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

Sadanandam Namsani[1$], Debabrata Pramanik[1,2,$], Mohd Aamir Khan[1,2], Sudip Roy*[1] and Jayant Kumar Singh[*1,2]

[1]Prescience Insilico Private Limited
Old Madras Road, Bangalore 560049, India

[2]Department of Chemical Engineering
Indian Institute of Technology, Kanpur, India

[$] 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 are highly specific in binding towards the 3-chymotrypsin-like cysteine protease (3CLpro) protein present in the novel SARS-CoV2 virus. The viral 3CLpro protein controls coronavirus replication. Therefore, 3CLpro is identified as a target for drug molecules. We have implemented an enhanced sampling method in combination with molecular dynamics and docking to bring down the computational screening search space to four molecules that could be synthesised and tested against COVID-19. 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 are able to perform large number of simulations within a month's time and converge to 4 most strongly bound ligands (by free energy and other scores) from a set of 17 ligands with lower docking scores. Based on our results and analysis, we claim with high confidence, that we have identified four potential ligands. Out of those, one particular ligand is the most promising candidate, based on free energy data, for further synthesis and testing against SARS-CoV-2 and might be effective for the cure of COVID-19.

1. Introduction

The current situation of the world is extraordinary due to Coronavirus Disease-2019 (COVID-19) pandemic. COVID-19 is caused by a new pathogen, severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) Virus which is from the family of betacoronavirus genus [1,2]. The infection with the new pathogenic SARS-CoV2 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. This family of SARS virus is different and thus there is a huge need for drug candidates and vaccines to be invented in a few months to tackle the pandemic. The current crisis is mainly because of the lack of any specific antiviral drugs that could function against the SARS-CoV2 or due to a lack of preparedness for finding and producing a new vaccine. To mitigate the risks posed by viruses, including the SARS-CoV-2, it is imperative that research efforts for the development of new antiviral agents targeting this virus be pursued with renewed invigoration. However, identification of a drug candidate is a time-consuming process and the final release of new drugs for patients takes a minimum of 10 years of research to identify potential efficacious molecules with less toxicity, testing in animals followed by human, and regulatory approvals. Moreover, the viruses have the tendency and ability to mutate rapidly in response to drug molecules, and thus it is imperative that an in-depth understanding of structure-activity relations with respect to the biology should evolve in order to combat current and future outbreaks.

The scientific world is responding to this pandemic by three major paths of innovation. The most focused research currently in progress is the development of vaccine candidates and clinical trials of existing FDA approved drugs for other relevant diseases, in order to repurpose them for the COVID-19. The third set of scientists are focusing on innovating new chemical entities (NCEs), as repurposing of drugs may fail and the reach of the vaccine could be limited all over the world in the initial 2-3 years. The research on NCEs is dependent on finding targets i.e., the proteins that are envisaged as moderators of functions, to help the virus propagate in human body. So, NCEs are designed to inhibit these proteins either on viruses or in human cells to stop the biological pathways hence control the disease. The ab initio design of NCEs predominantly starts from the computational screening of large sets e.g., millions of chemicals already available in chemical databases. So the key approach of this high-throughput computational screening[3] is to identify molecules from existing molecular databases that may have a therapeutic effect on 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 for mediating viral replication and transcription. The Mpro has similar cleavage-site specificity 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 the X-ray structures[4] of the SARS-CoV-2 Mpro and its complex with N3 inhibitor. The crystal structure of COVID-19 main protease in complex with an inhibitor N3 is reported[4] 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[5] for 2019-nCoV. They have used molecular docking for targeting Mpro. These predicted drugs formed more hydrogen bonds with 2019-nCoV Mpro compared to lopinavir/ritonavir. Walls et. al. recently showed that SARS-CoV-2 S uses membrane associated protein ACE2 of human cells to enter[6]. The receptor-binding domains of SARS-CoV-2 S and SARS-CoV S (SARS coronavirus identified in 2003) bind with similar affinities[6] to ACE2 of human cell. 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 sets the virus SARS-CoV2 different 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[7] their initial work on de novo design of NCEs for SARS-CoV-2 , targeting 3CLpro protein. They have performed deep neural network-based (DNL) generative and predictive methods for in silico design of NCEs. They have started with a dataset of ~1.6 million small molecules from the ChEMBL database[8] to train the DNL model. They filtered out stereochemistry, salts, undesirable atoms or groups, and SMILES string greater than 100 symbols. The DNL model and filtration method are explained in detail in their paper. Subsequently, they filtered artificial intelligence (AI) generated small molecules based on various physicochemical properties such as drug like ness[9] and synthetic accessibility[10]. Finally, these filtered small molecules were docked using AutoDock Vina[11] to the energy minimized 3CLpro structure (PDB ID: 6LU7) and ranked based on their virtual screening scores. They have docked 3960 molecules and obtained 1333 small molecules that have virtual screening scores below -7.0.

In the paper, Bung et.al. reported final high potential[7] (to qualify as drug candidates) 31 NCE molecules with virtual screening scores between -8.3 and -7.5. Out of these 31, they refer to 16 molecules which are similar to already FDA approved drugs darunavir, lopinavir, ritonavir, indinavir, saquinavir, and ASC09 and are currently in clinical trials for SARS-CoV-2. Moreover, reported virtual screening scores for their NCE molecules are better than the drugs in clinical trials. They have reported the rest 15 NCEs showed higher virtual screening score against 3CL protease of SARS-CoV2 than the other set of 16 molecules. The highest virtual screening score they reported is -9.1 and the highest Tanimoto coefficient[12] with the existing protease inhibitors among the top 15 molecules is 0.90.

While virtual screening tools are popular in use, the limitations of methods like docking with different variations in methodologies are well-established facts in the literature [13,14]. The target ligand docking often fails to produce or identify the right ligands (NCE) which could be the best possible bet i.e., high specificity towards the target. The failure is caused by multiple factors, e.g., proper sampling of the binding site, flexibility of the protein (change of conformation of protein due to binding of the ligand), non-existence of solvent (i.e., the solvent-ligand interaction) in the model and finally that results into the definition of the scoring function (missing entropic contributions). Therefore, often, docking is used for qualitative estimation of the chemical space of the target, and subsequently, medicinal chemists use their intuition (to design scaffolds) and synthesize large numbers of molecules (chemical library). This library of molecules is then tested in biological assays to screen further for better efficacy.

Therefore, here, we propose a much robust methodology to address some of the issues mentioned above and reduce the sub-set of NCEs for further synthesis and testing. Our methodology is to perform a large scale all-atom molecular simulation on the target-ligand complexes followed by enhanced free energy methods to identify the set of ligands with high specificity. In this method, we use molecular docking to select a set of chemical entities that shows significant interaction (high score) with the protein. These molecules then are subjected to molecular dynamics simulations with water as the explicit solvent. Solvation of molecules in water (or any other solvent) is critical for identifying right ligands that could bind at the binding site of the protein with high stability and not get 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 are 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. Also, defining a score associated with the binding for these ligands i.e., chemical entities is important to choose the best set of NCEs/potential drug candidates.

In this work, we have considered final molecules reported by Bung et. al.[7] for enhanced sampling at the binding site of the protein. We have rationally selected molecules that showed higher binding affinities toward the protein. Apart from this, 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 a clinical trial for COVID-19 (ClinicalTrials.gov Identifier: NCT04252274). The details of all ligand structures (in 2D and 3D optimized) structures are given in Table 1. The computational details which to the best of our knowledge are novel for selecting NCEs for SARS-CoV2 are described in section 2 of the paper. In section 3 we have provided the results along with discussions.

2. Computational Method

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

The geometry optimizations of all the ligands (listed in Table 1) are performed using a semi-empirical method at the PM6 level, followed by geometry optimization using density functional theory (DFT) with M06 functional and 6-311g (d,p) basis set. To account for the bulk solvent effects PCM method is used. Further, the partial atomic charges for the ligands are computed by fitting the electrostatic potential using CHELPG method as implemented in Gaussian09 code[15]. These charges are computed for the optimized structures using a single point calculation at the DFT with M06 functional with 6-311g (d,p) basis set and water as the solvent.

The virtual screening scores for binding are generated through docking to find out the affinity of all the ligands, with 3CL protease. In general, docking involves finding the optimal binding between protein and ligand. To get this optimal binding score the ligand conformational search is performed around the binding sites. Here, a genetic algorithm-based conformational search is employed to find the lowest energy conformation of the ligand. The ligand's conformational search is carried out by creating the 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[4]. The docking of ligands and protein is conducted using Autodock4[16] software. The docking is carried out using the Lamarckian genetic algorithm (LGA) and a total of 100 GA-LA hybrid runs are used to perform the conformational search for the ligand. Further, the lowest energy protein-ligand cluster is used to repeat the docking for twice, and the consistency of the results are combined to get the best score.

The lowest energy docked complexes, the protein-ligand systems obtained from docking, are used to perform MD simulations. The protein is modeled using CHARMM27 force-field[17] parameters. The CHARMM27 force-field is employed for all the ligands, and the force-field parameters are generated using SwissParam[18]. The ligand partial atomic charges are computed by fitting the electrostatic potential using CHELPG method[19] as implemented in Gaussian09 code. The protein-ligand systems are solvated in water and equilibrated using MD simulations at room temperature. At first, the systems are equilibrated using NVT ensemble at 300 K for 0.5 ns and extended to NPT ensemble at 300 K and 1 atm for another 1 ns. The temperature and the pressure during the simulations are maintained using velocity rescaling thermostat and Parrinello-Rahman barostat respectively. A time step of 2 fs is used to integrate the equation of motion and a non-bonded cut-off of 10 Å is used to perform the MD simulations. These simulations are used to understand the stability of the interaction of the ligand with respect to the protein binding site in explicit water. We have quantified the interactions between the amino acids in the binding pocket and the ligand using hydrogen bond analysis. All MD simulations are performed using GROMACS-5.1.4 simulation package [20,21]. Further, the equilibrated structure obtained from the 1ns MD simulations is used to perform the free energy analysis.

Since protein-ligand systems are complex in nature, exploring various important quantities like its thermodynamics, kinetics, microscopic description at the all-atom level, remain a challenge due to the length scales of the systems and also the time scales involved in the processes like dissociation or association[22,23,24] of the ligand from/to the protein binding pocket, etc. 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[25] (metaD) and it's variant well-tempered metadynamics[26] (wt-metaD) using Plumed 2.3.0[27] patched with MD engine GROMACS 5.1.4. 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, or else the system would have generally been trapped for a sufficiently long time. 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(σ)of 0.1 nm, bias-factor 15 and at temperature (T) 300 K. Once the system converges, the free energy F(s) (Eq. 1)[28] can be extracted by adding the deposited hills along the biased reaction coordinate (s). In a wt-metaD the amplitude of the bias is tuned such that the system converges smoothly. Here we used a tempering factor ΔT to tune the hills height and thus we achieved smooth convergence of the free energy landscape.

The wt-metaD simulations are started 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 here on the dissociation of ligands (Figure S2) in enhanced sampling simulations.

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

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

Since, a strong binding pose will be more stable and its root-mean square deviation (RMSD) will be lesser. Therefore, high RMSD values can be used as an indicator to the poor binding pose and higher RMSD meaning a stable binding pose. Hence, to find out the top binding ligands according to their binding specificity, we further performed wt-metaD simulations taking 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 collective variable of center of mass distance (d). We performed 20 independent wt-metaD simulations with RMSD as the collective variable for each ligand with each run extending up to 2 ns. As described here we have performed several independent short metadynamics simulations with RSMD as the collective variable. So, this is similar to doing a much longer unbiased MD simulation where the starting structure of ligand-protein complex could overcome the local barriers and reach global minimum. Therefore, these independent trajectories were used to evaluate the stability that are 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 3. The docking trend is found to be in qualitative agreement with the results obtained from Autodock Vina[11]. However, the absolute scores obtained from our simulations are different from that reported using Autodock Vina, due to the differences in force-fields used in Autodock 4 and Vina. The lowest energy docked complexes are examined to find the ligand location with respect to the protein binding site. The PI-06 ligand is found to exhibit high binding affinity with the protein. Among the 17 ligands, PI-04, PI-06, PI-10 and PI-12 ligands are found to exhibit higher binding scores with the protein. All ligands considered in this study are found to be in the binding pocket and interacting with HIS41 and CYS 148. The best docking pose of 4 ligands obtained from docking are shown Figure 1. The ligand position clearly indicates that the ligand prefers to stay at the binding site of the protein.

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

In case of docking, the protein is considered to be rigid and 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 are 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 are used as initial configurations to perform the MD simulations. The docked complex 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 are 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 are found less than 0.2 nm for all the protein-ligand complexes. This clearly shows that the protein-ligand systems are stable and the ligands prefer to be in the binding pocket.

Further to elucidate the main interactions of the ligand with the protein amino acids in the binding pocket hydrogen bonding scenarios are analyzed. The involved interacting groups of protein and ligands through hydrogen bonds are listed in Table 2. Almost all the ligands are found interacting with the -NH2 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 find out the contribution from all possible interactions one needs to explore the complete free energy surface associated with the 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 enough as conformations might get stuck in local minima. Therefore, enhanced sampling of ligand binding and change 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 is used as the starting configuration in the enhanced wt-metaD simulations. The average free energy of dissociation for all the ligands obtained from wt-metD simulations is reported in Table 3 and the corresponding free energy profiles are shown in Figure 2 (a). Here, the average free energy values are obtained from 20 independent dissociation simulations for each ligand to get better sampling and statistically reliable results. The free energy values are found to be in the range of from -22.7 to -4.8 kJ/mol for all the ligands (see Table 3). The PI-06, PI-08, PI-11 and PI-14 ligands are found to exhibit higher energy barriers in the same order compared to the other ligands. The maximum free energy of association is -22.7 kJ/mol, which is observed for PI-06 ligand. These four ligands clearly outperformed all other ligands. However, PI-06 is the best among these four with -8 kJ/mol lower free energy from the second best PI-14. To better understand the free energy behavior, the profiles for these four ligands are separately shown in Figure 2 (b). 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 i.e., at the binding site. This behavior represents multiple interactions between the ligands and the 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, where 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) is superimposed with the complex obtained from docking. Figure 3 presents the superimposed structures of free energy surface (FES) minimum configuration and the docked complex for the PI-06 ligand. In the wt-metD, the ligand position is found to be in the binding pocket marginally away from the residues HIS41 and CYS148. The ligand in the docked pose is shown as red sticks, whereas the FES minimum pose is shown as blue sticks. Further, to assess binding landscape of ligand-protein and to validate the binding pose from the FES, the RMSD based free energy is computed.

To understand the FES of binding poses of the ligands in detail we looked into the FES as a function of RMSD (collective variable) as described in the computational method section. For the poorly bound ligands, it is expected the RMSD (with respect to the lowest energy binding structure obtained from FES described in Figure 2) will be higher in comparison to 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 wt-metaD simulations and RSMD as the collective variable. In Figure 4a we present the free energy as a function of aligned RMSD for all the ligands. Here each FES is averaged over 20 independent runs. From FES of Figure 4 (a) it is evident there are stable conformations for all the ligands below 0.2 nm of RMSD. So, there is global minimum for all the ligands close to the starting conformation and almost no other local or global minimum are observed. However, there is some existence of metastable states after 0.3 nm of RMSD. Further, to quantify the binding of ligands to the protein we have computed the probability of the ligand-protein complex within 0.2 nm of RMSD from all the trajectories we obtained from FES calculation with RMSD as collective variable. The trajectories (RMSD as a function of time) for four ligand-protein complexes are shown in Figure 5 to elucidate the stability of the ligands in the binding site. In Figure 4 (b) we have depicted the distribution of the probability of RMSD for these four ligands (see Figure S3 for all the ligands). For ligands PI-06, PI-08 and PI-11 we observed sharp peaks for distribution of probability values for RMSD below 0.2 nm and for PI-14 it is slightly lesser. It signifies PI-06, PI-08, PI-11 and PI-14 ligands are strongly bound at the binding site of protein. The probability of the RMSD value below 0.2nm could, therefore, be an indicator of binding. Higher the probability, stronger the bonding will be. These values are reported in Table 3 along with the free energy change for all the ligands.

In a similar line to find the stability of ligand-protein complex we have calculated the average RMSD (s) from all the independent biased trajectories using following equation

$$\langle s \rangle = \frac{\int ds\, s\, e^{-(F(s)/k_BT)}}{\int ds\, e^{-(F(s)/k_BT)}} \qquad (2)$$

Here F(s) is the energy associated with the RMSD. A higher estimate of the average or thermodynamically preferred RMSD can then be considered an indication for poor instability of the complex. So higher the value (score) lower the stability and vice-a-versa. All these quantitative estimations of the stability (score) for each ligand using Eq. 2 are reported in Table 3.

We have calculated two types of scores (probability of RMSD below 0.2 nm and average RSMD as per Eq. 2) from the biased trajectories that are obtained from metadynamics simulations. Figure 6 shows the correlation of these two types of scores with the FESs for all the ligands. It is evident that for the ligands with 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 the ligands that show higher stability than others. We have used docking structures with similar docking scores and well separated 4 ligands that bind the 3CLpro with much higher affinity. These ligands are in the order of PI-06 > PI-14> PI-11>PI-08 according to the free energy barrier and average RSMD. However, if we consider the probability of RMSD values less than 0.2nm, then the resulting order is PI-06>PI-08>PI-11>PI-14. From all these scores. it is evident that PI-06 clearly 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 the free energy profile for PI-06 has much higher energy of solvation i.e., at dissociated state than others. And for all the ligands there are local minima present along with one global minimum in the free energy landscape. To understand this feature, we looked into the dissociation trajectory for PI-06 ligand (see Figure 7). We showed the 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 escapes 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 observed in Figure 7.

4. Conclusion

In this paper we have performed large scale all-atom molecular dynamics simulations with enhanced sampling for ligands that binds to the 3CL protease of SARS-CoV2. These calculations are robust and are modelled similar to the experimental system by incorporating explicit solvent molecules and considering all-atom molecular models and interactions. We have considered a set of 17 ligands with lower virtual screening score (for 3CLpro of SARS-CoV2) and high Tanimoto score with respect to known HIV inhibitor e.g., already FDA approved drugs darunavir, lopinavir, ritonavir, indinavir, saquinavir, and ASC09. Our method could distinctively isolate these 17 ligands into 4 possible NCEs and could even identify the best compound with very high confidence. Upon fruitful synthesis and testing, these four NCEs is expected have much higher probability of success in clinic trials.

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

Acknowledgement

JKS acknowledges the support of SERB for partial funding. This work is also supported by the CSR funding of GST IN.

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 RSMD of heavy atoms were considered as collective variables.

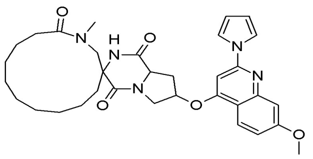
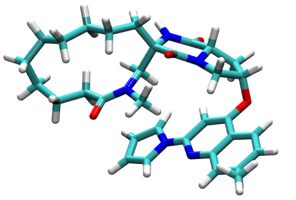
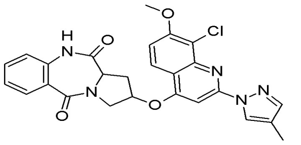
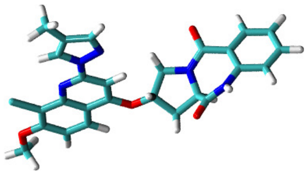
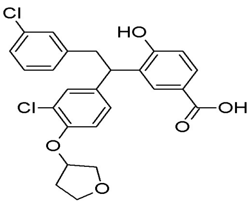
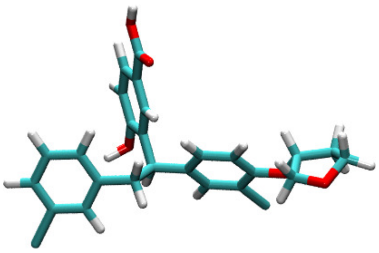
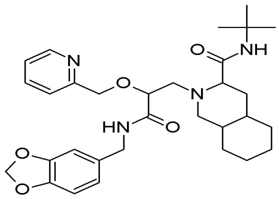
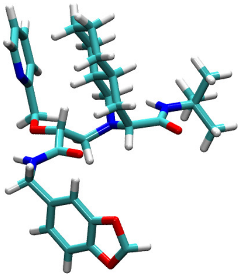
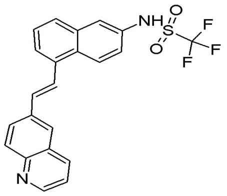
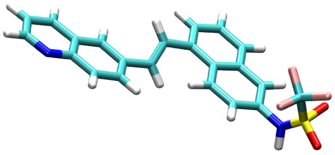
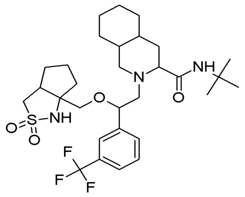
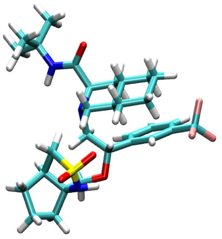
References

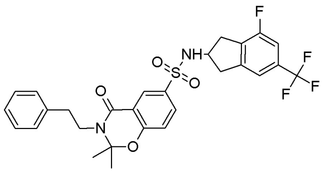
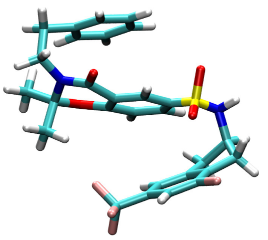
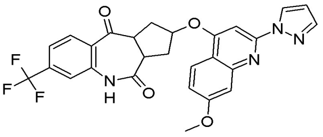
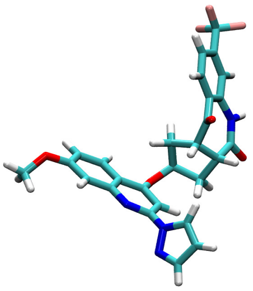
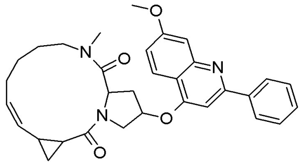
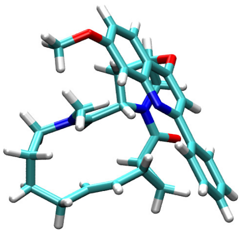
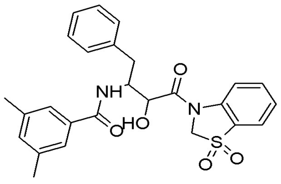
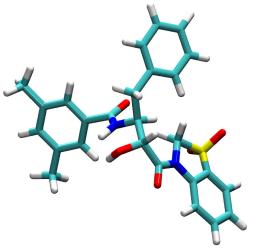
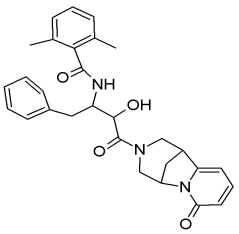
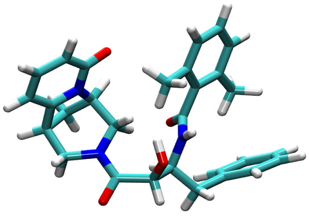
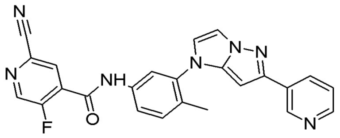
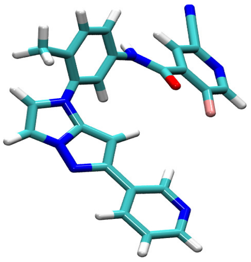
1. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
2. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
3. Zumla, A., Chan, J. F. W., Azhar, E. I., Hui, D. S. C. & Yuen, K. Y. Coronaviruses-drug discovery and therapeutic options. *Nature Reviews Drug Discovery* **15**, 327–347 (2016).
4. Jin, Z. *et al.* Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* (2020). doi:10.1038/s41586-020-2223-y
5. Liu, X. & Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. doi:10.1101/2020.01.29.924100
6. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
7. 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 Navneet. *Definitions* **2003**, (2020).
8. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, (2012).
9. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
10. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **1**, (2009).
11. 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.* NA-NA (2009). doi:10.1002/jcc.21334

12. Lipkus A H. A proof of the triangle inequality for the Tanimoto distance. *J. Math. Chem.* **26**, 263–265 (1999).
13. Hassan Baig, M. *et al.* Computer Aided Drug Design: Success and Limitations. *Curr. Pharm. Des.* **22**, 572–581 (2016).
14. Pons, C., Grosdidier, S., Solernou, A., Pérez-Cano, L. & Fernández-Recio, J. Present and future chanllenges and limitations in protein-Protein clocking. *Proteins Struct. Funct. Bioinforma.* **78**, 95–108 (2010).
15. 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).
16. 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).
17. 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 ALL-ATOM EMPIRICAL FORCE FIELD FOR NUCLEIC ACIDS. I. *J. Comput. Chem.* **21**, 86–104 (2000).
18. 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).
19. 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).
20. Pronk, S. *et al.* GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
21. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
22. 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. *bioRxiv* 558601 (2019). doi:10.1101/558601
23. 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).
24. 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).
25. Laio, A. & Parrinello, M. *Escaping free-energy minima*.
26. Barducci, A., Bussi, G. & Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. (2008). doi:10.1103/PhysRevLett.100.020603
27. Bonomi, M. *et al.* PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180**, 1961–1972 (2009).
28. 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

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		

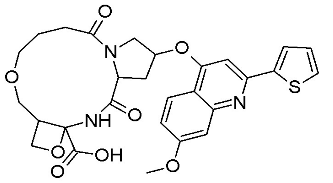
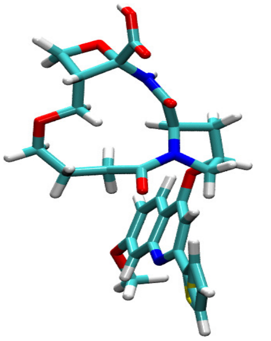
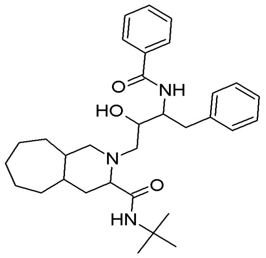
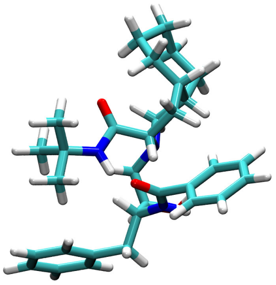
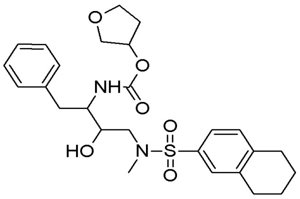
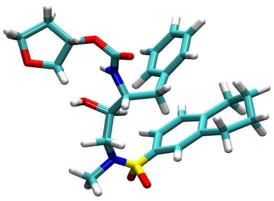
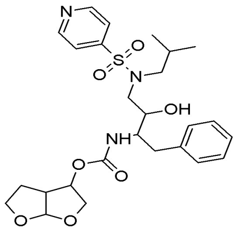
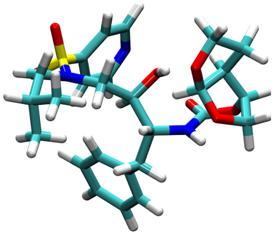
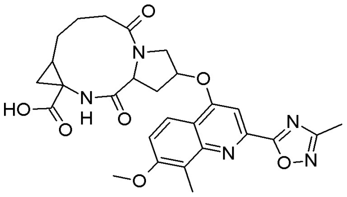
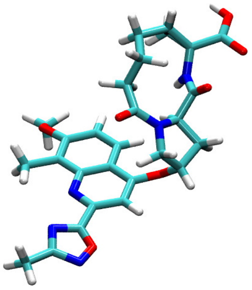
PI-13		
PI-14		
PI-15		
PI-16		
PI-17		

Table 2 Interaction residues and functional groups the protein with the ligands.

Ligand Code name	Residue	Residue-ligand interacting groups
PI-01	THR26	OH-C=O

PI-02	THR24
	ASN142
	GLN189
		OH-C=O
		NH2-C=O
		NH2-C=O

PI-03	HIS163 HIS16	NH2-C=0 NH2-OH
PI-04	GLY143 SER144 ASN142 GLN189	NH2-C=0 NH2-C=0 O-OH N-OH
PI-05	ASN142	NH2-N (-SO2)

PI-06	THR26
	ASN142
	GLY143
	CYS148
	GLY143
		NH2-O
		NH2-C=O; O-NH2
		NH2-N(-SO2)
		NH2-C=O
		NH2-O

PI-07	HIS41 ASN142 GLU166	NH2-C=O NH2-O O-NH
PI-08	THR26 ASN119 ASN142 GLY143 LEU27	NH2-O=C NH2-O=C NH2-O NH2-N O-NH (-NC=O)

PI-09	THR26
	SER46
	HIS143
		NH2-O=C
		OH-O=C
		NH2-O

PI-10	THR26 SER46 ASN142 GLN189 SER46	NH2-O=C OH-O=C NH2-OH NH2-N NH2-OH
PI-11	HIS41 ASN142	NH2-O=C NH2-OH

	GLN189	NH2-O=C; NH2-OH
PI-12	HIS41	NH2-O=C
PI-13	ASN142 GLY143 GLU166 CYS148	NH2-O=C NH2-O=C NH2-OH NH2-O=C
PI-14	ASN142 GLU166 GLN189	NH2-O=C NH2-O=C O-OH
PI-15	GLY143 GLU166 CYS148	NH2-O NH2-O NH2-O
PI-16	ASN142 GLU166	NH2-N NH2-O=C
PI-17	ASN142 GLY143 GLN189 SER46	NH2-O=C NH2-O=C NH2-N; NH2-O; NH2-O=C O-NH

Table 3 The docking score, free energies for dissociation, average RMSD values, probabilities (for RMSD < 0.2 nm) for all ligands.

Ligand code name	**Docking score**	**Free Energy (kJ/mol)**	**Average RMSD (nm) as per Eq. 2**	**Probability of RMSD (RMSD < 0.2 nm)**
PI-01	-8.51	-6.49 (2.55)	0.507	0.361
PI-02	-9.13	-7.6 (1.89)	0.450	0.254
PI-03	-9.43	-5.7 (2.5)	0.313	0.472
PI-04	-11.56	-8.1 (1.64)	0.222	0.594
PI-05	-10.85	-7.26 (2.4)	0.250	0.443
PI-06	-11.92	-22.7 (1.8)	0.152	0.842
PI-07	-10.40	-6.05 (2.06)	0.350	0.418
PI-08	-9.50	-11.27 (1.72)	0.184	0.722
PI-09	-10.30	-5.37 (2.16)	0.345	0.453
PI-10	-11.64	-5.2 (1.52)	0.389	0.317
PI-11	-10.54	-14.8 (1.64)	0.177	0.719
PI-12	-10.94	-10.89 (1.95)	0.237	0.522
PI-13	-10.22	-10.5 (1.7)	0.153	0.666
PI-14	-10.64	-14.86 (1.40)	0.160	0.642
PI-15	-9.68	-6.39 (1.67)	0.276	0.384
PI-16	-9.52	-10.6 (1.91)	0.170	0.604
PI-17	-9.04	-4.8 (2.15)	0.309	0.156

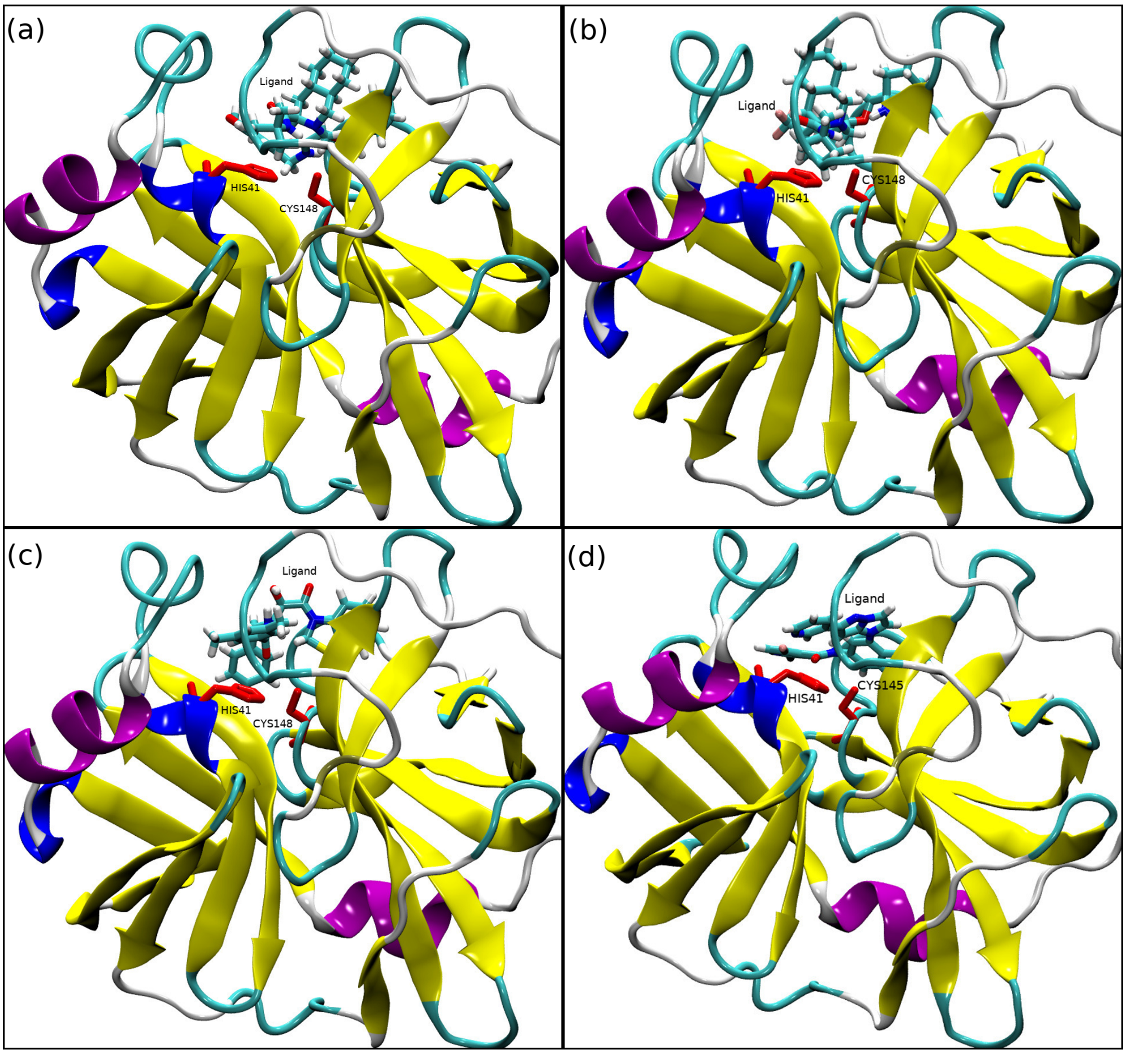


Figure 1 The best docking poses of lowest binding energy 4 ligands with protein (a) PI-04 (b) PI-06 (c) PI-10 and (d) PI-12 are shown here. The active site of the protein (HIS41 and CYS148) is shown as red sticks.

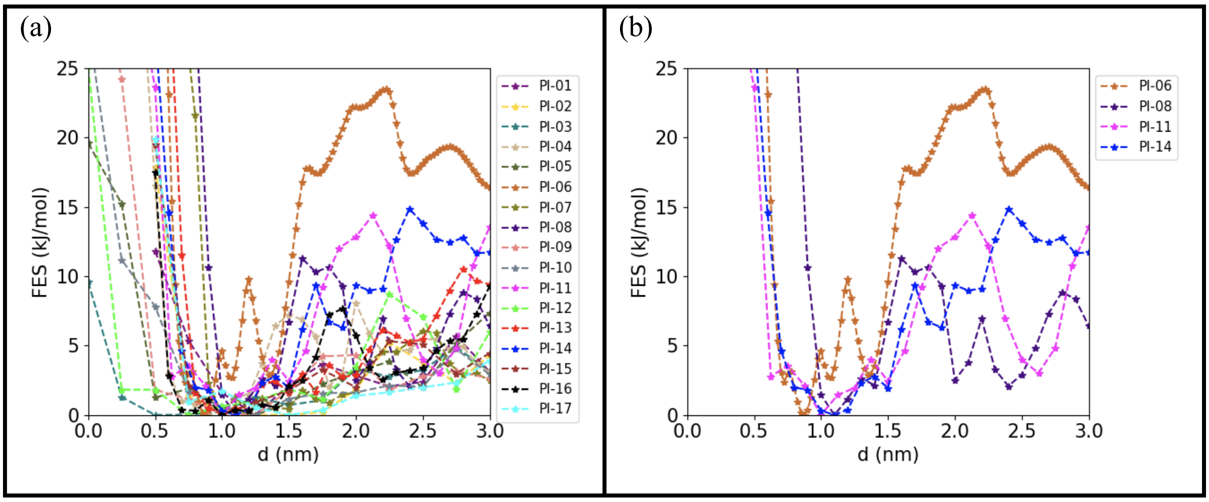


Figure 2: Average free energy with center of mass - center of mass distance (d) for dissociation of the ligands from the protein binding pocket. (a) Free energies for all the ligands. (b) Free energies for the top four ligands. For each ligand the errors in free energies are reported in Table 3.

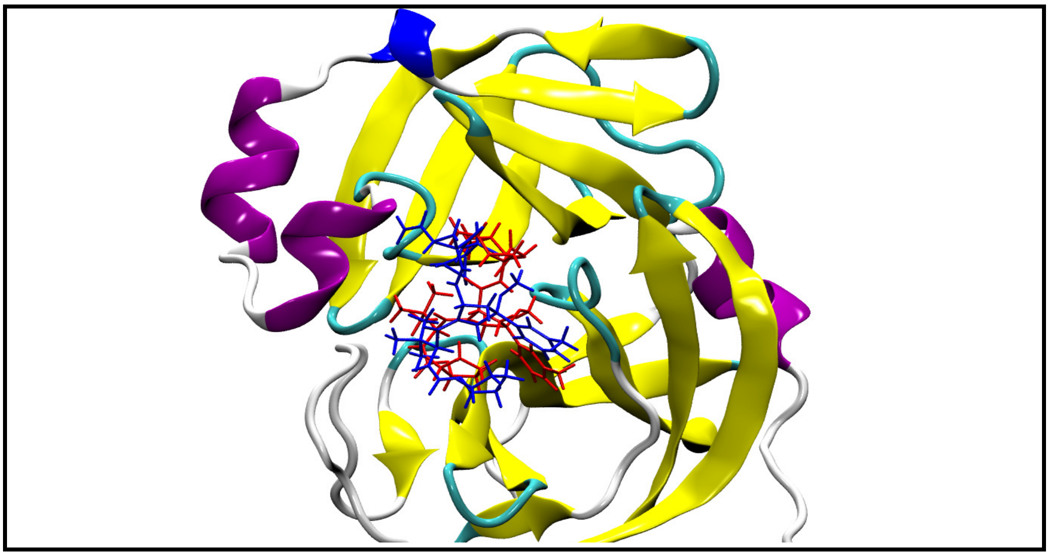


Figure 3: The zoom-in view for the superimposed structure of the PI-06 ligand docked pose and stable pose from the free energy minima. The Ligand in the docked pose is shown as red sticks and that of the free energy minimum structure is shown as blue sticks.

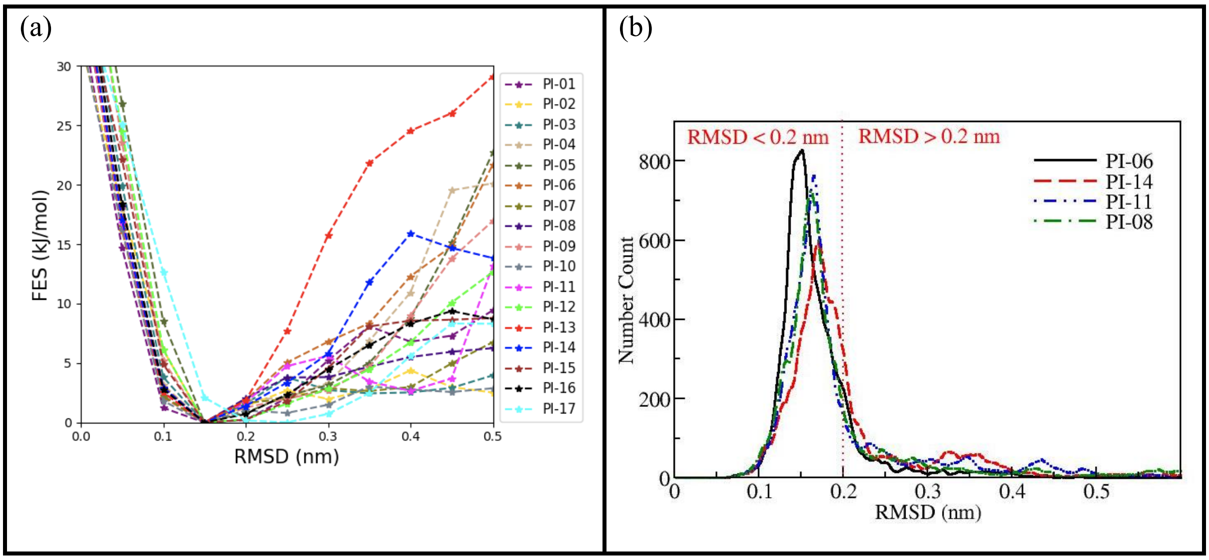


Figure 4: (a) Average free energy with aligned RMSD for all ligands. (b) The number count distributions for the probability to find a system within 0.2 nm of RMSD.

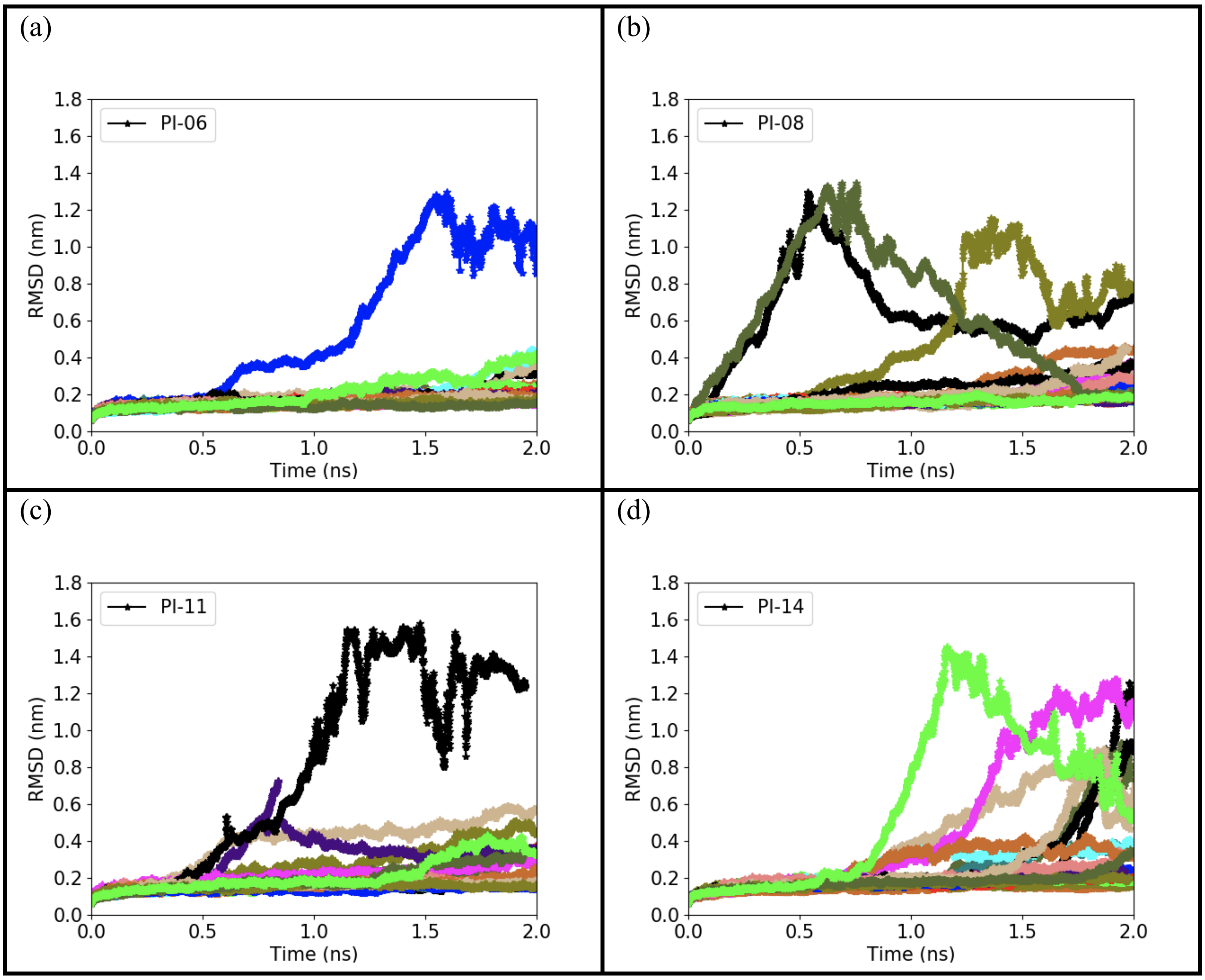


Figure 5: Time evolution of the RMSD for top four ligands, (a) PI-06, (b) PI-08, (c) PI-11 and (d) PI-14. In each plot we show RMSD from all independent runs.

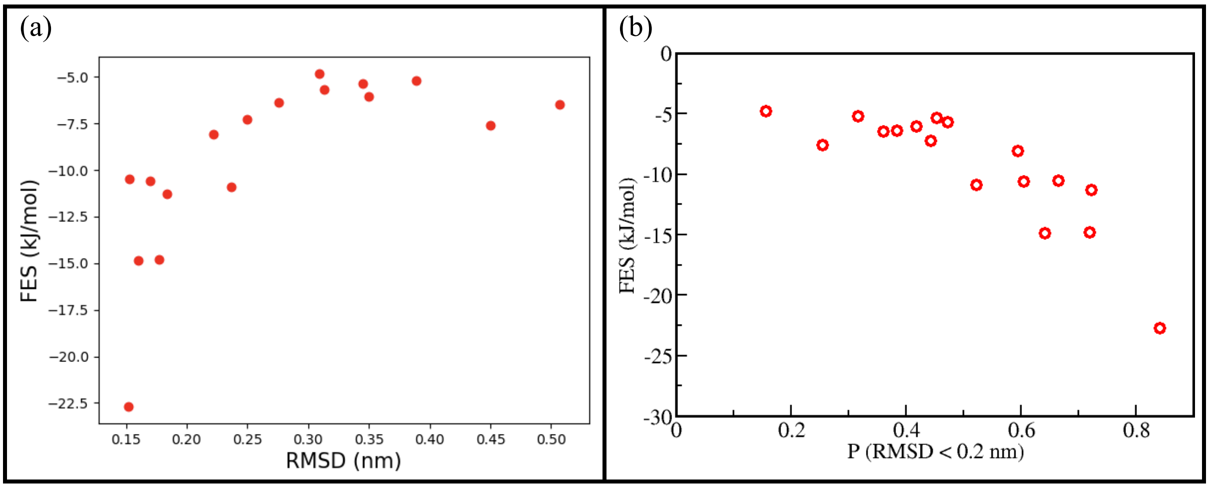


Figure 6 Average free energy of protein-ligand as a function of (a) average RMSD as per Eq. 2 and (b) probability of RMSD, here probability is calculated for the ligands which shows less than 0.2 nm RMSD.

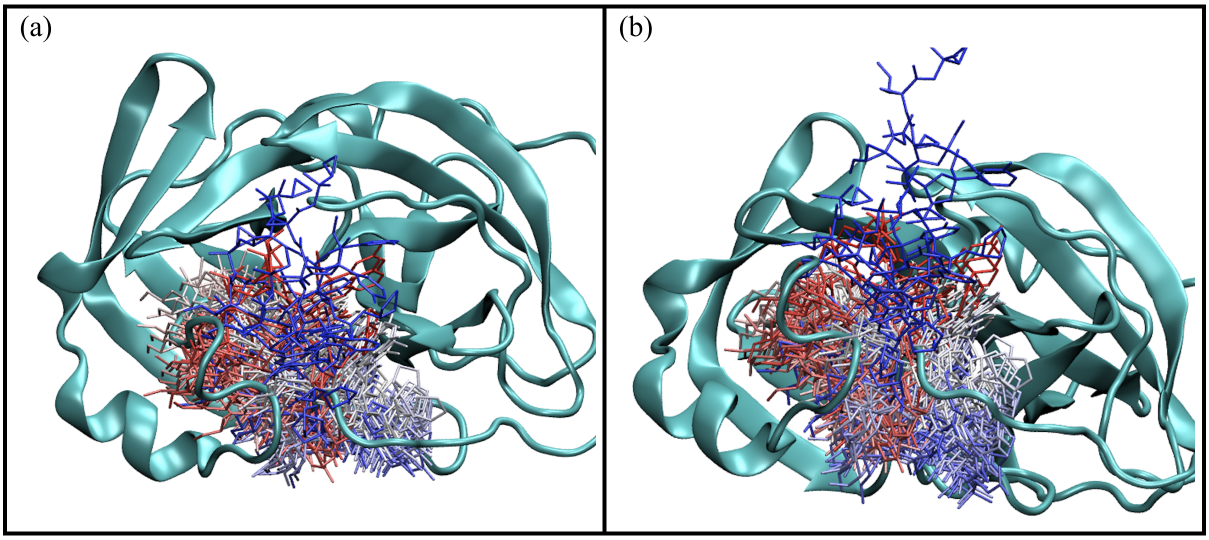


Figure 7: The trajectory of the ligand dissociation from the protein binding pocket. Two different views (a) left, and (b) right, show the full dissociation of the ligand from the binding pocket, interaction of the ligand with the protein backbone in the vicinity of the binding pocket for the ligand 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).