Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2022**; 27(4): 113
https://doi.org/10.31083/j.fbl2704113

Original Research

Potential SARS-CoV-2 nonstructural proteins inhibitors: drugs repurposing with drug-target networks and deep learning

Shayan K. Azmoodeh[1], Igor F. Tsigelny[2,3,4,*], Valentina L. Kouznetsova[2,4]

[1]REHS Program, San Diego Supercomputer Center, University of California, San Diego, CA 92093-0505, USA
[2]San Diego Supercomputer Center, University of California, San Diego, CA 92093-0505, USA
[3]Department of Neurosciences, University of California, San Diego, CA 92093-0505, USA
[4]Biomedical Information Analysis, La Jolla, CA 92038-2525, USA
*Correspondence: itsigeln@ucsd.edu (Igor F. Tsigelny)
Academic Editor: Sang Heui Seo
Submitted: 28 December 2021 Revised: 14 January 2022 Accepted: 14 January 2022 Published: 1 April 2022

Abstract

Background: In the current COVID-19 pandemic, with an absence of approved drugs and widely accessible vaccines, repurposing existing drugs is vital to quickly developing a treatment for the disease. **Methods**: In this study, we used a dataset consisting of sequences of viral proteins and chemical structures of pharmaceutical drugs for known drug–target interactions (DTIs) and artificially generated non-interacting DTIs to train a binary classifier with the ability to predict new DTIs. Random Forest (RF), deep neural network (DNN), and convolutional neural networks (CNN) were tested. The CNN and RF models were selected for the classification task. **Results**: The models generalized well to the given DTI data and were used to predict DTIs involving SARS-CoV-2 nonstructural proteins (NSPs). We elucidated (with the CNN) 29 drugs involved in 82 DTIs with a 97% probability of interaction, 44 DTIs of which had a 99% probability of interaction, to treat COVID-19. The RF elucidated 6 drugs involved in 17 DTIs with a 90% probability of interacting. **Conclusions**: These results give new insight into possible inhibitors of the viral proteins beyond pharmacophore models and molecular docking procedures used in recent studies.

Keywords: SARS-CoV-2; COVID-19; drug-target interactions; machine learning

1. Introduction

Since December 2019, COVID-19 has caused a global pandemic, affecting millions of lives in over 210 countries and territories. There are currently several vaccines available but there is an absence of other treatments for this virus. Due to the structural similarity between SARS-CoV-2 and other *Betacoronavirudae*, such as SARS-MERS and SARS-CoV (although it is much more similar in structure to SARS-CoV [1]), many previously established drugs are being researched to repurpose them for the current pandemic [2]. This allows for more rapid drug discovery and approval, which is vital in the current emergency.

There are many different *in-silico* methods by which this could be done. Docking and molecular screening/modeling have been widely used to discover potential treatments for the novel coronavirus as well as for other diseases in studies such as [3–7], among others. Additionally, several studies [8–15] have used machine learning and artificial intelligence to predict drug–target interactions (DTI) for various viruses, including SARS-CoV-2, with deep neural networks (DNN), support vector machines (SVM), and random forest (RF) classifiers, among others, as detailed by [16]. Studies such as [10] have employed methods like ours, using a convolutional neural network to predict drug target interactions; additionally, other studies have used other machine-learning methods such as Naïve Bayes to carry out the classification task [15]. On the other hand, studies such as [13] employed a regression model, as opposed to a binary classification model, to predict the binding scores of ligands against the SARS-CoV-2 viral proteins. Other similarity-based methods such as network-based inference and K nearest neighbor also have been utilized for this task, as they are often relatively less computationally intensive [17,18].

There have been efforts to repurpose currently approved drugs to inhibit the virus's structural and nonstructural proteins by preventing the virus from entering the cell, preventing it from activating, or preventing it from replicating itself (these are the preferred drugs) [19]. SARS-CoV-2 has four structural proteins and sixteen nonstructural proteins (NSP) that carry out various tasks essential to the virus's ability to infect individuals. In theory, all the NSPs can be exploited as drug targets, impeding the virus's ability to carry out its harmful functions in the host cell; however, some are more viable targets than others due to the availability of their crystal structures or their importance in the life cycle of the virus [1]. It is also possible to inhibit host-based targets that facilitate the virus's entry into the host cell, such as the angiotensin receptor enzyme 2 (ACE2), As a result, studies are emerging that consider this a potential way to treat the disease [1]. Recently it has been found that the TMPRSS2 enzyme in the host cell allows the virus to en-

ter the cell by priming the spike proteins, which is a promising target in developing/repurposing a drug [20,21]. Drugs that inhibit the RNA-dependent RNA polymerase (RdRp, NSP12) are also being considered as possible treatments of the virus; these include ribavirin, remdesivir, sofosbuvir, and IDX-184 [2,19,21]. NSP12 is an attractive target due to its role in RNA replication in the life cycle of the virus and the availability of its crystal structure [21]. However, it is also possible to inhibit the virus before the NSPs (including NSP12) have been cleaved from the polyproteins 1a and b (pp1a and pp1b). The 3-chymotrypsin-like protease ($3CL^{pro}$), also called the main protease (M^{pro}) or NSP5, and the papain-like protease (PL^{pro}, NSP3) of SARS-CoV-2, which are both encoded by the open reading frames (ORF 1a/b), have gathered a lot of attention as a possible target of an inhibitor due to their leading role in the replication and growth of the virus as they cleave the viral polyproteins to form the other nonstructural proteins [1,4,20]. Disrupting this process would affect the virus's life cycle, effectively disabling it from further infecting the host. Drugs such as fostamatinib, oberadilol, ribavirin, remdesivir, and itacitinib have been proposed as potential $3CL^{pro}$ inhibitors through computational methods [6,22,23]. Similarly, nilotinib, levomefolic acid, and glecaprevir have been predicted as possible inhibitors of PL^{pro} [3,24]. Interestingly, the drug ziprasidone, originally used to treat schizophrenia, has shown effectiveness against both the major viral proteins [3,22,25].

Considering this, in our study we look to repurpose approved drugs to inhibit *all* SARS-CoV-2 NSPs using a machine-learning approach that takes advantage of structural similarities between viral proteins and similarities between pharmaceutical drugs. This method allows for high-throughput DTI prediction greatly aiding the fight against the virus.

2. Methods

The methods used in this project are presented in Fig. 1. The programs developed are presented at https://github.com/Shkev/Sars-CoV-2-NSP-Predictions.

2.1 Datasets

2.1.1 Data collection

Experimentally verified DTIs for approved drugs, along with the drug SMILES and viral protein sequences, were downloaded from the DrugBank website [26] (release 5.1.7). In total, 19,242 DTIs were collected, involving 2468 drugs and 5177 proteins including those from various viruses (influenza, HIV, SARS-CoV, *etc.*) as well as human proteins.

All SMILES names were converted to canonical SMILES using Open Babel [27] to standardize them and allow easier handling of the data.

SARS-CoV-2 NSPs sequences were downloaded from the NCBI protein databank.

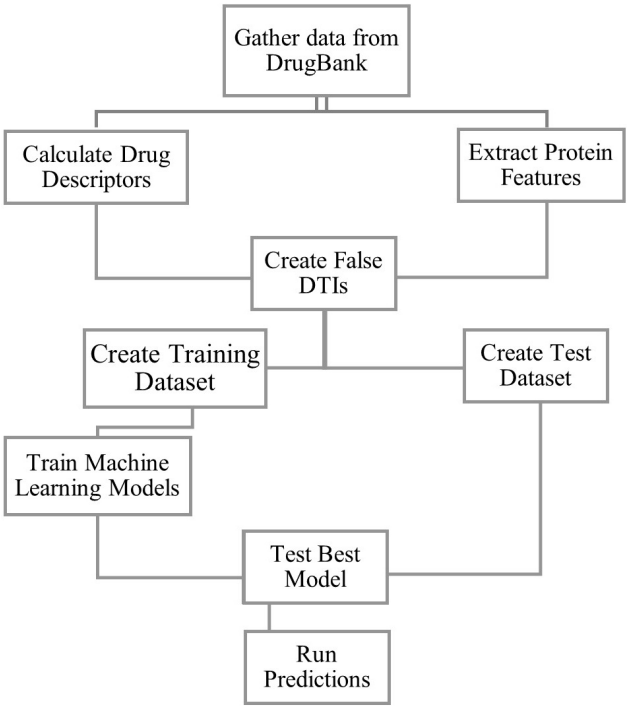

Fig. 1. Methods flowchart used in this study.

2.1.2 Extracting drug and protein features

The dataset contained chemical structures of the drugs in the form of SMILES names. Two-dimensional (2D) drug descriptors were calculated using the Online Chemical Database [28]. These descriptors contain the direct connections between the structures of the drugs and their properties, providing sufficient information to train the machine-learning model to recognize patterns in the data [29]. Drugs for which descriptors could not be calculated were removed along with any DTIs they were involved in, leaving 2444 drugs and 16,640 DTIs.

It is known that sufficient information about proteins is contained in the amino-acid sequences. Hence, we used common sequence descriptors and domain information to represent the proteins in our dataset [30]. The protein-sequence descriptors consisted of the amino-acid composition (AAC), dipeptide composition (DC), and tripeptide composition (TC). AAC is the frequency of each amino acid in the sequences. DC is the frequency of each possible pair of two amino acids in the sequences. TC is the frequency of each possible triplet of amino acids in the sequences. In addition to these, the domain information for each protein was obtained from the NCBI Batch Conserved Domain search and was used to construct an adjacency matrix. Each column and row represented one of the T target proteins, creating a $T \times T$ matrix of values (Fig. 2). Protein pairs that shared at least one domain were assigned a value of 1, and all other values were set to 0. Proteins that did not have any domain data were removed. Each row in the matrix, corresponding to the i-th target protein, is the domain data portion of that protein's feature vector. In total, we ex-

tracted 13,705 protein features consisting of 22 AAC, 441 DC, 8089 TC, and 5153 domain features. The domain data and sequence descriptors were combined for each protein, yielding its vector representation (Eqn. 1),

$$[\mathrm{P}_{\mathrm{AAC11}}, \ldots, \mathrm{P}_{\mathrm{AAC22}}, \mathrm{P}_{\mathrm{DC1}}, \ldots, \mathrm{P}_{\mathrm{DC441}}, \mathrm{P}_{\mathrm{TC1}}, \ldots, \mathrm{P}_{\mathrm{TC8089}}, \mathrm{P}_{\mathrm{Domain\ 1}}, \ldots, P_{\mathrm{Domain5153}}\,] \quad (1)$$

	t_1	t_2	t_3	$\ldots$	t_T
t_1	1	0	1	0	0
t_2	0	1	0	0	1
t_3	1	0	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_T	0	1	0	$\ldots$	1

Fig. 2. Example $T \times T$ matrix of domain features where t_i represents the *i*-th target protein. A 1 is assigned to a target–target pair if they share at least one domain, otherwise there is a 0. The row for target t_i is part of the feature vector for the *i*-th targets.

The respective protein and drug descriptors were combined to form one numerical vector for each DTI. Thus, each DTI was an array of approximately 25,000 values presented in the consequent set as in Eqn. 2.

$$[\mathrm{D}_1, \mathrm{D}_2, \ldots, \mathrm{P}_{\mathrm{AAC1}}, \ldots, \mathrm{P}_{\mathrm{AAC22}}, \mathrm{P}_{\mathrm{DC1}}, \ldots, \mathrm{P}_{\mathrm{DC441}}, \mathrm{P}_{\mathrm{TC1}}, \ldots, \mathrm{P}_{\mathrm{TC8089}}, \mathrm{P}_{\mathrm{Domain\ 1}}, \ldots, P_{\mathrm{Domain5153}}\,] \quad (2)$$

2.1.3 Negative DTI creation

Data from DrugBank supplied experimentally verified DTIs, however, to train a machine-learning model, we also need a set of false DTIs. To achieve this, all possible combinations between the D drugs and the T targets in the dataset were created, yielding $D \times T$ DTIs (approximately 12.7 million DTIs). This list was filtered to remove all such combinations that were contained in the true DTIs. These artificial false DTIs were randomly sub-sampled, and their number was the same as the number of true DTIs, giving approximately 32,000 data points in total. This sub-sampling also reduces the probability that the negative DTIs selected for our dataset are unidentified positive DTIs (that are yet to be experimented). The approximately 16,000 selected negative DTIs represent 0.126% of the total number of negative DTIs; therefore, there is a high degree of randomness in sub-sampling this sub-sampling process, making it unlikely that the selected negative DTIs are unidentified positive DTIs.

The artificially created set of negative DTIs was combined with the verified set. Positive DTIs were assigned a label of 1 and the artificial negative DTIs were assigned a label of 0, which allowed for binary classification. The data was randomly split into a training (70% of data) and a testing dataset (30%), ensuring that both datasets contained an approximately balanced number of both classes. The testing dataset was set aside and not considered in the development of the model as it was intended to represent a set of independent outside data.

2.1.4 Data preprocessing

The DTI data vectors were preprocessed to shift the mean of each feature to 0 and remove those that provide little information about patterns in the data. Doing so reduces bias in giving more importance to some features over others when training a neural network. Note that the protein features were not adjusted, as this would remove the value in the frequency counts and adjacency matrix since the units for these values are already standardized. The mean-adjusted value for the j-th value of the i-th descriptor was calculated using (Eqn. 3), where μ_i is the mean of the i-th descriptor across all data points (standard deviation scaling was not used in this process as this would remove the ability to reduce features using variance).

$$d_{i,j.\mathrm{standard}} = d_{i,j} - \mu_i. \quad (3)$$

Preliminary feature reduction was performed on the mean-adjusted drug data and the protein sequence descriptors using a variance threshold. Namely, all features with variance less than the chosen threshold of 0.01 were removed. Domain features were not reduced in this way, as they were sparse (more than half the values per feature are 0) and thus would result in a near 0 variance for all such features.

2.1.5 Lasso-based feature reduction

Feature reduction is a key step in using the data to train a machine-learning model effectively by removing the features with the least influence on the model. This improves the efficiency and effectiveness of the model, as processing high-dimensional data is computationally expensive [11]. In this study, we implemented a Lasso linear model using the SciKit-Learn Python library [31] to filter out the most informative features [30]. This algorithm is a modified linear regression that attempts to minimize the coefficients of terms that are least informative to the model to 0. Thus, features with coefficients lower than a chosen threshold in the trained model should be removed, leaving the most prominent features.

Drug features and protein features were considered separately using two Lasso models. Models were trained and validated using the training dataset. The Lasso model contains a regularization parameter, α, which effectively

controls how aggressively the model reduces feature coefficients to 0 (higher value results in more aggressive feature selection). To find the optimal value for α, we iterated through values from 1 to 10^{-8} using 5-fold cross validation to test accuracy and area under the receiver operator curve (AUROC) of the models. The chosen value for this parameter was the point where, as α decreased, there was minimal or no improvement in the AUC of the model. For the drug model, α was chosen as 10^{-4} and for the protein model it was chosen as 10^{-3}. Features in the trained protein model with a non-zero coefficient were selected giving 1228 values. Similarly, features with coefficients greater than 10^{-3} were selected from the trained drug model giving 1820 values; this threshold was selected to ensure a balance between the number of protein and drug features in the final dataset. The selected features were concatenated and used as the input data for our models.

2.2 Classification model

Classification models were trained for the DTI classification problem. All data handling in the process was done with the Pandas Python library [32]. The models were trained on the approximately 32,000 data points, half of which are the positive DTIs extracted from DrugBank and the other half are sub-sampled negative DTIs, each a vector with 3048 values along with a label, 0 or 1, to distinguish between positive and negative DTIs. The output of the model is a probability that the inputted DTI is positive.

Three machine-learning models were trained and tested. The best performing one was used in the final SARS-CoV-2 NSP DTI predictions. A deep neural network (DNN) [30], random forest classifier (RF) [33], and convolutional neural network (CNN) [34] were tested. A part of the training dataset (30%) was randomly separated from the rest of the data to create a validation dataset that was used to tune our models' hyperparameters and optimize metrics. This tuning was done manually, individually adjusting the various hyperparameters of the models (using guidance from [35]), until the desired training/validation metrics were obtained. The AUROC and the binary accuracy of the models was used to compare them. The binary accuracy was calculated as the percentage of predictions that were consistent with their corresponding known value in the testing/validation datasets using a threshold of 0.5 (model predictions greater than or equal to 0.5 were considered as 1 and all others as 0). Since accuracy only measures the performance of the model at a single threshold, we also utilize the AUROC score of the model, as it measures the performance of the model at various thresholds, in order to better judge the viability of the models.

Random Forest was implemented using the Scikit-Learn Python library (version 0.24) [31]. The model was trained with 100 trees with no maximum depth, a minimum of 2 samples to split an internal node, a minimum of 1 sample required to be at a leaf node, and all other default parameters (which can be seen in the documentation).

2.2.1 Deep-learning models

We implemented a DNN and CNN using the TensorFlow Python library [36].

The DNN architecture consisted of an input layer, two hidden dense layers with 4096 nodes each, and an output layer. The rectified linear activation function (ReLU) was applied to each hidden layer, and a sigmoid activation was used on the output layer to yield a value between 0 and 1. Additionally, dropout layers of 50% and Ridge regression (L2) regularization were used in the hidden layers to reduce overfitting [30]. Two hidden layers of equal size were used as recommended by [35]. Different numbers of nodes were experimented with, and the value that resulted in the least overfitting (as determined by comparing training and validation metrics) was used, namely 4096 nodes per layer. In general, as the number of nodes increased, the training accuracy and AUROC increased while these validation metrics suffered. Likewise, as the number of nodes decreased, the training metrics fell, but there was less overfitting. The model was trained with a learning rate of 0.00001 and a batch size of 32.

The CNN model is like the DNN architecture with the addition of 1D convolutional (Conv1D) layers. This allows the model to extract hidden patterns in the data that would otherwise not be recognized by the dense layers. We implemented a three-layer convolutional network that outputs to a fully connected dense layer (with 2048 nodes) that feeds into the output layer. The number of filters in the Conv1D layers increased from 16, 32, to 64 (each with kernel size 3) to progressively learn more features from the data. Each Conv1D layer fed into a Batch Normalization layer and a Max Pooling layer with pool size 3 to normalize weight values and prevent the model from overfitting. The ReLU activation function was also used on all Conv1D layers and hidden layers. The sigmoid activation function was applied to the output layer. A dropout of 50% was added to the flattened output of the final convolutional layer and dense layers, while L2 regularization was applied to all convolutional and dense layers. This model was also trained with a learning rate of 0.00001 and a batch size of 32.

As can be seen in Table 1, the CNN model performed best in both metrics, hence it was used in predicting DTIs involving SARS-CoV-2 nonstructural proteins.

Table 1. Validation metrics of the three tested models.

Machine-learning model	Validation AUROC	Validation accuracy
Random Forest	0.955	0.889
DNN	0.930	0.856
CNN	0.991	0.969

2.2.2 Testing the models

The testing dataset was used to test the best-performing model (CNN) on a partition of the DrugBank data that the model has not seen before (the labels for the testing input are known, so the accuracy and AUC of the model can be calculated). There was an approximately equal number of each class (positive and negative DTIs) in the test dataset, with 4992 negative DTIs and 4989 positive DTIs. The accuracy and AUC from these predictions give an accurate representation of how the model will perform when making predictions from the SARS-CoV-2 NSP data. The CNN scored very highly on this dataset, showing that it generalized well from the training data, which makes it viable to use in predicting new DTIs. The Random Forest classifier performed slightly worse in these metrics but outperformed the CNN in recall/true positive rate and F-measure, which are valuable metrics in this use-case as it is important that the predicted positive DTIs are predicted correctly (true positives).

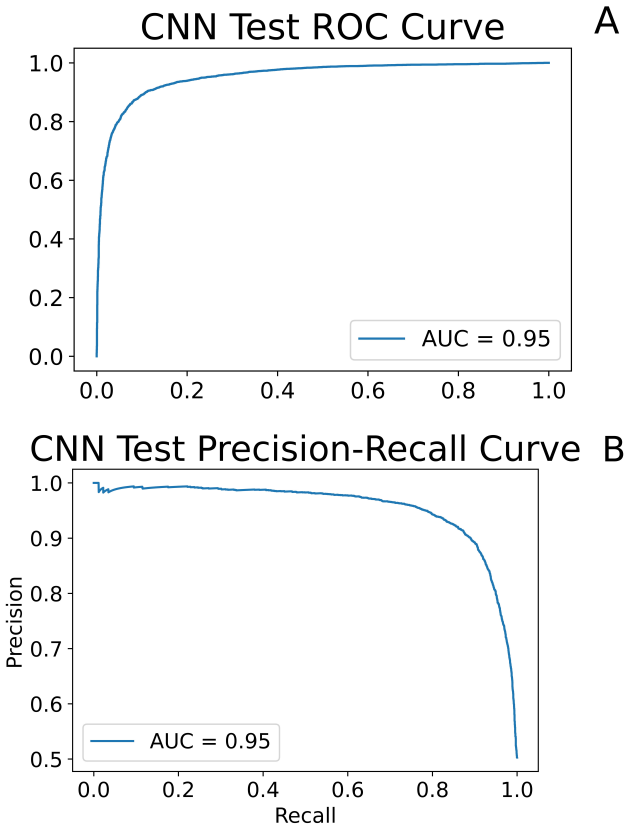

Fig. 3. Performance of a Convolution Neural Network (CNN) classification model. (A) ROC curve of CNN on the testing data. The area under the ROC curve is 0.95, which shows that our model generalized well to the training data and did not overfit. (B) Precision-Recall plot of CNN on the testing data. The area under the curve is 0.95, which shows that our model generalized well to the training data and performs well at different truth thresholds.

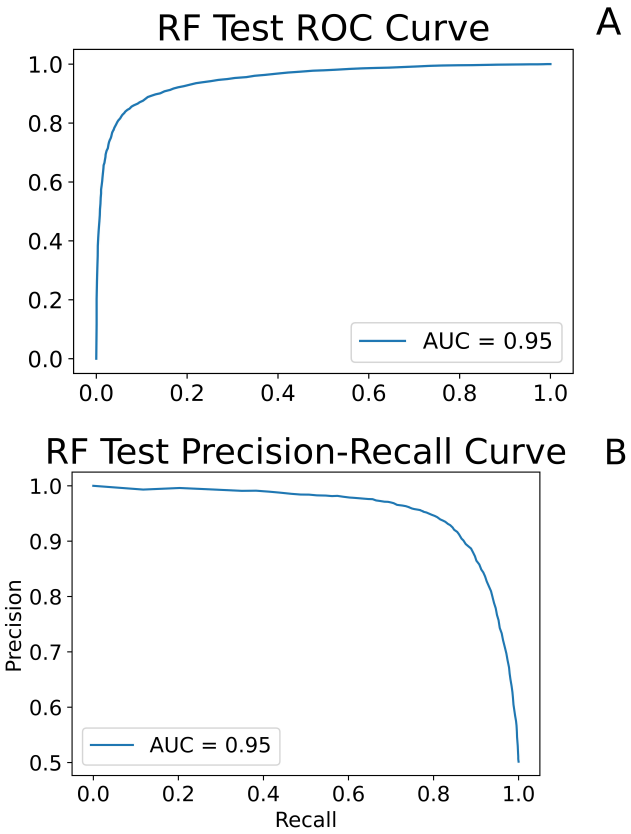

Fig. 4. Performance of a Random Forest (RF) classification model. (A) ROC curve of RF on the testing data. The area under the ROC curve is 0.95, which shows that our model generalized well to the training data and did not overfit. (B) Precision–Recall curve of RF on the testing data. The area under the curve is 0.95, which shows that our model generalized well to the training data and performs well at different prediction probability truth thresholds.

The CNN model performed with an AUROC of 0.954 (Fig. 3A), Precision-Recall AUC of 0.951 (Fig. 3B), and an accuracy of 0.895. The Random Forest classifier performed with an AUROC of 0.950 (Fig. 4A), Precision-Recall AUC of 0.950 (Fig. 4B), and an accuracy of 0.888. The test metrics of the CNN, along with those of the other models used (Random Forest and DNN), is shown in Table 2. The CNN confusion matrix for the predictions at a truth threshold of 0.97 can be seen in Fig. 5A. The confusion matrix for the RF classifier can be seen in Fig. 5B.

Table 2. Test metrics of the three tested models. Precision, recall, and F-measure calculated at a threshold of 0.97.

Model	AUROC	Accuracy	Precision	Recall	F-measure
Random Forest	0.950	0.888	0.921	0.848	0.883
DNN	0.920	0.846	0.971	0.406	0.573
CNN	0.954	0.895	0.965	0.704	0.814

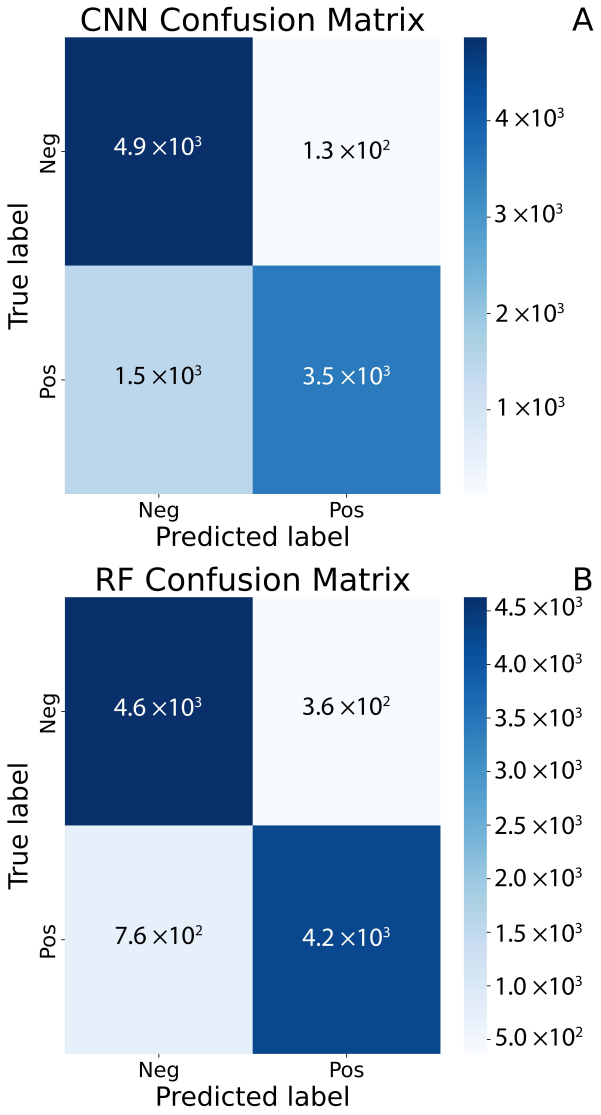

Fig. 5. Confusion matrices for machine-learning models test predictions with a truth threshold of 0.97. *True negatives* are represented by the top left square and *true positives* are represented by the bottom right square. *False positives* are seen in the top right square and *false negatives* are seen in the bottom left square. (A) Confusion matrix for CNN model test predictions. (B) Confusion matrix for the RF classifier test predictions.

2.3 Predictions

The CNN model trained on DTI data from the DrugBank website was used to predict potential interactions between drugs in the dataset and the 16 SARS-CoV-2 NSPs whose sequences were obtained from the NCBI protein databank. Each NSP was paired with all the drugs in the dataset and the same procedure presented above was followed to extract and reduce features from the proteins and drug sequences and create DTI vectors (Eqn. 2). The same features were chosen from the vectors as for the Lasso models. All the possible DTIs were inputted into the model, which calculated the probability that the input data correspond to a true DTI. DTIs with an output score greater than or equal to the thresholds of 0.97 and 0.99 were selected as potential DTIs between the repurposed DrugBank FDA-approved drugs and the viral proteins.

3. Results

We trained a convolutional deep-learning model and a random forest classifier to predict drugs that may inhibit SARS-CoV-2 viral proteins. See Table 2 for the performance metrics of these models.

The convolutional model reduced the inputted approximately 39,000 possible DTIs down to 82 of the most viable ones. We predicted 82 different drug-target interactions between FDA-approved (DrugBank) drugs and the SARS-CoV-2 NSPs with a probability of 97% or greater of interacting. Table 3 (Ref. [5,6,8,16,22,37–45]) shows the 29 unique drugs involved in these interactions. A subset of these results that met the threshold of 0.99 shown in Table 4 (Ref. [5,6,8,22,34,38,41,44]) was separated, which yielded 44 DTIs involving 13 unique drugs with a 99% probability of interaction with their respective proteins.

Similarly, the Random Forest classifier reduced the inputted DTIs down to 17 DTIs with probability greater than or equal to 90% of interacting, involving 6 unique drugs (**Supplementary Table 1**).

These results are summarized visually in Fig. 6 (Ref. [46]) and Fig. 7 (Ref. [46]). Figs. 8,9 also display the number of interactions for each drug and each target. NSP12, the RNA dependent RNA polymerase (RdRp), and NSP13, helicase, were the most targeted proteins in the 0.97 threshold result set with NSP12 being the most highly targeted protein overall. NSP12 was also the most targeted protein in the more restricted 0.99 threshold group, followed by NSP6 and NSP13, with NSP6 having two more inhibitors than NSP13 and four fewer inhibitors than NSP12. We also note that both fostamatinib, a tyrosine kinase inhibitor, and miconazole, an antifungal, were the drugs predicted to inhibit the most viral proteins overall, followed by the flavin adenine dinucleotide (in both result groups). Fostamatinib was also predicted as a potential inhibitor by the Random Forest classifier. The results for the inhibitors were consistent, as the drugs predicted in the 0.99 threshold group (excluding gabapentin enacarbil and vitamin A) were in the top half of those in the 0.97 group based on the number NSPs they were found to inhibit. Also, all but one of the drugs predicted by the Random Forest were in the 0.97 threshold group predicted by the CNN.

The majority of inhibitors were found to target less than or equal to four NSPs. Moreover, at least one drug was predicted to interact with all NSPs in both sets of results. However, not all the drugs in our dataset were found to interact with a SARS-CoV-2 viral protein.

There are many common compounds in our results that suggest the potential to quickly test and administer the drugs. These include the vitamins: vitamin A, pyridoxal

Table 3. Drugs in DTIs scoring above 0.97 from the CNN. Other studies finding the drugs as potential inhibitors of SARS-CoV-2 NSPs are indicated in the Theoretical Studies column and Clinical Trials for drugs are indicated in the Clinical Studies column. Drugs with NA in these columns are unique to this study.

DB ID[1]	Name	NSPs	Theoretical studies	Clinical studies (CT ID[2])
DB12010	Fostamatinib	1–16	[22]	NCT04352465
DB01110	Miconazole	2, 3, 5, 6, 7, 9–16	[8]	NA
DB03147	Flavin adenine dinucleotide	2, 3, 6, 12, 13, 14	[6,37]	NA
DB00114	Pyridoxal phosphate	3, 6, 12, 13	NA	NA
DB01987	Cocarboxylase	3, 12, 13	[38]	NA
DB06287	Temsirolimus	3, 12	NA	NA
DB09237	Levamlodipine	3	NA	NA
DB00132	Alpha-linolenic acid	6	NA	NCT04647604
DB00157	NADH	6, 9, 12, 13	NA	NA
DB00162	Vitamin A	6	NA	NA
DB00755	Tretinoin	6	[39]	NA
DB02659	Cholic acid	6, 9, 12, 13	NA	NA
DB03247	Flavin mononucleotide	6, 9, 11, 12, 13	[5]	NA
DB03796	Palmitic acid	6	[40]	NA
DB09061	Cannabidiol	6, 12, 13	[41]	NCT04647604
DB00143	Glutathione	9	NA	NCT04703036
DB03619	Deoxycholic acid	9	NA	NA
DB05154	Pretomanid	9	NA	NA
DB00144	Phosphatidyl serine	12	NA	NA
DB00563	Methotrexate	12	[42]	NCT04352465
DB01017	Minocycline	12	[16]	NA
DB01051	Novobiocin	12	[5]	NA
DB01329	Cefoperazone	12, 13	[43]	NA
DB08872	Gabapentin enacarbil	12	NA	NA
DB11901	Apalutamide	12, 13	NA	NA
DB14879	Cefiderocol	12, 13	[44]	NA
DB01117	Atovaquone	13	[45]	NCT04456153
DB01212	Ceftriaxone	13	NA	NA
DB08943	Isoconazole	13	NA	NA

[1] DrugBank ID.

[2] ClinicalTrials.gov ID.

Table 4. Drugs in DTIs scoring above 0.99 from the CNN. Other studies finding the drugs as potential inhibitors of SARS-CoV-2 NSPs are indicated in the Theoretical Studies column and Clinical Trials for drugs are indicated in the Clinical Studies column. Drugs with NA in these columns are unique to this study.

DB ID[1]	Name	NSPs	Theoretical studies	Clinical studies (CT ID[2])
DB12010	Fostamatinib	1–16	[22]	NCT04352465
DB03147	Flavin adenine dinucleotide	3, 12, 13, 14	[6,34]	NA
DB06287	Temsirolimus	3	NA	NA
DB01110	Miconazole	5, 6, 7, 9, 11–16	[8]	NA
DB00114	Pyridoxal phosphate	6, 12	NA	NA
DB00162	Vitamin A	6	NA	NA
DB03247	Flavin mononucleotide	6, 12, 13	[5]	NA
DB09061	Cannabidiol	6, 12	[41]	NA
DB02659	Cholic acid	9	NA	NA
DB00157	NADH	12	NA	NA
DB01987	Cocarboxylase	12	[38]	NA
DB08872	Gabapentin enacarbil	12	NA	NA
DB14879	Cefiderocol	12	[44]	NA

[1] DrugBank ID.

[2] ClinicalTrials.gov ID.

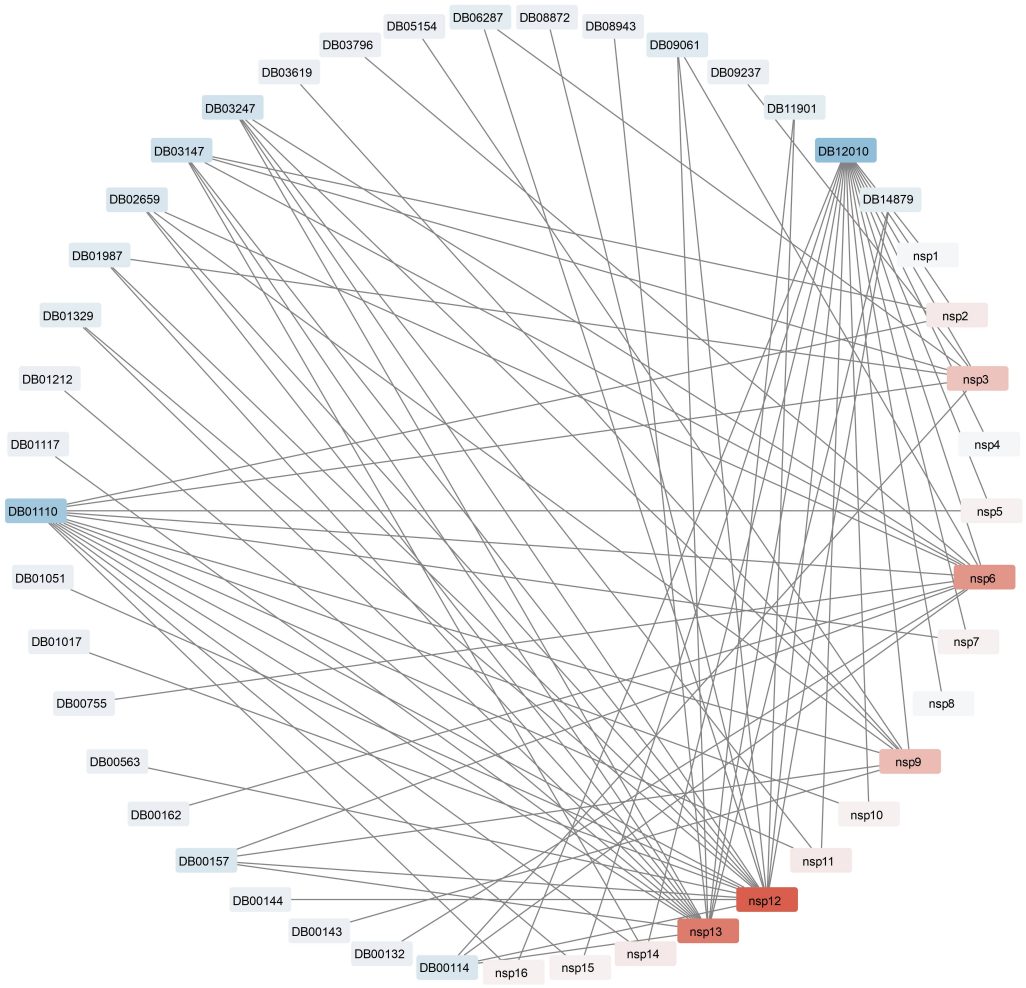

Fig. 6. Network of DTIs scoring above 0.97 from the CNN. Each edge between a drug node (blue) and a NSP node (red) represents a DTI. The intensity of the color of each node is directly proportional to its degree. Drawn with Cytoscape [46].

phosphate (a vitamin B6 derivative), cocarboxylase (vitamin B1), and tretinoin (a vitamin A derivative). The bile acids, cholic and deoxycholic acids, were also included in our results. Many antibacterial drugs were also predicted to be effective against SARS-CoV-2 such as isoconazole, atovaquone, cefoperazone, novobiocin, and ceftriaxone. **Supplementary Table 2** in Supplementary Materials shows all the compounds we elucidated and their current pharmaceutical applications.

4. Discussion

Based on the amino-acid sequences of viral proteins and chemical descriptors for various drugs, we trained a convolutional deep neural network and a Random Forest Classifier to predict new drug-target interactions. The results obtained give a starting point for selecting currently approved drugs that can be repurposed to inhibit the SARS-CoV-2 virus. The use of machine learning to make these predictions accelerates the search for a treatment and allows for high volume DTI classification that would not be possible with other techniques. Furthermore, the methods used are not specific to the SARS-CoV-2 virus and can be applied to predict DTIs in general, facilitating rapid drug discovery for other diseases as well.

As shown in Table 2, the CNN outperformed the Random Forest and DNN models in both AUROC and accuracy. This is most likely because the CNN can extract obscure relationships between the various features in the data due to its 1D convolutional layers in a way that the other models cannot. This property allows it to better generalize to the training data without overfitting. However, the Random Forest model had the highest F-measure and recall, indicating a high true-positive rate, which is valuable in predicting DTIs. Thus, we present the results from both models as both provide unique information about potential inhibitors of the SARS-CoV-2 virus. The CNN results, however, are more thoroughly analyzed as they were predicted with higher confidence (97% for the CNN as opposed to 90% for the RF; there are no RF predictions with a probability of interacting greater than 97%) and contained almost

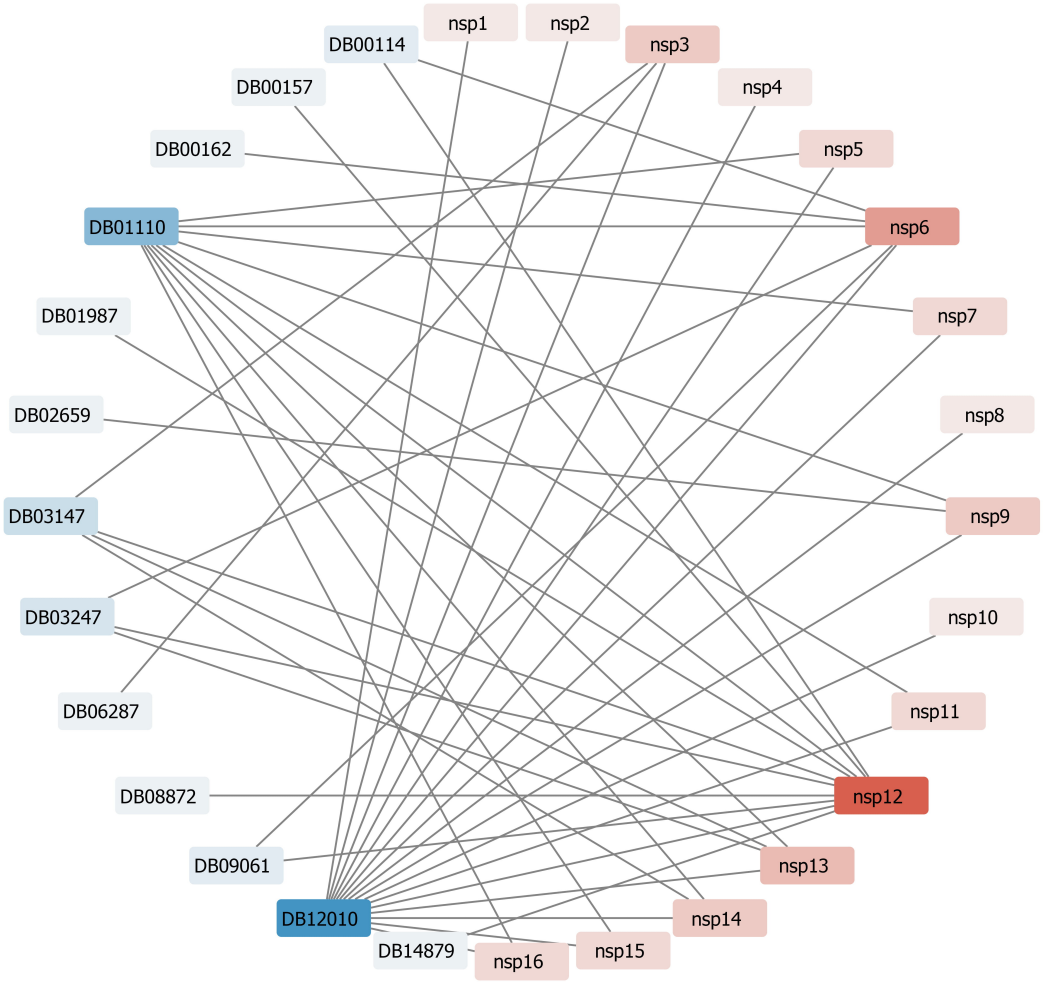

Fig. 7. Network of DTIs scoring above 0.99 from the CNN. Each edge between a drug node (blue) and a NSP node (red) represents a DTI. The intensity of the color of each node is directly proportional to its degree. Drawn with Cytoscape [46].

all the RF model predictions.

Our CNN model achieved similar accuracy, AUC, and F-measure score to other recent machine-learning based DTI prediction studies such as [14]. The performance of the models used in this study along with those of other studies are presented in Table 5 (Ref. [14,15,30,34]). Both of our models outscored all but one of the other models in AUROC and had much a higher precision score than [14]. The lower AUROC compared to [14] may be due to the restricted pool of drugs the study used as they only considered herbal drugs. The Random Forest also outperformed [14] in precision, recall, and F-measure (precision and recall scores were not available for the other studies). Note that we used a similar method to [30] in employing a Lasso model for feature selection as well as using the same protein features in the dataset, however we used a CNN as opposed to a DNN giving us more favorable metrics. This difference in performance can most likely be explained similarly to the difference in performance between the CNN and RF models. Overall, the relatively high performance of our models as compared to other studies can most likely be attributed to the unique DTI features used in this study, particularly the unique use of protein domain features as they are not widely used in DTI prediction studies. Additionally, the Lasso method for feature selection allows for highly effective dimensionality reduction. Thus, the models can learn relations among the data that would have otherwise been obscured or lost using other, less robust feature selection methods.

Eighty-two DTIs involving SARS-CoV-2 viral proteins were predicted using this model, forty-four of which had a 99% probability of interaction. There were 26 unique drugs within these DTIs, including fostamatinib, a tyrosine kinase inhibitor; miconazole, an antifungal; and ceftriaxone, an antibacterial.

We trained the model on data from known DTIs involving various proteins—including those of influenza and Ebola viruses—and FDA-approved drugs. This model generalized exceptionally well to this data as it learned the important patterns in the data to distinguish true and false DTIs, making it a strong choice to use in predicting new relationships. A Lasso model was applied to the data be-

Table 5. Comparison of our model's performance to those of other studies. Studies that did not include a metric have an NA marked in their respective field.

Study	Accuracy	AUROC	Precision	Recall	F-measure
This Study (CNN)	0.895	0.954	0.965	0.704	0.814
This Study (RF)	0.888	0.950	0.921	0.848	0.883
Semi supervised model [14]	0.940	0.970	0.817	0.830	0.822
CNN [34]	0.923	NA	NA	NA	0.895
Lasso-DNN [30]	0.81	0.89	NA	NA	NA
Naïve Bayes [15]	0.730	0.666	NA	NA	0.768

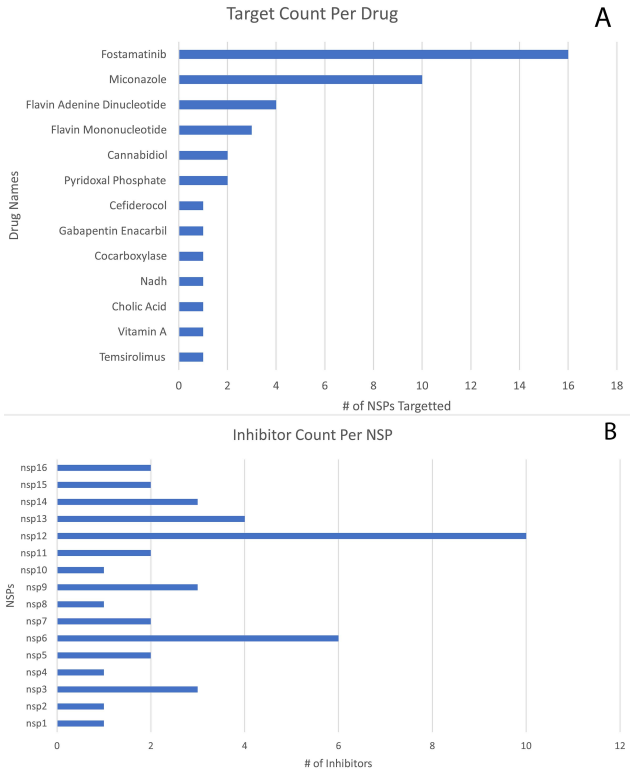

Fig. 8. Charts displaying the number of inhibitors and NSP targets in the 0.99 threshold group obtained from CNN. (A) Number of NSPs that each drug was found to inhibit. (B) Number of inhibitors that each NSP was found to have.

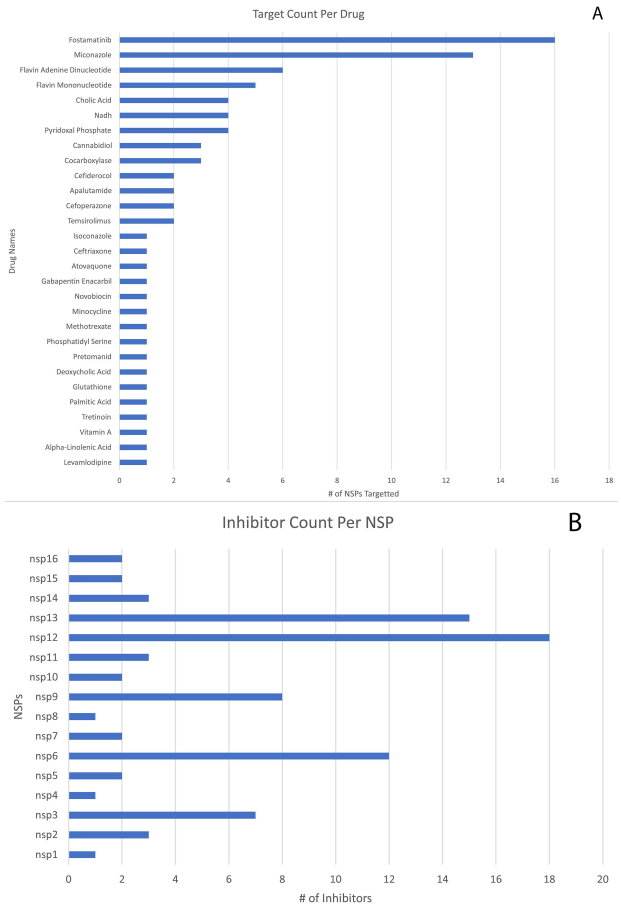

Fig. 9. Charts displaying the number of inhibitors and NSP targets in the 0.97 threshold group obtained from CNN. (A) Number of NSPs that each drug was found to inhibit. (B) Number of inhibitors that each NSP was found to have.

fore it was fed to the CNN to filter out the most informative features and improve the efficiency of our final model. The scores assigned to the COVID-19 DTIs by the model were compared and all pairs scoring above 0.97 were extracted as possible candidates.

Given that RdRp (NSP12), helicase (NSP13), and the main protease (NSP5) are viable and highly researched viral proteins for inhibition, it is of high interest that NSP12 and NSP13 are among the top 3 highly targeted proteins in both threshold sets resulting from our model's predictions. Given the significant role these proteins play in the life cycle of the virus, the drugs targeting them should be given the highest priority in testing.

It is interesting to consider why NSP12 is the most targeted protein. This may be due to the fact that the enzyme is conserved in structure among all RNA viruses [1]. Given that our model exploits similarities in structure between various proteins and drugs to make predictions, it is very likely that it took advantage of the recurring structure of the RdRp enzyme across various viruses to predict inhibitors for SARS-CoV-2. This pattern in the data most likely explains the large number of inhibitors predicted for NSP12 as its familiar structure links lots of other proteins, and thus inhibitors, to it.

It is interesting to note that among our results were common compounds such as vitamin A and the

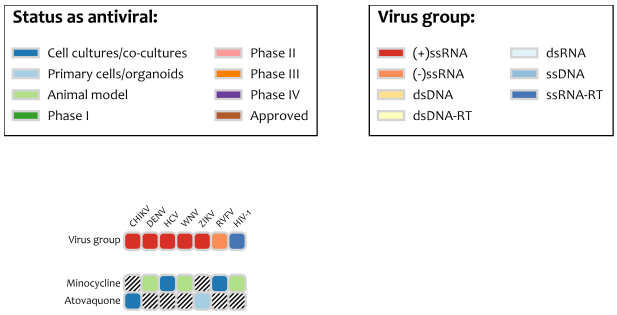

Fig. 10. Drugs among those involved in the predicted DTIs that have been tested for other drugs. Obtained from the DrugVirus.info database [48].

cholic and deoxycholic bile acids. In addition, clinical trials are currently in progress to test the efficacy of fostamatinib (CT ID: NCT04352465), cannabidiol (CT ID: NCT04647604), alpha-linolenic acid (omega-3 polyunsaturated fatty acid; CT ID: NCT04647604), glutathione (CT ID: NCT04703036), methotrexate (CT ID: NCT04352465), and atovaquone (CT ID: NCT04456153) in treating COVID-19. Furthermore, fostamatinib has been predicted as a potential inhibitor of NSP5 ($3CL^{pro}$) by [22] and, although less promising, has also been predicted to target NSP16 [47]. The flavin adenine dinucleotide was predicted to bind to NSP12 (RdRp) with a docking score of –11.8 kcal/mol and to NSP13 (helicase) with a score of –11.2 kcal/mol [37]. Wu and co-authors [5] proposed the antifungal novobiocin as a potential NSP12 inhibitor as well. Cefoperazone has also been found active in interaction with the apo-NSP13 ATP-binding sites (a Vina docking score of –10.2 kcal/mol [43]). All other predicted DTIs are unique to our study although many of the drugs *are not*. Namely, we predicted 15 inhibitors that have not yet been considered. Studies that examined our predicted drugs are indicated in Tables 3,4. A similar table for the results of the Random Forest can be seen in **Supplementary Table 1** in Supplementary Materials. We note that methotrexate has shown efficacy in inhibiting viral RNA replication, viral protein synthesis, and virus release in an *in-vitro* setting [42]. Fig. 10 (Ref. [48]) shows that atovaquone and minocycline, both antibacterial pharmaceuticals, are the only drugs from our results that have been tested for other viruses, suggesting the novelty of our predictions.

We note that although the results of this study partially overlap with those of other theoretical studies, the results should be further validated using other methods before administering the drugs in clinical trials. Further analysis of these results using docking simulations (which have already been used in some of the studies cited above) and pharmacophore models may be useful in determining which of the predicted DTIs are most likely to give positive results in a clinical environment.

5. Conclusions

We developed a machine-learning model to predict possible inhibitors of the 16 SARS-CoV-2 nonstructural proteins. A convolutional neural network with three convolutional layers and a Random Forest model were used. The CNN model, trained on 2444 drugs and 16,640 known drug-target interactions (DTIs) from DrugBank, was developed using the TensorFlow Python library. The best algorithm for the classification task was the CNN. A part of the training dataset (30%) was randomly separated from the rest of the data to create a validation dataset that was used to tune our models' hyperparameters and optimize metrics. The model predicted 29 COVID-19 drugs involved in 82 DTI with 97% probability.

Author contributions

IFT and SKA contributed to conception and study design, original manuscript preparation. SKA and VLK constructed the model and received the prediction data. SKA and VLK contributed to original manuscript preparation and final draft reviewing and editing.

Ethics approval and consent to participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This research received no external funding.

Conflict of interest

The authors declare no conflict of interest.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2704113.

References

[1] Gil C, Ginex T, Maestro I, Nozal V, Barrado-Gil L, Cuesta-Geijo MÁ, *et al*. COVID-19: Drug targets and potential treatments. Journal of Medicinal Chemistry. 2020; 63: 12359–12386.

[2] Elfiky AA. Anti-HCV, nucleotide inhibitors, repurposing against COVID-19. Life Sciences. 2020; 248: 117477.

[3] Kouznetsova VL, Zhang A, Tatineni M, Miller MA, Tsigelny IF. Potential COVID-19 papain-like protease PLpro inhibitors: repurposing FDA-approved drugs. PeerJ. 2020; 8: e9965.

[4] Muralidharan N, Sakthivel R, Velmurugan D, Gromiha MM. Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. Journal of Biomolecular Structure and Dynamics. 2020; 39: 2673–2678.

[5] Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y, *et al*. Analysis of therapeutic targets for SARS-CoV-2 and discovery of

potential drugs by computational methods. Acta Pharmaceutica Sinica B. 2020; 10: 766–788.
[6] Hall DC, Ji H. A search for medications to treat COVID-19 via *in silico* molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease. Travel Medicine and Infectious Disease. 2020; 35: 101646.
[7] Feng S, Luan X, Wang Y, Wang H, Zhang Z, Wang Y, *et al*. Eltrombopag is a potential target for drug intervention in SARS-CoV-2 spike protein. Infection, Genetics and Evolution. 2020; 85: 104419.
[8] Ke Y, Peng T, Yeh T, Huang W, Chang S, Wu S, *et al*. Artificial intelligence approach fighting COVID-19 with repurposing drugs. Biomedical Journal. 2020; 43: 355–362.
[9] Wang C, Wang W, Lu K, Zhang J, Chen P, Wang B. Predicting drug-target interactions with electrotopological state fingerprints and amphiphilic pseudo amino acid composition. International Journal of Molecular Sciences. 2019; 21: 5694.
[10] Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Computational Biology. 2019; 15: e1007129.
[11] Shi H, Liu S, Chen J, Li X, Ma Q, Yu B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics. 2019; 111: 1839–1852.
[12] Zhao T, Hu Y, Valsdottir LR, Zang T, Peng J. Identifying drug–target interactions based on graph convolutional network and deep neural network. Briefings in Bioinformatics. 2020; 22: 2141–2150.
[13] Majumdar S, Nandi SK, Ghosal S, Ghosh B, Mallik W, Roy ND, *et al*. Deep learning-based potential ligand prediction framework for COVID-19 with drug–target interaction model. Cognitive Computation. 2021; 2021: 1–13.
[14] Sulistiawan F, Kusuma WA, Ramadhanti NS, Tedjo A. Drug-target interaction prediction in coronavirus disease 2019 case using deep semi-supervised learning model. 2020 International Conference on Advanced Computer Science and Information Systems. 2020; 83–88.
[15] Mohapatra S, Nath P, Chatterjee M, Das N, Kalita D, Roy P, *et al*. Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking. PLoS ONE. 2020; 15: e0241543.
[16] Tripathi MK, Sharma S, Singh TP, Ethayathulla AS, Kaur P. Computational intelligence in drug repurposing for COVID-19. Studies in Computational Intelligence. 2021; 9: 273–294.
[17] Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Briefings in Bioinformatics. 2013; 15: 734–747.
[18] Wu Z, Li W, Liu G, Tang Y. Network-based methods for prediction of drug-target interactions. Frontiers in Pharmacology. 2018; 9: 1134.
[19] Yoo JH. Uncertainty about the efficacy of remdesivir on COVID-19. Journal of Korean Medical Science. 2020; 35: e221.
[20] Elmezayen AD, Al-Obaidi A, Şahin AT, Yelekçi K. Drug repurposing for coronavirus (COVID-19): *in silico* screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. Journal of Biomolecular Structure and Dynamics. 2020; 39: 2980–2992.
[21] Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic treatments for coronavirus disease 2019 (COVID-19): A review. Journal of the American Medical Association. 2020; 323: 1824–1836.
[22] Liu S, Zheng Q, Wang Z. Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus. Bioinformatics. 2020; 36: 3295–3298.
[23] Sternberg A, McKee DL, Naujokat C. Novel drugs targeting the SARS-CoV-2/COVID-19 machinery. Current Topics in Medicinal Chemistry. 2020; 20: 1423–1433.
[24] Mahdian S, Ebrahim-Habibi A, Zarrabi M. Drug repurposing using computational methods to identify therapeutic options for COVID-19. Journal of Diabetes & Metabolic Disorders. 2020; 19: 691–699.
[25] Liang H, Zhao L, Gong X, Hu M, Wang H. Virtual screening FDA approved drugs against multiple targets of SARS-CoV-2. Clinical and Translational Science. 2021; 14: 1123–1132.
[26] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, *et al*. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Research. 2018; 46: D1074–D1082.
[27] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. Journal of Cheminformatics. 2011; 3: 33.
[28] Sushko I, Pandey A, Novotarskyi S, Körner R, Rupp M, Teetz W, *et al*. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. Journal of Cheminformatics. 2011; 3: P20.
[29] Wang L, You Z, Chen X, Xia S, Liu F, Yan X, *et al*. A computational-based method for predicting drug-target interactions by using stacked autoencoder deep neural network. Journal of Computational Biology. 2018; 25: 361–373.
[30] You J, McLeod RD, Hu P. Predicting drug-target interaction network using deep learning model. Computational Biology and Chemistry. 2019; 80: 90–101.
[31] Varoquaux G, Buitinck L, Louppe G, Grisel O, Pedregosa F, Mueller A. Scikit-learn: Machine learning without learning the machinery. GetMobile: Mobile Computing and Communications. 2015; 19: 29–33.
[32] McKinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference. 2010; 1: 56–61.
[33] Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, *et al*. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLoS ONE. 2012; 7: e37608.
[34] Monteiro NRC, Ribeiro B, Arrais JP. Deep Neural Network Architecture for Drug-Target Interaction Prediction. Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. 2019; 11: 804–809.
[35] Bengio Y. Practical recommendations for gradient-based training of deep architectures. Lecture Notes in Computer Science. 2012; 10: 437–478.
[36] Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, *et al*. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. 2016. Available at: http://download.tensorflow.org/paper/whitepaper2015.pdf (Accessed: 25 January 2022).
[37] Anwaar MU, Adnan F, Abro A, Khan RA, Rehman AU, Osama M, *et al*. Combined deep learning and molecular docking simulations approach identifies potentially effective FDA approved drugs for repurposing against SARS-CoV-2. Computers in Biology and Medicine. 2021.
[38] Shankar U, Jain N, Majee P, Mishra SK, Rathi B, Kumar A. Potential drugs targeting Nsp16 protein may corroborates a promising approach to combat SARS-CoV-2 virus. ChemRxiv. 2020. (in press)
[39] Le BL, Andreoletti G, Oskotsky T, Vallejo-Gracia A, Rosales R, Yu K, *et al*. Transcriptomics-based drug repositioning pipeline identifies therapeutic candidates for COVID-19. bioRxiv. 2020. (in press)
[40] Elfiky AA. Natural products may interfere with SARS-CoV-2 attachment to the host cell. Journal of Biomolecular Structure & Dynamics. 2020; 39: 3194–3203.
[41] Raj V, Park JG, Cho K, Choi P, Kim T, Ham J, *et al*. Assessment

of antiviral potencies of cannabinoids against SARS-CoV-2 using computational and *in vitro* approaches. International Journal of Biological Macromolecules. 2021; 168: 474–485.

[42] Caruso A, Caccuri F, Bugatti A, Zani A, Vanoni M, Bonfanti P, *et al*. Methotrexate inhibits SARS-CoV-2 virus replication "*in vitro*". Journal of Medical Virology. 2020; 93: 1780–1785.

[43] White MA, Lin W, Cheng X. Discovery of COVID-19 inhibitors targeting the SARS-CoV-2 Nsp13 helicase. The Journal of Physical Chemistry Letters. 2020; 11: 9144–9151.

[44] Yadav R, Parihar RD, Dhiman U, Dhamija P, Upadhyay SK, Imran M, *et al*. Docking of FDA approved drugs targeting NSP-16, N-protein and main protease of SARS-CoV-2 as dual inhibitors. Biointerface Research in Applied Chemistry. 2021; 11: 9848–9861.

[45] Marak BN, Dowarah J, Khiangte L, Singh VP. Step toward repurposing drug discovery for COVID-19 therapeutics through *in silico* approach. Drug Development Research. 2020; 82: 374–392.

[46] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, *et al*. Cytoscape: A software environment for integrated models. Genome Research. 2003; 13: 2498–2504.

[47] Jiang Y, Liu L, Manning M, Bonahoom M, Lotvola A, Yang ZQ. Repurposing therapeutics to identify novel inhibitors targeting 2'-O-ribose methyltransferase Nsp16 of SARS-CoV-2. ChemRxiv. 2020. (in press)

[48] Andersen PI, Ianevski A, Lysvand H, Vitkauskiene A, Oksenych V, Bjørås M, *et al*. Discovery and development of safe-in-man broad-spectrum antiviral agents. International Journal of Infectious Diseases. 2020; 93: 268–276.