Extensive crystallographic fragment-based approach to design SARS CoV2 3CLpro main protease inhibitors and related metadata

Sarfraz Ahmad[1]*, Iskandar Abdullah[1], Lee Yean Kee[1], Mamoona Nazir[2], Muhammad Usman Mirza[3], John F. Trant[3], Noorsaadah Binti Abd Rahman[1].

[1]Drug Design Development Research Group, Department of Chemistry, Faculty of Science, Universiti Malaya, Kuala Lumpur 50603, Malaysia.

[2]Aiman Institute of Medical Science, Lahore, Pakistan.

[3]Department of Chemistry and Biochemistry, University of Windsor, Windsor, ON N9B 3P4, Canada.

*Corresponding author Email: sarfraz.ahmad@um.edu.my

Abstract

3CLpro is a vital protein for the SARS-CoV-2 replications and its inhibition using small molecules is a *bona fide* approach used to develop new drugs against the virus. In this study, a comprehensive crystallography-guided fragment-based drug discovery approach was employed to design new inhibitors for SARS-CoV-2 3CLpro. Protein Data Bank was explored to find small molecules cocrystallized with SARS-CoV-2 3CLpro. The fragments sitting in the binding pocket (87) were interactively coupled using various linkers with the intention to get molecules having the same orientation as those of the constituting fragments. In total, 1251 couples were prepared and converted to maximum possible stereoisomers using LigPrep for screening using Glide (standard precision and extra precision), AutoDock Vina, and Prime MMGBSA. Top 22 hits having conformations similar to their cocrystallized fragments were selected for MD simulation on Desmond. MD simulation suggested that 15 hits had conformations very close to their constituting fragments. Results indicated that these hits were computationally reliable and could be considered for further development. This suggests that the study could provide a benchmark starting point for the further design of SARS-CoV-2 3CLpro inhibitors with improved binding (data provided).

Keywords: Fragment-based drug discovery; coronavirus COVID-19; SARS-CoV-2; main protease Mpro 3CLpro; X-ray crystal structure.

Introduction

What is COVID-19:

Coronavirus disease 19 (COVID-19) is an ongoing pandemic announced by the World Health Organisation in March 2020 and a worldwide public health emergency. It is caused by a novel coronavirus and is termed as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1, 2]. The first recorded case of unknown pneumonia cases linked to a seafood market emerged in late December 2019 in Wuhan, Hubei province, China [3]. In the early stages of the research, it is hypothesized that the virus originated from bats. Later, claims towards pangolin as hosts might have passed the virus to humans [4].

Cases fatalities regions:

The world has seen far-reaching consequences of COVID-19 pandemic that has affected 216 countries globally [5, 6]. Over 64 million individuals were infected, claiming around 1.5 million lives until early December 2020 [6]. Geriatric and people possessing a history of chronic illness or compromised immune system are more likely to be severely affected [7, 8]. Its transmission can quickly occur through respiratory droplets via direct or indirect contact with the mucous membranes of eyes, mouth, and nose [9].

Socio-economic effects:

COVID-19 has transformed lives globally, giving a devastating impact on all aspects of human life, particularly the economy, health, and education. It has struck fear into our daily social life even to our neighbours and family members [10]. Along with the paralysis of socio-economic dynamics and vibrance of everyday life, mental health concerns associated with the pandemic are rising due to lockdowns, movement control order, social distancing, safety precautions, loss of income and the loss of lives.

Treatment strategies:

This virus's expeditious infection rate created global havoc triggering global efforts for the urgent identification of effective and safe vaccines and antiviral drug treatments [11-14]—the measures initiated in two fronts including vaccine development and antiviral drug therapy.

Vaccines:

Vaccines are considered the frontline strategy to curb the spread of viruses. But there is no successful example of vaccine development known in case of previously known coronaviruses. In a pursuit to produce the first COVID-19 vaccine, pharmaceutical companies such as Pfizer, Moderna, Astrazeneca and BioNtech, and research groups of China and Russia are reporting successful development of vaccines that could be considered a frontier against the pandemic.

Despite, there are some caveats while considering vaccines a putative solution to the problem. Vaccines are usually questioned for their efficacy and high cost. Furthermore, for how long the immunization period would last with the newly developed vaccines is still a question to be answered for these newly developed vaccines. Vaccines are not usually easy to handle due to their temperature-induced efficacy reduction. Personal compliance is also a concern, particularly when it is parenterally administered.

Furthermore, a vaccine is effective before the infection, and after the onset of the disease, antiviral drug treatment is mandatory and inevitable. As the newly developed vaccines are not 100% effective, so a drug is necessary for people who will not benefit from the vaccine.

Molecular mechanism:

Structurally, the coronaviruses are enveloped, single-stranded positive-sense RNA viruses with the largest documented genome size that could vary between 26 to 32 kb for different coronavirus types. They have a 5'-cap and a 3'- polyadenylate tail containing 6-12 open reading frames (ORFs) [15]. The first ORF (ORF 1a/b) measures for about two-thirds of the genome length and can undergo direct translation to yield two polyproteins, pp1a and pp1ab, according to an a-1 frameshift between ORF1a and ORF1b [16, 17]. These polyproteins are further cleaved into 16 functional non-structural proteins (nsps) by 3CLpro [18, 19].

The coronavirus 3CLpro is a cysteine protease with three domains made from approximately 300 amino acids [20, 21]. An active form of 3CLpro is required to appear as a homodimer with two promoters within [22, 23]. A non-classical catalytic dyad (Cys145, His41 residues) is positioned between domain I and II [24]. It has the capability to precisely recognize 11 cleavage sites of nsp4 to nsp16 [24]. In addition, it was found to exhibit self-hydrolytic cleavage activity [25, 26]. The nsp4-to-nsp16 protein fragments resulting from 3CLpro cleavage are essential for viral lifecycle processes, such as genome replication, transcription, protein translation, cleavage, modification, and nucleic acid synthesis [20, 27]. Since proteins analogues to 3CLpro are not found in humans, 3CLpro turns out to be an ideal specific antiviral target [28]. Moreover, a copious number of efforts are underway to develop anti-coronavirus drugs using 3CLpro as a target.

Hypothesis

Apart from protection measures, vaccines and antiviral drugs are available approaches that could be exploited to address human pathogenic viruses. A need for new strategies to deal with these pathogens would stay relevant due to the frequent mutation rates of these types of virus. A conventional *de novo* drug discovery pipeline takes several years for completion. Different relatively quicker approaches employed to develop SARS-CoV-2 3CLpro inhibitors include drug repurposing, structure-based drug design, and fragment-based drug design. To further contribute to the global efforts, the study could provide a comprehensive starting point for the rational fragment-based drug discovery for SARS-Cov-2 3CLpro inhibitors using all the available crystal data on PDB. A large body of metadata developed in this study is also provided to benefit further development.

Results and discussion

Binding site topology

Based on interactions, 23 residues were found interacting with ligands out of which 9 residues were presenting only one interaction. The binding site was found predominantly composed of polar amino acids. Leu27, Met49, and Met165 were among prominent lipophilic residues offering 4, 3, and 13 hydrophobic interactions, respectively. Gly143, Ser144, and Cys145 were found making H-bonds with 52, 32, and 49 ligands, respectively. The highest diversity of interactions was observed for His41, which presented 25 π-stacking, eight π-cationic, 5 H-bond, three hydrophobic, three water bridges, and one salt bridge.

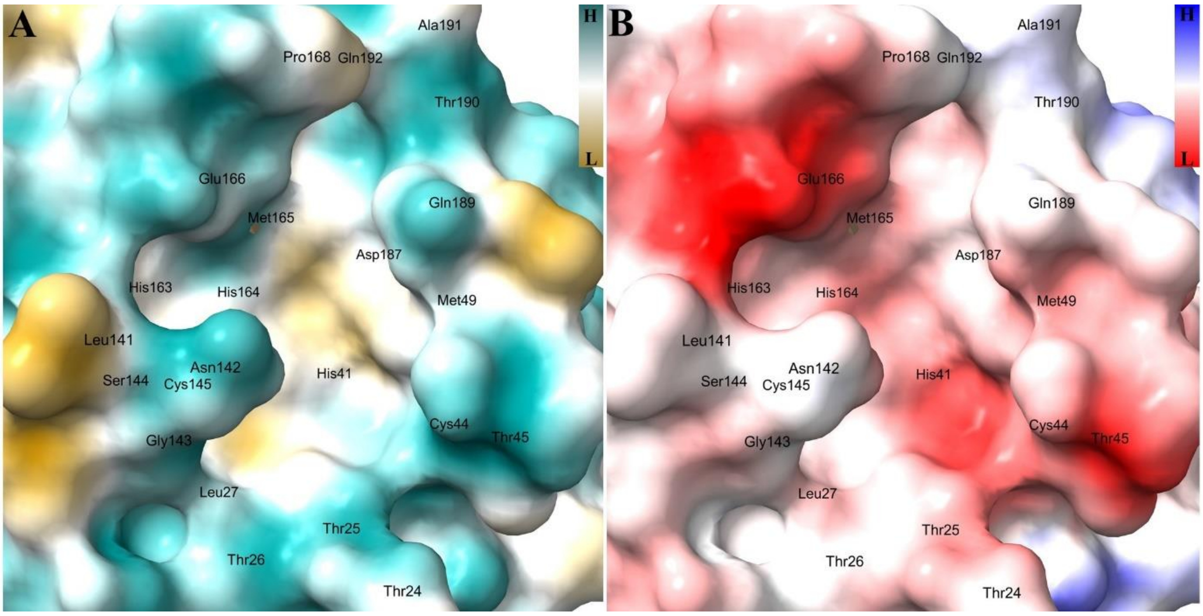

Figure 1: The interacting binding site residues of 3CLpro with selected 87 ligands (PDB ID 5R83): (A) atomic lipophilicity based molecular lipophilicity potential (B) amino acid lipophilicity based molecular lipophilicity potential, the Kyte-Doolittle scale. H and L in the top right of both images represent hydrophilicity and lipophilicity, respectively. The images were generated using UCSF ChimeraX (v 1.1).

Fragments

The selected fragment complexes were meticulously analyzed for ligand-protein interactions. A bar graph was plotted to elaborate on the frequency of binding site residues across 87 selected complexes (**Figure 2**). The complexes were keenly analyzed to categorize ligands based on their occupancy and ligand-amino acid interactions in the binding site, and the ligands were divided into eight sets **(Figure 3**).

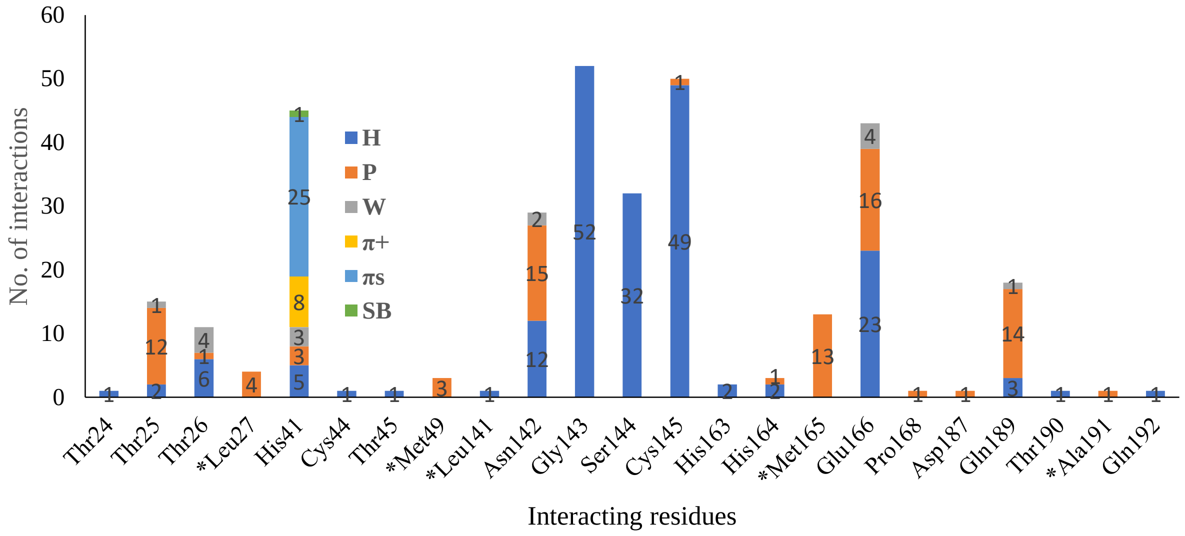

Figure 2: Number and type of interactions presented by the 3CLpro binding site amino acid residues with 87 selected fragments in terms of H (H-bond), P (hydrophobic interaction), W (water bridge), πs (π-stacking), π+ (π-cationic), and SB (salt bridge), while * represents lipophilic amino acids.

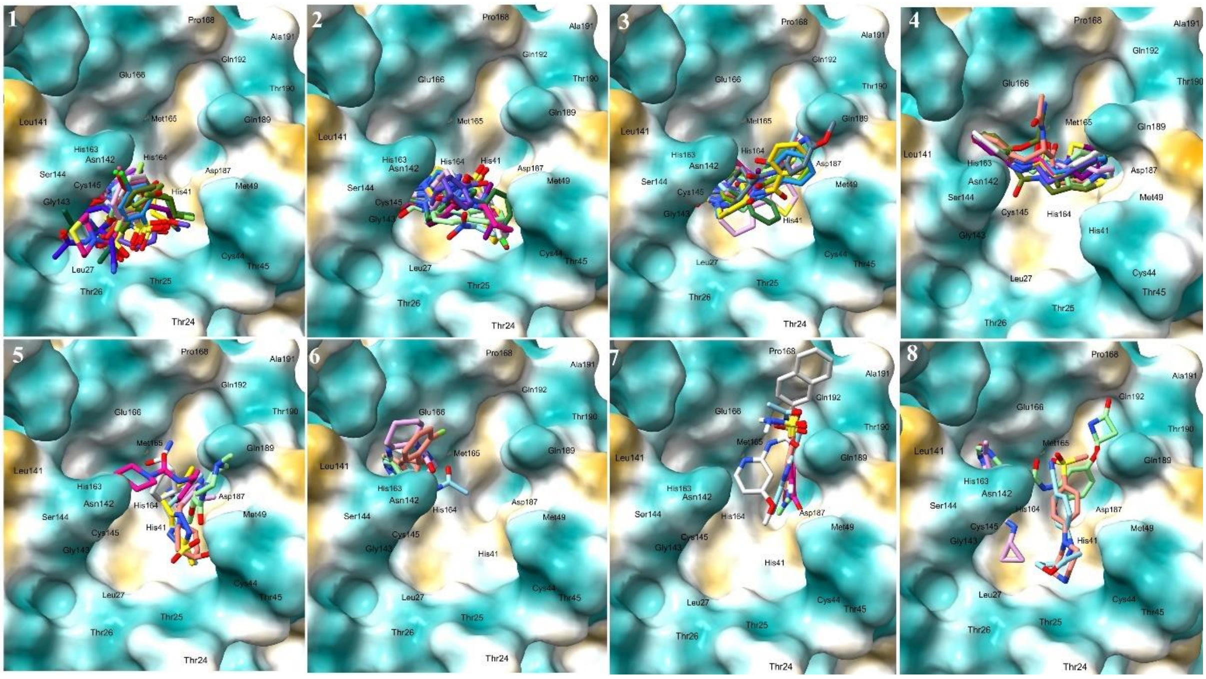

Figure 3: Pictorial representation of eight sets of 87 cocrystallized fragments in the binding site of 3CLpro based on their positions.

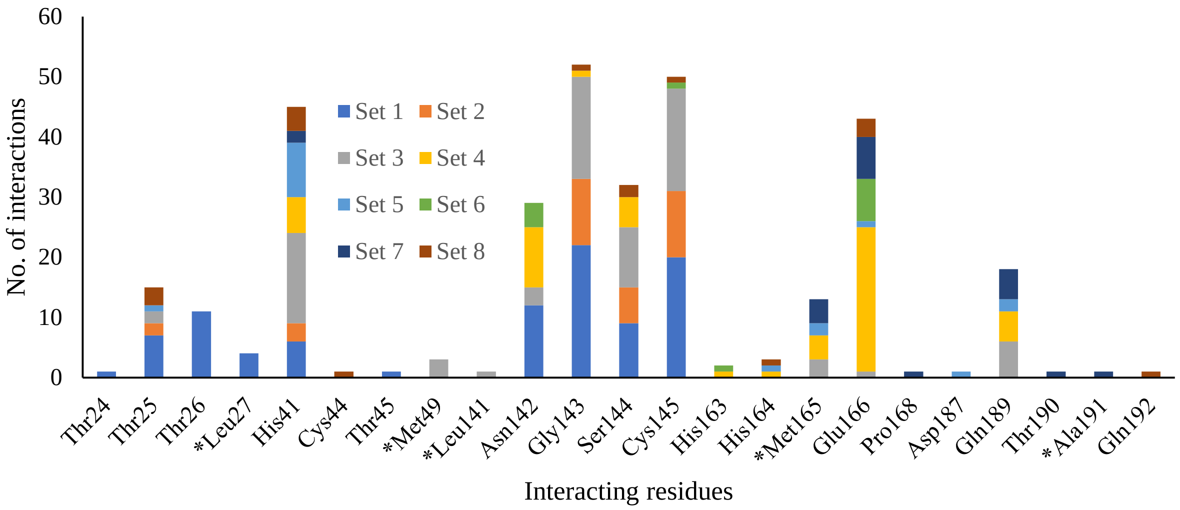

Figure 4: Graphical representation of eight sets of cocrystallized fragments based on their interactions with the binding site residues of 3CLpro. Lipophilic residues are labelled with *.

Set 1 is the largest category containing 24 compounds. These ligands were found residing in the left bottom of the binding pocket and were interacting with Thr25, Thr26, Leu27, His41, Asn142, Gly143, Ser144, and Cys145 (**Figure 4**). Set 2 consists of 11 compounds, residing slightly higher in the binding pocket compared to the compounds of set 1. Predominantly, these compounds were interacting with Gly143, Ser144, and Cys145. Compared to set 1, these compounds were not interacting with Thr26 and Asn142. Set 3 consists of 17 compounds. The primary difference between set 2 and set 3 is interaction with His41, where 15 compounds interacted with central His41. In this set, the ligand T8M was present in two complexes 5RFW and 5RHA. The interactions in the both complexes were the same except an H-bond with Ser144in 5RFW, which was replaced with a π-stacking interaction with His41 in 5RHA. Set 4 consists of 13 compounds. Here, Glu166 was interacting with most of the ligands via both H-bond and hydrophobic interactions. Set 5 contains eight compounds, which primarily reside in the centre of the pocket and interact with His41. Set 6 contains only five compounds uniquely residing in the left gorge of the binding pocket constituting Asn142 and Glu166. Set 7 carries five compounds. Set 8 carries only four compounds exhibiting diverse occupancies in the binding pocket. NTG and US7 share a similar binding site orientation while U0P and UGD occupy bottom left and top left gorges, respectively.

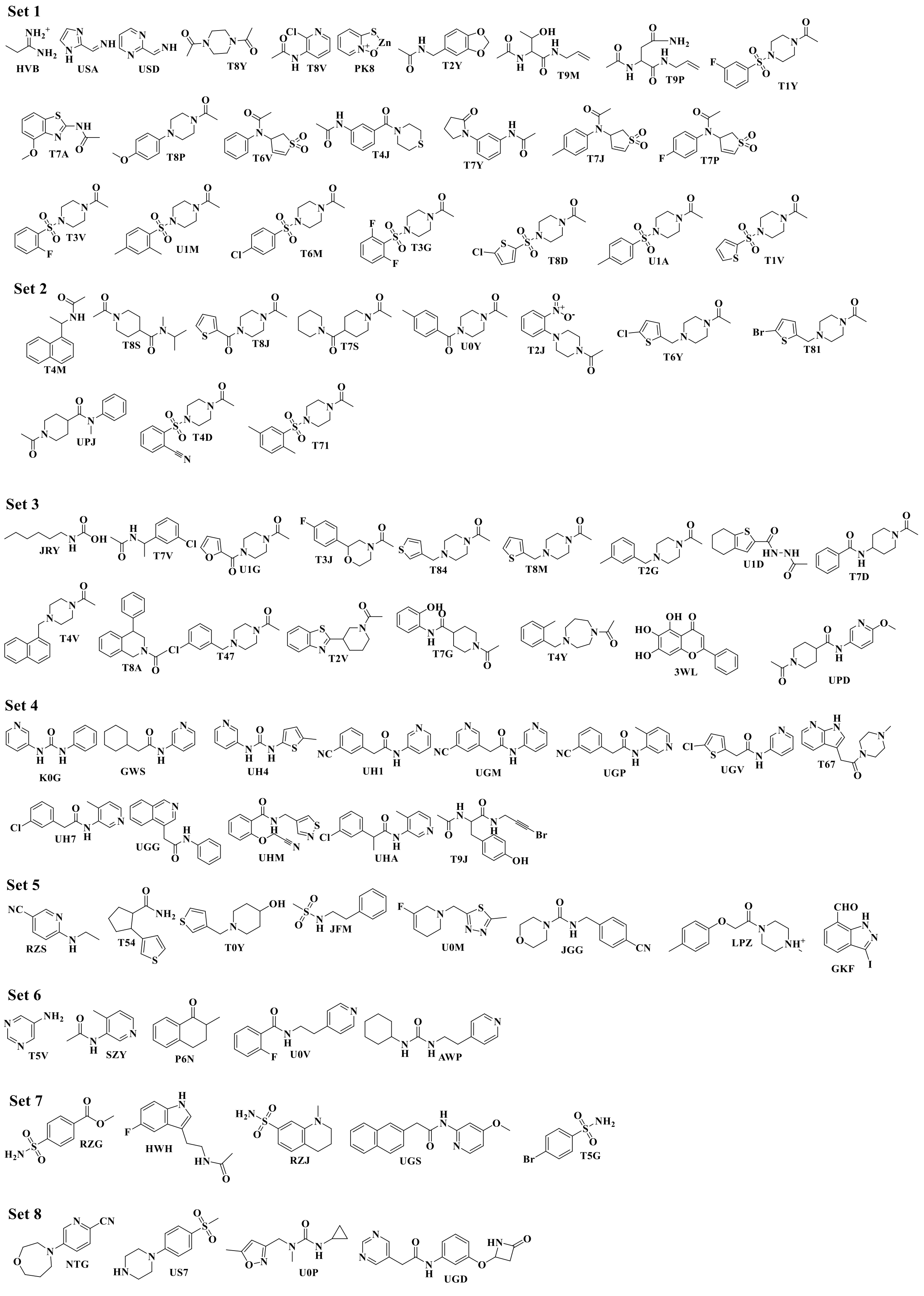

Figure 5: Structures of 87 3CLpro complexed cocrystallized fragments divided into eight sets based on their binding site positions.

Fragment coupling and linker selection

Although, fragment-based drug design is well-accepted and widely employed in various drug discovery pipelines. Several fragment libraries are available in online databases with large bodies of drug-like fragments. These fragments serve as a starting point to design larger molecules having satisfactory outcomes in *in silico* analysis against the target protein. The newly designed hits are screened for their in vitro efficacy, and selected leads are examined for their binding poses using X-ray crystallography, which further improves efficacy. The current study is unique in terms of fragment selection, where crystallographic binding poses of the selected fragments in the designated active site of SARS-CoV-2 3CLpro were used as a starting point to design larger molecules. It was hypothesized that using appropriate linkers to join the fragment poses could provide relatively similar poses as constituting fragments. A previous relatively similar but limited study has reported using only three cocrystallized fragments to design 19 ligands [29].

After selecting small drug-like fragments, the second step in the rational fragment-based drug discovery approach may involve the coupling of these fragments to build large molecules with improved affinity for the selected target protein. In this regard, the choice of the linker group connecting two fragments and the connection sites of the fragments is of critical importance. In this study, the selection of linkers was based on the desire to get ligands with a similar conformation of constituting fragments as those in their crystal structures. A rigorous campaign on ligand coupling was carried out. Some couplings were rejected due to the overlap of fragments, and some couples were linked through multiple linkers, resulting in 1251 structures. In case of closely situated fragments, direct connection, an amine, or maximum two-methylene linkers was used. Amine and amide containing linkers were used when three to four bond-lengths were possible. In some set 2 to set 4 couplings cyclizations were considered across two adjacent piperazine carbons of set 2 and atoms across carbonyl carbon in set 4. A smile file containing coupled structures along with 87 starting fragment and 25 standards is provided in the supplementary information.

Virtual screening

Ligand preparation

A reasonably high number of stereoisomers was taken (maximum 32 for each structure) to get maximum possible conformations screening, allowing to find ligand conformations that occupy similar poses like those of their constituting fragments. The cocrystallized ligands with Mr > 340 Da (25) were used as standards during docking. A smile file with 7158 structures is provided in the supplementary information.

Glide and AutoDock Vina

Table 1: Top 22 hits having one or both fragments aligned similar to their crystal structures. Glide extra precision (XP), AutoDock (AD) Vina scores are presented along with the number of fragments conserved according to their crystal structures.

Sr. no.	Title	Glide XP score	Conserved fragments	AD Vina score	Conserved fragments	MMGBSA ΔG_{bind}
1	T7Y_UGS-1	-7.42	2	-9	2	-76.06
2	T3V_UGD-4	-8.11	1	-8.1	1	-66.79
3	T3G_UGD-2	-7.02	1	-8.3	2	-67.35
4	T7G_HWH-4	-7.72	1	-9.0	1	-64.76
5	T9M_UH7-2	-7.89	2	-7.0	2	-63.48
6	T9P_UGV-1	-7.14	1	-7.3	2	-64.78
7	T9M_RZJ-8	-7.02	2	-7.4	1	-64.86
8	T9P_UHA-2	-8.01	2	-7.6	1	-62.55
9	T9P_RZJ-5	-6.94	2	-7.5	2	-64.02
10	T2Y_UHA-2	-7.43	1	-7.6	2	-61.52
11	T9M_HWH-1	-6.84	2	-7.9	2	-59.25
12	T9M_UHA-8	-7.40	2	-7.6	1	-58.02
13	3WL_UGD-1	-7.07	2	-10.1	2	-58.58
14	T9M_UGG-3	-7.07	2	-8.6	0	-57.41
15	T4M_UHA-3	-7.93	2	-7.8	2	-53.61
16	USD_T9J-3	-6.88	2	-7.9	2	-55.41
17	T2J_UHA-3	-6.95	2	-8.2	1	-55.26
18	T7S_T67-6	-6.98	1	-9.0	1	-55.18
19	T7A_UHA-2	-7.01	2	-7.9	2	-53.73
20	T9P_HWH-3	-7.20	2	-7.5	1	-52.09
21	T9P_T9J-3	-7.30	1	-7.7	0	-50.68
22	T9M_T9J-5	-7.22	2	-7.2	1	-46.63

The energy values are in kcal/mol.

MMGBSA

Molecular mechanics generalized Born surface area (MMGBSA) was used to predict binding free energy (ΔG_{bind}) of the ligand-receptor complex. The top 103 ligands were ranked according to their MMGBSA values, and top 22 ligands (**Table 1**) were selected for molecular dynamics simulation.

Molecular dynamics simulation

Due to the binding site's polar nature and being open to the solvent, Hbonds and water bridges were predominantly seen, while hydrophobic interactions were relatively less pronounced. The top 22-subset compounds were ranked in **Figure 6** based on their Hbond occupancy during 12 ns simulation.

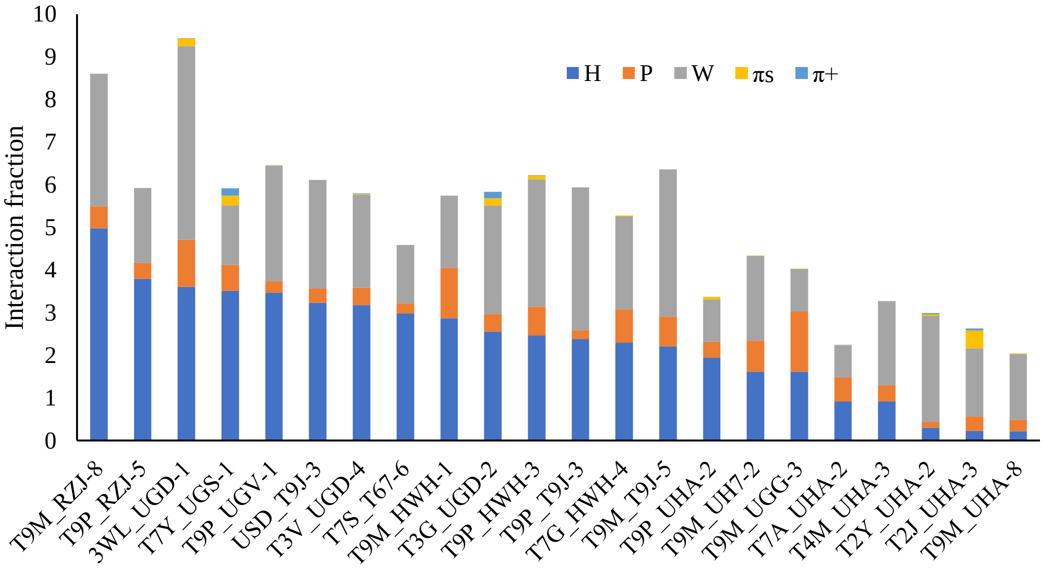

Figure 6: Interaction fraction (the number of interactions normalized throughout the trajectory) of the selected ligands for hydrogen bond (H), hydrophobic interaction (P), water bridge (W), π stacking (πs), and π cationic (π+) interactions during MD simulation.

RMSD graphs of 22 ligands with respect to the reference conformation (first frame at time t = 0) are given in **Figure 7**. It could be used to deduce the extent of ligand deviation from its first pose (at t = 0) before reaching a stable conformation during the simulation. Moreover, it can also explain ligand fluctuation or stability and the time taken by the ligand to achieve a stable conformation if there is one. For example, T2Y_UHA-2 stayed at a stable RMSD between 1.8 and 2 Å, T7A_UHA-2 was stable between 3 and 3.5 Å, and T9P_RZJ-5 was stable between 0.8 and 1 Å.

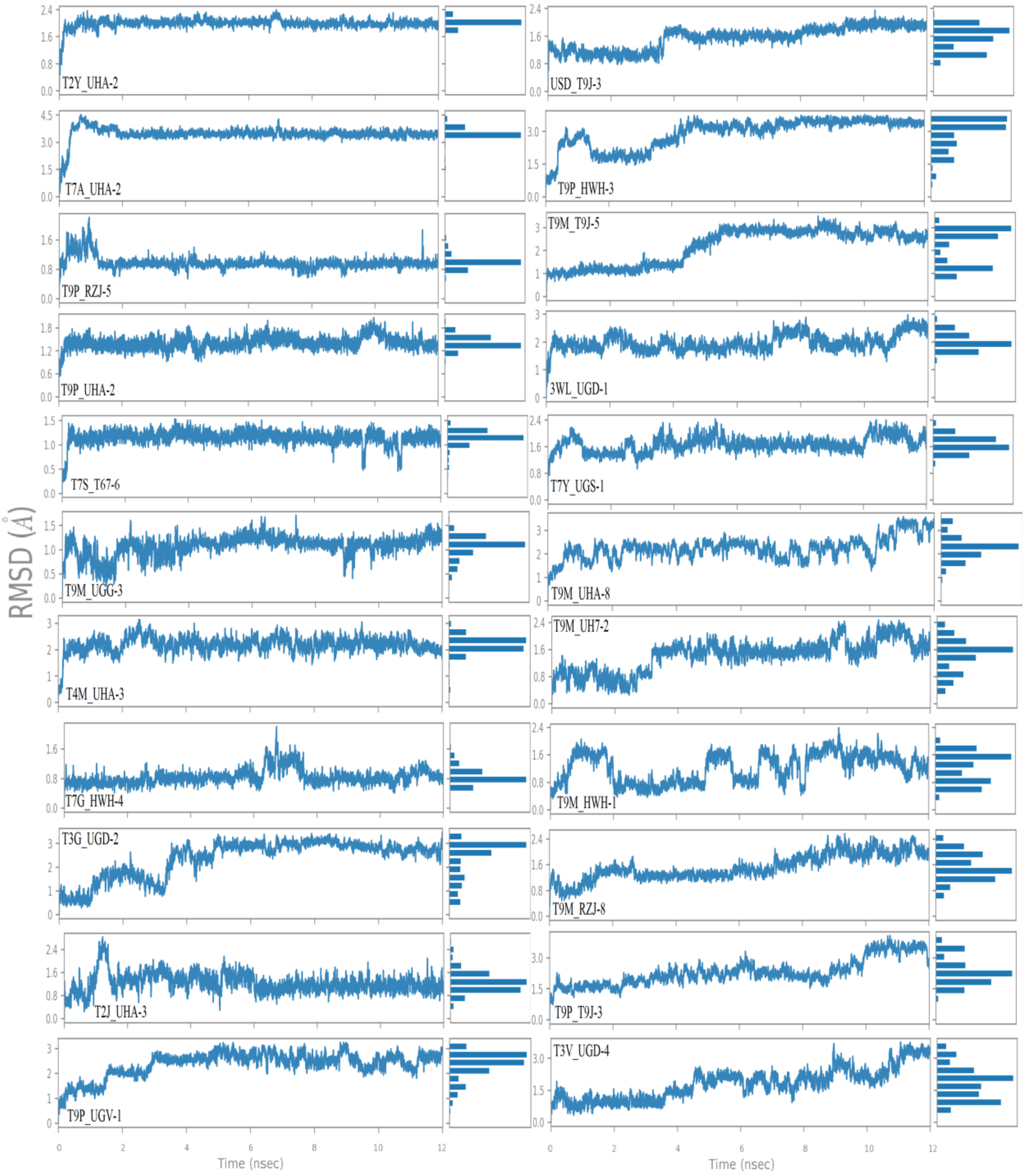

Figure 7: Root mean square deviation (RMSD) of 22 ligands with respect to the reference conformation (first frame at time t = 0).

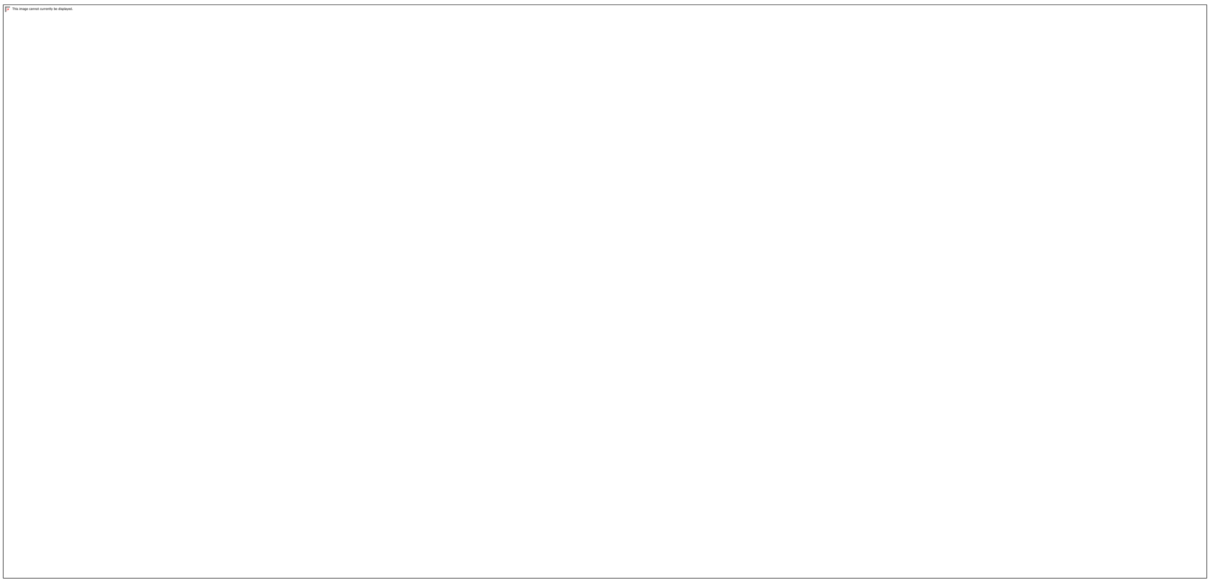

Figure 8: MD simulation data of T9P_RZJ-5. (A) Protein Cα RMSD on the left y-axis and ligand RMSD fit on protein on the right y-axis. (B) RMSF of ligand atoms. (C) MD simulated pose of the ligand in lime colour and cocrystallized poses of its fragments in teal and magenta colours. (D) Protein-ligand interactions that occur more than 30.0% of the simulation time. (E) A timeline representation of the interactions and contacts (Hbonds, hydrophobic, ionic, and water bridges). (F) Interaction fraction (the number of interactions normalized throughout the trajectory) of the ligand for hydrogen bond, hydrophobic interaction, ionic interaction, and water bridge.

T9P_RZJ-5 presented -6.94 kcal/mol XP score and -64.02 kcal/mol MMGBSA ΔG_{bind} (**Table 1**). It is an aromatic sulfonamide having 2nd highest occupancy of Hbonds among top 22 subset during 12 ns simulation. Thr25 and Gln189 were found to make single Hbond and His41 make two Hbonds during most of the simulations period. Gln189 was also making stable hydrophobic interactions throughout the period (**Figure 11F**). All the ligand atoms were having an RMSF of nearly 1 Å except one sulfonamide oxygen (2 Å) and CH_2 of terminal alkene (1.5 Å) (**Figure 11B**). It presented a stable RMSD of around 0.8 Å compared to the reference frame (t = 0), indicating that the pose obtained from XP docking was quite stable during the simulation (**Figure 7**). **Figure 8C** shows a comparison of cocrystallized poses of T9P and RZJ with MD simulated pose of T9P_RZJ-5. The RZJ is in close proximity to its corresponding part in T9P_RZJ-5, while T9P is drifted slightly into the pocket with flipped orientation. Furthermore, CHEMBL, PubChem, and SciFinder search for T9P_RZJ-5 yielded no output for reported activities of T9P_RZJ-5 or similar compounds. CHEMBL3467038, an N-acetyl derivative of RZJ has been found nontoxic towards HepG2 cells, and another closely resembling compound CHEMBL275919 is estrone sulfatase inhibitor (for details visit supplementary information).

This image cannot currently be displayed.

Figure 9: MD simulation data of T7Y_UGS-1. (A) Protein Cα RMSD on the left y-axis and ligand RMSD fit on protein on the right y-axis. (B) RMSF of ligand atoms. (C) MD simulated pose of the ligand in lime colour and cocrystallized poses of its fragments in teal and magenta colours. (D) Protein-ligand interactions that occur more than 30.0% of the simulation time. (E) A timeline representation of the interactions and contacts (Hbonds, hydrophobic, ionic, and water bridges). (F) Interaction fraction (the number of interactions normalized throughout the trajectory) of the ligand for hydrogen bond, hydrophobic interaction, ionic interaction, and water bridge.

T7Y_UGS-1 presented -7.42 kcal/mol XP score, -9 kcal/mole AD Vina score, and the highest MMGBSA ΔG_{bind} of -76.06 kcal/mol in the top-22 subset (**Table 1**). It is a naphthalene derivative having pyridine and pyrrolidone residues. It was found making two Hbonds with Gln166 and one Hbond with His41 (**Figure 9F**). RMSF plot of ligand fit on protein showed that naphthalene, methoxy N-acetyl residues fluctuated more than 1 Å, while the rest of the atoms were close to 1 Å (**Figure 9B**). Furthermore, **Figure 7** showed that the ligand mostly stayed at an RMSD between 1.2 and 2 Å compared to the initial pose (at t = 0). It is evident from **Figure 9B** that UGS part of T7Y_UGS-1 is very closely aligned with its cocrystallized posture, and T7Y part of T7Y_UGS-1 is drifted towards His41 as compare to its cocrystallized pose that is closer to Ans142. T7Y_UGS-1 or its closely related derivatives have not been reported in CHEMBL, PubChem, and SciFinder. In terms of fragments, a derivative of UGS having benzene in place of pyridine (CHEMBL1384462) has been reported for multiple biological activities, particularly against hepatitis C virus and influenza NS1. T7Y (CHEMBL1565601) has been reported against DNA polymerase iota, geminin, lysosomal α-glucosidase, and apurinic/apyrimidinic endonuclease-1. Detailed activity values could be accessed in the supplementary information.

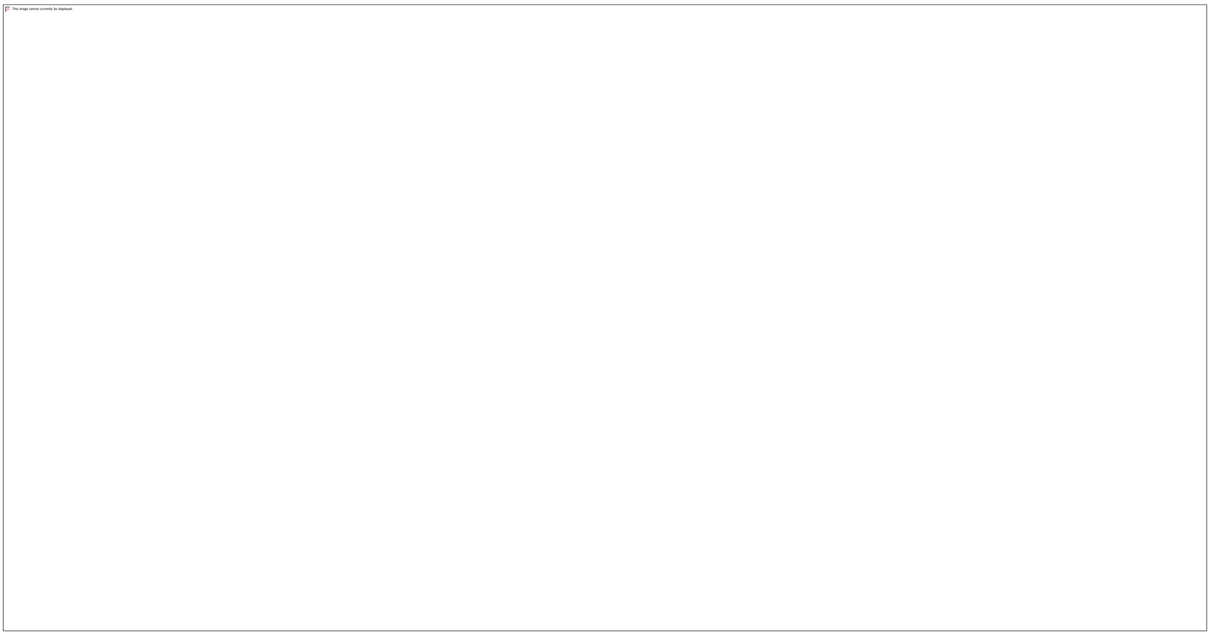

Figure 10: MD simulation data of T7S_T67-6. (A) Protein Cα RMSD on the left y-axis and ligand RMSD fit on protein on the right y-axis. (B) RMSF of ligand atoms. (C) MD simulated pose of the ligand in lime colour and cocrystallized poses of its fragments in teal and magenta colours. (D) Protein-ligand interactions that occur more than 30.0% of the simulation time. (E) A timeline representation of the interactions and contacts (Hbonds, hydrophobic, ionic, and water bridges). (F) Interaction fraction (the number of interactions normalized throughout the trajectory) of the ligand for hydrogen bond, hydrophobic interaction, ionic interaction, and water bridge.

T7S_T67-6 is a 7-azaindole derivative having one N-methylpiperazine and two piperidine residues. It has Glide XP score -6.98 kcal/mol, AD Vina score -9 kcal/mol, and MMGBSA ΔG_{bind} -55.18 kcal/mol (**Table 1**). It presented two Hbonds with Gln166 and one Hbond with Gln192 during most of the simulation period. A hydrophobic interaction was also observed with Gln189 (**Figure 10E and F**). The RMSF of ligand atoms was observed less than 1 Å except carbon 11 of the terminal piperidine, which was slightly higher than 1 Å (**Figure 10B**). In terms of ligand RMSD, it was stable between 1 and 1.4 Å when compared to the first frame (t = 0) (Figure 7). T7S part of T7S_T67-6 was found in the proximity of its cocrystallized pose while the pose of T67 fragment was not in agreement with its cocrystallized orientation (**Figure 10C**). Two closely related derivatives of T7S (CHEMBL1574087 and CHEMBL1541069) are potent inhibitors of hepatitis C virus, influenza NS1, TrxR1, SMN2, and IMPase in the low nanomolar range. Furthermore, close analogues of T67 (CHEMBL231023 and CHEMBL3754439) are weak anti-inflammatory agents in the rat model.

QikProp

ADME property assessment of the top-22 has been provided in the supplementary information.

Methods

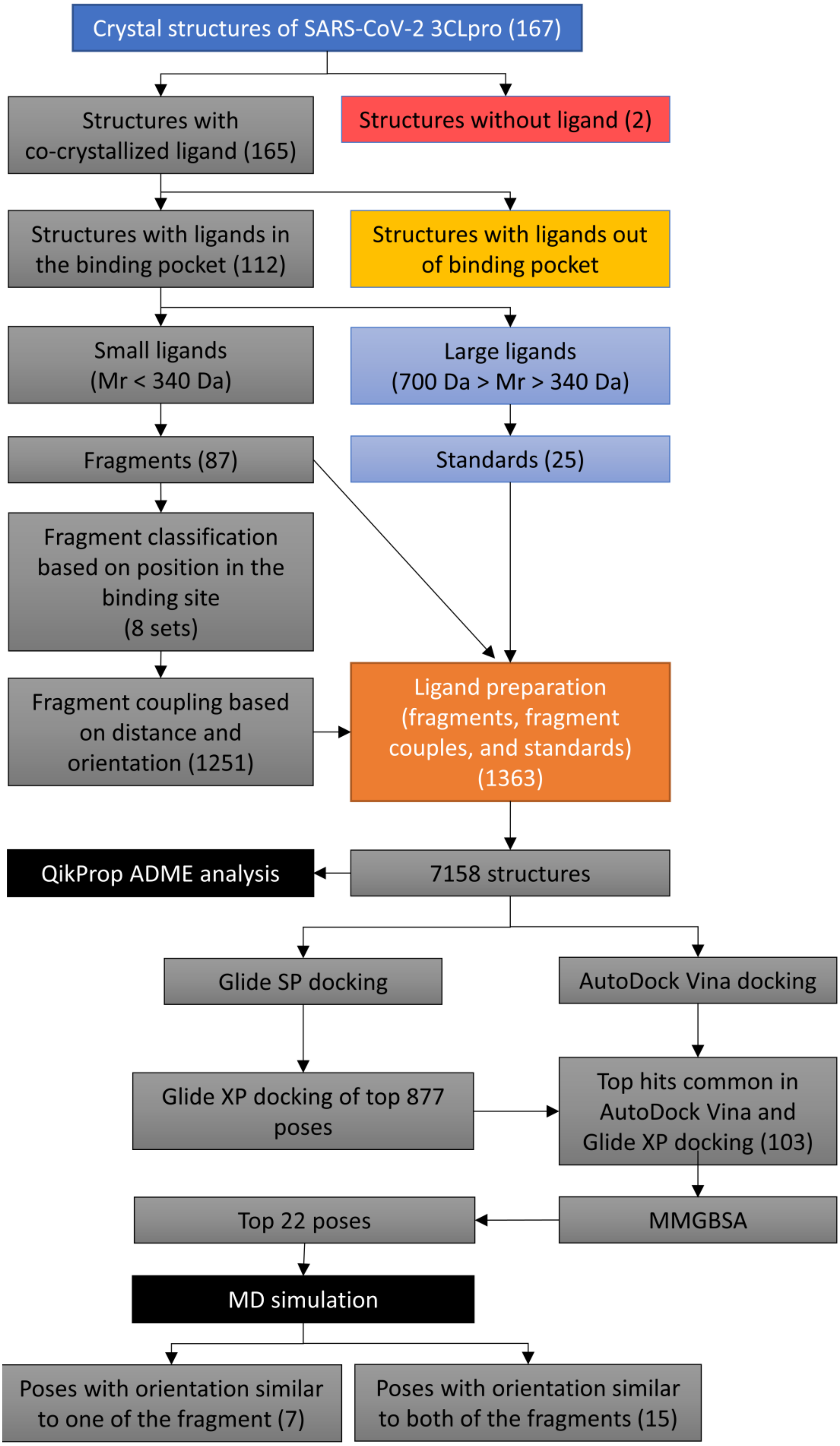
Crystal structures of SARS-CoV-2 3CLpro (167)
Structures with co-crystallized ligand (165)
Structures without ligand (2)
Structures with ligands in the binding pocket (112)
Structures with ligands out of binding pocket
Small ligands (Mr < 340 Da)
Large ligands (700 Da > Mr > 340 Da)
Fragments (87)
Standards (25)
Fragment classification based on position in the binding site (8 sets)
Fragment coupling based on distance and orientation (1251)
Ligand preparation (fragments, fragment couples, and standards) (1363)
QikProp ADME analysis
7158 structures
Glide SP docking
AutoDock Vina docking
Glide XP docking of top 877 poses
Top hits common in AutoDock Vina and Glide XP docking (103)
Top 22 poses
MMGBSA
MD simulation
Poses with orientation similar to one of the fragment (7)
Poses with orientation similar to both of the fragments (15)

Figure 11: A schematic block diagram of the workflow followed in the study.

Designations

All the ligands and protein structures are represented according to their PDB identifiers. The coupled ligands were named according to their PDB identifiers separated by an underscore. The ligands with 700>Mr>340 Dalton were treated as standards and their PDB identifiers were prefixed with L underscore (L_PDB ID).

Fragments

PDB (www.rcsb.org) housed 167 crystal structures of SARS-CoV-2 3CLpro at the time of data accumulation (10 November 2020). These structures were fetched and analyzed for initial filtration based on ligand molecular mass less than 700 Da and ligand residing in the *bona fide* binding pocket. This provided 112 crystal structures. The 87 ligands with molecular mass less than 340 Da were selected as fragments for further development and the remaining 25 ligands were selected as controls for the docking process.

Fragment coupling and linker selection

The selected 87 fragments were examined in detail for their binding site interactions and divided into eight groups based on their amino acid interactions and location in the binding site. In the next step, the fragments were coupled across different groups **(Table 2**). Coupling was ignored where more than two-atom overlap was found. During the linker design, it was kept in mind that the joining two ligands should occupy similar conformation in the binding pocket after the connection. In this regard, the distance between joining atoms was considered to find the number of new bonds and bond angles. Furthermore, it is plausible to find functional groups from other sets that occupy the empty space for the connection of any two sets. The type of amino acids in the vicinity was also considered while designing linkers.

Table 2: Tabulated representation of possible combinations of 8 sets of fragments. Green cells represent the potential coupling of two groups based on their positions. The digits above the red diagonal are the number of bonds that a linker could have. The numbers below the red diagonal represent the number of couples generated across two sets (sum 1251). In the case of set 8 two couplings were possible within the set.

	Set 1-24	Set 2-11	Set 3-17	Set 4-13	Set 5-8	Set 6-5	Set 7-5	Set 8-4
Set 1-24	🟥	🟨	🟨	1, 2	1	2	3	2
Set 2-11	🟨	🟥	🟨	1	🟨	2	4	1
Set 3-17	🟨	🟨	🟥	🟨	🟨	3	2	1
Set 4-13	333	143	🟨	🟥	🟨	🟨	2	🟨
Set 5-8	158	🟨	🟨	🟨	🟥	3	🟨	1
Set 6-5	120	55	65	🟨	37	🟥	3	3
Set 7-5	120	55	30	2	🟨	25	🟥	3
Set 8-4	53	11	11	🟨	8	10	13	2

Virtual screening

Ligand preparation

The smile strings of the 1363 ligands (87 fragments, 25 large molecules with Mr > 340 Da as standards, and 1251 combinations) were placed in the first column of the MS Excel and their corresponding unique identifiers were placed in the second column. Both columns were copied and pasted in the notepad, thereby giving a smile, tab space, and the identifier in each line. The contents of the notepad were converted into a single sdf file with the help of OpenBabel (v 3.1.1). The sdf file was loaded in Maestro, and 3D structures of the compounds were prepared and optimized using the LigPrep module of Maestro (Schrödinger Release 2020-4: Maestro, LigPrep, Schrödinger, LLC, New York, NY, 2020) [30]. Possible tautomers of the compounds were produced using Epik [31] at the target pH of 7.0±2.0. For stereoisomers, specified chiralities were retained, and other chiral centres were varied to get a maximum of 32 isomers. Finally, 7158 structures were obtained for the docking purpose.

Protein preparation

The crystal structure of SARS-CoV-2 3CLpro protease was retrieved in the form of biological assembly from the Protein Data Bank (PDB ID 5R83, resolution 1.58 Å). As 3CLpro exists as a homodimer, both chains were used for the further screening process. The structure was loaded in UCSF Chimera (v 1.15) and saved as a pdb file to get both chains. The Protein Preparation Wizard of the Maestro molecular modeling software (Schrödinger Release 2020-4: Protein Preparation Wizard, Schrödinger, LLC, New York, NY, 2020) was used to prepare protein after grouping both chains of the dimer. The preparation steps involved addition of hydrogens, optimization of the hydrogen bond networks, and assignment of protonation states of histidine residues. Furthermore, water molecules were removed, and a restrained minimization was performed using the OPLS3e force field [31]. Subsequently, the Receptor Grid Generation module of Maestro was employed to identify the docking site by generating a cubical grid box that was centred at (7.3, 0.9, 25.2) with 22 Å length.

Docking

Glide

Virtual screening in the specified search space of the SARS-CoV-2 3CLpro protease was performed by using the Glide docking tool of Maestro with SP (Standard Precision) mode. The top 877 poses were redocked using XP (Extra Precision) mode [32].

AutoDock Vina

For comparison purposes, 7158 ligand poses prepared by LigPrep in the previous step were exported and docked on 3CLpro (prepared previously) using an automated Mcule server [33]. Same search box parameters were used as previously defined.

MMGBSA

The results from Glide XP and AD Vina were compared to selected top 103 poses. The estimated binding free energy of the top ligand poses was calculated with the Prime/MMGBSA (molecular mechanics generalized Born surface area) module using the VSGB solvation model and the OPLS3e force field [32].

Molecular dynamics simulation

The Maestro System Builder module was used to prepare docked protein-ligand complexes for molecular dynamics (MD) simulation. Simple point-charge (SPC) water model was used with an orthorhombic box extending 10 Å from protein. The placement of ions was excluded within 20 Å of the ligand. The system was neutralized using sodium or chloride ions, and 0.15 M sodium chloride was added. OPLS3e force field was used to record 12 ns simulation with 3 ps recording interval. NPT ensemble was used with 300 K, and 1.01325 bar and the system was relaxed (100 ps) before simulation (Schrödinger Release 2020-4: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2020. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2020).

ADME prediction

ADME (absorption, distribution, metabolism, and excretion) properties of the 7158 ligands containing selected fragments, fragment couples, and standard molecules were predicted using QikProp tool of Maestro (Schrödinger Release 2020-4: QikProp, Schrödinger, LLC, New York, NY, 2020).

Conclusion and future directions

The study rigorously explored small molecule fragments cocrystallized with 3CLpro and employed crystal structure-guided fragment-based approach to finding its new inhibitors. Selected 87 fragments were interactively connected via linkers that could give the same conformations of the resulting molecules as those constituting fragments in the crystal structures. The resulting 1251 structures were screened using docking, MMGBSA, and MD simulation. Top 22 hits identified were having their poses closely aligned to at least one of the fragment's crystal pose after MD simulation. Based on MD simulation, pose, and binding site interactions, selected the top 3 compounds (T7Y_UGS-1, T9P_UHA-2, and T7S_T67-6) depicted computationally promising results that warrant their further development and biological testing. Furthermore, three groups linking, for example, set1+set5+set6 and set1+set6+set7 could be designed for the extended CLpro inhibitors with enhanced affinity.

Acknowledgement

The authors highly acknowledge the Universiti Malaya (grant no.: IIRG003A-2019) for the necessary resources to complete this work.

Conflict of interest

The authors declare no conflict of interest.

Author contribution

Study design, experimental work, and data collection: Sarfraz Ahmad; *in silico* analysis and interpretation of the data: Sarfraz Ahmad, Iskandar Abdullah, Muhammad Usman Mirza; drafting the manuscript and literature survey: Lee Yean Kee, Sarfraz Ahmad, Mamoona Nazir; critical revision of the manuscript: John F. Trant and Noorsaadah Binti Abd Rahman.

References

1. Mirza, M.U., S. Ahmad, I. Abdullah M. Froeyen. Identification of novel human USP2 inhibitor and its putative role in treatment of COVID-19 by inhibiting SARS-CoV-2 papain-like (PLpro) protease. Computational biology and chemistry 2020;89:107376-76.
2. Organization, W.H. *Coronavirus*. 2020 [cited 2020 29 November]; Available from: https://www.who.int/health-topics/coronavirus#tab=tab_1.
3. Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine 2020;382(8):727-33.
4. Lam, T.T.-Y., N. Jia, Y.-W. Zhang, M.H.-H. Shum, J.-F. Jiang, H.-C. Zhu, et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 2020:1-4.
5. Control, E.C.f.D.P.a. *COVID-19 situation update worldwide, as of 28 November 2020*. 2020 [cited 2020 29 November]; Available from: https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases.
6. https://www.who.int/data#reports , accessed on 3rd December.
7. Outbreak.MY. *COVID-19 Outbreak Live Updates* 2020 [cited 2020 29 November]; Available from: https://www.outbreak.my/.
8. Shah, A.U.M., S.N.A. Safri, R. Thevadas, N.K. Noordin, A.A. Rahman, Z. Sekawi, et al. COVID-19 outbreak in Malaysia: Actions taken by the Malaysian government. International Journal of Infectious Diseases 2020;97:108-16.
9. Karia, R., I. Gupta, H. Khandait, A. Yadav A. Yadav. COVID-19 and its Modes of Transmission. SN comprehensive clinical medicine 2020:1-4.
10. Barouki, R., M. Kogevinas, K. Audouze, K. Belesova, A. Bergman, L. Birnbaum, et al. The COVID-19 pandemic and global environmental change: Emerging research needs. Environment International 2021;146:106272.
11. Zhang, C., S. Huang, F. Zheng Y. Dai. Controversial treatments: an updated understanding of the Coronavirus Disease 2019. Journal of medical virology 2020.
12. Lurie, N., M. Saville, R. Hatchett J. Halton. Developing Covid-19 vaccines at pandemic speed. New England Journal of Medicine 2020;382(21):1969-73.
13. Li, G. E. De Clercq, *Therapeutic options for the 2019 novel coronavirus (2019-nCoV)*. 2020, Nature Publishing Group.
14. Zhang, J., H. Zeng, J. Gu, H. Li, L. Zheng Q. Zou. Progress and prospects on vaccine development against SARS-CoV-2. Vaccines 2020;8(2):153.
15. Lu, R., X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet 2020;395(10224):565-74.
16. Chen, Y., Q. Liu D. Guo. Emerging coronaviruses: genome structure, replication, and pathogenesis. Journal of medical virology 2020;92(4):418-23.
17. Hussain, S., Y. Chen, Y. Yang, J. Xu, Y. Peng, Y. Wu, et al. Identification of novel subgenomic RNAs and noncanonical transcription initiation signals of severe acute respiratory syndrome coronavirus. Journal of virology 2005;79(9):5288-95.
18. Yang, H., M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, et al. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. Proceedings of the National Academy of Sciences 2003;100(23):13190-95.
19. Liu, X. X.-J. Wang. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. Journal of Genetics and Genomics 2020;47(2):119.
20. Anand, K., J. Ziebuhr, P. Wadhwani, J.R. Mesters R. Hilgenfeld. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 2003;300(5626):1763-67.
21. Bacha, U., J. Barrila, A. Velazquez-Campoy, S.A. Leavitt E. Freire. Identification of novel inhibitors of the SARS coronavirus main protease 3CLpro. Biochemistry 2004;43(17):4906-12.

22. Lai, L., X. Han, H. Chen, P. Wei, C. Huang, S. Liu, et al. Quaternary structure, substrate selectivity and inhibitor design for SARS 3C-like proteinase. Current pharmaceutical design 2006;12(35):4555-64.

23. Xia, B. X. Kang. Activation and maturation of SARS-CoV main protease. Protein & cell 2011;2(4):282-90.

24. Fan, K., L. Ma, X. Han, H. Liang, P. Wei, Y. Liu, et al. The substrate specificity of SARS coronavirus 3C-like proteinase. Biochemical and biophysical research communications 2005;329(3):934-40.

25. Ramajayam, R., K.-P. Tan P.-H. Liang, *Recent development of 3C and 3CL protease inhibitors for anti-coronavirus and anti-picornavirus drug discovery*. 2011, Portland Press Ltd.

26. Yang, H., M. Bartlam Z. Rao. Drug design targeting the main protease, the Achilles' heel of coronaviruses. Current pharmaceutical design 2006;12(35):4573-90.

27. Berry, M., B.C. Fielding J. Gamieldien. Potential broad spectrum inhibitors of the coronavirus 3CLpro: a virtual screening and structure-based drug design study. Viruses 2015;7(12):6642-60.

28. Liu, Y., C. Liang, L. Xin, X. Ren, L. Tian, X. Ju, et al. The development of Coronavirus 3C-Like protease (3CLpro) inhibitors from 2010 to 2020. European journal of medicinal chemistry 2020:112711.

29. Luan, B. T. Huynh. Crystal-Structures-Guided Design of Fragment-Based Drugs for Inhibiting the Main Protease of SARS-CoV-2. 2020.

30. Sastry, G.M., M. Adzhigirey, T. Day, R. Annabhimoju W. Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. Journal of Computer-Aided Molecular Design 2013;27(3):221-34.

31. Roos, K., C. Wu, W. Damm, M. Reboul, J.M. Stevenson, C. Lu, et al. OPLS3e: Extending force field coverage for drug-like small molecules. Journal of Chemical Theory and Computation 2019;15(3):1863-74.

32. Li, J., R. Abel, K. Zhu, Y. Cao, S. Zhao R.A. Friesner. The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. Proteins: Structure, Function, and Bioinformatics 2011;79(10):2794-812.

33. Kiss, R., M. Sandor F.A. Szalai. http://Mcule.com: a public web service for drug discovery. Journal of cheminformatics 2012;4(S1):P17.