This article is made available via the ACS COVID-19 subset for unrestricted RESEARCH re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for the duration of the World Health Organization (WHO) declaration of COVID-19 as a global pandemic.

The Journal of Physical Chemistry Letters

pubs.acs.org/JPCL Letter

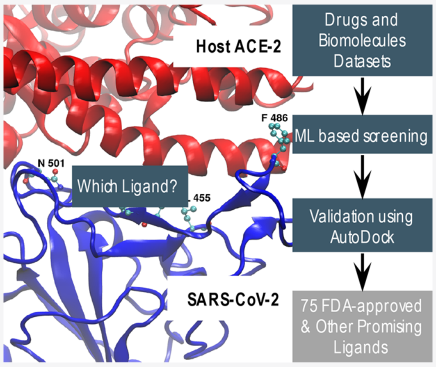
Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies

Rohit Batra,* Henry Chan, Ganesh Kamath, Rampi Ramprasad, Mathew J. Cherukara, and Subramanian K.R.S. Sankaranarayanan*

Cite This: J. Phys. Chem. Lett. 2020, 11, 7058−7065 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The current pandemic demands a search for therapeutic agents against the novel coronavirus SARS-CoV-2. Here, we present an efficient computational strategy that combines machine learning (ML)-based models and high-fidelity ensemble docking studies to enable rapid screening of possible therapeutic ligands. Targeting the binding affinity of molecules for either the isolated SARS-CoV-2 S-protein at its host receptor region or the S-protein:human ACE2 interface complex, we screen ligands from drug and biomolecule data sets that can potentially limit and/or disrupt the host−virus interactions. Top scoring one hundred eighty-seven ligands (with 75 approved by the Food and Drug Administration) are further validated by all atom docking studies. Important molecular descriptors ($^{2}\chi_{n}$, topological surface area, and ring count) and promising chemical fragments (oxolane, hydroxy, and imidazole) are identified to guide future experiments. Overall, this work expands our knowledge of small-molecule treatment against COVID-19 and provides a general screening pathway (combining quick ML models with expensive high-fidelity simulations) for targeting several chemical/biochemical problems.


Host ACE-2
F 486
N 501
Which Ligand?
455
SARS-CoV-2
Drugs and Biomolecules Datasets
ML based screening
Validation using AutoDock
75 FDA-approved & Other Promising Ligands

On March 11, 2020, the World Health Organization (WHO) declared the novel coronavirus disease, COVID-19, as a pandemic. More than 15 million people across 203 countries have already been affected by this disease, with more than half a million lives lost globally. In addition, daily lives of millions of people have been impacted because of the mandatory lock-downs observed across the world, let alone the economic cost of this adversity. The COVID-19 disease is caused by a new coronavirus SARS-CoV-2, belonging to the SARS family (SARS-CoV). SARS-CoV-2 has already been sequenced, and several ongoing studies are focusing on understanding its interaction with human cells (or receptors).[1−7] Small molecules or biomolecules with potential therapeutic ability against COVID-19 are also being screened using theoretical and machine learning (ML) methods.[8−12]

Initial reports on SARS-CoV-2, and previous works on the general SARS coronavirus, have suggested close interactions between the viral spike protein (S-protein) of coronavirus and specific human host receptors, such as the angiotensin-converting enzyme 2 (ACE2) receptor. It has been hypothesized that compounds that can weaken interactions between S-protein and ACE2 receptors could limit viral recognition of the host (human) cells and/or disrupt the host−virus interactions. To this end, Smith et al.[8] recently conducted virtual high-throughput screening of nearly 9000 small molecules that bind strongly to either (1) the isolated S-protein of SARS-CoV-2 at its host receptor region (thus, hindering the viral recognition of the host cells) or (2) the S-protein:human ACE2 receptor interface (thus, weakening the host−virus interactions). They successfully identified 77 ligands [24 of which have regulatory approval from the Food and Drug Administration (FDA) or similar agencies] that satisfied one of these two criteria. Despite the vast chemical space (millions to billions of biomolecules) that can be potentially explored, they were severely limited by the number of candidate compounds (nearly 9000) that were considered in their work owing to the high computational cost of the ensemble docking studies employed in their methodology.

Here, we present a general workflow that can be used for efficient screening of molecules with a target binding energy. We deploy this workflow to screen therapeutic molecules using their binding affinity for the S-protein and the S-protein:human ACE2 receptor interface. Specifically, we build on the work of Smith et al.[8] and use their data set generated from autodocking/molecular modeling for training and validating ML models. This allows us to significantly expand the search space and screen millions of potential therapeutic agents against COVID-19. Figure 1a presents the adopted screening workflow, while an illustration of the interface between coronavirus SARS-CoV-2 and the ACE2 receptor in presented

Received: July 24, 2020
Accepted: August 5, 2020
Published: August 5, 2020

Downloaded via Rampi Ramprasad on August 25, 2020 at 17:56:40 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

© XXXX American Chemical Society

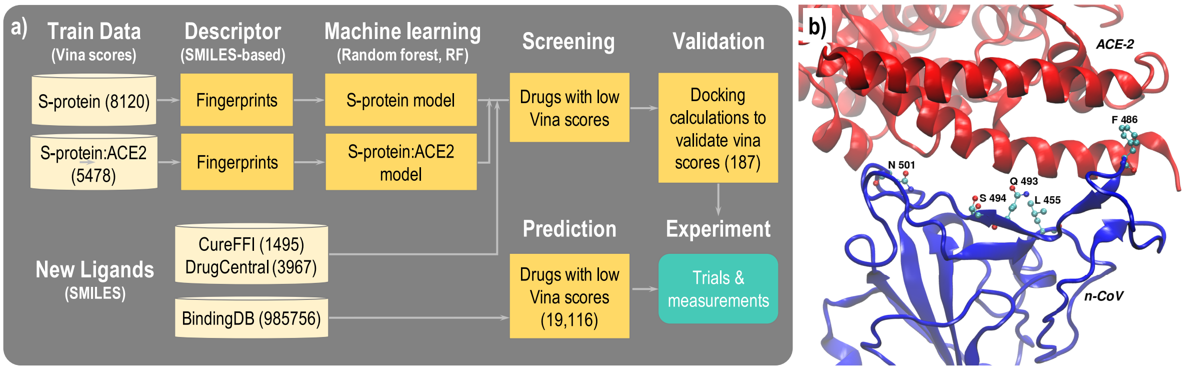


Figure 1. (a) Overview of the workflow adopted to screen drug active ingredients with potential therapeutic capability for COVID-19. The numbers in parentheses indicate the numbers of ligands in various data sets or stages of the workflow. (b) Representation of the interface between the coronavirus n-CoV or SARS-CoV-2 (blue) and the human ACE2 receptor (red). The mutations at a particular virus site are shown in CPK.

in Figure 1b. Two independent random forest (RF) regression models were trained to quickly estimate the Vina scores of a given candidate drug molecule (or ligand) for the isolated S-protein and the S-protein:human ACE2 receptor interface using the data sets provided by Smith et al.[8] The Vina score is an important physicochemical measure of the therapeutic process of a molecule and is used here as a hybrid (empirical and knowledge-based) scoring function that ranks molecular conformations and predicts the free energy of binding based on intermolecular contributions (e.g., steric, hydrophobic, hydrogen bonding, etc.).[13] A set of hierarchical descriptors (or features/fingerprints) that capture different geometric and chemical information at multiple length scales (atomic and morphological) were used to represent the molecules for successful application of the ML models. The models were validated by monitoring their performance on the validation set and against ensemble docking studies for 187 promising candidate ligands identified from the CureFFI and DrugCentral drug data sets, 75 of which are approved by the FDA. A list of ∼19000 biomolecules (from the BindingDB data set) satisfying the same screening criteria is also provided using the developed ML models. On the basis of the feature importance revealed by the ML models and a retrosynthesis analysis of the identified top candidates, we also provide key chemical trends and molecular fragments that are common across the top candidates. We note that this work not only expands our knowledge of potential small-molecule treatment against COVID-19 but also provides a powerful and efficient pathway, i.e., training ML on results of computationally expensive simulations, using ML to cast a wider net, down-selection followed by targeted computational studies, and finally chemical guidelines, for accelerating rational design of molecules/materials for other applications, including catalysis, energy storage, etc.

As depicted in Figure 1a, two training data sets were obtained from Smith et al.,[8] one corresponding to the Vina score of a molecule with the S-protein and other for the S-protein:ACE2 interface complex; among the six receptor conformations, the ones with the best Vina scores were used for training. Each of the data sets contains 9127 molecules from the SWEETLEAD database[14] along with their SMILES representations, which were used as input for our fingerprinting algorithm. For many molecules, the reported Vina scores were extremely high (reaching 1000000 kcal/mol), while those with favorable binding energetics ranged from −7 to 0 kcal/mol. To remove such skewness in the data and train models geared toward identifying favorable molecules, data points with only negative Vina scores were considered in this study. In addition, a few cases whose SMILES representation could not be resolved were filtered out. Overall, this resulted in 5478 and 8120 data points (from the original number of 9127) for the S-protein:ACE2 interface and the isolated S-protein system, respectively. Henceforth, we refer to this cleaned data set as the Smith data set. Its important to note that the Vina score is only an approximation of the experimental binding energies,[15] thereby limiting the accuracy of the results presented here. However, if and when more reliable data become available, perhaps using quantum mechanical treatment,[16] the general scheme presented here could be applied using improved ML models to achieve better accuracy.

To build accurate and reliable ML models, it is important to include relevant features that collectively capture the trends in the Vina scores of different molecules toward S-protein and the S-protein:ACE2 interface complex. The structural as well as physicochemical features should uniquely represent a molecule, be readily available for new cases, and, more importantly, capture the chemistry between the drug molecule and the virus. On the basis of our experience, a three-level hierarchical set of features capturing different geometric and chemical information about ligands at multiple length scales (atomic and morphological) were considered. Fingerprint details are provided in the Supporting Information. We note that the fingerprinting as well as the screening approach presented here can be used in other applications such as catalysis and energy storage. For instance, a key descriptor of the oxygen evolution reaction (OER) is the energy for binding of oxygen to the catalyst surface.[17] A similar ML procedure can be adopted to screen efficient OER catalysts by replacing Vina scores with oxygen binding energies.

The random forest (RF) regression algorithm, as implemented in scikit-learn,[18] was used to train the two Vina score models (S-protein and S-protein:ACE2 interface). RF is an ensemble of decision trees, which averages predictions from a large group of "weak models" to overall result in a better prediction. The RF hyperparameters, i.e., the number of weak estimators, were estimated by maximizing the validation error during 5-fold cross-validation (CV). The model performance was evaluated using the root-mean-square error (RMSE), mean absolute error (MAE), and correlation coefficient (R^2). To estimate prediction errors on unseen data, learning curves were generated by varying the sizes of the training and test sets, with results included in the Supporting Information. Statisti-

https://dx.doi.org/10.1021/acs.jpclett.0c02278

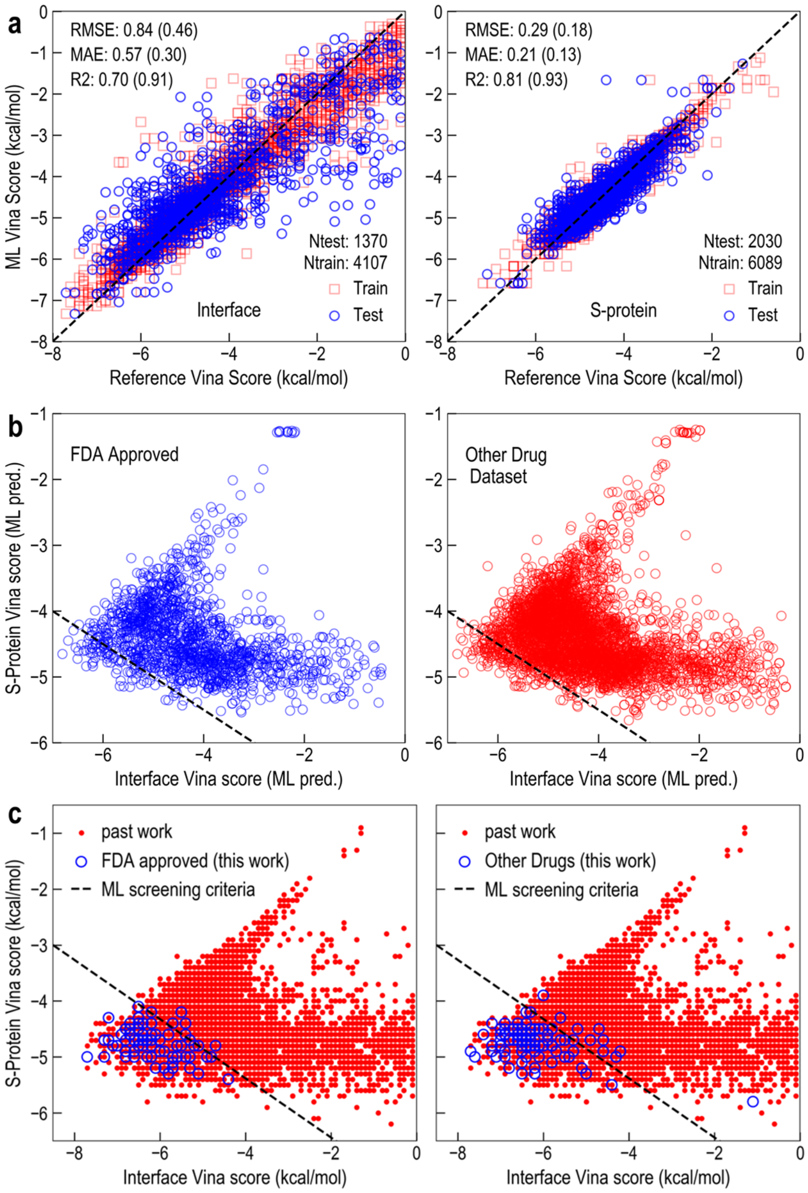
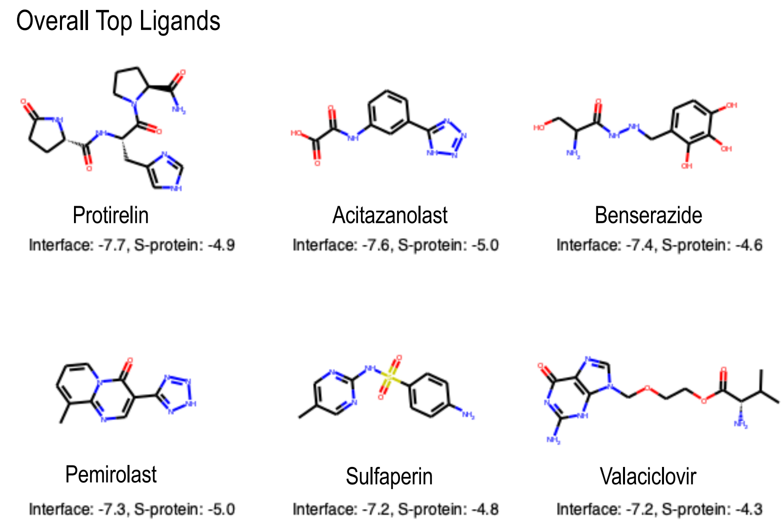


Figure 2. (a) Parity plot of the S-protein and interface ML models for the training and the test set, demonstrating the good prediction accuracy achieved by both ML models. Different error metrics for the test and training (within parentheses) set are also included. (b) ML predictions of Vina scores (in kcal/mol) for the isolated S-protein and S-protein:ACE2 receptor interface for FDA-approved (left) and other drug (right) candidates obtained from CureFFI and DrugCentral databases. Candidates with predictions below the dashed line were selected for further validation using docking studies. (c) Vina scores for the 187 selected candidates (blue) using the docking calculations. For comparison, previously considered candidates from an exhaustive past work are also included (red).

cally meaningful results were obtained by averaging over 10 different random test-train split. The final ML models used for prediction on the CureFFI, DrugCentral, and BindingDB data sets were trained on the entire Smith data set using 5-fold CV and consisted of 400 and 700 estimators for the S-protein and the interface data sets, respectively.

To validate our ML models, we performed docking calculations of the top candidates identified by the models based on their low Vina scores. The setups of the docking studies were kept consistent with the work of Smith et al.,[8] including the structure of the docking receptors (i.e., six conformations each for the S-protein:ACE2 interface complex and the isolated S-protein), and the binding search space of 1.2 nm × 1.2 nm × 1.2 nm. More computational details are provided in the Supporting Information. We note that 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), which is illustrated in Figure 1b. Docking studies are focused on this binding pocket region for evaluation of the binding affinities of different molecules. For each candidate, the docking procedure finds the top 10 optimized docking configurations and selects the one with the best Vina score.

While the Smith data set[8] was used to train and validate the ML models, three additional drug data sets were used to make predictions and identify ligand candidates that show high binding affinity for the viral S-protein or the S-protein:ACE2 interface. These include (1) an all FDA-approved CureFFI data set,[19] (2) a data set of common active ingredients from DrugCentral,[20] and (3) a BindingDB data set[21] of small


Overall Top Ligands
Protirelin
Interface: -7.7, S-protein: -4.9
Acitazanolast
Interface: -7.6, S-protein: -5.0
Benserazide
Interface: -7.4, S-protein: -4.6
Pemirolast
Interface: -7.3, S-protein: -5.0
Sulfaperin
Interface: -7.2, S-protein: -4.8
Valaciclovir
Interface: -7.2, S-protein: -4.3

Top FDA Approved Ligands

ID	General Name	Interface Vina Score	S-protein Vina Score	Source
1	Pemirolast	-7.3	-5	CureFFI
2	Sulfamethoxazole	-7.2	-4.7	CureFFI
3	Valaciclovir	-7.2	-4.3	CureFFI
4	Sulfamerazine	-7.1	-4.8	CureFFI
5	Tazobactam	-7	-4.8	CureFFI
6	Nitrofurantoin	-7	-4.8	CureFFI

Top Other Ligands

ID	General Name	Interface Vina Score	S-protein Vina Score	Source
1	Protirelin	-7.7	-4.9	DrugCentral
2	Acitazanolast	-7.6	-5	DrugCentral
3	Benserazide	-7.4	-4.6	DrugCentral
4	Sulfaperin	-7.2	-4.8	DrugCentral
5	Succinylsulfathiazole	-7.2	-4.4	DrugCentral
6	Uridine triphosphate	-7.2	-4.9	DrugCentral

Figure 3. Top candidates identified from this work along with their Vina scores for the S-protein:ACE2 interface (labeled, interface) and the S-protein systems using the ensemble docking studies.

molecules. SMILES representations of molecules were obtained from each of these data sets and, with some unprocessed candidates removed, resulted in 1495, 3967, and 985756 entries, respectively. The CureFFI data set consists of ligands approved by the FDA and specifically contains central nervous system drugs. DrugCentral is an open-access online drug compendium. It integrates the structure, bioactivity, regulatory, and pharmacologic actions and indications for active pharmaceutical ingredients approved by the FDA and other regulatory agencies. The BindingDB is a publicly accessible database based on measured binding affinities of drug-like molecules interacting with various protein targets and consists of >1 million entries of binding data and molecule data sets. The first two data sets were exclusively used to validate the ML models against docking studies, while the BindingDB data set was used for only ML predictions.

Figure 2a presents the performance results of the S-protein and S-protein:ACE2 interface RF models for the case in which 75% of Smith's data set was used for training (with 5-fold CV) and the remaining 25% as the test set. The overall model performance of the test set is a good indicator of the expected errors on new candidate drugs with unknown Vina scores. Both models can be seen to have good performance on the test set; a MAE of 0.21 kcal/mol was achieved for the S-protein model, while the S-protein:ACE2 model was only marginally worse with a MAE of 0.57 kcal/mol. Both of these errors are well within typical chemical accuracy of 1 kcal/mol, and we believe the ML models are acceptable for screening purposes. Even for the S-protein:ACE2 model, relatively smaller errors are observed for cases with low Vina scores, which are particularly more relevant to this study. See the Supporting Information for more detailed validation of the ML models using learning curves, including error convergence studies on the training and test sets.

These results clearly indicate that the developed surrogate ML models could be used to quickly screen new ligand candidates with low S-protein or S-protein:ACE2 interface Vina scores without exclusively performing computationally demanding docking studies. To this end, we use the ML models to make predictions for the FDA-approved active ingredients in the CureFFI data set and other ligands from the DrugCentral data set, presented in Figure 2b. Because the true Vina scores of these ligands are not known, here we show only their ML predictions. It has been hypothesized that a ligand could be effective against coronavirus if it either forms S-protein:ACE2 interface–ligand binding complexes (low S-protein:ACE2 Vina score) to disrupt the host–virus interaction or binds to the receptor recognition region of the S-protein (low S-protein Vina score) to reduce the extent of viral recognition of the host. Thus, we define a simple screening criterion for selecting top candidates having low Vina scores on both accounts. The dashed line in Figure 2b depicts the chosen screening criteria (given by the equation $y < -\frac{x}{2} - 7.5$, where x and y represent Vina scores for the S-protein:ACE2 interface–ligand complex and the S-protein–ligand system, respectively). We note that 187 ligands were selected, from which 80 are approved by the FDA (CureFFI data set), 107 are other drugs (DrugCentral data set), and 29 are common to the Smith data set. A list of all 187 drugs (including their generic name and SMILES representation) and their Vina score predictions are provided in the Supporting Information. In contrast to the screening criteria used here, Smith et al. used relatively higher threshold values: S-protein score < -6.2 or interface score < -7 kcal/mol. Because no molecule was found to satisfy the two criteria together, we adopted the selection definition as discussed above. In addition, we caution that molecules with a high level of binding to the interface may unintentionally stabilize it rather than disrupting the underlying interactions. Unfortunately, this cannot be known *a priori* and can only be resolved using experiments or exceptionally long time scale molecular dynamics simulations.

Results for the ensemble docking studies on the selected 187 drug candidates are presented in Figure 2c. For comparison, results from the Smith data set are also included. The purpose of these computations was threefold. First, a more accurate estimate of the Vina scores was obtained from these high-fidelity computations for the identified promising candidates; second, they provided new data points for further validation of the ML models, and third, for the 29 common candidate ligands (common to our top list and that of Smith), they help us to validate our docking studies against those performed in

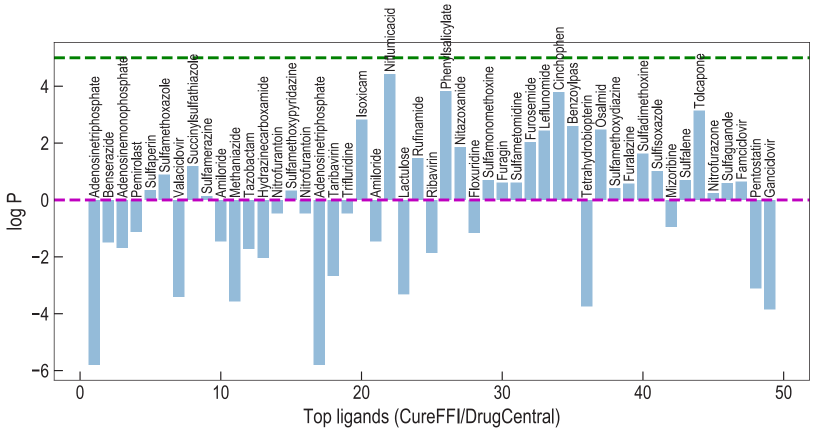


Figure 4. Use of physicochemical properties to assess the therapeutic prowess of ligands. 1-Octanol/water partition coefficients (log P) of the top candidates. These values were obtained from www.chemspider.com. The green dashed line indicates a log P value of 5. Most of the screened top candidates have log P values of <5.

the Smith paper[8] (see Figure S3 for a detailed comparison). From Figure 2c, it is evident that the ML models indeed helped us to screen candidates with favorable Vina scores; almost all screened candidates can be seen to be below the ML screening criterion line (dashed line), while only 12 of the identified 187 candidates were found to have Vina scores of >0 and did not show any binding affinity for the S-protein:ACE2 interface complex; such cases have relatively much higher Vina scores (>10) and are excluded from the plots for better readability. Thus, 175 of 187 (94%) of the screened candidates were indeed favorable. In comparison, Smith et al. needed to perform expensive docking studies for a large set of candidates, with many falling outside the screening boundary. This not only captures the efficiency of the procedure adopted here, i.e., the use of cheap surrogate models for quick screening followed by expensive high-fidelity docking studies for validation, but also provides further validation of the prediction accuracy of the developed ML models. Parity plots directly comparing the Vina score predictions from the ML models against their respective docking simulation results and example illustrations of the S-protein:ACE2 interface−ligand complex for the top candidates are included in the Supporting Information.

More importantly, our trained ML model predicts several ligands (including several FDA-approved active ingredients) with favorable Vina scores. The top six among the 187 candidates are presented in Figure 3 (see the Supporting Information for a complete list). The top FDA-approved ligand candidates include pemirolast (INN), which is a mast cell stabilizer used as an antiallergic drug therapy. It is marketed under the trade names Alegysal and Alamast. Sulfamethoxazole (SMZ or SMX), another FDA-approved ligand, is an antibiotic used for bacterial infections such as urinary tract infections, bronchitis, and prostatitis. Valaciclovir is another top candidate identified from our screening and is an antiviral drug used to treat herpes virus infections, including shingles, cold sores, genital herpes, and chickenpox. Sulfanilamide is used typically as an antibacterial agent to treat bronchitis, prostatitis, and urinary tract infections. Tzaobactum is another FDA-approved antibiotic and is typically combined with piperacillin to treat antibacterial infections such as cellulitis, diabetic foot infections, appendicitis, and postpartum endometritis infection. Nitrofurantoin is also an antibiotic and used to treat urinary tract infections.

Among the non-FDA-approved ligands, we find that the top candidate is Protirelin, which is a synthetic analogue of the endogenous peptide thyrotropin-releasing hormone (TRH). Benserazide (also called serazide) is another top ligand and is a peripherally acting aromatic L-amino acid decarboxylase or DOPA decarboxylase inhibitor that is used for Parkinson's disease. Other top candidates include sulfaperin (or sulfaperine), which is a sulfonamide antibacterial agent, and succinylsulfathiazole, which is a sulfonamide used as an intestinal bacteriostatic agent. Interestingly, one of the top candidates to emerge from our screening is uridine triphosphate (UTP), which is a nucleotide triphosphate and source of energy or an activator of substrates in metabolic reactions.

Once the top candidates are identified using our search procedure, we analyzed them using thermodynamic criteria other than binding energies to further screen the ligand candidates. For instance, other metrics developed by Lipinski and co-workers[22,23] could be used to understand the efficacy of a therapeutic molecule. Figure 4 shows the log P of the top 50 candidates identified (based on the lowest value of the Vina scores) from the CureFFI and DrugCentral databases. A ligand is most likely to have poor absorption when its *n*-octanol/water partition coefficient (log P) is >5, its molecular weight (MW) is >500, the number of H bond donors is >5, and the number of H bond acceptors is >10. Most of the top 50 ligands can be seen to have log P values of <5, which is consistent with Lipinski rules of five. In addition, the molecular weights of the compounds are <500 Da, as provided in the Supporting Information along with other properties, such as Henry's constant and the number of hydrogen bond acceptors and donors.

Henry's constant (or log H) measures the solubility of the compound in water. For a drug to be taken up by the cellular membrane, it is desirable for the drug to be soluble in water. The more negative Henry's constant, the more soluble the drug in the aqueous phase. However, a balance between desirable partitioning between the membrane and aqueous phase is generally sought. Thus, as presented in Table 1, the identified top candidates continue to satisfy all of the additional criteria mentioned above. Importantly, we note that more such constraints can be introduced in future work to further screen desirable candidate ligands. For instance, molecules with log P values of <0 are known to have high

Table 1. *n*-Octanol/Water Partition Coefficients (log P), Henry's Constants (log H), Average Molecular Weights, and Numbers of Hydrogen Bond Donors and Acceptors for the Top Ligands Identified in This Work[a]

	log P	log H	MW (Da)	no. of H bond donors	no. of H bond acceptors
FDA-Approved Ligands					
pemirolast	−1.12	−12.313	228.21	1	7
sulfamethoxazole	0.89	−10.408	253.278	3	6
valaciclovir	−3.41	−17.578	324.336	5	10
sulfamerazine	0.14	−8.145	264.304	3	3
tazobactam	−1.72	−14.714	300.291	1	9
Other Ligands					
proterelin	−2.46	−22.799	362.384	5	10
acitazanolast	−1.95	−16.014	233.184	3	8
sulfaperin	0.34	−8.145	264.304	3	6
benserazide	−1.49	−28.420	257.243	8	8
succinyl sulfathiozole	1.18	−19.117	355.389	3	8
uridine triphosphate	−4.09	−38.070	484.141	7	17

[a]These values were obtained from www.chemspider.com.

affinity for aqueous media and are poorly absorbed by the lipid bilayer of the cellular membranes. Many of the top candidates can be seen to fall under this category.

Beyond serving as a more computationally efficient alternative to drug docking studies, learned RF models can also be utilized to mine important chemical trends and extract simple chemical rules from the data. In RF, the relative importance of a feature can be defined using the relative rank (or depth) of that feature when used as a decision node in a tree, because features used at the top of a tree contribute to the final prediction for a larger fraction of the input samples. On the basis of this philosophy, we provide a list of the top 20 features that were found to be most relevant for the S-protein and the S-protein:ACE2 interface models in the Supporting Information. Importantly, we found that the $^{2}\chi_{n}$ score of a molecule correlates very well (with Pearson correlation coefficient R^2 = −0.67) with its S-protein Vina score; the higher the $^{2}\chi_{n}$ score, the lower the Vina score of the molecule:S-protein complex. As discussed here,[24,25] $^{2}\chi_{n}$ encodes the atomic identity and connectivity in a molecule by representing it as a graph. A variety of molecular quantum numbers (MQNs) were also found to be highly relevant: those that captured the number of five- or six-member rings, the topological surface area, cyclic trivalent and tetravalent nodes, and nodes and edges shared by more than two rings. The number of aliphatic rings was also among the important descriptors.

With the idea of identifying common molecular motifs that bind well to the S-protein and the interface systems, we performed a retrosynthesis analysis of the identified top candidates. The concept of breaking of retrosynthetically interesting chemical substructures (BRICS),[26] as implemented in RDKit,[25] was used to obtain common molecular fragments for both the CureFFI and DrugCentral data sets, as well as the screened 175 candidates with low Vina scores. Figure 5 compares the frequency (normalized with respect to the maximum value) of occurrence of a few representative fragments in the identified top ligands against that in the two drug data sets (see the Supporting Information for a complete list). A fragment displaying a higher (lower) occurrence frequency among top candidate ligands suggests it plausibly promotes (reduces) binding to the two systems. In particular, fragments involving oxolane-, hydroxy-, imidazole-, piperidine-, and benzenesulfonate-derived groups (also shown pictorially) are expected to promote binding of the ligand to the S-protein and the interface systems. In fact, an analysis of the docking poses shows that, in general, the N-ring-containing ligands interact with the side chain and backbone of Q493 and S494 sites (two of the five mutating sites from the SARS-CoV 2002 virus). For instance, the azole nitrogen in pemirolast interacts via a medium hydrogen bond with the side chain of Q493. The pyrimidine moiety in sulfamerazine interacts strongly with the side chain Q493 and the backbone oxygen of S494. These trends (also see the Supporting Information) suggest that the interactions with Q493 and S494 of the SARS-CoV-2 may be partly responsible for the efficacy of a ligand. The identified chemical fragments are also consistent with the important molecular descriptors mentioned above, which also

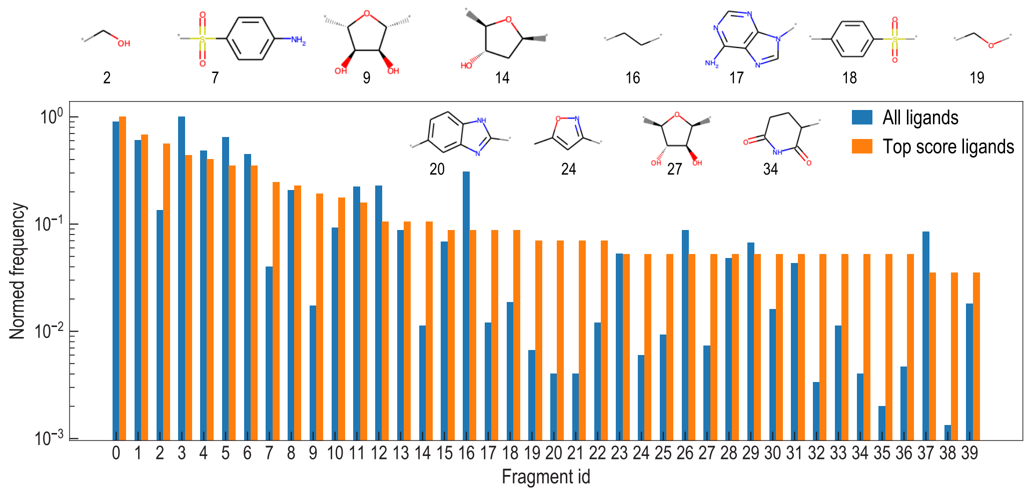


Figure 5. Normalized frequency of occurrence of a few representative molecular fragments in the CureFFI and DrugCentral drug data sets (blue) and the screened top 175 ligands (orange). Exemplary fragments with large frequency deviations in the two scenarios are displayed pictorially, along with their identifiers. Open bonds in the fragments are denoted by asterisks.

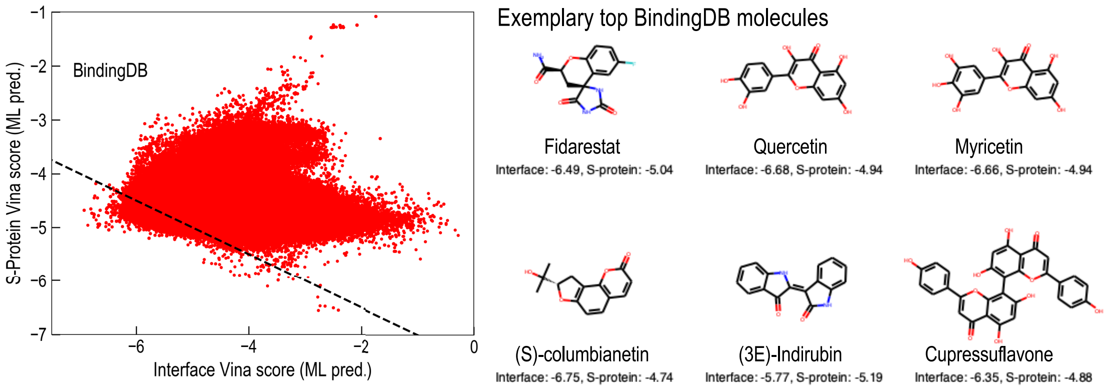


Figure 6. Vina score predictions for the isolated S-protein and S-protein:ACE2 receptor complex for all of the molecules in the BindingDB data set using ML models. More than 19000 molecules were found to satisfy the chosen screening criteria, shown using the dashed line in the plot.

involve the number of five- or six-member rings, cyclic trivalent and tetravalent nodes, etc. The identified top fragments in Figure 5 can drive more rigorous quantum mechanical studies of the interaction of these limited (and practically viable) cases, besides helping with the rational design of new drugs for COVID-19.

Next, we significantly expanded the search space of candidate molecules and made predictions for roughly 1 million molecules in the BindingDB data set, with the Vina score predictions presented in Figure 6. Nearly 19000 molecules were found to satisfy the previously chosen screening criteria (see the Supporting Information for the complete list), and a few exemplary cases are illustrated in the right panel of Figure 6. These results clearly demonstrate the power and efficiency of using surrogate models for preliminary screening. For instance, the docking studies for the identified 187 candidate active ingredients were completed in a period of around 2 days. In contrast, Vina score predictions from the ML model for the entire BindingDB data set were obtained within a day using similar computational resources, including the time required for fingerprinting and making the model predictions. Evidently, our ML strategy is efficiently able to screen millions of candidate biomolecules and make useful suggestions to aid the decision making process for expert biologists and medical professionals, who can focus on a much narrower subset of screened candidates and make more informed decisions by incorporating additional medical insights. More robust high-fidelity computations followed by synthesis and trial experiments should be performed to confirm the validity of these selected molecules.

Among the screened non-FDA-approved biomolecules, the top candidates include fidarestat (SNK-860), which is an aldose reductase inhibitor and is under investigation for the treatment of diabetic neuropathy. Quercetin is a plant flavonol from the flavonoid group of polyphenols, which also displayed high Vina scores among the screened candidates. Other top candidates include myricetin, which is a member of the flavonoid class of polyphenolic compounds, with antioxidant properties; S-columbianetin, which is used as an anti-inflammatory; indirubin, which has anti-inflammatory and anti-angiogenesis properties in vitro; and cupressuflavone, which has anti-inflammatory and analgesic properties.

In conclusion, we present an efficient virtual screening strategy for identifying ligands that can potentially limit and/or disrupt the host−virus interactions of SARS-CoV-2. Our hypothesis is that ligands that bind strongly to the isolated S-protein at its host (human) receptor region and to the S-protein:human ACE2 interface complex are likely to be the most effective. Our high-throughput screening strategy is based on using a combination of ML and high-fidelity docking studies to identify candidates that display such high binding affinities. We first train random forest models on results of computationally expensive studies and subsequently use the validated ML model to search a much larger chemical space (approximately thousands of FDA-approved ligands and approximately a million of biomolecules). Vina scores for the identified top ligands (based on ML predictions) are further confirmed using expensive docking studies, resulting in the identification of 75 FDA-approved and 100 other ligands from drug data sets. In addition, important chemical trends in terms of molecular fragments (e.g., oxolane-, imidazole-, and benzenesulfonate-derived groups) promoting binding affinities for the S-protein and the interface systems and determination of important molecular descriptors (e.g., ${}^{2}\chi_{n}$ and topological surface area) having strong correlations with binding affinities were also revealed. Finally, we note that the general scheme of ML-assisted discovery presented here, involving the use of surrogate models to search large chemical spaces or mine chemical guidelines through molecular descriptors and fragments, is equally useful in other areas of catalysis, energy storage, or corrosion, beyond accelerating the therapeutic cure of diseases.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c02278.

Details about molecular fingerprints, docking simulations, random forest model performance, validation against past docking studies, binding poses for the S-protein:ACE2 interface−ligand complex, and identified top ligand descriptors (PDF)

Other supporting data files (ZIP)

■ AUTHOR INFORMATION

Corresponding Authors

Rohit Batra − *Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-1098-7035; Email: rohitbatra1989@gmail.com

Subramanian K.R.S. Sankaranarayanan − *Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States; Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, Illinois 60607, United States;* orcid.org/0000-0002-9708-396X; Email: skrssank@uic.edu, skrssank@anl.gov

Authors

Henry Chan − *Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States; Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, Illinois 60607, United States;* orcid.org/0000-0002-8198-7737

Ganesh Kamath − *Dalzielfiver LLC, El Sobrante, California 94803, United States*

Rampi Ramprasad − *School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States;* orcid.org/0000-0003-4630-1565

Mathew J. Cherukara − *Center for Nanoscale Materials, Argonne National Laboratory, Lemont, Illinois 60439, United States;* orcid.org/0000-0002-1475-6998

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c02278

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Use of the Center for Nanoscale Materials and Advanced Photon Source, both Office of Science user facilities, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. The authors thank Naina Zachariah for useful discussions and for reviewing the manuscript. SKRS is thankful for a UIC Start-up faculty grant for supporting this work. The authors dedicate this research to the millions of lives affected by the COVID pandemic.

■ REFERENCES

(1) Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265−269.

(2) Yu, W.-B.; Tang, G.-D.; Zhang, L.; Corlett, R. T. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2) using whole genomic data. *Zool. Res.* **2020**, *41*, 247−257.

(3) Tang, X.; Wu, C.; Li, X.; Song, Y.; Yao, X.; Wu, X.; Duan, Y.; Zhang, H.; Wang, Y.; Qian, Z.; et al. On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **2020**, *7*, 1012−1023.

(4) Sun, P.; Lu, X.; Xu, C.; Sun, W.; Pan, B. Understanding of COVID-19 based on current evidence. *J. Med. Virol.* **2020**, *92*, 548−551.

(5) Bai, Y.; Yao, L.; Wei, T.; Tian, F.; Jin, D.-Y.; Chen, L.; Wang, M. Presumed asymptomatic carrier transmission of COVID-19. *Jama* **2020**, *323*, 1406−1407.

(6) Gralinski, L. E.; Menachery, V. D. Return of the Coronavirus: 2019-nCoV. *Viruses* **2020**, *12*, 135.

(7) Wan, Y.; Shang, J.; Graham, R.; Baric, R. S.; Li, F. Receptor recognition by novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS. *J. Virol.* **2020**, *94* (7), e00127-20.

(8) Smith, M.; Smith, J. C. Repurposing therapeutics for COVID-19: Supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface. *ChemRxiv* **2020**.

(9) Nguyen, D.; Gao, K.; Chen, J.; Wang, R.; Wei, G. Potentially highly potent drugs for 2019-nCoV. *BioRxiv* **2020**, DOI: 10.1101/2020.02.05.936013.

(10) Xu, Z.; Peng, C.; Shi, Y.; Zhu, Z.; Mu, K.; Wang, X.; Zhu, W. Nelfinavir was predicted to be a potential inhibitor of 2019-nCoV main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *BioRxiv* **2020**, DOI: 10.1101/2020.01.27.921627.

(11) Beck, B. R.; Shin, B.; Choi, Y.; Park, S.; Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (2019-nCoV) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 784.

(12) Bung, N.; Krishnan, S. R.; Bulusu, G.; Roy, A. De novo design of new chemical entities (NCEs) for SARS-CoV-2 using artificial intelligence. *ChemRxiv* **2020**.

(13) 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.* **2009**, *31*, 455−461.

(14) 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*, e79568.

(15) Nguyen, N. T.; Nguyen, T. H.; Pham, T. N. H.; Huy, N. T.; Bay, M. V.; Pham, M. Q.; Nam, P. C.; Vu, V. V.; Ngo, S. T. Autodock Vina adopts more accurate binding poses but Autodock4 forms better binding affinity. *J. Chem. Inf. Model.* **2020**, *60*, 204−211.

(16) Mardirossian, N.; Wang, Y.; Pearlman, D. A.; Chan, G. K.; Shiozaki, T. Novel algorithms and high-performance cloud computing enable efficient fully quantum mechanical protein-ligand scoring. *arXiv* **2020**, 2004.08725.

(17) Sen, F.; Kinaci, A.; Narayanan, B.; Gray, S.; Davis, M.; Sankaranarayanan, S.; Chan, M. Towards accurate prediction of catalytic activity in IrO2 nanoclusters via first principles-based variable charge force field. *J. Mater. Chem. A* **2015**, *3*, 18970−18982.

(18) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(19) CureFFI. https://www.cureffi.org/2013/10/04/list-of-fda-approved-drugs-and-cns-drugs-with-smiles/.

(20) Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C. G.; Yang, J. J.; Mathias, S. L.; Nelson, S. J.; Oprea, T. I. DrugCentral: Online drug compendium. *Nucleic Acids Res.* **2017**, *45*, D932−D939.

(21) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.

(22) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(23) Lipinski, C. A. Lead-and drug-like compounds: The rule-of-five revolution. *Drug Discovery Today: Technol.* **2004**, *1*, 337−341.

(24) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. *Rev. Comput. Chem.* **2007**, 367−422.

(25) RDKit open source toolkit for cheminformatics. http://www.rdkit.org/.

(26) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem* **2008**, *3*, 1503−1507.