Team name	KyuKen (九研 in kanji)
Team member(s) (firstname lastname; …)	Ashutosh Kumar(1), Francois Berenger(2), Yoshihiro Yamanishi(2), Kam Y. J. Zhang(1).
Affiliation	(1) Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. (2) Department of Bioscience and Bioinformatics, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan.
Contact email	beren314@bio.kyutech.ac.jp
Contact phone number (optional)	[redacted]
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	3CLpro, Human-ACE2/viral-spike-protein interface, PLpro.

Methods:

In essence, our strategy was two folds. First, fast ligand-based models specific to each target were applied on the whole 1B+ molecules library. Then, only molecules selected by the LBVS step were subjected to docking and structure-based rescoring procedures. This strategy was dictated by our docking capacity, which we estimated at a maximum of 10M molecules times three targets times three times per target. Our docking capacity was not only limited by the computational power we have access to (on the RIKEN Hokusai supercomputer), but also by the disk space required to store 3D conformers of so many molecules.
While for some targets we used ligand-data related to the SARS coronavirus (not Covid-19), we believe that our structure-based methods allow to precisely target Covid-19, since the ranking of molecules is determined with a weight of at least two thirds by structure-based methods (either docking or at least shape and pharmacophore features similarity to the 3D bioactive conformation of a known inhibitor for SARS-Cov-2).
Each scoring method assigns a score to a molecule. The final score of a molecule for a given target is its average normalized score (zero mean and unit standard deviation), over all models for this target.

Targets:

Target 1 (3CLpro PDB:6Y2F):

LBVS-1

For this target, we trained two classifiers and a regressor. The two classifiers can be considered as defining an applicability domain for the regressor. We used a SARS-coronavirus dataset to create ligand-based models for this target.
The corresponding CHEMBL page is:
https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL3927.
While CHEMBL claims that there are 133 molecules with an IC50 value for this target, our regressor model was trained using only the 91 molecules which have a pCHEMBL value (a kind of pIC50).

PLS regressor for this dataset. Molecular encoding: unfolded-counted atom pairs (there are 17368 distinct atom pairs in our encoding dictionary, which was created by encoding the whole

CHEMBL24 after standardization). The optimal number of PLS components was determined to be nine during a 10-fold cross-validation experiment (R2=0.53).

We also trained two classifiers for this protein target. The 91 CHEMBL molecules with pIC50 values were used as actives, then a random chemical background of 25 times more molecules was randomly drawn from the whole CHEMBL24 but not including those actives.

VRK classifier. A Vanishing Ranking Kernels [Berenger2020] model was also trained, using the probability read from the Kernel Density Estimate as the scoring function. The optimal kernel bandwidth was found to be 0.773 in a 10-fold cross-validation experiment. The corresponding area under the Receiver Operating Characteristic curve (AUC) was 0.99. An optimal classification threshold of 0.859 was determined by Matthews Correlation Coefficient (MCC) maximization in a 10-fold cross-validation experiment (MCC=0.732).

LRL2 classifier. Another random chemical background was drawn and another classification model was trained. Molecular encoding: 16384 bits long ECFP6 [Oboyle2016]. Model: L2-regularized Logistic Regressor (LRL2) with optimal parameter C=5. 10-fold cross-validation AUC=0.997. Classification threshold t=0.119; obtained via Matthews Correlation Coefficient (MCC) maximization in a 10-fold cross-validation experiment (MCC=0.91). Out of our 1B+ molecules collection, only 770,326 were selected for subsequent docking and ligand-based scoring by this classifier.

While only the fastest model (LRL2) has screened the whole database for this target, the three ligand-based models were used to rank-order the 770326 molecules predicted actives. Then, the average rank of each molecule was used as its ligand-based rank for this target.

LBVS-2

In the second LBVS strategy, we employed ligand 3D shape similarity calculations to prepare a ranked list. In this strategy, 770326 molecules obtained after ligand-based rank-ordering were screened for their shape similarities with a SARS CoV-2 3CLpro inhibitor (alpha-ketoamide 13b**)** [Zhang2020]. The bioactive conformation of alpha-ketoamide 13b was used as a query for shape comparisons (PDB code: 6Y2F). The shape similarity calculations were performed using ROCS program (OpenEye Scientific Software). Ligand conformations for shape similarity calculations were generated employing OMEGA program (OpenEye Scientific Software). A maximum number of 200 conformations per compound were generated. TanimotoCombo, a combined Tanimoto coefficient for the shape and chemical features was used to rank-order the screening library.

SBVS

Structure-based virtual screening (SBVS) was performed using molecular docking approach [Kumar2014, Jiang2016, Matsuoka2017, Jiang2020]. In this strategy, 770326 molecules obtained after LBVS-1 were docked to 3CLpro crystal structure using two separate docking programs FRED (OpenEye Scientific Software) and GOLD (CCDC). Conformational ensemble of compounds for FRED docking was generated using OMEGA(OpenEye Scientific Software). A maximum of 200 conformations per compound were generated. Receptor for FRED docking was prepared using the ‘make_receptor’ utility, which is a component of OpenEye’s OEDocking suite. SARS CoV-2 3CLpro crystal structure in complex with the inhibitor alpha-ketoamide 13b (PDB Code: 6Y2F) was employed in docking calculations. Ligand poses were scored using the Chemgauss4 scoring function and a single pose per ligand was saved. Protein structure for GOLD molecular docking calculations was prepared using the protein preparation utility in Maestro (Schrodinger Inc.) where hydrogens were added, bond-orders were assigned and protonation states of all charged residues were determined at the neutral pH. The bioactive conformation of alpha-ketoamide 13b (PDB code: 6Y2F) was used to define the binding pocket for GOLD molecular docking calculations. Ligands were prepared using OEChem toolkit (OpenEye Scientific Software) where hydrogens were added and charges were assigned using the AM1BCC force field. The PLP scoring function was used to score binding poses and a single pose per compound was saved.

Target 2 (ACE2 PDB:6LZG):

LBVS

Since we are targeting the Protein-Protein Interface (PPI) between the viral spike protein and the human ACE2 receptor, we did not find enough ligand data for this target (we were not looking for ACE2 inhibitors). However, since we knew we were targeting a PPI, we created a ligand-based model to detect potential PPI inhibitor molecules. First, the whole IPPIDB [Labbe2016] (https://ippidb.pasteur.fr) was downloaded. It contains 1756 molecules. Then, 25 times more molecules from the whole CHEMBL24 database not included in IPPIDB, were randomly drawn to create a random chemical background. Then, a classifier was trained on this dataset.

PPII LRL2 classifier. Molecular encoding: 16384 bits long ECFP6. Model: L2-regularized logistic regressor with optimal parameter C=0.5. Five-fold cross-validation AUC=0.99. Classification threshold t=0.31, obtained via Matthews Correlation Coefficient (MCC) maximization in a 10-fold cross-validation experiment (MCC=0.93). Out of our 1B+ molecules collection, only 2,461,774 were selected for subsequent docking by this classifier. However, we don't consider this classifier to be specific to the protein target at hand, so it was not used to rank order molecules for this target. For this target, our ranking is purely structure-based.

SBVS

We first utilized computational fragment mapping to identify small molecule binding hotspots on the binding interface between the human ACE2 receptor and the receptor binding domain of the viral spike protein. Computational fragment mapping calculation were performed using the FTMap server (https://ftmap.bu.edu/login.php). FTMap [Kozakov2015] uses molecular docking, energy minimization and clustering to predict clusters of probe molecules called consensus sites (CSs) as small molecule binding hotspots. Generally, the most populated CS is considered as the main hotspot, but here in this study we were interested in CS in the ACE2/Spike protein-protein interface (PPI). Three ACE2/Spike protein-protein complex structures were used for computational fragment mapping calculations (PDB code: 6M17, 6LZG and 6M0J). Computational fragment mapping was performed on ACE2 component of the ACE2/spike PPI. Computational fragment mapping calculation revealed several CSs on the binding interface of ACE2/Spike protein PPI. Combined hotspot from three structures occupies a region consisting of Phe32, Ala36, Glu37, Phe40, Ser43, Ser44, Ser47, Trp69, Leu73, Gln96, Gln98, Ala99, Leu100, Gln101, Gln102, His345, Pro346, Thr347, Ala348, Trp349, Asp350, Leu351, Gly352, Lys353, His374, Glu375, His378, Ile379, Asp382, Tyr385, Pro389, Phe390, Leu391, Leu392, Arg393, Asn394, Gly395, His401, Glu402, and Lys562. Computational fragment mapping calculation suggests that this region has high probability of small molecule binding, which may interfere with the binding of the viral spike protein to ACE2.

To prepare the ranked list of compounds, three separate structure-based virtual screening (SBVS) strategies were followed for this target utilizing computational fragment mapping.

First (FRED) and second (GOLD) strategy. 2,461,774 molecules obtained after ligand-based rank-ordering were docked to ACE2 using the two separate docking programs FRED (OpenEye Scientific Software) and GOLD (CCDC). Conformational ensemble of compounds for FRED docking was generated using OMEGA(OpenEye Scientific Software). A maximum of 200 conformations per compound were generated. Receptor for FRED docking was prepared using the ‘make_receptor’ utility which is a component of OpenEye’s OEDocking suite. Chain A of ACE2/spike PPI complex structure (PDB Code: 6LZG) was employed in docking calculations. Docking was performed at small molecule binding hotspot of the ACE2/spike protein binding interface, as identified by computational fragment mapping. Ligand poses were scored using the Chemgauss4 scoring function and a single pose per ligand was saved. Protein structure for GOLD molecular docking calculations was prepared using protein preparation utility in Maestro (Schrodinger Inc.) where hydrogens were added, bond-orders were assigned and protonation states of all charged residues were determined at neutral pH. Binding pocket for GOLD molecular docking calculations was defined using the list of the residues around 5 Å of FTMap CSs. Residues in binding pocket include Phe32, Ala36, Glu37, Phe40, Ser43, Ser44, Ser47, Trp69, Leu73, Gln96, Gln98, Ala99, Leu100, Gln101, Gln102, His345, Pro346, Thr347, Ala348, Trp349, Asp350, Leu351, Gly352, Lys353, His374, Glu375, His378, Ile379, Asp382, Tyr385, Pro389, Phe390, Leu391, Leu392, Arg393, Asn394, Gly395, His401, Glu402, and Lys562. Ligands were prepared using OEChem toolkit (OpenEye Scientific Software) where hydrogens were added and charges were assigned according to the AM1BCC force field. The PLP scoring function was used to score binding poses and a single pose per ligand was saved.

Third strategy. 2,461,774 molecules obtained after ligand-based rank-ordering were screened for their shape similarity with a query generated from FTMap probes bound at the ACE2/spike PPI. A simplified query was generated from bound chemical probes and probes occupying various regions of the binding pocket were selected. Shape similarity calculations were performed using ROCS (OpenEye Scientific Software). Ligand conformations for shape similarity calculations were generated with OMEGA (OpenEye Scientific Software). A maximum number of 200 conformations per ligand were generated. TanimotoCombo, a combined Tanimoto coefficient for the shape and chemical features was used to rank-order the chemical library. While this strategy could be thought of as ligand-based, it must be kept in mind that the query ligand is only virtual, and constructed using the targeted PPI hotspots.

Target 3 (Plpro PDB:6WUU):

LBVS-1

For this target also (as for 3CLpro) we used a SARS coronavirus dataset to train some ligand-based models. The PubChem assay id 1944 (https://pubchem.ncbi.nlm.nih.gov/bioassay/1944): "dose response biochemical high throughput screening assay to identify inhibitors of the papain-like protease (Plpro)" was used.

For this dataset, we could not obtain a good quality regression QSAR model. Indeed, this dataset only has 19 actives out of 101 assayed molecules. Instead, we created two distinct classifiers from this dataset.

- **1st model:** Molecular encoding: 16384 bits long ECFP6. L2-regularized logistic regressor with optimal parameters C=10 and bagging [Breiman1996] with 20 models. Five-fold cross-validation AUC=0.92. Classification threshold t=0.634, obtained via Matthews Correlation Coefficient (MCC) maximization in a 10-fold cross-validation experiment (MCC=0.71). Out of our 1B+ molecules collection, more than 20M were selected for subsequent screening by this classifier.
- **2nd model:** The dataset was encoded using counted atom pairs (17368 features). Then, a Vanishing Ranking Kernels model was trained, using the probability read from the Kernel Density Estimate as the scoring function. The optimal kernel bandwidth was found to be 1.0 (the maximum) in a five-fold cross-validation experiment (AUC=0.9). An optimal classification threshold of 0.576 was determined by MCC maximization in a 10-fold cross-validation experiment (MCC=0.59). The combination of this second ligand-based model with the previous one identified 4,001,268 molecules for subsequent docking.

The ranks calculated by both ligand-based models were combined to create a single LBVS-based rank for this target.

LBVS-2

We employed ligand 3D shape similarity calculations to prepare a ranked list. In this strategy, 4,001,268 molecules obtained after ligand-based rank-ordering were screened for their shape similarities with SARS CoV-2 PLpro peptide inhibitor VIR250 (PDB Code: 6WUU). The bioactive conformation of VIR250 was used as a query for shape comparisons (PDB code: 6WUU). The shape similarity calculations were performed using ROCS (OpenEye Scientific Software). Ligand conformations for shape similarity calculations were generated with OMEGA (OpenEye Scientific Software). A maximum number of 200 conformations per compound were generated. TanimotoCombo, a combined Tanimoto coefficient for the shape and chemical features was used to rank-order the chemical library.

SBVS

Two separate structure-based virtual screening (SBVS) strategies were followed for this target. 4,001,268 molecules obtained after ligand-based rank-ordering were docked to PLpro crystal structure using two separate docking programs FRED (OpenEye Scientific Software) and GOLD (CCDC). Conformational ensemble of compounds for FRED docking was generated using OMEGA(OpenEye Scientific Software). A maximum of 200 conformations per compound were generated. Receptor for FRED docking was prepared using the ‘make_receptor’ utility which is a component of OpenEye’s OEDocking suite. SARS CoV-2 Papain-like protease in complex with peptide inhibitor VIR250 (PDB Code: 6WUU) was employed in docking calculations. Ligand poses were scored using the Chemgauss4 scoring function and a single pose per ligand was saved. Protein structure for GOLD molecular docking calculations was prepared using the protein preparation utility in Maestro (Schrodinger Inc.) where hydrogens were added, bond-orders were assigned and protonation states of all charged residues were determined at the neutral pH. The bioactive conformation of peptide inhibitor VIR250 (PDB code: 6WUU; unpublished article) was used to define the binding pocket for GOLD molecular docking calculations. Ligands were prepared using OEChem toolkit (OpenEye Scientific Software) where hydrogens were added and charges were assigned according to the AM1BCC force field. The PLP scoring function was used to score binding poses and a single pose per ligand was saved.

Library:

Name	#Molecules / Fraction of the total	Source	Remark
ZINC15	997,402,117 / ~99 %	https://zinc15.docking.org/tranches/home/	2D, Standard reactivity, wait OK. Molecule names start with ‘ZINC’.
AMS	5,049,124 / ~5 per thousand	JEDI organizers	Not in ZINC, compiled by Merck. Molecule names start with ‘AMS_’.
CAS antivirals	47,488 / ~5 per 100,000	https://www.cas.org/covid-19-antiviral-compounds-dataset	Molecule names start with ‘CAS_’.
KEGG all drugs	7,557 / negligible	https://www.genome.jp/kegg/	Molecule names start with ‘D’.
SWEETLEAD	4,213 / negligible	https://simtk.org/home/sweetlead	Molecule names start with ‘SW’.

Total: 1,002,510,499 molecules.

Prior to LBVS, all molecules were subjected to standardization (Francis Atkinson's standardizer: https://github.com/flatkinson/standardiser; pip package chemo-standardizer). Molecules failing standardization (e.g. only salt, solvent, known toxic or mixtures) were removed. All SMILES files making our 1B+ molecules library can be downloaded from Google drive:

https://drive.google.com/open?id=1F9qfLwinygr2Xl5PHfHjhQ4zQSv4ewFR .

Results:

Top 5 ligands for Target 1 (3CLpro):

The chemical structures of top 5 ligands are shown in Figure 1. Three of the top 5 ligand (ZINC000641879661, ZINC000027399759 and ZINC000520995990) belong to 1-Anilino-2-(*o*-anilinocarbonylphenylthio)-1-ethanone scaffold. According to ChEMBL20, there is no known reported activity for any of these compounds. All of them are make-on-demand compounds from Enamine-Real database and had not been synthesized previously. We further noticed that among the top 10,000 best scoring molecules for 3CLpro, 96.5% are from ZINC database while the rest of the 3.5% are from AMS.

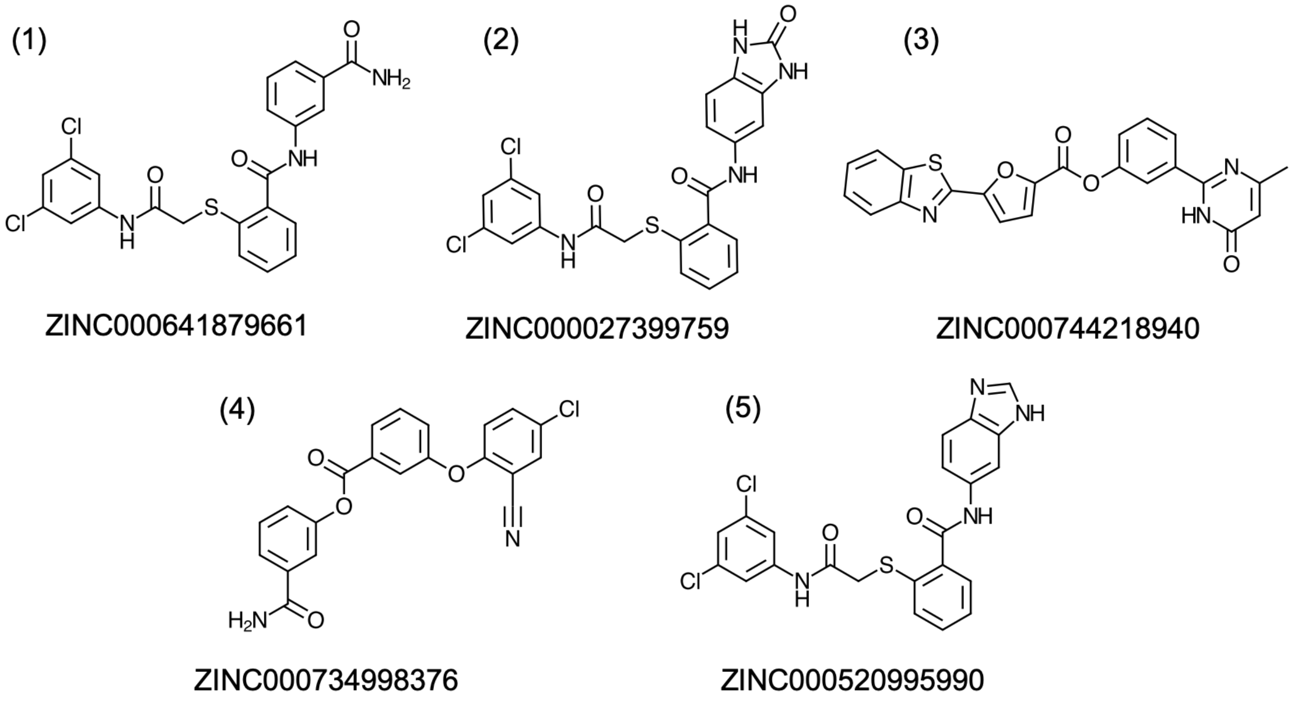

Figure 1. Chemical structures of top 5 compounds for 3CLpro target.

Docking predicted binding poses for the top 5 ligands are shown in Figure 2. All top 5 ligands are predicted to bind non-covalently in the substrate binding pocket occupying S1, S2 and S4 subsites. All top 5 ligands interact with S2 subsite residues, mainly to His41, Tyr54 and Met165 via hydrophobic contacts. Especially, the central aromatic ring in these ligands forms π-π stacking contacts with His41 which is one of the residues of the catalytic dyad consisting of His41 and Cys145. Various chemical moieties in most of these ligands also make key hydrogen bonding contacts with Gln192 and Phe140 of subsite S4 and S1 respectively. Apart from these interactions, the central carbonyl in most of these ligands also form a hydrogen bond with His164.

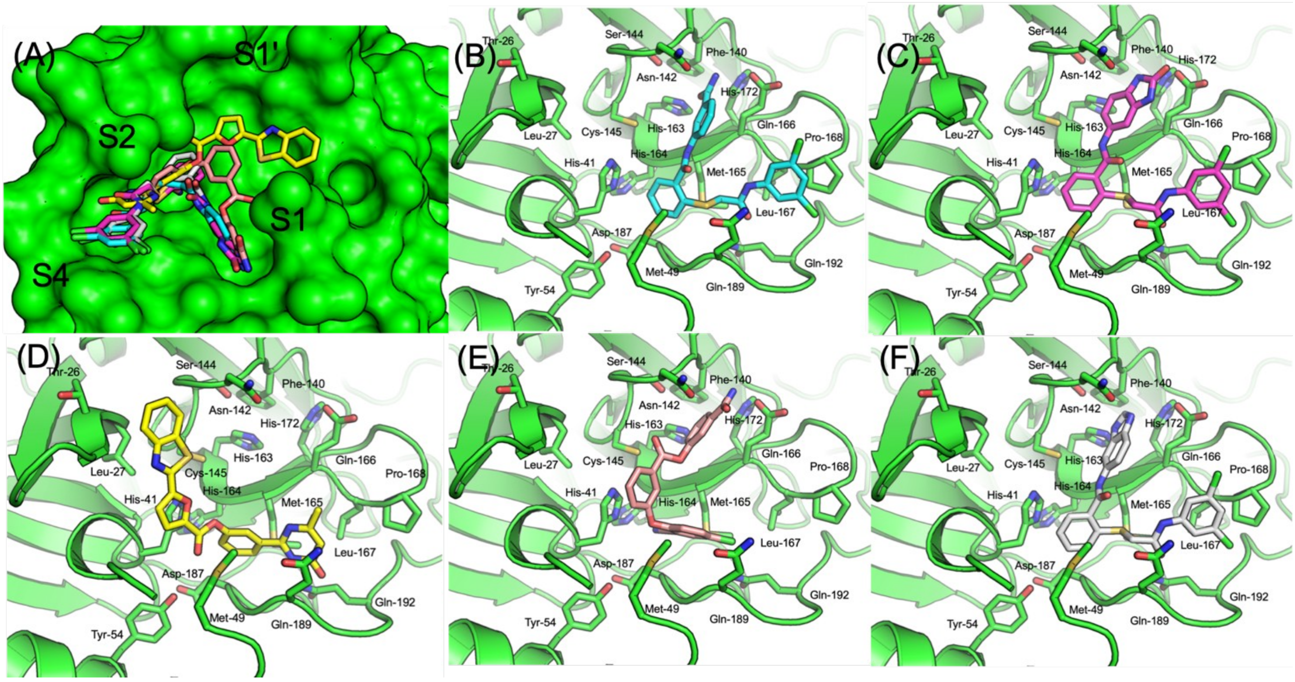

Figure 2. Docking predicted binding modes of top 5 compounds for 3CLpro. (A) Binding poses of top5 compounds in 3CLpro binding pocket in surface view. Various subsites are annotated (S1, S1', S2 and S4). (B) ZINC000641879661 (C) ZINC000027399759 (D) ZINC000744218940 (E) ZINC000734998376 (F) ZINC000520995990.

Top 5 ligands for Target 2 (ACE2/Spike protein):

The chemical structures of the top 5 ligands are shown in Figure 3. Three of the top 5 ligands (ZINC001811436992, ZINC001808095804 and ZINC001805011020) belong to the 1-[3-(Acetylaminomethyl)-2-(3-pyridyl)-1-pyrrolidinyl]-1-ethanone scaffold while the two others (ZINC000589881776 and ZINC000487361606) share a 2-(*o*-Chlorobenzoylamino)-3-(1*H*-indol-3-yl)-1-(propylamino)-1-propanone core. According to ChEMBL20, there is no known reported activity for any of these top5 compounds. All of these five compounds are make-on-demand compounds from WuXi and Enamine chemical vendors and have not been synthesized previously. For this target, we noticed a significant enrichment in compounds from the CAS antivirals library. Among the top 10,000 best scoring molecules, 3/10000 come from the CAS library while the random rate should be 5/100000 (i.e. there is an enrichment factor of 6 for this library and this target).

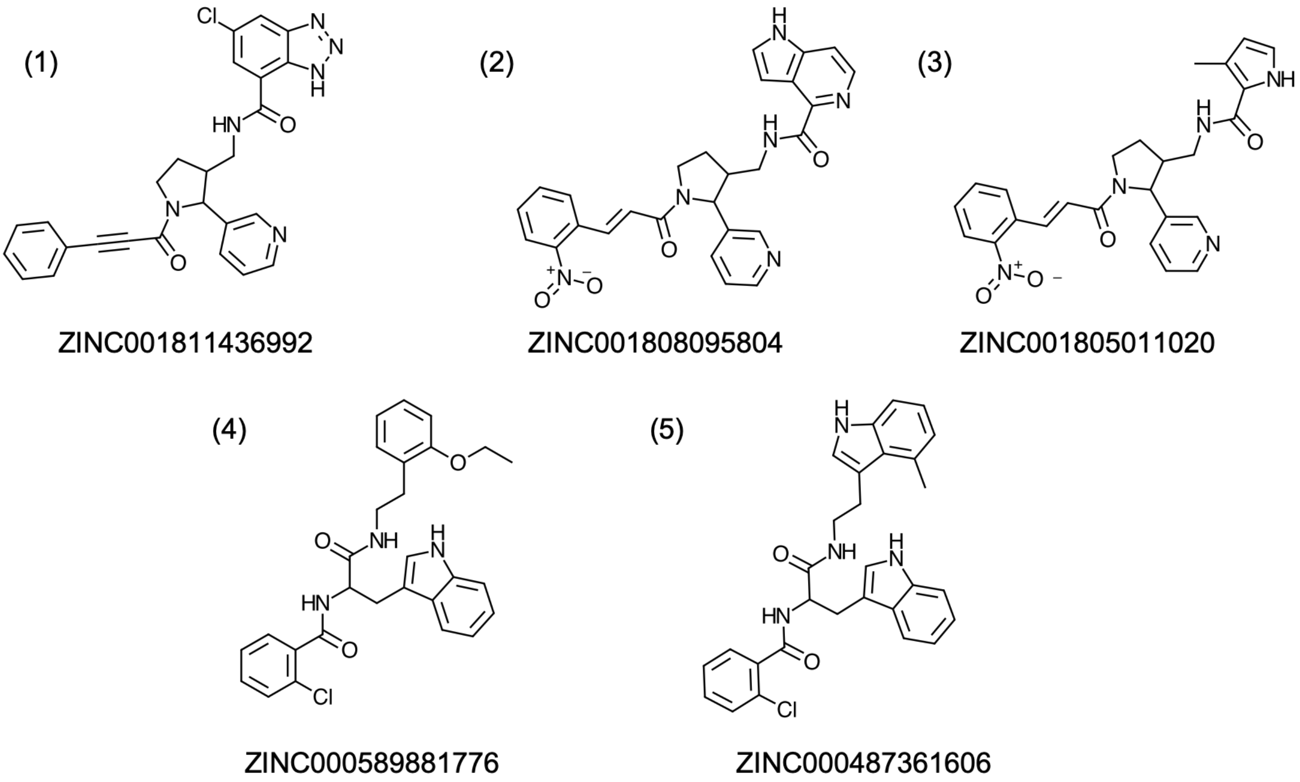

Figure 3. Chemical structures of top 5 compounds for ACE2/Spike protein-protein interface.

Docking predicted binding poses of top 5 ligands are shown in Figure 4. All top 5 ligands are predicted to bind in a binding pocket near to the ACE2 and receptor binding domain of spike protein binding interface as shown in Figure 4A. These ligands are expected to interrupt the ACE2/spike protein-protein interaction either by directly interfering with binding or by producing conformational changes unfavorable to spike protein binding. As the small molecule binding pocket near the interface of ACE2/spike is composed of hydrophobic amino acid residues including Phe 32, Phe40, Trp69, Leu73, Leu100 and Phe390, the top5 ligands interact with the pocket through hydrophobic contacts. Three of the top 5 ligands (ZINC001811436992, ZINC001808095804 and ZINC001805011020) that belong to 1-[3-(Acetylaminomethyl)-2-(3-pyridyl)-1-pyrrolidinyl]-1-ethanone scaffold bind to the pocket in a very similar manner as shown in Figure 4B, 4C and 4D. Benzotriazole, diaza indene and pyrrole ring in ZINC001811436992, ZINC001808095804 and ZINC001805011020 interact with Phe40, Trp69 and Phe390 via hydrophobic contacts. Additionally, these rings also form a hydrogen bond with the backbone of Phe390. Phenyl ring at the other end of this scaffold also interacts hydrophobically to Leu73 and Leu100. Other two compounds of top5 (ZINC000589881776 and ZINC000487361606) with 2-(*o*-Chlorobenzoylamino)-3-(1*H*-indol-3-yl)-1-(propylamino)-1-propanone core also bind in a similar manner as the other three except an indole ring that protrudes toward His378, Tyr385 and His401 is making hydrophobic interactions.

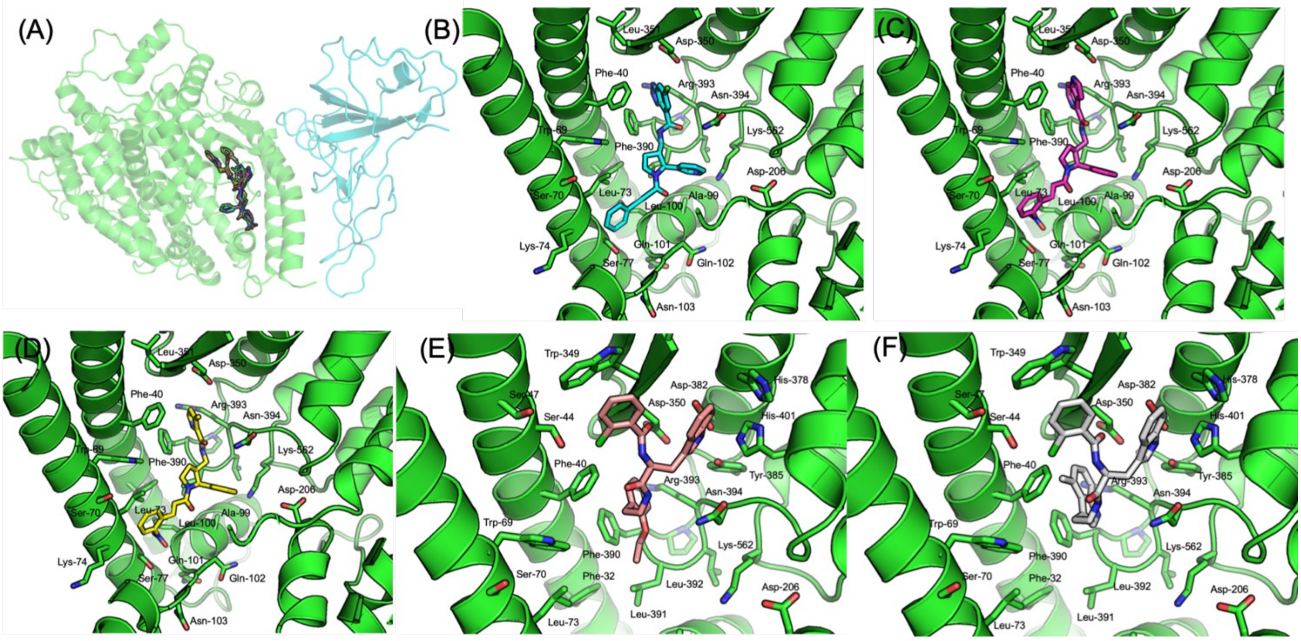

Figure 4. Docking predicted binding modes of top 5 compounds for ACE2/Spike PPI. (A) Binding poses of top5 compounds in ACE2/Spike protein binding interface pocket. ACE2 is shown as green cartoon while spike protein is shown as cyan cartoon. (B) ZINC001811436992 (C) ZINC001808095804 (D) ZINC001805011020 (E) ZINC000589881776 and (F) ZINC000487361606.

Top 5 ligands for Target 3 (PLpro):

The chemical structures of top 5 ligands are shown in Figure 5. According to ChEMBL20, there is no known reported activity for any of these top5 compounds. All of five compounds are make-on-demand compounds and have not been synthesized previously. It is interesting to note all top5 compounds harbors nitrogen heterocycles such as pyridine, benzimidazole, tetrazole, indole and thiazole. We further noticed that all of the compounds in top 10000 came from ZINC database.

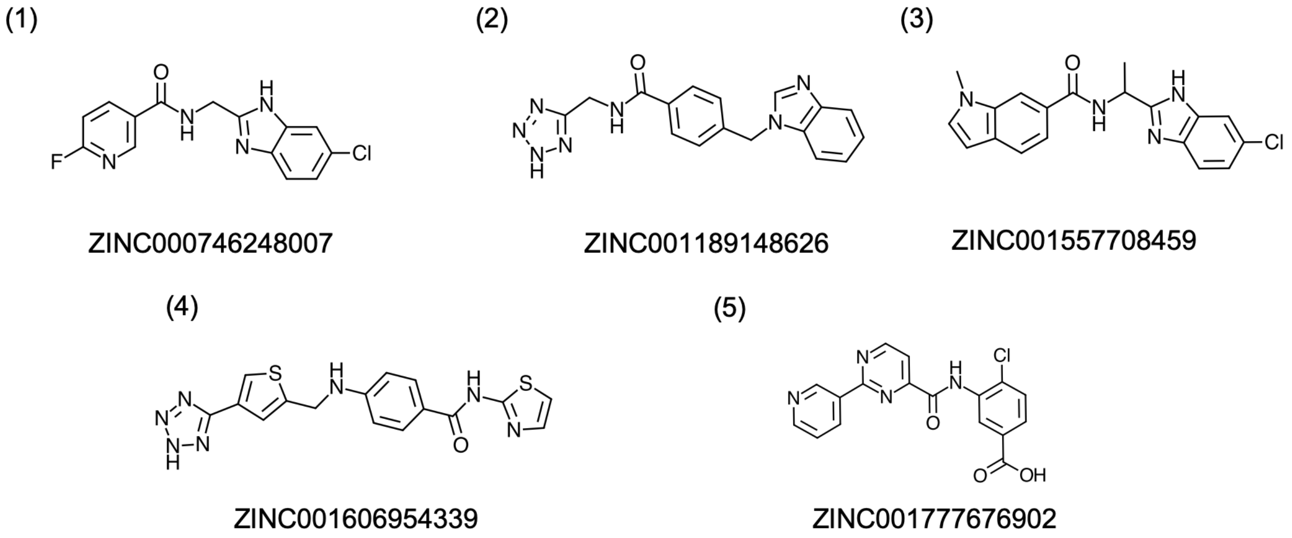

Figure 5. Chemical structures of top 5 compounds for PLpro protein target.

Docking predicted binding poses of top 5 ligands are shown in Figure 6. All top 5 ligands are predicted to bind to the more spacious S3/S4 pocket in a cleft leading to more restrictive S1/S2 pocket near to the catalytic triad residues (Cys111, His272 and Asp286). All top 5 ligands interact with S3/S4 subsite residues, mainly to Leu162, Met208, Pro247, Pro248, Tyr264, Tyr268, Cys270 and Tyr273 via hydrophobic contacts. Specifically, hydrophobic contacts with Tyr264 and Tyr268 are similar to the ones formed by several inhibitors of SARS-Cov PLpro. In addition to hydrophobic contacts, hydrogen bonding interactions with Asp164, Arg166, Thr301 and Asp302 were also prominent for these top 5 ligands.

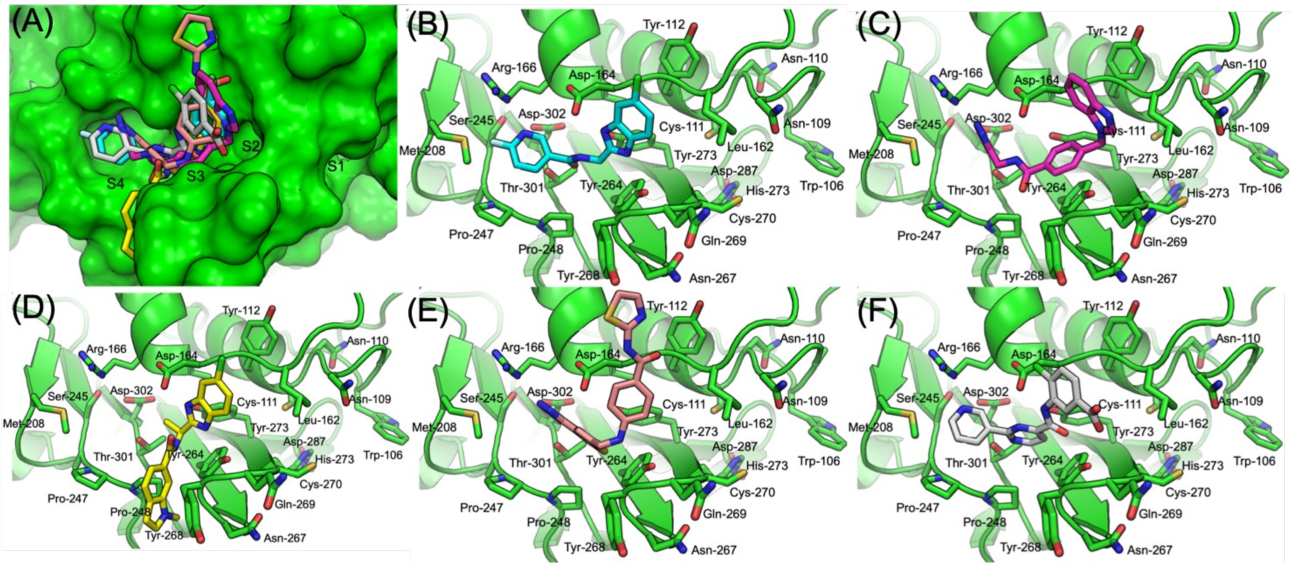

Figure 6. Docking predicted binding modes of top 5 compounds for PLpro. (A) Binding poses of top5 compounds in PLpro substrate binding pocket in surface view. Various subsites are annotated (S1, S2, S3 and S4). (B) ZINC000746248007 (C) ZINC001189148626 (D) ZINC001557708459 (E) ZINC001606954339 and (F) ZINC001777676902.

Other comments:

Most of the software used for Ligand-Based Virtual Screening are open-source and were written by us.

1. MolEnc: https://github.com/UnixJunkie/molenc a molecular encoder supporting Faulon fingerprints[Faulon2003] and atom pairs[Carhart1985]. Molenc is also a kind of swiss knife for chemoinformatics, with programs for clustering molecules (molenc_cluster), unique (molenc_uniq) or diversity filtering (molenc_filter), merging scored molecules (molenc_merge) and extracting selected molecules from SMI/SDF/MOL2 files (molenc_get).
2. LinWrap: https://github.com/UnixJunkie/linwrap a wrapper on top of liblinear[Fan2008]. LinWrap allows to train classification QSAR models using L2-regularized logistic regression and bagging[Breiman1996]. LinWrap models are fast to train and their screening frequency is also high.
3. RanKers: https://github.com/UnixJunkie/rankers reference implementation of the Vanishing Ranking Kernels method[Berenger2020]. A classifier method using a single bounded parameter in [0:1], optimized automatically and *including* an active applicability domain.
4. OPLSR: https://github.com/UnixJunkie/oplsr regression QSAR using Partial Least Squares[DeJong1993,Wehrens2007].

Acronyms:

LBVS: Ligand-Based Virtual Screening; PPI: Protein-Protein Interface. PPII: Protein-Protein Interface Inhibitor. SBVS: Structure-Based Virtual Screening.

Bibliography:

- [Carhart1985] Carhart, R. E., Smith, D. H., & Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. Journal of Chemical Information and Computer Sciences, 25(2), 64-73.
- [DeJong1993] De Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. Chemometrics and intelligent laboratory systems, 18(3), 251-263.
- [Breiman1996] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.
- [Faulon2003] Faulon, J. L., Visco, D. P., & Pophale, R. S. (2003). The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. Journal of chemical information and computer sciences, 43(3), 707-720.
- [Wehrens2007] Wehrens, R., & Mevik, B. H. (2007). The pls package: principal component and partial least squares regression in R.
- [Fan2008] Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A library for large linear classification. Journal of machine learning research, 9(Aug), 1871-1874.
- [Kumar2014] Kumar, A., Ito, A., Takemoto, M., Yoshida, M., & Zhang, K. Y. (2014). Identification of 1, 2, 5-oxadiazoles as a new class of SENP2 inhibitors using structure based virtual screening. Journal of Chemical Information and Modeling, 54(3), 870-880.
- [Kozakov2015] Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., ... & Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature protocols*, *10*(5), 733.
- [Labbe2016] Labbé, C. M., Kuenemann, M. A., Zarzycka, B., Vriend, G., Nicolaes, G. A., Lagorce, D., ... & Sperandio, O. (2016). iPPI-DB: an online database of modulators of protein–protein interactions. Nucleic acids research, 44(D1), D542-D547.
- [Jiang2016] Jiang, X., Kumar, A., Liu, T., Zhang, K. Y., & Yang, Q. (2016). A novel scaffold for developing specific or broad-spectrum chitinase inhibitors. Journal of chemical information and modeling, 56(12), 2413-2420.
- [Oboyle2016] O’Boyle, N. M., & Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. Journal of cheminformatics, 8(1), 1-14.
- [Matsuoka2017] Matsuoka, M., Kumar, A., Muddassar, M., Matsuyama, A., Yoshida, M., & Zhang, K. Y. (2017). Discovery of fungal denitrification inhibitors by targeting copper nitrite reductase from Fusarium oxysporum. Journal of chemical information and modeling, 57(2), 203-213.
- [Berenger2020] Berenger, F., & Yamanishi, Y. (2020). Ranking Molecules with Vanishing Kernels and a Single Parameter: Active Applicability Domain Included. Journal of Chemical Information and Modeling
- [Jiang2020] Jiang, X., Kumar, A., Motomura, Y., Liu, T., Zhou, Y., Moro, K., ... & Yang, Q. (2020). A series of compounds bearing a dipyrido-pyrimidine scaffold acting as novel human and insect pest Chitinase inhibitors. Journal of Medicinal Chemistry, 63(3), 987-1001.

- [Zhang2020] Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. & Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. Science, 368, 409-412.