

# Machine-Learning Driven Drug Repurposing for COVID-19

Semih Cantürk, Aman Singh, Jason Behrmann, Patrick St-Amant\*

Department of Mathematics and Physics, Zetane Systems, Quebec, Canada

## ABSTRACT

The integration of machine learning methods into bioinformatics provides particular benefits in identifying how therapeutics effective in one context might have utility in an unknown clinical context or against a novel pathology. We aim to discover the underlying associations between viral proteins and antiviral therapeutics that are effective against them by employing neural network models. Using the National Center for Biotechnology Information virus protein database and the Drug Virus database, which provides a comprehensive report of Broad-Spectrum Antiviral Agents (BSAAs) and viruses they inhibit, we trained ANN models with virus protein sequences as inputs and antiviral agents deemed safe-in-humans as outputs. Model training excluded SARS-CoV-2 proteins and included only Phases II, III, IV and approved level drugs. Using sequences for SARS-CoV-2 (the coronavirus that causes COVID-19) as inputs to the trained models produces outputs of tentative safe-in-human antiviral candidates for treating COVID-19. Our results suggest multiple drug candidates, some of which complement recent findings from noteworthy clinical studies. Our in-silico approach to drug repurposing has promise in identifying new drug candidates and treatments for other viral, bacterial and parasitic infections.

**Keywords:** Drug virus database; SARS-CoV-2; COVID-19; Nitazoxanide

## INTRODUCTION

Artificial Intelligence (AI) technology is a recent addition to bioinformatics that shows much promise in streamlining the discovery of pharmacologically active compounds [1]. Machine Learning (ML) provides particular benefits in identifying how drugs effective in one context might have utility in an unknown clinical context or against a novel pathology [2]. The application of ML in biomedical research provides new means to conduct exploratory studies and high-throughput analyses using information already available. In addition to deriving more value from past research, researchers can develop ML tools in relatively short periods of time.

Past research now provides a sizable bank of information concerning drug-biomolecule interactions. Using drug repurposing as an example, we can now train predictive algorithms to identify patterns in how antiviral compounds bind to proteins from diverse virus species. We aim to train an ML model so that when presented with the proteome of a novel virus, it will suggest antivirals based on the protein segments present in the proteome. The final output from the model is a best-fit prediction as to which known antivirals are likely to associate with those familiar protein segments.

These benefits are of particular interest for the current COVID-19 health crisis. The novelty of SARS-CoV-2 requires that we execute health interventions based on past observations. Grappling with an

unforeseen pandemic with no known treatments or vaccines, the potential for rapid innovation from ML is of utmost significance. The ability to conduct complex analyses with ML enables us to research insights quickly that can help steer us in the right direction for future studies likely to produce fruitful results. We present here multiple models that produced a number of antiviral candidates for treating COVID-19. Out of our 12 top predicted drugs, 6 of them have shown positive results in recent findings based on cell culture results and clinical trials. These promising antivirals are lopinavir, ritonavir, ribavirin, cyclosporine, rapamycin, and nitazoxanide [3-6]. For the 7 other predicted drugs, further research is needed to evaluate their effectiveness against SARS-CoV-2.

## METHODOLOGY

### Data

**Sourcing and preparation:** We used two main data sources for this study. The first database was the Drug Virus database; Drug Virus provides a database of Broad-Spectrum Antiviral Agents (BSAAs) and the associated viruses they inhibit [7]. The database covers 83 viruses and 126 compounds, and provides information on the status as antiviral of each compound-virus pair. These statuses fall into eight categories representing the progressive drug trial phases: Cell cultures/co-cultures, Primary cells/organoids, Animal model, Phases I-IV and Approved. See Appendix A for a more intuitive pivot table view of the database.

**Correspondence to:** Dr. Patrick St-Amant, Department of Mathematics and Physics, Zetane Systems, Quebec, Canada, E-mail: patrick@zetane.com

**Received date:** September 14, 2021; **Accepted date:** September 28, 2021; **Published date:** October 05, 2021

**Citation:** Cantürk S, Singh A, Behrmann J, St-Amant P (2021) Machine-Learning Driven Drug Repurposing for COVID-19. J Bacteriol Parasitol. S12: 004.

**Copyright:** © 2021 Cantürk S, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

The second database is the National Center for Biotechnology Information (NCBI) Virus Portal; as of April 2020, this database provides approximately 2.8 million amino-acid and 2 million nucleotide sequences from viruses with humans as hosts [8]. Each row of this database contains an amino acid sequence specimen from a study, as well as metadata that includes the associated virus species. In our work, we considered sequences only from the 83 virus species in the Drug Virus database or their subspecies in order to be able to merge the two data sources successfully. We also constrained ourselves to amino-acid sequences only in the current iteration. The main reasons for this are two-fold:

1. Amino-acid sequences are essentially derived from the DNA sequences, which may encode overlapping information on different levels. In somewhat simplified terms, amino-acid sequences are the outputs of a layer of preprocessing on genetic material (in the form of DNA/RNA).

2. Nucleotide triplets (codons) map to amino-acids, making amino-acid sequences much shorter and easier to extract features both in preprocessing and in the machine learning methods themselves. Shorter sequences also mean the ML pipeline will be more resource-efficient, i.e. easier to train.

The amino-acids were downloaded as three datasets: HIV types 1 and 2 (1,192,754 sequences), Influenza types A, B and C (644,483 sequences), and the “Main” dataset for all other types including SARS-CoV-2 (785,624 sequences). Each dataset came with two components. The “sequence” component is composed on Accession IDs and the amino-acid sequence itself, while the “metadata” component includes all other data (e.g. virus species, date specimen was taken, an identifier of the related study) as well as the Accession ID to enable merging the two components.

The amount of research with a focus on Influenza and HIV naturally lead to these viruses comprising most of the samples. In our experiments, we have excluded these viruses, and have worked only with dataset #3, though the other datasets can be integrated into the main one during the class balancing process, an idea we will discuss in Section 4, Future Work.

**Preprocessing:** The first step of the preparation phase was to merge the “sequence” and “metadata” components into

a single NCBI dataset based on sequence IDs. Afterwards, we mapped the “Species” column in this main dataset to the Virus Name column in the Drug Virus database. This step was required as these two columns that denote the virus species in the respective datasets did not match due to subspecies present in the sequence dataset and alternative naming of some viruses.

Afterwards, we processed the Drug Virus dataset to a format suitable for merging with the NCBI data frame. Every row of the Drug Virus dataset consists of a single drug-virus pairing and their respective interaction/drug trial phase, meaning any given drug and virus appeared in multiple rows of the dataset. We derived a new Drug Virus dataset that functioned as a dictionary where each unique virus was a key, and the interactions with antivirals encoded as a multi-label binary vector (1 if viable antiviral according to the original dataset, 0 if not) of length 126 (the number of antivirals) which corresponded to the value. We came up with three “versions” depending on how we decided an antiviral was a viable candidate to inhibit a virus. The criteria depended on drug trial phases:

1. In the first version, any interaction between a drug-virus pair is designated by a 1. This means drugs that did not go past cell cultures/co-cultures or primary cells/organoids testing are still considered viable candidates.

2. This second version expands upon the first stemming from our discovery that an attained trial phase in the database does not necessarily mean previous phases were also listed in the database. For example, we found that for a given virus, a given drug had undergone Phase III testing, designated by a 1, but Phase I and II were listed as 0s. This undermined our assumption that drug trials are hierarchical; though, in reality this is usually the case. This can be caused by missing data reporting or possibly skipped phases. We proceeded with the hierarchy assumption, and extended the database in (1) to account for the previous phases. This meant that in this second version, an Approved drug will have all phases designated with 1s, for example. Keeping track of the 8 phases meant that the size of the database also grew by 8.

3. In the third version, we considered a drug-virus pair as viable only if it has attained Phase II or further drug trials, signifying some success with human trials have been observed. In the results presented in Section 3, our training database was based on this third version of the Drug Virus database.

The full dataset was then generated by merging this “new” version of the Drug Virus dataset with the NCBI dataset. We then generated two versions of this full dataset: one that consists of all SARS-CoV-2 sequences and one that consists of all other viruses available. This enabled us to compare how successful our models are in a case when they have not been trained on the virus species at all and have to detect peptide substructures in the sequences to suggest antivirals. A sample of this final database (with some columns excluded for brevity) is available in Appendix B.

Upon inspection of the data, we found that there were replete of duplicate or extremely similar virus sequences. To reduce this exploitability and pose a more challenging problem, we removed the duplicate sequences that belonged to the same species and had the exact same length. This reduced the size of the dataset by approximately 98%. The counts for each virus species before and after dropping duplicate viruses is available in Appendix C1 and C2.

**Balancing:** Our main database also contained a class imbalance in the number of times certain virus species appeared in the database. We oversampled rare viruses (e.g., West Nile virus: 175 sequences) and excluded the very rare species which compose less than 0.5% of the available unique samples in the dataset (e.g., Andes virus: 4 sequences), and under sampled the common viruses (e.g., Hepatitis C: 16,040 sequences). This produced a more modest database of 30,479 amino acid sequences, with each virus having samples in the 400–900 range (see Appendix C3). We kept the size of the dataset small both to enable easier model training and validation in early iterations and to handle data imbalance more smoothly.

The class imbalance problem also presented itself in the antiviral compounds. Even with balanced virus classes, the number of times each drug occurred within the dataset varied, simply because some drugs apply to more viruses than others. To alleviate this, we computed class weights for each drug, which we then provided to the models in training. This enabled a fairer assessment and a more varied distribution of antivirals in predicted outputs.

**Train/test splitting:** The final step of data processing involved generating the training and validation sets. We split the data in two different ways, resulting in two different experiments (see Section 2.3, Experiment Setup for the full experiment pipeline). Experiment I is based on a standard, randomized an 80% training/20% validation split on the main dataset.

For Experiment II, we split the data on virus species, meaning the models were forced to predict drugs for a species that it was not trained on, and have to detect peptide substructures in the amino-acid sequences to suggest drugs. In this setup we also guaranteed that the SARS-CoV-2 sequences were always in the test set, in addition to three other viruses randomly picked from the dataset. We used a variant of this setup that trains on all virus sequences except SARS-CoV-2 and is validated on SARS-CoV-2 only to generate the results presented in Section 3.

## Models

A growing number of studies demonstrate the success of using Artificial Neural Networks (ANN) in evaluating biological sequences in drug repositioning and repurposing [9,10]. Previous work on training neural networks on nucleotide or amino-acid sequences have been successful with recurrent models such as Gated Recurrent Units (GRU), Long Short-Term Memory Networks (LSTM) and bidirectional LSTMs (biLSTM), as well as 1D convolutions and 2D Convolutional Neural Networks (CNN) [11,12]. We have therefore focused on these network architectures, and conducted our experiments with an LSTM with 1D convolution and bidirectional layers as well as a CNN. The network architectures are explained briefly below.

**LSTM and 1D convolutions:** For the LSTM, A character-level tokenizer was used to encode the FASTA sequences into vectors consumable by the network. The sequences were then padded with zeros or cut off to a fixed length 500 to maintain a fixed input size. The network architecture consisted of an embedding layer, followed by 1D convolution and bidirectional LSTM layers (each followed by maxpooling), and two fully connected layers. A more detailed architecture diagram is available in Appendix D.

**Convolutional Neural Network (CNN):** For the CNN, the input features were one-hot encoded based on the FASTA alphabet/charset, which assisted in interpretability when examining the 2D input arrays as images. The inputs are also fixed at a length of 500, resulting in  $500 \times 28$  images, where 28 is the number of elements in the FASTA charset. The network architecture consists of four 2D convolutions with filter sizes of  $1 \times 28$ ,  $2 \times 28$ ,  $3 \times 28$  and  $5 \times 28$  respectively, which are maxpooled, concatenated and passed through a fully connected layer. A more detailed architecture diagram is available in Appendix E.

## Experiment setup

The experiments were run on a computer with a 2.7 GHz Intel Broadwell CPU (61 GB RAM) and NVIDIA K80 GPU (12 GB). Both models completed a 20-epoch experiment in 60-90 minutes. One to three training and evaluation runs were made for each setup during model and hyperparameter selections, and ten training and evaluation runs were done to produce the average metrics in Section 3.

The experiments start by determining the model to use and apply the appropriate preprocessing steps mentioned in Section 2.2. We then proceed with determining the dataset to train and validate on. This part of the experiment setup is more extensively covered in Section 2.1.4, Train/Test Splitting. We used Binary Cross Entropy (BCE) loss, Adam optimizer, precision, recall and F1-score as metrics since accuracy tends to be an unreliable metric given the class imbalance and the sparse nature of our outputs. After training and validation, predictions were done on the validation set and the results were post-processed for interpretability.

In post-processing, we applied a threshold to the sigmoid function outputs of the neural network, where we assigned each drug a probability of being a potential antiviral for a given amino acid sequence. After experimenting with different values, we settled on a threshold value of 0.2. Postprocessing outputs a list of drugs that were selected along with the respective probabilities for the drugs being “effective” against the virus with the given amino acid sequence. For other hyperparameters involved as well as information on hyperparameter tuning, see Appendix F.

## RESULTS

Here we present the results for the two experiments described in Section 2.1.4, Train/Test Splitting. The figures and tables presented in this section are based on the LSTM and CNN architectures described in Section 2.2, which were trained on 128 batch size and 0.01 and 0.001 learning rates respectively for 20 epochs with an Adam optimizer.

### Experiment I: train/test split

In the regular setup, we performed an 80%/20% train-test split on our data of 30,479 sequences. The metrics for the best set of hyperparameters (based on validation set F1-score) for both the CNN and LSTM architectures respectively are presented in Table 1. Similarly, plots for the same set of models and hyperparameters over 20 epochs are presented in Figures 1 and 2.

Our models handled the task successfully, achieving 0.958 F1-score in a multi-label multi-class problem setting. This means that the models were able to match the virus species with the sequence substructures and appropriately assign the inhibiting antivirals with accuracy. These satisfactory results led to us implementing Experiment II.

### Experiment II: predictions on unseen virus species

In Experiment II, the models predicted antiviral drugs for virus species they haven't been trained on. This meant the models were not able to recommend drugs by “recognizing” the virus from the sequence and therefore had to rely only on peptide substructures in the sequences to assign drugs. In the results presented below, the test set consists of SARS-CoV-2, Herpes simplex virus 1, Human Astrovirus and Ebola virus, whose sequences were removed from the training set.

We see here that the CNN (and the LSTM) had issues with convergence, and the accuracies are clearly below their counterparts in the regular setup, though this is certainly expected. We now turn to the actual predictions on the sequences and attempt to interpret them (Table 2).

Upon examination of drug predictions for Herpes Simplex Virus 1 (HSV-1), however, we see that our CNN was in fact quite successful. In Tables 3 and 4, Count represents how many times each drug was flagged as potentially effective for HSV-1 sequences, and Mean Probability denotes the average confidence predicted over all instances of the drug. A sample of the outputs where these metrics are derived from is available in Appendix G. Antivirals used for Phase II and further trials for HSV-1 are highlighted in bold, meaning all six drugs in the database that are used for Phase II and further trials are predicted by our model. Three of the top five predictions are approved antivirals for HSV-1 and the only remaining one is predicted 11th among 126 antivirals. This high level of accuracy is remarkable given that our model has not been trained on HSV-1 sequences.

Table 1: Experiment I, metrics with 95% confidence intervals for optimal hyperparameters tested.

	Model	Accuracy	Precision	Recall	F1-score	Loss
Training set	LSTM	$0.998 \pm 2.3e-4$	$0.951 \pm 0.0059$	$0.914 \pm 0.01$	$0.932 \pm 0.0081$	0.0126
	CNN	$0.999 \pm 1e-4$	$0.992 \pm 0.0023$	$0.989 \pm 0.0047$	$0.990 \pm 0.0035$	0.00265
Validation set	LSTM	$0.998 \pm 2.1e-4$	$0.956 \pm 0.011$	$0.892 \pm 0.012$	$0.923 \pm 0.0074$	0.0134
	CNN	$0.999 \pm 6e-5$	$0.986 \pm 0.0019$	$0.948 \pm 0.0039$	$0.967 \pm 0.0022$	0.00767

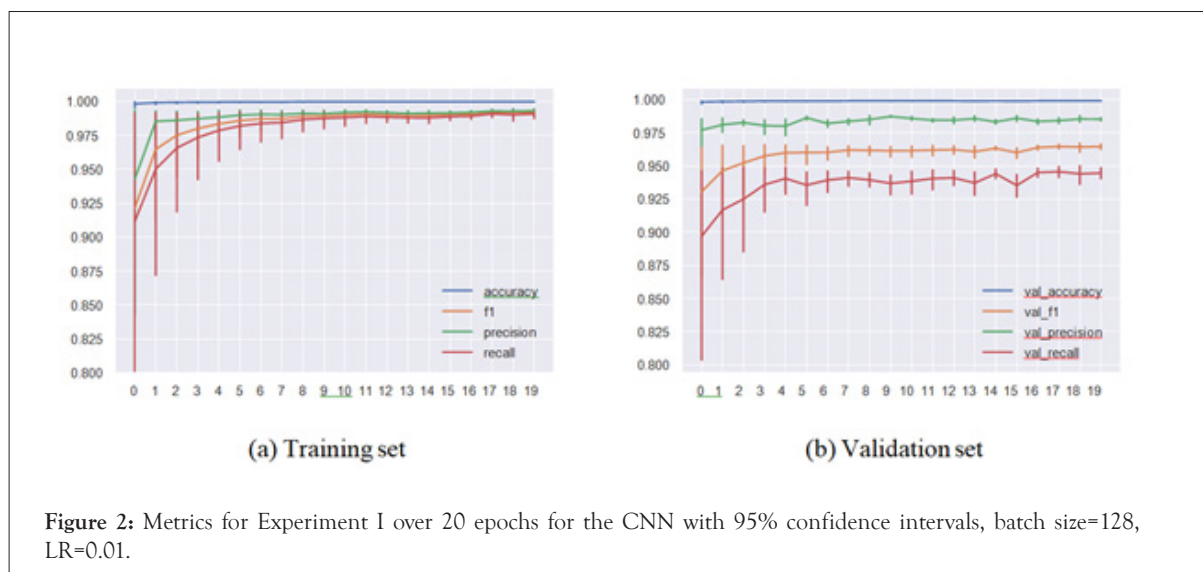
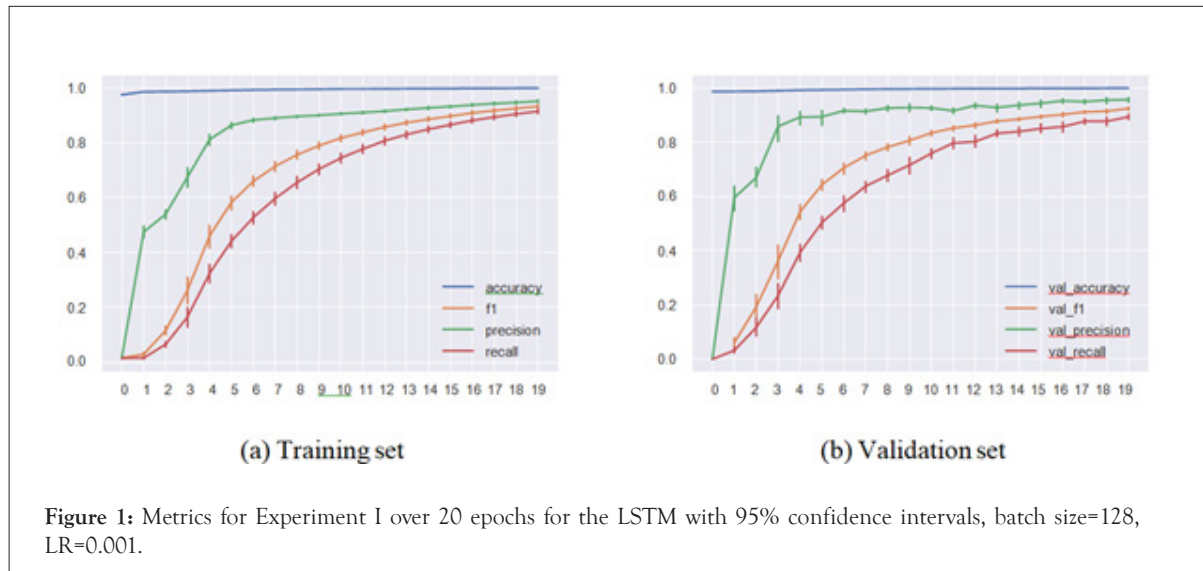


Table 2: Experiment II, metrics with 95% confidence intervals for optimal hyperparameters tested.

	Model	Accuracy	Precision	Recall	F1-score	Loss
Training set	LSTM	$0.998 \pm 5e-4$	$0.942 \pm 0.011$	$0.877 \pm 0.029$	$0.908 \pm 0.021$	0.00369
	CNN	$0.999 \pm 1.6e-4$	$0.990 \pm 0.0038$	$0.983 \pm 0.0085$	$0.986 \pm 0.0063$	0.00325
Validation set	LSTM	$0.958 \pm 0.0019$	$0.530 \pm 0.048$	$0.227 \pm 0.014$	$0.318 \pm 0.0099$	3.764
	CNN	$0.960 \pm 7.9e-4$	$0.588 \pm 0.026$	$0.239 \pm 0.0068$	$0.340 \pm 0.0033$	7.562

**Table 3:** Experiment II with CNN, Count/Mean Probability table for 588 HSV-1 sequences.

<b>Antiviral</b>	<b>Count</b>	<b>Mean probability</b>
Valacyclovir	578	0.899 ± 0.0047
Cidofovir	567	0.737 ± 0.017
Foscarnet	566	0.904 ± 0.0066
Brincidofovir	559	0.646 ± 0.048
Aciclovir	525	0.759 ± 0.0081
Rapamycin	419	0.540 ± 0.017
Ganciclovir	410	0.552 ± 0.064
Artesunate	333	0.453 ± 0.026
Cyclosporine	328	0.449 ± 0.094
Letermovir	317	0.430 ± 0.02
Tilorone	243	0.433 ± 0.071
Tenofovir	173	0.573 ± 0.017

Drugs with more than 100 counts are listed. Antivirals used for Phase II and further trials for HSV-1 are highlighted in bold. Mean probabilities are provided with 95% confidence intervals.

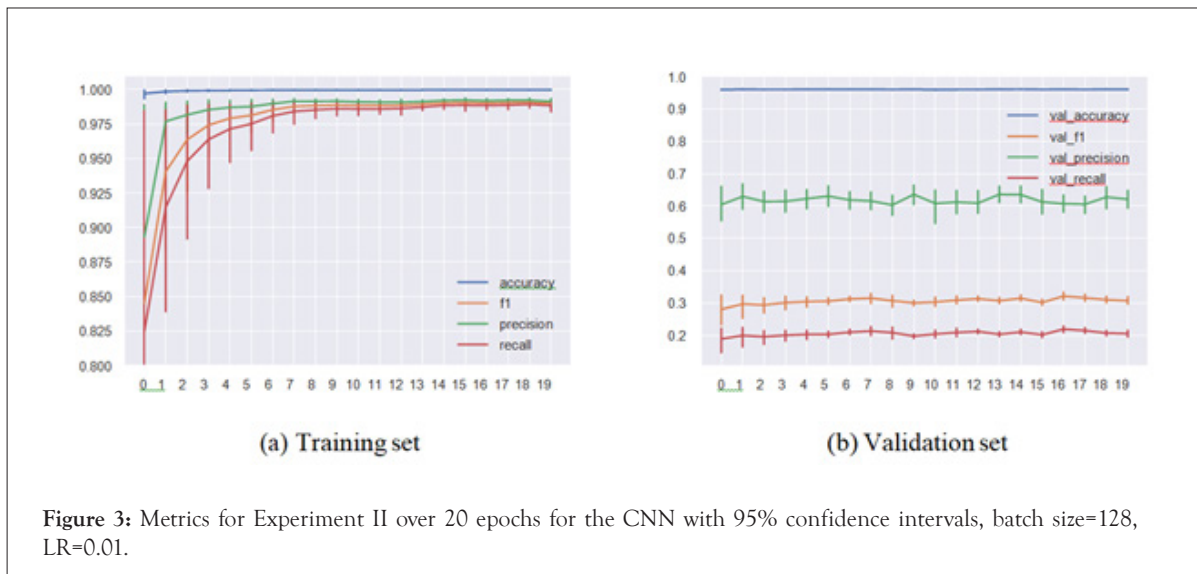
**Table 4:** Experiment II, Count/Mean Probability table for 97 SARS-CoV-2 sequences, compounded results of 10 trials. Top 20 antivirals shown in a and b.

<b>(a) LSTM</b>			<b>(b) CNN</b>		
Antiviral	Count	Mean probability	Antiviral	Count	Mean probability
Tilorone	321	0.627 ± 0.029	Tilorone	177	0.885 ± 0.031
Brincidofovir	197	0.457 ± 0.023	Brincidofovir	116	0.807 ± 0.044
Ganciclovir	113	0.469 ± 0.040	Ritonavir	114	0.808 ± 0.046
Ritonavir	106	0.589 ± 0.049	Lopinavir	93	0.881 ± 0.030
Lopinavir	106	0.588 ± 0.049	Valacyclovir	35	0.748 ± 0.085
Valacyclovir	93	0.522 ± 0.046	Ganciclovir	31	0.793 ± 0.089
Cidofovir	83	0.400 ± 0.039	Cidofovir	26	0.694 ± 0.10
Foscarnet	58	0.473 ± 0.047	Nitazoxanide	24	0.767 ± 0.13
Artesunate	55	0.385 ± 0.048	Artesunate	23	0.740 ± 0.10
Nitazoxanide	48	0.359 ± 0.048	Rapamycin	12	0.769 ± 0.14
Rapamycin	32	0.399 ± 0.065	Foscarnet	18	0.809 ± 0.11
Pleconaril	30	0.626 ± 0.059	Cyclosporine	10	0.780 ± 0.22
Cyclosporine	24	0.429 ± 0.084	Ribavirin	9	0.731 ± 0.23
Chloroquine	21	0.499 ± 0.11	Aciclovir	8	0.671 ± 0.20
Letermovir	19	0.395 ± 0.089	Chloroquine	6	0.938 ± 0.064
Aciclovir	19	0.347 ± 0.077	Thymalfasin	5	0.826 ± 0.38
Ribavirin	17	0.415 ± 0.10	Azithromycin	5	0.648 ± 0.49
CYT107	5	0.369 ± 0.39	CYT107	3	0.540 ± 0.94
Tenofovir	5	0.340 ± 0.20	Tenofovir	2	0.990 ± 0.11
Alisporivir	3	0.550 ± 0.80	Letermovir	2	0.550 ± 0.88

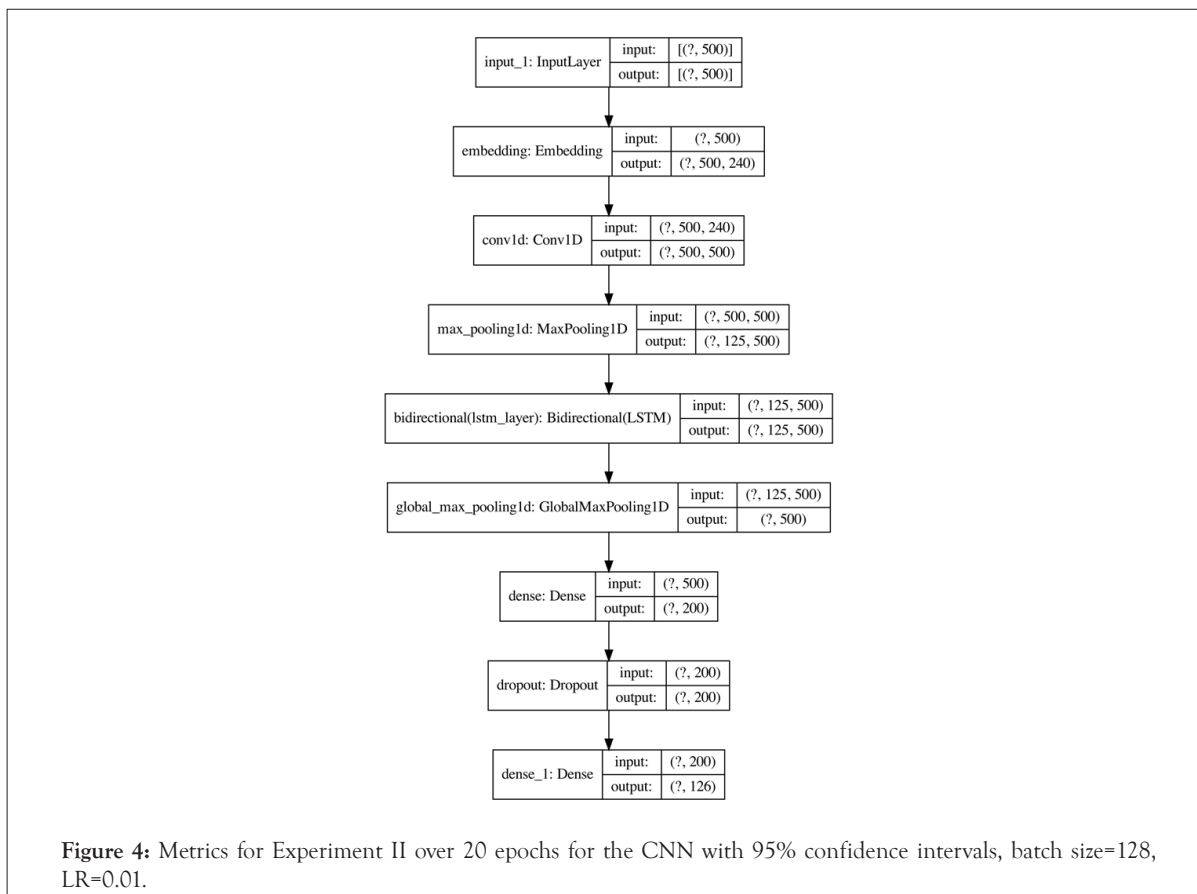
**Predictions for SARS-CoV-2:** With some variation between the two, both the LSTM (Table 4a) and the CNN (Table 4b) seem to converge on a number of drugs: ritonavir, lopinavir (both Phase III for MERS-CoV), tilorone (Approved for MERS-CoV) and brincidofovir are in the top five candidates in both, while valacyclovir, ganciclovir, rapamycin and cidofovir rank high up in both lists. Most of the remaining drugs are present in both lists as well. The LSTM is more conservative in its predictions than the CNN, and the overall counts for SARS-CoV-2 are significantly lower than for

Herpes simplex virus 1 for both, pointing a comparable lack of confidence on the models' part in predicting SARS-CoV-2 sequences (Figures 3-8).

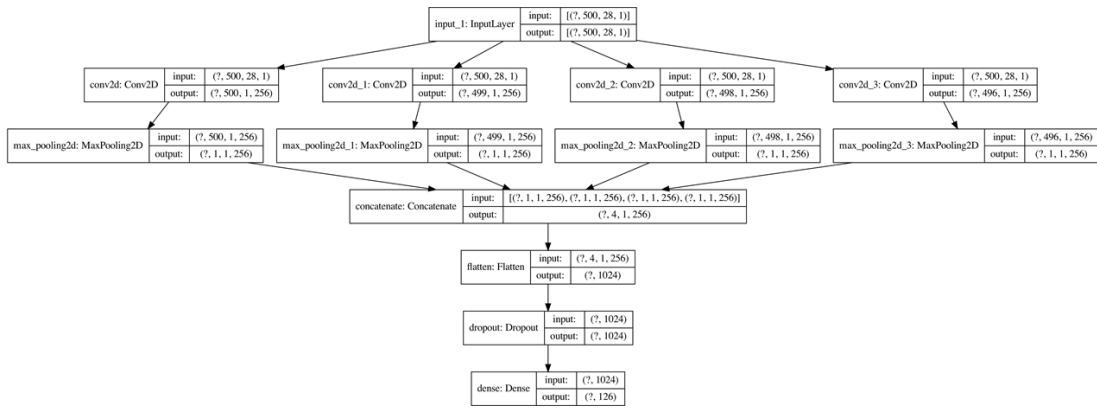
A further step we took for the SARS-CoV-2 sequences was visualizing the layer activations in the Zetane Engine to validate that the model was processing the data at a fine-grained level. This was done in similar fashion to a study where integrated gradients were used to generate attributions on a neural network performing molecule classification [13]. The layer activations in both models showed that different antivirals activated different subsequences of a given sequence at the amino acid level, thus validating our approach. The filter activations are available in Appendix H (Tables 4-10).



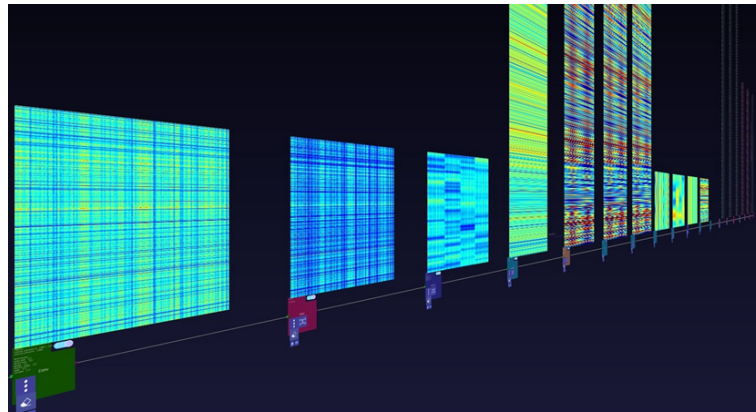
**Figure 3:** Metrics for Experiment II over 20 epochs for the CNN with 95% confidence intervals, batch size=128, LR=0.01.



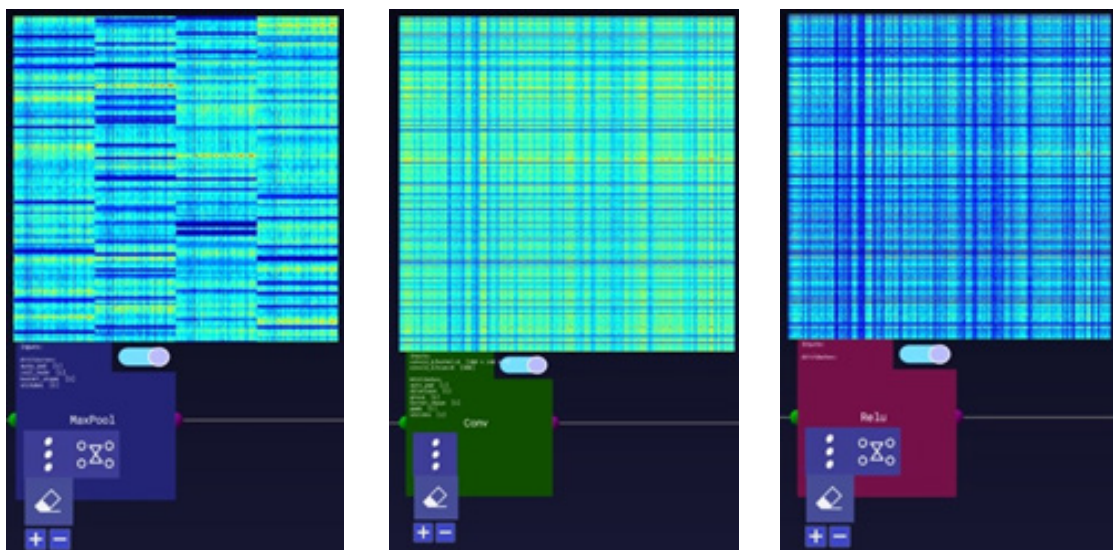
**Figure 4:** Metrics for Experiment II over 20 epochs for the CNN with 95% confidence intervals, batch size=128, LR=0.01.



**Figure 5:** CNN architecture used. Number of trainable parameters: 209,022. "?" denotes the batch dimension. For more information on architecture selection, see Section F, Hyperparameter and Architecture Selection.



**Figure 6:** Activation visualizations of the LSTM layers for a given sequence.

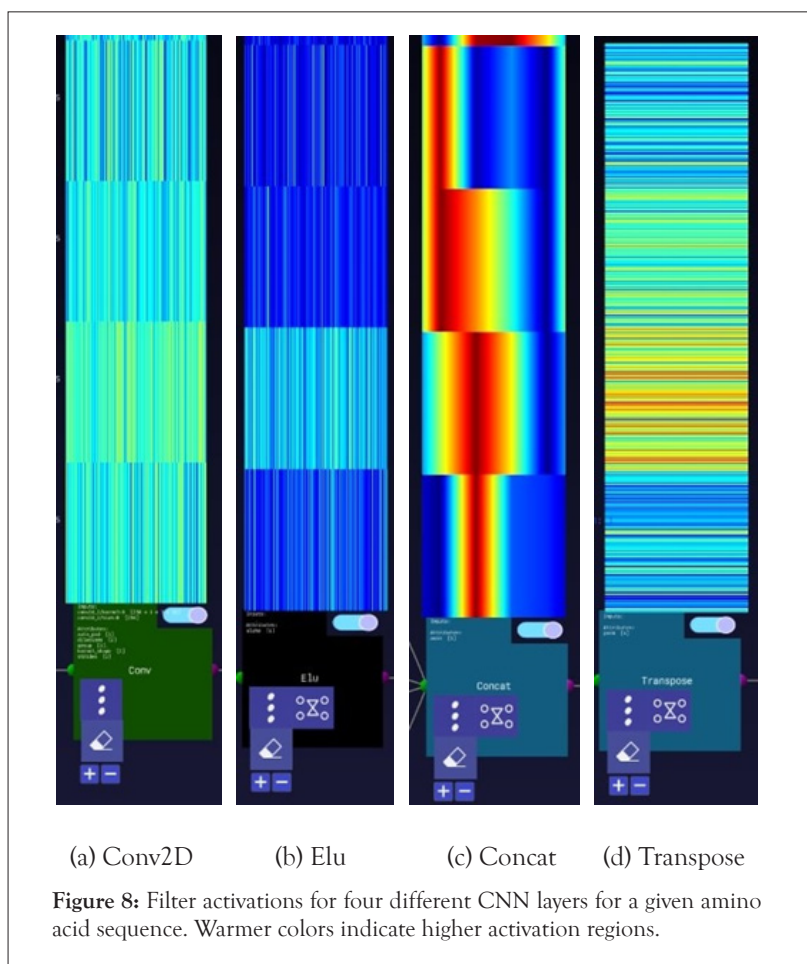


(a) MaxPool

(b) Conv1D

(c) ReLU

**Figure 7:** Filter activations for three different LSTM layers for a given amino acid sequence. Warmer colors indicate higher activation regions.



**Table 5:** A section of the merged database used in model training. Each sequence is associated with a virus species and paired with a length-126 binary vector of antiviral drugs (only 4 elements of the vector are visible), where a 1 denotes effectiveness linked with a given amino acid sequence. Only the first 10 elements of the amino acid sequence are displayed for brevity.

FASTA sequence	Virus name	Genbank_title	4-HPR	ABT-263	Aciclovir	Alisporivir
YIDPVVVLDF	Varicella zoster virus	DNA polymerase catalytic subunit, partial	0	0	1	0
YIDPVVVLDF	Varicella zoster virus	DNA polymerase catalytic subunit, partial	0	0	1	0
TGFYIDPVVV	Varicella zoster virus	DNA polymerase catalytic subunit, partial	0	0	1	0
ANYNHSDEYL	Hepatitis A virus	VP1, partial	0	0	0	0
STTQPQLQTT	Herpes simplex virus 1	glycoprotein G, partial	0	0	1	0
NDPITNAVES	Enterovirus A	VP1, partial	0	0	0	0
MGAQVSTQKT	Enterovirus B	polyprotein, partial	0	0	0	0
IFEKHSRYKY	Norovirus	RNA replicase, partial	0	0	0	0
GCSFSIFLLA	Hepatitis C virus	polyprotein, partial	0	0	0	1
KKDLKPQTT	Respiratory syncytial virus	glycoprotein, partial	0	0	0	0

**Table 6:** Virus counts before dropping duplicate sequences..

Virus name	Count
Hepatitis C virus	29620
Epstein-Barr virus	16040
Cytomegalovirus	13089
Norovirus	7924
Dengue virus	7720



Hepatitis B virus	7642
Herpes simplex virus 1	7039
Ebola virus	6250
Respiratory syncytial virus	6067
Varicella zoster virus	6061
Herpes simplex virus 2	5890
Human papillomavirus	4613
Adenovirus	4394
Human herpesvirus 6	3622
Human parainfluenza virus 2	2518
Enterovirus A	2511
Chikungunya virus	2105
Measles virus	1672
Lassa virus	1222
MERS coronavirus	1171
Enterovirus C	1095
Enterovirus D	1063
SARS coronavirus	867
Zika virus	845
HHV-8	838
SFTS virus	835
Enterovirus B	831
Human coronavirus OC43	795
Human metapneumovirus	761
Hepatitis E virus	694
Human parainfluenza virus 1	591
Molluscum contagiosum virus	491
Parvovirus B19	435
Human coronavirus strain NL63	317
Human rhinovirus A	295
Parechovirus A3	289
Crimean-Congo hemorrhagic fever virus	238
Rabies virus	220
Yellow fever virus	198
Ross River virus	189
West Nile virus	175
Marburg virus	168
Human coronavirus strain 229E	156
Japanese encephalitis virus	153
Hepatitis A virus	149
Rubella virus	131
Human rhinovirus A B	110
Vaccinia virus	107
Human herpesvirus 7	107
Rift Valley fever virus	100
Nipah virus	81
Hantavirus	76
Tick-borne encephalitis virus	76
John Cunningham virus	50
Saffold virus	49
Human Astrovirus	39
Junin virus	28
LCM virus	21
Sindbis virus	12
Bunyamwera virus	8
Powassan virus	5
Andes virus	4
Sin Nombre virus	3

Table 7: Virus counts after dropping duplicate sequences.

Virus name	Count
Hepatitis C virus	870
Dengue virus	691
Adenovirus	657
Cytomegalovirus	654
Hepatitis B virus	644
Herpes simplex virus 2	635
Herpes simplex virus 1	588
Norovirus	579
Epstein-Barr virus	545
Respiratory syncytial virus	474
Enterovirus B	468
Enterovirus A	417
Human papillomavirus	386
Chikungunya virus	358
Vaccinia virus	348
Human herpesvirus 6	322
Enterovirus D	301
Hepatitis E virus	293
Human rhinovirus A	289
Human metapneumovirus	284
Molluscum contagiosum virus	284
Enterovirus C	271
HHV-8	264
Parechovirus A3	247
Zika virus	223
Varicella zoster virus	213
MERS coronavirus	211
Hepatitis A virus	207
Parvovirus B19	200
Lassa virus	188
Ebola virus	164
Hantavirus	162
Human parainfluenza virus 2	149
Human rhinovirus B	134
Human Astrovirus	132
Measles virus	120
Crimean-Congo hemorrhagic fever virus	118
Saffold virus	112
Human coronavirus OC43	104
Human coronavirus strain NL63	87
Human herpesvirus 7	87
John Cunningham virus	83
Rabies virus	81
SFTS virus	79
West Nile virus	75
Rubella virus	73
Human parainfluenza virus 1	73
Hepatitis D virus	65
Human coronavirus strain 229E	57
Rift Valley fever virus	47
Yellow fever virus	40
Tick-borne encephalitis virus	36
Japanese encephalitis virus	32
Nipah virus	29
SARS coronavirus	22
Andes virus	20
Junin virus	15
LCM virus	15
Ross River virus	12
Marburg virus	12
Sin Nombre virus	11
Sindbis virus	4
Bunyamwera virus	4

**Table 8:** Virus counts after dropping duplicate sequences and balancing.

Virus name	Count
Enterovirus B	936
Hepatitis C virus	870
Human metapneumovirus	852
Enterovirus A	834
Human parainfluenza virus 1	803
HHV-8	792
Human papillomavirus	772
Parechovirus A3	741
Chikungunya virus	716
Human coronavirus strain NL63	696
Vaccinia virus	696
Human herpesvirus 7	696
Dengue virus	691
Zika virus	669
John cunningham virus	664
Adenovirus	657
Ebola virus	656
Cytomegalovirus	654
Hantavirus	648
Rabies virus	648
Hepatitis B virus	644
Human herpesvirus 6	644
Varicella zoster virus	639
Herpes simplex virus 2	635
MERS coronavirus	633
SFTS virus	632
Hepatitis A virus	621
Enterovirus D	602
Measles virus	600
West Nile virus	600
Parvovirus B19	600
Human parainfluenza virus 2	596
Crimean-Congo hemorrhagic fever virus	590
Herpes simplex virus 1	588
Hepatitis E virus	586
Rubella virus	584
Norovirus	579
Human rhinovirus A	578
Molluscum contagiosum virus	568
Lassa virus	564
Saffold virus	560
Epstein-Barr virus	545
Enterovirus C	542
Human rhinovirus B	536
Human Astrovirus	528
Human coronavirus OC43	520
Respiratory syncytial virus	474

**Table 9:** List of hyperparameters experimented with. The hyperparameters used for generating the results in the article are highlighted in bold.

Hyperparameter	LSTM	CNN
Learning rate	<b>1e-1, 1e-2, 1e-3, 1e-4</b>	1e-1, 1e-2, 1e-3, 1e-4
Batch size	<b>32, 64, 128, 256, 512</b>	32, 64, 128, 256, 512
Threshold	<b>0.9, 0.7, 0.5, 0.2, 0.1</b>	0.9, 0.7, 0.5, 0.2
Epochs	<b>5 to 30 (20)</b>	5 to 30 (20)
Sequence length cutoff	<b>500, 1000</b>	500, 1000

**Table 10:** A section of sample outputs for amino acid sequences and their associated antivirals. Post-processing outputs a list of drugs that were selected along with the respective probabilities of the drugs being “effective” against the virus with the given amino acid sequence.

FASTA sequence	Virus name	Genbank title	Antivirals	Probabilities
MKFLVFLGII	SARS-CoV-2	ORF8 protein	Rapamycin, Mitoxantrone	0.994, 0.351
MGYINVFAPFT	SARS-CoV-2	ORF10 protein, partial	Brincidofovir, Nitazoxanide, Tilorone	0.959, 0.865, 0.452
MESLVPGFNE	SARS-CoV-2	orf1ab polyprotein	Lopinavir, Ritonavir, Tilorone,	0.761, 0.653, 0.280

## DISCUSSION

The preliminary results of our experiments show promise and merit further investigation. We note that our ML models predict that some antivirals that show promise as treatments against MERS-CoV may also be effective against SARS-CoV-2. These include the broad-spectrum antiviral tilorone and the drug lopinavir, the latter of which is now in Phase IV clinical trials to determine its efficacy against COVID-19 [14-16]. Such observations suggest with confidence that our models can recognize reliable patterns between particular antivirals and species of viruses containing homologous amino acid sequences in their proteome.

Additional observations that support our findings have come to light from a study in *The Lancet* published shortly before this article [3]. This open-label, randomized, Phase II trial observed that the combined administration of the drugs interferon beta-1b, lopinavir, ritonavir and ribavirin provides an effective treatment of COVID-19 in patients with mild to moderate symptoms. Both of our models flagged three of the drugs in that trial (note that interferon was not part of our datasets). In terms of number of occurrences aka Count, ritonavir, lopinavir and ribavirin were ranked 4<sup>th</sup>, 5<sup>th</sup> and 11<sup>th</sup> by the LSTM, while the CNN model ranked them 3<sup>rd</sup>, 4<sup>th</sup> and 10<sup>th</sup>, respectively.

Other studies have also focused on the treatment of SARS-CoV-2 by drugs predicted in our experiments. Wang et al. discovered that nitazoxanide (LSTM rank 10<sup>th</sup>, CNN rank 8<sup>th</sup>) inhibited SARS-CoV-2 at a low-micromolar concentration [6]. Gordon et al. suggest promising results in viral growth and cytotoxicity assays with rapamycin as an inhibitor (11<sup>th</sup> and 10<sup>th</sup>), while de Wilde et al. point out that cyclosporine (13<sup>th</sup> and 12<sup>th</sup>) is known to be effective against diverse coronaviruses [4,5].

Such observations are encouraging. They demonstrate that predictive models may have value in identifying potential therapeutics that merit priority for advanced clinical trials. They also add to growing observations that support using ML to streamline drug discovery. From that perspective, our models suggest that the broad spectrum antiviral tilorone, for instance, may be a top candidate for COVID-19 clinical trials in the near future. Other candidates highlighted by our results and may merit further studies are brincidofovir, foscarnet, artesunate, cidofovir, valacyclovir and ganciclovir.

The antivirals identified here have some discrepancies with emerging research findings as well. For instance, our models did not highlight the widely available anti-parasitic ivermectin. One research study observed that ivermectin could inhibit the replication of SARS-CoV-2 *in vitro* [17]. Another large-scale drug repositioning survey screened a library of nearly 12,000 drugs and identified six candidate antivirals for SARS-CoV-2: PIKfyve kinase inhibitor Apilimod, cysteine protease inhibitors MDL-28170, Z LVG CHN2, VBY-825, and ONO 5334, and the CCR1 antagonist MLN-3897 [18]. It comes as no surprise that our models did not identify these compounds as our data sources did not contain them. Future efforts to strengthen our ML models will thus require us to integrate a growing

bank of novel data from emerging research findings into our ML pipeline.

In terms of our machine learning models, better feature extraction can improve predictions drastically. This step involves improvements through better data engineering and working with domain experts who are familiar with applied bioinformatics to better understand

the nature of our data and find ways to improve our data processing pipeline. Some proposals for future work that could strengthen the performance of our machine learning process are as follows:

1. Deeper interaction with domain experts and further lab testing would lead to a better understanding of the antivirals and the amino-acid sequences they target, leading to building better ML pipelines for drug repurposing.
2. Better handling of duplicates can improve the quality of data available. The current approach (which is based on species and sequence length) can be improved through using string similarity measures such as Dice coefficient, cosine similarity, Levenshtein distance etc.
3. Influenza and HIV datasets should be integrated into the data generation and processing pipeline to enhance available data.
4. Vectorizers can be used to extract features as n-grams (small sequences of chars), which has attained success in similar problems [19]. Other unsupervised learning methods such as singular value decomposition also may be applicable to our study [20].

### Broader impact

We hope that the machine learning approaches and pipelines developed here may provide longterm benefit to public health. The fact that our results show much promise in streamlining drug discovery for SARS-CoV-2 motivates us to adapt our current models so we can conduct identical drug repurposing assessments for other known viruses. Moreover, experimental data suggests that our approaches are generalizable to other viruses (see the HSV-1 example in Section 3.2, Experiment II)-we are therefore confident that we could adapt our models to conduct equivalent studies during the next outbreak of a novel virus. This also means our methods can be used to repurpose existing drugs in order to find more potent treatments for known viruses.

The direct beneficiaries of our findings are members of the clinical research community. Using relatively few resources, ML-guided drug repurposing technology can help prioritize clinical investigations and streamline drug discovery. In addition to reducing costs and expediting clinical innovation, such efficiency gains may reduce the number of clinical trials-and thus human subjects used in risky research-needed to find effective treatments (this pertains to the ethical imperative to avoid harm when possible). Also of importance is that in-silico analyses using machine learning provide yet another means to employ past research findings in new investigations. ML-guided drug repurposing thus provides means to obtain further value from knowledge on-hand; maximizing value in this case is laudable on many fronts, especially in terms of providing maximum benefit from publicly-funded research.

The negative consequences that could arise should our models fail appear limited but are noteworthy nonetheless. Note that our models aim to only indicate possible therapeutics that merit further clinical investigation in order to prove any antiviral activity against SARS-CoV-2. Should our models fail by recommending spurious treatments, these incorrect predictions may divert limited time and resources towards frivolous investigations. It should also be noted that our methods aim to primarily work as guidance for medical experts, and not as a be-all-end-all solution. And any incorrect inferences made by our models are bound to be detected early by medical experts.

## CONCLUSION

Communicating any machine-learning predictions of tentative

antiviral drugs from this study requires much caution. The current pandemic continues to demonstrate how fear, misinformation and a lack of knowledge about a novel communicable disease can encourage counterproductive health-seeking behaviour amongst the public. Soon after the coronavirus became a widely understood threat, the internet was awash in false—sometimes downright harmful—information about preventing and treating COVID-19. Included within this misleading health information were premature claims by some prominent government officials that therapeutics like chloroquine and hydroxychloroquine might hold promise as a repurposed drug for COVID-19. Such unfounded advice caused avoidable poisonings from people self-medicating with chloroquine. Subsequent clinical investigations demonstrated no notable benefit and potential adverse reactions to chloroquine when used to treat COVID-19. Such unfortunate events remind us that preliminary findings may be misinterpreted as conclusive treatments or as evidence to support inconclusive health claims.

## ACKNOWLEDGMENTS AND DISCLOSURE OF FUNDING

We would like to thank the administrators of the Drug Virus and the NCBI Virus Portal for providing the datasets that are central to this study. We appreciate comments on preliminary drafts of this manuscript from Dr Tariq Daouda from the Massachusetts General Hospital, Broad Institute, Harvard Medical school.

The authors declare they will not obtain any direct financial benefit from investigating and reporting on any given pharmaceutical compound. The following study is funded by the authors' employer, Zetane Systems, which produces software for AI technologies implemented in industrial and enterprise contexts.

## REFERENCES

- Stephenson N, Shane E, Chase J, Rowland J, Ries D, Justice N, et al. Survey of machine learning techniques in drug discovery. *Curr Drug Metab.* 2019;20(3):185-193.
- Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: A machine-learning approach through data integration. *J Cheminform.* 2013;5(1):30.
- Hung IFN, Lung KC, Tso EYK, Liu R, Chung TWH, Chu MY, et al. Triple combination of interferon beta-1b, lopinavir-ritonavir, and ribavirin in the treatment of patients admitted to hospital with COVID-19: an open-label, randomised, phase 2 trial. *The Lancet.* 2020;395(10238):1695-1704.
- de Wilde AH, Zevenhoven-Dobbe JC, van der Meer Y, Thiel V, Narayanan K, Makino S, et al. Cyclosporin A inhibits the replication of diverse coronaviruses. *J Gen Virol.* 2011;92(11):2542-2548.
- Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature.* 2020;583(7816):459-468.
- Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. *Cell Res.* 2020;30(3):269-271.
- Andersen PI, Ianevski A, Lysvand H, Vitkauskienė A, Oksenyč V, Bjørås M, et al. Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int J Infect Dis.* 2020;93(1):268-276.
- Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. *Nucleic Acids Res.* 2015;43(Database issue):D571-D577.
- Donner Y, Kazmierczak S, Fortney K. Drug repurposing using deep embeddings of gene expression profiles. *Mol Pharm.* 2018;15(10):4314-4325.
- Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to *in silico* drug repositioning. *Bioinformatics.* 2019;35(24):5191-5198.
- Lee TK, Nguyen T. Protein Family Classification with Neural Networks. *Semantic Scholar.* 2016.
- Hou J, Adhikari B, Cheng J. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics.* 2018;34(8):1295-1303.
- McCloskey K, Taly A, Monti F, Brenner MP, Colwell LJ. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc Natl Acad Sci U S A.* 2019;116(24):11624-11629.
- Ekins S, Lane TR, Madrid PB. Tilorone: A broad-spectrum antiviral invented in the USA and commercialized in Russia and beyond. *Pharm Res.* 2020;37(4):71.
- Yao TT, Qian JD, Zhu WY, Wang Y, Wang GQ. A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus—A possible reference for coronavirus disease-19 treatment option. *J Med Virol.* 2020;92(6):556-563.
- Basha SH. Corona virus drugs—a brief overview of past, present and future. *Semantic Scholar.* 2020.
- Caly L, Druce JD, Catton MG, Jans DA, Wagstaff KM. The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 *in vitro*. *Antiviral Res.* 2020;178(1):104787.
- Riva L, Yuan S, Yin X, Martin-Sancho L, Matsunaga N, Burgstaller-Muehlbacher S, et al. A large-scale drug repositioning survey for SARS-CoV-2 antivirals. *bioRxiv.* 2020;2020(1):143.
- Szalkai B, Grolmusz V. Near Perfect Protein Multi-Label Classification with Deep Neural Networks. *arXiv.* 2017.
- Wu C, Berry M, Shivakumar S, McLarty J. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning.* 1995;21(1):177-193.