Machine-Learning Driven Drug Repurposing for COVID-19

Semih Cantürk *
Zetane Systems
1870 Boulevard des Sources, Suite 307
Pointe-Claire, Canada H9R 5N4
semih@zetane.com

Aman Singh *
Zetane Systems
1870 Boulevard des Sources, Suite 307
Pointe-Claire, Canada H9R 5N4
aman@zetane.com

Patrick St-Amant *
Zetane Systems
1870 Boulevard des Sources, Suite 307
Pointe-Claire, Canada H9R 5N4
patrick@zetane.com

Jason Behrmann
Zetane Systems
1870 Boulevard des Sources, Suite 307
Pointe-Claire, Canada H9R 5N4
jason@zetane.com

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 DrugVirus 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 viruses.

---
*Equal contributions

1   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 [3], cyclosporine, [4], rapamycin [5], and nitazoxanide [6]. For the 7 other predicted drugs, further research is needed to evaluate their effectiveness against SARS-CoV-2.

2   Method

2.1   Data

2.1.1   Sourcing and Preparation

We used two main data sources for this study. The first database was the DrugVirus database [7]; DrugVirus provides a database of broad-spectrum antiviral agents (BSAAs) and the associated viruses they inhibit. 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.

The second database is the National Center for Biotechnology Information (NCBI) Virus Portal [8]; as of April 2020, this database provides approximately 2.8 million amino-acid and 2 million nucleotide sequences from viruses with humans as hosts. 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 DrugVirus 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 & 2 (1,192,754 sequences), Influenza types A, B & 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.

2.1.2 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 DrugVirus 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 DrugVirus dataset to a format suitable for merging with the NCBI data frame. Every row of the DrugVirus 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 DrugVirus 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 & 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 DrugVirus database.

The full dataset was then generated by merging this "new" version of the DrugVirus 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.

2.1.3 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 undersampled 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.

2.1.4 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.

2.2 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 convolutions 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 x 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 1x28, 2x28, 3x28 and 5x28 respectively, which are maxpooled, concatenated and passed through a fully connected layer. A more detailed architecture diagram is available in Appendix E.

4

2.3 Experiment Setup

The experiments were run on a computer with an 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 five 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. Post-processing 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.

3 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.

3.1 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*.

3.2 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

Table 1: *Experiment I*, metrics for the best set of hyperparameters tested

Training set					
Model	Accuracy	Precision	Recall	F1-score	Loss
LSTM	0.998	0.936	0.877	0.904	0.0156
CNN	0.999	0.992	0.983	0.988	0.00268
Validation set					
Model	Accuracy	Precision	Recall	F1-score	Loss
LSTM	0.997	0.958	0.809	0.875	0.0259
CNN	0.999	0.978	0.936	0.958	0.0117

(a) Training set	(b) Validation set

Figure 1: Metrics for *Experiment I* over 20 epochs for the LSTM, learning rate = 0.001

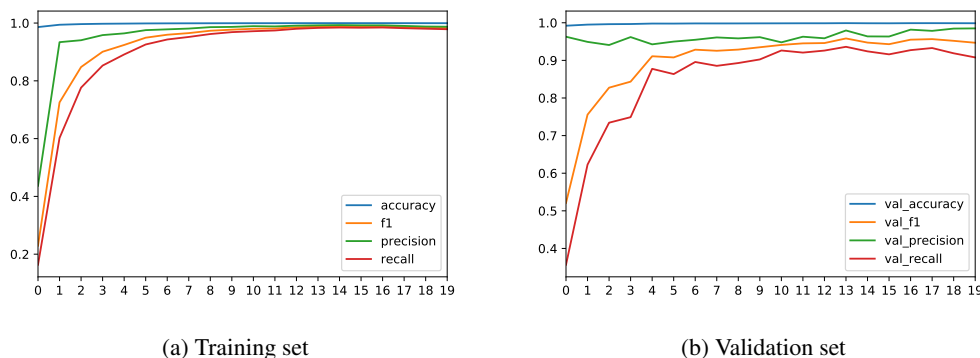|	
(a) Training set	(b) Validation set

Figure 2: Metrics for *Experiment I* over 20 epochs for the CNN, learning rate = 0.01

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.

Upon examination of drug predictions for Herpes simplex virus 1 (HSV-1), however, we see that our CNN was in fact quite successful. In Table 3 and Table 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.

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) and tilorone (Approved for MERS-CoV) are the top three candidates in both, while brincidofovir, rapamycin, cidofovir, valacyclovir and ganciclovir rank high up in both lists.

6

Table 2: *Experiment II*, metrics for the best set of hyperparameters tested

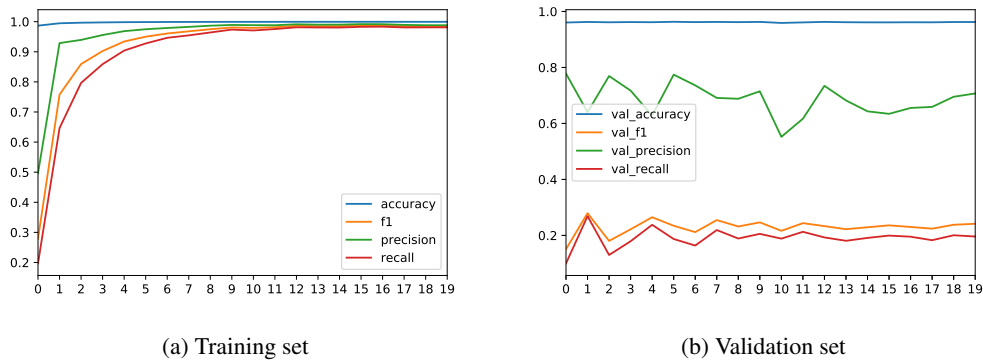| Training set					
Model	Accuracy	Precision	Recall	F1-score	Loss
LSTM	0.998	0.936	0.877	0.904	0.0156
CNN	0.999	0.992	0.983	0.988	0.00268
Validation set					
Model	Accuracy	Precision	Recall	F1-score	Loss
LSTM	0.997	0.958	0.809	0.875	0.0259
CNN	0.999	0.978	0.936	0.958	0.0117

(a) Training set   (b) Validation set

Figure 3: Metrics for *Experiment II* over 20 epochs for the CNN, learning rate = 0.01

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

Antiviral	Count	Mean Probability
Valacyclovir	**578**	**0.899**
Cidofovir	**567**	**0.737**
Foscarnet	**566**	**0.904**
Brincidofovir	**559**	**0.646**
Aciclovir	**525**	**0.759**
Rapamycin	419	0.540
Ganciclovir	410	0.552
Artesunate	333	0.453
Cyclosporine	328	0.449
Letermovir	317	0.430
Tilorone	**243**	**0.433**
Tenofovir	173	0.573

Drugs with more than 100 counts are listed.
Antivirals used for Phase II and further trials
for HSV-1 are highlighted in **bold**.

Table 4: *Experiment II*, Count/Mean Probability table for 97 **SARS-CoV-2** sequences

Antiviral	Count	Mean Probability
Ritonavir	19	0.475
Lopinavir	19	0.443
Tilorone	16	0.639
Brincidofovir	7	0.507
Valacyclovir	6	0.654
Ganciclovir	6	0.563
Cidofovir	6	0.457
Rapamycin	5	0.367
Foscarnet	4	0.736
Nitazoxanide	4	0.506
Ribavirin	4	0.344
Cyclosporine	2	0.387
Aciclovir	2	0.359
Artesunate	1	0.549
Letermovir	1	0.396
Tenofovir	1	0.272
Chloroquine	1	0.240

(a) LSTM

Antiviral	Count	Mean Probability
Tilorone	64	0.472
Lopinavir	56	0.790
Ritonavir	55	0.740
Brincidofovir	17	0.729
Nitazoxanide	16	0.673
Rapamycin	12	0.783
Artesunate	11	0.781
Cyclosporine	9	0.790
Cidofovir	8	0.812
Valacyclovir	7	0.656
Ganciclovir	7	0.641
Letermovir	6	0.607
Ribavirin	4	0.713
Chloroquine	4	0.697
Foscarnet	3	0.450
Pleconaril	2	0.967
Thymalfasin	1	0.930
CYT107	1	0.917
Mitoxantrone	1	0.351
Erlotinib	1	0.208

(b) CNN

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.

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.

4   Discussion and Future Work

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 [14] and the drug lopinavir [15], the latter of which is now in Phase IV clinical trials to determine its efficacy against COVID-19 [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 1st, 2nd and 11th by the LSTM, while the CNN model ranked them 2nd, 3rd and 13th, 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 10th, CNN rank 5th) 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 (8th & 6th) [5], while de Wilde et al. point out that cyclosporine (12th & 8th) is known to be effective against diverse coronaviruses [4].

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 long-term 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.

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

References

[1] N. Stephenson, E. Shane, J. Chase, *et al.*, "Survey of Machine Learning Techniques in Drug Discovery," en, *Current Drug Metabolism*, vol. 20, no. 3, pp. 185–193, May 2019, ISSN: 13892002. DOI: 10.2174/1389200219666180820112457.

[2] F. Napolitano, Y. Zhao, V. M. Moreira, *et al.*, "Drug repositioning: A machine-learning approach through data integration," en, *Journal of Cheminformatics*, vol. 5, no. 1, p. 30, Dec. 2013, ISSN: 1758-2946. DOI: 10.1186/1758-2946-5-30.

[3] I. F.-N. Hung, K.-C. Lung, E. Y.-K. Tso, *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," en, *The Lancet*, S01406736203104024, May 2020, ISSN: 01406736. DOI: 10.1016/S0140-6736(20)31042-4.

[4] A. H. de Wilde, J. C. Zevenhoven-Dobbe, Y. van der Meer, *et al.*, "Cyclosporin A inhibits the replication of diverse coronaviruses," en, *Journal of General Virology*, vol. 92, no. 11, pp. 2542–2548, Nov. 2011, ISSN: 0022-1317, 1465-2099. DOI: 10.1099/vir.0.034983-0.

[5] D. E. Gordon, G. M. Jang, M. Bouhaddou, *et al.*, "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," en, *Nature*, Apr. 2020, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2286-9.

[6] M. Wang, R. Cao, L. Zhang, *et al.*, "Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro," en, *Cell Research*, vol. 30, no. 3, pp. 269–271, Mar. 2020, ISSN: 1001-0602, 1748-7838. DOI: 10.1038/s41422-020-0282-0.

[7]   P. I. Andersen, A. Ianevski, H. Lysvand, *et al.*, "Discovery and development of safe-in-man broad-spectrum antiviral agents," en, *International Journal of Infectious Diseases*, vol. 93, pp. 268–276, Apr. 2020, ISSN: 12019712. DOI: 10.1016/j.ijid.2020.02.018.

[8]   J. R. Brister, D. Ako-adjei, Y. Bao, and O. Blinkova, "NCBI Viral Genomes Resource," en, *Nucleic Acids Research*, vol. 43, no. D1, pp. D571–D577, Jan. 2015, ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gku1207.

[9]   Y. Donner, S. Kazmierczak, and K. Fortney, "Drug Repurposing Using Deep Embeddings of Gene Expression Profiles," en, *Molecular Pharmaceutics*, vol. 15, no. 10, pp. 4314–4325, Oct. 2018, ISSN: 1543-8384, 1543-8392. DOI: 10.1021/acs.molpharmaceut.8b00284.

[10]   X. Zeng, S. Zhu, X. Liu, *et al.*, "deepDR: A network-based deep learning approach to in silico drug repositioning," en, *Bioinformatics*, vol. 35, no. 24, L. Cowen, Ed., pp. 5191–5198, Dec. 2019, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz418.

[11]   T. K. Lee and T. Nguyen, "Protein Family Classification with Neural Networks," en, 2016.

[12]   J. Hou, B. Adhikari, and J. Cheng, "DeepSF: Deep convolutional neural network for mapping protein sequences to folds," en, *Bioinformatics*, vol. 34, no. 8, A. Valencia, Ed., pp. 1295–1303, Apr. 2018, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btx780.

[13]   K. McCloskey, A. Taly, F. Monti, *et al.*, "Using attribution to decode binding mechanism in neural network models for chemistry," en, *Proceedings of the National Academy of Sciences*, p. 201 820 657, May 2019, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1820657116.

[14]   S. Ekins, T. R. Lane, and P. B. Madrid, "Tilorone: A Broad-Spectrum Antiviral Invented in the USA and Commercialized in Russia and beyond," en, *Pharmaceutical Research*, vol. 37, no. 4, p. 71, Apr. 2020, ISSN: 0724-8741, 1573-904X. DOI: 10.1007/s11095-020-02799-8.

[15]   T. Yao, J. Qian, W. Zhu, *et al.*, "A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus—A possible reference for coronavirus disease-19 treatment option," en, *Journal of Medical Virology*, vol. 92, no. 6, pp. 556–563, Jun. 2020, ISSN: 0146-6615, 1096-9071. DOI: 10.1002/jmv.25729.

[16]   S. H. Basha, "Corona virus drugs – a brief overview of past, present and future," en, vol. 2, no. 2, p. 16, 2020.

[17]   L. Caly, J. D. Druce, M. G. Catton, *et al.*, "The FDA-approved drug ivermectin inhibits the replication of SARS-CoV-2 in vitro," en, *Antiviral Research*, vol. 178, p. 104 787, Jun. 2020, ISSN: 01663542. DOI: 10.1016/j.antiviral.2020.104787.

[18]   L. Riva, S. Yuan, X. Yin, *et al.*, "A Large-scale Drug Repositioning Survey for SARS-CoV-2 Antivirals," en, Microbiology, preprint, Apr. 2020. DOI: 10.1101/2020.04.16.044016.

[19]   B. Szalkai and V. Grolmusz, "Near Perfect Protein Multi-Label Classification with Deep Neural Networks," en, *arXiv:1703.10663 [cs, q-bio, stat]*, Mar. 2017, arXiv: 1703.10663.

[20]   C. Wu, M. Berry, S. Shivakumar, and J. McLarty, "Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition," en, *Machine Learning*, vol. 21, no. 1-2, pp. 177–193, 1995, ISSN: 0885-6125, 1573-0565. DOI: 10.1007/BF00993384.
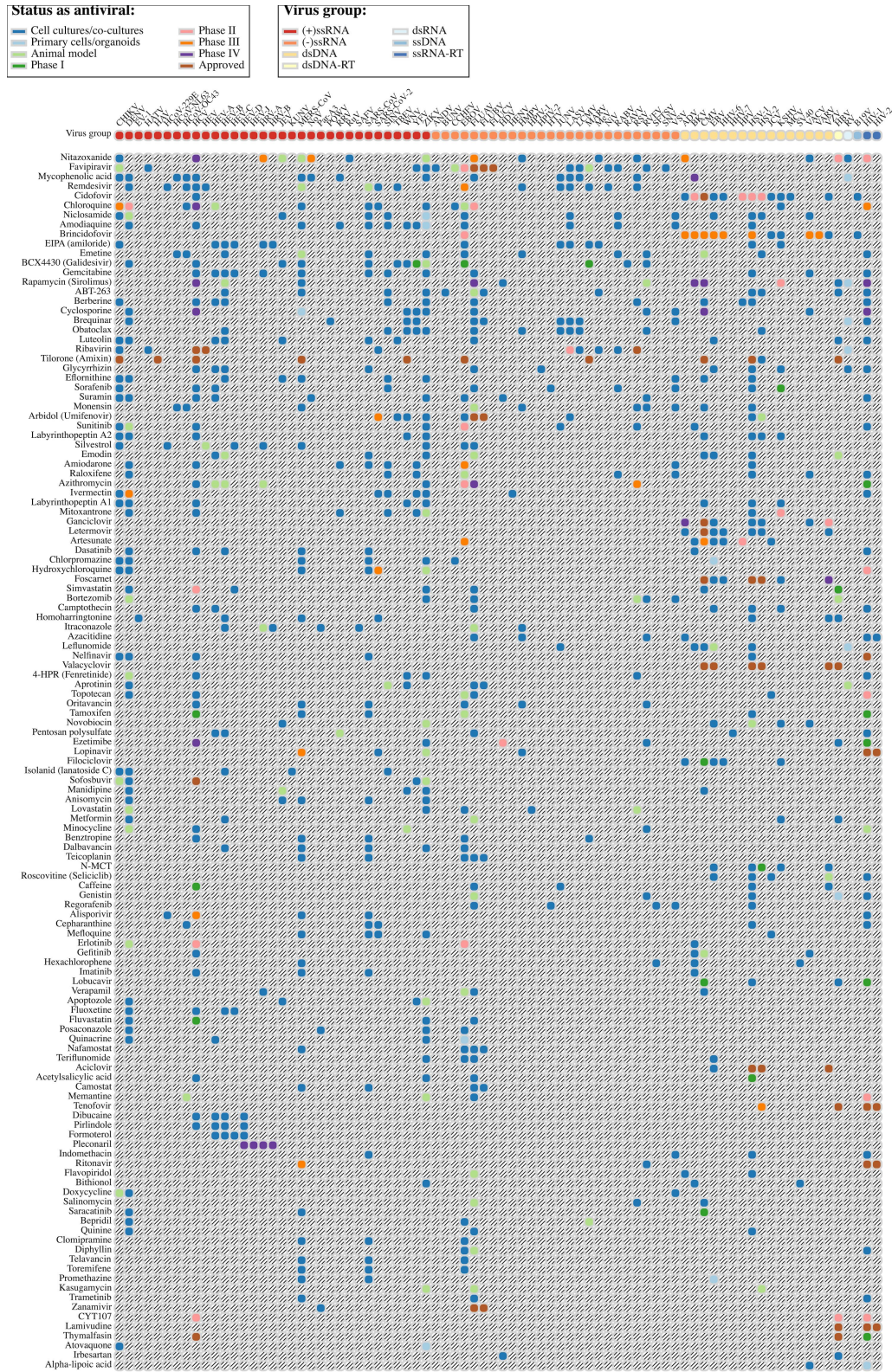
Appendices

Table of Contents

A The DrugVirus database displayed in a pivot table

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

C   Database profile

C.1   Virus counts before dropping duplicate sequences

Table 6: Virus counts before dropping duplicate sequences

Virus Name	Count	Virus Name	Count
Hepatitis C virus	29620	Parvovirus B19	435
Epstein-Barr virus	16040	Human coronavirus strain NL63	317
Cytomegalovirus	13089	Human rhinovirus A	295
Norovirus	7924	Parechovirus A3	289
Dengue virus	7720	Crimean-Congo hemorrhagic fever virus	238
Hepatitis B virus	7642	Rabies virus	220
Herpes simplex virus 1	7039	Yellow fever virus	198
Ebola virus	6250	Ross River virus	189
Respiratory syncytial virus	6067	West Nile virus	175
Varicella zoster virus	6061	Marburg virus	168
Herpes simplex virus 2	5890	Human coronavirus strain 229E	156
Human papillomavirus	4613	Japanese encephalitis virus	153
Adenovirus	4394	Hepatitis A virus	149
Human herpesvirus 6	3622	Rubella virus	131
Human parainfluenza virus 2	2518	Human rhinovirusÂ B	110
Enterovirus A	2511	Vaccinia virus	107
Chikungunya virus	2105	Human herpesvirus 7	107
Measles virus	1672	Rift Valley fever virus	100
Lassa virus	1222	Nipah virus	81
MERS coronavirus	1171	Hantavirus	76
Enterovirus C	1095	Tick-borne encephalitis virus	76
Enterovirus D	1063	John Cunningham virus	50
SARS coronavirus	867	Saffold virus	49
Zika virus	845	Human Astrovirus	39
HHV-8	838	Junin virus	28
SFTS virus	835	LCM virus	21
Enterovirus B	831	Sindbis virus	12
Human coronavirus OC43	795	Bunyamwera virus	8
Human metapneumovirus	761	Powassan virus	5
Hepatitis E virus	694	Andes virus	4
Human parainfluenza virus 1	591	Sin Nombre virus	3
Molluscum contagiosum virus	491		

C.2 Virus counts after dropping duplicate sequences

Table 7: Virus counts after dropping duplicate sequences

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

C.3 Virus counts after dropping duplicate sequences and balancing

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

D LSTM Architecture

input_1: InputLayer	input:	[(?, 500)]
	output:	[(?, 500)]

embedding: Embedding	input:	(?, 500)
	output:	(?, 500, 240)

conv1d: Conv1D	input:	(?, 500, 240)
	output:	(?, 500, 500)

max_pooling1d: MaxPooling1D	input:	(?, 500, 500)
	output:	(?, 125, 500)

bidirectional(lstm_layer): Bidirectional(LSTM)	input:	(?, 125, 500)
	output:	(?, 125, 500)

global_max_pooling1d: GlobalMaxPooling1D	input:	(?, 125, 500)
	output:	(?, 500)

dense: Dense	input:	(?, 500)
	output:	(?, 200)

dropout: Dropout	input:	(?, 200)
	output:	(?, 200)

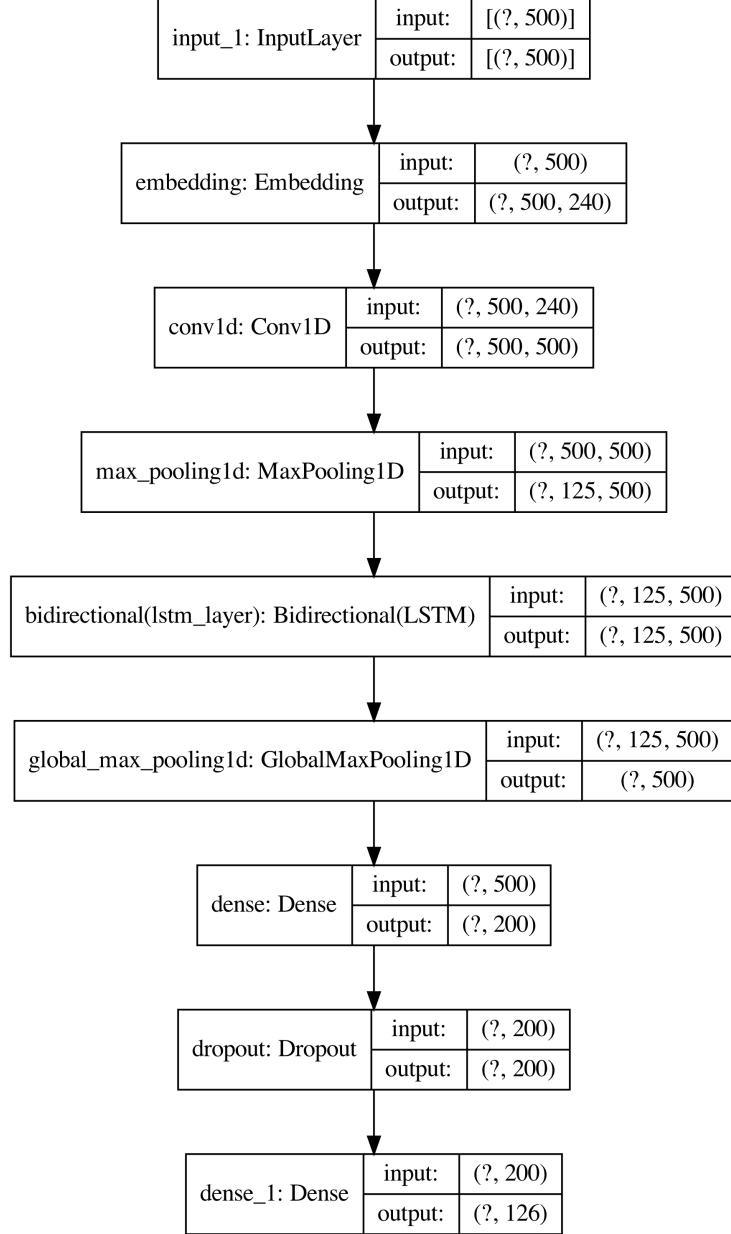| dense_1: Dense	input:	(?, 200)
	output:	(?, 126)

Figure 4: LSTM architecture used. Number of trainable parameters: 1,740,266. "?" denotes the batch dimension. For more information on architecture selection, see Section F, Hyperparameter and Architecture Selection.
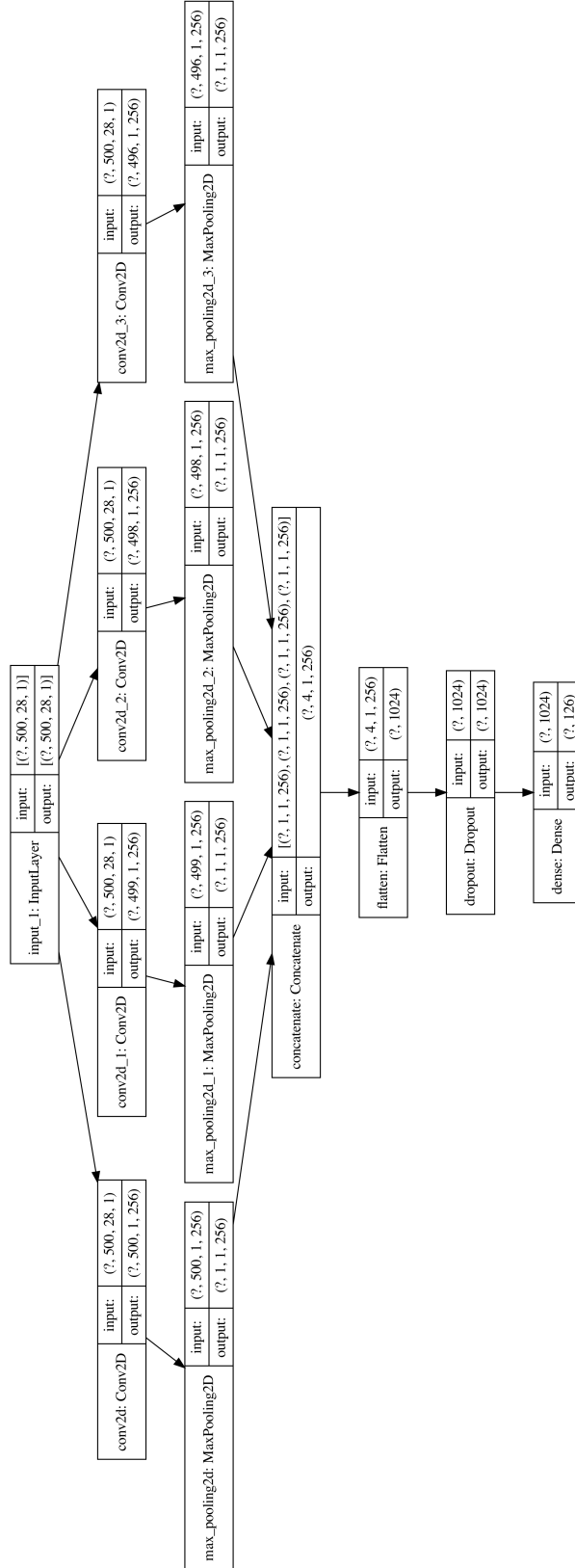
E   CNN Architecture

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.

F   Hyperparameter and Architecture Selection

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	1e-1, 1e-2, **1e-3**, 1e-4	1e-1, **1e-2**, 1e-3, 1e-4
Batch size	32, 64, **128**, 256, 512	32, 64, **128**, 256, 512
Threshold	0.9, 0.7, 0.5, **0.2**, 0.1	0.9, 0.7, 0.5, **0.2**
Epochs	5 to 30 **(20)**	5 to 30 **(20)**
Sequence length cutoff	**500**, 1000	**500**, 1000

The hyperparameters tested in our experiments are presented in Section F It is certainly possible to improve the accuracies of our experiments by conducting a vaster coverage of the loss landscape through more extensive training (e.g. running longer experiments with smaller learning rates on more complex network architectures), especially for results in *Experiment II*. However, due to performance constraints, the scope of hyperparameter tuning as well as the ANN architectures experimented on are relatively constrained as we focused on the methodology as opposed to optimal performance in this study. It should be noted that much improvement is possible in this front, as pointed out in Discussion and Future Work. Additional notes regarding our observations during hyperparameter tuning are presented below.

- For the threshold, we wanted to predict eagerly, i.e. we considered false negatives more costly errors than false positives. A high threshold would mean the outputs would be composed only of the antivirals our models are very confident about per amino acid sequence. This we deem undesirable, as while we do hope these outputs narrow the scope of antivirals to focus on, over-restricting could prevent antivirals that are predicted frequently yet with low probability be detected. A low threshold such as 0.2 filtered the number of antivirals sufficiently, but also left enough breathing room for the domain experts to draw their own conclusions on a per-drug basis.

- While a larger sequence length cutoff was possible and not detrimental to the results, we deemed 500 a suitable trade-off in terms of performance versus accuracy, as many sequences do not reach lengths in the thousands to begin with.

- As mentioned, the number of epochs trained could be increased, as we did not see dramatic signs of overfitting at 20 epochs or further. However, a flattening of the metrics were evident around 20 epochs with the hyperparameters listed, which therefore was selected a suitable stopping point.

G   Output table sample

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
MGYINVFAFPFT	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
MFVFLVLLPL	SARS-CoV-2	surface glycoprotein	Pleconaril	0.963

H   Filter Activation Visualizations
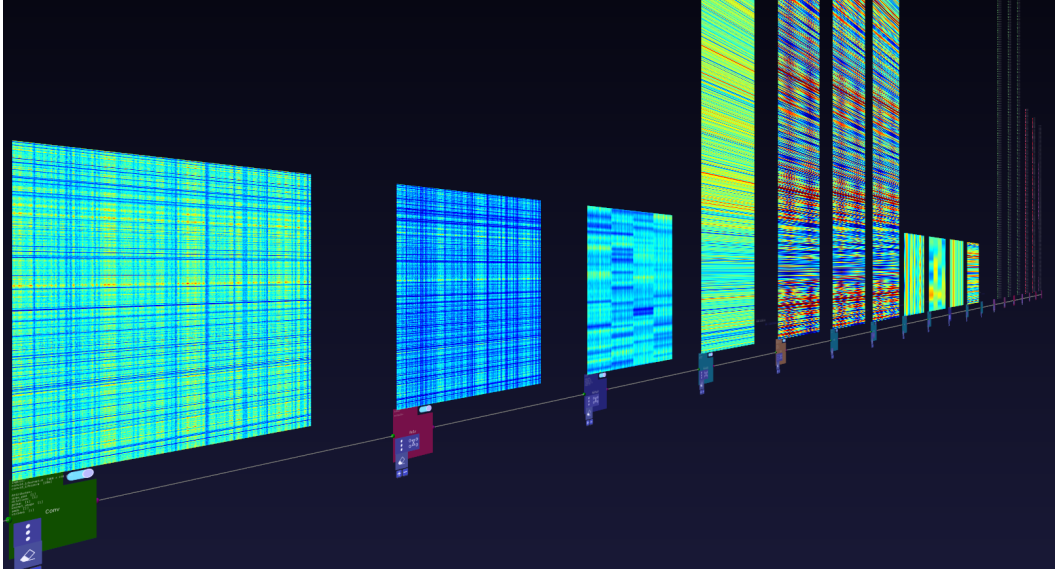
Figure 6: Activation visualizations of the LSTM layers for a given sequence
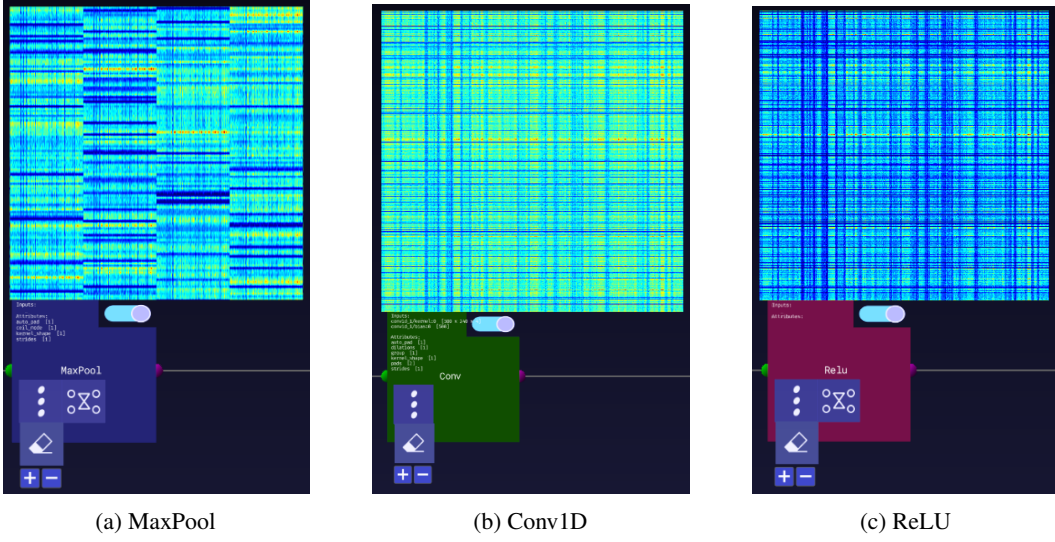

(a) MaxPool

(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.
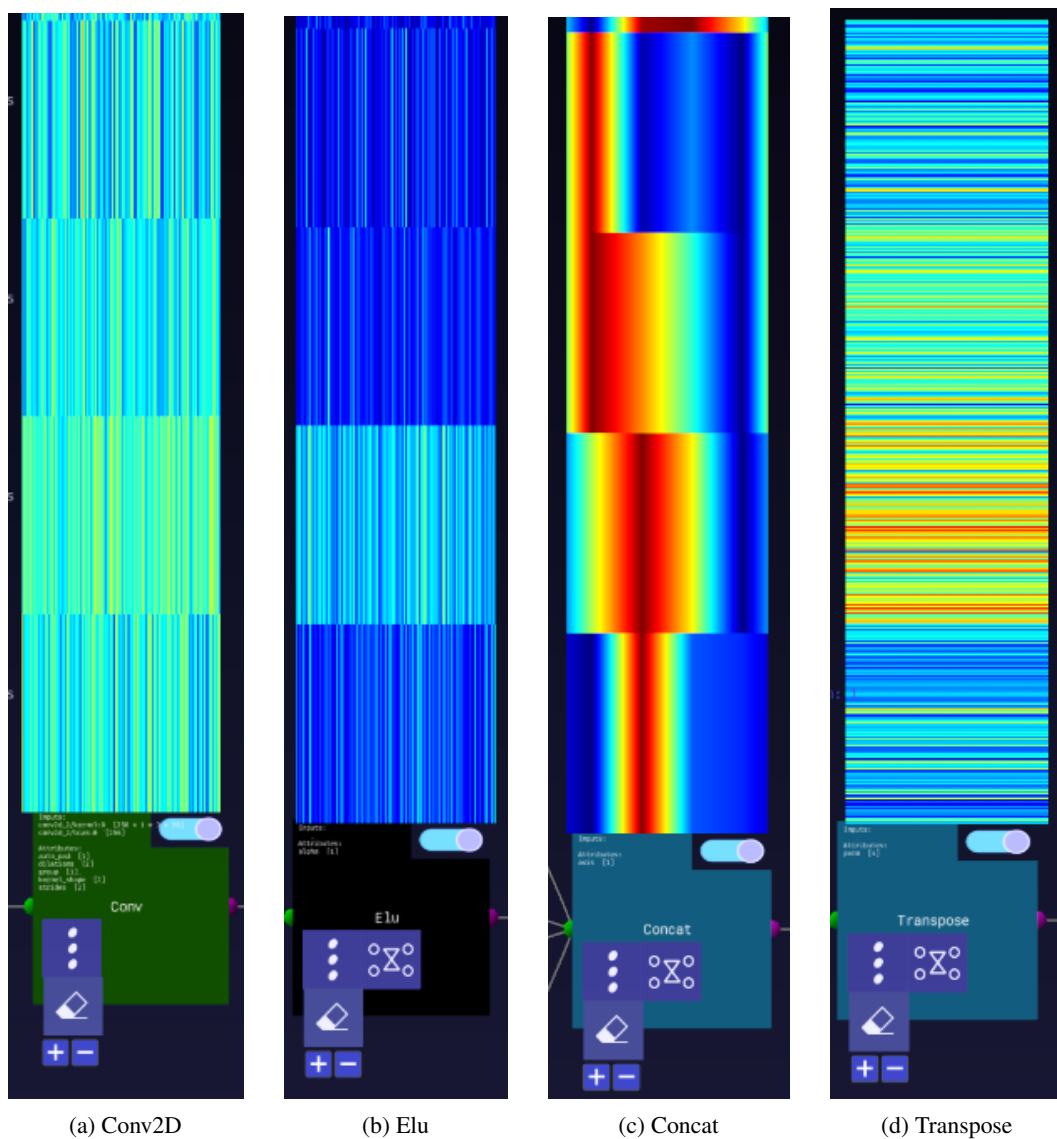
Figure 8: Filter activations for four different CNN layers for a given amino acid sequence. Warmer colors indicate higher activation regions.