Team name	**AI Winter is Coming**
Team member(s) (firstname lastname; …)	Roman Zubatyuk, Evgeny Gutkin, Phil Gusev, Chamali Narangoda, Hatice Gokcan, Shuhao Zhang, Zhen Liu, Maria Kurnikova, Olexandr Isayev
Affiliation	Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA
Contact email	olexandr@olexandrisayev.com
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	Mpro, PLpro, TMPRSS2

Section 1: methods & metrics

Describe what methods you have used, how they are independent from one another, what your workflow was, how you performed the cross-correlation between your methods. If applicable, please report estimated performance metrics of your methods, such as accuracy, sensitivity, false-discovery rate, etc., and how those metrics were obtained (e.g. cross-validation). Please provide key references if available.

We screened 4.59 billion of molecules against three (Mpro, PLpro, TMPRSS2) targets. All molecules were triaged according the following protocol:

1. All 4.59B molecules were screened with three independent methods: AutoDock Vina, OpenEye Fred and Global QSAR model (see detailed description below)
2. Hits were triaged through 21 ML models for ADME, off-target activities (kinases & GPCRs) and solubility (LogS). Molecules with low solubility and/or more than four liabilities were discarded.
3. Remaining molecules were filtered through Local QSAR model (see detailed description below)
4. Final ranking was determined by consensus of Local QSAR model and penalized composite docking score.

Docking

Ligands were docked into multiple protein targets using AutoDock Vina[1] software with exhaustiveness setting of 16. For each docking experiment, 5 highest score poses were saved. The poses were re-scored with OpenEye ChemGauss4 (CG4) score.[2] For each pose, the consensus score was calculated as average of Vina and CG4 scores.

We have observed that for some protein targets and ligands, docked poses occupy different regions of the pocket. Therefore, we have identified active site residues that interact with native ligands. Several X-ray structures of Papain-Like protease (PLpro) and main protease (Mpro) with different ligands have been investigated. Two-dimensional ligand-protein interaction diagrams of ligand bound proteins (PDB IDs for PLpro: 6WRH, 3E9S, 3MJ5, 6WX4, 4OVZ, 4OW0, 6WUU, and PDB IDs for Mpro: 6Y7M, 6LU7, 7BQY, 6WNP) have been plotted using LigPlot^{+}.[3] It was observed that interactions with four residues in PLpro (Gly163, Asp164, Tyr268, and Gln269) have been conserved in all cases. In case of main protease, four residues (Gly143, Cys145, His163 and His 164) have been observed to interact with the ligands. A similar procedure is followed for TMPRSS2, but with proteins that share structural similarities with it, such as Complement factor D or Factor XIa. Six different protein-ligand complexes have been selected (PDB IDs: 2VNT, 6FUG, 4Y8Y, 5WCM, 5QCN, 5UGD) and

two-dimensional ligand-protein interaction diagrams have been compared with TMPRSS2 model. Three residues (Asp176, Ser177, and Ser182) have been identified as possible regions to interact with a ligand.

For each docked pose we calculated the minimum distance from an atom of ligand to the centroids of all active residues and applied penalty to the consensus docking score and calculated penalized composite docking score S^* using following formula:

$$S^* = S + 2.0 \sum_{k}^{M} \frac{n(d_k)}{N}$$

where S is consensus docking score (average of Vina and CG4 scores), M is number of active residues (see Table 2 in Targets section), $n(d_k)$ is rank of minimum distance d_k from the ligand to residue for this particular protein target, and N is number of ligands docked. This ensures that the poses in which molecule is located further from one or several active residues compared other ligands, will get penalty to the consensus docking score up to 2 kcal/mol per active residue, e.g. up to 8 kcal/mol for MPro and PLPro, and up to 6 kcal/mol for TMPRSS2. For each protein conformation, the best penalized score among 5 poses was selected. The final composite docking score was calculated as average of scores for several protein conformations.

In order to accelerate screening of billions-size database with our composite docking method, we have empoyed ML model to predict docking score for particular target from 2D molecular structure of ligand. This approach is known as Deep Docking[4]. We have developed AIMNet-2D deep neural network model, which based on AIMNet neural network potential model[5], which is capable to predict energies, charges and volumes for non-equilibrium conformers of organic molecules. Original AIMNet model builds atomic features based on distances to the neighbors, iteratively updates the features using message passing, and predicts atom property from the atomic features. The adapted AIMNet-2D model operates on molecular graph and uses topological distances instead, up to 5 bonds. The composite docking score is predicted as a sum of atomic contributions. Similar to Deep Docking approach, we have applied active learning technique to construct dataset iteratively, selecting molecules with high docking score predicted on the model trained on previous iteration. As a seed dataset, we used random sample of 100k molecules from *in-virto* subset of ZINC dataset. For every subsequent iteration we have added 100k random molecules from combined 4.5B dataset with predicted docking score less then 10th percentile of docking score distribution in the current version of dataset. Figure 1 shows enrichment of training dataset with molecules which have high docking scores with AL iterations. For all three targets, the AIMNet-2D model is capable to predict composite docking score with RMS error of 0.75 kcal/mol, and in evaluation mode can prediction for 50M molecules per hour on a Nvidia P100 GPU.

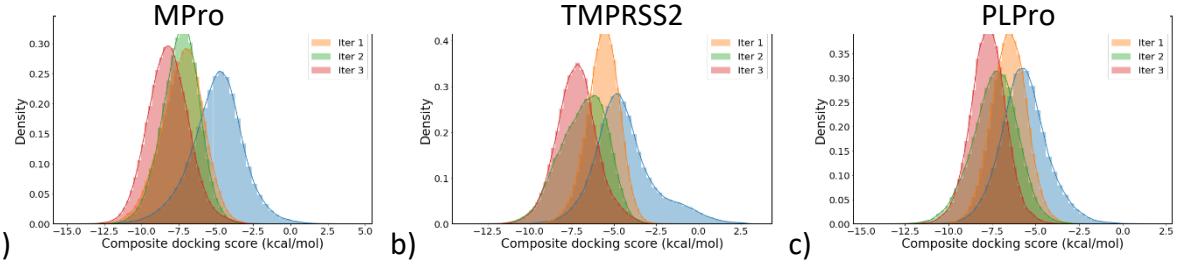

Figure 1. The distribution docking scores in final training datasets, which were selected on every active learning iteration.

ML models for bioactivity (QSAR)

ML model building.
All models were trained within two nested five-fold cross validation loops using xGBoost. All splits were random. Final scoring were done by simple averaging of five predictions from the external CV loop. Internal CV loop was used to perform hyperparamters search and variable selection for the corresponding fold using the protocol described our recent paper.[6]

Mpro and PLpro data preparation. To model MPro and PLpro we used historical data of putative inhibitors screening SARS-CoV from publicly available databases: BindingDB, PubChem, ChEMBL.

Proteins of SARS-CoV and SARS-CoV-2 are encoded as a polyprotein 1ab which is digested to active forms of proteins (including Mpro and PLpro) during viruses' live cycle. But binding affinity measurements performed on isolated Mpro or PLpro or other SARS-CoV proteins thus a common misannotation presented in those databases which occurs due to automatic cross-annotation by sequence similarity: it is common that protein record of SARS-CoV's data has correct description but as a sequence it stores full 1ab sequence or it stores correct sequence of isolated protein but due to high similarity of some regions of PLpro and Mpro it still leads to misannotation.

To resolve this issue, we performed search by sequence similarity with query sequence of Mpro, PLpro, 1ab of SARS-CoV and SARS-CoV-2 in BindingDB and a full text search for terms "SARS", "Coronavirus", "SARS-CoV" in PubChem and ChEMBL. After that we manually verified and resolved each hit (molecule) in BindingDB and hit (assay record) for PubChem and ChEMBL.

TMPRSS2 data preparation. TMPRSS2 is a member of a family of serine proteases, thus despite lack of data for the TMPRSS2, based on sequence similarity among presented in BindingDB we've selected set of representative proteins (Suppressor of tumorigenicity 14 protein (ST14_HUMAN),Tryptase alpha/beta-1 (TRYB1_HUMAN),Coagulation factor XI (FA11_HUMAN),Serine protease hepsin (HEPS_HUMAN), Plasma kallikrein (KLKB1_HUMAN), Plasminogen (PLMN_HUMAN), Transmembrane protease serine 6 (TMPS6_HUMAN) such each pair of them share at least one inhibitor with pIC50 10nM or less (pX 6 or more) whose data from BindingDB we used to build a QSAR model.

Global and local QSAR models.
For each collected dataset (to be referred as *local*) we build a complementary one, referred as *global*. Global QSAR model is well suited for ultra-large screening campaigns. While local is more suited for refinement of hits. To construct a global dataset we augmented local dataset with synthetic presumed inactive molecules. Using the MinMax picker algorithm several thousands of presumed inactive molecules were selected from libraries ZINC-diverse and ChEMBL. The size of each global dataset was selected in a way to maximize its size yet keep QSAR model's performance on a reasonable level.

Table 1. Statistics of datasets for QSAR modeling

Target	Local dataset size	Global dataset size
Mpro	615	5125

PLpro	205	1576
TMPRSS2	1358	24909

Data Curation

For the targets of interest data was integrated from BindingDB, PubChem and ChEMBL 25. Bioactivities were extracted for pChEMBL activity values -Log(IC50/EC50/Ki/Kd) of 10 µM or better, with ChEMBL CONFIDENCE_SCORE of 6 or greater for 'binding' or 'functional' assays. Due to conflicting naming schemes all target datasets were integrated by Uniprot IDs.

Each target dataset was curated according to well-established best practices.[7] Structural standardization, the cleaning of salts, and the removal of mixtures, inorganics, and organometallics was performed using ChemAxon software. In the case of replicate compounds, InChI Keys were generated. For replicates with the same activities in a given assay, a single representative compound was selected for inclusion into the training set. For replicates with the different activities (> 1 log unit) in a given assay, all compounds were excluded.

ADME/tox

In drug development, early assessments of pharmacokinetic and toxic properties are important stepping stones to avoid costly and unnecessary failures. We have used our ML method to develop 13 absorption, distribution, metabolism, excretion, and toxicity (ADMET) prediction models.

These models quickly assess some of the most important properties of potential drug candidates, including their cytotoxicity, mutagenicity, cardiotoxicity, drug-drug interactions, microsomal stability, and likelihood of causing drug-induced liver injury.

Additionally we considered solubility (logS) and off target activity (9 models). Off targets include human kinases (AKT1, AKT2, AKT3, AURKA, AURKB) and GPCRs (CHRM1, CHRM2, CHRM3, HRH1)

Section 2: targets

Describe for each protein target: why you chose it, from which source you obtained it (e.g., insidecorona.net / covid.molssi.org / rcsb.org) and why this is the best quality structure, if any pre-processing (e.g., energy minimization, residue correction, alternative folding, …) was performed.

The following SARS-CoV-2 proteins were considered as potential molecular targets: 3C-like protease, also referred to as main protease (Mpro), papain-like protease (PLpro), spike glycoprotein (S), and RNA-dependent RNA-polymerase (RdRp). These proteins are essential for the virus life cycle and some compounds binding to these proteins were found to inhibit viral entry (S inhibitors) and replication (Mpro, PLpro and RdRp inhibitors).[8] Cellular transmembrane serine protease 2 (TMPRSS2) was also considered since its inhibitors were demonstrated to block SARS-CoV-2 cell entry.[9] It should be noted that the presence of oligosaccharide chains and RNA nucleotides in the structures of S and RdRp complicates modeling of these systems by using molecular dynamics compared to apoproteins and protein-ligand complexes. Based on an evaluation of protein crystal structures presented in Protein Data Bank (PDB), three molecular targets were selected for conducting structure-based virtual screening: Mpro, PLpro and TMPRSS2.

Target 1: Mpro. The crystal structures of MPro were initially ranked by their resolution and the models with the resolution > 3A were excluded from further consideration. The retained structures were structurally aligned on the binding site residues and root-mean-square deviation (RMSD) on Cα and all non-hydrogen atoms of the binding site residues was calculated. The following structures were selected to provide conformational diversity of protein backbone and sidechains of the binding site residues for ensemble docking: an apo-form of the protein (PDB ID 6M03), the protein bound covalently to FDA-approved antiviral drug Boceprevir (PDB ID 6WNP) and the protein bound covalently to a peptide-like inhibitor N3 (PDB ID 7BQY). To obtain equilibrated structures, explicit solvent molecular dynamics (MD) simulations were performed for each of these crystal structures. Only the ligand-bound domain of subunit 1 (chain A; residues 5 to 198) and the fragment of N-terminal loop from subunit 2 (chain B; residues 1 to 10) were retained and all other protein atoms were removed. The truncated structures were protonated and prepared for MD simulations using tLeap in Amber 18. In structures 6WNP and 7BQY, the covalent bond between Cys145 and the inhibitor was removed and the ligand was converted to its original state. All structures were solvated with water using a periodic cuboid box with the minimum distance between the edge of the box and any solute atom set to 10 A. AMBER ff99SB-ILDN[10], GAFF[11] and TIP3P[12] parameters were used for protein, ligands and water correspondingly. Ligand charges were assigned using AM1-BCC method.[13] All simulations were performed in Amber 18 with GPU acceleration on NVIDIA GeForce RTX 2080 graphics card. A simulation time step of 2 fs was used, and all hydrogen bonds were constrained via SHAKE.[14] Periodic boundary conditions were applied in all directions with a cutoff radius of 8 Å. Electrostatic interaction calculations were performed using Particle Mesh Ewald (PME) method. Langevin thermostat and Berendsen barostat, as implemented in AMBER, were used to maintain temperature and pressure respectively.

MD simulations were performed using the following protocol:

1. 1000 steps of energy minimization with the steepest descent method;
2. 250 ps of heating in canonical ensemble (NVT) with harmonic restraints (40 kcal mol^{-1} $Å^{-2}$) imposed on all heavy atoms;
3. 1 ns of equilibration in isothermal-isobaric ensemble (NPT) with harmonic restraints (40 kcal mol^{-1} $Å^{-2}$) imposed on protein Cα atoms;
4. 1 ns of simulations in canonical ensemble (NVT) with harmonic restraints (40 kcal mol^{-1} $Å^{-2}$) imposed on protein Cα atoms.

Obtained MD trajectories were visually inspected and evaluated in terms of stability of protein binding site residues and ligand binding mode. For production trajectories, the frame with the minimum RMSD on CA atoms with respect to the average structure was selected as the equilibrated structures and were used as receptors for molecular docking.

Target 2: PLpro. An apo structure of SARS-CoV-2 PLpro (PDB ID 6WRH) and four structures of SARS-CoV PLpro in complex with non-covalent inhibitors were selected (PDB IDs 3E9S, 3MJ5, 4OVZ, and 4OW0). Since the binding site residues located within 6 A of ligands are identical between SARS-CoV and SARS-CoV-2 PLpros, we consider it reasonable to use SARS-CoV structures as receptors for docking. The residue 111 which is presented by serine in the SARS-CoV-2 PLpro structure and by cysteic acid in SARS-CoV PLpro structures was mutated to cysteine according to the SARS-CoV-2 PLpro sequence using PYMOL 1.8.4. The preparation of structures to MD simulations was performed as described in the previous section. Four coordinate bonds between sulfur atoms of the Cys189,

Cys192, Cys224, Cys226 and zinc ion were created in tleap and zinc AMBER force field was used to obtain parameters for the zinc metal center. For all structures, the MD simulations and evaluation of trajectories were performed using the protocol described in the previous section. In addition, 100 ns-long MD trajectories in NVT were conducted to ensure that the closed state of a so-called blocking loop is stable in the presence of ligands for the structures 3E9S and 3MJ5. For three structures (3E9S, 4OVZ and 4OW0) three water molecules located in between ligand and the residues Asp165, Arg167, Tyr274, Thr302, and Asp303 were found to retain their positions and hydrogen bonds with these protein residues during MD trajectories. The equilibrated structures 3E9S with and without these water molecules were used as two separate receptors for docking while a single equilibrated structure was used for each of the structures 4OVZ, 4OW0 (with the water molecules) and 6WRH and 3MJ5 (without them) resulting in six PLpro receptors for docking in total.

Target 3: TMPRSS2. The amino acid sequence of the catalytic chain of TMPRSS2 (residues 256 to 492) was retrieved from the UniProt database (Accession No. O15393) and the Protein Data Bank (PDB) was searched for available three-dimensional structures of homologous serine proteases. The Protein BLAST program at NCBI was used for the sequence similarity analysis [ref]. Two serine proteases that belong to the same subfamily as TMPRSS2 – hepsin/TMPRSS1 (sequence similarity: 58%) and enteropeptidase/TMPRSS15 (sequence similarity: 61%) – were selected as templates for homology modeling. Two homology models of residues 260-489 of TMPRSS2 were generated in SwissModel[15] using hepsin (PDB ID 5CE1) and endopeptidase (PDB ID 4DGJ) as templates. The models were protonated and prepared for MD simulations using tLeap in Amber 18. Disulfide bonds were enforced between residues Cys281 and Cys297, Cys410 and Cys426, and Cys437 and Cys465. His296 was protonated at the delta nitrogen in order to maintain the geometry of the preserved catalytic triad (residues His296, Asp345 and Ser441) of serine proteases. The models were solvated, minimized, and heated using the protocol described in the previous sections, followed by a 6 ns equilibration in NPT while slowly releasing the restraints on the protein from 40 to 0.2 kcal mol^{-1} $Å^{-2}$. Unrestrained NVT simulations were then carried out for 20 ns. The stability of the models was assessed by monitoring the RMSD of protein Cα atoms with respect to initial structures. Based on the stability of the overall protein and the binding site residues during unrestrained MD simulations, the hepsin-based model was selected as the receptor for docking. In order to account for variations in side-chain conformations, two representative structures were selected from the unrestrained MD trajectory.

Table 2.

Target	Protein structures	Protein conformations	Active residues
Mpro	3M03, 6WNP, 7BQY	3M03, 6WNP, 7BQY	Gly143, Cys145, His163, His 164
PLPro	6WRH	6WRH, 6WRH (apo)	Gly163, Asp164, Tyr268, Gln269
TMPRSS2	5CE1 (Homology model)	$TMPRSS2^{(h)}$_1, $TMPRSS2^{(h)}$_2	Asp435, Ser436, Ser441

Section 3: libraries

Describe which libraries you have used, how they were combined, if any compounds were removed / added, why additions are relevant, any unique features of your library, etc. Please provide the sources you obtained the libraries from (if publicly available). Describe the procedure of data preparation (removal of duplicates, standardization, etc). Indicate if different libraries were used for different targets, and why. If possible, provide a download link to your version of the library.

Our overall screening library was 4.59B of molecules, see Table 3. All SMILES were curated according to our well-established best practices.[7] Structural standardization, the cleaning of salts, and the removal of mixtures, inorganics, and organometallics was performed using ChemAxon software.

Table 3. List of molecule databases used for virtual screening.

Name	Number of molecules	Source URL
WuXi GalaXi	1,763,730,226	
ZINC	1,488,277,372	http://files.docking.org/2D
Enamine REAL	1,211,723,743	https://enamine.net/library-synthesis/real-compounds/real-database
PubChem & ChEMBL	74,877,957	ftp://ftp.ncbi.nlm.nih.gov/pubchem
Mcule Purchasable (Full)	45,472,755	https://mcule.com/database
Merck	5,049,120	From Challenge organizers
ChemSpace «In Stock»	3,147,253	https://chem-space.com/compounds
CAS Antiviral DB	27,525	https://www.cas.org/covid-19-antiviral-compounds-dataset
Total:	4,592,305,951	

Section 4: results

Briefly describe you key findings, any interesting trends in your data, a description of your top 5 compounds for each target. If possible, provide a link to a code and/or data repository. Please do not submit randomly selected compounds!

Results:

The set of hits selected with combination of composite docking and QSAR modeling (40k for MPro, 40k for PLPro and 25k for TMPRSS2) we have performed final docking experiment. The composite docking score used in Deep Docking approach was based on ligand poses produced Autodock Vina. Now, we utilize OpenEye FRED poses and docking score, as another independent method. Similar to Deep Docking campaign, we used several protein conformations for each target, and for each ligand selected best CG4 score among conformations. Final ranking of hits was performed using consensus of composite docking score, FRED docking score and QSAR score. For each score, we calculated ligand rank (from best to worst) and percentile in score distribution for each target. Figure 2 shows rank correlation for all three scores. Overall, there is little, if any, rank correlation between the scores. The final score was calculated as average of ranks for Deep Docking, FRED docking and QSAR.

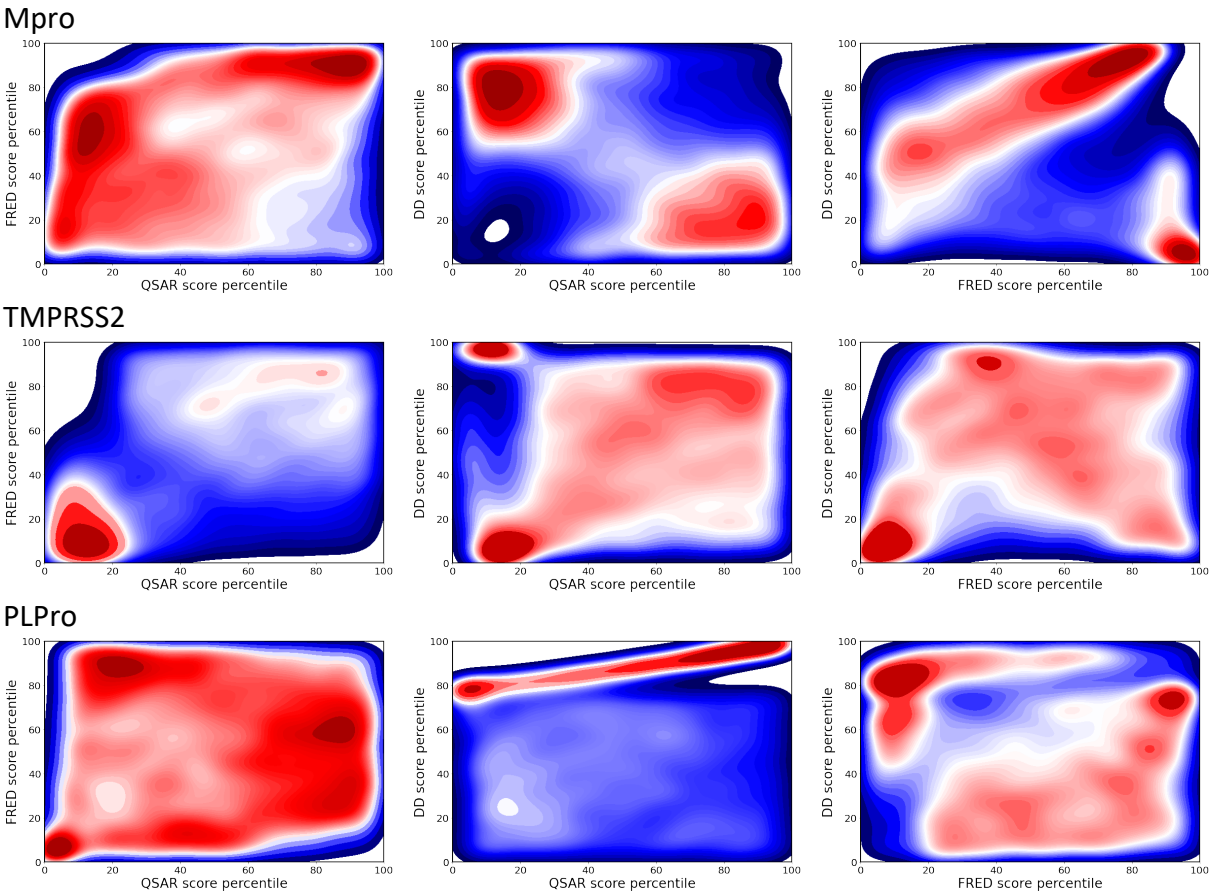

Figure 2. Pair density plots of rank scores for three protein targets, smoothed with Gaussian kernel.

Other comments:

Supplementary Material

Homology modeling of TMPRSS2

A

```
TMPRSS2   IVGGESALPGAWPWQVSLHVQNVHVCGGSIITPEWIVTAAHCVEKPLNNPWHWTAFAGIL 315
5ce1.1.A  IVGGRDTSLGRWPWQVSLRYDGAHLCGGSLLSGDWVLTAAHCFPERNRVLSRWRVFAGAV 222

TMPRSS2   RQSFMFYGAGYQVEKVISHPNY------DSKTKNNDIALMKLQKPLTFNDLVKPVCLPNP 369
5ce1.1.A  AQASP-HGLQLGVQAVVYHGGYLPFRDPNSEENSNDIALVHLSSPLPLTEYIQPVCLPAA 281

TMPRSS2   GMMLQPEQLCWISGWGATEEKGKTSEVLNAAKVLLIETQRCNSRYVYDNLITPAMICAGF 429
5ce1.1.A  GQALVDGKICTVTGWGNTQYYGQQAGVLQEARVPIISNDVCNGADFYGNQIKPKMFCAGY 341

TMPRSS2   LQGNVDSCQGDSGGPLV----TSKNNIWLIGDTSWGSGCAKAYRPGVYGNVMVFTDWIY 485
5ce1.1.A  PEGGIDACQGDSGGPFVCEDSISRTPRWRLCGIVSWGTGCALAQKPGVYTKVSDFREWIF 401

TMPRSS2   RQMR 489
5ce1.1.A  QAIK 405
```

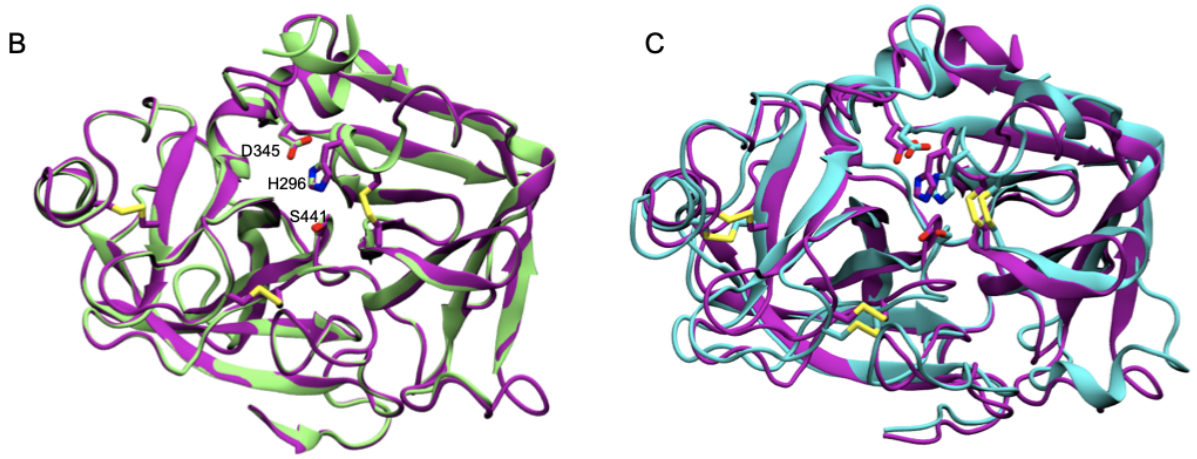

Figure: Homology modeling of TMPRSS2 using TMPRSS1 (PDB ID 5CE1) as the template. A) Sequence alignment between residues the catalytic chains of TMPRSS2 and hepsin/TMPRSS1. B) Overlap of the homology model (purple) and the template (green). The model contains residues 260 to 489 of TMPRSS2. Residues of the catalytic triad of TMPRSS2 are labeled. RMSD of all overlapping Cα atoms is 1.35 Å. C) Overlap of the initial (purple) and simulated representative (cyan) structures of the homology model. RMSD of all Cα atoms: 1.98 Å, RMSD of Cα atoms excluding flexible loop regions (RMSD of residues 260 to 298, 306 to 313, 327 to 333, 345 to 351, 378 to 384, 397 to 405, and 425 to 489): 1.55 Å.

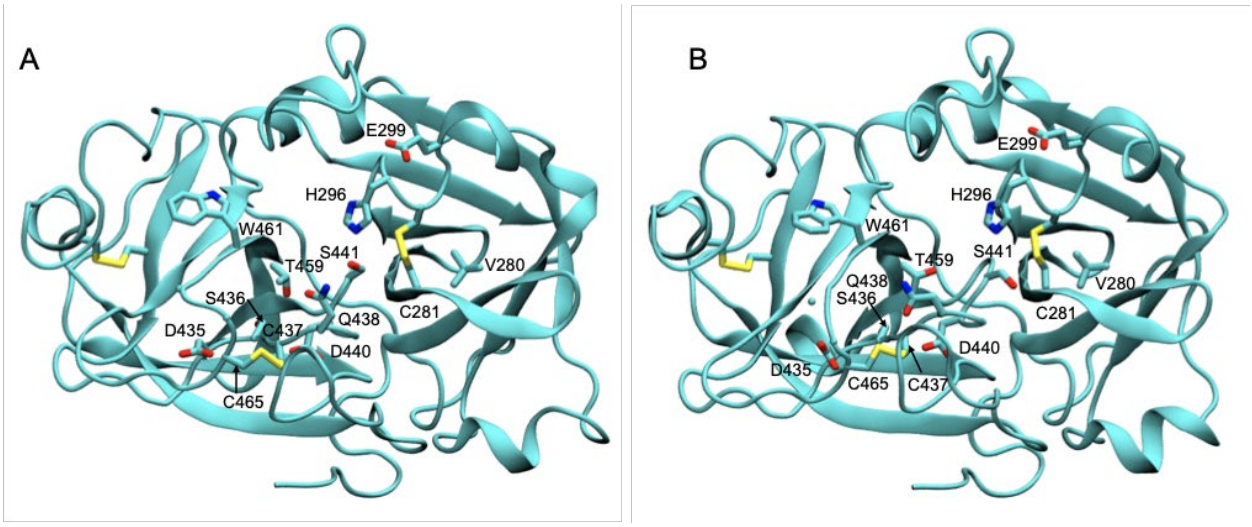

Figure: TMPRSS2 conformations used for docking. Two representative structures A) TMPRSS2$^{(h)}$_1 and B) TMPRSS2$^{(h)}$_2 were selected from the MD simulation trajectory of the TMPRSS2 homology model to represent variations in side-chain conformations of binding site residues.

References

1. 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.
2. McGann, M. FRED Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **51**, 578–596 (2011).
3. Laskowski, R. A. & Swindells, M. B. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. *J. Chem. Inf. Model.* **51**, 2778–2786 (2011).
4. Gentile, F. *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).
5. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **5**, eaav6490 (2019).
6. Cichonska, A. *et al.* Crowdsourced mapping of unexplored target space of kinase inhibitors. *bioRxiv* (2020) doi:10.1101/2019.12.31.891812.
7. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J Chem Inf Model* **50**, 1189–1204 (2010).
8. McKee, D. L., Sternberg, A., Stange, U., Laufer, S. & Naujokat, C. Candidate drugs against SARS-CoV-2 and COVID-19. *Pharmacol. Res.* **157**, 104859 (2020).
9. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
10. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins Struct. Funct. Bioinforma.* NA-NA (2010) doi:10.1002/prot.22711.
11. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
12. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
13. Jakalian, A., Bush, B. L., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **21**, 132–146 (2000).
14. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. . Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
15. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).