molecular informatics

Supporting Information

A community effort in SARS-CoV-2 drug discovery

Johannes Schimunek | Philipp Seidl | Katarina Elez | Tim Hempel | Tuan Le | Frank Noé | Simon Olsson | Lluís Raich | Robin Winter | Hatice Gokcan | Filipp Gusev | Evgeny M. Gutkin | Olexandr Isayev | Maria G. Kurnikova | Chamali H. Narangoda | Roman Zubatyuk | Ivan P. Bosko | Konstantin V. Furs | Anna D. Karpenko | Yury V. Kornoushenko | Mikita Shuldau | Artsemi Yushkevich | Mohammed B. Benabderrahmane | Patrick Bousquet-Melou | Ronan Bureau | Beatrice Charton | Bertrand C. Cirou | Gérard Gil | William J. Allen | Suman Sirimulla | Stanley Watowich | Nick Antonopoulos | Nikolaos Epitropakis | Agamemnon Krasoulis | Vassilis Pitsikalis | Stavros Theodorakis | Igor Kozlovskii | Anton Maliutin | Alexander Medvedev | Petr Popov | Mark Zaretckii | Hamid Eghbal-Zadeh | Christina Halmich | Sepp Hochreiter | Andreas Mayr | Peter Ruch | Michael Widrich | Francois Berenger | Ashutosh Kumar | Yoshihiro Yamanishi | Kam Y. J. Zhang | Emmanuel Bengio | Yoshua Bengio | Moksh J. Jain | Maksym Korablyov | Cheng-Hao Liu | Gilles Marcou | Enrico Glaab | Kelly Barnsley | Suhasini M. Iyengar | Mary Jo Ondrechen | V. Joachim Haupt | Florian Kaiser | Michael Schroeder | Luisa Pugliese | Simone Albani | Christina Athanasiou | Andrea Beccari | Paolo Carloni | Giulia D'Arrigo | Eleonora Gianquinto | Jonas Goßen | Anton Hanke | Benjamin P. Joseph | Daria B. Kokh | Sandra Kovachka | Candida Manelfi | Goutam Mukherjee | Abraham Muñiz-Chicharro | Francesco Musiani | Ariane Nunes-Alves | Giulia Paiardi | Giulia Rossetti | S. Kashif Sadiq | Francesca Spyrakis | Carmine Talarico | Alexandros Tsengenes | Rebecca C. Wade | Conner Copeland | Jeremiah Gaiser | Daniel R. Olson | Amitava Roy | Vishwesh Venkatraman | Travis J. Wheeler | Haribabu Arthanari | Klara Blaschitz | Marco Cespugli | Vedat Durmaz | Konstantin Fackeldey | Patrick D. Fischer | Christoph Gorgulla | Christian Gruber | Karl Gruber | Michael Hetmann | Jamie E. Kinney | Krishna M. Padmanabha Das | Shreya Pandita | Amit Singh |

Supporting Information:
A community effort in SARS-CoV-2 drug discovery

Johannes Schimunek, Philipp Seidl, Katarina Elez, Tim Hempel, Tuan Le, Frank Noe, Simon Olsson, Lluís Raich, Robin Winter, Hatice Gokcan, Filipp Gusev, Evgeny M. Gutkin, Olexandr Isayev, Maria G. Kurnikova, Chamali H. Narangoda, Roman Zubatyuk, Ivan P. Bosko, Konstantin V. Furs, Anna D. Karpenko, Yury V. Kornoushenko, Mikita Shuldau, Artsemi Yushkevich, Mohammed B. Benabderrahmane, Patrick Bousquet-Melou, Ronan Bureau, Beatrice Charton, Bertrand C. Cirou, Gérard Gil, William J. Allen, Suman Sirimulla, Stanley Watowich, Nick A. Antonopoulos, Nikolaos E. Epitropakis, Agamemnon K. Krasoulis, Vassilis P. Pitsikalis, Stavros T. Theodorakis, Igor Kozlovskii, Anton Maliutin, Alexander Medvedev, Petr Popov, Mark Zaretckii, Hamid Eghbal-zadeh, Christina Halmich, Sepp Hochreiter, Andreas Mayr, Peter Ruch, Michael Widrich, Francois Berenger, Ashutosh Kumar, Yoshihiro Yamanishi, Kam Y.J. Zhang, Emmanuel Bengio, Yoshua Bengio, Moksh J. Jain, Maksym Korablyov, Cheng-Hao Liu, Gilles Marcou, Enrico Glaab, Kelly Barnsley, Suhasini M. Iyengar, Mary Jo Ondrechen, V. Joachim Haupt, Florian Kaiser, Michael Schroeder, Luisa Pugliese, Simone Albani, Christina Athanasiou, Andrea Beccari, Paolo Carloni, Giulia D'Arrigo, Eleonora Gianquinto, Jonas Goßen, Anton Hanke, Benjamin P. Joseph, Daria B. Kokh, Sandra Kovachka, Candida Manelfi, Goutam Mukherjee, Abraham Muñiz-Chicharro, Francesco Musiani, Ariane Nunes-Alves, Giulia Paiardi, Giulia Rossetti, S. Kashif Sadiq, Francesca Spyrakis, Carmine Talarico, Alexandros Tsengenes, Rebecca C. Wade, Conner Copeland, Jeremiah Gaiser, Daniel R. Olson, Amitava Roy, Vishwesh Venkatraman, Travis J. Wheeler, Haribabu Arthanari, Klara Blaschitz, Marco Cespugli, Vedat Durmaz, Konstantin Fackeldey, Patrick D. Fischer, Christoph Gorgulla, Christian Gruber, Karl Gruber, Michael Hetmann, Jamie E. Kinney, Krishna M. Padmanabha Das, Shreya Pandita, Amit Singh, Georg Steinkellner, Guilhem Tesseyre, Gerhard Wagner, Zi-Fu Wang, Ryan J. Yust, Dmitry S. Druzhilovskiy, Dmitry A. Filimonov, Pavel V. Pogodin, Vladimir Poroikov, Anastassia V. Rudik, Leonid A. Stolbov, Alexander V. Veselovsky, Maria De Rosa, Giada De Simone, Maria R. Gulotta, Jessica Lombino, Nedra Mekni, Ugo Perricone, Arturo Casini, Amanda Embree, D. Benjamin Gordon, David Lei, Katelin Pratt, Christopher A. Voigt, Kuang-Yu Chen, Yves Jacob, Tim Krischuns, Pierre Lafaye, Agnès Zettor, M. Luis Rodríguez, Kris M. White, Daren Fearon, Frank Von Delft, Martin A. Walsh, Dragos Horvath, Charles L. Brooks III, Babak Falsafi, Bryan Ford, Adolfo García-Sastre, Sang Yup Lee, Nadia Naffakh, Alexandre Varnek*, Günter Klambauer*, Thomas M. Hermans*

* corresponding authors; contact: hermans@unistra.fr
See all affiliation details and author contributions:
https://github.com/hermanslab/COVID-19/tree/main/AuthorContributions_Acknowledgements

Sections

Section 1: methods of each team

See the originally submitted team reports and compound lists on the following repository: https://github.com/hermanslab/COVID-19. Below are the ~1 page summaries of the method of each participating team.

ai4science (1)

Team members. Katarina Elez, Tim Hempel, Robin Winter, Tuan Le, Lluís Raich, Simon Olsson and Frank Noé.

Target structures. We conduct all-atom molecular dynamics (MD) simulations of apo and holo models of the full activated form of the protease domain of TMPRSS2 (residues 256 to 490) taken from[1] to account for possible artifacts of the homology model[2] and target flexibility. We initialize our apo simulations from the homology model based on Enteropeptidase-1 (PDB: 3W94) and our holo simulations from docked poses with the confirmed inhibitors Camostat and Nafamostat[3,4]. We run a standard MD protocol mirroring physiological conditions with OpenMM 7.4.0[5] using the CHARMM 36 force field[6]. Using a dataset of 50 microseconds and hidden Markov models[7], we selected a total of 20 representative protein conformations - 10 from apo simulations, 4 from Nafamostat7-bound and 6 from Camostat-bound simulations.

Drug libraries. We considered the following libraries: (1) ZINC ("standard"-reaction database with purchasability status "wait OK", ~997.4M compounds); (2) *DrugBank* (Mw <= 550 Da) and (3) *ChEMBL* (different assays of serine protease inhibitors clustered to obtain diverse compounds with strong binding affinity to Trypsin).

Initial drug library. We first gathered a total of ~24,000 drug candidates as follows: (1) ~9,000 diverse 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[1,8–11]; (2) ~6,000 diverse ZINC compounds containing a guanidinium group; (3) ~6,500 molecules from DrugBank and (4) ~2,500 ChEMBL molecules that strongly inhibit serine proteases. All ~24,000 initial compounds were docked and scored to the 20 target structures before entering the active learning cycle.

Docking. We retrieved 3D structures (as mol2 files) from the ZINC database (reference molecule) when they were available. Otherwise, we generated them from SMILES strings with the optimal ionization states at pH 7.05 using LigPrep[12]. To prepare the receptor and the ligand structures for docking we used MGLTools[13].
We docked each ligand against each of the 20 receptor structures using the *smina* software package[14,15]. We defined the search space as a box of size 30 Å^3, centered on the catalytic serine (SER441). We used the Vinardo[16] scoring function with the exhaustiveness of 10. Side chains of GLU299, LYS300, ASP435, GLN438 and TRP461 were kept flexible throughout the docking run.

Scoring. We collected raw docking scores of the best binding poses for each receptor-ligand pair. For each receptor structure, we normalized the ligand scores and computed the distance maps for each pose to retain only those in which the ligand formed at least 2 contacts (based on heavy atom distance, threshold = 3.5 Å) with residues 435-441 and 459-464. Based on the retained poses, we computed the mean normalized score across the receptor structures for each of the ligands.

Active learning cycle for drug library expansion. We performed three cycles of active learning as follows: (1) define a set of new molecules from ZINC that are structurally close to the already scored compounds (Tanimoto similarity > 0.4) but diverse to each other; (2) using a machine-translation-based autoencoder model[17], encode these molecules to a continuous latent space representation; (3) train a kernel Support Vector Regressor (sklearn) on these latent representations and available scores to predict the score of all new molecules; (4) assemble a list of ~7,000 diverse compounds with the highest predicted scores for another docking round. 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.

Compound grouping and ranking strategy. We grouped the compounds into 3 distinct groups: DrugBank (6490 compounds), covalent (6010 compounds), and non-covalent (31485 compounds). 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 together by docking score. We considered a compound as covalent if it had at least one of the following groups: esters, aldehydes, trifluoromethylketones, chloromethylketones or sulfonyl fluorides. We compile a ranked list for each of the three groups and a final ranked list of compounds. For the final list we took the top-100 compounds from the DrugBank list, all compounds from the covalent list and as many compounds from the noncovalent list as needed to reach a total of 10,000 compounds.

aiwinter (2)

Team members. Filipp Gusev , Evgeny Gutkin, Roman Zubatyuk, Chamali Narangoda, Hatice Gokcan, Maria Kurnikova, Olexandr Isayev

We screened 4.59 billion of molecules against three (Mpro, PLpro, TMPRSS2) targets. All molecules were triaged according to 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 (receptors, ion channels, 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.

Protein system preparation. X-ray crystal structures for Mpro (PDB IDs 6M03, 6WNP, and 7BQY) and PLpro (6WRH, 3E9S, 3MJ5, 4OVZ, and 4OW0) were obtained from PDB. For TMPRSS2, a homology model of the catalytic chain residues 260-489 (UniProt accession number O15393) was generated in SwissModel15[18] using hepsin (PDB ID 5CE1) as a template. All structures were protonated, solvated, and parametrized using tleap program of AMBER 18[19]. AMBER ff99SB-ILDN[20], GAFF[21] and TIP3P[22] parameters were used for protein, ligands and water correspondingly. Molecular dynamics (MD) simulations were performed for all structures using pmemd.cuda program of AMBER 18[19]. A detailed description of the system preparation and simulation protocol is provided elsewhere[23]. Representative structures extracted from production trajectories were used as receptors for docking.

Virtual chemical library preparation. Our overall screening library was 4.59B small-molecule ligands. Molecules were obtained from 8 chemical databases (WuXi GalaXi, ZINC, Enamine REAL, PubChem & ChEMBL, Mcule Purchasable (Full), Merck, ChemSpace «In Stock», and CAS Antiviral DB). All SMILES were curated according to our well-established best practices[24,25] including standardization, the cleaning of salts, and the removal of mixtures, inorganics, and organometallics using ChemAxon 21 software[26].

Docking protocol. For each target, ligands were docked to several receptors using AutoDock Vina software[15] with exhaustiveness setting of 16. For each docking experiment, five highest score ligand poses were saved and re-scored with OpenEye ChemGauss4 (CG4) score[27]. For each target, residues that have conserved interactions with native ligands (key residues) were identified by inspection of crystal structures. Penalized composite docking score S* was computed for each pose using the following formula:

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

where S is the consensus docking score (average of Vina and CG4 scores), M is number of key residues, n(d_k) is rank of minimum distance d_k from the ligand to a key residue for this particular protein target, and N is the number of ligands docked. The final composite docking score was calculated as the average of scores for several receptors. To accelerate the screening of billions-size database, we have employed AIMNet-2D deep neural network model, a modification of the original AIMNet[28] adjusted to operate on molecular graphs and use topological distance, to predict docking scores from 2D molecular structure of ligands. Similar to the Deep Docking approach[29], we have applied an active learning technique to construct a dataset iteratively, selecting molecules with high docking score predicted on the model trained on the previous iteration. As a seed dataset, we used a random sample of 100k molecules from an in-vitro subset of the ZINC dataset. For every subsequent iteration we have added 100k random molecules from combined 4.5B dataset with predicted docking score less than 10th percentile of docking score distribution in the current version of the dataset.

QSAR modeling protocol. All models were trained within two nested five-fold cross validation loops using xGBoost[ref]. All splits were random. Final scoring was done by simple averaging of five predictions from the external CV loop. Internal CV loop was used to perform hyperparameters search and variable selection for the corresponding fold using the protocol described elsewhere[24]. Each target dataset was curated according to well-established best practices[25].

To model MPro and PLpro we used historical data of putative inhibitors screening SARS-CoV from publicly available databases: BindingDB, PubChem, ChEMBL. To model TMPRSS2 due to lack of data for the TMPRSS2, we've selected set of representative proteins ST14_HUMAN, TRYB1_HUMAN, FA11_HUMAN, HEPS_HUMAN, KLKB1_HUMAN, PLMN_HUMAN, 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.

For each collected dataset (to be referred as local) we build a complementary one, referred as global. We augmented each of the 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.

As part of the pipeline we have used our ML method to develop 13 absorption, distribution, metabolism, excretion, and toxicity (ADMET) prediction models. 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).

Novelty. To assess the novelty of a reported aiwinter hit compound (nsp5-14) we performed a search for a nearest neighbor based on Tanimoto similarity score (Ts). Tanimoto similarity was computed using ECFP with radius 2 and bit vector length 2048. The closest analog had similarity score Ts=0.51 with annotated active molecules from nsp5 (MPro) training dataset. The closest analog from Chembl 29 (published in July 2021) has Ts=0.54. Overall this suggests that the observed hit was novel at the time of the competition

belarus (3)

Team members. Alexander Tuzikov , Alexander Andrianov, Yuri Kornoushenko, Hanna Karpenka, Ivan Bosko, Nikita Shuldov, Artsemi Yushkevich, Konstantin Furs

Target structures. The target proteins were taken from the Protein Data Bank[30]. We used the main protease of the SARS-CoV-2 virus (3CLpro) with the X77 inhibitor (PDB ID: 6W63), the main protease of the SARS-CoV2 virus (3CLpro) (PDB ID: 6Y84), the S1 protein receptor-binding domain in complex with the ACE2 protein (PDB ID: 6M0J).

Virtual screening. Pharmacophore-based virtual screening was performed using the Pharmit server software[31], utilizing all the nine available molecular libraries (Pubchem, Molport, ZINC, ChemDiv, etc.), containing over 213.5 million chemical structures. For the 3CLpro targets, several pharmacophore models were built on the basis of the interaction of the X77 inhibitor with the SARS-CoV2 main protease. For S1 protein, the pharmacophore models were built on the basis of the interaction of Lys353 ACE2 receptor residue with S1 protein residues. The selected databases were searched for compounds that match the specified pharmacophore using the Pharmit web server[31]. The best screening compound was used to create a new pharmacophore model. A repeated search was performed on the selected databases. Energy minimization of results was used to optimize both the pose and conformation of identified hits with respect to the provided receptor using the AutoDock Vina[15] scoring function and Smina[14], a fork of AutoDock Vina with enhanced minimization functionality. Minimized results were sorted by affinity and saved as a compressed SDF structure file.

Molecular docking. Before docking, hydrogen atoms were added to the ligand and receptor structures and their optimization was performed in the UFF force field[32]. For this purpose, the Open Babel[33] program was used. Molecular docking of all compounds was performed using the Quick Vina 2 program[34] with a conformational coverage parameter of 10.

Molecular dynamics. The values of binding free energy were calculated with Amber18[19] using the MM/GBSA method. The calculations were made for 200 snapshots extracted from the final 40 ns of the MD trajectories, by keeping the snapshots every 0.2 ns. The polar solvation energies were computed in a continuum solvent using the Poisson-Boltzmann continuum-solvation model with ionic strength of 0.10 M. The non-polar terms were estimated using solvent accessible surface areas.

Machine learning. Molecular ACCess System (MACCS) fingerprints[35] were obtained from SDF structure files and binding free energy was parsed using docking results. Fingerprints for each compound were matched with corresponding binding energy to form a dataset used for training of an adversarial generative autoencoder[36]. The training was conducted in a semi-supervised manner, where fingerprints were fed to the encoder, while binding energy was fed to the special neuron on the latent layer, and the decoder used both the compressed representation obtained by the encoder and the binding energy to restore original fingerprints. Data on the latent layer were additionally discriminated with a normal distribution, to enforce the encoder to make meaningful compressed representation. Later, the model was used to generate new molecular descriptors (fingerprints) with a preset property of the binding energy. Namely, in order to generate new fingerprints, numbers from normal distribution were sampled to the latent layer, while the neuron responsible for binding energy was given a threshold value to generate new fingerprints with (e.g. -10 kcal/mol). For the generated fingerprints, a similarity search was conducted among ZINC library compounds, using L1 distance as a metric. Closest compounds were subjected to docking procedures. Python package RDKit[37] was used for preprocessing and MACCS fingerprints generation. Tensorflow 2.1 for Python[38] was used as a deep learning framework.

cermn (4)

Team members. Ronan Bureau, Patrick Bousquet-Melou, Beatrice Charton, Bertrand Cirou, Mohammed Benabderrahmane

Targets. Three targets were considered : 1) COVID-19 main protease[39,40] (3C-like proteinase, pdb: 6W63), 2) ADP ribose phosphatase of NSP3[41] (Papain-like protease (PLpro)), pdb: 6W02),

3) Human transmembrane protease serine 2[13] (TMPRSS2): Source: covid.molssi.org., Homology model based on the structure of TMPRSS15/enteropeptidase11(pdb: 4DGJ).

Chemical libraries. Five chemical libraries were considered in the overall studies with Real database as the main basis: 1) Real database from HTVS project[42] (MW between 375 and 500 daltons / logP between 1 and 4,5 (382 millions compounds) and MW between 325 to 375 daltons / logP between 2,5 and 4,5 (240 millions compounds)), 2) Sweetlead library[43] from SimTK (selection of 7134 compounds with MW between 200 and 700 Daltons) 3) Merck library (more than 5 millions compounds) 4) ChEMBLDB[44] for proteases (53092 compounds for which at least a pKi/pIC50 value is superior to 6), 5) Chembl4303835 (SARS-CoV-2, 5653 compounds).

Docking. Virtual Screening (VS) was based on docking approaches by two methods: High-Throughput Virtual Screening (HTVS[42]: VIRTUAL FLOW with QVINA2, AutoDockTools[13] for the definition of the configuration file) for big dataset and classical VS (Glide[45] from Schrodinger) for a small database (sweetlead).

Data analysis. A clustering of the HTVS / VS results was done with kmeans (Pipeline Pilot) with a selection/ranking of the best ligands. This was carried in four steps : a) clustering with ECFP4 as descriptors (1024 bit) / Coefficient of Tanimoto for the distances / 3000 clusters b) For each cluster, extraction of the centroid and the 5 best ligands (QVINA2 score) c) For the selection, determination of a druglikeness with calculation of the Quantitative Estimate of Druglikeness (QED, the ligands should have a value superior or equal to 0.5) and d) Ranking of the final selection in function of the QVINA2 score.

Results. Target 1. 3C-like proteinase. Scoring data for 559887710 compounds (Min : -11.5 / Max : 400.3 / Median : -7.1). Cutoff at -9.0 for the first selection : 930619 compounds. Second HTVS on the 930619 compounds (flexibility of the active site and the ligands with an exhautiveness level of 8). Cutoff at -9.5 for the second selection : 57038 compounds selected. Clustering of the final results (towards 10000 compounds) : 10613 compounds selected (classification in function of the docking score). Similarities between the 57038 compounds and the 271 compounds described as active towards SARS-COV2 (Chembl4303835) : 32 compounds with similarities >= 0.5. Target 2. ADP ribose phosphatase of NSP3. Scoring data for 567746375 compounds (Min : -13.6 / Max : 355.7 / Median : -4.8). Cutoff at -10.1 for the first selection : 1106522 compounds. Second HTVS on the 1106522 compounds (flexibility of the active site and the ligands with an exhautiveness level of 8). Cutoff at -10.7 for the second selection : 50616 compounds selected. 3. Clustering of the final results : 12227 compounds selected (classification in function of the docking score). Similarities between the 50616 compounds and the 271 compounds described as active towards SARS-COV2 (Chembl4303835) : 13 compounds with similarities >= 0.5. Target 3. TMPRSS2. Scoring data for 586252945 compounds (Min : -11.8 / Max : 127.5 / Median : -6.7). Cutoff at -9.0 for the first selection : 1185453 compounds. Second HTVS on the 1185453 compounds (flexibility of the active site and the ligands with an exhautiveness level of 8). Cutoff at -9.6 for the second selection : 58407 compounds selected. Clustering of the final results : 9009 compounds selected (NSP5-5 is in this set). Similarities between the 58407 compounds and the

271 compounds described as active towards SARS-COV2 (Chembl4303835) : 23 compounds with similarities >= 0.5.

covid19ddc (5)

Team members. Stan Watowich, Suman Sirimulla, William Allen, Xiaodong Cheng, Robert Davey, Andrea Dimet, Francisco Enguita, Amit Gupta, Yurii Moroz, Pei-Yong Shi, Clifford Stephan, Adrian Varela-Alvarez, Jin Wang, Mark White

Protein targets. ADP ribose phosphatase (MAC1 or X domain) (Nsp3), Main protease (Mpro; also called 3CL-pro) (Nsp5), RNAase (Nsp15), methyltransferase (Nsp16 / Nsp10), papain-like protease (PLpro) (Nsp3 domain), RNA dependent RNA polymerase (RdRp) (Nsp12)

Method. Six different SARS-CoV-2 proteins (see above list) were prioritized as small molecule drug targets and their 3-dimensional crystallographic structures retrieved from the Protein Data Bank. For each protein chain in the available unique PDB structures, active sites conducive for inhibitor binding were identified. For example, two PDB (Protein Data Bank) entries (6W02, 6W6Y) were available for the SARS-CoV-2 nsp3 ADP ribose phosphatase protein, and each PDB file contained two structures in the crystallographic asymmetric unit, which resulted in four unique structures for use in the virtual screening calculations. All unique protein chains (a total of 17 different structures) for each SARS-CoV-2 protein target, together with chemical libraries containing 2.6 million drug-like, commercially-available small molecules, were optimized and parameterized for very large-scale virtual screening. Scripts were streamlined to efficiently perform large-scale Vina-based virtual screening calculations on TACC supercomputing systems (Frontera, Stampede2, Lonestar5, and BOINC@TACC). Vina-based virtual screening calculations against the SARS-CoV-2 proteins were rapidly completed. The virtual screening phase of this project systematically calculated optimal binding structures and energies between each small molecule in a 2.6 million-member library (encompassing Enamine Ltd's HTS, Advanced, and Premium chemical collections) and each SARS-CoV-2 non-structural protein (e.g., Mpro, RdRp, helicase, RNAase, papain-like protease). These calculations identified several thousand small molecules ("hits") predicted to tightly bind and inhibit critical active sites within each of the SARS-CoV-19 non-structural proteins. This list of several thousand hits for each protein target was subjected to a second round of virtual screening using the Schrodinger Glide algorithm. Results from the Vina and Glide scoring calculations were re-ordered using ordinal averaging to produce a consensus list of compounds ranked by average consensus ordinal score.

deeplab (6)

Team members. Nick Antonopoulos, Nikolaos Epitropakis, Agamemnon Krasoulis, Vassilis Pitsikalis, and Stavros Theodorakis.

Target structures and preprocessing. We downloaded all targets from the Protein Data Bank (PDB)[30]. We used the following three targets: spike protein S (PDB ID: 6M0J), 3CLpro / Nsp5 (PDB ID: 6Y7M) and PLpro / Nsp3 (PDB ID: 6WUU). We extracted a total of 15 pockets using the P2Rank algorithm[46], three for the spike protein and six for each of the 3CLpro and PLpro targets. We extracted both atomic[47] and pocket surface[48] features from the protein pockets.

Ligand databases and preprocessing. We used four small molecule databases for virtual screening: ZINC15, MERC, CAS and SWEETLEAD. We only used ligand atomic features[47].

Binding affinity prediction. For each target we approached the problem of ligand ranking via binding affinity prediction using graph neural networks (GNNs). The pocket and ligand feature graphs were transformed into fixed-vector representations using dedicated GNNs. The two vectors were then combined using an outer-product layer and a dense layer was finally used to predict binding affinity (K_d or K_i). We used two distinct types of networks, each one using as input either the atomic or pocket surface feature representation.

Model training. We trained the two binding affinity networks on the refined set of the PDBbind v.2019 database (4,852 protein-ligand pairs). The mean-squared-error loss function was used for model training. We generated artificial negative samples during training by combining protein pockets and ligands from different complexes in the training set. We assigned negative samples a zero log-affinity value and the positive to negative sample ratio was set to unity.

Final selection. We screened all ligands against all 15 pockets. For each target, we aggregated estimates made for all corresponding pockets by considering for each target-pair combination the pocket that yielded the highest binding affinity score. For each target we then compiled two lists of the 10K top-ranked (i.e. highest binding affinity score) ligands, one for each type of network. All ligands included in both lists were selected for the final submission. The remaining ligands for each target were selected based on their maximum estimated score across the two models. A total of 10K ligands were selected for each of the three targets.

Further information. Our submission was based on our Deep Neural Virtual Screening (DENVIS) algorithm described in ref.[49]. An inference API using our models is publicly available[50].

imolecule (7)

Team members. Petr Popov, Alexander Medvedev, Igor Kozlovkii, Mark Zaretckii, Anton Maliutin.

Method. Our goal was to investigate cryptic binding sites in the COVID targets followed by the virtual ligand screening campaigns (VLS) against the identified binding sites. Therefore, on the first stage we analyzed the three-dimensional structures of the SARS-CoV-2 proteins using our recently developed deep learning approach called BiteNet[51] for spatiotemporal identification of binding sites in protein structures. BiteNet is applicable to molecular dynamics trajectories and capable to identify the most promising protein conformation for VLS. We applied BiteNet to all available three-dimensional structures, models, and available molecular dynamics trajectories (for

3CLPro, RNAPolymerase, and Spike) retrieved from the molssi resource[52]. All the protein conformations for each target were superimposed, the predicted binding sites were clustered, and the top 3 binding sites per target were selected for VLS campaigns.

On the next stage we used the ligand-based screening of the REAL Enamine library consisting of ~1.4B small molecules. As the first filter we used the 'rule of three' criteria, thus focusing on ~70M lead-like molecules. As the second filter we used the derived target-specific QSAR models. More precisely, to derive QSAR models we have used SARS-CoV datasets from collected from AICure (https://www.aicures.mit.edu) and Chembl (http://chembl.blogspot.com/ 2020/05/chembl27-sars-cov-2-release.html?m=1) resources. In total we derived three independent QSAR models: general, 3CLPro-specific, and PLPro-specific. The general model was directly trained on the set of anticovid compounds, while the 3CLPro- and PLPro-specific models was trained on SARS-CoV datasets and tested on SARS-CoV-2 datasets in order to obtain optimal hyperparameters. We used Morgan fingerprints, as the molecular descriptors, and Xgboost, as the machine learning algorithm. We used precision as the target metric in order to minimise number of false positive predictions. The best precision on the test sets was ~0.25 for all datasets. Overall by applying the QSAR filters we further narrowed down the chemical libraries to the ~1.5M of the most promising compounds.

Given the identified binding sites and the filtered chemical libraries, we run structure-based VLS campaigns. We used VirtualFlow (https://virtual-flow.org) installed to the Skoltech HPC cluster Zhores[53] and the Smina docking software (https://sourceforge.net/projects/smina/). In total we used ~500K CPU hours to dock ~1.5M compounds in 15 docking runs: single docking run per binding site, 3 binding sites per protein target, 5 protein targets.

In order to take into account the FDA-approved small molecules as well as the investigational drugs, we used the ICM-Pro docking software (http://molsoft.com/) for molecular docking of the DrugBank library (https://www.drugbank.ca/). We used semi-empirical quantum mechanics calculations to generate 3D conformers for this chemical library. Similarly we used 15 independent docking runs: single docking run per binding site, 3 binding sites per protein target, and 5 protein targets.

Finally, to explore possibility to experimentally test custom small molecules, we used structure-based deep learning approach for de novo drug design for the RNA polymerase and the 3CLPro targets starting from known active compounds docked into the corresponding structures. We applied our recently developed 3D shape generator (unpublished) to the voxelized representation of small molecules bound to proteins, thus, generating ~1K de novo small molecules. We ranked the designed compounds using the derived QSAR models and selected up to 10 best molecules into the final list.

As we used different scoring functions and different binding sites within each target, we cannot rigorously estimate the binding affinities. Instead we firstly ranked predictions with respect to each binding site giving priority to the Drug Bank compounds with the ICM score < -30.0. Then from each binding site we selected top 4000 hits and combined the results, hence list of 12K compounds. Then, we multiply the inner rank of each compound by the rank of the corresponding binding site. Finally, we removed the duplicated compounds and manually rearranged the top 50 compounds to ensure that top compounds for each binding site present within top 50.

Our main idea was to investigate vulnerable regions on the SARS-CoV-2 protein structures, that are 'druggable'. For this purposed we applied BiteNet, state-of-the-art binding site detection approach to the conformational ensembles of SARS-COV-2 protein targets : Spike, 3CLPro, PLPro, Nucleocapsid, RNA polymerase. In general we seek for the binding sites that are hidden from the naked human eye using artificial intelligence approach. The main resource we used for the structures and MD simulations was covid.molssi.org. In case of experimentally determined three-dimensional structures we used the refined structures from insidecorona.net. Routine standartization of the structures, e.g. adding hydrogens, missing heavy atoms, rotamer optimizaiton, etc, was done using ICM-Pro; converting to the .pdbqt format was done using the Open Babel suite (http://openbabel.org).

Library 1 - REAL Enamine. The REAL Enamine contains more than 1.4B compounds. On the first stage we applied RO3 filter and keep only compounds with i) octanol-water partition coefficient log P not greater than 3, ii) molecular mass less than 300 daltons, iii) not more than 3 hydrogen bond donors, iv) not more than 3 hydrogen bond acceptors, v) not more than 3 rotatable bonds, vi) Polar surface area no greater than 140 A^2, yielding 71.3M compounds. On the next step we applied the derived QSAR models in order to obtain 3CLPro-specific, PLPro-specific and general anticovid chemical libraries (for N protein, Spike, and RNA polymerase) of ~1.1M, ~1.5M, and ~1.0M compounds respectively.

Library 2 - DrugBank. We retrieved small molecules from DrugBank and applied sanitazing procedure according to the Chembl structure standartization pipeline yielding 8282 compounds. Then we generated 3D conformers and assign partial charges for each molecule using semi-empirical quantum mechanics approach implemented in the ORCA software (https://orcaforum.kofo.mpg.de).

jku (8)

Team members. Starting with the lead and continuing then in alphabetical order, the team members are Peter Ruch, Hamid Eghbal-zadeh, Christina Halmich, Helga Ludwig, Andreas Mayr, Philipp Renz, Elisabeth Rumetshofer, Johannes Schimunek, Philipp Seidl, Andreu Vall and Michael Widrich.

Method. The ligand-based screening[54,55] consists of five independent Machine Learning methods for compound selection namely RandomForest, Gradient Boosting, Self-Normalizing Networks (SNN)[56] and two LSTM-based sequence models[57]. For the first LSTM-based approach, several models were trained with a Multi-Task objective, followed by a target-specific model selection to identify the models which work best for a specific target. For the second LSTM-based-approach, the model was pretrained on a large dataset and then fine-tuned in a transfer-learning scheme. The RandomForest, Gradient Boosting and the SNNs are descriptor-based, whereas the LSTM-based models directly operate on the molecular SMILES representation. Merging the results of the different models, the output of the SNNs are weighted higher, since SNNs showed high-robustness against domain shifts.

kyuken (9)

Team members. Ashutosh Kumar, Francois Berenger, Yoshihiro Yamanishi, Kam Y. J. Zhang

Method. Screened virtual chemical library: 1,002,510,499 molecules in total. Made of ZINC15, AMS, CAS antivirals, KEGG drugs, SWEETLEAD. Full details available in our report, included in the supplementary materials. Here is only a high level view of our massive virtual screening effort, dictated by the challenge rules (need to screen 1B+ molecules, using three different ranking methods each time). Our strategy was to use first fast ligand-based models to trim down the whole 1B+ molecules library to our docking capacity, estimated at maximum 90M molecules (10M molecules per protein target, three times per target, three targets). This approach was later refined by the same authors as the "lean-docking" protocol[58] to make docking scale to large chemical libraries.

Target 1 - 3CLpro (PDB:6Y2F). A SARS-Cov-1 dataset (CHEMBL3927) was used to train one regressor and two classifiers. Those models reduced the whole library down to 770,326 molecules (ranking 1). Those 770,326 molecules were rank-ordered by an OpenEye ROCS query using the bioactive conformation of alpha-ketoamide 13b from PDB:6Y2F (ranking 2).

A docking screen using OpenEye FRED and CCDC Gold provided the third ranking for this target.

Target 2 - ACE2 (PDB:6LZG). We targeted on human ACE2 the Protein-Protein Interface (PPI) with the receptor binding domain of the viral spike protein. First, a classifier was trained using IPPIDB plus a 25 times larger random chemical background drawn from ChEMBL-24 not included in IPPIDB. This model trimmed down the chemical library to 2,461,774 molecules. The FTMap server was used to define a probable ligand-binding site on this PPI. Docking was performed using OpenEye FRED (1st ranking) and CCDC Gold (2nd ranking). OpenEye ROCS was used using a query generated from FTMap probes located on the ACE2/spike PPI (3rd ranking).

Target 3 - PLpro (PDB:6WUU). Two ligand-based classifiers were trained using PubChem assay 1944. They trimmed down the chemical library to 4,001,268 molecules (ranking 1). OpenEye ROCS was used to rank-order those molecules using shape and chemical similarity to the bioactive conformer of peptide inhibitor VIR250 from PDB:6WUU (ranking 2). Docking was performed using OpenEye FRED and CCDC Gold to obtain a 3rd ranking for this target.

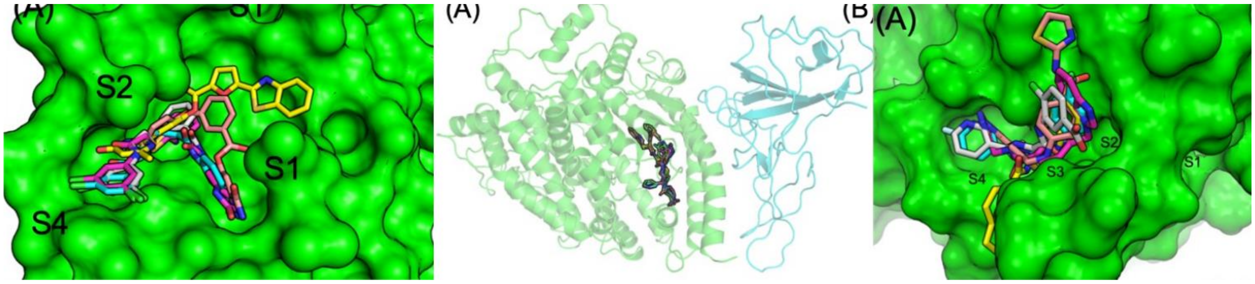


Figure S1: top five predicted ligands and their binding mode for 3CLpro (left), hACE2/viral-spike PPI (middle) and PLpro (right).

lambdazero (10)

Team members. Brooks Paige, Jose Miguel Hernandez Lobato, John Bradshaw, Joanna Chen, Bianca Dumitrascu, Matt Kusner, Marwin Segler, Jarrid Rector-Brooks, Paul Bertin, George

Lamb, Simon Verret, Jian Tang, Will Hamilton, Chenghao Liu, Bruno Rosseau, Kostiantyn Lapchevsky, Aga Slowik, Michael Bronstein, Emmanuel Bengio, Doina Precup, Pierre-Luc Bacon, Yoshua Bengio, Scott Fujimoto, Pierre Thodoroff, Clement Gehring, Shivam Patel, Victor Butoi, Samira Kahou, Riashat Islam, Howard Huang, Evgenii Nikishin, Moksh Jain, Sumana Basu

Method. This project leverages a previously tested, novel and yet unpublished deep reinforcement learning algorithm derived from the recent successes of DeepMind's AlphaZero and of MIT's approach to represent drug molecules. It also uses a new approach to search molecular space using reinforcement learning algorithms developed by the team at MILA and collaborating universities led by prof. Yoshua Bengio (whole team listed at: https://mila.quebec/en/ai-society/exascale-search-of-molecules/). This new approach is based on the definition of a set of molecular building blocks which can be combined to form molecules and search the space at a more abstract level, as opposed to searching by modifying individual atoms.

The research plan consists of the following main steps. (1) Run docking simulations on 200M randomly chosen molecules to bind on virus targets. (2) Keep the highest score as training examples for the initial value function network which approximates the docking scoring function. (3) Train Lambda-Zero (our novel RL algorithm) using the docking score, drug-likeness, and predicted cost of synthesis as the reward function and select top-scoring molecules. (4) Send these molecules to simulated retrosynthesis collaborators (Molecule.one, Dr. Piotr Byrski) (which will identify which of these molecules can be fabricated) and actual chemical synthesis. (5) Send best molecules designed by RL algorithm to computational biophysics collaborators (DE Shaw Research) to perform (a) long-term molecular dynamics and (b) Free Energy Perturbation (6) Send the best performing molecules in the MD simulations to our experimental collaborators for biological assays (IRIC, Dr. Mike Tyers) (measuring actual biological binding to the target) and biomedical assays (measuring protein binding energy, and effectiveness in cells and then animal models). (7) Feed back the results of these assays as additional data to constrain and improve the search, iterating back to step (3) but with a modified reward function which incorporates both docking and empirical data from (5) as part of a trained reward function neural network.

The project delivers molecular structures deemed good candidates for antiviral drugs. If these molecules have good binding energy to the targets, this should convince biomedical collaborators to pursue the pipeline of evaluation of these molecules, leading eventually to clinical trials as well as compassionate use distribution. The project also delivers code for efficient implementation and parallelization of the previously developed LambdaZero algorithm honed in for the search of SARS-CoV-2 antivirals.

Ici (11)

Team members. Dragos Horvath, Gilles Marcou, Alexandre Varnek

Method. The following chemoinformatics strategies were used in our approach:

1. Database standardization: The ~1.3 billion ZINC compounds (zinc.docking.org; highest reactivity: “standard”, minimum purchasability “boutique”) plus the the 5M compounds of the AMS Merck library were standardized on the chemoinformatics web server http://infochim.u-strasbg.fr/webserv/VSEngine.html.

2. Chemoinformatics-driven preselection.

a. 55K ZINC molecules present within the Relevant Antiviral Spaces (RAS) of at minimum three of the seven Universal Maps[59] were selected following the already published procedure[60] which originally reported the >400 compounds residing in 4/7 of these RAS.

b. 58K ZINC molecules were selected by a consensus similarity screening using as references 84 antiviral compounds from the DrugBank (compound names ending in "-vir"), using the seven ISIDA fragment descriptor[61,62] spaces at the basis of above-mentioned Universal maps and Tanimoto score. Only ZINC compounds which "made it" into the top 1000 neighbors in two or more of the seven descriptor spaces (consensus neighbors) were kept

c. The published audit[60] of coronavirus-related structure-activity data revealed a very small series of compounds confirmed to be active on the SARS-CoV 3CL protease in ChEMBL (as of March 2020). These 25 molecules were at their turn employed as references for similarity-based virtual screening of ZINC, as described above. A pool of 24885 compounds resulted.

d. In absence of any RNA polymerase structure-activity data on coronaviruses, a search for RNA polymerase inhibitors of any viruses was conducted in ChEMBL, and 29 most potent nanomolar inhibitors were selected. They too were used in the similarity search-driven virtual screening.The best-ranked 10K of these were reported as such as RNA polymerase inhibitor candidates (no further docking).

3. The program PLANTS was used for docking (in "precision mode" speed1, with the ChemPLP score) of above preselected sets for the two viral proteases 3CLpro (PDB ID 6W63) and papain-like protease (6W9C). In addition, machine-learning-enhanced docking (protocol not published) was used to process the entire AMS Merch database. The final selections were compounds containing between 20 and 40 heavy atoms located at the "Pareto front" of optimal ChemPLP docking score and ChemPLP ligand efficacy (i.e. docking score/number of heavy atoms).

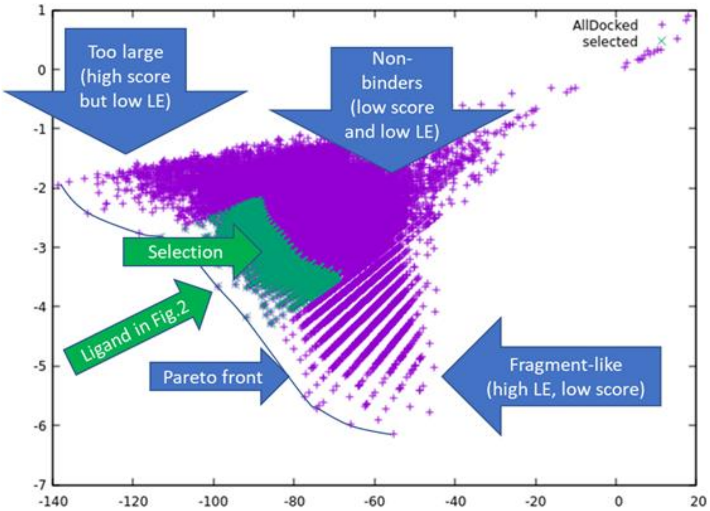


Figure S2: Enlarged Pareto front (Global ChemPLP docking score:X, ligand efficiency:Y) for the 108K (ZN-RAS+ZINC-VIR) compounds docked into 6W9C. In green: selected 10K compounds, of intermediate sizes (between 20 and 40 heavy atoms).

luxscreen (12)

Team members. Enrico Glaab

Targets considered. 1) 3CLPro/Nsp5 (PDB: 5R8T); 2) helicase / Nsp13 (PDB: 6JYT); 3) TMPRSS2 (homology model template: PDB: 5TJX).

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. For proteins with multiple chains, the chain with the highest Verify3D score was chosen for further analysis.

Ligand pre-processing, feature tree search & 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”. These compounds were downloaded as SMILES, using the “ZINC-downloader-2D-smi.wget” script derived from the “Tranches” web-page on ZINC.

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; 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.

nuwave (13)

NUWAVE (Northeastern University Warriors of the Anti-Viral Enterprise).

Team Members. PI: Mary Jo Ondrechen; PhD students: Suhasini M Iyengar, Kelly K Barnsley; Undergraduates: Hoang Yen Vu, Ian Jef A. Bongalonta, Alyssa S. Herrod, Jasmine A. Scott

Methods. For the binding site prediction, Partial Order Optimum Likelihood (POOL)[63–65] was used. Partial Order Optimum Likelihood (POOL) is a machine learning method that predicts biochemically active amino acids using the three-dimensional structure of the query protein as input. POOL predicts multiple types of binding sites in proteins which include catalytic sites,

allosteric sites and other sites, some of which may not be detected by other predictive methods because POOL is based primarily on computed electrostatic and chemical properties of the query protein. POOL points to the residues involved in reversible binding, including catalytic sites and non-catalytic binding sites such as allosteric sites, ligand transport sites, and some protein-protein interaction sites. The other input features for POOL consist of properties of the local environment and surface topological metrics.
Molecular Docking was performed using Schrödinger Glide 2019-3 on the Discovery Cluster at the Massachusetts Green High-Performance Computing Center. Glide Standard Precision (SP) was used as an initial screen and top predicted ligands with docking score of <=-7 kcal/mol were used for Glide Extra Precision (XP).

Targets. The target proteins were downloaded from the protein data bank, except for NSP1, for which a comparative model structure was built. The N Protein model structure was built in YASARA[66] using a series of structures from the Protein Data Bank. These structures were obtained after a BLAST search of the N Protein sequence. The model was built by manually providing template structures with sequence homology to N Protein. These templates are Crystal Structure of SARS-CoV-2 nucleocapsid protein N-terminal binding domain (PDB ID:6M3M), Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS-CoV-2 (PDB ID: 6VYO) crystal structure of C-terminal dimerization domain of Nucleocapsid Phosphoprotein from SARS-CoV-2 (PDBID: 6WJI) and the N-Terminal binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein (PDBID: 6YI3). Using these three structures as templates, YASARA built a hybrid model for the N protein. Details of our method, the screened ligand libraries, and the types of interactions between active amino acids in the target proteins have been reported[67].

pharmAI (14)

Team members. Joachim Haupt, Florian Kaiser, Michael Schroeder

Method. PharmAI uses own proprietary, knowledge-based algorithms that exploit hidden information in protein structures. It is a combination of several methods, including: protein binding site similarity[68], protein-ligand interaction similarity[69], and sophisticated chemical similarity. The PharmAI DiscoveryEngine allows to quickly select from millions of compounds to obtain a ranked list specific for the given target. The results from the independent methods are ranked and combined using an empirical P-value. In a recent benchmark study, this so-called Focused Library approach, achieved a hit rate of 6% and identified 7 new lead compounds cGMP-dependent 3',5'-cyclic phosphodiesterase inhibition. For more details see ref.[70]. The unique approach of our DiscoveryEngine guarantees a maximal scaffold diversification to discover new chemical entities. Please understand that we cannot disclose any more details about the algorithm as this is our main IP.

safan (15)

Team members. Luisa Pugliese[1]

[1]S.A.F.AN. BIOINFORMATICS, Turin, Italy

Methods. The computational profiling was carried out using our proprietary technology SAFAN-ISP. It is a ligand based method calculating the binding affinities between each ligand and more than 4500 targets from 15 different protein classes. As many ligand based methods SAFAN-ISP is based on the molecular similarity evaluation between the submitted molecule(s) and those in an active compound database containing experimental data concerning the interactions between molecules and protein targets. SAFAN-ISP calculates similarities using newly derived fingerprints based on small substructures. The similarity (Tanimoto) is computed matching atoms to substructures and evaluating common atoms.

For the experimental data we use a refactored bioactivity database derived from the CHEMBL25 database.

The technology involves three steps:

1. Molecule fragmentation: The input molecule is split into fragments following a newly derived scheme. Each fragment is compared to SAFAN-ISP database of fragments, derived from the compounds included in our bioactivities database. Fragments are used to :
    - select targets sharing similar fragments with the input compounds,
    - calculating affinities combining concerning different compounds binding the same target
2. Affinity calculation based on compound similarity: The similarity of the input compound is derived with all the compounds present in SAFAN-ISP database. Next the binding constant on all targets outputted from step 1. is computed combining similarities and experimental data using four different schemes, two including chiral fingerprints and two excluding them.
3. Weka Machine Learning approach: The REPTree algorithm, available from the WEKA open source package, is used to combine all binding constants derived in steps 1. and 2. in a single value that will be the final output. RepTree uses the regression tree logic and creates multiple trees in different iterations. After that it selects the best one from all generated trees.

sarstroopers (16)

Team members. Goutam Mukherjee, Giulia D’Arrigo, Eleonora Gianquinto, Sandra Kovachka, S. Kashif Sadiq, Daria B. Kokh, Alexandros Tsengenes, Christina Athanasiou, Abraham Muniz Chicharro, Ariane Nunes-Alves, Anton Hanke, Giulia Paiardi, Jonas Gossen, Simone Albani, Benjamin Philipp Joseph, Francesco Musiani, Candida Manelfi, Carmine Talarico, Andrea Beccari, Paolo Carloni, Giulia Rossetti, Francesca Spyrakis and Rebecca C. Wade

Method. The computational virtual screening pipeline was carried out by three research groups, who independently conducted different parts of the pipeline and then together merged the results and ranked the compounds. The pipeline included rapid virtual screening using several methods (scaffold searching, RASPD+[71], ligand similarity searches and ROCS ligand pharmacophore-based screening[72]), molecular docking with three independent methods (GOLD[73,74], Glide[45,75], FRED[76]), molecular dynamics simulation of docked complexes and trajectory analysis using the MD-IFP[77] workflow, and ADMET screening with the Schrodinger Qikprop[78] module. The pipeline allowed us to assign compounds to different classes according to their ranking as potential binders to the four SARS-CoV-2 protein targets studied: Nsp3, Nsp5, RdRp and S receptor binding domain. All data for this virtual screening, including a list of compound libraries screened and the scaffolds used for screening, have been deposited on Sciebo: https://fz-juelich.sciebo.de/s/j3S598yadmudMow

sarswars (17)

Team members. Vishwesh Venkatraman, Daniel R. Olson, Jeremiah Gaiser, Conner J. Copeland, Travis J. Wheeler, and Amitava Roy

Methods. We targeted only viral proteins (NSP12, 3CLPro, and N), as this aligns with our long-term research goal. As a pre-processing step, we identified potential binding pockets for the target proteins through a combination of the cavity detection tools FTMAP[79] and POCASA[80], literature review, and visual inspection. Six pocket-like regions were identified: 2 each for N, NSP12 and 3CLpro. Some of the pocket-like regions were too large to be occupied by a typical sized ligand, so these larger pocket-like regions were subdivided into smaller pockets. In total, 22 pockets (8 each for N and NSP12 and 6 for 3CLPro) were finalized as targets.

To identify candidate ligands, we developed a multi-stage screening pipeline that can rapidly explore drug candidates from a library of billions of molecules. The pipeline, drugsniffer, has since been released[81], and required ~40,000 total compute hours to screen for potential ligands for the 3 target proteins among a library of ~3.7 billion candidate molecules.

For each pocket, drugsniffer designs a collection of ligand molecules from scratch, using the software AUTOGROW4[82]. Unique structures were generated over 5 independent runs of AUTOGROW against each pocket, resulting in a total of 30,000 designed molecules. Using these molecules as seeds, the union of several chemical libraries was queried to identify library-sourced compounds similar to the seeds. Similarity was assessed using 1024-bit ECFP4 fingerprints, which were computed on all AUTOGROW4 molecules, and also for all ~3.7 billion library compounds (fingerprints computed using RDKIT[83]). Among all library compounds, ~97,000 had Tanimoto similarity greater than 0.6 to some seed ligand, and another ~955,000 had Tanimoto similarity of 0.5-0.6. Among the 97K closer neighbors: ~43K were identified for nsp12, ~34k for N, ~20k for 3CLPro.

All ~1.05M neighbors were docked into the respective protein targets using AUTODOCK VINA[84]. Up to 4 docking poses for each ligand were retained. The drugsniffer pipeline includes a deep learning module, dock2bind, that is trained to discriminate between binding and non-binding compounds based on pose data (see ref.[81] for details). This module serves as a method for re-scoring the results of docking, and was applied to each of the molecule neighbor molecules, and produced a ranked set of candidates. This list was passed through drugsniffer‘s custom ADMET filter to produce the final candidate list.

virtualflow (18)

Team members: Christoph Gorgulla, Amit Singh, Georg Steinkellner, Karl Gruber, Klara Blaschitz, Konstantin Fackeldey, Marco Cespugli, Michael Hetmann, Vedat Durmaz, Christian C. Gruber, Haribabu Arthanari

Methods: We used the recently developed ultra-large in silico screening platform VirtualFlow[42] to screen for inhibitors of SARS-CoV-2. In this unprecedented structure-based virtual campaign, we screened approximately 1 billion molecules against each of 40 different target sites on 17 different potential virus and host targets. We targeted not only the active sites of viral enzymes, but also critical auxiliary sites such as functionally important protein-protein interactions. Target preparation was done in part using molecular dynamics simulations and binding site characterization using Innophore's CavitOmiX and Catalophore platform. Further methodological details have been published by the VirtualFlow team in ref.[85].

way2drug (19)

Team members: Vladimir Poroikov; Dmitry Filimonov; Dmitry Druzhilovskiy; Alexander Veselovsky; Alexey Lagunin; Vladlen Skvortsov, Anastasia Rudik; Alexander Dmitriev; Pavel Pogodin; Leonid Stolbov; Olga Tarasova; Sergey Ivanov; Boris Sobolev; Dmitry Karasev; Tatyana Gloriozova; Kirill Shcherbakov, Polina Savosina; Nikita Ionov; Nadezhda Biziukova; Vladislav Sukhachev

Methods: Taking into account the incomplete and sometimes contradictory information about SARS-CoV-2 virus and its interaction with the host cell, which was available on May 6th, 2020 when the project began, we decided to apply the approach for virtual screening of anticoronavirus hits in the big chemical libraries included three sequential stages.

(1) Selection of the potential hits among the 1+ billion compounds based on the similarity assessment using as the “reference drugs” molecules with experimentally determined anticoronavirus action.

(2) Further filtration and ranking of the selected hits using machine-learning methods implemented in our software PASS (Naïve Bayes Classifier) and GUSAR(Self-Consistent Regression). The training sets were extended and improved permanently in the framework of the whole project.

(3) Verification of the representative examples of the selected compounds using the molecular modeling approach.

General workflow for selection of hits with potential anti-SARS-CoV-2 activity is given in Figure S3.

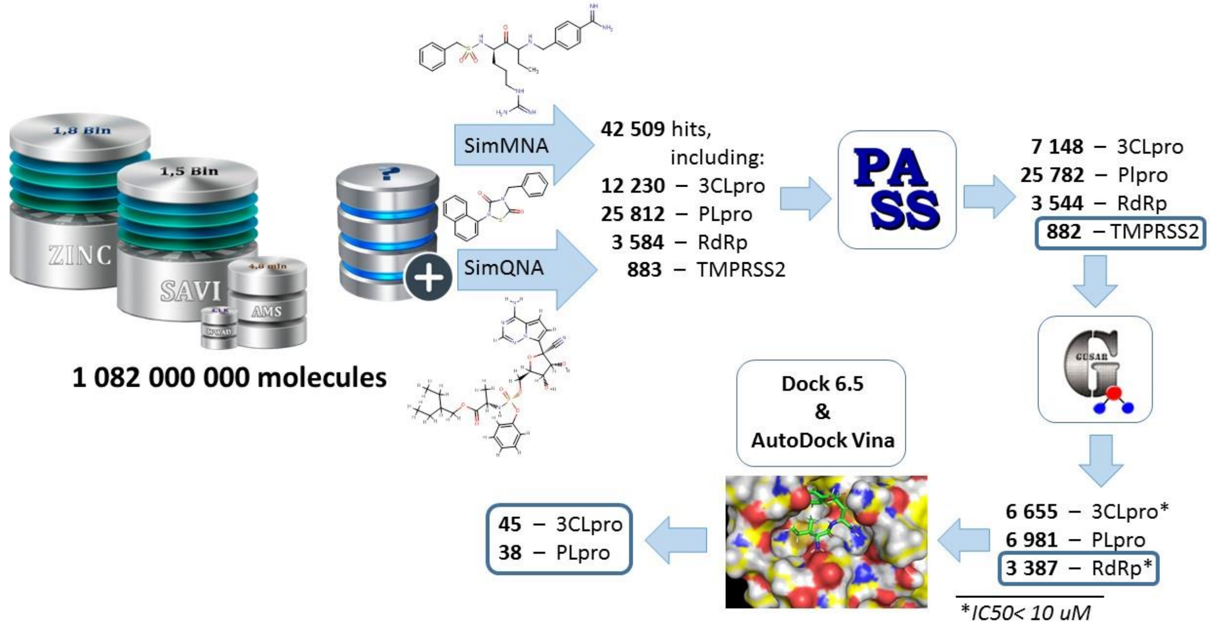


Figure S3. General workflow and results of selection of anti-SARS-CoV-2 hits.

The similarity estimates were carried out using the method based on Multilevel Neighborhoods of Atoms (MNA) and Quantitative Neighborhoods of Atoms (QNA) descriptors (for details see: ref.[86]). Machine learning approaches were applied using our software PASS[87] and GUSAR[88]. Molecular modeling was performed using Dock 6.5[89] and AutoDock Vina[90].

yoda (20)

Team members. Maria Rita Gulotta, Nedra Mekni, Maria De Rosa, Giada De Simone, Jessica Lombino, Ugo Perricone

Methods. SPIKE GLYCOPROTEIN. A preliminary computational alanine scanning of the complex between spike (S) receptor-binding domain (RBD) and angiotensin-converting enzyme 2 (ACE2) peptidase domain (PD), using the PDB structures 6M17[91] and 6M0J[92], respectively was done, then two molecular dynamics (MD) simulations using Desmond (each lasting for 200 ns) were performed on both PDBs. This analysis allowed identifying the key residues of S RBD, and putative hot-spots guided the virtual screening campaign (ZINC12 and SWEATLEAD libraries) based on docking and pharmacophore screenings using Glide (Docking)[45,75,93] and Ligandscout (structure-based pharmacophore)[94]. From the consensus retrieved molecules (43,800), PAINS and REOS filters were applied and finally, the first 10,000 consensus compounds were ranked and retained according to the best score values.

3-CHYMOTRYPSIN-LIKE PROTEASE (3CLpro or M^{pro}). Several SARS-CoV-2 3-chymotrypsin-like proteases (3CL^{pro} /M^{pro}), X-ray structures were available. Particularly, Diamond Light Source synchrotron facility has largely contributed to enrich this collection with numerous M^{pro} complexes co-crystallized with fragments of covalent and non-covalent nature. From the analysis of these PDBs, we focused on the key residues non-covalently bound to the substrate with the help of Maestro GUI and short MD of 50 ns for molecular contacts exploration. X-ray resolution and chemical diversity of co-crystallized ligands were considered to select three M^{pro} PDB structures including non-covalent inhibitors. PDB structures 6W63, 5RF7 and 5R83 were chosen in order to build docking models[45,75,93] to screen ZINC12 and SWEATLEAD libraries on each of these models. The final outcomes were filtered from PAINS and REOS groups and only the consensus molecules from the three models were retained.

PAPAIN-LIKE PROTEASE (PL^{pro}). In order to identify the key residues for SARS-CoV-2 papain-like protease (PL^{pro}), a similarity analysis between SARS-CoV PL^{pro} and SARS-CoV-2 PL^{pro} was performed and a high identity percentage of 86% was obtained. Furthermore, the alignment of the C-α, performed with the Maestro superimposition tool, highlighted that the catalytic sites were conserved on the same region of the two proteases. Therefore, through the study of PL^{pro}crystal structures (PDB IDs: 6WUU and 6WX4), we analyzed the crucial residues for the binding of inhibitors (Thr264, Asp164, Pro248, Gly171) also through MD simulation.

On the basis of the information retrieved from PDBs analysis and the MD simulation(Desmond), a docking model[45,75,93] was built to virtually screen ZINC12 and SWEETLEAD. Finally, the best ranked compounds by docking were filtered from PAINS and REOS groups.

Section 2: Consensus ranking

Method 1: Generative Topographic Mapping

Generative Topographic Maps[95–97] (GTMs) are grid-based chemical space mapping tools, which project the chemical entities represented in the high-dimensional descriptor space onto (square) a 2D-grid of nodes, by fuzzily associating each compound to one or more of these nodes. Association is based on the closeness in descriptor space between the image of a compound (its descriptor vector) and the node coordinates. A compound is said to “reside” in a node, with a “residence time” proportional to its distance to the node. Thus, GTMs are fuzzy Self-Organizing Maps[98,99] (SOMs) – the latter adopting the stricter rule “a compound resides 100% in the node closest to it”. On a GTM, the compound may “spend most of its time” in the closest node, but also be marginally associated with slightly more distanced ones. The sum of these residence “times” – technically called responsibilities – must equal one. Compounds are said to have distinct Responsibility Patterns[59,100] (RP), the concept of which is a natural, intuitive approach to explore chemical space. Each RP corresponds to a well-defined chemical space zone populated by molecules of near-identical responsibility vectors, a “cell” in chemical space. Interesting RPs (for a given target) should be both a) highly populated, and b) “cosmopolitan”, in the sense that its residents have different origins (were proposed by different participants, based on different virtual screening protocols). There is one criterion combining both aspects: the total Shannon entropy .

$$S_{RP} = -N_{RP} \sum_{g} \frac{N_{RP}(g)}{N_{RP}} ln \frac{N_{RP}(g)}{N_{RP}}$$

Above, N_{RP} is the total number of residents in the RP, the sum of RP residents proposed by each group g. In short, for each protein target, each skeleton SMILES is (if ever predicted to concern that target) is defined by the sum of the target-specific Shannon entropies of the RPs it is associated with. Ranking these entries according to the sum of entropies leads to a ranked list by consensuality index (i.e., top-ranked is the most consensual), which was done for each of the protein targets.

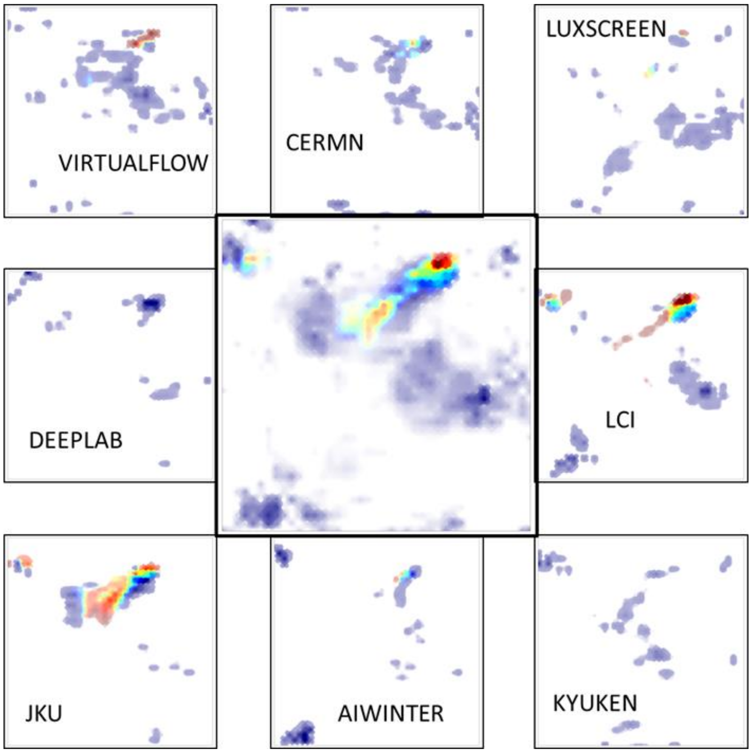


Figure S4: Center, large landscape: global GTM landscape of all Nsp5-selected compounds, all contributors confounded, colored by consensus rank: red - consensus zones, blue – “non-cosmopolitan” areas proposed by a single contributor. Surrounding example landscapes detail the chemical spaces selected by the given contributors for Nsp5 (not all shown), following the same consensus rank-based color code.

Method 2: k-medoids clustering.

The cluster analysis was done separately for each target-protein. For each team and each protein target, the top-1000 valid molecules of each team were collected. Then the extended-connectivity fingerprints[101] or, synonymous Morgan Fingerprints, of size 2048, with radius 2 for each molecule were calculated. This means that each molecule was represented by a vector of substructures or a set of substructures. Next, the Tanimoto similarity matrix, or equivalently the Jaccard index, was calculated. Different Clustering methods with different parameters were then compared (e.g., Agglomerative Clustering, k-medoids, DBSCAN, Affinity propagation...) and compared with respect to their ability to produce meaningful and even-sized clusters of molecules. Based on these criteria, k-medoids clustering with 700 pre-defined cluster-centers was selected. The 700 cluster-medoids were sorted by cluster size, considering the largest cluster as the most important one (i.e., top-ranked). An advantage of k-medoids clustering is that each cluster is represented by an exemplar or a prototype molecule. Thus, the procedure allowed us to provide a ranked list of exemplar molecules that represent a cluster of similar molecules, where the ranking is in decreasing order by cluster size. The lists contained for each molecule also information about each team having proposed that molecule, for which target they proposed it, their position, the cluster size, as well as the molecule encoded as Inchii and SMILES.

Selected molecules list.

For each target, the list of cluster medoids was fused with the top N ranked molecules proposed by every team, where N was chosen such as to accumulate ~1300 different entries in the resulting fused pool. N is therefore dependent on the number of groups having proposed molecules for the given target and the degree of mutual overlap of their proposed lists. The fused pool of each target was further enriched by (not yet included) entries ranked by consensuality index, and herewith expanded to include 2700 distinct molecules (adding the ~1400 most consensual picks not amongst the ~1300 selected at the first step). A global merit index was calculated for each of the 2700 pooled items, as the unbiased average of the “clustering”, “top ranking” and “consensuality” scores (clustering score being proportional to the size of the cluster represented by the medoid – minimum 3).

The pool of 2700 was eventually ranked by global merit, and a procedure to remove lower-merit redundant structural analogues was applied. For each target, the 2700 compounds were encoded as IIRAB-FF-1-2 ISIDA descriptors (these are like Morgan circular fragment fingerprints but use force field types instead of element symbols to “color” atoms) and the full inter-compound similarity matrix was computed. Compound pairs were sorted by their degree of redundancy (Tanimoto similarity), and out of each pair of close neighbors the compound with the lesser global merit score was withdrawn - iteratively, until the selection size of 2700 was reduced to the desired 2000. The final distribution of compounds was 38% top-ranked, 47% GTM, and 15% k-medoids and called the “selected compound lists” (one for each of the 6 protein targets).

Section 3: Detailed team comparison

Hit rate. Table 1 and Table 2 in Section 2.3 and Section 2.4 of the main text show selected molecule counts, synthesized molecule counts and hit rates whereby the numbers were computed only considering compounds which were submitted for the target on which it was tested. Due to the molecule selection procedure, in which some molecules were chosen based on consensus scores, some teams submitted compounds which were not tested on the predicted target but on another one. Table S1 and Table S3 present selected molecule counts, synthesized molecule counts and hit rates irrespective of the predicted target.

Table S1. Overview of selected and synthesized molecules across teams (rows) and drug targets (columns). Molecules which were selected or tested for a specific target but submitted for another target contribute - in contrast to Table 1 - to the team counts.

	Selected molecules (section 2.3 main text)							Synthesized molecules (section 2.4 main text)						
	N	Nsp3	Nsp5	Nsp12	S	TMPRSS2	SUM	N	Nsp3	Nsp5	Nsp12	S	TMPRSS2	SUM
ai4science (1)	60	19	14	25	39	499	656	1	0	0	1	0	16	18
aiwinter (2)	0	64	88	1	1	63	217	0	12	8	0	0	0	20
belarus (3)	0	0	32	1	68	0	101	0	0	0	0	0	0	0
cermn (4)	9	77	82	5	4	626	803	0	8	3	1	0	49	61
covid19ddc (5)	0	79	55	82	1	0	217	0	6	5	10	0	0	21
deeplab (6)	0	60	69	1	160	2	292	0	8	7	0	11	0	26
iMolecule (7)	1013	81	71	358	403	17	1943	73	4	6	42	22	0	147
jku (8)	0	86	259	63	0	0	408	0	0	62	6	0	0	68
kyuken (9)	0	81	52	0	424	0	557	0	15	0	0	41	0	56
lambdazero (10)	0	0	32	0	0	0	32	0	0	0	0	0	0	0
lci (11)	0	1150	700	60	1	2	1913	0	86	54	5	0	0	145
luxscreen (12)	3	5	76	331	9	256	680	0	0	2	14	0	5	21
nuwave (13)	0	39	26	9	2	4	80	0	0	0	0	0	0	0
pharmai (14)	5	16	43	15	293	3	375	0	2	0	0	35	0	37
safan (15)	3	67	81	209	14	7	381	0	2	17	15	1	0	35
sarstroopers (16)	20	61	58	215	132	29	515	0	5	0	19	2	0	26
sarswars (17)	472	2	86	299	0	0	859	8	0	2	9	0	0	19
virtualflow (18)	547	53	112	372	219	487	1790	46	9	2	45	24	44	170
way2drug (19)	3	78	59	57	11	98	306	0	13	3	0	2	21	39
yoda (20)	0	69	90	1	341	2	503	0	3	11	0	24	1	39
SUM	2135	2087	2085	2104	2122	2095	12628	128	173	182	167	162	136	948

Table S2. Number of active molecules, i.e. hits, confirmed with in-vitro testing and hit-rates (ratio of active against tested molecules). The best hit-rate is marked bold. The number of tested compounds is taken from Table 3. Analogous to Table 3, hit counts include compounds for which the predicted target and the tested target is different.

				Hits						
	N	Nsp3	Nsp5	Nsp12	S	TMPRSS2	hits	tested	Hit Rate[1] %	95% conf-int
jku	0	0	14	0	0	0	**14**	89	**15.7**	[8.9-25.0]
covid19ddc	0	0	1	**1**	0	0	2	21	9.5	[1.2-30.4]
kuyken	0	**2**	1	0	**2**	0	5	74	6.8	[2.2-15.1]
aiwinter	0	0	1	0	0	0	1	20	5.0	[0.1-24.9]
ai4sciences	0	0	1	0	0	0	1	34	2.9	[0.1-15.3]
deeplab	0	0	0	0	1	0	1	47	2.1	[0.1-11.3]
cermn	0	0	1	0	0	0	1	64	1.6	[0.0-8.4]
way2drug	0	1	0	0	0	0	1	76	1.3	[0.0-7.1]
iMolecule	**1**[2]	0	0	**1**	0	0	2	314	0.6	[0.1-2.3]
virtualflow	0	0	1	0	0	0	1	225	0.4	[0.0-2.5]
Ici	0	0	0	0	0	0	0	157	0.0	[0.0-2.3]
yoda	0	0	0	0	0	0	0	54	0.0	[0.0-6.6]
safan	0	0	0	0	0	0	0	42	0.0	[0.0-8.4]
sarstroopers	0	0	0	0	0	0	0	29	0.0	[0.0-11.9]
luxscreen	0	0	0	0	0	0	0	24	0.0	[0.0-14.2]
sarswars	0	0	0	0	0	0	0	19	0.0	[0.0-17.6]
belarus	0	0	0	0	0	0	0	0	NA	NA
lambdazero	0	0	0	0	0	0	0	0	NA	NA]
nuwave	0	0	0	0	0	0	0	0	NA	NA
pahrmai	0	0	0	0	0	0	0	0	NA	NA
All teams	1	3	19[3]	2	3	0	28[5]	878	**3.2**	[2.2-4.8]

[1] Hit rate from the pooled analysis described in this paper.

[2] The related compound is not considered a hit due to its low confidence and therefore is not included in the main text.

[3] One hit was found by two teams

Table S3. SMILES ID of hit compounds

Hit information

ID	SMILES	Target
S-1	CC(C(=O)NC(Cc1c[nH]c2ccccc12)C(O)=O)c1ccc(Cl)cc1	S
S-2	CCc1cc(NC(=O)C2(CC2)c2cccc(F)c2)[nH]n1	S
S-3	COC(=O)C(Cc1c[nH]c2cc(F)ccc12)NC(=O)C(C)Oc1ccccc1	S
Nsp5-1	CC(N(C)C(=O)Cn1nnc2ccccc12)c1cccc(Cl)c1	Nsp5
Nsp5-2	O=C(Nc1nc2CCN(Cc3ccccc3)Cc2s1)c1ccc2C(=O)N3CCCCCC3=Nc2c1	Nsp5
Nsp5-3	Fc1ccc2c(CCNC(=O)c3cccc(Nc4ccc(cc4)C#N)c3)c[nH]c2c1	Nsp5
Nsp5-4	O=C(Nc1nc2cc3OCCOc3cc2s1)C1CCN(CC1)S(=O)(=O)c1ccc2CCCc2c1	Nsp5
Nsp5-5	O=C(Nc1nc2CN(Cc3ccccc3)CCc2s1)c1ccc2C(=O)N3CCCCCC3=Nc2c1	Nsp5
Nsp5-6	Cc1ccc(Nc2cccc(c2)C(=O)Nc2cc([nH]n2)C(=O)OCc2ccccc2)nn1	Nsp5
Nsp5-7	CC(C)N(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-8	CN(Cc1cccc(F)c1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-9	O=C(Cn1nnc2ccccc12)N1CCCCCC1c1ccccc1	Nsp5
Nsp5-10	CCN(C(c1ccccc1)c1ccccc1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-11	CC(C)CN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-12	CC(N(C1CC1)C(=O)Cn1nnc2ccccc12)c1ccccc1	Nsp5
Nsp5-13	CC(N(C)C(=O)Cn1nnc2ccccc12)c1ccccc1	Nsp5
Nsp5-14	CC(NC(=O)Cn1nnc2ccccc12)c1cccc(Cl)c1	Nsp5
Nsp5-15	COc1ccc(Cl)cc1CN(C)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-16	CCN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-17	CN(Cc1ccccc1)C(=O)Cn1nnc2ccccc12	Nsp5
Nsp5-18	CC(N(C1CC1)C(=O)Cn1nnc2ccccc12)c1cccc(c1)C(F)(F)F	Nsp5
Nsp5-19	CCC(N(CC)C(=O)Cn1nnc2ccccc12)c1ccccc1	Nsp5
Nsp3-1	COc1ccc(cc1OC)C(C)NC(=O)c1cc(N)ccc1C	Nsp3
Nsp3-2	Fc1ccc2[nH]c(CNC(=O)c3ccc(F)nc3)nc2c1	Nsp3
Nsp3-3	Fc1ccc(cn1)C(=O)NCc1nc2ccc(Cl)cc2[nH]1	Nsp3
Nsp12-1	O=C(CC1=NNC(=O)c2ccccc12)Nc1ccc(NS(=O)(=O)c2ccc3OCCOc3c2)cc1	Nsp12
Nsp12-2	Oc1nc(no1)-c1ccc(NC(=O)C2CCCS2)cc1	Nsp12

Further experimental tests and hit rate evaluation by individual teams. The SARStrooper team experimentally tested a few of the compounds that the team had ranked highly and found 7 inhibitors with IC50 <10 uM (Mukherjee et al., in preparation).

Pharm.ai compared their top 100 predictions for Nsp5 against public data published after the competition deadline and obtained a hit rate of 17% on a highly diverse set of scaffolds.[102]

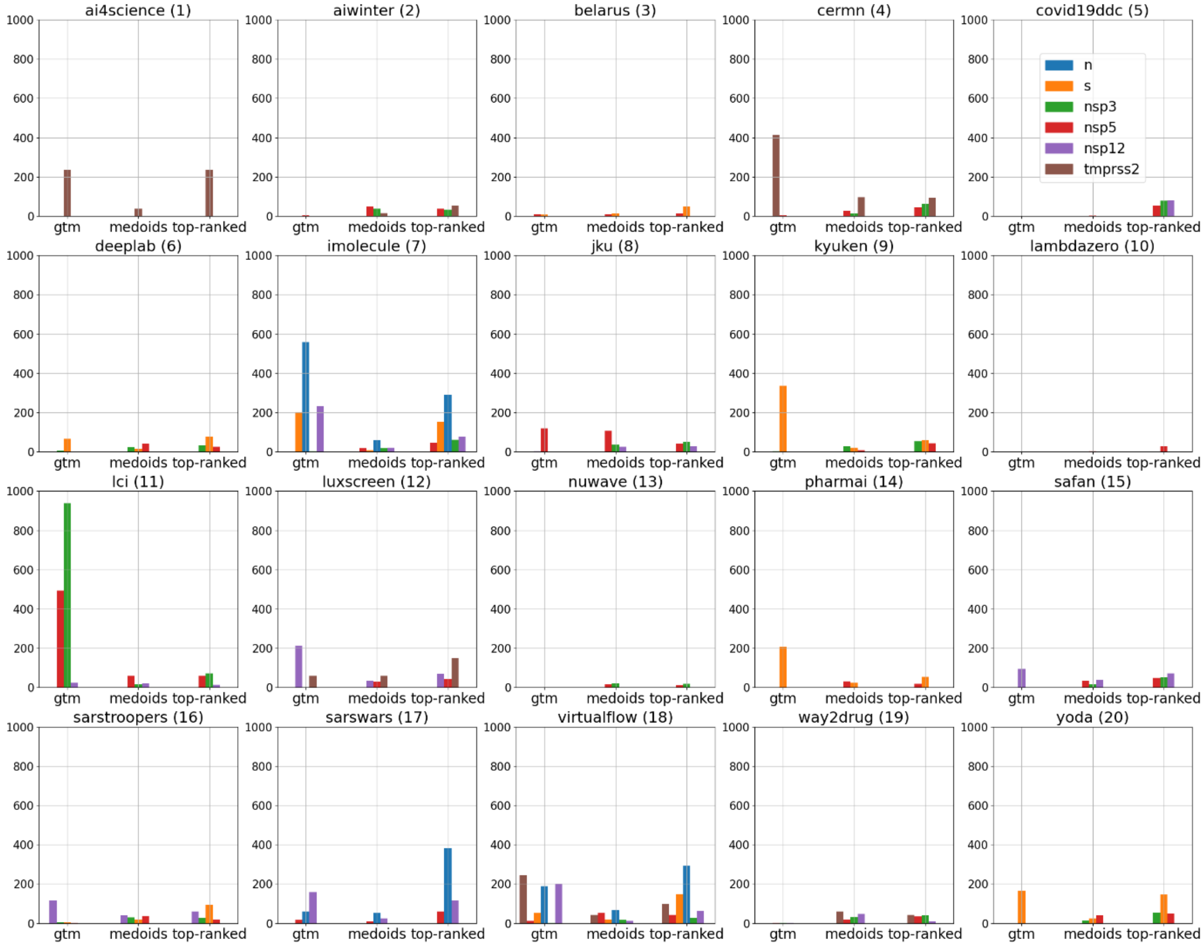


Figure S5a. Selected compounds list (~2000 compounds for each of the 6 targets) broken down by team, target, and selection method. See the main text 'Biases in compound selection and synthesis' paragraph (section 2.4 of the main text).

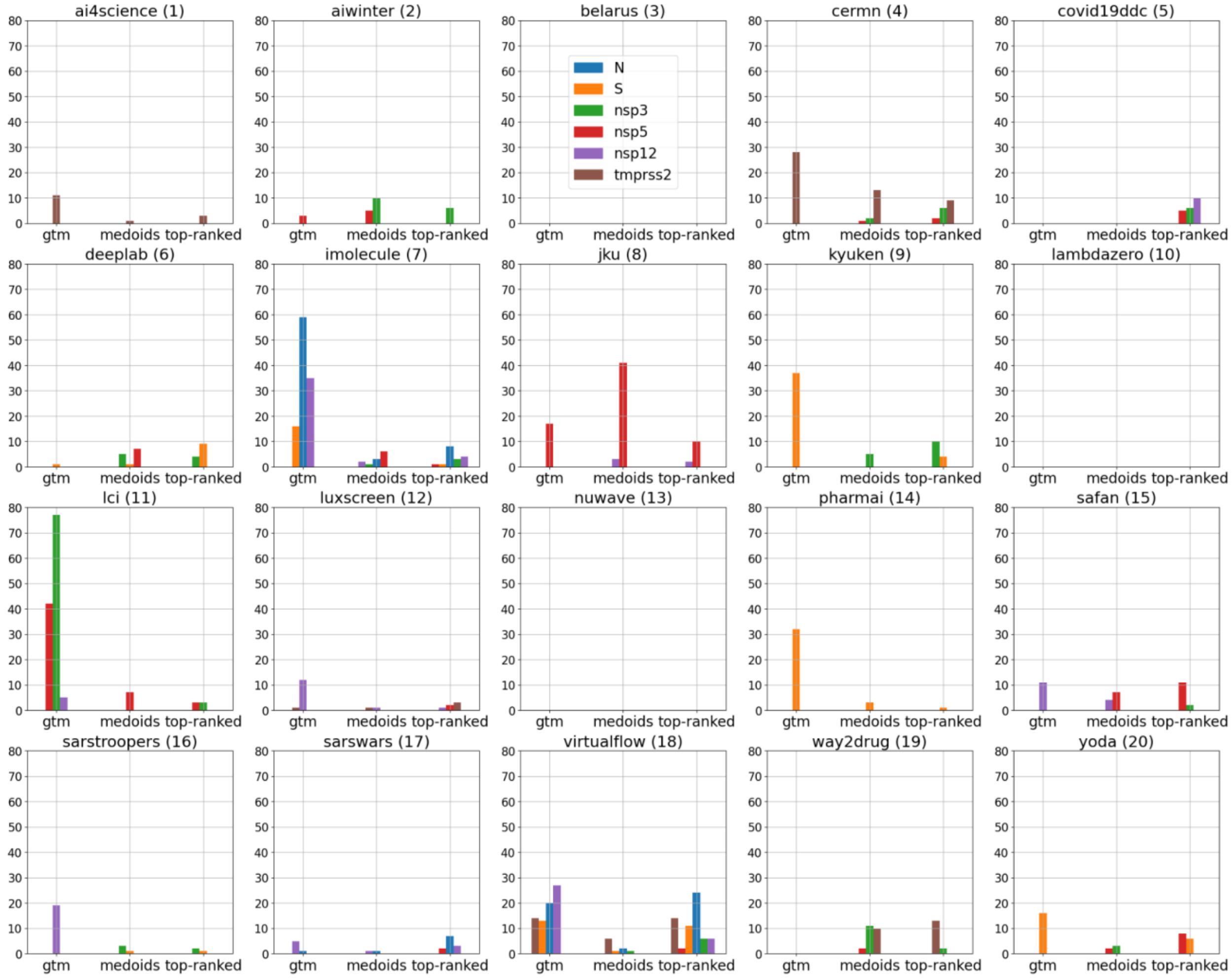


Figure S5b: Synthesized compounds list (878 compounds for all of the 6 targets combined) broken down by team, target, and selection method. See the main text ‘Biases in compound selection and synthesis’ paragraph (section 2.4 of the main text).

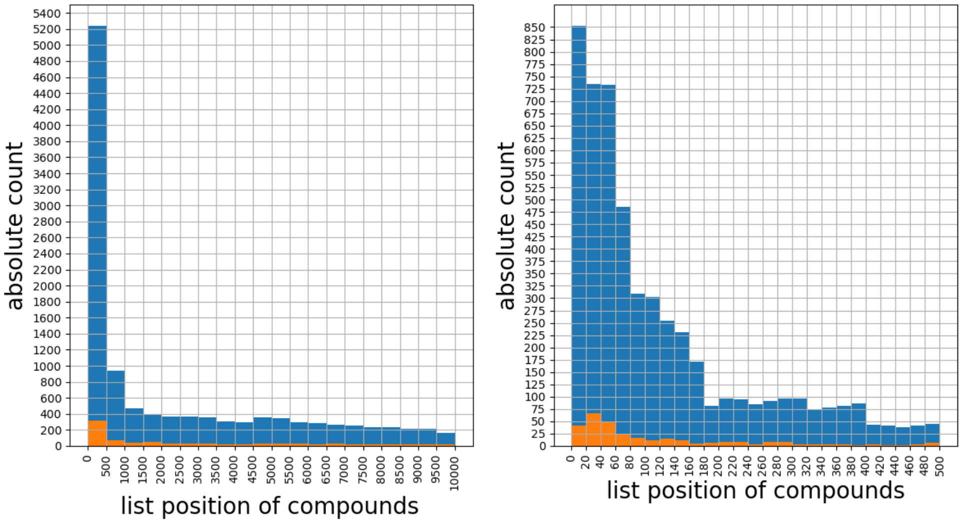


Figure S6. Histograms of selected compounds (blue) and synthesized compounds (orange) as a function of the ranking of each individual team that submitted to any of the 6 protein targets. Left: full view up to (the maximum allowed number of) 10000 compounds. Right: zoomed in section up to the top 500 ranked compounds.

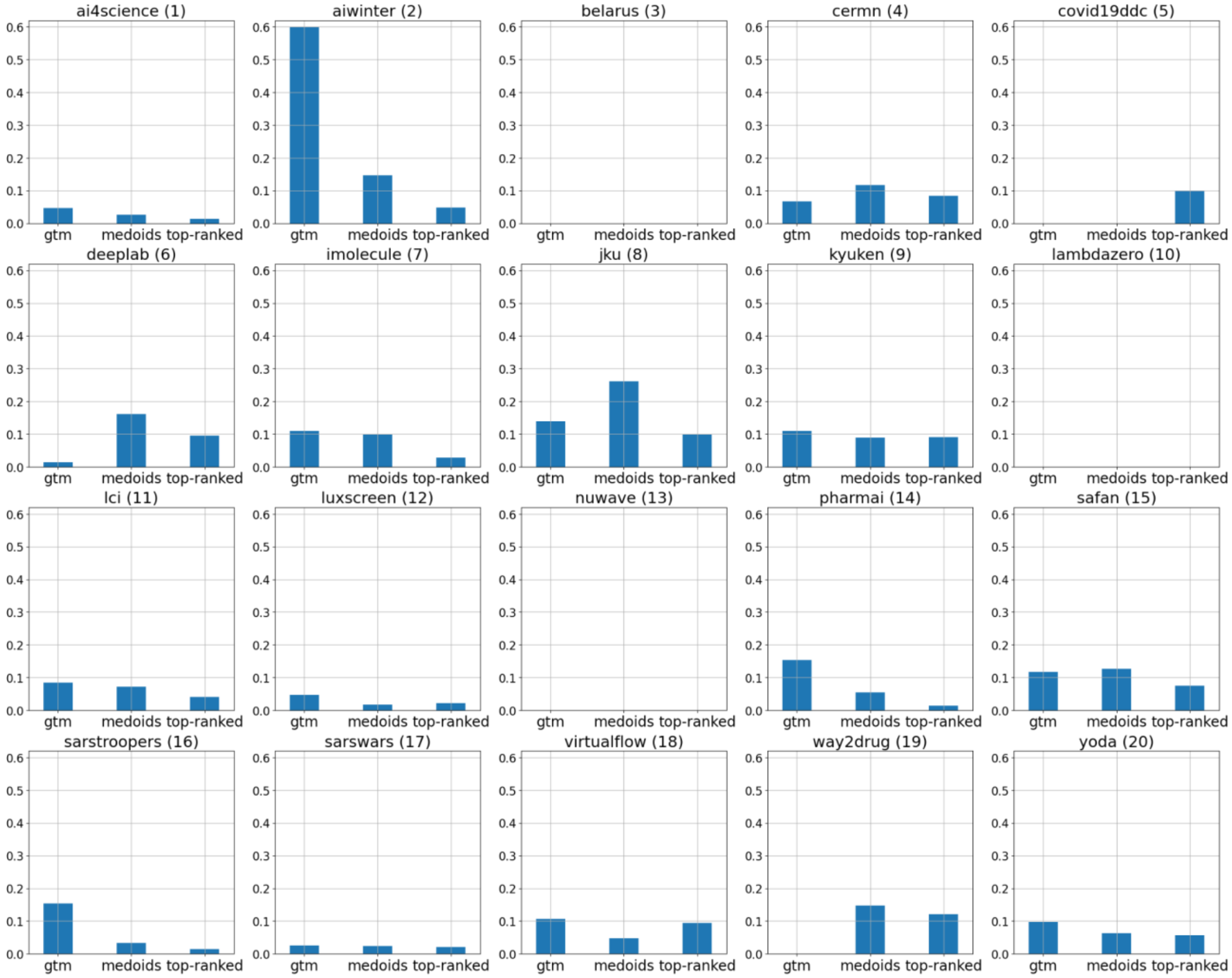


Figure S7: Ratio synthesized compounds list (878 total) versus selected compounds list (12082 compounds; ~2000 for each of the 6 targets).

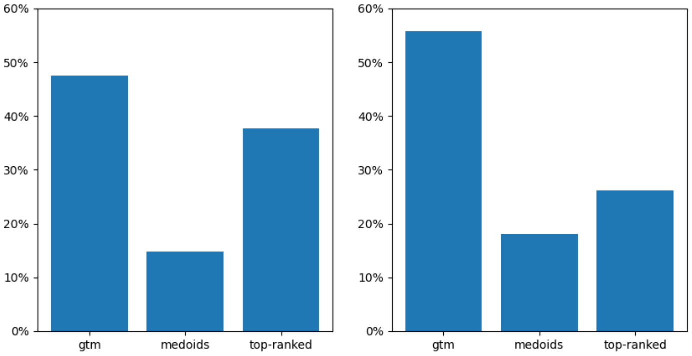


Figure S8. Overall percentages of origin by selection method. Left: selected molecules list (6 protein targets with ~2000 compounds each), right: synthesized compound list (878 molecules in total over 6 targets)

Details on t-SNE embeddings. All compounds displayed in Figure S9a and Figure S9b (and main text Figure 2a) were mapped to a vector with dim 1024, using folded ECFPs. Based on these features, a two-dimensional t-SNE embedding was created, whereby the manifold package by scikit-learn was used and the t-SNE algorithm was run with default hyperparameters.

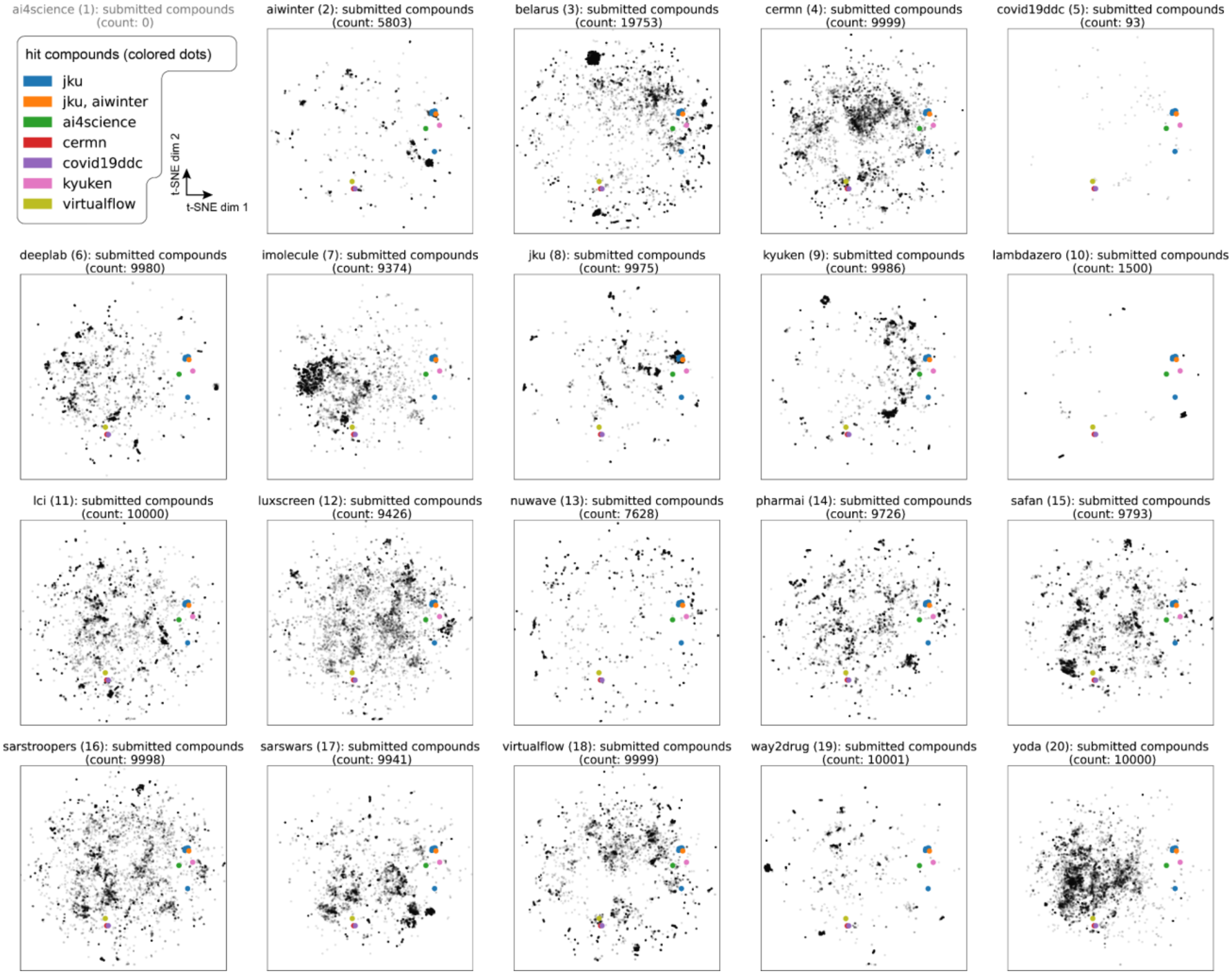


Figure S9a: For the Nsp5 target, the different hits (colored dots) are t-SNE embedded into the space of submitted compounds for the Nsp5 target. The black dots (in each figure panel) show single team submissions which contributed to the Nsp5 list. The directions of t-SNE dimension 1 and 2 are indicated by arrows (see top left, next to the legend).

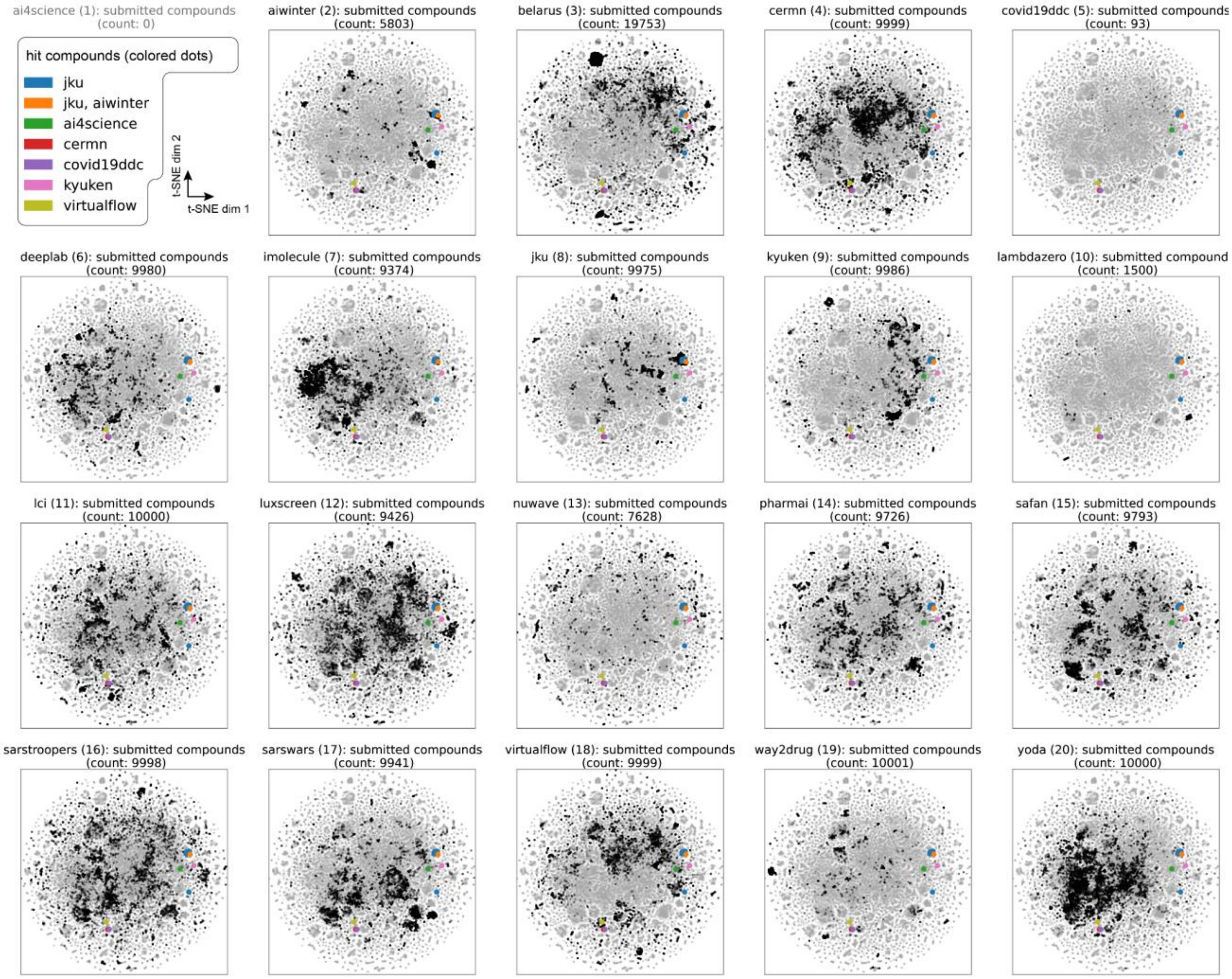


Figure S9b: The same plot as Figure S9a, but in gray dots all team submissions to the Nsp5 are shown in the background.

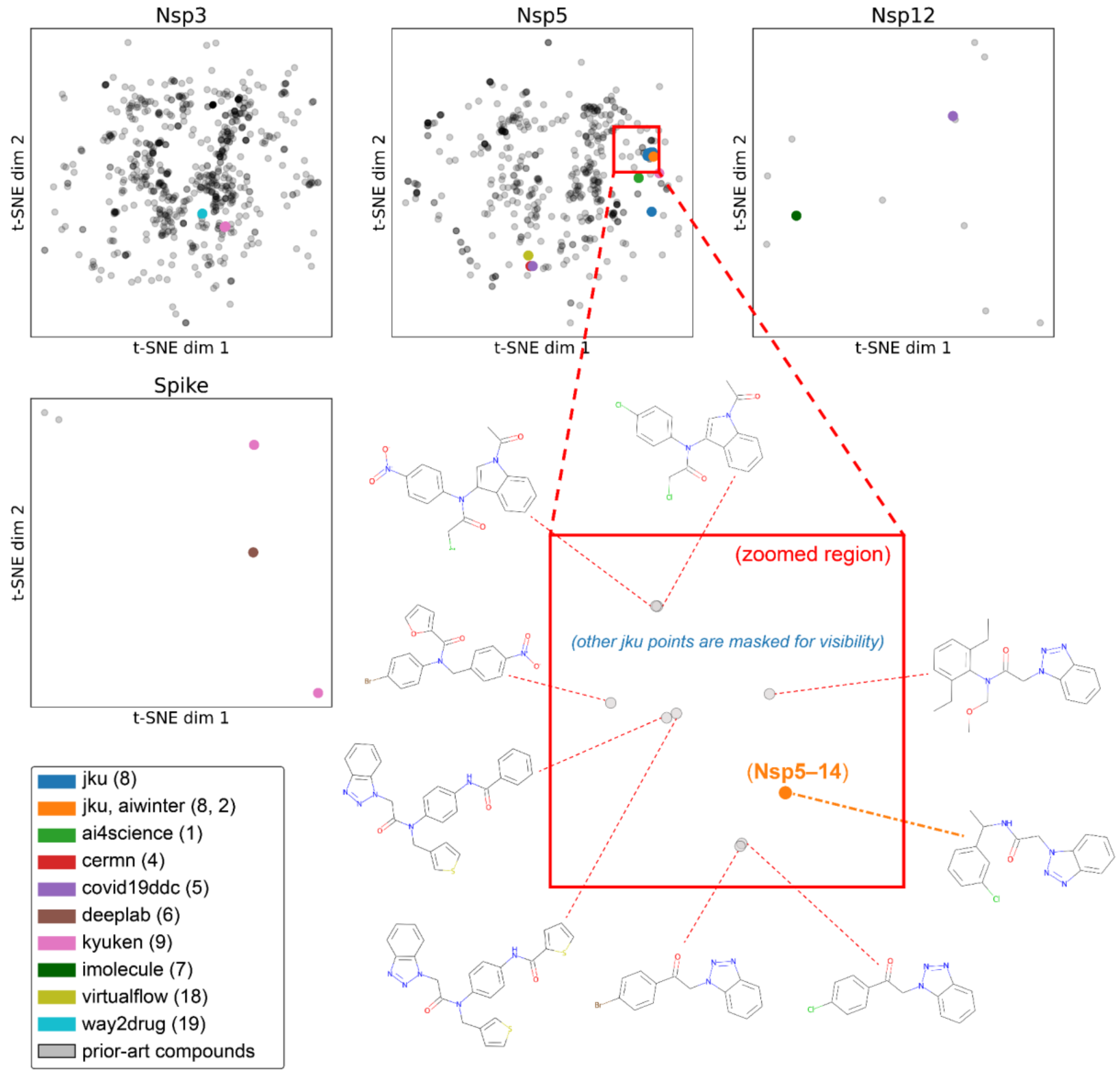


Figure S10: Scatterplots in t-SNE coordinates which show the hits and the prior-art compounds (gray/black circles). Colored dots are hits found in the current work. For Nsp5 a zoomed-in region is shown in the red box, with the chemical structures of proximal prior-art compounds, where several benzotriazolyl acetamides can be identified.

Section 4: Details on biological assays and testing procedures

4.1 Compound feasibility, synthesis and purity.

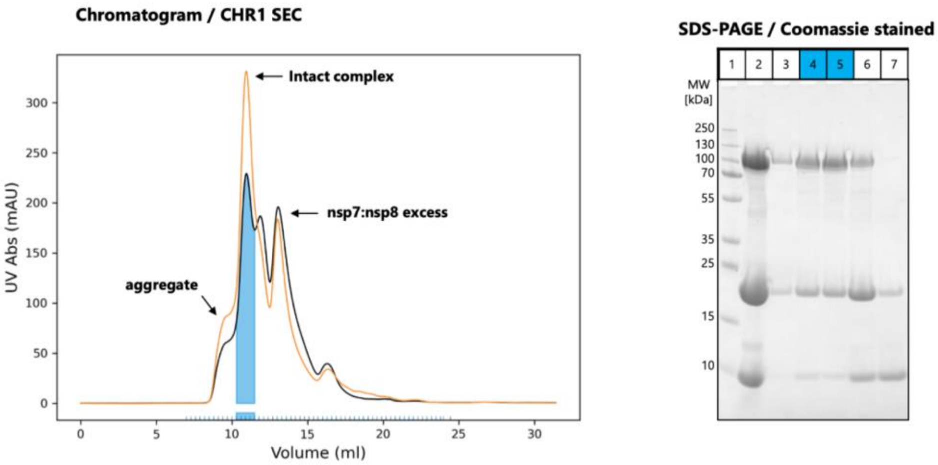
All compounds were synthesized by WuXi Apptec (China) upon instructions from the organizing team. The final compounds were selected based on 3 criteria by WuXi Apptec using proprietary methods: 1) cADME (computational absorption, distribution, metabolism, and excretion) filtering was done to arrive at compounds with molecular weight (MW) below 500 g mol^{-1}, CLogP < 5, Hydrogen bond acceptors HBA < 10, Hydrogen bond donors HBD < 5, TPSA < 140, Number of rotatable bond < 5. In addition, possible PAINS (Pan-assay interference compounds) were removed; 2) Chemical feasibility: a similarity search versus the WuXi Apptec virtual library was performed to assess feasibility; 3) reagent availability and cost was considered (approx. 245 000 $ in compound costs alone). The organizing team supplied the top-ranked ~2000 compounds from the 'selected compounds list' of each of the 6 protein targets to WuXi Apptec (i.e., 11440 total compounds), which then used the 3 criteria to select 1414 feasible compounds in total. The synthesis period lasted from November 2020 to February 2021, and 878 compounds were delivered as 20 mM DMSO (dimethylsulfoxide) stock solution on well-plates. The compound purity was determined by LC-MS and has been reported previously.[103] Of all 878 compound, 58 (i.e., 6.6%) had a purity below 90%. The latter data set also includes information on solubility, and on compound chirality. Duplicate compound plates with DMSO stock solutions were shipped to the MIT-Broad institute (USA), Crelux GmbH (Germany), Pasteur Institute (France), and the Diamond Light Source (UK), for further experiments. For purity of the compounds in the main text see the caption of Table 3 for details.

4.2 Binding assays using microscale thermophoresis.

Binding assays were performed by Crelux GmbH.

4.2.1 Binding assays Nsp12 – Materials and Methods

Attempts to stabilize the Nsp7/8/12 complex using a literature procedure[104] were only partially successful (see Figure S11). The complex could be obtained, but was found to destabilize over time and with (gentle) heating. This is sufficient for structure determination (in ref.[104]), but not for MST measurements. Therefore only the Nsp12 domain was used instead.



Lane	Sample
1	Marker: PageRuler Plus
2	Input
3	Aggregate
4	Complex #1
5	Complex #2
6	Shoulder
7	nsp7:nsp8 excess

Figure S11. Purification and stabilization of Nsp7/8/12 complex.

Nsp12 containing a C-terminal His-Strep tag (lot PC13929-1, aa 4393-5324) was purchased from Crelux GmbH - a WuXi Apptec company. Suramin was purchased from Sigma Aldrich (574625). Nsp12 was labeled in an assay buffer using 25 nM protein and 12.5 nM dye, following the labeling protocol as specified in NanoTemper Technologies RED-NHS 2nd Generation labeling kit. The assay buffer for all experiments was 20 mM Hepes pH 7.5, 150 mM NaCl, 1 mM MgCl2, 2.5 mM DTT, 0.005% TWEEN® 20, and a final volume of 20 µL per datapoint was used for both the 8-pt-screen and the affinity screen (12 points) in Dianthus 384-well microwell plates.

The compound library and Suramin were dissolved in 100% DMSO at a concentration of 20 mM. For the 8-pt-screen, we diluted the compounds from 20 mM stock solutions to 4 mM and Suramin to 0.8 mM in 100% DMSO and transferred the first dilution into a conventional microwell plate. The second and all following 3-fold dilution steps were prepared in DMSO in the microwell plate, using at each step 5 µl of the preceding dilution and 10 µl of DMSO. Then, 19.5 µl of labeled nsp12 at a final concentration of 25 nM was transferred into the Dianthus 384-well microplate, and 0.5 µl of compound dilution series was added and mixed thoroughly. Final concentrations were 2.5% DMSO and 100 µM – 45.7 nM compound, 3-fold dilution series. Final Suramin concentrations were 20 µM – 9.14 nM, 3-fold dilution series. For the 12-pt affinity screen, concentrations were adapted to obtain final compound concentrations of 250 µM – 1.41 nM and Suramin concentrations of 100 µM – 0.56 nM, 3-fold dilution series for both.

After the Dianthus 384-well microplates were loaded with the compounds + nsp12 mix, they were equilibrated for 30 min at RT and centrifuged for 30 sec at 400 x g before loading into the Dianthus NT.23PicoDuo. The system was set to 25°C as set temperature. The samples were first measured for 1 sec without heating and for 5 sec with the IR-laser turned on. The two optical systems in Dianthus were used in parallel. Measured fluorescence values collected are displayed as relative fluorescence, where the fluorescence obtained at ambient temperature is normalized to one, and as normalized fluorescence (F_{norm}) which describes the ratio between fluorescence values (F_1) after and the fluorescence values (F_0) prior to IR laser activation and is typically given in ‰. The dissociation constant or K_D, is obtained by fitting a dose-response curve to a plot of Fnorm vs. ligand concentration.

References for Dianthus:

- APPLICATION NOTE “Fast molecular interaction screening of epigenetic gene regulator G9a with fragments from a large chemical space”, ©2019 NanoTemper Technologies, Inc. South San Francisco, CA, USA. All Rights Reserved.

4.2.2 Binding assays Nucleocapsid – Materials and Methods

His-tagged SARS-CoV-2 (COVID-19) Nucleocapsid protein was purchased from Acro Biosystems (NUN-C51H9). Nanobodies against the N- and C-terminal domain of Nucleocapsid were provided by Pasteur institute (by Dr. Pierre Lafaye).

Nucleocapsid was labeled in assay buffer at a concentration of 100 nM using 25 nM dye, following the labeling protocol as specified in NanoTemper Technologies RED-NHS 2nd Generation labeling kit. The assay buffer for all experiments was 20 mM Hepes pH 7.5, 150 mM NaCl, 2 mM DTT, 0.05% TWEEN® 20, 0.1% PEG-8000.

The compound library was dissolved in 100% DMSO at a concentration of 20 mM. For the 8-pt-screen, compounds were diluted to 4 mM in 100% DMSO and transferred into well 1 of a conventional 384-well plate. The second and all following 3-fold dilution steps were prepared in DMSO in the microwell plate, using at each step 5 µl of the preceding dilution and 10 µl of DMSO. Then, labeled Nucleocapsid at a final concentration of 50 nM was transferred into another 384-well microplate, and appropriate amounts of compound dilution series were added and mixed thoroughly. Final concentrations were 2.5% DMSO and 100 µM – 45.7 nM compound, 3-fold dilution series. For the 12-pt affinity screen, concentrations were adapted to obtain final compound concentrations of either 200 µM – 1.13 nM or 40 µM – 0.23 nM, 3-fold dilution series for both.

As a positive control, the nanobody against NTD was titrated to Nucleocapsid from 500 nM – 0. 23 nM, 3-fold dilution series (8 concentrations) or 500 nM – 0.24 nM, 2-fold dilution series (12 concentrations). Final DMSO concentrations in the Nucleocapsid-nanobody mixtures were 2.5%.

Compounds (or nanobody) + Nucleocapsid mixtures were equilibrated for 30 min at RT before filling of samples into MonolithTM NT.115 Series MST Premium Coated Capillaries and loading into a MonolithTM NT.115 instrument.[105] Measurements were performed at 25°C. The samples were first measured for 3 sec without heating, then for 20 sec with the IR-laser turned on at high MST power, followed by 1 sec without heating. Measured fluorescence values collected are displayed as relative fluorescence, where the fluorescence obtained at ambient temperature is normalized to one. For analysis, the change in thermophoresis is expressed as the change in the normalized fluorescence (ΔF_{norm}), which is defined as the ratio between fluorescence values (F_{hot}) after and the fluorescence values (F_{cold}) prior to IR laser activation and is typically given in ‰. The dissociation constant or K_D, is obtained by fitting a dose-response curve to a plot of ΔF_{norm} vs. ligand concentration.

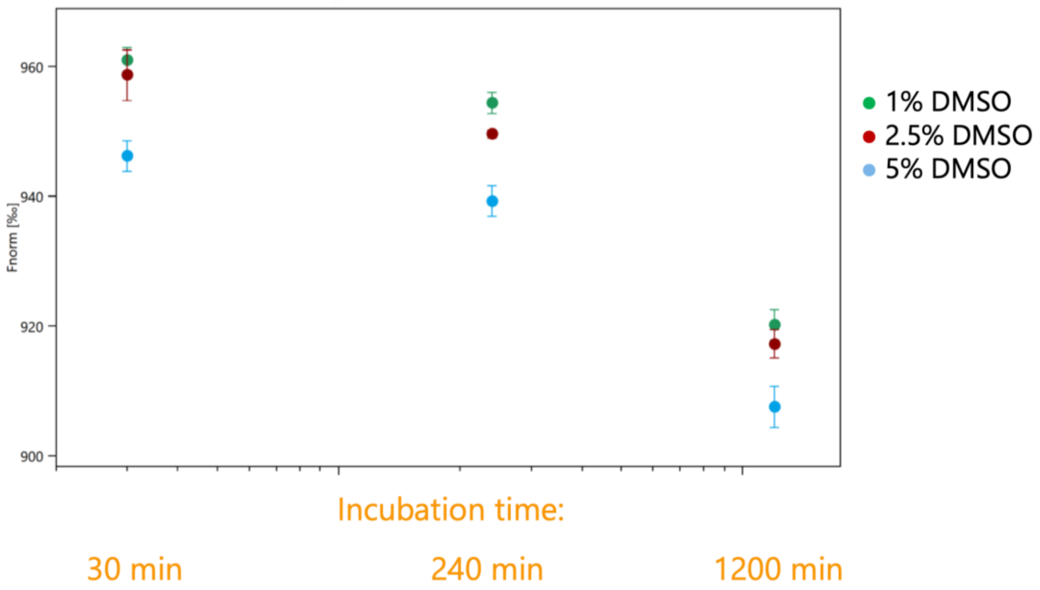


Figure S12. A gradual decrease in F_{norm} over time during N assays upon addition of 1–5% DMSO made K_D determination impossible.

4.2.3 Binding assays Spike – Materials and Methods

His-tagged SARS-CoV-2 (COVID-19) S protein (R683A, R685A, K986P, V987P) active trimer was purchased from Acro Biosystems (SPN-C52H2). Human Fc-tagged ACE2 was purchased from Acro Biosystems (AC2-H5257).

Spike and ACE2 were dialyzed into assay buffer using XPRESS MicroDialyzer tubes (6-8 kDa MWCO). Spike was labeled in assay buffer at a concentration of 100 nM using 50 nM dye, following the labeling protocol as specified in NanoTemper Technologies RED-NHS 2nd Generation labeling kit. The assay buffer for all experiments was 20 mM Hepes pH 7.5, 150 mM NaCl, 0.05% TWEEN® 20, 0.1% PEG-8000.

The compound library was dissolved in 100% DMSO at a concentration of 20 mM. For the 8-pt-screen, compounds were diluted to 4 mM in 100% DMSO and transferred into well 1 of a conventional 384-well plate. The second and all following 3-fold dilution steps were prepared in DMSO in the microwell plate, using at each step 5 µl of the preceding dilution and 10 µl of DMSO. Then, labeled Spike at a final concentration of 50 nM was transferred into another 384-well microplate, and appropriate amounts of compound dilution series were added and mixed thoroughly. Final concentrations were 2.5% DMSO and 100 µM – 45.7 nM compound, 3-fold dilution series. For the 12-pt affinity screen, concentrations were adapted to obtain final compound concentrations of either 200 µM – 1.13 nM or 40 µM – 0.23 nM, 3-fold dilution series for both. As a positive control, ACE2 was titrated to Spike using maximum possible concentrations after dialysis, typically 350 nM – 0.16 nM, 3-fold dilution series (8 concentrations) or 350 nM to 0.17 nM, 2-fold dilution series (12 concentrations). Final DMSO concentrations in the Spike-ACE2 mixtures were 2.5%.

Compounds (or ACE2) + Spike mixtures were equilibrated for 30 min at RT before filling of samples into MonolithTM NT.115 Series MST Premium Coated Capillaries and loading into a MonolithTM NT.115 instrument. Measurements were performed at 25°C. The samples were first measured for 3 sec without heating, then for 20 sec with the IR-laser turned on at medium MST power, followed by 1 sec without heating. Measured fluorescence values collected are displayed as relative fluorescence, where the fluorescence obtained at ambient temperature is normalized to one. For analysis, the change in thermophoresis is expressed as the change in the normalized fluorescence (ΔF_{norm}), which is defined as the ratio between fluorescence values (F_{hot}) after and the fluorescence values (F_{cold}) prior to IR laser activation and is typically given in ‰. The dissociation constant or K_D, is obtained by fitting a dose-response curve to a plot of ΔF_{norm} vs. ligand concentration.

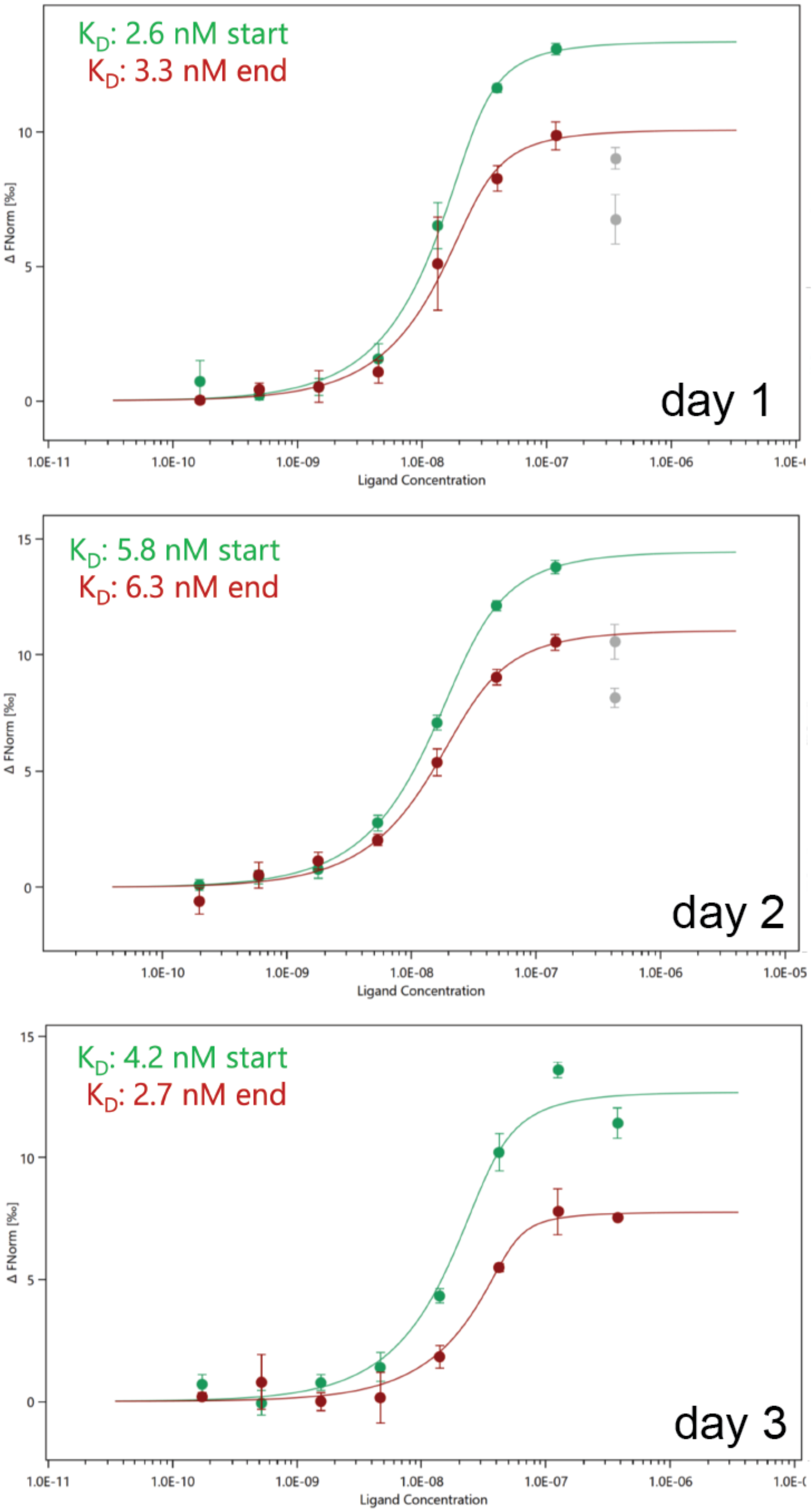


Figure S13. Control experiments of Fc-tagged Ace2 binding to (stable trimer) S using microscale thermophoresis. On the three days that compounds were measured, control runs were done before (“start”) and after (“end”).

4.3 Other assays

4.3.1 Nsp5 Luciferase assays.

The viral protease activity assay was performed as described in Antiviral Research 2022, PMID: 35278581. Briefly, ~10^7 HEK-293T cells were transfected in 50 cm2 dishes with 1 µg of the Rev-Nluc-CoV reporter plasmid and 15 µg of nsp4-5-6 expression plasmids encoding the wild-type nsp5 or catalytically inactive (Cys145Ala) nsp5 protein. Increasing amounts of the tested compounds, corresponding to final concentrations of 0.1 to 50 µM with a constant DMSO concentration of 0.5% were distributed in 384-well white opaque plates using an Echo 555 Liquid Handler (Labcyte). Transfected HEK-293T cells were trypsinized at 6 hours post-transfection (hpt) and distributed in 384-well plates ($2x10^4$ cells per well in 50 µL). The Nanoluciferase activity was measured at 24 hpt. DMSO and GC376 were used as negative and positive controls. The data were expressed as % of restored Nanoluc activity. Curve fitting was performed with the assumption that 0% is equal to the lowest value measured with the wild-type nsp5, and 100% is the value measured with the catalytically inactive control. For some compounds, the 50 µM concentration appeared to be toxic, and therefore it was excluded from the curve fitting.

4.3.2 Enzymatic protease (cleavage) activity

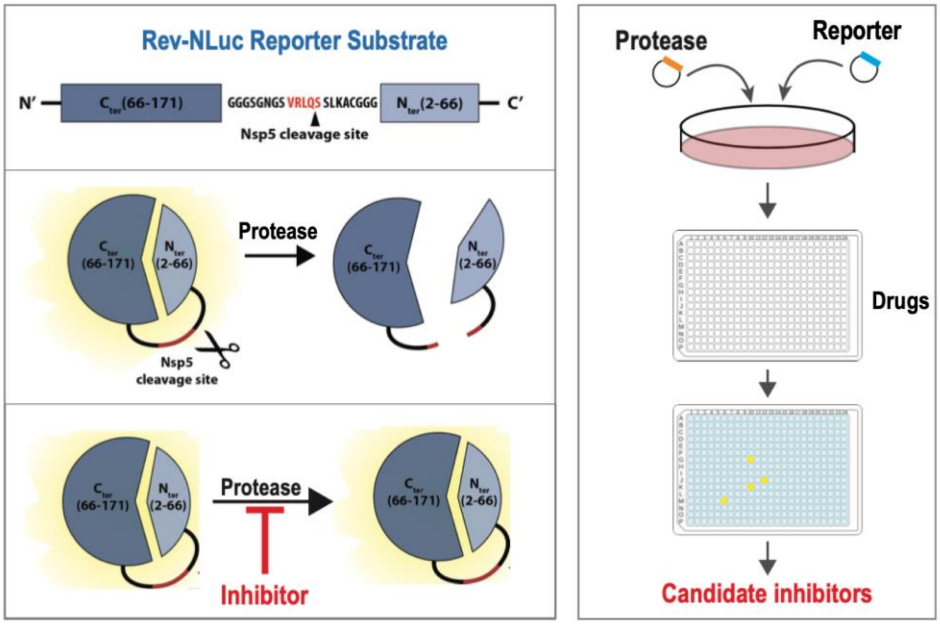


Figure S14 schematic representation of the cell-based assay for nsp5 activity. Cleavage of the reporter Rev-Nluc protein by the nsp5 protease decreases the luminescence signal. In the presence of an inhibitor of the Nsp5 protease, the luminescence signal is restored. The assay is performed in 384-well plates and automated. Cells that co-express the nsp5 protease and the Rev-Nluc reporter are seeded into wells that contain the compounds to be tested. The luminescence signal is measured 24 hours later. In parts reproduced from ref.[106]

Nsp5: The gene for SARS-CoV-2 M^{pro} (orf1ab polyprotein residues 3264-3569, GenBank: MN908947) was codon optimized for Escherichia coli and subsequently cloned into a pET28a backbone containing an N-terminal 6x His tag with an enterokinase cleavage site. In order to evaluate the assay initially and then as positive control in every experiment 111 nM of cleaved and purified M^{pro} was incubated with GC376 at various concentrations ranging 100 pM to 100 µM in buffer containing 20 mM Tris pH = 7.3, 100 mM NaCl, 1 mM EDTA, 1% DMSO. Cleavage of fluorescent peptide substrate Dabcyl-KTSAVLQSGFRKME-Edans at 26°C was monitored by fluorescence increase ($\lambda_{excitation}$ = 360 nm, $\lambda_{emission}$ = 460 nm) for one hour. Initial velocities (V_o) were calculated and normalized to the V_o at lowest inhibition. IC50 value (mean ± 1 SD, N = 3) was determined via nonlinear regression using Graphpad Prism 8 (see figure X a). The same protocol was used for each test sample.

Nsp3: The gene for SARS-CoV-2 PL^{pro} (orf1ab polyprotein residues 1564-1880, GenBank: MN908947) was codon optimized for E. coli and inserted into a pET28a backbone containing an N-terminal 6x His tag with an enterokinase cleavage site. In order to evaluate the assay initially and then as positive control in every experiment 12 nM of purified His-PL^{pro} was incubated with GRL0617 at various concentrations ranging 100 pM to 100 µM in buffer containing 50 mM Tris pH = 8, 50 mM NaCl, 1% DMSO. Cleavage of peptide substrate Z-RLRGG-AMC at 37°C was monitored by fluorescence increase ($\lambda_{excitation}$ = 360 nm, $\lambda_{emission}$ = 460 nm) for one hour. V_o were calculated and normalized to the V_o at lowest inhibition. IC50 value (mean ± 1 SD, N = 3) was determined via nonlinear regression using Graphpad Prism 8 (see figure X b). The same protocol was used for each test sample.

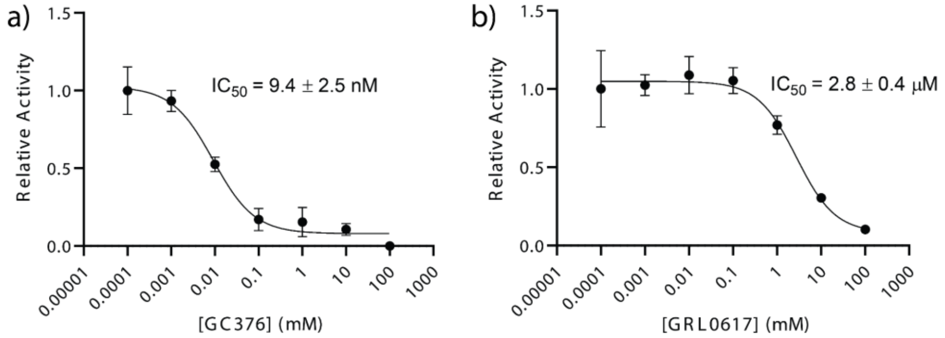


Figure S15. Validation of protease inhibition assays for SARS-CoV-2 Main Protease (Nsp5 / M^{pro}) and SARS-CoV-2 Papain-like Protease (Nsp3 / PL^{pro}) via replication of IC50 values of known inhibitors. a. Calculation of IC50 value of SARS-CoV-2 M^{pro} inhibitor GC376: b. Calculation of IC50 value of SARS-CoV-2 PL^{pro} inhibitor GRL0617:

4.3.3 Viral reduction assays

Two thousand Vero-TMPRSS2 or HeLa-ACE2 cells (BPS Bioscience) were seeded into 96-well plates in DMEM (10% FBS) and incubated for 24 hours at 37°C, 5% CO_2. Two hours before infection, the medium was replaced with 100 µL of DMEM (2% FBS) containing the compound of interest at concentrations 50% greater than those indicated, including a DMSO control. Plates were then transferred into the BSL3 facility and 100 PFU (MOI = 0.025) was added in 50 µL of DMEM (2% FBS), bringing the final compound concentration to those indicated. Plates were then incubated for 48 hours at 37°C. After infection, supernatants were removed and cells were fixed with 4% formaldehyde for 24 hours prior to being removed from the BSL3 facility. The cells were then immunostained for the viral N protein (an inhouse mAb 1C7, provided by Dr. Thomas Moran, Thomas.Moran@mssm.edu) with a DAPI counterstain. Infected cells (488 nm) and total cells (DAPI) were quantified using the Celigo (Nexcelcom) imaging cytometer. Infectivity was measured by the accumulation of viral N protein (fluorescence accumulation). Percent infection was quantified as ((Infected cells/Total cells) - Background) *100 and the DMSO control was then set to 100% infection for analysis. Data was fit using nonlinear regression and IC50s for each experiment were determined using GraphPad Prism version 8.0.0 (San Diego, CA). Cytotoxicity was also performed using the MTT assay (Roche), according to the manufacturer's instructions. Cytotoxicity was performed in uninfected cells with same compound dilutions and concurrent with viral replication assay. All assays were performed in biologically independent triplicates.

4.3.4 Crystallographic screening

Mpro protein expression, purification and crystallization was carried out as previously described.[107] Crystals were grown with a reservoir solution containing 100 mM MES pH 6.5, 15% PEG4K, 5% DMSO and drop ratios of 150 nl protein solution, 300 nl reservoir solution and 50 nl seed stock.

Nsp3-macrodomain protein expression, purification and crystallization was carried out as previously described.[108] Crystals were grown with a reservoir solution containing 100 mM CHES (pH 9.5) and 30% PEG-3000 and drop ratios of 150 nl reservoir solution plus 150 nl protein solution.

20 mM stock solutions of compounds in DMSO were directly added to the crystallization drops giving a final compound concentration of 2 mM and DMSO concentration of 10%. The crystals were incubated in the presence of the compounds for 1–2 hours before being harvested and flash cooled in liquid nitrogen without the addition of further cryoprotectant. X-ray diffraction data were collected on beamline I04-1 at Diamond Light Source and automatically processed using the Diamond automated processing pipelines.[109] Analysis was performed as outlined previously.[107] Briefly, electron density maps were generated with Dimple[110], ligand-binding events were identified using PanDDA[111], and ligands were modeled into PanDDA-calculated event maps using Coot[112]. Ligand restraints were calculated with GRADE[113] structures were refined with Refmac[114] and Buster[115] and models and quality annotations cross-reviewed.

Supporting references

(1) Rensi, S.; Keys, A.; Lo, Y.-C.; Derry, A.; McInnes, G.; Liu, T.; Altman, R. Homology Modeling of TMPRSS2 Yields Candidate Drugs That May Inhibit Entry of SARS-CoV-2 into

Human Cells. *ChemRxiv* **2020**.

(2) Raval, A.; Piana, S.; Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Refinement of Protein Structure Homology Models via Long, All-Atom Molecular Dynamics Simulations. *Proteins Struct. Funct. Bioinforma.* **2012**, *80* (8), 2071–2079. https://doi.org/10.1002/prot.24098.

(3) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; others. 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.

(4) Hoffmann, M.; Schroeder, S.; Kleine-Weber, H.; Müller, M. A.; Drosten, C.; Pöhlmann, S. Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19. *Antimicrob. Agents Chemother.* **2020**, *64* (6), e00754-20.

(5) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; others. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13* (7), e1005659.

(6) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Jr. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain X1 and X2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8* (9), 3257–3273. https://doi.org/10.1021/ct300400x.

(7) Noé, F.; Wu, H.; Prinz, J.-H.; Plattner, N. Projected and Hidden Markov Models for Calculating Kinetics and Metastable States of Complex Molecules. *J. Chem. Phys.* **2013**, *139* (18), 184114. https://doi.org/10.1063/1.4828816.

(8) Huggins, D. Structural Analysis of Experimental Drugs Binding to the COVID-19 Target TMPRSS2. ChemRxiv May 18, 2020. https://doi.org/10.26434/chemrxiv.12315449.v1.

(9) Bestle, D.; Heindl, M. R.; Limburg, H.; Van, T. V. L.; Pilgram, O.; Moulton, H.; Stein, D. A.; Hardes, K.; Eickmann, M.; Dolnik, O.; Rohde, C.; Becker, S.; Klenk, H.-D.; Garten, W.; Steinmetzer, T.; Böttcher-Friebertshäuser, E. 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. bioRxiv April 15, 2020, p 2020.04.15.042085. https://doi.org/10.1101/2020.04.15.042085.

(10) Nimishakavi, S.; Raymond, W. W.; Gruenert, D. C.; Caughey, G. H. Divergent Inhibitor Susceptibility among Airway Lumen-Accessible Tryptic Proteases. *PLOS ONE* **2015**, *10* (10), e0141169. https://doi.org/10.1371/journal.pone.0141169.

(11) Katz, B. A.; Clark, J. M.; Finer-Moore, J. S.; Jenkins, T. E.; Johnson, C. R.; Ross, M. J.; Luong, C.; Moore, W. R.; Stroud, R. M. Design of Potent Selective Zinc-Mediated Serine Protease Inhibitors. *Nature* **1998**, *391* (6667), 608–612. https://doi.org/10.1038/35422.

(12) Schrödinger Release 2020-2: LigPrep; Schrödinger, LLC: New York, NY, 2020.

(13) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.

(14) Koes, D. R.; Baumgartner, M. P.; Camacho, C. J. 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.

(15) 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.* **2010**, *31* (2), 455–461. https://doi.org/10.1002/jcc.21334.

(16) Quiroga, R.; Villarreal, M. A. 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.

(17) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. 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.

(18) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296–W303. https://doi.org/10.1093/nar/gky427.

(19) Case, D.; Ben-Shalom, I.; Brozell, S.; Cerutti, D.; Cheatham III, T.; Cruzeiro, V.; Darden, T.; Duke, R.; Ghoreishi, D.; Gilson, M.; others. AMBER 2018; 2018. *Univ. Calif. San Franc.* **2018**.

(20) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved Side-Chain Torsion Potentials for the Amber Ff99SB Protein Force Field. *Proteins Struct. Funct. Bioinforma.* **2010**, *78* (8), 1950–1958. https://doi.org/10.1002/prot.22711.

(21) 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.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(22) 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.* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(23) Gusev, F.; Gutkin, E.; Kurnikova, M. G.; Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *J. Chem. Inf. Model.* **2023**, *63* (2), 583–594. https://doi.org/10.1021/acs.jcim.2c01052.

(24) Cichońska, A.; Ravikumar, B.; Allaway, R. J.; Wan, F.; Park, S.; Isayev, O.; Li, S.; Mason, M.; Lamb, A.; Tanoli, Z.; Jeon, M.; Kim, S.; Popova, M.; Capuzzi, S.; Zeng, J.; Dang, K.; Koytiger, G.; Kang, J.; Wells, C. I.; Willson, T. M.; Oprea, T. I.; Schlessinger, A.; Drewry, D. H.; Stolovitzky, G.; Wennerberg, K.; Guinney, J.; Aittokallio, T. Crowdsourced Mapping of Unexplored Target Space of Kinase Inhibitors. *Nat. Commun.* **2021**, *12* (1), 3307. https://doi.org/10.1038/s41467-021-23165-1.

(25) 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.* **2010**, *50* (7), 1189–1204. https://doi.org/10.1021/ci100176x.

(26) Marvin Version 21.17.0, ChemAxon. https://www.chemaxon.com.

(27) McGann, M. FRED and HYBRID Docking Performance on Standardized Datasets. *J. Comput. Aided Mol. Des.* **2012**, *26* (8), 897–906. https://doi.org/10.1007/s10822-012-9584-8.

(28) 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.* **2019**, *5* (8), eaav6490. https://doi.org/10.1126/sciadv.aav6490.

(29) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, *6* (6), 939–949. https://doi.org/10.1021/acscentsci.0c00229.

(30) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *The Protein Data Bank*. Nucleic Acids Research. https://www.rcsb.org.

(31) Sunseri, J.; Koes, D. R. Pharmit: Interactive Exploration of Chemical Space. *Nucleic Acids Res.* **2016**, *44* (W1), W442–W448. https://doi.org/10.1093/nar/gkw287.

(32) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A. I.; Skiff, W. M. UFF, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114* (25), 10024–10035. https://doi.org/10.1021/ja00051a040.

(33) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R.

Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3*, 33. https://doi.org/10.1186/1758-2946-3-33.

(34) Alhossary, A.; Handoko, S. D.; Mu, Y.; Kwoh, C.-K. Fast, Accurate, and Reliable Molecular Docking with QuickVina 2. *Bioinformatics* **2015**, *31* (13), 2214–2216. https://doi.org/10.1093/bioinformatics/btv082.

(35) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(36) Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. *ArXiv Prepr. ArXiv151105644* **2015**.

(37) Landrum, G. RDKit: Open-Source Cheminformatics Software.

(38) Martín Abadi; Ashish Agarwal; Paul Barham; Eugene Brevdo; Zhifeng Chen; Craig Citro; Greg S. Corrado; Andy Davis; Jeffrey Dean; Matthieu Devin; Sanjay Ghemawat; Ian Goodfellow; Andrew Harp; Geoffrey Irving; Michael Isard; Jia, Y.; Rafal Jozefowicz; Lukasz Kaiser; Manjunath Kudlur; Josh Levenberg; Dandelion Mané; Rajat Monga; Sherry Moore; Derek Murray; Chris Olah; Mike Schuster; Jonathon Shlens; Benoit Steiner; Ilya Sutskever; Kunal Talwar; Paul Tucker; Vincent Vanhoucke; Vijay Vasudevan; Fernanda Viégas; Oriol Vinyals; Pete Warden; Martin Wattenberg; Martin Wicke; Yuan Yu; Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. https://www.tensorflow.org/.

(39) Ghosh, A. K.; Brindisi, M.; Shahabi, D.; Chapman, M. E.; Mesecar, A. D. Drug Development and Medicinal Chemistry Efforts toward SARS-Coronavirus and Covid-19 Therapeutics. *ChemMedChem* **2020**, *15* (11), 907–932. https://doi.org/10.1002/cmdc.202000223.

(40) Jacobs, J.; Grum-Tokars, V.; Zhou, Y.; Turlington, M.; Saldanha, S. A.; Chase, P.; Eggler, A.; Dawson, E. S.; Baez-Santos, Y. M.; Tomar, S.; Mielech, A. M.; Baker, S. C.; Lindsley, C. W.; Hodder, P.; Mesecar, A.; Stauffer, S. R. Discovery, Synthesis, And Structure-Based Optimization of a Series of N-(Tert-Butyl)-2-(N-Arylamido)-2-(Pyridin-3-Yl) Acetamides (ML188) as Potent Noncovalent Small Molecule Inhibitors of the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) 3CL Protease. *J. Med. Chem.* **2013**, *56* (2), 534–546. https://doi.org/10.1021/jm301580n.

(41) Shimizu, J. F.; Martins, D. O. S.; McPhillie, M. J.; Roberts, G. C.; Zothner, C.; Merits, A.; Harris, M.; Jardim, A. C. G. Is the ADP Ribose Site of the Chikungunya Virus NSP3 Macro Domain a Target for Antiviral Approaches? *Acta Trop.* **2020**, *207*, 105490. https://doi.org/10.1016/j.actatropica.2020.105490.

(42) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.-F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; Fackeldey, K.; Hoffmann, M.; Iavniuk, I.; Wagner, G.; Arthanari, H. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668. https://doi.org/10.1038/s41586-020-2117-z.

(43) Novick, P. A.; Ortiz, O. F.; Poelman, J.; Abdulhay, A. Y.; Pande, V. S. SWEETLEAD: An In Silico Database of Approved Drugs, Regulated Chemicals, and Herbal Isolates for Computer-Aided Drug Discovery. *PLOS ONE* **2013**, *8* (11), e79568. https://doi.org/10.1371/journal.pone.0079568.

(44) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(45) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.;

Banks, J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47* (7), 1750–1759. https://doi.org/10.1021/jm030644s.
(46) Krivák, R.; Hoksza, D. P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. *J. Cheminformatics* **2018**, *10*, 1–12.
(47) Hu, W.; Fey, M.; Zitnik, M.; Dong, Y.; Ren, H.; Liu, B.; Catasta, M.; Leskovec, J. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 22118–22133.
(48) Gainza, P.; Sverrisson, F.; Monti, F.; Rodola, E.; Boscaini, D.; Bronstein, M.; Correia, B. Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat. Methods* **2020**, *17* (2), 184–192.
(49) Krasoulis, A.; Antonopoulos, N.; Pitsikalis, V.; Theodorakis, S. DENVIS: Scalable and High-Throughput Virtual Screening Using Graph Neural Networks with Atomic and Surface Protein Pocket Features. *J. Chem. Inf. Model.* **2022**, *62* (19), 4642–4659.
(50) Deep Neural Virtual Screening (DENVIS), 2023. https://github.com/deeplab-ai/denvis (accessed 2023-02-28).
(51) Kozlovskii, I.; Popov, P. Spatiotemporal Identification of Druggable Binding Sites Using Deep Learning. *Commun. Biol.* **2020**, *3* (1), 1–12. https://doi.org/10.1038/s42003-020-01350-0.
(52) *COVID-19 Molecular Structure and Therapeutics Hub*. https://covid.molssi.org/ (accessed 2023-03-01).
(53) Zacharov, I.; Arslanov, R.; Gunin, M.; Stefonishin, D.; Bykov, A.; Pavlov, S.; Panarin, O.; Maliutin, A.; Rykovanov, S.; Fedorov, M. “Zhores”—Petaflops Supercomputer for Data-Driven Modeling, Machine Learning and Artificial Intelligence Installed in Skolkovo Institute of Science and Technology. *Open Eng.* **2019**, *9* (1), 512–520.
(54) Hofmarcher, M.; Mayr, A.; Rumetshofer, E.; Ruch, P.; Renz, P.; Schimunek, J.; Seidl, P.; Vall, A.; Widrich, M.; Hochreiter, S.; Klambauer, G. Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks. arXiv August 17, 2020. https://doi.org/10.48550/arXiv.2004.00979.
(55) Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; K. Wegner, J.; Ceulemans, H.; Clevert, D.-A.; Hochreiter, S. Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction on ChEMBL. *Chem. Sci.* **2018**, *9* (24), 5441–5451. https://doi.org/10.1039/C8SC00148K.
(56) Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; Vol. 30.
(57) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.
(58) Berenger, F.; Kumar, A.; Zhang, K. Y. J.; Yamanishi, Y. Lean-Docking: Exploiting Ligands’ Predicted Docking Scores to Accelerate Molecular Docking. *J. Chem. Inf. Model.* **2021**, *61* (5), 2341–2352. https://doi.org/10.1021/acs.jcim.0c01452.
(59) Sidorov, P.; Gaspar, H.; Marcou, G.; Varnek, A.; Horvath, D. Mappability of Drug-like Space: Towards a Polypharmacologically Competent Map of Drug-Relevant Compounds. *J. Comput. Aided Mol. Des.* **2015**, *29*, 1087–1108.
(60) Horvath, D.; Orlov, A.; Osolodkin, D. I.; Ishmukhametov, A. A.; Marcou, G.; Varnek, A. A Chemographic Audit of Anti-Coronavirus Structure-Activity Information from Public Databases (ChEMBL). *Mol. Inform.* **2020**, *39* (12), 2000080.
(61) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov’ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA-Platform for Virtual Screening Based on

Fragment and Pharmacophoric Descriptors. *Curr. Comput. Aided Drug Des.* **2008**, *4* (3), 191.

(62) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. ISIDA Property-Labelled Fragment Descriptors. *Mol. Inform.* **2010**, *29* (12), 855–868.

(63) Tong, W.; Wei, Y.; Murga, L. F.; Ondrechen, M. J.; Williams, R. J. Partial Order Optimum Likelihood (POOL): Maximum Likelihood Prediction of Protein Active Site Residues Using 3D Structure and Sequence Properties. *PLoS Comput. Biol.* **2009**, *5* (1), e1000266.

(64) Somarowthu, S.; Yang, H.; Hildebrand, D. G.; Ondrechen, M. J. High-Performance Prediction of Functional Residues in Proteins with Machine Learning and Computed Input Features. *Biopolymers* **2011**, *95* (6), 390–400.

(65) Somarowthu, S.; Ondrechen, M. J. POOL Server: Machine Learning Application for Functional Site Prediction in Proteins. *Bioinformatics* **2012**, *28* (15), 2078–2079.

(66) Krieger, E.; Koraimann, G.; Vriend, G. Increasing the Precision of Comparative Models with YASARA NOVA—a Self-Parameterizing Force Field. *Proteins Struct. Funct. Bioinforma.* **2002**, *47* (3), 393–402.

(67) Iyengar, S. M.; Barnsley, K. K.; Vu, H. Y.; Bongalonta, I. J. A.; Herrod, A. S.; Scott, J. A.; Ondrechen, M. J. Identification and Characterization of Alternative Sites and Molecular Probes for SARS-CoV-2 Target Proteins. *Front. Chem.* **2022**, *10*.

(68) Haupt, V. J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLOS ONE* **2013**, *8* (6), e65894. https://doi.org/10.1371/journal.pone.0065894.

(69) Salentin, S.; Adasme, M. F.; Heinrich, J. C.; Haupt, V. J.; Daminelli, S.; Zhang, Y.; Schroeder, M. From Malaria to Cancer: Computational Drug Repositioning of Amodiaquine Using PLIP Interaction Patterns. *Sci. Rep.* **2017**, *7* (1), 11401. https://doi.org/10.1038/s41598-017-11924-4.

(70) *Rapid Identification of Novel PDE2 Inhibitors – PharmAI*. https://www.pharm.ai/2020/04/27/rapid-identification-of-novel-pde2-inhibitor/ (accessed 2023-03-01).

(71) Holderbach, S.; Adam, L.; Jayaram, B.; Wade, R. C.; Mukherjee, G. RASPD+: Fast Protein-Ligand Binding Free Energy Prediction Using Simplified Physicochemical Features. *Front. Mol. Biosci.* **2020**, *7*, 601065.

(72) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.

(73) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A New Test Set for Validating Predictions of Protein–Ligand Interaction. *Proteins Struct. Funct. Bioinforma.* **2002**, *49* (4), 457–471.

(74) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein–Ligand Docking Using GOLD. *Proteins Struct. Funct. Bioinforma.* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.

(75) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.

(76) Mcgann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian Docking Functions. *Biopolym. Orig. Res. Biomol.* **2003**, *68* (1), 76–90.

(77) Kokh, D. B.; Doser, B.; Richter, S.; Ormersbach, F.; Cheng, X.; Wade, R. C. A Workflow for Exploring Ligand Dissociation from a Macromolecule: Efficient Random Acceleration Molecular Dynamics Simulation and Interaction Fingerprint Analysis of Ligand Trajectories.

J. Chem. Phys. **2020**, *153* (12), 125102.

(78) Schrödinger Release 2022-4: QikProp, Schrödinger, LLC, New York, NY, 2021.

(79) Kozakov, D.; Grove, L. E.; Hall, D. R.; Bohnuud, T.; Mottarella, S. E.; Luo, L.; Xia, B.; Beglov, D.; Vajda, S. The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins. *Nat. Protoc.* **2015**, *10* (5), 733–755. https://doi.org/10.1038/nprot.2015.043.

(80) Yu, J.; Zhou, Y.; Tanaka, I.; Yao, M. Roll: A New Algorithm for the Detection of Protein Pockets and Cavities with a Rolling Probe Sphere. *Bioinformatics* **2010**, *26* (1), 46–52. https://doi.org/10.1093/bioinformatics/btp599.

(81) Venkatraman, V.; Colligan, T. H.; Lesica, G. T.; Olson, D. R.; Gaiser, J.; Copeland, C. J.; Wheeler, T. J.; Roy, A. Drugsniffer: An Open Source Workflow for Virtually Screening Billions of Molecules for Binding Affinity to Protein Targets. *Front. Pharmacol.* **2022**, *13*.

(82) Spiegel, J. O.; Durrant, J. D. AutoGrow4: An Open-Source Genetic Algorithm for de Novo Drug Design and Lead Optimization. *J. Cheminformatics* **2020**, *12* (1), 25. https://doi.org/10.1186/s13321-020-00429-4.

(83) *RDKit*. https://rdkit.sourceforge.net/ (accessed 2023-03-01).

(84) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61* (8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203.

(85) Gorgulla, C.; Das, K. M. P.; Leigh, K. E.; Cespugli, M.; Fischer, P. D.; Wang, Z.-F.; Tesseyre, G.; Pandita, S.; Shnapir, A.; Calderaio, A.; Gechev, M.; Rose, A.; Lewis, N.; Hutcheson, C.; Yaffe, E.; Luxenburg, R.; Herce, H. D.; Durmaz, V.; Halazonetis, T. D.; Fackeldey, K.; Patten, J. J.; Chuprina, A.; Dziuba, I.; Plekhova, A.; Moroz, Y.; Radchenko, D.; Tarkhanova, O.; Yavnyuk, I.; Gruber, C.; Yust, R.; Payne, D.; Näär, A. M.; Namchuk, M. N.; Davey, R. A.; Wagner, G.; Kinney, J.; Arthanari, H. A Multi-Pronged Approach Targeting SARS-CoV-2 Proteins Using Ultra-Large Virtual Screening. *iScience* **2021**, *24* (2). https://doi.org/10.1016/j.isci.2020.102021.

(86) Druzhilovskiy, D. S.; Stolbov, L. A.; Savosina, P. I.; Pogodin, P. V.; Filimonov, D. A.; Veselovsky, A. V.; Stefanisko, K.; Tarasova, N. I.; Nicklaus, M. C.; Poroikov, V. V. Computational Approaches To Identify A Hidden Pharmacological Potential In Large Chemical Libraries. *Supercomput. Front. Innov.* **2020**, *7* (3). https://doi.org/10.14529/jsfi200306.

(87) Poroikov, V. V.; Filimonov, D. A.; Gloriozova, T. A.; Lagunin, A. A.; Druzhilovskiy, D. S.; Rudik, A. V.; Stolbov, L. A.; Dmitriev, A. V.; Tarasova, O. A.; Ivanov, S. M.; Pogodin, P. V. Computer-Aided Prediction of Biological Activity Spectra for Organic Compounds: The Possibilities and Limitations. *Russ. Chem. Bull.* **2019**, *68* (12), 2143–2154. https://doi.org/10.1007/s11172-019-2683-0.

(88) Filimonov, D. A.; Zakharov, A. V.; Lagunin, A. A.; Poroikov, V. V. QNA-Based ‘Star Track’ QSAR Approach. *SAR QSAR Environ. Res.* **2009**, *20* (7–8), 679–709. https://doi.org/10.1080/10629360903438370.

(89) *UCSF DOCK*. https://dock.compbio.ucsf.edu/ (accessed 2023-03-01).

(90) *AutoDock Vina*. https://vina.scripps.edu/ (accessed 2023-03-01).

(91) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q. Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2. *Science* **2020**, *367* (6485), 1444–1448. https://doi.org/10.1126/science.abb2762.

(92) Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X. Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. *Nature* **2020**, *581* (7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5.

(93) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T.

A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. *J. Med. Chem.* **2006**, *49* (21), 6177–6196. https://doi.org/10.1021/jm051256o.

(94) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45* (1), 160–169. https://doi.org/10.1021/ci049885e.

(95) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inform.* **2012**, *31* (3–4), 301–312. https://doi.org/10.1002/minf.201100163.

(96) Bishop, C. M.; Svensén, M.; Williams, C. K. Developments of the Generative Topographic Mapping. *Neurocomputing* **1998**, *21* (1–3), 203–224.

(97) Bishop, C. M.; Svensén, M.; Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **1998**, *10* (1), 215–234. https://doi.org/10.1162/089976698300017953.

(98) Kohonen, T. *Self-Organization and Associative Memory*; Springer Berlin Heidelberg, 1989.

(99) Kohonen, T. *Self-Organizing Maps*; Huang, T. S., Kohonen, T., Schroeder, M. R., Series Eds.; Springer Series in Information Sciences; Springer: Berlin, Heidelberg, 2001; Vol. 30. https://doi.org/10.1007/978-3-642-56927-2.

(100) Kayastha, S.; Horvath, D.; Gilberg, E.; Gütschow, M.; Bajorath, J.; Varnek, A. Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps. *J. Chem. Inf. Model.* **2017**, *57* (5), 1218–1232.

(101) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.

(102) Schake, P.; Dishnica, K.; Kaiser, F.; Leberecht, C.; Haupt, V. J.; Schroeder, M. An Interaction-Based Drug Discovery Screen Explains Known SARS-CoV-2 Inhibitors and Predicts New Compound Scaffolds. *Scientific Reports* **(under review)**.

(103) Le, T.; Hempel, T.; Winter, R.; Olsson, S.; Raich, L.; Elez, K.; Noé, F.; Narangoda, C.; Gokcan, H.; Gusev, F.; Zubatiuk, R.; Kurnikova, M.; Gutkin, E.; Bosko, I. P.; Yushkevich, A.; Shuldau, M.; Karpenko, A. D.; Kornoushenko, Y. V.; García-Sastre, A.; Furs, K.; Bureau, R.; Benabderrahmane, M.; Naffakh, N.; Cirou, B.; Bousquet-Melou, P.; Charton, B.; Ford, B.; Gil, G.; Epitropakis, N.; Krasoulis, A.; Pitsikalis, V.; Antonopoulos, N.; Theodorakis, S.; Schimunek, J.; Widrich, M.; Eghbal-zadeh, H.; Lee, S. Y.; Seidl, P.; Ruch, P.; Halmich, C.; Zhang, K.; Berenger, F.; Yamanishi, Y.; III, C. L. B.; Kumar, A.; Jain, M.; Bengio, E.; Bengio, Y.; Marcou, G.; Popov, P.; Haupt, J.; Schroeder, M.; Kaiser, F.; Pugliese, L.; Paiardi, G.; Wade, R.; Hanke, A.; Goßen, J.; D'Arrigo, G.; Rossetti, G.; Albani, S.; Spyrakis, F.; Mukherjee, G.; Kokh, D.; Sadiq, S. K.; Nunes-Alves, A.; Carloni, P.; Musiani, F.; Gianquinto, E.; Athanasiou, C.; Kovachka, S.; Tsengenes, A.-A.; Joseph, B.; Talarico, C.; Manelfi, C.; Beccari, A.; Venkatraman, V.; Ondrechen, M. J.; Olson, D.; Copeland, C.; Roy, A.; Wheeler, T.; Tesseyre, G.; Gorgulla, C.; PadmanabhaDas, K.; Wagner, G.; Fackeldey, K.; Gruber, C. C.; Fischer, P. D.; Yust, R.; Pandita, S.; Wang, Z.-F.; Veselovsky, A.; Poroikov, V.; Druzhilovskiy, D.; Stolbov, L.; Pogodin, P.; Sobolev, B.; Barnsley, K.; Gulotta, M. R.; Lombino, J.; Simone, G. D.; Perricone, U.; Mekni, N.; Rosa, M. D.; Iyengar, S.; Watowich, S.; Falsafi, B.; Steinkellner, G.; Durmaz, V.; Cespugli, M.; Singh, A.; Gruber, K.; Hetmann, M.; Kozlovskii, I.; Zaretckii, M.; Medvedev, A.; Blaschitz, K.; Korablyov, M.; Allen, W.; Loesekrug-Pietri, A.; Hermans, T. JEDI Billion Molecules against Covid-19: Compounds Synthesized. **2021**. https://doi.org/10.6084/m9.figshare.14458896.v3.

(104) Hillen, H. S.; Kokic, G.; Farnung, L.; Dienemann, C.; Tegunov, D.; Cramer, P. Structure of Replicating SARS-CoV-2 Polymerase. *Nature* **2020**, *584* (7819), 154–156. https://doi.org/10.1038/s41586-020-2368-8.

(105) Jerabek-Willemsen, M.; André, T.; Wanner, R.; Roth, H. M.; Duhr, S.; Baaske, P.; Breitsprecher, D. MicroScale Thermophoresis: Interaction Analysis and Beyond. *J. Mol. Struct.* **2014**, *1077*, 101–113. https://doi.org/10.1016/j.molstruc.2014.03.009.

(106) Chen, K. Y.; Krischuns, T.; Varga, L. O.; Harigua-Souiai, E.; Paisant, S.; Zettor, A.; Chiaravalli, J.; Delpal, A.; Courtney, D.; O'Brien, A.; Baker, S. C.; Decroly, E.; Isel, C.; Agou, F.; Jacob, Y.; Blondel, A.; Naffakh, N. A Highly Sensitive Cell-Based Luciferase Assay for High-Throughput Automated Screening of SARS-CoV-2 Nsp5/3CLpro Inhibitors. *Antiviral Res.* **2022**, *201*, 105272. https://doi.org/10.1016/j.antiviral.2022.105272.

(107) Douangamath, A.; Fearon, D.; Gehrtz, P.; Krojer, T.; Lukacik, P.; Owen, C. D.; Resnick, E.; Strain-Damerell, C.; Aimon, A.; Ábrányi-Balogh, P.; Brandão-Neto, J.; Carbery, A.; Davison, G.; Dias, A.; Downes, T. D.; Dunnett, L.; Fairhead, M.; Firth, J. D.; Jones, S. P.; Keeley, A.; Keserü, G. M.; Klein, H. F.; Martin, M. P.; Noble, M. E. M.; O'Brien, P.; Powell, A.; Reddi, R. N.; Skyner, R.; Snee, M.; Waring, M. J.; Wild, C.; London, N.; von Delft, F.; Walsh, M. A. Crystallographic and Electrophilic Fragment Screening of the SARS-CoV-2 Main Protease. *Nat. Commun.* **2020**, *11* (1), 5047. https://doi.org/10.1038/s41467-020-18709-w.

(108) Schuller, M.; Correy, G. J.; Gahbauer, S.; Fearon, D.; Wu, T.; Díaz, R. E.; Young, I. D.; Carvalho Martins, L.; Smith, D. H.; Schulze-Gahmen, U.; Owens, T. W.; Deshpande, I.; Merz, G. E.; Thwin, A. C.; Biel, J. T.; Peters, J. K.; Moritz, M.; Herrera, N.; Kratochvil, H. T.; QCRG Structural Biology Consortium; Aimon, A.; Bennett, J. M.; Brandao Neto, J.; Cohen, A. E.; Dias, A.; Douangamath, A.; Dunnett, L.; Fedorov, O.; Ferla, M. P.; Fuchs, M. R.; Gorrie-Stone, T. J.; Holton, J. M.; Johnson, M. G.; Krojer, T.; Meigs, G.; Powell, A. J.; Rack, J. G. M.; Rangel, V. L.; Russi, S.; Skyner, R. E.; Smith, C. A.; Soares, A. S.; Wierman, J. L.; Zhu, K.; O'Brien, P.; Jura, N.; Ashworth, A.; Irwin, J. J.; Thompson, M. C.; Gestwicki, J. E.; von Delft, F.; Shoichet, B. K.; Fraser, J. S.; Ahel, I. Fragment Binding to the Nsp3 Macrodomain of SARS-CoV-2 Identified through Crystallographic Screening and Computational Docking. *Sci. Adv.* **2021**, *7* (16), eabf8711. https://doi.org/10.1126/sciadv.abf8711.

(109) Douangamath, A.; Powell, A.; Fearon, D.; Collins, P. M.; Talon, R.; Krojer, T.; Skyner, R.; Brandao-Neto, J.; Dunnett, L.; Dias, A.; Aimon, A.; Pearce, N. M.; Wild, C.; Gorrie-Stone, T.; von Delft, F. Achieving Efficient Fragment Screening at XChem Facility at Diamond Light Source. *J. Vis. Exp. JoVE* **2021**, No. 171. https://doi.org/10.3791/62414.

(110) Winn, M. D.; Ballard, C. C.; Cowtan, K. D.; Dodson, E. J.; Emsley, P.; Evans, P. R.; Keegan, R. M.; Krissinel, E. B.; Leslie, A. G. W.; McCoy, A.; McNicholas, S. J.; Murshudov, G. N.; Pannu, N. S.; Potterton, E. A.; Powell, H. R.; Read, R. J.; Vagin, A.; Wilson, K. S. Overview of the CCP4 Suite and Current Developments. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67* (Pt 4), 235–242. https://doi.org/10.1107/S0907444910045749.

(111) Pearce, N. M.; Krojer, T.; Bradley, A. R.; Collins, P.; Nowak, R. P.; Talon, R.; Marsden, B. D.; Kelm, S.; Shi, J.; Deane, C. M.; von Delft, F. A Multi-Crystal Method for Extracting Obscured Crystallographic States from Conventionally Uninterpretable Electron Density. *Nat. Commun.* **2017**, *8* (1), 15123. https://doi.org/10.1038/ncomms15123.

(112) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and Development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66* (Pt 4), 486–501. https://doi.org/10.1107/S0907444910007493.

(113) Smart, O.; Sharff, A.; Womack, T.; Flensburg, C.; Keller, P.; Paciorek, W.; Vonrhein, C.; Bricogne, G. Grade, 2021. https://www.globalphasing.com/ (accessed 2023-03-09).

(114) Murshudov, G. N.; Vagin, A. A.; Dodson, E. J. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr. D Biol. Crystallogr.* **1997**, *53* (Pt 3), 240–255. https://doi.org/10.1107/S0907444996012255.

(115) Bricogne, G.; Blanc, E.; Brandl, M.; Flensburg, C.; Keller, P.; Paciorek, W.; Roversi, P.;

Sharff, A.; Smart, O. S.; Vonrhein, C.; Womack, T. O. BUSTER, 2017.