Team name	AI4Science
Team member(s) (firstname lastname; ...)	Katarina Elez; Tim Hempel; Robin Winter; Tuan Le; Lluís Raich; Simon Olsson; Frank Noé
Affiliation	Freie Universität Berlin, Department of Mathematics and Computer Science, Germany
Contact email	frank.noe@fu-berlin.de
Contact phone number (optional)	+49 178 6074029
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc...) | 3 required	TMPRSS2, TMPRSS2, TMPRSS2

Section 1: methods & metrics

High-throughput Molecular dynamics and Markov modelling:

As no experimental structure is available for our target TMPRSS2, we initialize our simulations using a homology model based on an X-ray structure of Enteropeptidase-1 (PDB: 3W94) presented in [1]. To relax possible artifacts of the initial homology model [13], a total of approximately 0.5 milliseconds of all-atom molecular dynamics (MD) simulations were run with the apo protein, and complexes with the confirmed SARS-CoV-2 inhibitors Camostat and Nafamostat [15-16] using docked poses from [1]. Our target construct includes the full activated form of the protease domain of TMPRSS2 (residues 256 to 490), responsible for the function used by SARS-CoV-2 during cell-entry. We run a standard MD protocol mirroring physiological conditions with OpenMM 7.4.0 [12] using the CHARMM 36 force field (2019 version) [14]. This approach accounts for both target flexibility and possible induced fit effects.

In order to enter a rapid production cycle, we used the first 50 microseconds of simulations to select a total of 20 representative protein conformations from *apo* and drug-bound set of MD trajectories, as receptor structures for drug docking. The remaining 450 microseconds were used for other purposes, but we confirmed that the protein structure was overall stable and the 20 target structures could be kept. To select target structures, we built hidden Markov models (HMM) from the simulation data, using structural features (distances and torsions), including the enzyme charge relay system, and the S1 binding pocket. For simulations with camostat/nafamostat we also included distances between their chemical moieties and the target residues from above. We select representative conformations from the *apo* simulation HMM with 10 states and from each of the camostat and nafamostat HMM, with 5 states each. All docking runs were done with all 20 target structures.

Initial drug library*:* We first gathered a total of ~24,000 drug candidates as follows (see Section 3 for ZINC, DrugBank, ChEMBL library description):

1. ~9,000 ZINC compounds that were structurally most similar (Tanimoto similarity > 0.6 based on extended connectivity fingerprints) to 18 lead compounds reported to inhibit TMPRSS2 or other Trypsins [1-5]. To ensure diversity, the 9,000 compounds were obtained by

choosing k-means cluster representatives with highest Tanimoto similarity to the 18 lead compounds.

2. ~6,000 structurally diverse ZINC compounds containing a guanidinium group.
3. ~6,500 molecules from DrugBank (all molecules with molecular weight < 550 Da)
4. ~2,500 ChEMBL molecules that strongly inhibit serine proteases.

All these ~24,000 initial compounds were docked and scored (see below) to all 20 target structures before entering the active learning procedure to expand the library (see below).

Active learning cycle for library expansion*:* We performed three cycles of active learning as follows:

1. Define a set of new molecules under consideration from ZINC that are structurally close to the already scored molecules (Tanimoto similarity > 0.4) but diverse to each other.
2. Using the machine-translation-based autoencoder model described in [6], encode these molecules under consideration to a continuous latent space representation.
3. Using the available scored compounds, train a kernel Support Vector Regressor (sklearn) in this latent space to predict the docking score of all new molecules under consideration.
4. Assembled a list of ~7,000 diverse compounds with the highest predicted scores for another docking round.

The Pearson correlation between predicted and actual docking scores of compounds varied between 0.69 and 0.77. Overall, a total of ~10,000,000 compounds were considered by the machine learning model and ~21,000 new compounds were added to the drug library, resulting in a total library size of ~45,000 compounds.

Structure preparation and flexible docking:

We retrieved 3D structures (as mol2 files) from the ZINC database (reference molecule) when available. We generated 3D structures of other candidates from SMILES strings, as a single low energy conformer, with the optimal ionization states at pH 7.05 using LigPrep [7]. To prepare the receptors and the ligands for docking, we assigned partial charges and AutoDock 4 atom types to the structures using MGLTools [8].

We docked each candidate against each of the 20 receptor structures (see MD section) using the *smina* software package [9], a fork of AutoDock Vina [10]. We defined the search space as a box of size 30 $Å^3$, centered on the catalytic serine (SER441) of the protease. We used the Vinardo [11] scoring function with the exhaustiveness of 10. We keep other docking parameters at their default values. Based on analysis of our MD simulations (see above), we identify 5 residues (GLU299, LYS300, ASP435, GLN438 and TRP461) that are essential to the binding of ligands to the active site and the S1 pocket. The side chains of these residues are kept flexible during the docking run.

We computed the distance maps for each of the docking poses and retained only those poses in which the ligand formed at least 2 contacts (based on heavy atom distance, threshold = 3.5 Å) with the residues around the S1 pocket (residues 435-441 and 459-464).

Scoring:

We collected the raw docking scores (predicted "binding free energies") of the best binding poses for each receptor-ligand pair. For each receptor structure, we normalized the ligand scores by subtracting the mean and dividing by the standard deviation to correct the differences between the receptor structures. We computed the mean normalized score across the receptor structures for each of the ligands based on the retained poses.

Below we report the mean binding free energies for the 15 top-ranked compounds for illustration. Note that the score and the binding free energies are a meaningful way to rank within groups of compounds, but not between covalent and non-covalent binders as they only capture the non-covalent protein-drug interactions.

Compound grouping and ranking strategy:

We grouped the compounds into 3 distinct groups based on their properties and general availability: **DrugBank** (6490 compounds), **covalent** (6010 compounds), and **non-covalent** (31485 compounds). Compounds were assigned to the DrugBank group if they fell into multiple of these categories. We employ this distinction as DrugBank molecules may be preferable for treatment even if they have inferior scores and covalent/non-covalent binders cannot be ranked with respect to another by docking score. We compile 3 ranked lists and 1 final list of compounds in the following way:

Drugbank: We distinguish between approved (1522) and not-approved (4968) compounds. We pick the top-10 compounds in this list by pooling the top-5 approved and the top-5 not-approved compounds together and ordering them by their mean normalized score. The remainder of the Drugbank list is formed by pooling the remaining compounds and ordering them using the score described above.

Covalent and non-covalent: For both lists, we distinguish between serine protease inhibitors (832 covalent and 1183 non-covalent) and other compounds (5178 covalent and 30302 non-covalent). We considered a compound as covalent if it had at least one of the following groups:

- esters ([*]-C(=O)O-[*])
- aldehydes ([*]-C-[CH]=O)
- trifluoromethylketones ([*]-C(=O)C(F)(F)F)
- chloromethylketones ([*]-C(=O)CCl)
- sulfonyl fluorides (O=S(=O)(F)[*]).

We pick the top-10 compounds in both lists by pooling the top-3 serine protease inhibitors and the top-7 other compounds and ordering them by their mean normalized score. In order to reduce the redundancy of the rest of the lists, we first cluster the other compounds based on extended-connectivity fingerprints and take the best scorers from each cluster. This results in 1899 ordered other covalent compounds and 8855 ordered other non-covalent compounds which are then merged with the respective ordered serine protease inhibitors (829 and 1180 compounds, respectively) while ensuring a uniform distribution of the latter group throughout the list. The resulting lists contain 2738 and 10045 compounds, respectively.

Final list of 10,000 drug candidates: For the final list we took the top-100 compounds from the DrugBank group, all compounds from the covalent group and as many compounds from the noncovalent group as needed to reach a total of 10,000 compounds. In order to ensure that all three compound categories are found in the top ranks the list was ordered as follows: 1st of DrugBank, 1st of covalent, 1st of non-covalent, 2nd of DrugBank, etc.

Section 2: targets

We focused on one target only, TMPRSS2, but on different strategies to inhibit this protein.

TMPRSS2 is a host single-pass membrane protein with a serine protease domain exposed on several human cells, e.g. in the upper respiratory tract, and is essential to SARS-CoV-2 cell-entry. During infection, TMPRSS2 activates the viral spike protein, thereby enabling the virus to enter the cell where it replicates [15]. Besides SARS-CoV-2, TMPRSS2 is also required by other Coronaviruses and several strands of Influenza [15]. The physiological function of TMPRSS2 is, as yet, unclear, but a TMPRSS2 knock-out mouse model displays a wild-type phenotype, indicating that side-effects of inhibiting TMPRSS2 may be mild [18]. Therefore, TMPRSS2 is a very interesting Covid-19 target: it is pharmacologically accessible due to its exposed location, its inhibition appears to have few

side-effects, it will effectively stifle viral infection and since it is a host protein required by the virus, SARS-CoV-2 cannot easily avoid the effect of a TMPRSS2 inhibitor by its own genetic variability.

Section 3: libraries

ZINC: We used the ZINC "standard"-reaction database (http://zinc.docking.org/tranches/home/) with purchasability status "wait OK" which consists of ~997.4M compounds. To obtain the initial - and expanded library (see Section 1), we filtered the downloaded ZINC database according to structurally similar compounds with respect to the initial lead compounds and best predicted compounds from the QSAR model, which we then clustered to obtain reduced libraries while maintaining diversity.

DrugBank: We used the whole database excluding compounds with a molecular weight greater than 550 Dalton.

ChEMBL: We collected different assays from ChEMBL of Trypsin serine protease inhibitors and merged all compounds with unique ChEMBL-ID into one dataset. We then applied molecular weight filters and clustered the resulting subset of ~2.5K samples to obtain diverse compounds that have strong binding affinity, i.e. high pChEMBL value, against Trypsin.

Section 4: results

As described in the Methods section, we distinguish three groups of compounds that should be all considered in an *in vitro* assay as they cannot be meaningfully compared with a single score and represent different treatment strategies:

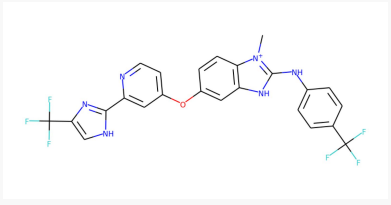
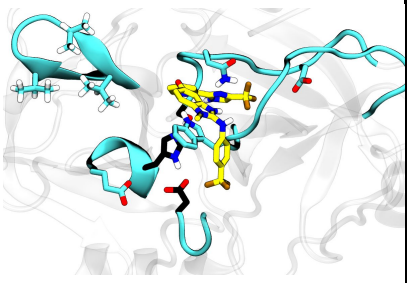
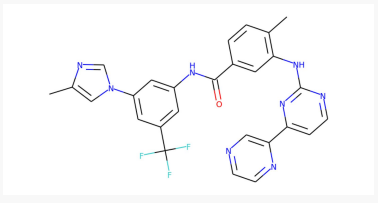
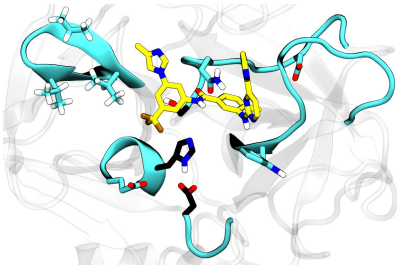
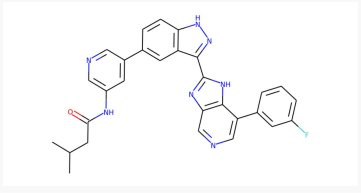
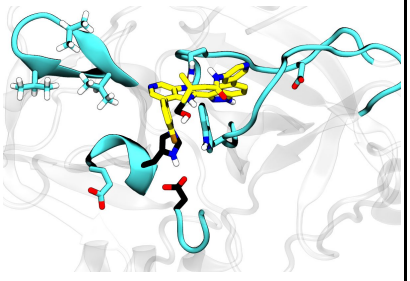
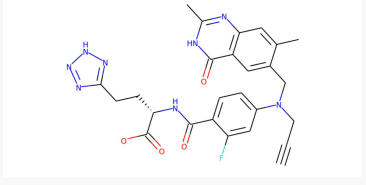
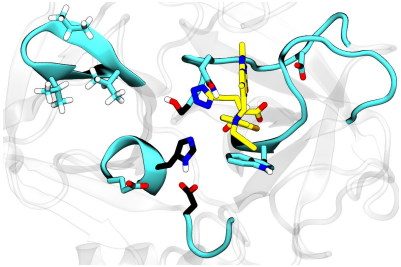
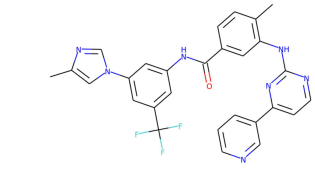
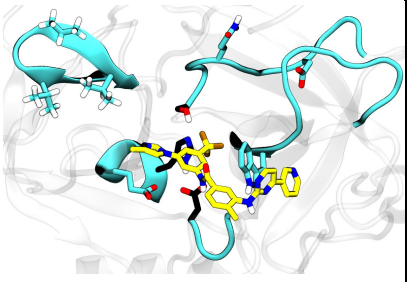
(1) Drugbank: As expected, molecules from drugbank do not score best in absolute terms, but these molecules are either already approved or in clinical trials, so they may be readily available for compassionate use. Even if not optimized for TMPRSS2, these compounds represent a possible short-time treatment venue.

(2) Covalent inhibitors: As TMPRSS2 is a protease, it can cleave certain substrates and one inhibition strategy is thus to design a drug where part of the catalyzed substrates stays attached to the protein and thus blocks its activity. Camostat and Nafamostat are examples for such covalent inhibitors [15-16]. As docking scores only predict the binding affinity of the non-covalent complex preceding the covalent inhibitory state, and most likely underestimate the potency of covalent inhibitors significantly, covalent inhibitor candidates should be considered separately. The stability of the noncovalent complex may still contribute to overall drug potency and also to the rate to enter the covalent state, and therefore it is a useful proxy to sort within the group of covalent inhibitors.

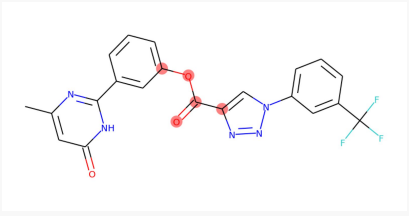
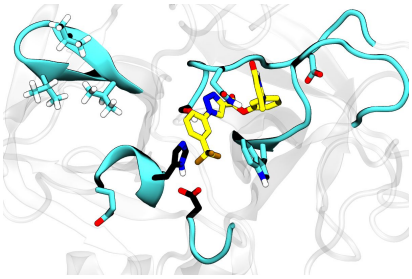
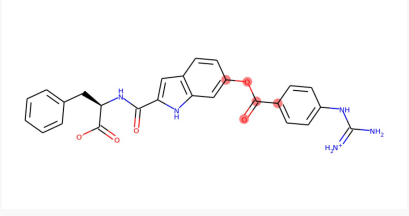
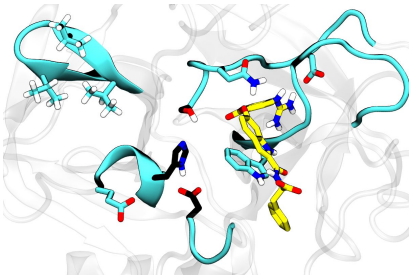
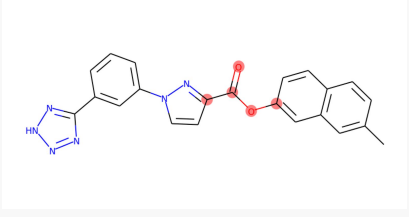
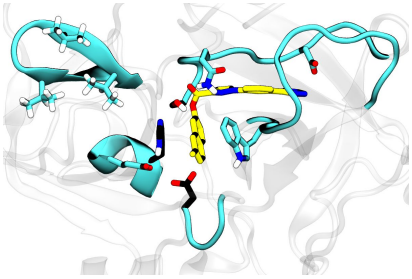
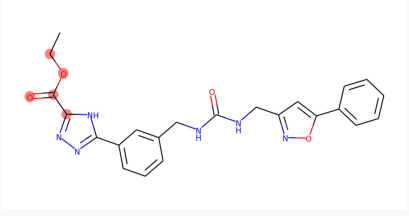
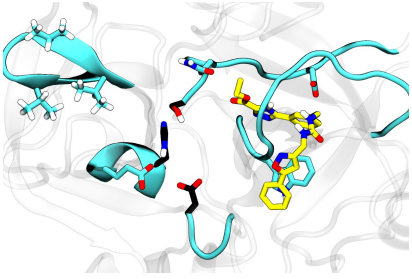
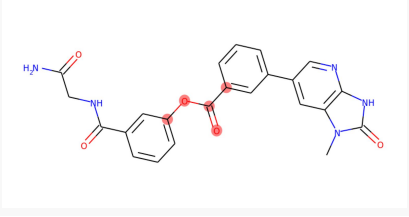
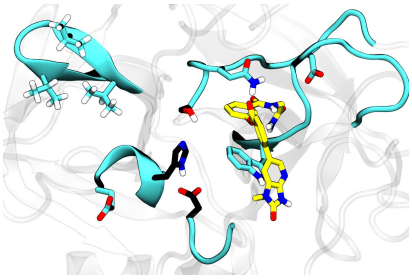
(3) Non-covalent inhibitors: These inhibit the protein by sticking to it via non-covalent interactions and the docking score is a meaningful way to rank them.

In our submitted *csv* file, the top 30 compounds contain 10 from each of these three groups. Below we report details for the top 5 of each group:

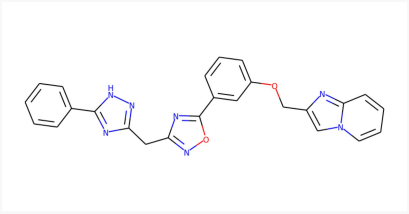
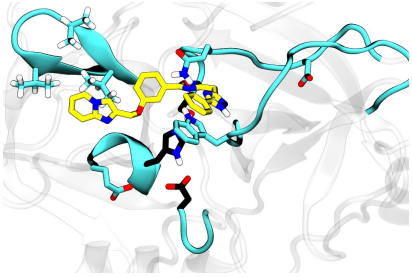
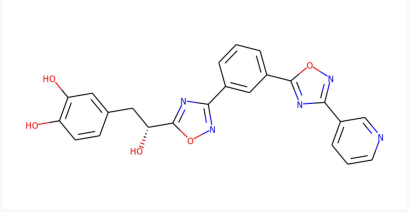
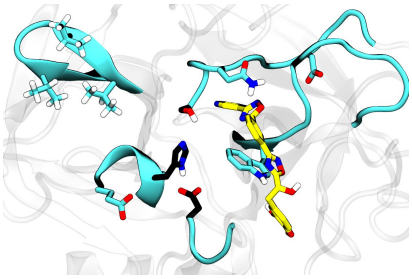
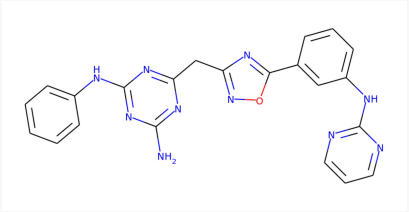
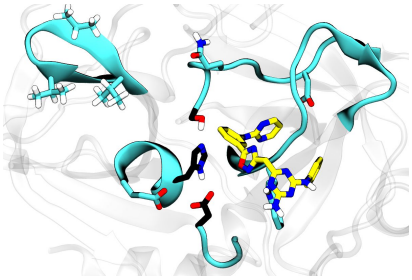
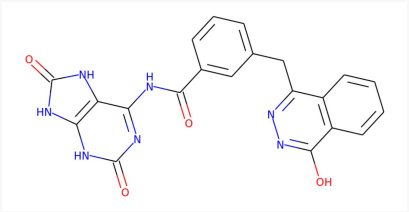
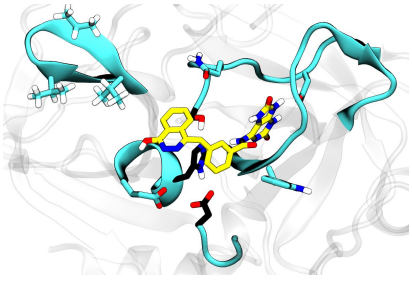
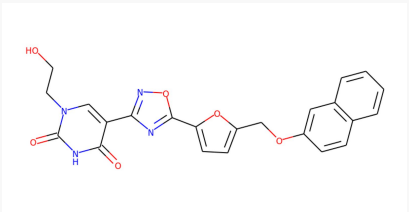
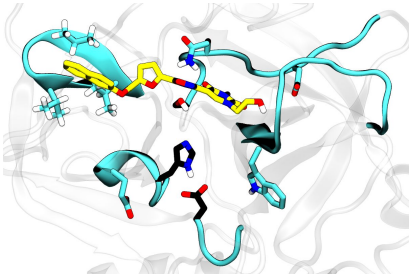
<u>Drugbank top 5</u> (see full list for more <u>approved</u> drugs):

ZINC ID, binding energy, structure	Best-scoring docked pose	Approved/developed for purpose, clinical stage, Drugbank ID
ZINC000018710085 -12.1 kcal / mol 		RAF-265, Raf kinase inhibitor, <u>investigational</u> (clinical trial phase 2 completed), melanoma treatment DB05984
ZINC000059749972 -12.1 kcal / mol 		Radotinib, tyrosine kinase inhibitor, <u>investigational</u> (clinical trial phase 3 completed / recruiting), leukemia treatment DB12323
ZINC000642771770 -12.1 kcal / mol 		Lorecivivint, <u>investigational</u> (clinical trial phase 3), Intervertebral disc degeneration DB14883
ZINC000001654736 -11.7 kcal / mol 		Plevitrexed, <u>investigational</u> (clinical trial phase 2 completed), cancer treatment DB06163
ZINC000006716957 -11.9 kcal / mol 		Nilotinib, tyrosine kinase inhibitor, <u>approved</u>, leukemia treatment DB04868

Covalent inhibitors top 5:

ZINC ID, binding energy, structure	Best-scoring docked pose	Comments
ZINC000744218107 -12.0 kcal / mol 		
ZINC000147666687 -12.0 kcal / mol 		derived from trypsin-1 inhibitor lead
ZINC001592207500 -11.9 kcal / mol 		
ZINC000492744745 -11.9 kcal / mol 		
ZINC000777318295 -11.8 kcal / mol 		

Non-covalent inhibitors top 5:

ZINC ID, binding energy, structure	Best-scoring docked pose	Comments
ZINC000545085033 -12.6 kcal / mol 		
ZINC001211665215 -12.6 kcal / mol 		
ZINC000680669974 -12.4 kcal / mol 		
ZINC001193488436 -12.3 kcal / mol 		
ZINC000904863425 -12.3 kcal / mol 		

Bibliography

[1] Rensi, S. et al. Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into Human Cells; *preprint*; **2020**. https://doi.org/10.26434/chemrxiv.12009582.
[2] Huggins, D. Structural Analysis of Experimental Drugs Binding to the COVID-19 Target TMPRSS2; *preprint*; **2020**. http://doi.org/10.26434/chemrxiv.12315449.
[3] Bestle, D. et al. TMPRSS2 And Furin Are Both Essential For Proteolytic Activation And Spread Of SARS-Cov-2 In Human Airway Epithelial Cells And Provide Promising Drug Targets; *preprint*; **2020**. https://doi.org/10.1101/2020.04.15.042085.
[4] Nimishakavi, S. et al. Divergent Inhibitor Susceptibility Among Airway Lumen-Accessible Tryptic Proteases. *PLoS ONE* **2015**, 10 (10), e0141169. https://doi.org/10.1371/journal.pone.0141169.
[5] Katz, B. et al. Design Of Potent Selective Zinc-Mediated Serine Protease Inhibitors. *Nature* **1998**, 391 (6667), 608-612.. https://doi.org/10.1038/35422
[6] Winter, R., et al. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations, *Chem. Sci.* **2019**:10.6 1692-1701. https://doi.org/10.1039/C8SC04175J
[7] *Schrödinger Release 2020-2: LigPrep*; Schrödinger, LLC: New York, NY, 2020.
[8] Morris, G. et al. Autodock4 And Autodocktools4: Automated Docking With Selective Receptor Flexibility. *J. Comp. Chem.* **2009**, *30* (16), 2785-2791. https://doi.org/10.1002/jcc.21256.
[9] Koes, D. et al. Lessons Learned In Empirical Scoring With Smina From The CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **2013**, 53 (8), 1893-1904. https://doi.org/10.1021/ci300604z.
[10] Trott, O. et al. Autodock Vina: Improving The Speed And Accuracy Of Docking With A New Scoring Function, Efficient Optimization, And Multithreading. *J. Comp. Chem.* **2009**, 31, 455-461. https://doi.org/10.1002/jcc.21334.
[11] Quiroga, R.; Villarreal, M. Vinardo: A Scoring Function Based On Autodock Vina Improves Scoring, Docking, And Virtual Screening. *PLoS ONE* **2016**, 11 (5), e0155183. https://doi.org/10.1371/journal.pone.0155183.
[12] Eastman, P. et al. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comp. Biol.* **2017**, *13* (7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659.
[13] *Raval, A. et al. Refinement of Protein Structure Homology Models via Long, All-Atom Molecular Dynamics Simulations. Proteins* ***2012****, 80, 2071–2079. https://doi.org/10.1002/prot.24098.*
[14] Best, R. B. et al. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. https://doi.org/10.1021/ct300400x.
[15] Hoffmann, M. et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181* (2), 271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052.
[16] Hoffmann, M. et al. Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19. *Antimicrob. Agents Chem.* **2020**, 64 (6). http://dx.doi.org/10.1128/AAC.00754-20
[18] Kim, T. S. et al. Phenotypic Analysis of Mice Lacking the Tmprss2-Encoded Protease. *Mol. Cell. Biol.* **2006**, *26* (3), 965–975. https://doi.org/10.1128/mcb.26.3.965-975.2006.