Team name	SARSWARS
Team member(s) (firstname lastname; …)	Amitava Roy; Vishwesh Venkatraman; Travis Wheeler; Daniel Olson; Conner Copeland; Jeremiah Gaiser
Affiliation	Amitava Roy Contractor, MSC. Bioinformatics and Computational Biosciences Branch (BCBB) NIH/NIAID/OD/OSMO/OCICB Rocky Mountain Laboratories, NIH, USA Vishwesh Venkatraman Department of Chemistry, Norwegian University of Science and Technology, Trondheim, Norway Travis Wheeler Department of Computer Science Division of Biological Sciences University of Montana, USA Daniel Olson Department of Computer Science University of Montana, USA Conner Copeland Department of Computer Science Division of Biological Sciences University of Montana, USA Jeremiah Gaiser Department of Computer Science Division of Biological Sciences University of Montana, USA

Contact email	amitava.roy@nih.gov
Contact phone number (optional)	
Protein targets (for example: 3CLPro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, Nspx, OrfXx, N, E, etc…) | 3 required	NSP12, N, 3CLPro

Section 1: methods & metrics

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

Methods:

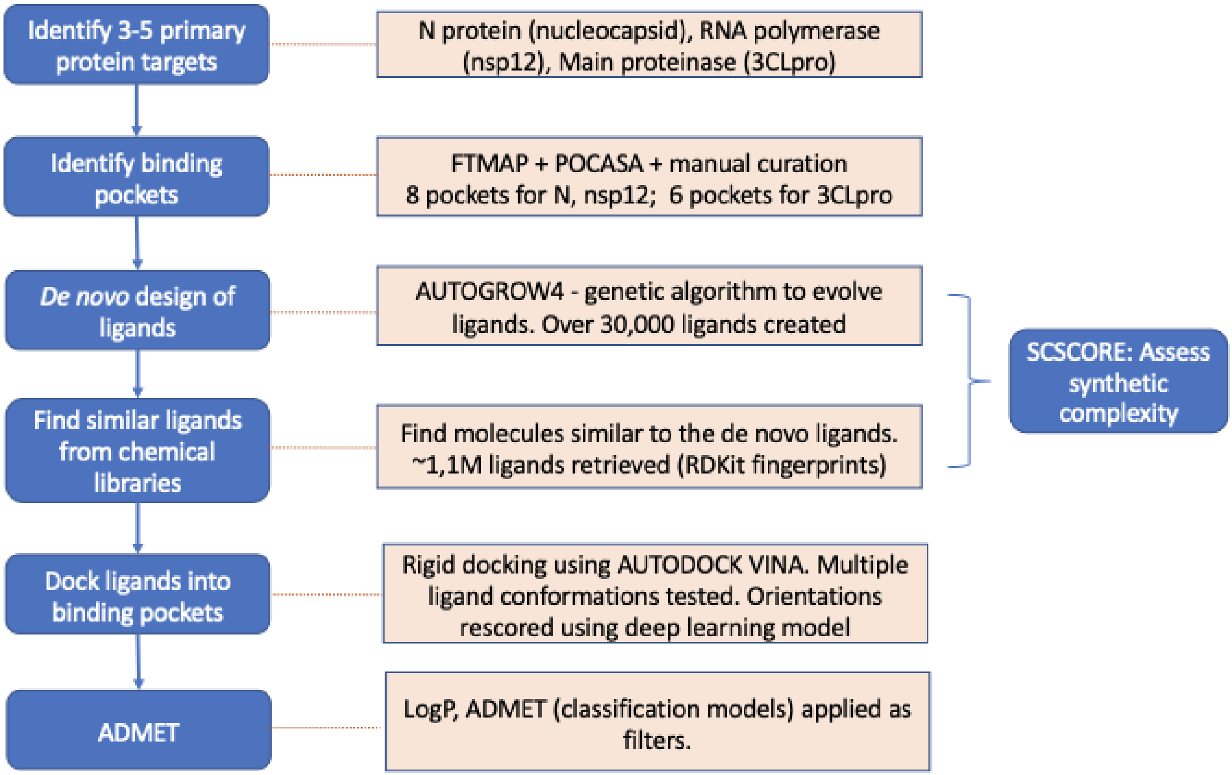
Identify 3-5 primary protein targets
N protein (nucleocapsid), RNA polymerase (nsp12), Main proteinase (3CLpro)
Identify binding pockets
FTMAP + POCASA + manual curation
8 pockets for N, nsp12; 6 pockets for 3CLpro
De novo design of ligands
AUTOGROW4 - genetic algorithm to evolve ligands. Over 30,000 ligands created
Find similar ligands from chemical libraries
Find molecules similar to the de novo ligands. ~1,1M ligands retrieved (RDKit fingerprints)
SCSCORE: Assess synthetic complexity
Dock ligands into binding pockets
Rigid docking using AUTODOCK VINA. Multiple ligand conformations tested. Orientations rescored using deep learning model
ADMET
LogP, ADMET (classification models) applied as filters.

Figure 1. Schematic shows the workflow to identify suitable drugs for the chosen protein targets.

An overview of our approach is provided in Figure 1. Details about selection, modeling of coordinates and glycosylation of the targets are described in the next section. The binding pockets for 3 selected targets: N protein (nucleocapsid), RNA polymerase (nsp12), main proteinase (3CLpro), were identified from literature search and using cavity detection software FTMAP (10.1038/nprot.2015.043) and POCASA (10.1093/bioinformatics/btp599). The results from these software were further examined visually. Six pocket-like regions were identified: 2 each for N, NSP12 and 3CLpro. Some of the pocket-like regions were too big to be occupied by a typical sized ligand. Consequently, larger pocket-like regions were subdivided into smaller pockets. Twenty two pockets, 8 each for N and NSP12 and 6 for 3CLPro, were finalized as targets. For each pocket, a number of ligand molecules were designed from scratch using the software AUTOGROW4 (10.1186/s13321-020-00429-4). The program uses a genetic algorithm to evolve ligands from seed fragments (obtained from ZINC). The molecules are subsequently docked into the pockets of the specified target protein and ranked based on the docking score. Additional filters based on the Lipinski RO5 are employed to exclude candidate structures that do not satisfy drug-like criteria. AUTOGROW4 performs *in silico* chemical reactions to generate new child compounds derived from a parent molecule. This also ensures that the generated molecules are likely to be synthesizable. Over 30000 unique structures were generated over 5 independent runs of the *de novo* software.

Using the *de novo* designed molecules as seeds, various chemical libraries (see Table 1) were queried to identify library-sourced lead compounds similar to the molecules developed with AUTOGROW4 (see below). In order to ensure that the molecules are synthetically feasible, the structures were assessed using the software SCSCORE (10.1021/acs.jcim.7b00622) which provides an estimate (ranging between 1 and 5) of the expected number of reaction steps required to produce a target molecule. For a majority of the molecules identified by the fingerprint search, the number of steps was predicted to be within 3.

To identify library-sourced lead compounds, 1024-bit ECFP4 fingerprints were computed for de novo molecules produced by AUTOGROW4, and also for all ~3.7 billion library compounds (fingerprints computed using RDKIT, https://www.rdkit.org). Among the all library compounds, ~97,000 had Tanimoto similarities greater than 0.6 to some *de novo* ligand, and another ~955,000 had Tanimoto similarities of 0.5-0.6. Among the 97K closer neighbours: ~43K were identified for nsp12, ~34k for N, ~20k for 3CLPro. All ~1.05M neighbours were docked into the respective protein targets using AUTODOCK VINA (http://vina.scripps.edu/). Up to 4 docking poses for each ligand were retained. A deep learning model was then trained to discriminate between binding and non-binding compounds. The model takes as input, descriptors calculated by SMINA (https://sourceforge.net/projects/smina/) and a statistical knowledge-based potential term

(10.1186/s13321-019-0373-4). To train the model, docking poses were generated for a number of targets contained in the DUD-E (**10.1021/jm300687e**): ~60000 complexes and LIT-PCBA (10.1021/acs.jcim.0c00155): ~70000 complexes. For each target, a number docked poses for a set of active and decoy ligands were analysed. Docked poses for the DUD-E were taken from http://bits.csb.pitt.edu/files/docked_dude.tar. In the case of LIT-PCBA, since the number of available decoy structures for docking were quite large, we only performed docking on a selected set of decoys.

The trained model was then applied to the nearest neighbour candidates to identify potential active molecules, and ranked according to predicted probability of binding. In addition, fingerprint based classification models were trained for Drug Solubility, Blood Brain Barrier Permeability, Human Intestinal Absorption, AMES Mutagenicity, HERG Cardiac Toxicity ,Hepatotoxicity, cytochrome p450 interaction, Metabolic Stability and EPA LD50 Toxicity. These models were used as additional filters to arrive at the final list of candidates.

Section 2: targets

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

Zhanglab published models, created by I-TASSER suit, of all full length viral proteins of SARS-CoV-2 (https://zhanglab.ccmb.med.umich.edu/COVID-19/). Initial models were downloaded from there. The modeled coordinates have estimated TM-score of 0.67 to 0.96 with their native structure. TM-score is a quantitative measurement of similarity between two proteins and has the value [0,1], where 1 indicates a perfect match between two structures. A TM-score of greater than 0.5 indicates similar folding patterns between the proteins. Hydrogen coordinates were initialized and were rebuilt using molecular modeling software CHARMM.

We targeted only viral proteins as that strategy aligns with our long-term research goal. Out of the four possible viral protein targets, we did not choose S protein for the reasons below.

a) Epitopes of S^1, the domain that initiates binding with ACE2 protein, are highly variable among different serotypes.
b) S protein is highly glycosylated. We were not confident about modeling glycosylation of S protein within the timeframe of the competition.
c) During cell entry, S protein goes through large conformational change. Information about intermediate conformations was not available when we started working on the project.

Target 1:   NSP12/RDRP , estimated TM-score 0.80.

Target 2:   3CLPro, estimated TM-score 0.96.

Target 3:   N, estimated TM-score 0.67.

Glycosylation sites for the above targets were determined by searching literature and using N-GlyDe (http://bioapp.iis.sinica.edu.tw/N-GlyDE/) for *in silico* prediction of N-linked glycans. We did not use *in silico* methods for prediction of O-linked glycans, as we could not find any reliable method for such predictions. Our predicted glycosylation sites are residue 269 of N and residues 767 and 911 of NSP12. As none of the glycosylation sites were near any of the possible binding pockets we identified, we did not consider glycosylation for our docking exercises.

Section 3: libraries

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

The selection of the screening libraries was based on the synthetic feasibility of the compounds. To this end, we compiled a large collection of molecules from suppliers such as Enamine and Molport. Since databases such as SweetLead, Drugbank, CAS antivirals contain drugs in use/early-late stage clinical trials, our initial assumption was that these could be repurposed. Both Pubchem and Zinc are popular databases that offer collections from different vendors. The only exceptions to the list include SAVI and GDB and contain compounds that require synthesis routes to be established.

Database	Ligands
Sweetlead (https://simtk.org/projects/sweetlead)	~4K
Drugbank (https://www.drugbank.ca/releases/latest)	~10K
CAS Antivirals	~50K

Merck	~5.0M
MOLPORT (https://www.molport.com/shop/libraries-collections)	~7.6M
PUBCHEM (ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/)	~103M
ZINC15 (http://files.docking.org/catalogs/)	~417M
GDB (http://gdb.unibe.ch/downloads/)	~1.03B
SAVI (https://cactus.nci.nih.gov/download/savi_download/)	~1.09B
ENAMINE (https://enamine.net/library-synthesis/real-compounds/real-database)	~1.2 B
Total	**~3.7B unique**

Table 1. Molecule libraries included in the fingerprint search.

Section 4: results

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

Results: We have compiled a list of 30000 ligands (10 K per protein) in a tab separated file. Inchi keys include both comma and semicolon, which made it difficult to create comma or semicolon separated file of the list. The ligands are ranked based on a probability of binding calculated from different components interaction energy between ball-and-stick with partial charges representations of ligands and proteins. In our opinion, such representations of molecules did not capture the full aspects of interactions between ligands and proteins (see the discussion in the next section). Consequently, estimates of binding affinity were not reliable in our current effort. Efforts to provide more reliable binding affinity estimates are currently in progress.

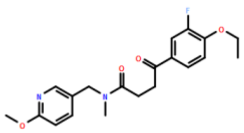
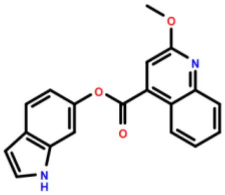
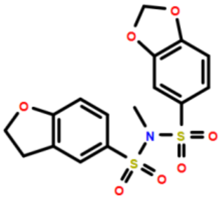
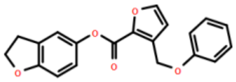
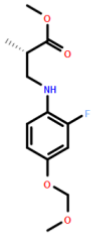

Figure 2: Top 5 compounds for N-protein.

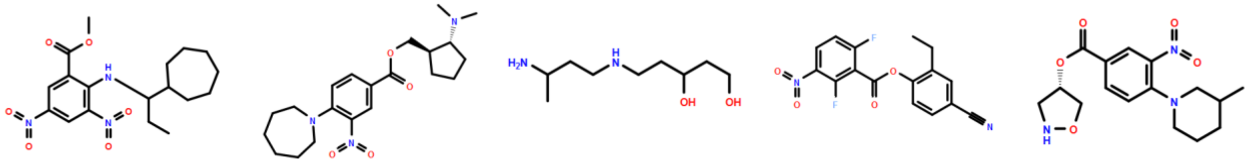

Figure 3: Top 5 compounds for NSP-12.

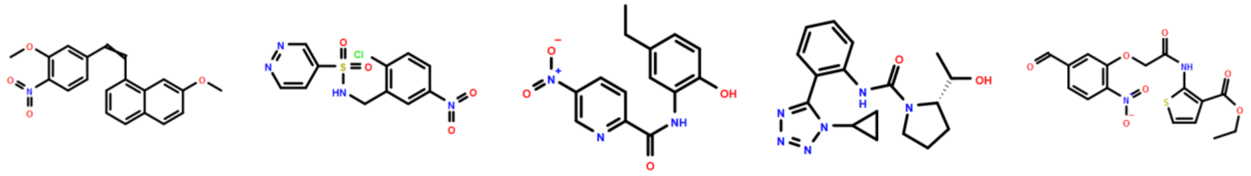

Figure 4: Top 5 compounds for 3clpro.

Figures 2-4 list the top 5 compounds for the chosen targets. A majority of the selected compounds were seen to satisfy Lipinski RO5 rules with predicted LogP values in the 2-4 range. Alongside solubility and other ADMET criteria, for the most part, the compounds were seen to satisfy drug likeness criteria. The ADMET models however suggested some level of hepatotoxicity.

Other comments:

In agreement with some of the recent publications (10.1371/journal.pone.0220113; 10.1021/acs.jcim.0c00155), we found machine learning (ML) approaches to identify true ligands learn properties of decoys as discriminators, rather than learning properties of interactions between ligands and proteins. While such learning schemes can be useful for screening a smaller number of ligands, the high false positive rate of such approaches makes them unsuitable for screening billions of molecules. The failure of docking+ML approaches to capture interactions between ligands and protein as discriminators has been attributed to biases in training datasets. In our opinion, inability of currently available docking scores to capture all aspects of interactions between proteins and ligands also contributes to the high false positive rate of docking+ML approaches as well. An alternative approach to popular scoring schemes is to calculate various descriptors based on molecule surface properties to represent interactions between ligands and proteins (e.g. surface autocorrelations, min, max, range, skewness, kurtosis of the properties mapped to the surface). Given the time frame of the competition, we are yet to fully optimize our pipeline to include our surface descriptors, so the model used to produce the results presented here does not include these features. We expect to extend our model to effectively incorporate these features in the coming months. The software we have used is open source with some in-house custom scripts. We are currently working on refining the software protocol which will be made available open source.