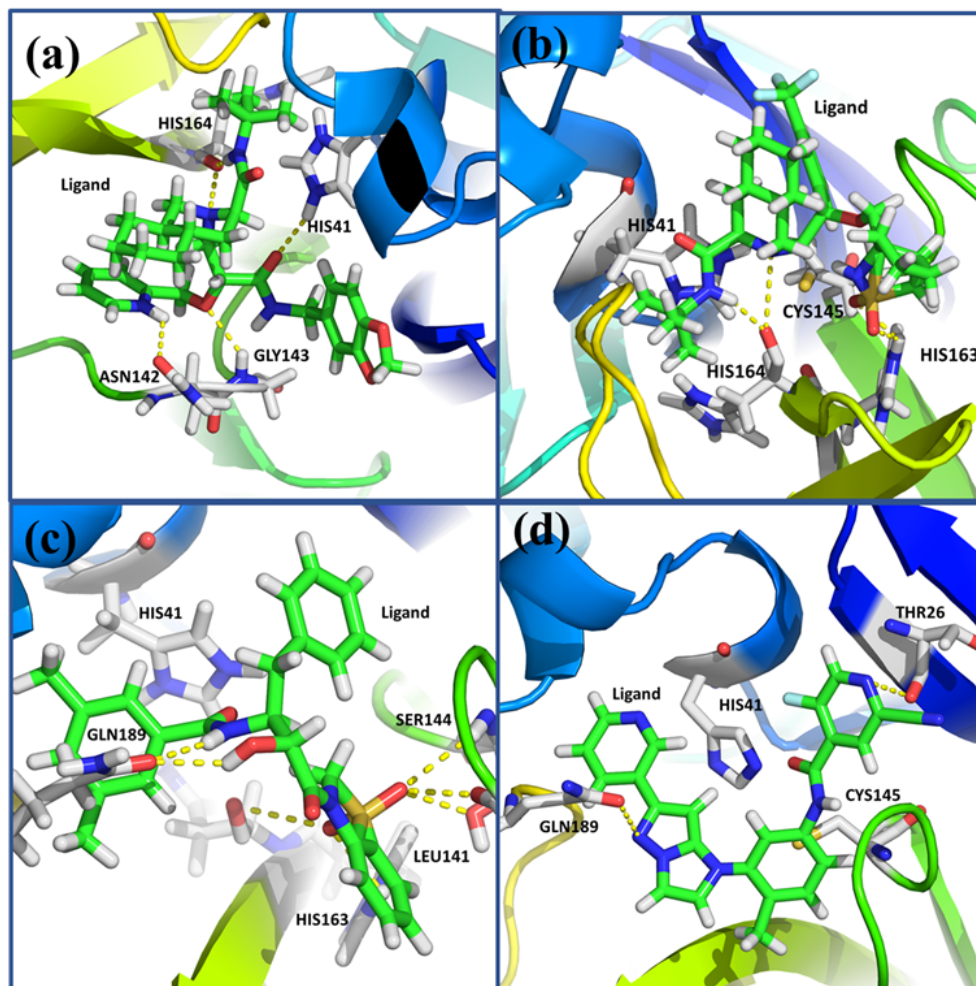


Table 2 The docking score, free energies for dissociation, average RMSD values, probabilities (for RMSD < 0.2 nm), for all ligands. The interaction residues and functional groups the protein with the ligands.

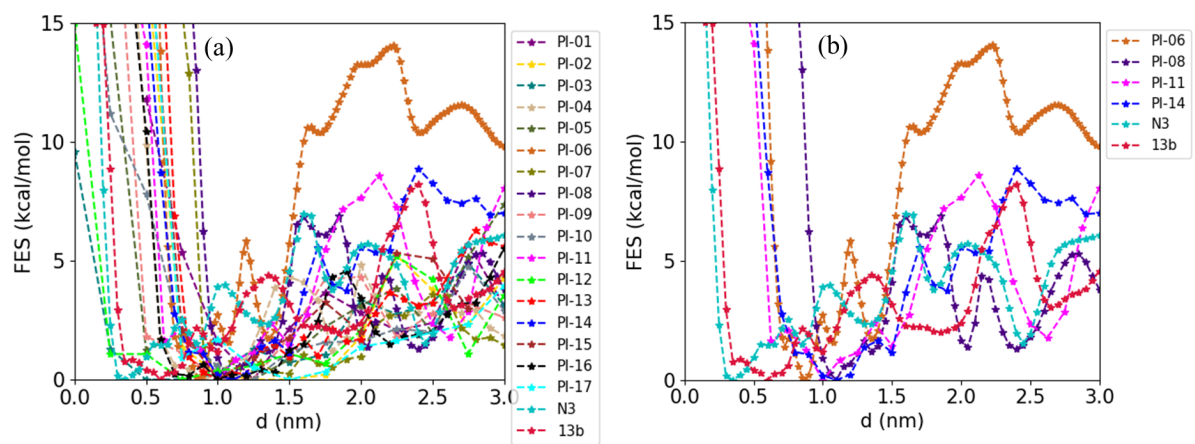
Ligand code name	Docking score	Free Energy (kcal/mol)	Average RMSD (nm) as per Eq. 2	Probability of RMSD(RMSD < 0.2 nm)	Interacting Residue	Residue-ligand interacting groups
PI-01	-8.51	-3.83 (0.609)	0.359 (0.068)	0.361	THR26	OH-C=O
PI-02	-9.13	-4.49 (0.451)	0.401 (0.072)	0.254	THR24 ASN142 GLN189	OH-C=O NH2-C=O NH2-C=O
PI-03	-9.43	-3.36 (0.598)	0.344 (0.058)	0.472	HIS163 HIS16	NH2-C=O NH2-OH
PI-04	-11.56	-4.78 (0.392)	0.207 (0.035)	0.594	GLY143 SER144 ASN142 GLN189	NH2-C=O NH2-C=O O-OH N-OH
PI-05	-10.85	-4.29 (0.574)	0.214 (0.028)	0.443	ASN142	NH2-N (-SO2)
PI-06	-11.92	-13.40 (0.430)	0.143 (0.004)	0.842	THR26 ASN142 GLY143 CYS148 GLY143	NH2-O NH2-C=O; O-NH2 NH2-N(-SO2) NH2-C=O NH2-O
PI-07	-10.40	-3.57 (0.492)	0.424 (0.078)	0.418	HIS41 ASN142 GLU166	NH2-C=O NH2-O O-NH
PI-08	-9.50	-6.65 (0.411)	0.227 (0.051)	0.722	THR26 ASN119 ASN142 GLY143 LEU27	NH2-O=C NH2-O=C NH2-O NH2-N O-NH (-NC=O)
PI-09	-10.30	-3.17 (0.516)	0.395 (0.081)	0.453	THR26 SER46 HIS143	NH2-O=C OH-O=C NH2-O
PI-10	-11.64	-3.07 (0.363)	0.394 (0.068)	0.317	THR26 SER46 ASN142	NH2-O=C OH-O=C NH2-OH

					GLN189 SER46	NH2-N NH2-OH
PI-11	-10.54	-8.74 (0.392)	0.174 (0.016)	0.719	HIS41 ASN142 GLN189	NH2-O=C NH2-OH NH2-O=C; NH2-OH
PI-12	-10.94	-6.43 (0.466)	0.214 (0.036)	0.522	HIS41	NH2-O=C
PI-13	-10.22	-6.20 (0.406)	0.260 (0.055)	0.666	ASN142 GLY143 GLU166 CYS148	NH2-O=C NH2-O=C NH2-OH NH2-O=C
PI-14	-10.64	-8.77 (0.334)	0.160 (0.005)	0.642	ASN142 GLU166 GLN189	NH2-O=C NH2-O=C O-OH
PI-15	-9.68	-3.77 (0.399)	0.217 (0.030)	0.384	GLY143 GLU166 CYS148	NH2-O NH2-O NH2-O
PI-16	-9.52	-6.26 (0.456)	0.198 (0.025)	0.604	ASN142 GLU166	NH2-N NH2-O=C
PI-17	-9.04	-2.83 (0.514)	0.426 (0.067)	0.156	ASN142 GLY143 GLN189 SER46	NH2-O=C NH2-O=C NH2-N; NH2- O; NH2-O=C O-NH
N3	-7.75	-6.91 (0.478)	0.244 (0.086)	0.66	HIS41 ASN142 GLY143 SER144 HIS172 GLU166 GLN189	NH2-OR NH2-O=C NH2-OR NH2-O=C NH2-N NH2-O=C NH2-OR
13b	-8.54	-8.15 (0.382)	0.171 (0.008)	0.57	HIS41 ASN142  GLY143 CYS145 GLU166 GLN189 HIS164	NH2-O=C NH2-OR; O- NH; N-NH2 NH2-O=C NH2-O=C NH2-O=C NH2-O=C O-OH

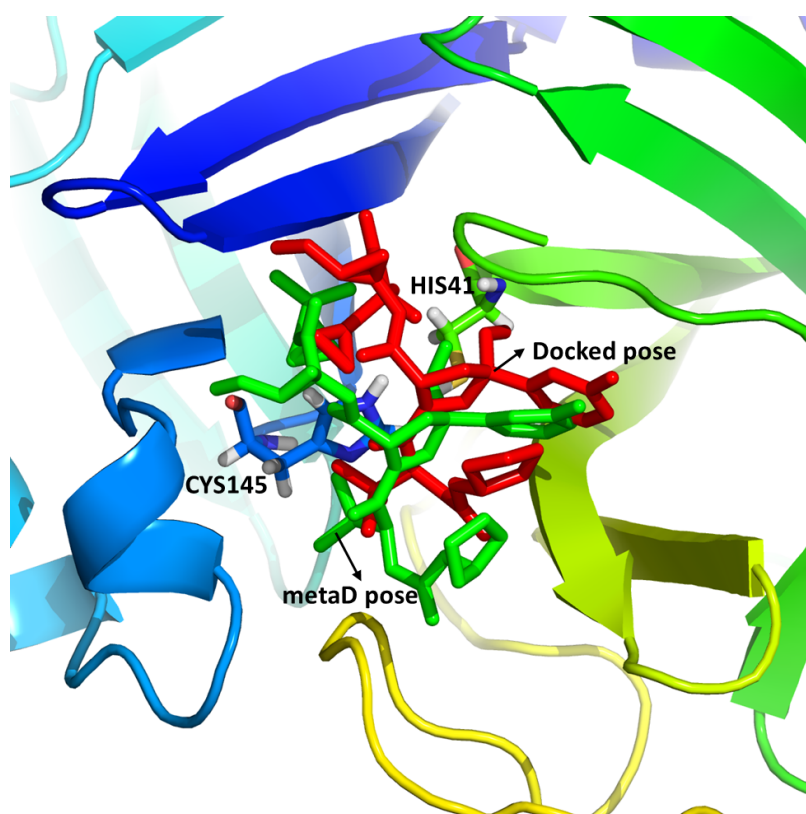
					THR26	O-NH
--	--	--	--	--	-------	------



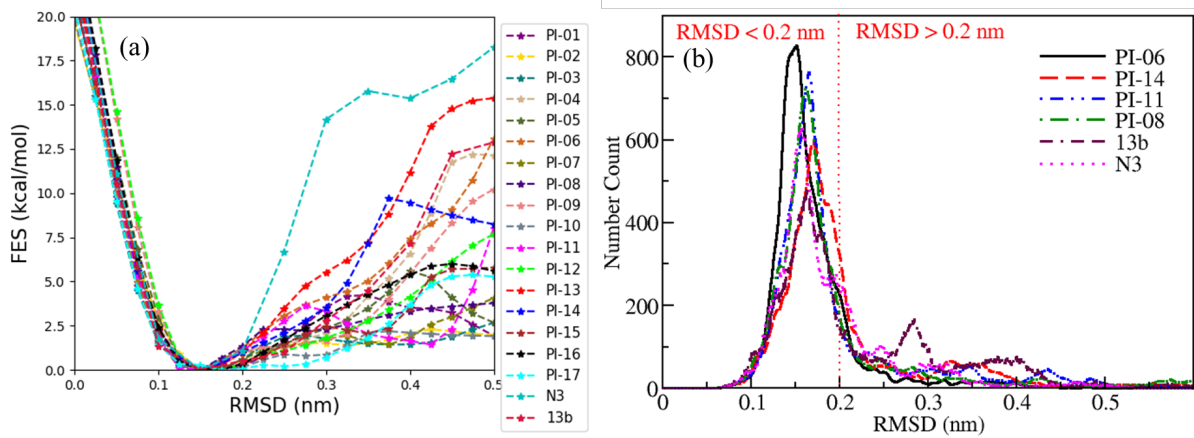
**Figure 1** The best docking poses of lowest binding energy 4 ligands with protein (a) PI-04 (b) PI-06 (c) PI-10 and (d) PI-12 are shown here. The active site of the protein (HIS41 and CYS148) is shown as red sticks. The stabilizing polar interaction contacts are shown in yellow dotted lines in all the figures.



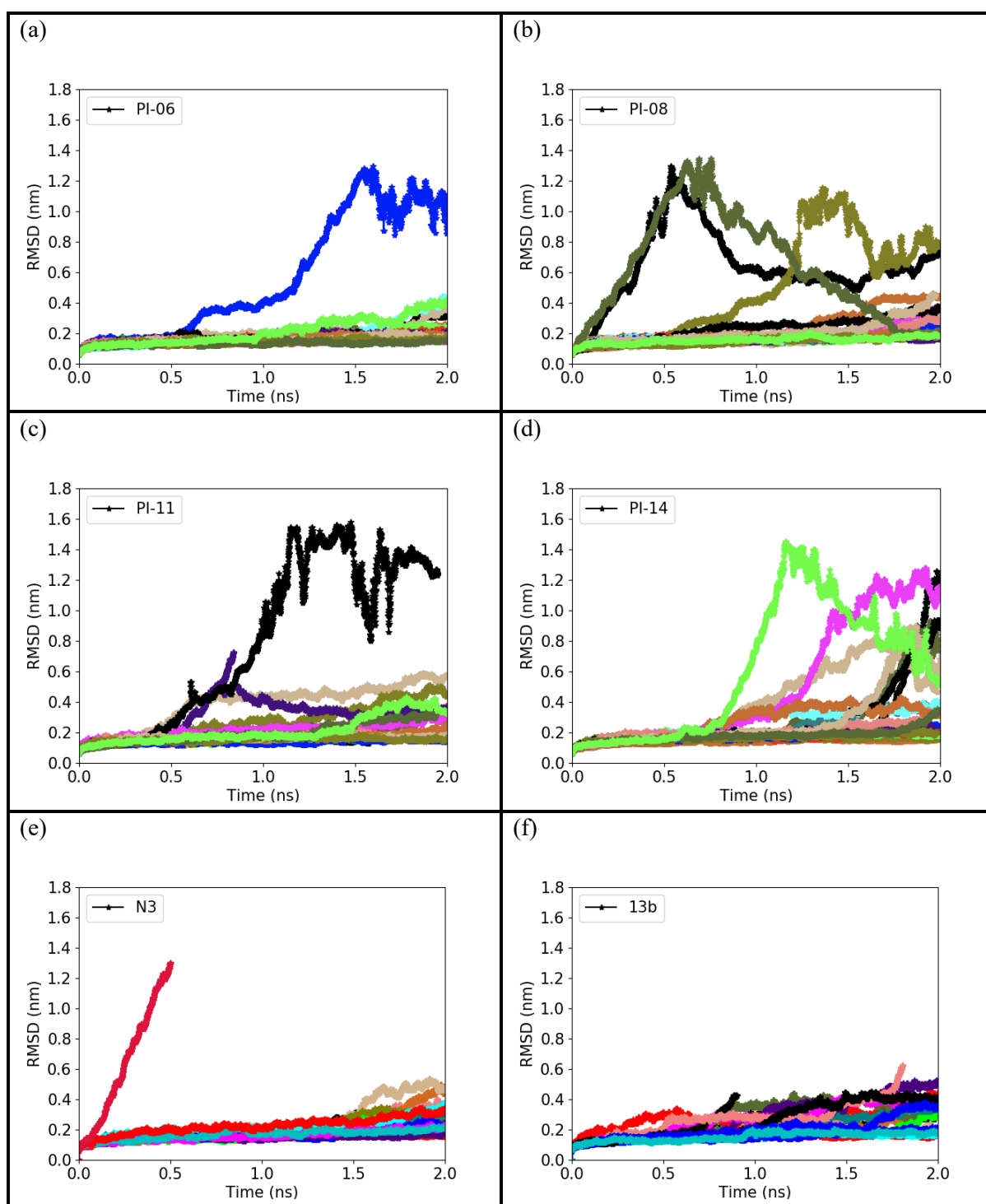
**Figure 2:** Average free energy with center of mass - center of mass distance ( $d$ ) for dissociation of the ligands from the protein binding pocket. (a) Free energies for all the ligands. (b) Free energies for the top four ligands including N3 and 13b. For each ligand the errors in free energies are reported in Table 2.



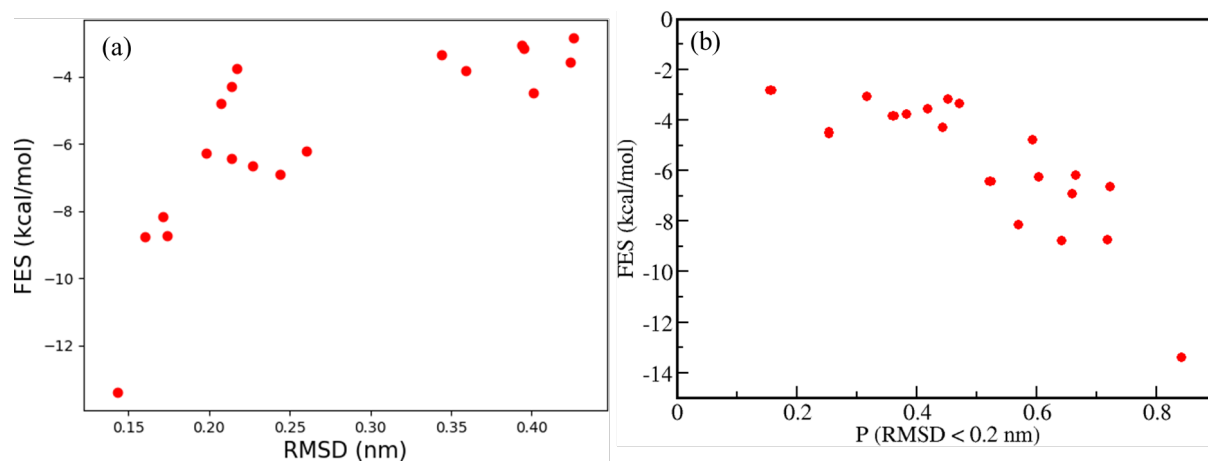
**Figure 3:** The zoom-in view for the superimposed structure of the PI-06 ligand docked pose and stable pose from the free energy minima. The Ligand in the docked pose is shown as red sticks and that of the free energy minimum structure is shown as green sticks.



**Figure 4:** (a) Average free energy with aligned RMSD for all ligands. (b) The number count distributions for the probability to find a system within 0.2 nm of RMSD.

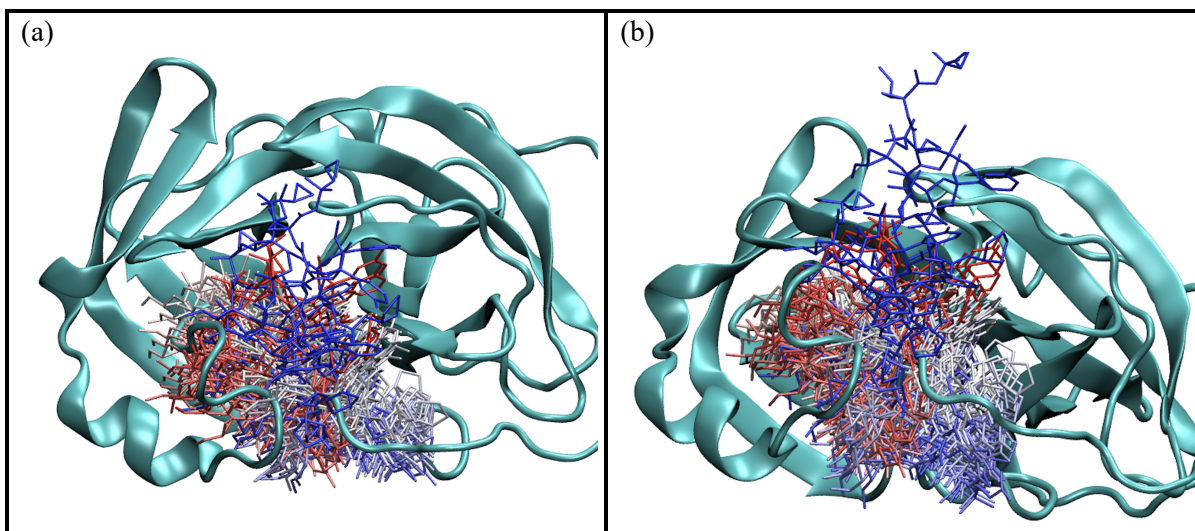


**Figure 5:** Time evolution of the RMSD for top four ligands, (a) PI-06, (b) PI-08, (c) PI-11 (d) PI-14, (e) N3 and (f) 13b. In each plot, we show RMSD from all independent runs.

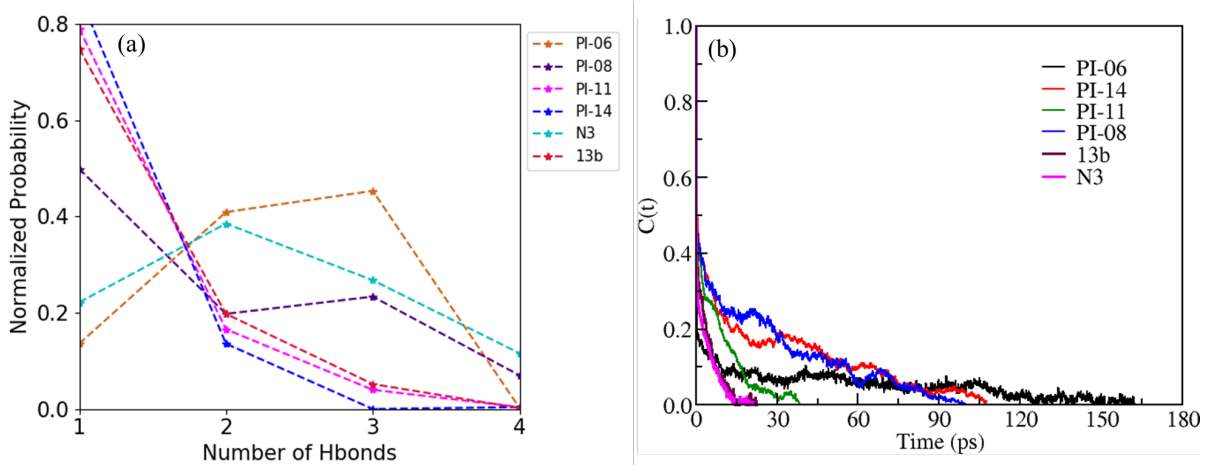


**Figure 6:** Average free energy of protein-ligand as a function of (a) average RMSD as per Eq. 2 and (b) probability of RMSD, here probability is calculated for the ligands which shows less than 0.2 nm RMSD.





**Figure 7:** The trajectory of the ligand dissociation from the protein binding pocket. Two different views (a) left, and (b) right, show the full dissociation of the ligand from the binding pocket, interactions of the ligand with the protein backbone in the vicinity of the binding pocket for the ligand PI-06 from an independent simulation. The colors of the ligand wire frames are from red (inside the binding pocket) to gray (in between) to blue (outside of the pocket).



**Figure 8:** (a) Normalized probability of the formation of hydrogen bonds (b) The hydrogen bond correlation functions, for the top four ligands including controls N3 and 13b.