THE JOURNAL OF PHYSICAL CHEMISTRY Letters

pubs.acs.org/JPCL

Perspective

Repositioning of 8565 Existing Drugs for COVID-19

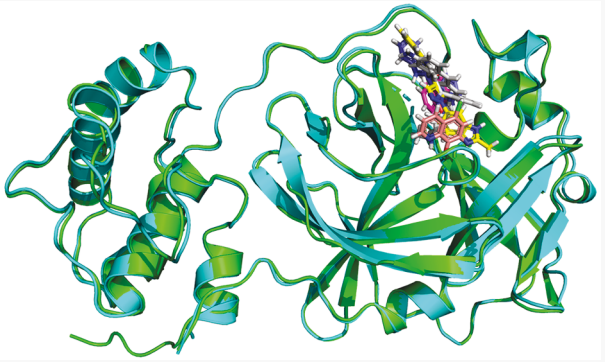
Kaifu Gao, Duc Duy Nguyen, Jiahui Chen, Rui Wang, and Guo-Wei Wei*

Cite This: J. Phys. Chem. Lett. 2020, 11, 5373−5382

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected over 7.1 million people and led to over 0.4 million deaths. Currently, there is no specific anti-SARS-CoV-2 medication. New drug discovery typically takes more than 10 years. Drug repositioning becomes one of the most feasible approaches for combating COVID-19. This work curates the largest available experimental data set for SARS-CoV-2 or SARS-CoV 3CL (main) protease inhibitors. On the basis of this data set, we develop validated machine learning models with relatively low root-mean-square error to screen 1553 FDA-approved drugs as well as another 7012 investigational or off-market drugs in DrugBank. We found that many existing drugs might be potentially potent to SARS-CoV-2. The druggability of many potent SARS-CoV-2 3CL protease inhibitors is analyzed. This work offers a foundation for further experimental studies of COVID-19 drug repositioning.



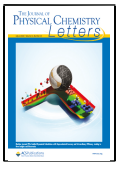
Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) appeared in Wuhan, China, in late December 2019 and has rapidly spread around the world. By June 11, 2020, over 7.1 million individuals were infected, and more than 408 000 fatalities had been reported. Currently, there is no specific antiviral drug for this epidemic. It is worth noting that recently, an experimental drug, Remdesivir, has been recognized as a promising anti-SARS-CoV-2 drug. However, the high experimental value of IC_{50} (11.41 μM)[1] indicates that it must be used in a large dose in treating COVID-19, which is subject to side effects.

Considering the severity of this widespread dissemination and health threats, panicked patients misled by media flocked to pharmacies for Chinese medicine herbs, which were reported to "inhibit" SARS-CoV-2, despite no clinical evidence supporting the claim. Although there is also no evidence for Chloroquine's claimed curing effect, some desperate people take it as "prophylactic" for COVID-19. Many researchers are engaged in developing anti-SARS-CoV-2 drugs.[2,3] However, new drug discovery is a long, costly, and rigorous scientific process. A more effective approach is to search for anti-SARS-CoV-2 therapies from existing drug databases.

Drug repositioning (also known as drug repurposing), which concerns the investigation of existing drugs for new therapeutic target indications, has emerged as a successful strategy for drug discovery because of the reduced costs and expedited approval procedures.[4−6] Several successful examples reveal its great value in practice: Nelfinavir, initially developed to treat the human immunodeficiency virus (HIV), is now being used for cancer treatments. Amantadine was first designed to treat the influenza caused by type A influenza viral infection and is being used for the Parkinson's disease.[7] In recent years, the rapid growth of drug-related data sets, as well as open data initiatives, has led to new developments for computational drug repositioning, particularly structural-based drug repositioning (SBDR). Machine learning, network analysis, and text mining and semantic inference are three major computational approaches commonly applied in drug repositioning.[8] The rapid accumulation of genetic and structural databases (https://www.rcsb.org/ and https://www.ncbi.nlm.nih.gov/genbank/), the development of low-dimensional mathematical representations of complex biomolecular structures,[9] and the availability of advanced deep learning algorithms have made machine learning-based drug repositioning a promising

New drug discovery is a long, costly, and rigorous scientific process. A more effective approach is to search for anti-SARS-CoV-2 therapies from existing drug databases.

Received: May 21, 2020
Accepted: June 16, 2020
Published: June 16, 2020



approach.[8] Because of the urgent need for anti-SARS-CoV-2 drugs, a computational drug repositioning is one of the most feasible strategies for discovering SARS-CoV-2 drugs.

In SBDR, one needs to select one or a few effective targets. Study shows that the SARS-CoV-2 genome is very close to that of the severe acute respiratory syndrome (SARS)-CoV.[10] The sequence identities of SARS-CoV-2 3CL protease, RNA polymerase, and the spike protein with corresponding SARS-CoV proteins are 96.08%, 96%, and 76%, respectively[11] (see Figure S1). We, therefore, hypothesize that a potent SARS 3CL protease inhibitor is also a potent SARS-CoV-2 3CL protease inhibitor. Unfortunately, there is no effective SARS therapy at present. Nevertheless, the X-ray crystal structures of both SARS and SARS-CoV-2 3CL proteases have been reported.[12,13] Additionally, the binding affinities of SARS-CoV or SARS-CoV-2 3CL protease inhibitors from single-protein experiments are available in various databases or the original literature. Moreover, the DrugBank contains about 1600 drugs approved by the U.S. Food and Drug Administration (FDA) as well as more than 7000 investigational or off-market drugs.[14] The aforementioned information provides a sound basis for developing an SBDR machine learning model for SARS-CoV-2 3CL protease inhibition. It is worth clarifying that SBDR machine learning models are driven by data and do not explicitly form the energy terms related to some biophysical characteristics such as electrostatics and hydrogen bonding. Instead, these biophysical interactions are implicitly encoded in the fingerprints, and their impacts on the binding affinity are regulated by machine learning scoring functions.

In responding to the pressing need for anti-SARS-CoV-2 medications, we have carefully collected 314 bonding affinities for SARS-CoV or SARS-CoV-2 3CL protease inhibitors, which is the largest set available to date for this system. Machine learning models are built for these data points.

Unlike most earlier COVID-19 drug repositioning works that did not provide a target-specific cross-validation test, we have carefully optimized our machine learning model with a 10-fold cross-validation test on SARS-CoV-2 3CL protease inhibitors. We achieve a Pearson correlation coefficient of 0.78 and a root-mean-square error (RMSE) of 0.79 kcal/mol on the test sets of 10-fold cross validation tasks, which is much better than that of similar machine learning models for standard training sets in the PDBbind database (around 1.9 kcal/mol).[15] We systematically evaluate the binding affinities (BAs) of 1553 FDA-approved drugs as well as 7012 investigational or off-market drugs in the DrugBank by our 2D-fingerprint-based machine learning model. In addition, a three-dimensional (3D) pose predictor named MathPose[16] is also applied to predict the 3D binding poses. With these models, we report the top 20 potential anti-SARS-CoV-2 3CL inhibitors from the FDA-approved drugs and another top 20 from investigational or off-market drugs. We also discuss the druggability of some potent inhibitors in our training set. The information provides timely guidance for the further development of anti-SARS-CoV-2 drugs.

With the SARS-CoV-2 3CL protease as the target, we predict the binding affinities of 1553 FDA-approved drugs using our machine learning predictor. Given these predicted affinities, the top 20 potential SARS-CoV-2 inhibitors from the FDA-approved drugs are shown in Table 1. We also supply the corresponding IC_{50} (μM) derived from the binding affinity X (kcal/mol) via the following conversion: $IC_{50} = 10^{X/1.3633} \times 10^{-6}$. A complete list of the predicted values for 1553 FDA-approved drugs is given in the Supporting Tables (FDA_approved) in Supporting Information.

Table 1. Summary of the Top 20 Potential Anti-SARS-CoV-2 Drugs from 1553 FDA-Approved Drugs with Their Predicted Binding Affinities (unit: kcal/mol), IC_{50} (μM), and Corresponding Brand Names

DrugID	name	brand name	predicted binding affinity	IC_{50}
DB01123	Proflavine	Bayer Pessaries, Molca, Septicide	−8.37	0.72
DB01243	Chloroxine	Capitrol	−8.24	0.89
DB08998	Demexiptiline	Deparon, Tinoran	−8.14	1.06
DB00544	Fluorouracil	Adrucil	−8.11	1.11
DB03209	Oteracil	Teysuno	−8.09	1.16
DB13222	Tilbroquinol	Intetrix	−8.08	1.18
DB01136	Carvedilol	Coreg	−8.06	1.22
DB01033	Mercaptopurine	Purinethol	−8.04	1.26
DB08903	Bedaquiline	Sirturo	−8.02	1.29
DB00257	Clotrimazole	Canesten	−8.00	1.35
DB00878	Chlorhexidine	Betasept, Biopatch	−8.00	1.35
DB00666	Nafarelin	Synarel	−8.00	1.35
DB01213	Fomepizole	Antizol	−7.98	1.39
DB01656	Roflumilast	Daxas, Daliresp	−7.97	1.41
DB00676	Benzyl benzoate	Ascabin, Ascabiol, Ascarbin, Tenutex	−7.96	1.45
DB06663	Pasireotide	Signifor	−7.95	1.47
DB08983	Etofibrate	Lipo Merz Retard, Liposec	−7.94	1.48
DB06791	Lanreotide	Somatuline	−7.94	1.48
DB00027	Gramicidin D	Neosporin Ophthalmic	−7.94	1.48
DB00730	Thiabendazole	Mintezol, Tresaderm, and Arbotect	−7.93	1.51

We briefly describe the top 10 predicted potential anti-SARS-CoV-2 drugs from the FDA-approved set. The most potent one is Proflavine, an acriflavine derivative. It is a disinfectant bacteriostatic against many Gram-positive bacteria. Proflavine is toxic and carcinogenic in mammals and so it is used only as a surface disinfectant or for treating superficial wounds. Under the circumstance of the SARS-CoV-2, this drug might be used to clean skin or SARS-CoV-2 contaminated materials, offering an extra layer of protection. The second drug is Chloroxine, also an antibacterial drug, which is used in infectious diarrhea, disorders of the intestinal microflora, giardiasis, and inflammatory bowel disease. It is notable that this drug belongs to the same family with Chloroquine, which was once considered for anti-SARS-CoV-2. However, according to our prediction, Chloroquine is not effective for SARS-CoV-2 3CL protease inhibition (BA: −6.92 kcal/mol). The third one, Demexiptiline, a tricyclic antidepressant, acts primarily as a norepinephrine reuptake inhibitor. The next one, Fluorouracil, is a medication used to treat cancer. By injection into a vein, it is used for colon cancer, esophageal cancer, stomach cancer, pancreatic cancer, breast cancer, and cervical cancer. The fifth drug, Oteracil, is an adjunct to antineoplastic therapy, used to reduce the toxic side effects associated with chemotherapy. The next one, Tilbroquinol, is a medication used in the treatment of intestinal amoebiasis. The seventh drug, Carvedilol, is a medication used to treat high blood pressure, congestive heart failure, and left ventricular

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382

dysfunction. The number eight drug, Mercaptopurine, is a medication used for cancer and autoimmune diseases. Specifically, it treats acute lymphocytic leukemia, chronic myeloid leukemia, Crohn's disease, and ulcerative colitis. The next one is Bedaquiline, which is a medication used to treat active tuberculosis, specifically multidrug-resistant tuberculosis along with other tuberculosis. The number ten drug, Clotrimazole, is an antifungal medication, which is used to treat vaginal yeast infections, oral thrush, diaper rash, pityriasis versicolor, and types of ringworm including athlete's foot and jock itch.

Using our validated machine learning model, we present the binding affinity prediction and ranking of 7012 investigational or off-market drugs. We list the top 20 from the investigational or off-market drugs in Table 2. A complete list of the predicted values can be found in the Supporting Tables (Other_drugs) in Supporting Information.

Table 2. Summary of Top 20 Potential Anti-SARS-CoV-2 Drugs from 7012 Investigational or Off-Market Drugs with Predicted Binding Affinities (BAs) (unit: kcal/mol), IC_{50} (μM), and Corresponding Trade Names

DrugID	name	predicted BA	IC_{50}
DB12903	Debio-1347	−9.02	0.24
DB07959	3-(1H-benzimidazol-2-yl)-1H-indazole	−9.01	0.24
DB07301	9H-carbazole	−8.96	0.27
DB07620	2-[(2,4-dichloro-5-methylphenyl) sulfonyl]-1,3-dinitro-5-(trifluoromethyl) benzene	−8.89	0.30
DB08036	6,7,12,13-tetrahydro-5H-indolo[2,3-a] pyrrolo[3,4-c]carbazol-5-one	−8.89	0.30
DB08440	N-1,10-phenanthrolin-5-ylacetamide	−8.83	0.33
DB01767	Hemi-Babim	−8.80	0.35
DB06828	5-[2-(1H-pyrrol-1-yl)ethoxy]-1H-indole	−8.73	0.39
DB14914	Flortaucipir F-18	−8.69	0.42
DB15033	Flortaucipir	−8.69	0.42
DB13534	Gedocarnil	−8.67	0.44
DB02365	1,10-Phenanthroline	−8.64	0.45
DB09473	Indium In-111 oxyquinoline	−8.64	0.45
DB08512	6-amino-2-[(1-naphthylmethyl)amino]-3,7-dihydro-8H-imidazo[4,5-g] quinazolin-8-one	−8.60	0.48
DB01876	Bis(5-Amidino-2-Benzimidazolyl) Methanone	−8.60	0.49
DB07919	7-methoxy-1-methyl-9H-β-carboline	−8.59	0.49
DB02089	CP-526423	−8.59	0.50
DB07837	[4-(5-naphthalen-2-yl-1H-pyrrolo[2,3-b] pyridin-3-yl)phenyl]acetic acid	−8.53	0.55
DB08073	(2S)-1-(1H-indol-3-yl)-3-{[5-(3-methyl-1H-indazol-5-yl)pyridin-3-yl]oxy} propan-2-amine	−8.53	0.55
DB08267	6-amino-4-(2-phenylethyl)-1,7-dihydro-8H-imidazo[4,5-g]quinazolin-8-one	−8.52	0.56

In comparison to FDA-approved drugs, investigational or off-market drugs might be more promising SARS-CoV-2 inhibitors. Among them, Debio-1347 has a binding affinity of −9.02 kcal/mol (IC_{50}: 0.24 μM). Another top-ranking drug is 3-(1H-benzimidazol-2-yl)-1H-indazole, which has a binding affinity of −9.01 kcal/mol (IC_{50}: 0.24 μM). However, we note that drug discovery is a complex and challenging issue. Having a favorable binding affinity is a necessary but not sufficient condition. Many FDA-approved drugs are selected for their other characteristics, including toxicity, partition coefficient (log P), solubility (log S), synthesizability, pharmacodynamics, pharmacokinetics, etc. These are the issues that prevent many investigational/experimental drugs from becoming FDA-approved drugs. Many off-market drugs might have toxicity and/or side-effect issues, in addition to the availability of better alternatives.

In comparison to FDA-approved drugs, investigational or off-market drugs might be more promising SARS-CoV-2 inhibitors.

The prediction of binding poses is another important task in drug discovery. The goal pose prediction is to determine the binding conformations of small-molecule ligands to the appropriate target binding site. The availability of binding poses enables researchers to understand the molecular mechanism of protein−drug interactions and elucidate fundamental biochemical processes. For example, protein−ligand pose and binding affinity predictions are major tasks in D3R Grand Challenges.[16] Molecular docking is one of the most frequently used methods for pose predictions. In this work, utilizing MathPose developed in recent work,[16] we predict and analyze the binding poses of our predicted top 3 FDA-approved drugs and predicted top 3 investigational or off-market drugs. More detail of the MathPose is given below. The predicted poses are described in the next section.

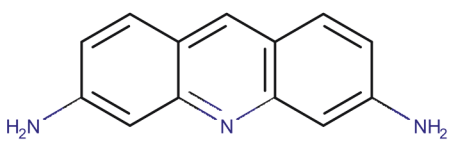
The first-ranking candidate from the FDA-approved drugs is Proflavine (see Figure 1a), with a predicted binding affinity to the SARS-CoV-2 3CL protease of −8.37 kcal/mol. The predicted binding pose using our MathPose[16] is illustrated in Figure 1b. It reveals that there are two hydrogen bonds formed between the drug and the SARS-CoV-2 3CL protease. The first one is between one amino of Proflavine and the O atom in the main chain of the residue Glu166 of the protease. The second one is between the other amino of the drug and the five-member ring in the side chain of the residue His41 of the protease. As a result, the binding affinity is promising.

The predicted second-best drug is Chloroxine (see Figure 1c). Its predicted binding affinity is −8.24 kcal/mol. Between the drug and the protease, there are two hydrogen bonds (see Figure 1d): One is formed by the H atom of the hydroxy of the drug with the main-chain O atom of the residue Leu141. The other one is between the hydroxy O atom of the drug and the amino in the main chain of Cys145.

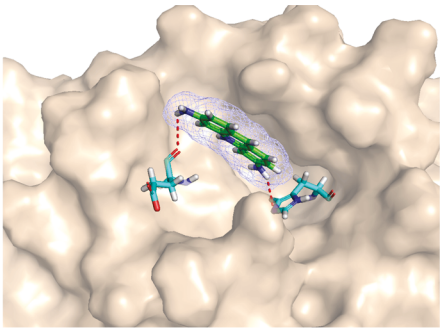
The third one, Demexiptiline (see Figure 1e), has a predicted binding affinity of −8.14 kcal/mol. The hydrogen bonds between this drug and the protease are formed by the H atom of the amino on the tail of the drug with the side-chain O atom of Ser144. Hydrophobic interactions also play a critical role in the binding.

It is interesting to analyze the binding affinities of the existing drugs developed as protease inhibitors. Table 3 shows their predicted binding affinities. The predicted values by a recent study[17] are given in parentheses, and it appears that these values are overestimated. Notably, the current protease inhibitors do not have a substantial effect on the SARS-CoV-2 3CL protease. A possible reason is that SARS-CoV-2 3CL protease is genetically and structurally different from most other known proteases.

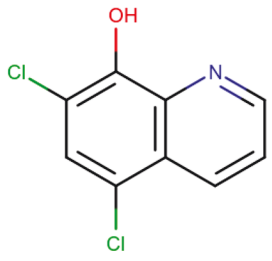
https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382



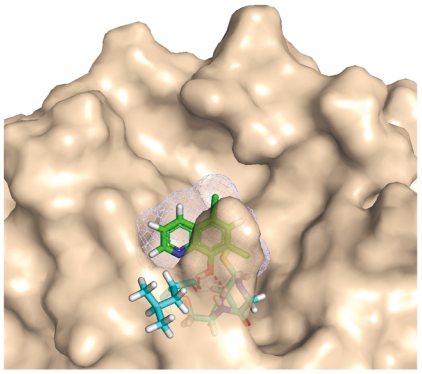
(a) Proflavine, -8.37 kcal/mol



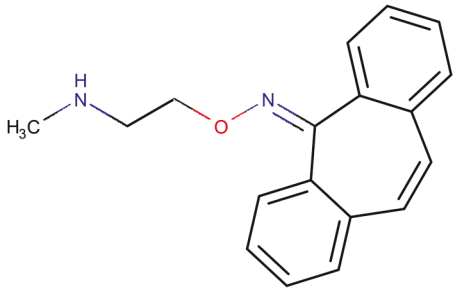
(b) SARS-CoV-2 protease and Proflavine complex



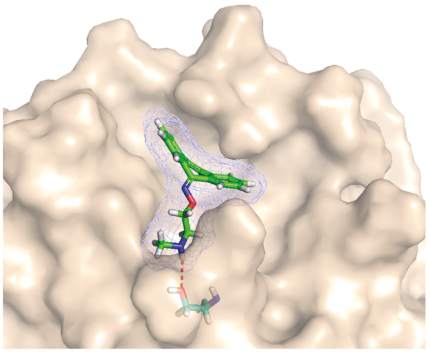
(c) Chloroxine, -8.24 kcal/mol



(d) SARS-CoV-2 protease and Chloroxine complex



(e) Demexiptiline, -8.14 kcal/mol



(f) SARS-CoV-2 protease and Demexiptiline complex

Figure 1. Proflavine, Chloroxine, Demexiptiline, and their complexes with SARS-CoV-2 3CL protease.

In this section, we are interested in comparing our predicted binding affinities to the corresponding experimental ones of some existing drugs outside our training set. Table 4 lists our predictions along with the experimental values of these drugs. These experimental data are extracted from the recent literature.[1,18,19] The RMSE of experimental values and predicted ones is 0.87 kcal/mol, showing a good agreement. It is worth noting that all these data were obtained from cell-culture experiments, leading to discrepancies when comparing these experimental values to our results only tailoring to the inhibition of the SARS-CoV-2 3CL protease. For example, the target of Remdesivir is the RNA-dependent RNA polymerase rather than the 3CL protease.

Among the investigational or off-market drugs, the top-ranking candidate is Debio-1347 (see Figure 2a). Its binding affinity with the SARS-CoV-2 3CL protease is predicted to be −9.02 kcal/mol. The MathPose-predicted pose is illustrated in Figure 2b. It indicates a hydrogen bond network formed between the drug and the protease leads to the moderately high binding affinity. This network consists of two hydrogen bonds: the first hydrogen bond is between one N atom in the Pyrazole of the drug and the main-chain amino of the residue Glu166 of the protease; the second one is between one N atom in the 1H-1,3-benzodiazole of the drug and the main-chain amino of the residue Gly143 of the protease.

The second-best investigational drug is 3-(1H-benzimidazol-2-yl)-1H-indazole (Figure 2c) with a predicted binding affinity of −9.01 kcal/mol. Figure 2d reveals that the drug forms two hydrogen bonds with the protease. One is between one N atom in the 1H-1,3-benzodiazole of the drug and the main-chain O atom of the residue Glu166 of the protease. The other

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382

Table 3. Summary of the Predicted Binding Affinities (BAs) (unit: kcal/mol) and IC_{50} (μM) of the Existing Protease Inhibitors[a]

DrugID	predicted binding affinity	IC_{50}	DrugID	predicted BA	IC_{50}
Remikiren	−7.42	3.57	Moexipril	−6.55	15.63
Candoxatril	−7.22	5.05	Trandolapril	−6.54	17.70
Darunavir	−7.16	5.55	Lopinavir	−6.50	16.92
Isofluorophate	−7.09	6.28	Spirapril	−6.49	17.16
Atazanavir	−7.03 (−9.57)	6.96	Dabigatran etexilate	−6.46	17.96
Argatroban	−7.02	6.98	Apixaban	−6.44	18.84
Sitagliptin	−6.93	8.22	Tipranavir	−6.39	20.36
Fosamprenavir	−6.92	8.26	Lisinopril	−6.35	21.87
Quinapril	−6.91	8.45	Perindopril	−6.34	22.10
Amprenavir	−6.82	9.83	Cilazapril	−6.31	23.36
Benazepril	−6.81	10.05	Ritonavir	−6.26 (−8.47)	25.50
Rivaroxaban	−6.74	11.21	Ximelagatran	−6.24	26.14
Fosinopril	−6.74	11.28	Vildagliptin	−6.15	30.38
Telaprevir	−6.73	11.54	Cilastatin	−6.15	30.40
Captopril	−6.72	11.68	Indinavir	−6.11	32.91
Ramipril	−6.66	12.84	Saxagliptin	−6.07	35.27
Enalapril	−6.66	12.93	Nelfinavir	−6.05	36.23
Alogliptin	−6.62	13.90	Boceprevir	−6.00	39.16
Linagliptin	−6.58	14.73	Simeprevir	−5.77 (−8.29)	58.25
Saquinavir	−6.56	15.26	Ecabet	−5.71	64.15

[a]Numbers in parentheses are predictions from the literature.[17]

Table 4. Summary of our Predicted Binding Affinities (BAs) and the Corresponding Experimental Values of Some Existing Drugs against SARS-CoV-2[a]

DrugID	experiment	prediction	DrugID	experiment	prediction
Remdesivir	−6.74[1]	−6.29	Perhexiline	−7.08[1]	−6.67
Chloroquine	−7.00[1]	−6.92	Loperamide	−6.86[1]	−6.98
Lopinavir	−6.87[1]	−6.51	Mefloquine	−7.31[1]	−6.89
Niclosamide	−8.93[1]	−7.66	Amodiaquine	−7.21[1]	−6.93
Proscillaridin	−7.75[1]	−6.50	Phenazopyridine	−6.21[1]	−7.51
Penfluridol	−7.23[1]	−6.54	Clomiphene	−7.19[1]	−7.12
Toremifene	−7.42[1]	−7.20	Digoxin	−9.16[1]	−7.00
Hexachlorophene	−8.24[1]	−7.37	Thioridazine	−7.05[1]	−6.96
Salinomycin	−9.02[1]	−7.00	Pyronaridine	−6.13[1]	−6.68
Ciclesonide	−7.31[1]	−7.04	Ceritinib	−7.56[1]	−6.77
Osimertinib	−7.48[1]	−6.62	Lusutrombopag	−7.39[1]	−6.78
Gilteritinib	−7.05[1]	−5.57	Berbamine	−6.96[1]	−6.87
Ivacaftor	−7.07[1]	−6.74	Mequitazine	−7.00[1]	−6.41
Dronedarone	−7.37[1]	−6.19	Eltrombopag	−6.93[1]	−6.17
Fluphenazine	−7.08[18]	−6.29	Benztropine	−6.63[18]	−6.94
Chlorpromazine	−7.50[18]	−7.00	Terconazole	−6.71[18]	−7.18
Simeprevir	−6.67[19]	−5.77	Boceprevir	−7.34[19]	−6.00
Narlaprevir	−7.14[19]	−6.38			

[a]All numbers are in kcal/mol.

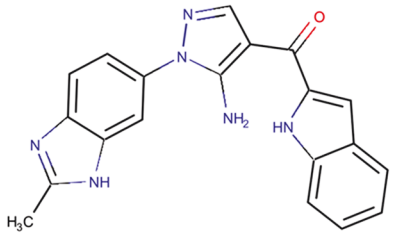
is between one N atom in the 1H-indazole of the drug and the main-chain O atom of the residue His164 of the protease.

The third one, 9H-carbazole (see Figure 2e), also has a promising predicted affinity of −8.96 kcal/mol. As one can see from Figure 2f, a strong hydrogen bond is formed between the N atom of the drug and the main-chain O atom of the residue His164 of the protease. The hydrophobic interactions play an essential role in the binding as well.

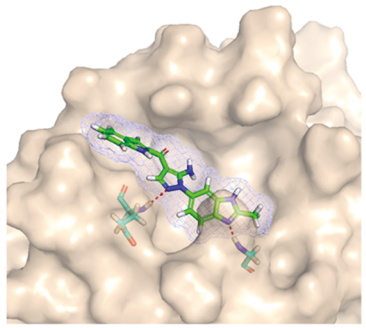
Note that in our training set collected from the existing experimental data, 21 samples have binding affinity values lower than −9 kcal/mol. Table 5 provides a list of the top 20 SARS-CoV/SARS-CoV-2 3CL-protease inhibitors with their experimental binding affinities and estimated druggable properties. Moreover, 4 of these 21 samples have 3D experimental structures available. Although these inhibitors are not on the market yet, they serve as good starting points for the design of anti-SARS-CoV-2 drugs. A full list of our training compounds is given in the Supporting Tables (Training set) in Supporting Information.

Among the SARS-CoV/SARS-CoV-2 3CL-protease complexes with their 3D experimental structures available, the one with the PDB ID 2zu4[20] is the most potent one with a binding affinity over −10 kcal/mol. This high binding affinity is due to a strong hydrogen bond network between the inhibitor and the protease, which consists of as many as 7 hydrogen bonds. These 7 hydrogen bonds are formed by the inhibitor with

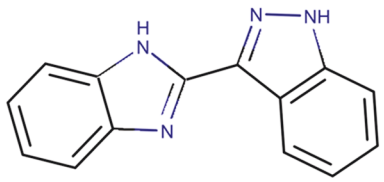
https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382



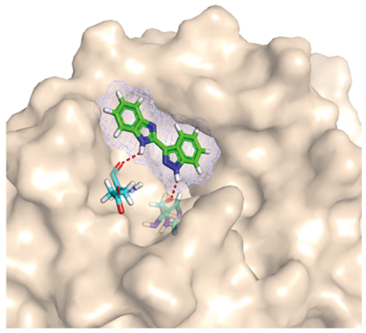
(a) DEBIO-1347, -9.02 kcal/mol



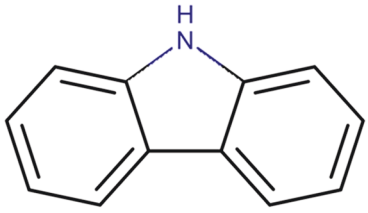
(b) SARS-CoV-2 protease and DEBIO-1347 complex



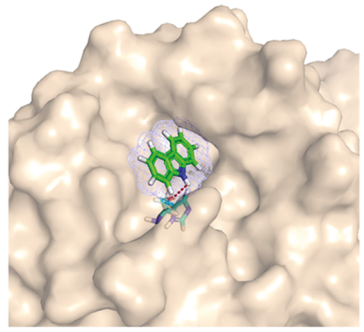
(c) 3-(1H-BENZIMIDAZOL-2-YL)-1H-INDAZOLE, -9.01 kcal/mol



(d) SARS-CoV-2 protease and 3-(1H-BENZIMIDAZOL-2-YL)-1H-INDAZOLE complex



(e) 9H-CARBAZOLE, -8.96 kcal/mol



(f) SARS-CoV-2 protease and 9H-CARBAZOLE complex

Figure 2. Debio-1347, 3-(1H-benzimidazol-2-yl)-1H-indazole, 9H-carbazole, and their complexes with SARS-CoV-2 3CL protease.

protease residues Gln189, Gly143, His163, His164, and Glu166 of the protease.

The second-best 3D-experimental structure is the one with the PDB ID 3avz,[21] and its binding affinity is −9.80 kcal/mol. A hydrogen bond network, including 7 bonds, plays an essential role in this strong binding. This network is between the inhibitor and protease residues Gln192, Thr190, His164, His163, Glu166, and Gly143.

The PDB ID of the third one is 2zu5[20] with a binding affinity of −9.56 kcal/mol. A strong hydrogen bond network with 7 bonds can also be found in the structure. The protease residues in the network are Glu166, Phe148, His163, His164, Gly143, and Gln189.

Because His163, His164, and Glu166 emerge in the hydrogen bond networks of all three structures, it suggests that these three residues are critical to inhibitor binding.

The partition coefficient (log P), aqueous solubility (log S), and synthesizability are also critical medical chemical properties for deciding whether a compound can be a drug or not. Notably, synthesizability is always in terms of synthetic accessibility score (SAS), for which 1 indicates the easiest, 10 indicates the hardest. Here, we first calculate the log P's, log S's, and SASs of the 1553 FDA-approved drugs (see the Supporting Tables (FDA_approved) in Supporting Information); we then investigate whether the three properties of the inhibitors in the top three 3D experimental structures (Figure 3) are in the preferred ranges of the FDA-approved drugs.

According to the log P distribution of the FDA-approved drugs in the Supporting Tables (FDA_approved) in Supporting Information, the log P interval with a large population of the FDA-approved drugs is between −0.14 and 4.96. The log P values of the top 3 inhibitors are 2.35, −1.35, and −0.46, respectively.

The log S distribution reveals that the preferred range of log S is between −5.12 and 1.76. The log S values of the top 3 inhibitors are −3.53, −2.33, and −4.39.

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382

Table 5. Summary of Top 20 SARS-CoV-2 3CL Protease Inhibitors in the Training Set with Experimental Binding Affinities (unit: kcal/mol), IC_{50} (μM), as Well as Calculated Synthesizability, log P, and log S

ID	binding affinity	IC_{50}	synthesizability	log P	log S
CHEMBL497141	−11.08	0.01	2.4	2.18	−3.65
PDB ID 2zu4	−10.12	0.04	4.04	2.35	−3.53
CHEMBL222234	−9.95	0.05	2.26	2.66	−3.59
CHEMBL2442057	−9.94	0.05	2.26	5.39	−6.22
CHEMBL213054	−9.92	0.05	4.2	3.15	−3.81
CHEMBL212080	−9.87	0.06	4.25	3.15	−3.76
CHEMBL222840	−9.85	0.06	2.23	2.55	−3.37
CHEMBL398437	−9.85	0.06	2.29	4.12	−5.39
CHEMBL222769	−9.82	0.06	2.16	4.87	−5.73
PDB ID 3avz	−9.80	0.07	4.65	−1.35	−2.33
CHEMBL225515	−9.80	0.07	2.22	3.44	−4.28
CHEMBL1929019	−9.80	0.07	4.23	−0.77	−2.41
CHEMBL222893	−9.57	0.10	2.21	4.17	−5.01
PDB ID 2zu5	−9.56	0.10	4.27	3.79	−4.39
PDB ID 3atw	−9.55	0.10	4.63	−0.46	−2.47
CHEMBL334399	−9.50	0.11	2.20	3.06	−4.17
CHEMBL253905	−9.43	0.12	2.43	4.78	−5.45
CHEMBL403932	−9.42	0.12	1.94	4.11	−4.97
CHEMBL254103	−9.25	0.16	2.10	2.35	−3.34
CHEMBL426898	−9.23	0.17	2.17	3.70	−4.72

In the SAS distribution, most of the FDA-approved drugs have SASs between 1.84 and 3.94. The SAS values of the top 3 inhibitors are 4.04, 4.65, and 4.27.

In summary, for the inhibitor in the first ranking PDB structure 2zu4, its log P and log S are quite good for a drug. The SAS is a little higher, but it is still not too difficult to synthesize: 344 of the 1553 FDA-approved drugs have larger SASs than this inhibitor, and 56 of them even have SASs over 6.

Similarly, for the 3avz and 2zu5 inhibitors, their log S's are very promising. Some of the log P's and SASs are out of the preferred ranges, but many FDA-approved drugs still have worse log Ps and SASs. As a result, these top 3 inhibitors, especially the first one, could be good starting points for developing anti-SARS-CoV2 drugs. Obviously, their toxicity will be a major concern for any further development.

We collect the training set from single-protein experimental data of SARS/SARS-CoV-2 3CL protease in public databases or the related literature.

ChEMBL is a manually curated database of bioactive molecules.[22] Currently, ChEMBL contains more than 2 million compounds only in the SMILES string format. In ChEMBL, we find 277 SARS-CoV or SARS-CoV-2 3CL protease inhibitors with reported K_d/IC_{50} from single-protein experiments.

Another database is PDBbind. The PDBbind database includes all the protein−ligand complexes with the crystal structures deposited in the Protein Data Bank (PDB) and their binding affinities in the form of K_d, K_i, or IC_{50} reported in the literature.[23] The newest PDBbind v2019 consists of 17 679 complexes as well as the binding affinities. We find another 30 inhibitors in the PDBbind v2019.

Additionally, binding affinities for four other SARS-CoV 3CL protease inhibitors and three other SARS-CoV-2 3CL protease inhibitors are extracted from refs 24 and 13, respectively. Therefore, we collected 314 SARS-CoV/SARS-CoV-2 3CL protease inhibitors with available experimental binding affinities.

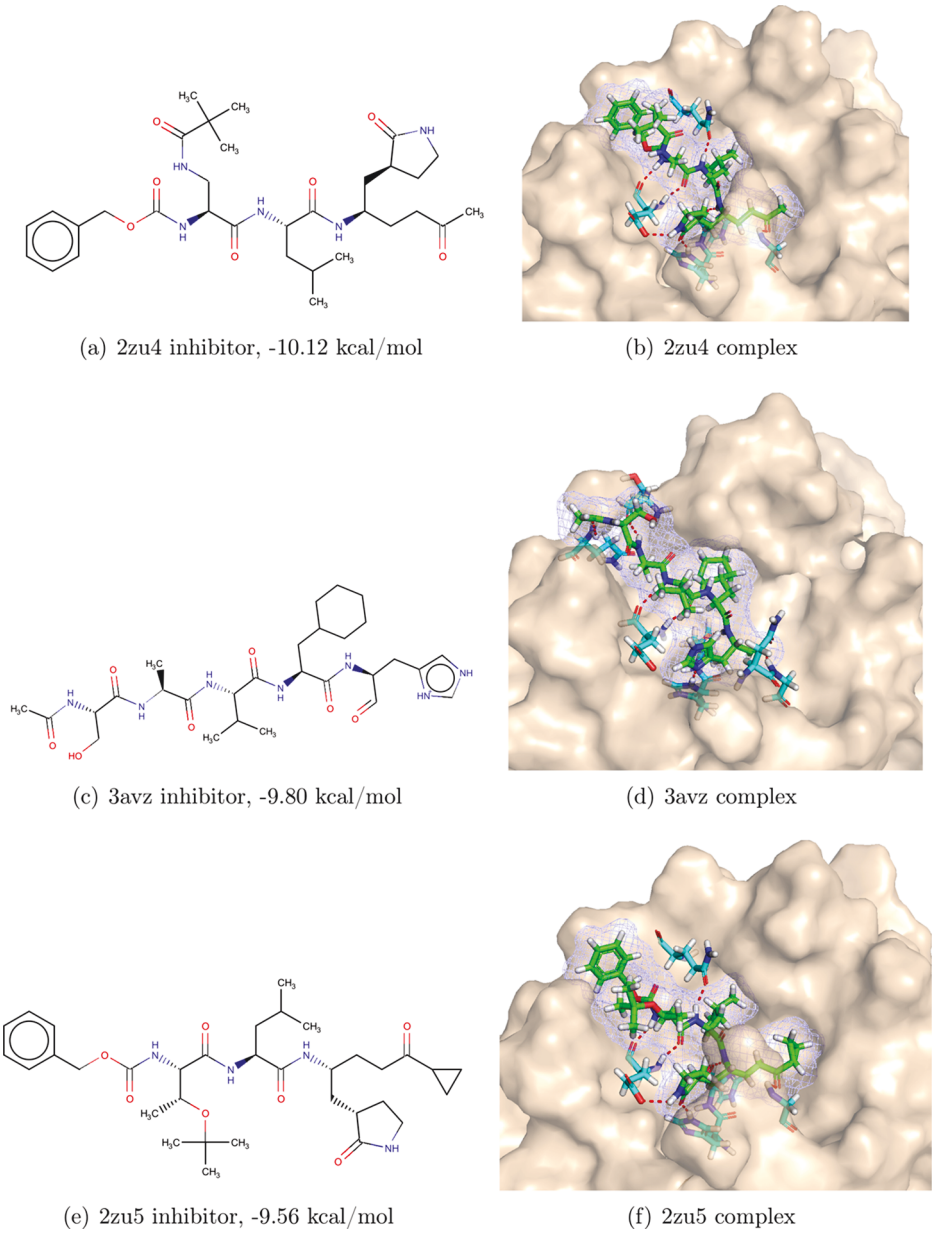
The binding affinity range in this set is from −3.68 kcal/mol to −11.08 kcal/mol. The distribution is depicted in Figure S3. The top 20 inhibitors in the training set are summarized in Table 5.

DrugBank (www.drugbank.ca)[14] is a richly annotated, freely accessible online database that integrates massive drug, drug target, drug action, and drug interaction information about FDA-approved drugs as well as investigational or off-market drugs. Because of the high quality and sufficient information contained in it, the DrugBank has become one of the most popular reference drug resources used all over the world. In the current work, we extract 1553 FDA-approved drugs and 7012 investigational or off-market drugs from DrugBank and evaluate their binding affinities to the SARS-CoV-2 3CL protease.

In this work, the log P and synthesizability values are calculated by RDKit (http://www.rdkit.org); the synthesizability in RDkit is reported in terms of synthetic accessibility score (1 means the easiest, and 10 means the hardest). The log S values are obtained via Alog PS 2.1.[25]

The 3D binding poses in this work are predicted by the MathPose, a 3D pose predictor which converts SMILES strings into 3D poses with references of target molecules. It was the top performer in D3R Grand Challenge 4 in predicting the poses of 24 beta-secretase 1 (BACE) binders.[16] For one SMILES string, around 1000 3D structures can be generated by a common docking software tool, i.e., GLIDE. Moreover, a selected set of known complexes is redocked by the three docking software packages mentioned above to generate 100 decoy complexes per input ligand as a machine learning training set. All of those structures are optimized by a minimization component in GLIDE with the OPLS3 force field.[26] The machine learning labels will be the calculated root mean squared deviations (RMSDs) between the decoy and native structures for this training set. Furthermore, MathDL models[16] are set up and applied to select the top-ranked pose for the given ligand.

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382



(a) 2zu4 inhibitor, -10.12 kcal/mol (b) 2zu4 complex

(c) 3avz inhibitor, -9.80 kcal/mol (d) 3avz complex

(e) 2zu5 inhibitor, -9.56 kcal/mol (f) 2zu5 complex

Figure 3. Inhibitors and complexes from the top three PDBbind structures, 2zu4, 3avz, and 2zu5.

In the current work, we develop a machine learning model for predicting the binding affinities of SARS-CoV-2 inhibitors. Our current model is classified as a ligand-based approach, the most popular framework in computer-aided drug design owing to its simplicity in data preparation while still delivering satisfactory performance.[27−30] Because the size of the training set in our current case study is only 314, we apply the gradient-boosting decision tree (GBDT) model because of its accuracy for handling small data sets. This GBDT predictor is constructed using the gradient boosting regressor module in scikit-learn (version 0.20.1).

The 2D fingerprints of compounds are used as the input features to our GBDT predictor. Previous study shows that the consensus of ECFP4, Estate1, and Estate2 fingerprints performs the best on binding-affinity prediction tasks.[31] In this work, we also make use of this consensus. The 2D fingerprints are calculated from SMILES strings using RDKit software (version 2018.09.3) (http://www.rdkit.org).

We validate the performance of our machine learning predictor for the 314 inhibitors in the SARS-CoV-2 BA training set. We use 10-fold cross-validation, which is carried out using 50 random splittings. In Table S1 we show that our machine learning predictor is trained with the average Pearson correlation coefficient (R_p) of 0.997, the Kendall's τ (τ) of 0.972, and RMSE of 0.095 kcal/mol. These metrics are based on the averaged values across 10 folds, and these results indicate our model is well-trained. Their averaged test performances across the 10 folds of the whole SARS-CoV-2 BA set are found to be R_p = 0.777, τ = 0.586, and RMSE = 0.792 kcal/mol. These results endorse the reliability of our model in the binding affinity prediction of SARS-CoV-2 inhibitors.

The current pneumonia outbreak caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has evolved into a global pandemic. Although currently there is no effective antiviral medicine against SARS-CoV-2, the 3CL

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382

Although currently there is no effective antiviral medicine against SARS-CoV-2, the 3CL proteases of SARS-CoV-2 and SARS-CoV have a sequence identity of 96.1% and the binding-site RMSD of 0.42 Å, which provide a solid basis for us to hypothesize that all potential anti-SARS-CoV chemotherapies are also effective anti-SARS-CoV-2 molecules.

proteases of SARS-CoV-2 and SARS-CoV have a sequence identity of 96.1% and the binding-site RMSD of 0.42 Å, which provide a solid basis for us to hypothesize that all potential anti-SARS-CoV chemotherapies are also effective anti-SARS-CoV-2 molecules. In this work, we curate 314 SARS-CoV-2/SARS-CoV 3CL protease inhibitors with available experimental binding data from various sources to form a machine learning training set. Using this training set, we develop gradient-boosted decision trees (GBDT) model to predict the binding affinities of potential SARS-CoV-2 3CL protease inhibitors. The 10-fold cross-validation shows our model has a Pearson correlation coefficient of 0.78 and a relatively low root-mean-square error of 0.80 kcal/mol. A total of 8565 drugs from DrugBank are evaluated by their predicted binding affinities. We highlight 20 FDA-approved drugs as well as 20 investigational or off-market drugs as potentially potent medications against SARS-CoV-2. We also analyze the druggability of some potent inhibitors in our training set. This work serves as a foundation for further experimental development of anti-SARS-CoV-2 drugs.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c01579.

3CL protease sequence identity and 3D structure similarity analysis, machine learning details, the MathDL model, and the list of nonpolar 3CL protease binding site residues (PDF)

Tables of experimental binding affinities for 314 SARS-CoV-2 3CL protease inhibitors, the predicted binding affinities of 1553 FDA-approved drugs, and 7012 investigational or off-market drugs (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Guo-Wei Wei − *Department of Mathematics, Department of Biochemistry and Molecular Biology, and Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0002-5781-2937; Email: wei@math.msu.edu

Authors

Kaifu Gao − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

Duc Duy Nguyen − *Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506, United States;* orcid.org/0000-0002-5921-8851

Jiahui Chen − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

Rui Wang − *Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c01579

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported in part by NIH Grant GM126189; NSF Grants DMS-1721024, DMS-1761320, and IIS1900473; Michigan Economic Development Corporation; Bristol-Myers Squibb; and Pfizer. The authors thank The IBM TJ Watson Research Center, The COVID-19 High Performance Computing Consortium, NVIDIA, and MSU HPCC for computational assistance supporting this work.

■ REFERENCES

(1) Jeon, S.; Ko, M.; Lee, J.; Choi, I.; Byun, S. Y.; Park, S.; Shum, D.; Kim, S. Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs. *Antimicrob. Agents Chemother.* **2020**, DOI: 10.1128/AAC.00819-20.

(2) MacIntyre, C. R. Wuhan novel coronavirus 2019nCoV−update January 27th 2020. *Glob. Biosecur.* **2019**, *1*, 1 DOI: 10.31646/gbio.51.

(3) 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**.

(4) Brown, A. S.; Patel, C. J. A standard database for drug repositioning. *Sci. Data* **2017**, *4*, 1−7.

(5) Amelio, I.; Gostev, M.; Knight, R.; Willis, A.; Melino, G.; Antonov, A. DRUGSURV: a resource for repositioning of approved and experimental drugs in oncology based on patient survival information. *Cell Death Dis.* **2014**, *5*, e1051−e1051.

(6) Jin, G.; Wong, S. T. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discovery Today* **2014**, *19*, 637−644.

(7) Patwardhan, B.; Chaguturu, R. *Innovative Approaches in Drug Discovery: Ethnopharmacology, Systems Biology and Holistic Targeting*; Academic Press, 2016.

(8) Li, J.; Zheng, S.; Chen, B.; Butte, A. J.; Swamidass, S. J.; Lu, Z. A survey of current trends in computational drug repositioning. *Briefings Bioinf.* **2016**, *17*, 2−12.

(9) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **2018**, *14*, No. e1005929.

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

(11) Xu, X.; Chen, P.; Wang, J.; Feng, J.; Zhou, H.; Li, X.; Zhong, W.; Hao, P. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China: Life Sci.* **2020**, *63*, 457−460.

(12) Lee, T.-W.; Cherney, M. M.; Huitema, C.; Liu, J.; James, K. E.; Powers, J. C.; Eltis, L. D.; James, M. N. Crystal structures of the main peptidase from the SARS coronavirus inhibited by a substrate-like aza-peptide epoxide. *J. Mol. Biol.* **2005**, *353*, 1137−1151.

(13) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **2020**, *368*, 409−412.

(14) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.
(15) Li, H.; Sze, K.-H.; Lu, G.; Ballester, P. J. Machine-learning scoring functions for structure-based drug lead optimization. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, No. e1465.
(16) Nguyen, D. D.; Gao, K.; Wang, M.; Wei, G.-W. Mathdl: Mathematical deep learning for d3r grand challenge 4. *J. Comput.-Aided Mol. Des.* **2020**, *34*, 131−147.
(17) Beck, B. R.; Shin, B.; Choi, Y.; Park, S.; Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 784−790.
(18) Weston, S.; Haupt, R.; Logue, J.; Matthews, K.; Frieman, M. FDA approved drugs with broad anti-coronaviral activity inhibit SARS-CoV-2 in vitro. *bioRxiv* **2020**, DOI: 10.1101/2020.03.25.008482.
(19) Ma, C.; Hurst, B.; Hu, Y.; Szeto, T.; Tarbet, B.; Wang, J. Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res.* **2020**, DOI: 10.1038/s41422-020-0356-z.
(20) Lee, C.-C.; Kuo, C.-J.; Ko, T.-P.; Hsu, M.-F.; Tsui, Y.-C.; Chang, S.-C.; Yang, S.; Chen, S.-J.; Chen, H.-C.; Hsu, M.-C.; et al. Structural basis of inhibition specificities of 3C and 3C-like proteases by zinc-coordinating and peptidomimetic compounds. *J. Biol. Chem.* **2009**, *284*, 7646−7655.
(21) Akaji, K.; Konno, H.; Mitsui, H.; Teruya, K.; Shimamoto, Y.; Hattori, Y.; Ozaki, T.; Kusunoki, M.; Sanjoh, A. Structure-based design, synthesis, and evaluation of peptide-mimetic SARS 3CL protease inhibitors. *J. Med. Chem.* **2011**, *54*, 7962−7973.
(22) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(23) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein- ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(24) Bacha, U.; Barrila, J.; Gabelli, S. B.; Kiso, Y.; Mario Amzel, L.; Freire, E. Development of Broad-Spectrum Halomethyl Ketone Inhibitors Against Coronavirus Main Protease 3CLpro. *Chem. Biol. Drug Des.* **2008**, *72*, 34−49.
(25) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; et al. Virtual computational chemistry laboratory−design and description. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 453−463.
(26) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; et al. OPLS3: a force field providing broad coverage of drug-like small molecules and proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281−296.
(27) Soufan, O.; Ba-alawi, W.; Magana-Mora, A.; Essack, M.; Bajic, V. B. DPubChem: a web tool for QSAR modeling and high-throughput virtual screening. *Sci. Rep.* **2018**, *8*, 1−10.
(28) Peón, A.; Li, H.; Ghislat, G.; Leung, K.-S.; Wong, M.-H.; Lu, G.; Ballester, P. J. MolTarPred: a web tool for comprehensive target prediction with reliability estimation. *Chem. Biol. Drug Des.* **2019**, *94*, 1390−1401.
(29) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme gradient boosting as a method for quantitative structure−activity relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353−2360.
(30) Sidorov, P.; Naulaerts, S.; Ariey-Bonnet, J.; Pasquier, E.; Ballester, P. Predicting synergism of cancer drug combinations using NCI-ALMANAC data. *Front. Chem.* **2019**, *7*, 509.
(31) Gao, K.; Nguyen, D. D.; Sresht, V.; Mathiowetz, A. M.; Tu, M.; Wei, G.-W. Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **2020**, *22*, 8373−8390.

https://dx.doi.org/10.1021/acs.jpclett.0c01579
J. Phys. Chem. Lett. 2020, 11, 5373−5382