Team name	MLINCH
Team member(s) (firstname lastname; …)	Petr Popov
Affiliation	Skoltech
Contact email	p.popov@skoltech.ru
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	Nucleocapsid, Papain-like protease, 3C-like protease, RNA-dependent RNA polymerase, Spike

Section 1: methods & metrics

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 analysed the three-dimensional structures of the SARS-CoV-2 proteins using our recently developed deep learning approach called BiteNet (Kozlovksii & Popov, 2020) 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 (https://covid.molssi.org). 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 (Zacharov et al 2019) 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.

Section 2: targets

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).

Section 3: libraries

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).

Section 4: results

We are exciting to share the results and looking forward for the experimental validation of the most promising compounds. Currently only few conventional binding sites are considered for SARS-CoV-2 protein targets. Interestingly, we observed several compounds docking with very high score and reasonable binding poses (e.g. comparable to known drug-target protein complexes) with respect to the previously unseen and/or unexplored binding sites. We hope that identified binding pockets presents pharmacologically relevant regions on the SARS-Cov-2 proteins. For each out of 5 targets we provided list with 10K compounds by means of canonical smiles and InChI identifiers. For the Drug Bank compounds we provide the PubChem-ID, and for the Enamine Real Library compounds we provide compound name as downloaded from the VirtualFlow web-server (https://virtual-flow.org/real-library). The compounds with no identifier correspond to the de novo designed small molecules and present only for the 3CLPro and RNA polymerase targets. We manually rearranged the top 50 compounds to ensure that at least one top compound for each binding site is presented and taking into account the ADME properties, drug-likeness, synthetic accessibility and toxicity. We are currently preparing repository with the obtained data and manuscript describing the obtained results.

NB: We observed that different software produce different “canonical smiles”. Particularly, we observed that PubChem canonical smiles in general do not correspond to the RDKit canonical smiles. In the provided lists, canonical smiles were obtained using the RDKit software (https://www.rdkit.org). We also observed that for some RDKit smiles InChI cannot be generated; in such cases we left the InChI field blank. Please contact us if you need more details, e.g. 3D conformers of these compounds.

Petr Popov,
Assistant Professor,
Skoltech,
p.popov@skoltech.ru
17 July 2020