Briefings in Bioinformatics, 00(00), 2021, 1–11

doi: 10.1093/bib/bbab211
Problem Solving Protocol

A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2

Shiwei Wang, Qi Sun, Youjun Xu, Jianfeng Pei and Luhua Lai

Corresponding authors: Jianfeng Pei, Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, PR China. Tel./Fax: +86 10 6275 9669. E-mail: jfpei@pku.edu.cn; Luhua Lai, BNLMS, Peking-Tsinghua Center for Life Sciences at the College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, PR China. Tel./Fax: +86 10 6275 1725. E-mail: lhlai@pku.edu.cn

Abstract

The COVID-19 pandemic calls for rapid development of effective treatments. Although various drug repurpose approaches have been used to screen the FDA-approved drugs and drug candidates in clinical phases against SARS-CoV-2, the coronavirus that causes this disease, no magic bullets have been found until now. In this study, we used directed message passing neural network to first build a broad-spectrum anti-beta-coronavirus compound prediction model, which gave satisfactory predictions on newly reported active compounds against SARS-CoV-2. Then, we applied transfer learning to fine-tune the model with the recently reported anti-SARS-CoV-2 compounds and derived a SARS-CoV-2 specific prediction model COVIDVS-3. We used COVIDVS-3 to screen a large compound library with 4.9 million drug-like molecules from ZINC15 database and recommended a list of potential anti-SARS-CoV-2 compounds for further experimental testing. As a proof-of-concept, we experimentally tested seven high-scored compounds that also demonstrated good binding strength in docking studies against the 3C-like protease of SARS-CoV-2 and found one novel compound that can inhibit the enzyme. Our model is highly efficient and can be used to screen large compound databases with millions or more compounds to accelerate the drug discovery process for the treatment of COVID-19.

Key words: COVID-19; SARS-CoV-2; deep learning; drug repurposing; virtual screening

Introduction

COVID-19 is a newly emerged infectious disease that becomes a worldwide pandemic. According to the World Health Organization (WHO) statistics, tens of millions of confirmed cases of COVID-19 and millions of deaths have been reported [1]. SARS-CoV-2, a new coronavirus, has been identified to cause COVID-19 [2, 3]. Coronaviruses (CoVs) are a group of enveloped single stranded positive-sense RNA viruses, which are able to infect many animals and humans and cause a wide range of diseases [4]. Whole genome sequencing showed that SARS-CoV-2 shares 79.6% sequence identify to SARS-CoV [5]. SARS-CoV-2 appears to have relatively high transmission rate among humans and causes severe and fatal pneumonia and other damages, threatening people at all ages, especially senior ones [6]. As the number of infections and deaths are rapidly increasing, there is an urgent call for drug and vaccine development against COVID-19.

Though immediately needed, developing new drugs within a short period of time is unpractical. Repurposing of clinically approved drugs provides a fast and effective strategy to identify antiviral drugs for immediate use. Several drugs such as remdesivir [7], chloroquine [7, 8] and lopinavir [9] have shown antiviral activity *in vitro* and been tested in clinical trials. FDA-approved drugs as well as those that were previously found to inhibit SARS-CoV and MERS-CoV have been screened for their anti-SARS-CoV-2 activities [10–13]. Nine approved HIV-1

Shiwei Wang is a PhD student in the Academy for Advanced Interdisciplinary Studies, Peking University.
Qi Sun is a research scientist in the College of Chemistry and Molecular Engineering, Peking University.
Youjun Xu is a postdoc fellow in the College of Chemistry and Molecular Engineering, Peking University.
Jianfeng Pei is an associate professor in the Academy for Advanced Interdisciplinary Studies, Peking University.
Luhua Lai is a professor in the College of Chemistry and Molecular Engineering, Peking University.
Submitted: 11 February 2021; **Received (in revised form):** 29 April 2021

protease inhibitors were also evaluated for their anti-SARS-CoV-2 activities *in vitro* and nelfinavir was found to be active [14]. Traditional Chinese medicines provide rich resources for developing antiviral drugs. Baicalein, an ingredient isolated from *Scutellaria baicalensis* (Huangqin in Chinese), has been reported to inhibit the 3C-like protease (3CL^{pro}) of SARS-CoV-2 and SARS-CoV-2 replication *in vitro*, which provides potential treatment of COVID-19 [15, 16]. The crystal structure of 3CL^{pro} from SARS-CoV-2 was quickly solved [17] and docking-based virtual screening can be applied to discover more potential 3CL^{pro} inhibitors.

Compared to experimental screening and docking-based screening, deep learning based virtual screening provides a new approach in drug discovery. It generally encodes molecules into vectors and then constructs a mapping relationship from these vectors to their properties. Artificial intelligence (AI)-based virtual screening methods enable rapid search against large molecular libraries containing 10^6–10^9 molecules, which is usually time-consuming for traditional methods. Stokes *et al*. [18] discovered a new antibiotic with a broad-spectrum bactericidal activity by combining *in silico* predictions and experimental investigations. Ton *et al*. [19] applied deep docking model to screen all the 1.3 billion compounds from ZINC15 library and recommended the top 1000 hits as potential SARS-CoV-2 3CL^{pro} inhibitors, though no experimental testing has been reported. In the present study, we combined *in silico* methods and *in vitro* studies to screen anti-SARS-CoV-2 drugs. We used a directed message passing neural network to learn the structure–activity relationship from a collection of anti-beta-CoV active and inactive compounds. The first model (COVIDVS-1) trained on experimental data on several beta-CoVs gave good predictions for the recently identified anti-SARS-CoV-2 compounds when screening the Drug Repurposing Hub (DRH) library containing 6235 FDA-approved drugs, clinical trial drugs and pre-clinical tool compounds [20]. We then fine-tuned COVIDVS-1 successively with recently reported active and inactive compounds against SARS-CoV-2 and derived the COVIDVS-3 model. We applied COVIDVS-3 model to screen a large compound library with 4.9 million drug-like molecules from ZINC15 database [21]. We suggested a list of potential anti-SARS-CoV-2 compounds for further experimental testing. As a proof-of-concept, we experimentally tested the activities of seven molecules with high prediction scores and good binding affinities from docking studies against 3CL^{pro} of SARS-CoV-2 and found one non-covalent inhibitor with novel chemical scaffold.

Material And Methods

Data

Training dataset is essential for deep learning methods. In order to train a robust model that can predict new antiviral drugs against SARS-CoV-2, an ideal training set should contain sufficient positive and negative compounds for SARS-CoV-2. Unfortunately, SARS-CoV-2 is a newly emerged coronavirus, and only limited information is available now. SARS-CoV-2, as well as HCoV-OC43, SARS-CoV and MERS-CoV, belongs to beta-coronaviruses [3, 22]. They share a high degree of conservation in essential functional proteins, including the 3CL^{pro}, the RNA-dependent RNA polymerase, the RNA helicase, etc. [23] For example, the 3CL^{pro} in SARS-CoV and SARS-CoV-2 share a sequence identity of 96.1%, indicating that these CoVs share potential targets for broad-spectrum anti-CoV drugs. Potent MERS-CoV inhibitors identified by screening an FDA-approved drug library also inhibit the replication of SARS-CoV and HCoV-229E [24]. Shen *et al*. [23] found seven broad-spectrum antiviral inhibitors through a high-throughput screening of a 2000-compound library against HCoV-OC43. These studies provide a list of antivirals for beta-CoVs that can be used to train a model for screening SARS-CoV-2 antiviral candidates.

We collected a set of inhibitors against HCoV-OC43, SARS-CoV and MERS-CoV from literatures with a cutoff of $EC_{50} < 10$ μM and selective index >10 [22, 23, 25–27]. All the inhibitors were identified by screening libraries including FDA-approved drugs and pharmacologically active compounds. After applying the cutoff filter, 90 compounds were selected as antivirals and each of them can inhibit at least one of the three CoVs. The remaining compounds were regarded as negative data. This primary training dataset (Training Set 1) containing 90 positives and 1862 negatives was used to train the deep learning classification model for screening anti-beta-coronavirus compounds.

We also constructed an independent dataset containing a collection of experimentally tested active and inactive molecules against SARS-CoV-2 [10–14]. In general, compounds were labeled as actives if their EC_{50} against SARS-CoV-2 are <50 μM. Several well tested compounds, including indinavir, were also labeled as positive data with EC_{50}'s a little bit higher than 50 μM [14]. This gave a dataset (Fine-tuning Set 1) with 70 actives and 84 inactives. We applied this SARS-CoV-2-specific dataset to train the SARS-CoV-2-specific antiviral prediction model. In addition, an independent test set (Test Set 1) including 33 actives and 38 inactives was constructed by removing the compounds that also present in the Training Set 1 from Fine-tuning Set 1.

A recent study experimentally screened the ReFRAME library, which collects a large number of clinical-phase or FDA-approved drugs, against SARS-CoV-2 and reported 20 active compounds (hereafter referred as ReFRAME actives) [28, 29]. Among the 20 active compounds, 3 were already included in our Training Set 1 and Fine-tuning Set 1. We added the 17 newly discovered actives to Fine-tuning Set 1 to construct a new dataset (Fine-tuning Set 2). For simplicity, we called the combination of Training Set 1 and Fine-tuning Set 1 as Training Set 2 and the combination of Training Set 1 and Fine-tuning Set 2 as Training Set 3.

The DRH is a curated and annotated collection of FDA-approved drugs, clinical trial drug candidates and pre-clinical compounds with a companion information resource [20]. We applied our model to this library to identify potential antiviral molecules. Compounds overlapping with the training dataset were removed and the rest compounds were used to screen potential antivirals.

ZINC15 is a free database designed for virtual screening, containing ∼1.5 billion molecules [21]. We extracted a subset database containing ∼4.9 million molecules that are drug-like and in stock. Virtual screening was applied to this library to discover potential antiviral molecules.

A summary of all the datasets used in this study were listed in Table 1.

Model

In this work, we developed a series of COVIDVS models which showed satisfied performances and then applied them to screening potential antiviral drugs from a large compound library. Our study contains three steps (Figure 1A): (1) we trained a broad-spectrum anti-beta-coronavirus compounds prediction model (COVIDVS-1) with Training Set 1; (2) we fine-tuned COVIDVS-1 model with SARS-CoV-2 specific datasets Fine-tuning Sets 2 and 3, and obtained the SARS-CoV-2 specific prediction model COVIDVS-2 and COVIDVS-3 and (3) we applied

Table 1. Summary of datasets

Dataset	Description	Source
Training Set 1	Contains 90 antiviral compounds against HCoV-OC43, SARS-CoV or MERS-CoV and 1862 inactive compounds	Collected from refs 22, 23, 25, 26, 27
Fine-tuning Set 1	A collection of experimentally tested active and inactive molecules against SARS-CoV-2	Collected from refs 10-14
Test Set 1	A subset of fine-tuning Set 1 with compounds presented in Training Set 1 removed	–
ReFRAME actives	Active compounds against SARS-CoV-2 obtained by experimentally screening the ReFRAME library	ref. 29
Fine-tuning Set 2	Combination of Fine-tuning Set 1 and ReFRAME actives	–
Drug Repurposing Hub	A curated and annotated collection of FDA-approved drugs, clinical trial drug candidates and pre-clinical compounds	https://clue.io/repurposing
Drug-like library from ZINC15	A subset database of ZINC15 containing ~4.9 million molecules that are drug-like and in stock	https://zinc15.docking.org/

COVIDVS-3 to screen potential antiviral molecules from large compound library and selected the molecules with best scores to make further evaluation. A summary of the training/test data sets for each COVIDVS model was listed in Table 2 and the details for each model will be discussed in a later section.

Our COVIDVS models implemented the framework of Chemprop model which has been used to predict molecular properties directly from the graph structure of molecules [30]. Based on Chemprop, we constructed a classifier containing a message-passing neural network (MPNN) module [31] and a feed-forward neural network (FNN) module [32]. The classifier takes molecular SMILES as input and converts it to a graph representation internally. Atoms and bonds are regarded as graph nodes and edges, respectively and a related feature vector is assigned to each atom and bond. The MPNN module aggregates all information from atoms and bonds to a molecule-level representation. The learned molecular featurization was then fed to the FNN module to make final prediction. The architectures of MPNN module and FNN module are shown in Figure 1B.

The ensemble method has been shown to be able to improve the performance of machine learning models [33]. An ensemble of N models is constructed by training the same model architecture for *N* times with different random initial weights. The predictions of the *N* models are usually averaged as the ensemble's prediction. Here, we applied the ensemble method to improve the performance of COVIDVS models.

Transfer learning (TL) is an AI technology that can be applied to resolve problems of data scarcity by leveraging existing knowledge from source tasks to a target task with low data [34]. Transfer learning have achieved success on low data tasks in many fields including computer vision [35], natural language processing [36, 37] and drug discovery [38, 39]. In the present study, we have only 154 data for SARS-CoV-2, which is obviously insufficient to train a model. We implemented fine-tuning technique, which is one of the most commonly used transfer learning techniques to deal with the data scarcity problem. We trained a source model from scratch with Training Set 1 and regarded it as the source task. Then we constructed the target model, whose weights were inherited from the source model and fine-tuned it with Fine-tuning Set. More details about the transfer learning are provided in Supplementary Data.

Results

The broad-spectrum anti-beta-coronavirus compound prediction model

We used the Training Set 1, which consists of 1952 compounds labeled by their activities against SARS-CoV, MERS-CoV or HCoV-OC43 to train a general classification model for anti-beta-CoV activity. Test Set 1 was used to evaluate the model's performance. The hyperparameters were defined by Bayesian hyperparameter optimization method. The MPNN part and the classification part were trained together, so that the model can learn to extract the molecular features and classify antiviral molecules automatically. However, the model showed a receiver operating characteristic curve-area under the curve (ROC-AUC) of 0.99 for Training Set 1 and an AUC of 0.71 for Test Set 1, which was obviously overfitting. In order to solve this problem, we concatenated an additional vector containing molecular features computed by RDKit [40] (Supplementary Table S6) to the molecular representation generated by MPNN module. The new model showed an AUC of 0.97 for Training Set 1 and an AUC of 0.89 for Test Set 1, demonstrating that the augmentation of data representation can significantly improve the model's performance. In order to further evaluate the performance of model, we trained models on the training data from each of the 10 different random splits of Training Set 1, each with 80% training data, 10% validation data and 10% test data, resulting in an average of AUC of 0.96 on the training data and 0.83 on the testing data (Figure S1). However, the test performance varies a lot when training with different split of data due to the limited training data size. In order to make the model more robust, we constructed ensembles containing 5, 10 and 20 models, respectively and tested their performance on Test Set 1. We found that applying the ensemble technique significantly improve the performance of model. The ensemble of 20 models achieved the best performance with a ROC-AUC of 0.89 on Test Set 1 (Figure 2A and Supplementary Table S1), indicating that this model can efficaciously discriminate actives and inactives for SARS-CoV-2. Therefore, we selected the ensemble of 20 models (COVIDVS-1) for further prediction.

We applied COVIDVS-1 to predict the anti-SARS-CoV-2 activity of compounds in a library containing 1417 launched drugs extracted from the DRH. Figure 2B gives the distribution of the predicted scores. Most of the launched drugs (89.3%) have scores less than 0.2. Among the 70 top-ranking (5%) drugs, 6 have

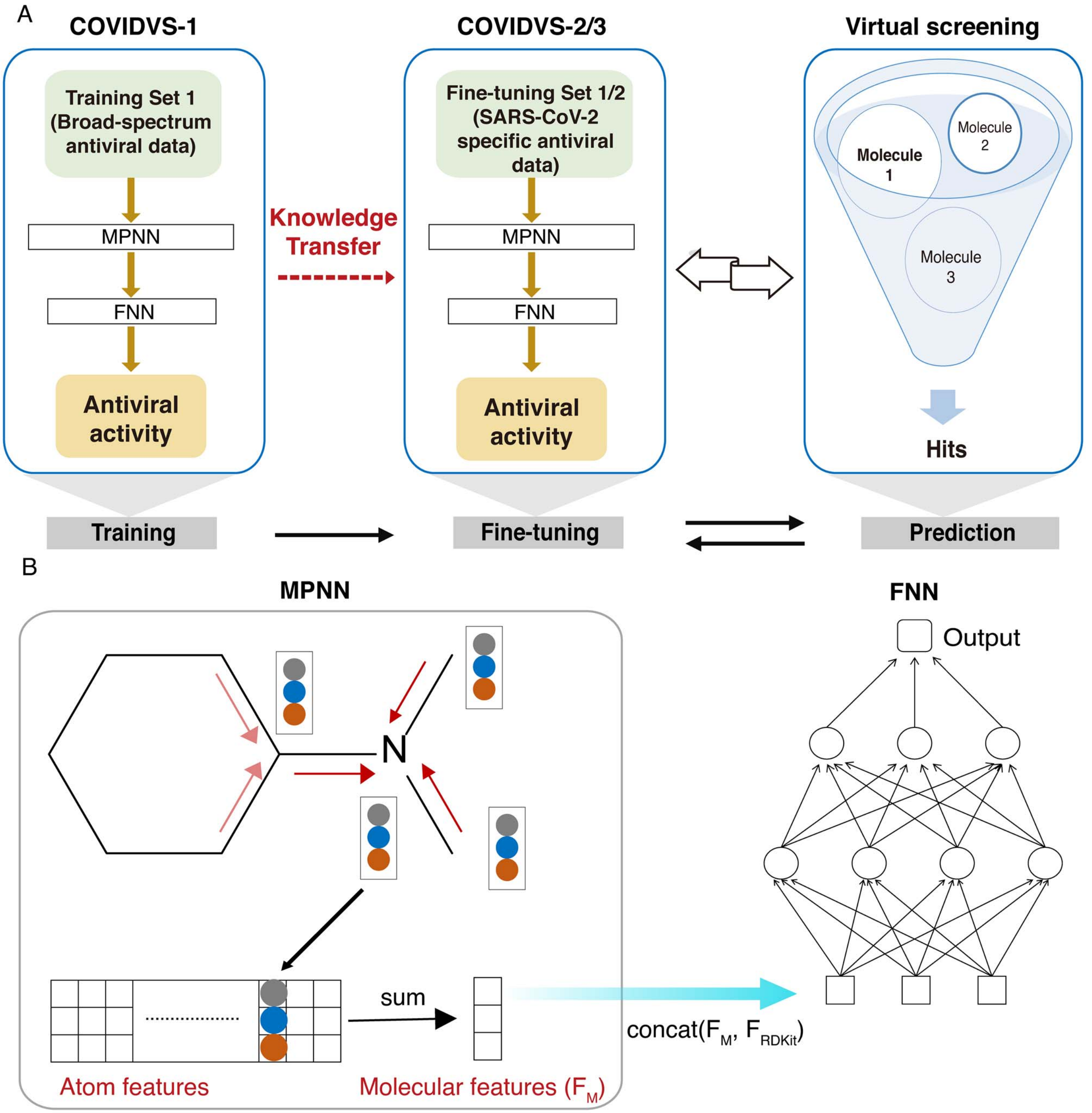

Figure 1. (**A**) The pipelines of our transferable deep learning based virtual screening model. (**B**) The architectures of MPNN module and FNN module. F_{RDKit} represents the additional vector containing 200 molecular features computed by RDKit. The molecular features calculated by MPNN and the F_{RDKit} were concatenated as the input of FNN module. MPNN: message passing neural network; FNN: feed-forward neural network.

been reported to be active against SARS-CoV-2 (Table 3). Ceritinib (also named LDK378), a drug that is used for the treatment of non-small-cell lung cancer, ranked at position 11 and has been reported to inhibit the replication of SARS-CoV-2 with an IC_{50} of 2.86 μM [10]. Terconazole, an antifungal drug that ranked at position 24, showed an IC_{50} of 11.92 μM [11]. Osimertinib, an anti-cancer drug that is used to treat non-small-cell lung carcinomas with a specific mutation, ranked at position 35 and has been shown to be active against SARS-CoV-2 with an IC_{50} of 3.26 μM [10]. Ritonavir, an antiretroviral medication used along with other medications to treat AIDS, showed 8.63 μM of EC_{50} and ranked at position 42 [14]. Abemaciclib, a drug for the treatment of breast cancers that showed potency against SARS-CoV-2 with an IC_{50} of 6.62 μM, ranked at position 46 [10]. Indinavir, a protease inhibitor used as a component of highly active antiretroviral therapy to treat AIDS, ranked at position 60 with a reported anti-SARS-CoV-2 EC_{50} of 59 μM [14]. These results demonstrated that COVIDVS-1 can successfully screen out potential antiviral drugs against SARS-CoV-2, even though the active compounds in the training set were only tested on beta coronaviruses other than SARS-CoV-2. We analyzed the chemical structure similarity between the six drugs and the active compounds in the Training

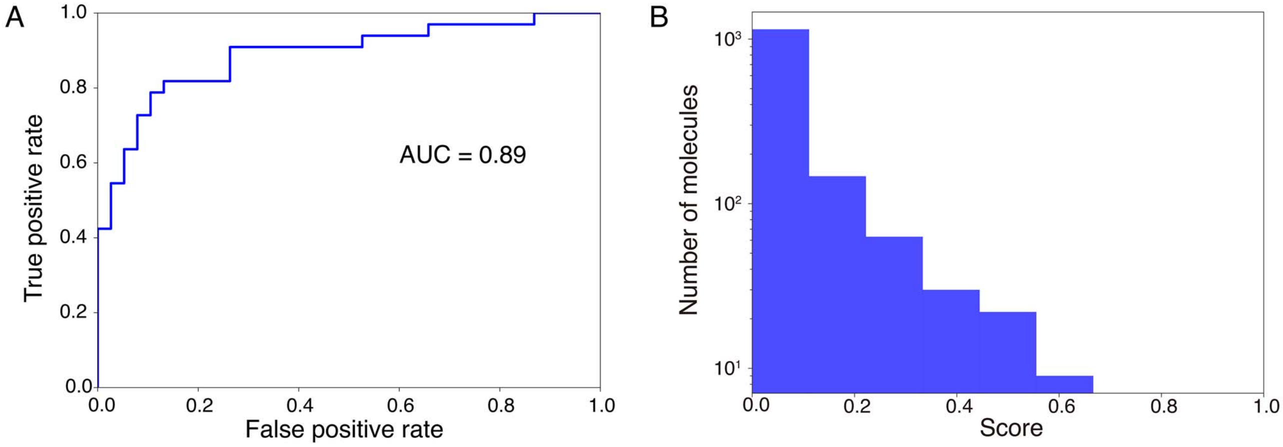

Figure 2. The performance and prediction results of COVIDVS-1 model. (**A**) ROC curve showing the performance of COVIDVS-1 on Test Set 1. (**B**) Histogram showing the distribution of predicted scores of the launched drugs library extracted from DRH. Molecules that are in Training Set 1 have been removed.

Table 2. Datasets used for constructing the COVIDVS models

	Training	Fine-tuning	Prediction
COVIDVS-1	Training Set 1	–	Test Set 1 and Launched drugs in DRH [20]
COVIDVS-2	–	Fine-tuning Set 1	ReFRAME actives [28] and DRH
COVIDVS-3	–	Fine-tuning Set 2	Drug-like library from ZINC15 [21]

Table 3. COVIDVS-1 predicted ranking positions of the six launched drugs reported to have antiviral activity against SARS-CoV-2 [10, 11, 14] among the 1417 launched drugs extracted from the DRH

Name	Experimental activity (EC_{50})(μM)	Ranking position
Ceritinib	2.86	11
Terconazole	11.92	24
Osimertinib	3.26	35
Ritonavir	8.63	42
Abemaciclib	6.62	46
Indinavir	59	60

Set 1 by Tanimoto similarity coefficient with Morgan Fingerprint (Calculated by RDKit). All these six retrieved active drugs have maximum similarity <0.4 to the active molecules in Training Set 1, indicating that the model can identify potential candidates with novel structures.

Development of anti-SARS-CoV-2 compound prediction model

As we have only 154 data for SARS-CoV-2, we applied transfer learning to develop the SARS-CoV-2 specific model. We used the Fine-tuning Set 1 to fine-tune our COVIDVS-1 model, resulting in the second-generation model, COVIDVS-2. Considering that the data size of Fine-tuning Set 1 is quite limited and overfitting may easily happen, we decided to freeze the MPNN parameters and only fine-tune the classification part. This strategy reduced the number of trainable parameters and we have demonstrated that it is less likely to cause overfitting than fine-tuning the whole COVIDVS-1 model (see Supplementary Data for details). COVIDVS-2 contains information from Training Set 2, which was constructed by adding all data in the Fine-tuning Set 1 to the Training Set 1. Molecules already existed in the Training Set 1 were relabeled according to their activity against SARS-CoV-2 and molecules not presented in the Training Set 1 were directly added. Training Set 2 contains 133 positive data and 1890 negative data. We analyzed the chemical space distribution of the Training Set 2 and data from the DRH using the t-distributed stochastic neighbor embedding (t-SNE) dimension reduction method. The t-SNE plots were created using the scikit-learn tools [41]. Tanimoto similarity was utilized to quantify the chemical distance. The 'perplexity' parameter, which controls the balanced attention between local and global aspects of the data, was set to 30 for an appropriate data projection. All other parameters were set to the scikit-learn's default values. The t-SNE plot showed that the positive data in the training set largely overlaps with the data from DRH in chemical space (Figure 3A). We then used this model to screen the full DRH dataset. The distribution of the predicted scores is given in Figure 3B. There are 280 molecules with score > 0.8 and 55 molecules with score > 0.9. The chemical structures of the 55 high-scored molecules are highly diverse (Figure 3A). About half of the top 55 molecules were reported kinase inhibitors, demonstrating the potential of using kinase inhibitors as anti-SARS-CoV-2 drugs. Six anaplastic lymphoma kinase (ALK) tyrosine kinase receptor inhibitors, three cyclin-dependent kinase (CDK) inhibitors and eight epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors were enriched in the top 55 list, which have the same targets to Ceritinib, Abemaciclib and Osimertinib that are active on SARS-CoV-2, respectively. We noticed a newly reported work that carried out a mass spectrometry-based phosphoproteomics survey of SARS-CoV-2 early infection. Dramatic rewiring of phosphorylation on host and viral proteins and altered activities of kinases were observed during the SARS-CoV-2 infection, making kinases to be ideal drug targets [42]. Compared to the ~40 known targets of the 55 molecules and the ~60 known targets of active molecules in Training Set 2, only 8 targets are the same,

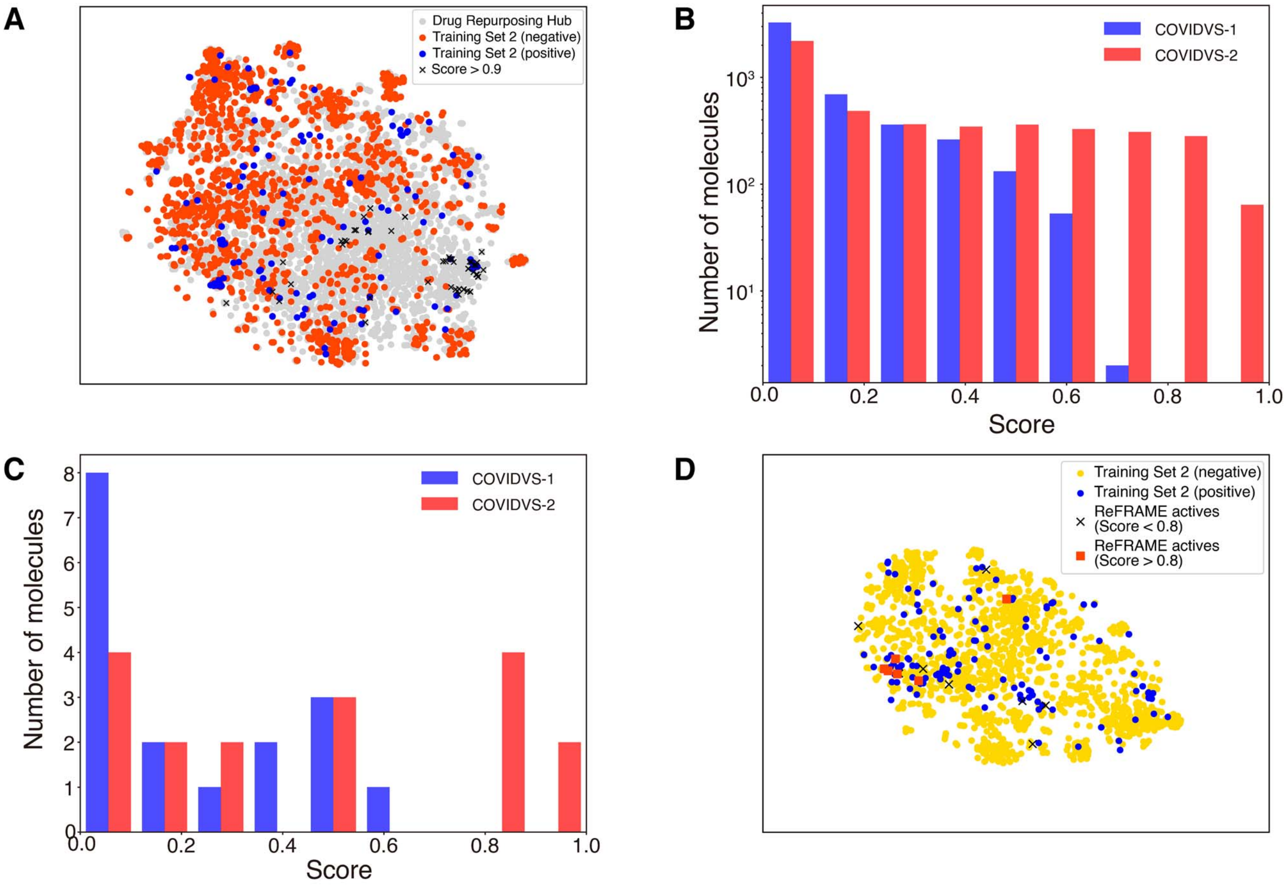

Figure 3. Comparison the prediction results between COVIDVS-1 and COVIDVS-2 (**B**, **C**) and the data distribution of Training Set 2, DRH and ReFRAME actives (**A**, **D**). (**A**) t-SNE of all molecules from the Training Set 2 (red: negative data, blue: positive data) and DRH (gray). The top 55 high-score molecules were also plotted (black cross). (**B**) Histogram showing the difference between prediction results of COVIDVS-1 and COVIDVS-2 on DRH. (**C**) Distribution of scores for 17 ReFRAME actives predicted with COVIDVS-1 and COVIDVS-2, respectively. (**D**) t-SNE of all molecules from the Training Set 2 (blue: positive data, gold: negative data) and ReFRAME actives (black cross: 11 molecules with score < 0.8, red squares: six molecules with score > 0.8).

demonstrating that the molecular targets of predicted results were not constrained by the training set. We listed all the 55 molecules with score > 0.9 in Supplementary Table S2 and grouped them according to their clinical study states.

To test the power of the fine-tuned model, we used the 17 newly identified ReFRAME actives to evaluate our COVIDVS-1 and COVIDVS-2. We predicted the ReFRAME actives with COVIDVS-2 and 6 of them have scores >0.8. Compound KW 8232 ($EC_{50} \sim 1.2$ μM) has a score of 0.94, which is above most of the 4711 DRH molecules. This suggests that our model can successfully discover novel antiviral drugs against SARS-CoV-2. We also compared the performance of COVIDVS-2 and COVIDVS-1 on ReFRAME actives (Figure 3C). Among the 17 compounds, six got predicted scores > 0.8 by COVIDVS-2, while no molecule got predicted score > 0.8 by COVIDVS-1. Of course, higher scores may not guarantee true activity. We mixed these 17 molecules into the 4711 molecules from the DRH and ranked all of them by their predicted scores. The top-2 compounds among the 17 ReFRAME actives ranked in 28th and 58th among all the 4728 compounds when predicted with COVIDVS-1, while the ranking raised to 4th and 29th when predicted with COVIDVS-2. We posted these ReFRAME actives onto the t-SNE plot of the Training Set 2. Although all the 6 compounds with good predictions are close to active compounds in the training set, some of the 11 compounds with predicted scores less than 0.8 are relatively far from the active compounds in the Training Set 2 (Figure 3D). This demonstrates that the diversity of active compounds limits the model's performance, which can be improved by increasing the number and chemical diversity of active compounds in the training set data.

Screening ZINC database to identify novel anti-SARS-CoV-2 compounds

Though several FDA-approved drugs have shown anti-SARS-CoV-2 activities using drug repurposing approaches, none of them were highly effective in clinical trials. Highly effective novel anti-SARS-CoV-2 drugs need to be developed. Deep learning models can be easily applied to deal with big data, which allows us to screen large chemical libraries. We subsequently applied our method to screen ZINC15 database. We fine-tuned the COVIDVS-1 model with the Fine-tuning Set 2 to derive the third-generation model, COVIDVS-3. Similar to Training Set 2, we constructed Training Set 3 by combining Fine-tuning Set 2 and Training Set 1, which contains all data that contributed to the training of COVIDVS-3. As all known active data are included in Training Set 3, COVIDVS-3 model is the best model we can obtain with the current data.

We applied COVIDVS-3 to screen the 4.9 million drug-like molecules selected from ZINC15. This screen run was finished

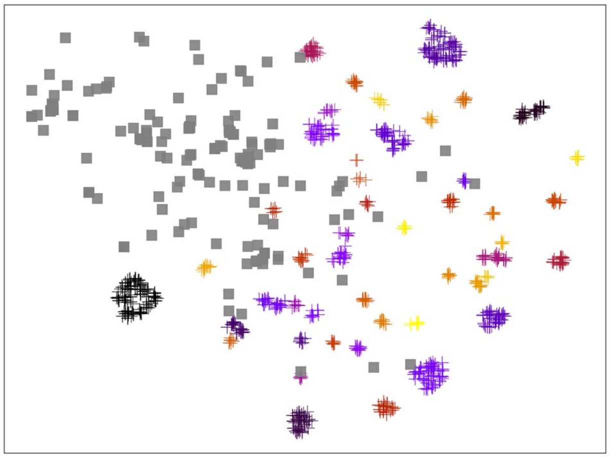

Figure 4. Clustering results of high-score ZINC15 molecules. Molecules from positive data of the Training Set 3 and the predicted ZINC15 molecules with score > 0.9 were plotted on t-SNE distribution plot. Gray squares represent the positive data of the Training Set 3. Plus markers with different colors represent different clusters. Noisy samples which are far away from any clusters are not plotted for clarity.

within 6 h using 200 CPUs for feature generation and 4 NVIDIA GPUs for prediction, which can be easily speeded up. We ranked all the 4.9 million molecules by their predicted scores. This gave 3641 molecules with score > 0.9, and 94.6% of them have the maximum similarity < 0.4 to positive data of Training Set 3. In order to understand the structure distribution relationship of the high-score molecules, we analyzed the chemical space distribution of the 3641 molecules and the positive data in Training Set 3. A density-based spatial clustering of applications with noise (DBSCAN) method [43, 44] was performed to cluster the 3641 ZINC molecules with score > 0.9 (see Supplementary Data for details). There are 46 clusters that have at least 10 molecules, and 8 clusters with at least 50 molecules. Figure 4 showed the clustering results on t-SNE plot, as well as the distribution of positive data of Training Set 3, revealing that the predicted compounds locate in different chemical space, showing the diversity of results. For each of the 8 clusters with at least 50 molecules, we selected one molecule with the best score as a representative compound (shown in Figure 5). The top 100 molecules with best prediction scores and representative molecules from the 46 clusters were given in Supplementary Tables S3 and S4, respectively. We suggest that these compounds can be tested for their anti-SARS-CoV-2 activities in future experimental studies. Our method can be easily applied to screen other large library with millions, even billions of compounds.

Identifying novel SARS-CoV-2 3C-like protease inhibitors

We have applied our COVIDVS model to screen ZINC15 database and predicted a set of potential antiviral molecules, which may act on different targets, including the SARS-CoV-2 3CL^{pro}. 3CL^{pro} plays an important role in mediating viral replication and transcription [45]. The sequence identity of 96.1% between 3CL^{pro} in SARS-CoV and SARS-CoV-2 makes it an ideal target for developing broad-spectrum anti-CoV drugs. In order to further screen 3CL^{pro} inhibitors from the prediction results, we performed molecular docking using Autodock Vina software [46]. The structure of SARS-CoV-2 3CL^{pro} (PDB ID 6 LU7) and candidate molecules were prepared with AutodockTools [47]. All the 3641 ZINC15 molecules with prediction score > 0.9 from the previous section were subjected to docking. Their docking scores ranges from −10.5 to −6.3 kcal/mol. From the top 40 results, we manually selected seven compounds to experimentally evaluate their activities.

We purchased these seven compounds and tested their SARS-CoV-2 3CL^{pro} inhibition activity (see Supplementary Data for experimental details). The prediction scores, docking scores and inhibition rates at 50 μM of the seven molecules were shown in Supplementary Table S5. Among all the seven compounds tested, ZINC000017053528 showed strong inhibition at 50 μM and has an IC$_{50}$ of 37.0 μM (Figure 6). To the best of our knowledge, no bioactivity of this molecule has been reported before. We calculated the 2D structure similarity between the active compound and the 405 reported SARS-CoV 3CL^{pro} inhibitors from PubChem AID1706 assay [48] with ECFP4 fingerprint [49]. All the known active compounds have the similarities less than 0.4.

In the past two decades, many synthetic compounds and natural products with inhibitory activity against coronaviruses' 3CL^{pro} have been reported [50, 51]. Most of them are covalent inhibitors targeting the active site Cys145. The currently reported SARS-CoV-2 3CL^{pro} inhibitors were mainly discovered by testing the previously developed 3CL^{pro} inhibitors or by drug repurposing screen. Only a few non-covalent inhibitors were reported with modest activity. For example, Jin *et al*. predicted that cinanserin as a potential inhibitor by docking-based virtual screening and the experimentally measured IC$_{50}$ value was 125 μM. They also found seven hits by high-throughput experimental screening method. Three of them are non-covalent inhibitors and their IC$_{50}$ values are in the range of 1.55–15.75 μM [17]. Alice *et al*. identified 23 non-covalent hits by performing a large-scale screen of fragments through a combined mass spectrometry and X-ray approach against the SARS-CoV-2 3CL^{pro}. These hits can be used to rapidly develop more potent inhibitors [52]. Recently, Yang *et al*. developed a ligand-based method named D3Similarity to evaluate molecular similarity between a submitted molecule and molecules in the active compound database containing all the reported bioactive molecules against the coronaviruses [53]. We submitted ZINC000017053528 to the D3Similarity web server to evaluate its 2D and 3D molecular similarity to the known bioactive molecules. ZINC000017053528 is not similar in chemical structure to all the bioactive molecules in the database (with 2D similarity < 0.4), although it showed certain degree of three-dimensional similarity with the 3D similarity scores between 0.64 and 0.75. Molecules in the top 10 similarity rankings (sorted by the product of 2D and 3D similarity) have seven different targets and four molecules are 3CL^{pro} inhibitors. As the chemical structure of ZINC000017053528 is different from all the reported 3CL^{pro} inhibitors as well as anti-coronavirus compounds, it provides a novel scaffold for further development and optimization of anti-SARS-CoV-2 compounds targeting 3CL^{pro}.

Discussion

The data scarcity problem is the major problem for developing anti-SARS-CoV-2 molecules prediction model. In this study, we successfully extended the data source from anti-SARS-CoV-2 molecules to anti-beta-coronavirus molecules with the help of transfer learning. Our model was firstly trained with a collection of broad-spectrum anti-beta-coronavirus compounds against SARS-CoV, MERS-CoV or HCoV-OC43 and then migrated the extracted knowledge to anti-SARS-CoV-2 prediction model

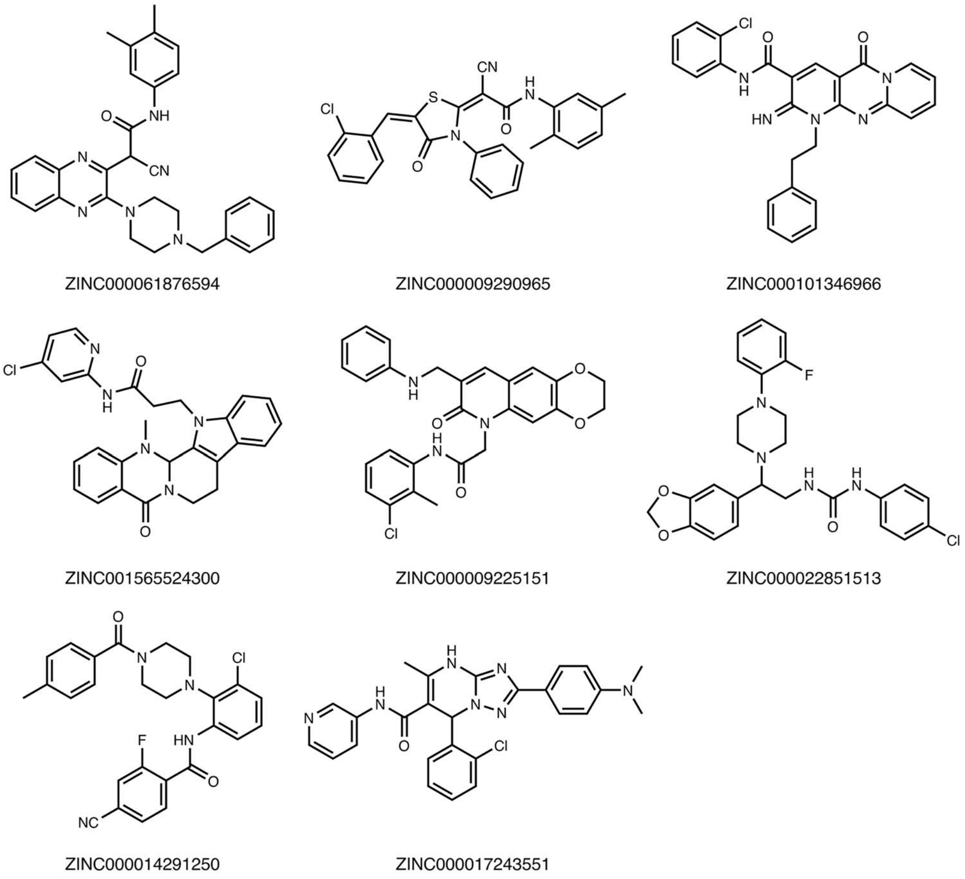

Figure 5. 2D structures of the eight ZINC15 molecules with best score from each of clusters that have at least 50 molecules.

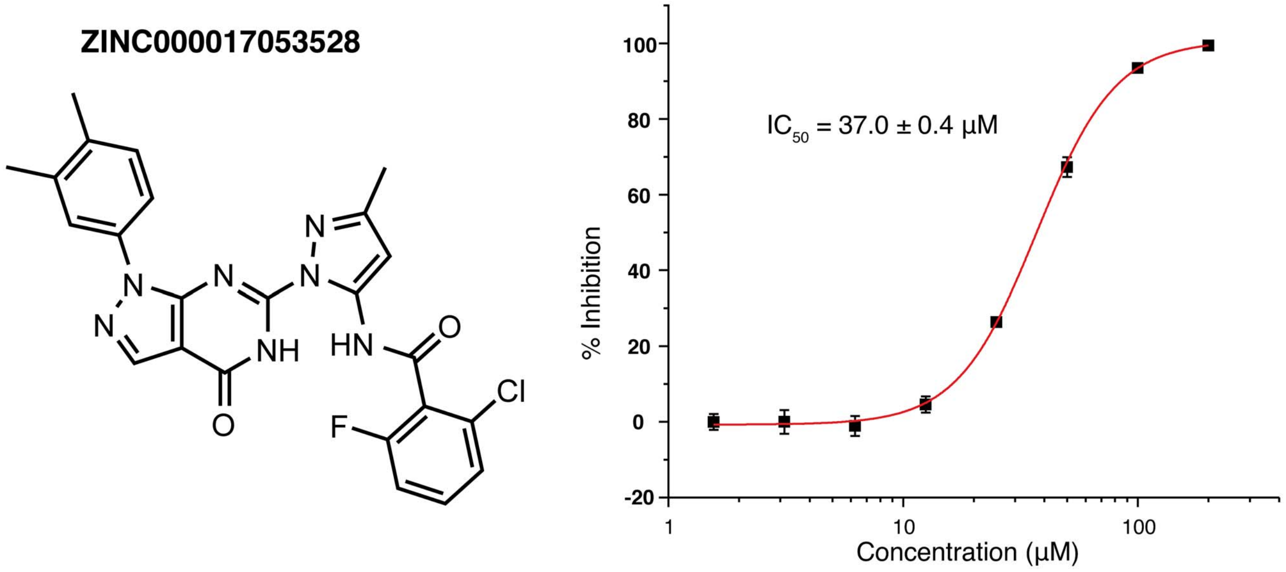

Figure 6. The chemical structure (left) and the *in vitro* anti-SARS-CoV-2 $3CL^{pro}$ activity (right) of molecule ZINC000017053528. Values represent the mean ± SE of two independent experiments, each based on three biological replicates.

through fine-tuning technique. The three types of beta-coronaviruses have been widely studied. Molecules screened for one of the CoVs often showed broad-spectrum anti-CoVs activities, indicating that these data can help us to discover new antiviral compounds for SARS-CoV-2. We have demonstrated that the broad-spectrum antiviral prediction model COVIDVS-1 trained with Training Set 1 can successfully screen out potential active molecules against SARS-CoV-2 in the top list of prediction results. Fine-tuning COVIDVS-1 with SARS-CoV-2 related data introduced more task-specific information and allow the model more suitable for our target task. It is expected that more data will further improve the prediction ability of transfer learning-based model. The experimental testing results of high score molecules predicted by fine-tuned model can in turn be used to fine-tune new model. The model performance can be improved by iterating the fine-tuning, model prediction and experimental estimation process. This strategy can be easily achieved when we are facing an interesting target lacking enough data. The newly discovered active compounds targeting SARS-CoV-2, as well as the other three coronaviruses, can also be used to enhance the performance of the broad-spectrum model, which can be applied to screen potential broad-spectrum drugs for newly emerged coronaviruses in the future.

We utilized ensemble technique to improve the performance and robustness of our COVIDVS model, however, it also increased the computational cost proportionally. Therefore, the balance between performance and cost should be taken into consideration. Although the ensemble of 20 models showed the best performance, we have demonstrated that the ensemble of 5 or 10 models are quite effective to improve the model's performance (Supplementary Table S1) and can be applied when screening ultra-large compound libraries to reduce the computational cost.

Due to experimental limitations, we tested if our predicted compounds contain inhibitors that specifically targeting SARS-CoV-2 3CL^{pro} *in vitro*. We used our COVIDVS prediction together with protein-ligand docking to screen potential SARS-CoV-2 3CL^{pro} inhibitors. The screening process on the 4.9 million drug-like compounds from ZINC15 took about 6 hours with 100 CPUs and 4 GPUs, and only the 3641 top-ranking molecules were subject to docking calculations. In a similar setting, traditional docking-based virtual screening method would need 2 days with 1000 CPUs. The difference of time cost will further be enlarged when screening larger compound library. Although many 3CL^{pro} inhibitors have been reported, most of them only showed activity in *in vitro* enzyme assay. As our COVIDVS models were trained with antiviral activity data, compounds with *in vitro* 3CL^{pro} inhibition activity and good COVIDVS prediction scores may have high probability of anti-viral activity. Similar to COVIDVS-2 and 3, a target-specific model for 3CL^{pro} can be trained by fine-tuning COVIDVS-1 with known 3CL^{pro} inhibitors and non-inhibitors, which is expected to increase the success rate of prediction.

COVID-19 remains as a global pandemic that is waiting for effective vaccines and drugs. A number of FDA-approved drugs and clinical-phase molecules are being tested in clinical trials. However, no magic bullets have been found yet. More efforts are necessary to identify safe and efficacious therapeutic solutions for COVID-19 and emerging CoV related diseases in the future. Here, we combine *in silico* method and *in vitro* experimental test to screen potential antiviral compounds from large virtual screening library and successfully identified an inhibitor against 3CL^{pro} target. We hope our method can help to accelerate the drug discovery process for SARS-CoV-2 and other challenged targets and diseases.

Availability of Data

The Training Set, Fine-tuning Set, Test Set and other related data can be downloaded via https://disk.pku.edu.cn:443/link/6F8366DBBA21B841CB7C8844E54471E8. The source code can be accessed via https://github.com/pkuwangsw/COVIDVS.

Key Points

- A broad-spectrum anti-beta-coronavirus drug prediction model was developed based on the experimentally identified SARS-CoV, MERS-CoV and HCoV-OC43 inhibitors.
- Specific anti-SARS-CoV-2 drug prediction model was derived by fine-tuning the broad-spectrum model with SARS-CoV-2 specific antiviral compounds.
- Potential anti-SARS-CoV-2 compounds were suggested for experimental testing by screening the ZINC drug-like library containing 4.9 million compounds.
- Several possible SARS-CoV-2 3C-like protease inhibitors were predicted from the suggested anti-SARS-CoV-2 compound list and one 3C-like protease inhibitor with novel chemical scaffold was found.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

We would like to thank Dr Xiaobing Deng, Dr Weilin Zhang, Dr Hao Liang, Weixin Xie and Hongyi Wang for collecting and sharing anti-beta-CoV actives/inactives. We would like to thank Chenjing Cai for her help in transfer learning technique. Parts of the computation and analysis were performed on the high-performance computing platform of the Peking-Tsinghua Center for Life Sciences, Peking University.

Funding

This work was supported in part by the Ministry of Science and Technology of China (2016YFA0502303), the National Natural Science Foundation of China (22033001, 21633001) and the Fundamental Research Funds for the Central Universities of China.

References

1. World Health Organization. *WHO Coronavirus Disease (COVID-19) Dashboard*. Available at: https://www.who.int/.
2. Zhu N, Zhang DY, Wang WL, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**(8):727–33.
3. Wu F, Zhao S, Yu B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 2020;**579**(7798):265–9.
4. Zumla A, Chan JFW, Azhar EI, *et al.* Coronaviruses - drug discovery and therapeutic options. *Nat Rev Drug Discov* 2016;**15**(5):327–47.
5. Zhou P, Yang XL, Wang XG, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 2020;**579**(7798):270–3.
6. Li Q, Guan XH, Wu P, *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. *N Engl J Med* 2020;**382**(13):1199–207.

7. Wang ML, Cao RY, Zhang LK, *et al*. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. *Cell Res* 2020;**30**(3):269–71.
8. Gao JJ, Tian ZX, Yang X. Breakthrough: chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Biosci Trends* 2020;**14**(1):72–3.
9. Cao B, Wang Y, Wen D, *et al*. A trial of lopinavir–ritonavir in adults hospitalized with severe COVID-19. *N Engl J Med* 2020;**382**:1787–99.
10. Jeon S, Ko M, Lee J, *et al*. Identification of antiviral drug candidates against SARS-CoV-2 from FDA-approved drugs. *Antimicrob Agents Chemother* 2020;**64**(7):e00819–20.
11. Weston S, Coleman CM, Haupt R, *et al*. Broad anti-coronavirus activity of Food and Drug Administration-approved drugs against SARS-CoV-2 *in vitro* and SARS-CoV *in vivo*. *J Virol* 2020;**94**(21):e01218–20.
12. Touret F, Gilles M, Barral K, *et al*. *In vitro* screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. *Sci Rep* 2020;**10**(1):13093.
13. Fintelman-Rodrigues N, Sacramento CQ, Ribeiro Lima C, *et al*. Atazanavir, alone or in combination with ritonavir, inhibits SARS-CoV-2 replication and proinflammatory cytokine production. *Antimicrob Agents Chemother* 2020;**64**(10):e00825.
14. Yamamoto N, Matsuyama S, Hoshino T, *et al*. Nelfinavir inhibits replication of severe acute respiratory syndrome coronavirus 2 *in vitro*. *bio Rxiv* 2020. doi: 10.1101/2020.04.06.026476.
15. H-x S, Yao S, W-f Z, *et al*. Anti-SARS-CoV-2 activities *in vitro* of Shuanghuanglian preparations and bioactive ingredients. *Acta Pharmacol Sin* 2020;**41**(9):1167–77.
16. Liu H, Ye F, Sun Q, *et al*. Scutellaria baicalensis extract and baicalein inhibit replication of SARS-CoV-2 and its 3C-like protease in vitro. *J Enzyme Inhib Med Chem* 2021;**36**(1):497–503.
17. Jin Z, Du X, Xu Y, *et al*. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* 2020;**582**: 289–93.
18. Stokes JM, Yang K, Swanson K, *et al*. A deep learning approach to antibiotic discovery. *Cell* 2020;**180**(4):688–702 e13.
19. Ton AT, Gentile F, Hsing M, *et al*. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 13 Billion compounds. *Mol Inform* 2020;**39**(8):e2000028.
20. Corsello SM, Bittker JA, Liu ZH, *et al*. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med* 2017;**23**(4):405–8.
21. Sterling T, Irwin JJ. ZINC 15–ligand discovery for everyone. *J Chem Inf Model* 2015;**55**(11):2324–37.
22. Pillaiyar T, Meenakshisundaram S, Manickam M. Recent discovery and development of inhibitors targeting coronaviruses. *Drug Discov Today* 2020;**25**(4):668–88.
23. Shen L, Niu J, Wang C, *et al*. High-throughput screening and identification of potent broad-spectrum inhibitors of coronaviruses. *J Virol* 2019;**93**(12):e00023–19.
24. de Wilde AH, Jochmans D, Posthuma CC, *et al*. Screening of an FDA-approved compound library identifies four small-molecule inhibitors of middle east respiratory syndrome coronavirus replication in cell culture. *Antimicrob Agents Chemother* 2014;**58**(8):4875–84.
25. Ko M, Chang SY, Byun SY, *et al*. Screening of FDA-approved drugs using a MERS-CoV clinical isolate from South Korea identifies potential therapeutic options for COVID-19. *Viruses* 2021;**13**(4):651.
26. Liang R, Wang L, Zhang N, *et al*. Development of small-molecule MERS-CoV inhibitors. *Viruses* 2018;**10**(12):721.
27. Shin JS, Jung E, Kim M, *et al*. Saracatinib inhibits middle east respiratory syndrome-coronavirus replication *in vitro*. *Viruses* 2018;**10**(6):283.
28. Riva L, Yuan S, Yin X, *et al*. Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* 2020;**586**(7827):113–9.
29. Janes J, Young ME, Chen E, *et al*. The ReFRAME library as a comprehensive drug repurposing library and its application to the treatment of cryptosporidiosis. *Proc Natl Acad Sci* 2018;**115**(42):10750–5.
30. Yang K, Swanson K, Jin WG, *et al*. Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 2019;**59**(8):3370–88.
31. Gilmer J, Schoenholz SS, Riley PF, *et al*. *Neural message passing for quantum chemistry*. In: *Proceedings of the 34th International Conference on Machine Learning, P. Doina and T. Yee Whye (eds) 2017*. PMLR – Proceedings of Machine Learning Research, 1263–72.
32. Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. *Chemom Intell Lab Syst* 1997;**39**(1):43–62.
33. Dietterich TG. Ensemble methods in machine learning. In: *International Workshop on Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.
34. Pan SJ, Yang QA. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010;**22**(10):1345–59.
35. Kan MN, Wu JT, Shan SG, *et al*. Domain adaptation for face recognition: targetize source domain bridged by common subspace. *Int J Comput Vis* 2014;**109**(1–2):94–109.
36. Li T, Zhang Y, and Sindhwani V. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*. 2009. *Suntec*, Singapore: Association for Computational Linguistics.
37. Dai W, Xue G-R, Yang Q, *et al*. Co-clustering based classification for out-of-domain documents. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2007. San Jose, California, USA: Association for Computing Machinery.
38. Yang X, Wang YF, Byrne R, *et al*. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem Rev* 2019;**119**(18):10520–94.
39. Cai C, Wang S, Xu Y, *et al*. Transfer learning for drug discovery. *J Med Chem* 2020;**63**(16):8683–94.
40. RDKit: open-source cheminformatics. Available at: https://www.rdkit.org/docs/index.html.
41. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
42. Bouhaddou M, Memon D, Meyer B, *et al*. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell* 2020;**182**(3):685–712.
43. Martin Ester, Hans-Peter Kriegel, Jiirg Sander, *et al*. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press* 1996: 226–31.
44. Schubert E, Sander J, Ester M, *et al*. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans Database Syst* 2017;**42**(3):1–21.

45. Anand K, Palm GJ, Mesters JR, *et al*. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *EMBO J* 2002;**21**(13):3213–24.
46. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem* 2010;**31**(2):455–61.
47. Morris GM, Huey R, Lindstrom W, *et al*. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009;**30**(16):2785–91.
48. National Center for Biotechnology Information. *PubChem Bioassay Record for AID 1706, Source: The Scripps Research Institute Molecular Screening Center*. PubChem. Available at: https://pubchem.ncbi.nlm.nih.gov/bioassay/1706.
49. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010;**50**(5):742–54.
50. Li C, Sun Q, Lu Y, *et al*. Advances in enzymatic activity regulation mechanism and inhibitor discovery of Coronavirus 3C-like protease. *Sci Sin Chim* 2020;**50**:1250–79.
51. Liu Y, Liang C, Xin L, *et al*. The development of Coronavirus 3C-Like protease (3CL(pro)) inhibitors from 2010 to 2020. *Eur J Med Chem* 2020;**206**:112711.
52. Douangamath A, Fearon D, Gehrtz P, *et al*. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat Commun* 2020;**11**(1):1–11.
53. Yang Y, Zhu Z, Wang X, *et al*. Ligand-based approach for predicting drug targets and for virtual screening against COVID-19. *Brief Bioinform* 2021;**22**(2):1053–64.