Team name	LuxScreen
Team member(s) (firstname lastname; …)	Enrico Glaab
Affiliation	Luxembourg Centre for Systems Biomedicine, University of Luxembourg
Contact email	enrico.glaab@uni.lu
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	3CLPro/Nsp5, helicase/Nsp13, 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.

Methods:

Protein structure pre-processing: The initial structures for 3CLPro (PDB: 5R8T) and the viral helicase Nsp13 (PDB: 6JYT) were obtained from the Protein Data Bank (rcsb.org). For the human protein TMPRSS2 no public crystal structure was available, but high-quality template structures could be obtained for homology modeling. For this purpose, the software PRIMO (primo.rubi.ru.ac.za) was used, creating the homology model by using the crystal structure for human plasma kallikrein as template (PDB: 5TJX; this template shows a 42.6% sequence identity to TMPRSS2, and provided the best DOPE Z-Score in the PRIMO software in comparison to other candidate template structures with high sequence similarity).
The receptor structures were pre-processed using the Schroedinger Maestro software by adding hydrogens, generating protonation states, and optimizing hydrogen positions. The quality of the original and final structures was assessed using Verify3D, WHATCHECK and PROCHECK (https://servicesn.mbi.ucla.edu/Verify3D/). For proteins with multiple chains, the chain with the highest Verify3D score was chosen for further analysis.

Ligand pre-processing & docking: All ligands from the Merck AMS and SWEETLEAD libraries were preprocessed using the AutoDock ligand preparation script, and docked using AutoDock-GPU. For the ZINC database, in order to focus on compounds that are commercially available and have drug-like chemical and ADMET properties, the ZINC database was filtered to download all compounds with the properties “drug-like”, “purchasable” (minimum purchasability = “Wait OK”) and reactivity = “clean” (1,276,766,435 substances before filtering, and 898,838,573 substances after the first-step filtering). These compounds were downloaded in SMILES-format, using the “ZINC-downloader-2D-smi.wget” script derived from the “Tranches” web-page on ZINC (https://zinc.docking.org/tranches/home/).
The collection of ZINC compounds was further filtered using a feature trees representation and similarity assessment approach (software: BioSolveIT Ftrees v6.2) to score the topological and physicochemical similarity to small-molecule binders reported in the literature and top-scored candidate inhibitors from the AutoDock-GPU screening on the Merck AMS and SWEETLEAD libraries for each of the chosen target proteins. Specifically, the literature-derived query compounds for 3CLPro include: ebselen, amentoflavone, hesperetine, pectolinarin, baicalein and dieckol; for the helicase they include: SSYA10-001 and FSPA ((E)-3-(furan-2-yl)-N-(4-sulfamoylphenyl) acrylamide); for TMRPSS2 they include: nafamostat, camostat and bromhexine hydrochloride. These literature-derived query compounds for the feature tree search were completed by the top-ranked compounds from the Merck AMS and SWEETLEAD libraries obtained from the AutoDock-

GPU screening as further query compounds, such that 10 query compounds in total were obtained for each target protein. All ZINC compounds exceeding a minimum similarity threshold of 0.8 in the FTrees screen to the query compounds were retained for further docking analyses.
For the final docking runs, the prefiltered compounds for each of the target proteins were docked by three different approaches: AutoDock-GPU, OpenEye FRED/HYBRID, and BioSolveIT FlexX+HYDE (initially, Schroedinger Glide was included as a fourth docking approach, but due to the availability of only a single-user license and long runtimes, this additional screening was not completed and not taken into consideration for the final ranking). Since FlexX+HYDE was the most time-consuming docking approach among the three methods considered, to save time, it was first run complete on the SWEETLEAD library, and for the larger Merck AMS and ZINC-derived compounds, only run on compounds with higher than average scores from the AutoDock-GPU and OpenEye FRED/HYBRID screens in the order of their rankings (since compounds with lower scores in the first two docking approaches would in any case not reach the top 10k of the final combined ranking).
The final ranking was determined by the sum-of-ranks across the three docking tools, and the binding energy estimates were obtained from the best predicted binding affinity by the BioSolveIT HYDE software from among the top 30 docking poses.

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.

Target 1: 3CLPro/Nsp5 (PDB: 5R8T)
Reason of choice: 3CLPro was chosen due to the public availability of high-quality crystal structures for molecular docking, its key role as viral protease in the SARS-CoV-2 replication cycle, and the information first reported small-molecule inhibitors in the literature (covering the compounds ebselen, amentoflavone, hesperetine, pectolinarin, baicalein and dieckol), which served as query structures for ligand-based screening using the Feature Trees approach (see above). A further reason for choosing 3CLPro was that a druggability analysis using the software DoGSiteScorer (https://proteins.plus/) confirmed that the protein contains a druggable binding pocket (DrugScore: 0.81).

Source: The initial structure was obtained from rcsb.org (PDB: 5R8T) but underwent further pre-processing steps (see below).

Quality: The PDB structure 5R8T has a resolution of 1.27 Å and an R-free value of 0.208, and provides complete coverage of the amino acid sequence (see rcsb.org/structure/5r8t). Only one structure with a slightly superior resolution and R-free value was identified (PDB: 6YB7); however, after pre-processing both these structures using the software Maestro Schroedinger (see below), the evaluation results with PROCHECK were superior for the preprocessed 5R8T structure in comparison to the preprocessed 6YB7 structure and to both unprocessed structures. Therefore, the preprocessed 5R8T structure was used for all following analyses.

Pre-processing: The receptor structure was pre-processed using the Schroedinger Maestro software by adding hydrogens, generating protonation states, and optimizing hydrogen positions.

Target 2: helicase / Nsp13 (PDB: 6JYT)
Reason of choice: The viral helicase Nsp13 was selected because of its central role in viral replication, due to the availability of a suitable crystal structure for molecular docking and first

known inhibitors (SSYA10-001 and FSPA) for ligand-based screening. Moreover, similar to 3CLpro, a druggability analysis using the software DoGSiteScorer (https://proteins.plus/) confirmed that the protein contains a druggable binding pocket (DrugScore again: 0.81).

Source: The initial structure was obtained from rcsb.org (PDB: 6JYT) but underwent further pre-processing steps (see below).

***Quality**:* The PDB structure 6JYT has a resolution of 2.8 Å and an R-free value of 0.292, and provides almost complete coverage of the amino acid sequence (see rcsb.org/structure/6jyt). No structure with a superior resolution or R-free value was identified. Subsequent quality checks of the preprocessed structure (see Pre-processing description below) using Verify3D, WHATCHECK and PROCHECK confirmed the suitability of the preprocessed structure for further analyses.

***Pre-processing**:* The receptor structure was pre-processed using the Schroedinger Maestro software by adding hydrogens, generating protonation states, and optimizing hydrogen positions.

Target 3: TMPRSS2 (homology model template: PDB: 5TJX)

Reason of choice: The human protein TMPRSS2 was mainly chosen, because it has been reported to prime the SARS-CoV-2 spike protein for viral entry into the host cell via the receptor ACE2 (Hoffmann et al., Cell, 2020). Moreover, the same publication reported that TMPRSS2 blockage via the protease inhibitor camostat inhibited SARS-CoV-2 cell entry. Apart from providing a promising target for cell entry inhibition, as a human protein TMPRSS2 does not involve the risk of drug resistance development, as opposed to viral targets. Although no public crystal structure for TMPRSS2 was available at the time of analysis, high-quality template structures could be obtained for homology modeling. Specifically, the human plasma kallikrein was chosen as template (PDB: 5TJX), because this template has a 42.6% sequence identity to TMPRSS2 and provided the best DOPE Z-Score in in comparison to other candidate template structures with high sequence similarity (moreover, the template PDB structure has high quality, see Quality section below). A druggability analysis for the homology model using the software DoGSiteScorer (https://proteins.plus/) also confirmed that the protein contains a druggable binding pocket (DrugScore: 0.8).

Source: The software PRIMO (primo.rubi.ru.ac.za) was used to create the homology model by using the crystal structure for human plasma kallikrein as template (PDB: 5TJX; this template shows a 42.6% sequence identity to TMPRSS2, and provided the best DOPE Z-Score in the PRIMO software in comparison to other candidate template structures with high sequence similarity).

***Quality**:* The PDB structure of the used template 5TJX has a resolution of 1.41 Å and an R-free value of 0.184 (see https://www.rcsb.org/structure/5TJX). When assessing the quality of the homology model obtained using the software PRIMO, as well as a further processed version of this model using the software Maestro Schroedinger (see below), the evaluation results confirmed the utility of the model for further analysis, and the quality did not improve after the processing with the Maestro software. Therefore, the homology model derived from PRIMO was directly used for the subsequent analyses.

***Pre-processing**:* The homology model structure was pre-processed using the Schroedinger Maestro software by adding hydrogens, generating protonation states, and optimizing hydrogen positions.

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.

Library 1: SWEETLEAD

The SWEETLEAD library was obtained in sdf-format after registration from https://simtk.org/projects/sweetlead. The library was converted into mol2-format using the software PyMOL 2.3.5. The complete library was used for screening, i.e. without removing any compounds. Ligand pre-processing was performed using tools specific to the particular molecular docking approaches used. Specifically, for docking with the OpenEye software, ligand pre-processing and conformer generation was performed using the oeomega classic tool; for the docking with AutoDock-GPU the compounds were preprocessed using the script "prepare_ligand4.py" from the MGLTools package (v1.5.6); for docking with BioSolveIT FlexX+HYDE the ligand pre-processing was performed using the LeadIT command line tool.

Library 2: Merck AMS

The Merck AMS library was obtained in smi-format via a request to the JEDI challenge organizer. The conversion of this library into mol2-format and generation of 3D coordinates was obtained using the software Open Babel v2.3.1. The complete library was used for screening, i.e. without removing any compounds. The ligand pre-processing was again performed using tools specific to the particular molecular docking approaches used, following the same procedure as for the SWEETLEAD library (see above).

Library 2: ZINC

Compounds from the ZINC database were pre-filtered in order to focus on those that are commercially available and have drug-like chemical and ADMET properties. Specifically, the ZINC database was filtered to download all compounds with the properties "drug-like", "purchasable" (minimum purchasability = "Wait OK") and reactivity = "clean" (1,276,766,435 substances before filtering, and 898,838,573 substances after the first-step filtering). These compounds were downloaded in SMILES-format, using the "ZINC-downloader-2D-smi.wget" script derived from the "Tranches" web-page on ZINC (https://zinc.docking.org/tranches/home/).
After further filtering the compounds using feature tree searches specific to each target protein (see Methods section above), the corresponding mol2-files for the selected compounds were downloaded from the ZINC database. These compounds were used for screening without further filtering. The ligand pre-processing was again performed using tools specific to the particular molecular docking approaches used, following the same procedure as for the SWEETLEAD library (see above).

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:

Target 1: 3CLPro/Nsp5 (PDB: 5R8T)

Overall findings/trends: For the 3CLPro target the top-scoring compounds were all obtained from the ZINC database, i.e. derived from the feature tree similarity search within this database. This matches with the fact that more already known binding compounds could be derived from the

literature for this target, resulting in a better performance of the ligand-based screening component within the overall screening pipeline. The smaller Merck AMS and SWEETLEAD libraries, from which all compounds were docked, did not contain as many structurally and physicochemically similar compounds to the known binders as the similarity-based pre-filtered compounds from the larger ZINC database, which may explain why the filtered ZINC compounds ranked higher.

Top 5 compounds:

Rank	Canonical.Smiles	Estimated binding affinity (kJ/mol)
1	O=C(Nc1cccnc1)C(=O)N1C[C@H](O)[C@H](O)C1	-39.871
2	C[C@@H]1CNC[C@H]1Oc1cc(O)c(O)c2c1C(=O)c1ccccc1C2=O	-31.161
3	COc1ccc(-c2c(C)oc3c(CN4CCN(CCO)CC4)c(O)ccc3c2=O)cc1	-49.084
4	O=c1c2cccc(O)c2oc2cc(O)cc(OC[C@H]3CC[C@@]4(CCCCO4)O3)c12	-30.341
5	NC[C@H](N)C(=O)NCCCNC(=O)c1c2ccccc2nc2ccccc12	-40.976

The top 5 compounds have estimated binding affinities in the range between -30 to -40 kJ/mol. Due to the prior filtering of the ZINC compounds, these substances all fulfil the predetermined criteria of having drug-like structures, being (minimum purchasability = "Wait OK") and having "clean" reactivity (= PAINS-ok).
In terms of the structural and chemical properties of the top 5 compounds, a common characteristic appears to be a three-ring structures, which can be seen in compounds 2, 4 and 5. In general, all top 5 compounds have at least two ring sub-structures, and multiple hydrogen bond donors and acceptors. The molecular weights are in a range between 250 to 500 g/mol. No clear trend could be seen for the n-octanol/water partition coefficient (logP), which varied in a range between approx. -1.4 to 3.8.

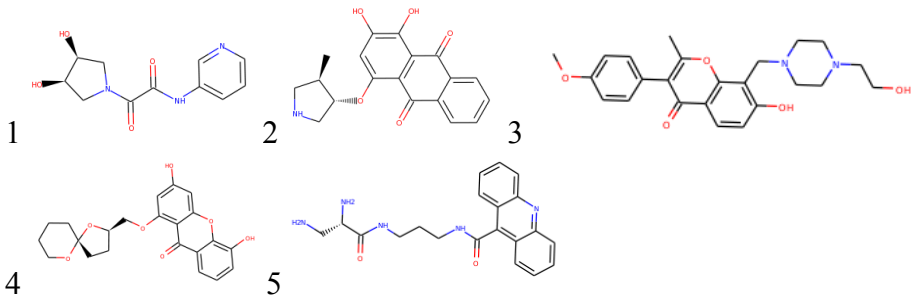

Fig. 1: Structures of the top 5 compounds selected for the 3CLPro target.

Target 2: helicase / Nsp13 (PDB: 6JYT)

Overall findings/trends: For the helicase, the top-scoring compounds came mostly from the Merck AMS database and the SWEETLEAD library. The compounds from the ZINC database, which had been pre-filtered using ligand-based screening, tended to provide inferior docking results, which may be explained by the fact that fewer known binding compounds could be derived from the literature for the ligand-based feature tree search as compared to the 3CLpro target, resulting in a comparatively weaker performance of the ligand-based pre-screening for the ZINC database. Overall, potentially due to the lower quality of the 6JYT crystal structure as compared to the structure available for the 3CLpro target, a lower agreement between the rankings was observed for the different docking tools, likely resulting in sum-of-rank scores were lower reliability, which is also reflected by the lower estimated binding affinities.

Top 5 compounds:

Rank	Canonical.Smiles	Estimated binding affinity (kJ/mol)
1	C/C(=C\CNC(=O)OCC1c2ccccc2-c2ccccc12)/C(=O)N1CCCC(C)(C(=O)O)C1	-30.113
2	O=C(NC1CCC(C(=O)N2CCC[C@@H](C(=O)O)C2)CC1)OCC1c2ccccc2-c2ccccc12	-22.196
3	O=C(N[C@H]1CC[C@@H](C(=O)N2CCC[C@@H](C(=O)O)C2)CC1)OCC1c2ccccc2-c2ccccc12	-23.862
4	C[C@@H](CC(=O)NC(C)(C(=O)O)c1ccccc1)NC(=O)OCC1c2ccccc2-c2ccccc12	-20.764
5	O=C(N[C@@H]1C=CC[C@@H](C(=O)N2CCC(O)(C(=O)O)C2)C1)OCC1c2ccccc2-c2ccccc12	-28.489

The top 5 compounds have estimated binding affinities in the range between -20 to -30 kJ/mol, which may result from the limited agreement between the scores across the different docking softwares, and fact that the compounds were ordered by the sum-of-ranks, and not by the estimated binding affinity using BioSolveIT HYDE (resorting the compounds by estimated affinity provides top 5 affinities in the range between -35 to -70 kJ/mol; however, the sum-of-ranks across the scores from the different docking tools was considered as the more reliable information source for ordering the compounds, as it exploits information from three different scoring approaches).
Since the compounds are from the Merck AMS library, they fulfil common properties of drug-likeness, such as having molecular weights below 500 g/mol.
In terms of the structural and chemical properties of the top 5 compounds, they show high structural similarity, with a shared 3-ring sub-structure (see Fig. 2, right part of the displayed structures) and a shared carboxyl group on the other side of the structure (see Fig. 2, left part of the displayed structures).
In general, all top 5 compounds have at least four ring sub-structures, and display a similar arrangement of hydrogen bond donors and acceptors. The molecular weights are in a range between 400 to 500 g/mol. The n-octanol/water partition coefficient (logP) lies in a range between approx. 3 to 4.5.

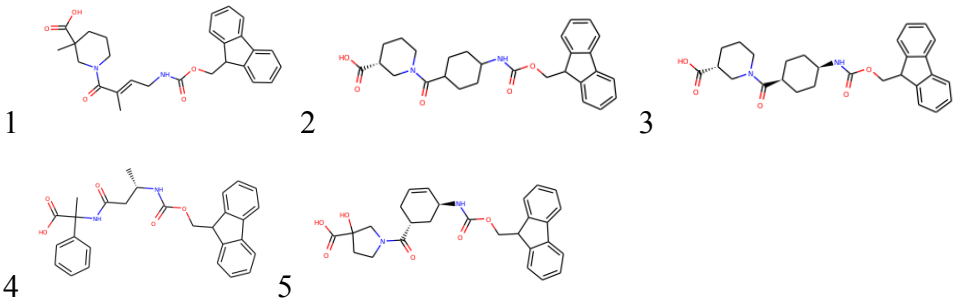

Fig. 2: Structures of the top 5 compounds selected for the helicase target.

Target 3: TMPRSS2 (homology model template: PDB: 5TJX)
Overall findings/trends: For the target TMPRSS2, similar as for the helicase, the top-scoring compounds mainly came from the Merck AMS database and the SWEETLEAD library. The compounds from the ZINC database, which had been pre-filtered using ligand-based screening, mostly provided inferior docking results, since in this case again fewer known binding compounds could be derived from the literature for the ligand-based feature tree search as compared to the 3CLpro target. However, in contrast to the binding affinity estimation results for the helicase, higher affinities were predicted, which may be explained by the higher quality of the template crystal

structure used for homology modeling (however, since only a homology model was used in this case, the affinity estimates may be less reliable than for the other two chosen targets).

Top 5 compounds:

Rank	Canonical.Smiles	Estimated binding affinity (kJ/mol)
1	N#Cc1c(Cl)cccc1Oc1ccc(S(=O)(=O)N2CCCN(c3nc4ccc(Cl)cc4s3)CC2)cc1	-45.829
2	O=C(CCCn1nc(c2ccccc2)c2cc(Cl)ccc12)NC1CCN(Cc2ccccc2)CC1	-45.98
3	Cc1cc(C)c(/C=C/c2onc(C)c2S(=O)(=O)N2CCCC(C(=O)N3CCc4ccccc34)C2)c(C)c1	-45.676
4	C=C1Nc2cc(C(=O)NCCc3ccc(Cl)cc3)ccc2Sc2ccc(C)cc12	-44.326
5	COc1ccccc1CNC(=O)c1ccc(CN2C(=O)CCn3nc(c4ccccc4)cc23)cc1	-40.751

The top 5 compounds have estimated binding affinities in the range between -40 to -50 kJ/mol. Since the compounds are from the Merck AMS library, they fulfil common properties of drug-likeness, such as having molecular weights below 500 g/mol.
In terms of the structural and chemical properties of the top 5 compounds, a common characteristic was a bent, V-like shape with aromatic rings at the outer ends and in the center (see Fig. 3). Three of the compounds contained at least one chloride atom (see structures 1, 2, and 4). In general, all top 5 compounds have at least three ring sub-structures, and various hydrogen bond donors and acceptors. The molecular weights are in a range between 400 to 500 g/mol. The n-octanol/water partition coefficient (logP) tended to be relatively high, in a range between approx. 4 to 6.

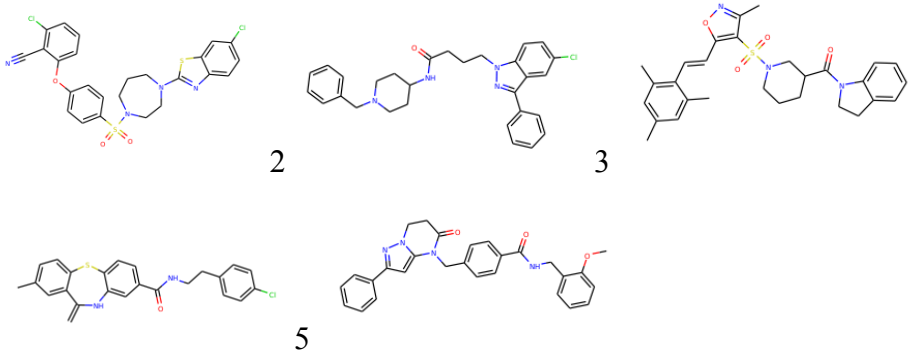

Fig. 3: Structures of the top 5 compounds selected for the TMPRSS2 target.

Other comments:
Since the final compound rankings were determined by the sum-of-ranks across the different rankings obtained from the three docking tools, and not by the estimated binding affinity using the BioSolveIT HYDE software, the compounds are not sorted by this binding affinity estimate in the 5th column (the sum-of-ranks, which determined the ordering is not included in the table, as it does not correspond to a binding affinity prediction, but was considered as the most reliable information to determine the final ranking). However, if an ordering of the compounds in terms of estimated binding affinity is preferred, a reordering of the rows by increasing values of the predicted binding affinities in column 5 is sufficient for this purpose (if needed, I can provide corresponding alternative versions of the submitted tables upon request).