Supplementary Information for Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies

Rohit Batra$^1$, Henry Chan$^{1,2}$, Ganesh Kamath$^3$, Rampi Ramprasad$^4$, Mathew J. Cherukara$^5$, and Subramanian K.R.S. Sankaranarayanan$^{1,2}$

$^1$Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States
$^2$Department of Mechanical and Industrial Engineering, University of Illinois, Chicago, Illinois 60607, United States
$^3$Dalzielfiver LLC, 3500 Carlfield St., El Sobrante, CA 94803, United States
$^4$School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, Georgia 30332, United States
$^5$Advanced Photon Source, Argonne National Laboratory, Lemont, Illinois 60439, United States

August 7, 2020

1 Molecular Fingerprints

Based on our past experience on fingerprinting organic materials, three hierarchical levels of features were considered, capturing different geometric and chemical information about ligands at multiple length-scales. At the atomic scale, a count of a predefined set of motifs is included. The motifs are specified by the generic label "A$^i$B$^j$C$^k$", representing an i-fold coordinated A atom, a j-fold coordinated B atom, and a k-fold coordinated C atom, connected in the specified order. For example, N3-C3-C4 represents a three-fold coordinated N, a three-fold coordinated carbon and a four-fold coordinated carbon [1]. At a slightly larger length-scale, quantitative structure-property relationship (QSPR) descriptors, often used in chemical and biological sciences, and implemented in the RDKit Python library, were used [2, 3, 4]. Lastly, at the highest length-scale, ‘morphological descriptors’, such as length of the largest side-group, shortest topological distance between rings, etc. were considered. More details on the different hierarchical descriptors can be found in our previous works [5].

2 Docking Simulations

Docking calculations were performed for 187 top candidates identified based on their ML predicted Vina scores. In-line with the works of Smith et al. [6], we used the Autodock Vina software [7] to compute binding affinities between the top candidates and the SARS-CoV-2 S-protein:ACE2 complex. Initial structures of the complex were generated from homology models (SWISSMODEL), where NCBI Reference Sequence of the SARS-CoV-2 S-protein is YP_009724390.1 and protein data bank ID of the ACE2 receptor is PDB:2AJF. Details regarding the construction and modeling of the S-protein:ACE2 complex are described here [6]. The S-protein has the necessary mutations from its predecessor SARS variety SARS-CoV, namely at L(455),
F(486), Q(493), S (494), and N(501), respectively, which is illustrated in Figure 1b of the main manuscript. Smith et al. [6] focused on this binding pocket region to evaluate the binding affinities of different molecules.

The docking receptors obtained from Smith et al. [6] consists of six conformations of S-protein:ACE2 interface as well as isolated SARS-CoV-2 S-protein receptor (i.e., with the ACE2 receptor removed), which were sampled using root mean squared displacement (RMSD) based clustering from 1.3 microsecond long all-atom Temperature Replica Exchange GROMACS simulations of the S-protein:ACE2 complex in water. Setup of the docking calculations is similar to that described by Smith et al. [6], which defines a 1.2 nm × 1.2 nm × 1.2 nm search space that encompasses the binding pocket located at the S-protein:ACE2 interface shown in Figure 1b in the main manuscript. The same search space was explored for the isolated S-protein receptor cases as well. For each candidate, the docking procedure finds the top 10 optimized docking configurations and selects one with the best Vina score. The docking ligands were prepared from SMILES strings of the candidates using the Open Babel software [8].

3 Learning Curves for Random Forest Models

![Learning curves for the isolated S-protein and the S-protein:ACE2 interface models](image)

Figure S1: Learning curves for the the isolated S-protein (left panel) and the S-protein:ACE2 interface (right panel) models developed in this work. Mean absolute errors (MAE) on the test set can be seen to converge around 0.2 and 0.55 kcal/mol for the S-protein and the S-protein:ACE2 interface models, respectively. This level of accuracy is below the acceptable level of chemical accuracy (1 kcal/mol) and thus the ML models are suitable for screening purposes. More details on the ML training procedure and datasets are provided in the main article.
4 Validation of ML models

Figure S2: Parity plot of the ML Vina score predictions for the S-protein:ACE2 interface (left panel) and the isolated S-protein (right panel) systems against those obtained from Autodock simulations for the identified top 187 molecules. ML predictions can be seen to match well, demonstrating the accuracy of the ML models. We note that 12 of the 187 candidates had greater than 0 Vina score (i.e., no binding affinity) for the S-protein:ACE2 interface complex, and are not included in this plot for better readability. These 12 cases highlight limitations of the ML model employed and are available in the Supporting files. Further, owing to the small dataset size, which also belongs to a region of molecular distribution with extreme Vina scores, the correlation coefficient were computed to be 0.16 and 0.40 for interface and S-protein systems, respectively.

5 Validation of Docking Simulations

Figure S3: Parity plot of the Vina scores for the S-protein:ACE2 interface (left panel) and the isolated S-protein (right panel) systems computed in this work against those obtained from Smith dataset for the common 29 molecules. Our docking simulation results can be seen to match well with the past work. ML predictions for these 29 cases have also be included.
6 Illustrations of S-protein:ACE2 Interface-ligand Complex

The binding poses for the top FDA approved ligands with high Vina scores for the S-protein:ACE2 interface complex are illustrated in Fig. S4. As reflected in Fig. 5 of the main manuscript, chemical fragments, such as oxolane, hydroxy, imidazole, piperidine and benzenesulphonate-derived groups, were identified to promote ligand binding to the S-protein and the interface systems. Other molecular fragments include appendages, such as alkyl, alcohol chains which are generally introduced as substituent or linker to ring systems often to alter the entropic behavior of the scaffold while trying to tailor the delicate balance of hydrophobic/hydrophilic nature of the ligand. Another functional moiety occurring is the fluoromethyl group (CF3) which are known to improve adsorption digestion metabolism and extraction (ADME) and pharmaco-kinetic properties. Further, nitrogen was identified as a key atom in fused rings and heterocycles with a large number of top ligands consisting of theazole moiety. As seen from the docking poses in Fig. S4, N-ring containing ligands, in general, interact with the side chain and backbone of Q493 and S494 respectively; note S494 and Q493 are two of the five mutating sites from the SARS-CoV 2002 virus. For example in pemrolast, the azole nitrogen interacts via a medium hydrogen bond with the side chain hydrogens of QM493. The pyrimidine moiety in Sulfamerazine likes to interact strongly with the side chain Q493 and the backbone oxygen of S494. In Valaciclovir, the isopropyl groups interacts with the backbone oxygen of Q493, with a bonding distance of around 2.8 Å. Thus, we postulate that ligand interactions with Q493 and S494 of the SARS-CoV-2 may be partly responsible for their efficacy.

Figure S4: Cartoon illustrations of the binding pose of the top FDA approved ligands with low Vina scores for the S-protein:ACE2 interface complex.
7 Important Ligand Descriptors

Figure S5: Relative feature importance (for top 20 descriptors) as extracted from the two random forest (RF) models. The feature importance in case of RF is based on the concept of mean decrease in impurity (MDI), as explained here (G. Louppe, Understanding Random Forests: From Theory to Practice, PhD Thesis, U. of Liege, 2014). The first three letters of the descriptor, i.e., afp or mfp, denotes atomic triples or QSPR features, respectively (see section "Molecular Fingerprints" for details).

8 Details of Supporting Files

Following are the details of the files provided with this work:

1. FDA Approved Top Candidates, 80 entries (includes Generic names, SMILES, Docking results and ML predictions; source CureFFI dataset)
2. Other Drugs Top Candidates, 107 entries (includes SMILES, Docking results and ML predictions; source DrugCentral dataset)
3. Small Bio-molecules Top Candidates, 19,116 entries (includes SMILES and ML predictions; source BindingDB dataset)
4. Chemical fragments and their frequency distribution in the two drug datasets, and the identified top scoring 187 ligands.
5. n-octanol/water partition coefficient (log P), Henry’s constant (log H), average molecular weight, and number of hydrogen bond donors and acceptors for the top ligands identified in this work. 50 entries (source www.chemspider.com)

References

[1] Tran Doan Huan, Arun Mannodi-Kanakkithodi, and Rampi Ramprasad. Accelerated materials property predictions and design using motif-based fingerprints. Phys. Rev. B, 92(1):014106, 2015.

[2] Chartchalerm Isaranakura-Na-Ayudhya, Thanakorn Naenna, Chanin Nantasenamat, and Virapong Prachayasittikul. A practical overview of quantitative structure-activity relationship. 2009.

[3] Chanin Nantasenamat, Chartchalerm Isaranakura-Na-Ayudhya, and Virapong Prachayasittikul. Advances in computational methods to predict the biological activity of compounds. Expert Opin. Drug Discov., 5(7):633–654, 2010.
[4] RDKit open source toolkit for cheminformatics.

[5] Chiho Kim, Anand Chandrasekaran, Tran Doan Huan, Deya Das, and Rampi Ramprasad. Polymer genome: A data-powered polymer informatics platform for property predictions. *J. Phys. Chem. C*, 122(31):17575–17585, 2018.

[6] Micholas Smith and Jeremy C. Smith. Repurposing therapeutics for COVID-19: Supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface. *ChemRxiv*, 3 2020.

[7] Oleg Trott and Arthur J Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, 2010.

[8] Noel M O’Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *J. Cheminformatics*, 3(1):33, 2011.