Data-driven Drug Repurposing for COVID-19

Raghvendra Mall *[1], Abdurrahman Elbasir[2], Hossam Al Meer[1], Sanjay Chawla[1], and Ehsan Ullah †[1]

[1] *Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, 34110, Qatar*
[2] *ICT Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, 34110, Qatar*

Abstract

Motivation: A global effort is underway to identify drugs for the treatment of COVID-19. Since *de novo* drug design is an extremely long, time-consuming, and expensive process, efforts are underway to discover existing drugs that can be repurposed for COVID-19.
Model: We propose a machine learning representation framework that uses deep learning induced vector embeddings of drugs and viral proteins as features to predict drug-viral protein activity. The prediction model in-turn is used to build an ensemble framework to rank approved drugs based on their ability to inhibit the three main proteases (enzymes) of the SARS-COV-2 virus.
Results: We identify a ranked list of 19 drugs as potential targets including 7 antivirals, 6 anticancer, 3 antibiotics, 2 antimalarial, and 1 antifungal. Several drugs, such as Remdesivir, Lopinavir, Ritonavir, and Hydroxychloroquine, in our ranked list, are currently in clinical trials. Moreover, through molecular docking simulations, we demonstrate that majority of the anticancer and antibiotic drugs in our ranked list have low binding energies and thus high binding affinity with the 3CL-pro protease of SARS-COV-2 virus.
Availability: All code is available at: `https://github.com/raghvendra5688/Drug-Repurposing`

1 Introduction

The breakout of Covid-19 started in December 2019, in China's Hubei province (Dong *et al.*, 2020), and to date, this pandemic has caused over 10 million infections and over 500k deaths worldwide in just eight months (Organization, 2020). There is an immediate need for effective treatment and vaccines to contain the spread of this pandemic. Based on the time and resources required to develop new drugs to treat Covid-19, it is not feasible to rely completely on the traditional process of drug discovery, which takes an average 15 years and costs $2-3 billion to bring a new compound to market (Pushpakom *et al.*, 2019). A more pragmatic approach would be to perform drug repurposing.

Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of their original medical usage (Ashburn and Thor, 2004; Pushpakom *et al.*, 2019). Given a large number of already approved drugs, there is a significant chance that one or several of these could help to treat Covid-19. There are several advantages of drug repurposing compared to de novo drug design. The repurposed drug will have a low risk of failure on several critical criteria as it has already been tested in preclinical models and humans (Pushpakom *et al.*, 2019). The time frame for testing is relatively compact because its safety assessment has already been conducted (Pushpakom *et al.*, 2019; Cheng *et al.*, 2016, 2017; Cheng, 2019). The investment required is low for repurposing of a drug (Pushpakom *et al.*, 2019; Cheng *et al.*, 2016, 2017; Cheng, 2019). The repurposed drugs can also reveal new targets for a given medical condition (Oprea *et al.*, 2011; Pushpakom *et al.*, 2019). For drug repurposing, computational approaches offer advantages over costly and time-consuming experimental techniques (Cheng *et al.*, 2018) and can be the only feasible solution given the large number of candidate drugs which makes in vitro or in vivo testing impractical.

In this paper, we present an integrative computational approach, which combines data from a variety of sources to identify already known drugs as candidates for viral diseases, using Covid-19 as a specific use case. The crux of our approach is based on the observations that a) we can map drug and virus information to the small molecule activity such that drugs with similarities in structure and physio-chemical properties tend to have similar activities for given protein targets, and b) proteins serving as drug targets for other similar viral diseases may be potential targets for Covid-19. For our use case, we focus on primary protein targets of severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) to identify potential drugs that can inhibit these target proteins to prevent viral activity.

Our analysis follows a data-driven perspective. We collect information about various viral organisms, their main proteases and their known (published) small molecule interactions from plethora of resources including ChEMBL (Gaulton *et al.*, 2017), PubChem (Kim *et al.*, 2016), NCBI (Wheeler *et al.*, 2007), UniProt (uni, 2017), DrugBank (Wishart *et al.*, 2018) etc. **In this work, we used the term drugs for small molecules and compounds interchangabley**. The traditional approach for estimating drug (ligand) activity for a particular viral protein (enzyme) is through molecular docking (Kitchen *et al.*, 2004). For performing molecular docking, an inherent requirement is the availability of high-quality 3d crystal structure of the protein of interest as well as annotation information about the presence of active sites (Chakraborti and Srinivasan, 2020). Moreover, it is computationally expensive to perform the docking simulations for a large number of drugs in combination with many viral proteins. However, it is relatively easy to collect information about the primary structure (linear chain of amino acids) for proteases associated with viruses from resources such as UniProt. Moreover, structural and chemical information for drugs in the form of SMILES strings is readily

[*] Corresponding author: rmall@hbku.edu.qa
[†] Corresponding author: eullah@hbku.edu.qa

available in resources such as DrugBank and ChEMBL. Finally, standardized activity (inhibition/potency/affinity) information for a plethora of drug-viral protein combinations is available in PubChem, ChEMBL, and NCBI. These are essential to building data-driven supervised predictive models. The primary notion is that by providing a large dataset of drug-viral activity, machine learning (ML) models can identify frequently occurring patterns in the form of presence of k-mers in the viral protein sequences and subsequences in SMILES representation of drugs that together drive the activity values to be high or low. Our primary contributions are:

- Collection, curation, and assimilation of drug-viral protein activity from resources such as PubChem, ChEMBL, and NCBI leading to ¿60k interactions between ¿50k drugs and ≈ 100 viral organisms.
- Propose autoencoder frameworks to obtain numeric vector representations for drugs and viral proteins respectively, which can be utilized by state-of-the-art traditional supervised ML techniques.
- Propose 4 different end-to-end deep learning techniques to predict drug-viral protein activity scores based on SMILES strings and primary structure of viral proteins.
- Identify a ranked list of 19 drugs as potential therapeutic agents for COVID-19 by targeting the three main proteases of the SARS-COV-2 virus using our data-driven approach. These include 7 antivirals, 3 antibiotics, 6 anticancer, 2 antimalarial, and 1 antifungal drugs, several of which are currently involved in clinical trials.
- Showcase efficiency of the predicted anticancer and antibiotic drugs for inhibiting the 3CL-Pro protease (low binding energy) of the SARS-COV-2 virus through molecular docking simulations.

Figure 1 provides a flowchart of our proposed drug-viral activity prediction framework.

2 MATERIALS

In order to build data-driven predictive models, we collected information about drugs, viral protein sequences, and drug-viral protein interactions (activity scores) from resources such as MOSES (Polykovskiy *et al.*, 2018), ChEMBL, UniProt, PubChem and NCBI. Below we describe in detail the data collection and curation steps required for the preparation of quality data, essential for downstream predictive models.

2.1 Data Collection & Curation

2.1.1 Drugs:

We initially collected 556,134 SMILES strings for drugs used in (Gupta *et al.*, 2018). However, in order to have more robust and realistic molecules, the dataset was augmented with 1,936,962 drugs available in the MOSES dataset (Polykovskiy *et al.*, 2018). Together these two datasets represented ≈ 2.5 million SMILES for drugs. We then filtered this dataset to remove salts and stereochemical information and confined the length of the SMILES strings in the range

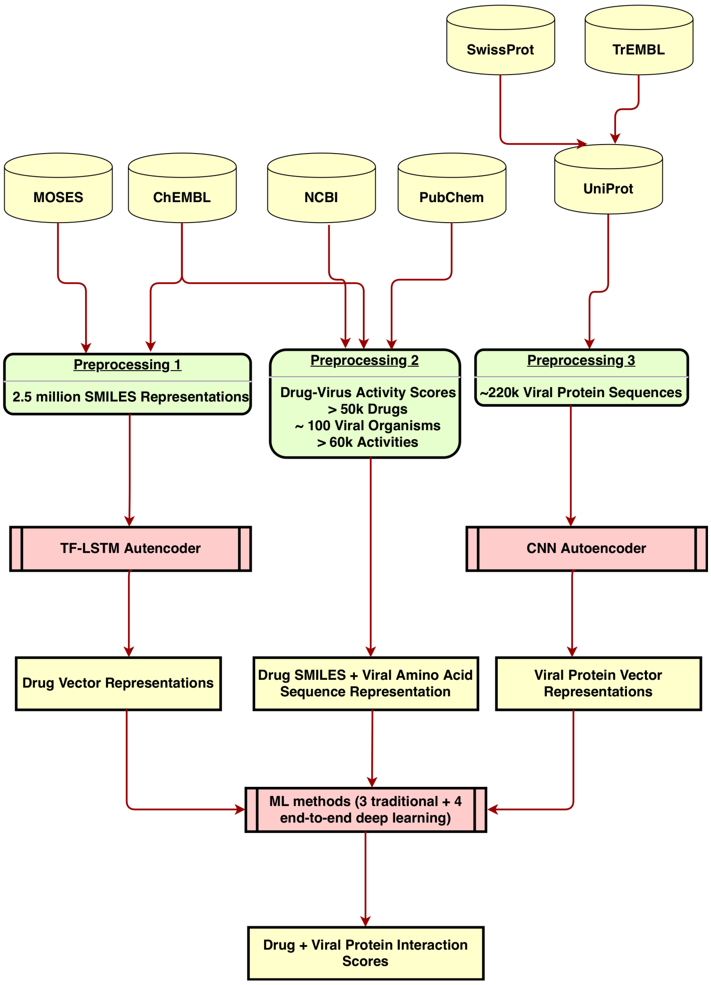

SwissProt
TrEMBL
MOSES
ChEMBL
NCBI
PubChem
UniProt
Preprocessing 1
2.5 million SMILES Representations
Preprocessing 2
Drug-Virus Activity Scores
> 50k Drugs
~ 100 Viral Organisms
> 60k Activities
Preprocessing 3
~220k Viral Protein Sequences
TF-LSTM Autencoder
CNN Autoencoder
Drug Vector Representations
Drug SMILES + Viral Amino Acid Sequence Representation
Viral Protein Vector Representations
ML methods (3 traditional + 4 end-to-end deep learning)
Drug + Viral Protein Interaction Scores

Figure 1: Flowchart of our proposed drug-viral protein activity predictor.

[34, 128] characters resulting in a final set $\mathcal{S}$ of 2,459,695 canonical SMILES for small molecules.

To train the majority of ML algorithms, it is essential to have numeric vector representation for drugs. We used the set $\mathcal{S}$ to train a teacher forcing long short term memory neural network (TF-LSTM) (Gers *et al.*, 1999; Lamb *et al.*, 2016) based autoencoder (Kramer, 1991) which generates a low dimensional vector representation (LS_d) for each drug. A detailed description of the TF-LSTM model is provided in the Methods section.

2.1.2 Viral Proteins:

We downloaded all the viral protein sequences available in UniProt (uni, 2017) comprising a total of 2,684,733 protein sequences. Among these 10,685 are deposited in SwissProt (Boeckmann *et al.*, 2003) i.e. are manually curated and functionally annotated, whereas the remaining 2,674,048 are obtained from TrEBML (Boeckmann *et al.*, 2003) and are not well-curated. These viral proteins span over 2,742 viral organisms. We then perform two preprocessing steps as utilized in (Rawi *et al.*, 2018; Khurana *et al.*, 2018) to avoid any unwanted bias and to ensure heterogeneity of sequences within the dataset. Similar to previous works (Smialowski *et al.*, 2007; Elbasir *et al.*, 2019), we first used CD-HIT (Li and Godzik, 2006; Fu *et al.*, 2012) method to decrease sequence redundancy within the dataset with a maximum sequence identity of 90%. This resulted in a reduced set of

214,915 primary viral protein sequences. We finally removed all protein sequences of length $L¿2,000$ resulting in a final set $\mathcal{V}$ of 212,057 viral protein sequences.

In order to train most ML methods, it is essential to have numeric vector representation for viral protein sequences. We utilized the set $\mathcal{V}$ to train a convolutional neural network (CNN) (LeCun *et al.*, 1995) based autoencoder which generates a low dimensional vector representation (LS_v) for each viral protein sequence. A detailed description of the CNN autoencoder (CANN) is provided in the Methods section.

2.1.3 Drug-Viral Protein Activities:

The primary focus of our work is the 3 main proteases of the SARS-COV-2 virus including papain-like proteinase (PL-PRO), 3C-like proteinase (3CL-PRO also referred as cleavage protein) and the Spike glycoprotein (S glycoprotein). We centered our work on these SAR-COV-2 proteins due to the following reasons: a) availability of high-quality 3d structures deposited in protein data bank (Bank, 1971) (PDB Ids: **6W02**, **5R7Y**, **6M0J** respectively); b) for several other viral organisms, the PL-PRO and 3CL-PRO are the main proteases that have been targeted by drugs; c) It has been shown (Lan *et al.*, 2020), that spike protein attaches the virion to the cell membrane by interacting with host receptor, initiating the infection. It binds to human ACE2 and CLEC4M/DC-SIGNR receptors and the internalization of the virus into the endosomes of the host cell induces conformational changes in the S glycoprotein.

As the SARS-COV-2 is a new virus, it is harder to get quality data about drug-viral protein activity. However, information about similar viruses, their main proteases and small molecules used to target these viral proteins are available in repositories such as NCBI, PubChem, ChEMBL, BindingDB (Liu *et al.*, 2007). We initially searched for compound activity information related to SARS-COV-1 (SARS-1), Middle East Respiratory Syndrome (MERS), Human Immunodeficiency Virus (HIV) and Hepacivirus C (HepC) using the "PUG-REST" API of NCBI (Wheeler *et al.*, 2007) which was used to download raw information from various NCBI records. We further processed only those records which contain Assay Id's (AID). A given assay can report different kinds of compound bioactivities depending on the objective of the study. These bioactivities include measurements such as IC_{50}, EC_{50}, AC_{50}, K_i, K_d, Potency etc. as described in (Haas *et al.*, 2017). These biological activities are standard potency measures that are derived from dose-response assays at different concentrations designed to measure activation, inhibition of targets, and pathways of pharmacological significance (Haas *et al.*, 2017) for a drug. We further filter those records which don't contain a PubChem standard value for activity (as otherwise, it makes it difficult to have an unbiased comparison of compound activities).

In this work, we pivoted on IC_{50} value as done by (Ullah *et al.*, 2017), which is based on the concentration of a compound at which 50% inhibition of a viral protein is observed. The PubChem standard value for IC_{50} is reported in micromolar (10^{-3}) concentration (Kim *et al.*, 2016). Furthermore, it is known from enzyme kinetics (Cheng-Prusoff Equation (Beck *et al.*, 2017)) that when a ligand (drug) binds to a protein in an uncompetitive scenario i.e. an assay, the K_i value determined is equal to IC_{50} value. Thus, we can augment our dataset with records containing K_i values. We then removed records where drugs contain salt and those whose SMILES string exceeds 128 characters, resulting in an interaction set of 13,763 drug-viral protein activities.

We next downloaded all drugs and viral protein interaction information available in ChEMBL (Gaulton *et al.*, 2017) repository. As a part of internal quality checks provided by ChEMBL, we include only those drug-viral protein interactions which have a confidence score of at least 5. The confidence score value reflects both the type of target assigned to a particular assay and the confidence that the target assigned is the correct target for that assay. As stated in (Gaulton *et al.*, 2017), assays assigned a non-molecular target type, e.g. a cell-line or an organism, receive a confidence score of 1, while assays with assigned protein targets receive a confidence score of at least 5. Moreover, we remove those activities for which a standard pChEMBL value is not available. The myriad published activities from heterogeneous resources utilized by ChEMBL are converted into a standardized activity, namely, the pChEMBL value. This value allows us to compare different measures of half-maximal response (concentration/potency/affinity) on a negative logarithmic scale. For instance, an IC_{50} value of 1 nano-molar (nM) would have a pChEMBL value of 9. The standard pChEMBL value is associated with standard PubChem value through a simple mathematical formulae ($\mathrm{pChEMBL} = -\log_{10}(\mathrm{PubChem} \times 10^{-3}) + 6$).

We initially obtain a set of 92,638 such drug-viral protein activities and after filtering for only those records which contain IC_{50}, K_i, and Potency as standard types, we limit the set to 62,219 interactions. We then remove records where the compounds contain salt and their corresponding SMILES string exceeds 128 characters. We truncated viral protein sequences to have a maximal length L=2000 amino acids in the interaction set. This results in a final set of 54,756 drug-viral protein interactions obtained and curated via ChEMBL.

We finally perform a union of the set of drug-viral protein activities obtained via PubChem and ChEMBL, resulting in the dataset $\mathcal{D}$ consisting of 60,195 such interactions. These interactions comprise 54,617 unique drugs, 153 unique viral protein sequences (based on Uniprot accession ids), and span over 97 different viral organisms. We randomly split the dataset $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ (54,175 interactions) and $\mathcal{D}_{\text{test}}$ (6,020 activities) in the ratio of 0.9 : 0.1, which are used as the training and independent test set respectively for the purpose of predictive modelling.

3 Methods

3.1 Overview

Drug-viral protein activity prediction can be modeled as a regression task. We learn a mapping function g that takes as input a joint drug and viral protein representation, (x_d, x_v) and outputs the activity score y_{dv}. Here y_{dv} corresponds to the -$\log_{10}(\mathrm{IC}_{50})$ and is used as standardized pChEMBL activity score. If ℓ is the model-specific loss function, then the regression task reduces to estimating the parameters w which minimizes

$$\min_{w} \sum_{d,v} \ell(y_{dv}, g(x_d, x_v; w))$$

In this work, the mapping function g is a ML method including Random Forests (Breiman, 2001), XGBoost (Chen

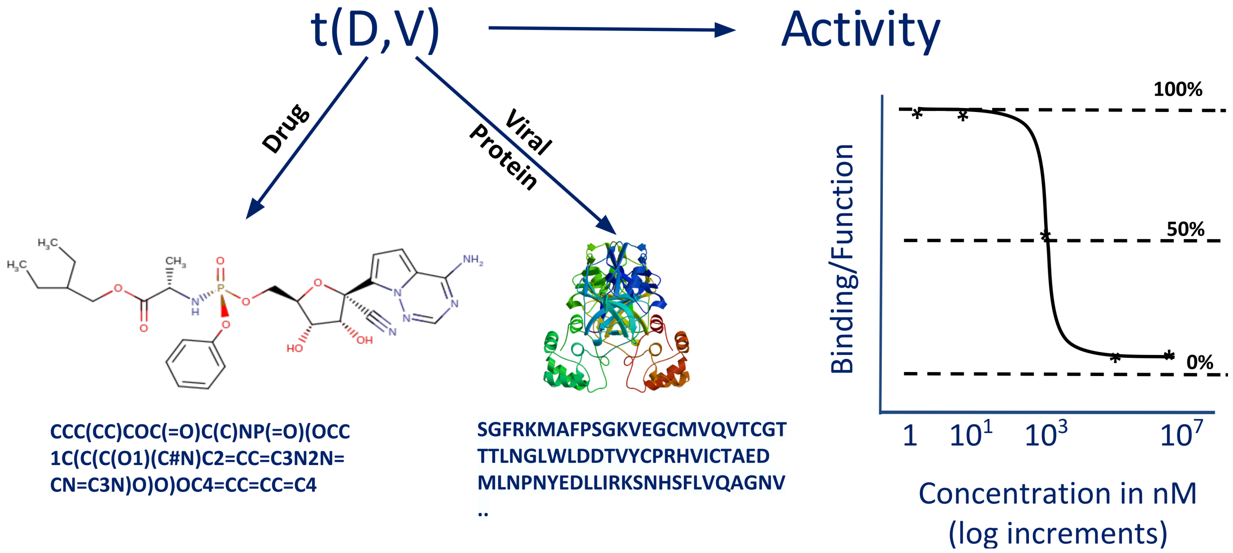


Figure 2: **Overview figure depicting our predictive modelling process.** For each drug d and each viral protein v, we use representations x_d and x_v based on SMILES strings and primary structure respectively. For each drug-viral protein interaction, the activity value used in the training set is obtained from resources such as PubChem, ChEMBL and NCBI. We used the -$\log_{10}(\text{IC}_{50})$ value measured in nano-molar units i.e. -$\log_{10}(10^3 \times 10^{-9})$=6 as the standardized pChEMBL activity score (y_{dv}) in Figure 2.

and Guestrin, 2016), Support Vector Machines (Suykens and Vandewalle, 1999; Mall and Suykens, 2015) and ℓ is the squared loss function.

For these techniques, x_d and x_v are passed to a TF-LSTM (Gers *et al.*, 1999) and a CANN (LeCun *et al.*, 1995) respectively to generate numeric vector representations LS_d and LS_v which are utilized by the aforementioned ML models to estimate activity scores, such that $\hat{y}_{dv} = g(\text{LS}_d, \text{LS}_v; w)$.

Furthermore, we also considered end-to-end deep learning models using CNN, LSTM, CNN-LSTM and Graph Attention Network (GAT)-CNN as function g, where x_d corresponds to SMILES representation for drugs and x_v reflects the primary structure or linear chain of amino acids for viral protein sequences and $\hat{y}_{dv} = g(x_d, x_v; w)$. The SMILES representation, is parameterized by a sequence of vectors, $x_d = \{x_{d,1}, x_{d,2}, \ldots, x_{d,l}\}$, where $x_{d,i}$ is a one-hot coded vector (Harris and Harris, 2010) i.e. a binary vector of length 51 (51 unique characters appears in SMILES of all possible drugs) with 1 bit active for i^{th} character in the SMILES string and $l = 128$. Similarly, for each viral protein sequence, $x_v = \{x_{v,1}, x_{v,2}, \ldots, x_{v,L}\}$, where $x_{v,j}$ is a one-hot coded vector of length 22 (20 for amino acids, 1 for gap and 1 for ambiguous amino acids) and L=2000. Figure 2 provides an overview of our modeling process.

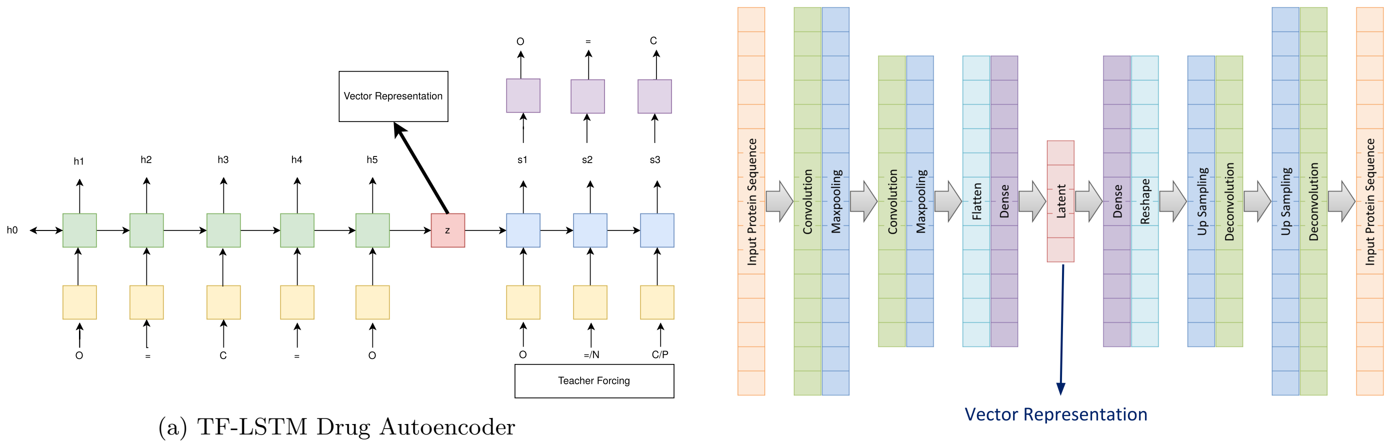
3.2 Drug Autoencoder: TF-LSTM

The goal of a drug autoencoder model (Kramer, 1991) is to learn the innate low dimensional representation LS_d from SMILES strings of drugs (x_d) in an unsupervised setting such that compounds with similar patterns tend to be closer in the low dimensional space. Our drug autoencoder framework consists of an encoder, a decoder, and a sequence to sequence (seq2seq) model which encapsulates the encoder and decoder and provides a way to interface with each. The encoder consists of a multi-layered LSTM (Gers *et al.*, 1999) which overcomes limitations like vanishing gradients experienced by a traditional recurrent neural network (RNN) models (Dupond, 2019). The output of LSTM encoder can be represented as $(h, c) = \text{EncoderLSTM}(e(x_d))$. Here $e(x_d)$ represents the embedding representation for drug, h and c correspond to hidden state representations encapsulating sequential information.

The decoder component does a single step of decoding i.e. it outputs single $(\hat{y}^t_{dv})$ token per time-step t. Since, we are building a drug autoencoder model, $\hat{y}^t_{dv} = x^t_d$ i.e. the vector corresponding to the t^{th} character in the drug representation x_d. The decoder can mathematically be depicted as $s^t = \text{DecoderLSTM}(x^t_d, (h, c))$. The hidden state s^t obtained from DecoderLSTM is passed through a linear layer f to make a prediction for the next token in the target sequence i.e. $\hat{y}^{t+1} = f(s^t)$.

Our seq2seq method takes the source drug representation (x_d), target drug representation (x_d) and a teacher-forcing ratio. The teacher forcing ratio is used when training our model. When decoding, at each time-step we predict what the next token in the target sequence will be from the previous tokens decoded, $\hat{y}^{t+1}_{dv} = f(s^t)$. With probability 1 - teacher forcing ratio, we will use the token that the model predicted as the next input to the model, even if it doesn't match the actual next token in the sequence. The latent space representation LS_d for a given drug is equivalent to the hidden state representation h for our TF-LSTM model.

We trained this TF-LSTM model on ≈ 2.5 million SMILES strings for small molecules. During the training phase, the teacher forcing ratio is set to 0.5 and during the test phase of our experiments, it is set to 0. Interestingly, 96.7% of the SMILES generated by our TF-LSTM model were valid small molecules (tested using RDKit (Landrum, 2013) package) and had a mean categorical cross-entropy (Goodfellow *et al.*, 2016) error of 0.001. The convergence of the reconstruction error for our TF-LSTM model is depicted in Supplementary Figure 1a. Figure 3a illustrates our TF-LSTM drug autoencoder model.



(a) TF-LSTM Drug Autoencoder

(b) Protein Convoluational Autoencoder Neural Network

Figure 3: Drug and Protein autoencoder models designed to generate numeric vector representations LS_d and LS_v from drugs and viral proteins.

3.3 Protein Autoencoder: CNN

The goal of the viral protein autoencoder model is to learn a low dimensional representation LS_v from the amino acid sequences of viral proteins x_v. We used a convolutional autoencoder neural network for this purpose. Our protein autoencoder framework consists of two main components: an encoder and a decoder as shown in Figure 3b. The autoencoder was trained in an unsupervised fashion to learn a low dimensional space (LS_v).

The encoder consists of multi-layered convolution and subsampling layers followed by a fully connected layer. The purpose of using the convolution layers is to extract features that preserve input neighborhood interactions and spatial locality, which is important to capture local protein structures and frequently occurring k-mers. The max-pooling layers are used for subsampling to obtain translation-invariant representations and reduce the number of convolution filters required resulting in a lesser number of trainable parameters. The max-pooling layers also perform regularization and help to generalize the learned latent space. The decoder consists of multi-layered deconvolution and upsampling layers preceded by a fully connected layer. These layers perform the inverse function of the encoder layers in the reverse order to generate the initial input.

We trained our autoencoder on $212,057$ viral proteins. The mean categorical cross-entropy (Goodfellow *et al.*, 2016) error for the autoencoder was 0.1. The convergence of the reconstruction error for the autoencoder is depicted in Supplementray Figure 1b.

3.4 Traditional Machine Learning Models

We used three state-of-the-art ML models, namely, Random Forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016) and Support Vector Machines (SVM) (Suykens and Vandewalle, 1999; Mall and Suykens, 2015) as mapping function g. Thus, our predicted activity score can be represented as $\hat{y}_{dv} = g(\mathrm{LS}_d, \mathrm{LS}_v; w)$ for a given drug d and viral protein v. It has been shown that Random Forests, XGBoost and SVMs can be used efficiently for a variety of bioinformatics problems (Mall *et al.*, 2017; Rawi *et al.*, 2018; Mall *et al.*, 2018; Ullah *et al.*, 2018; Palotti *et al.*, 2019; Elbasir *et al.*, 2020).

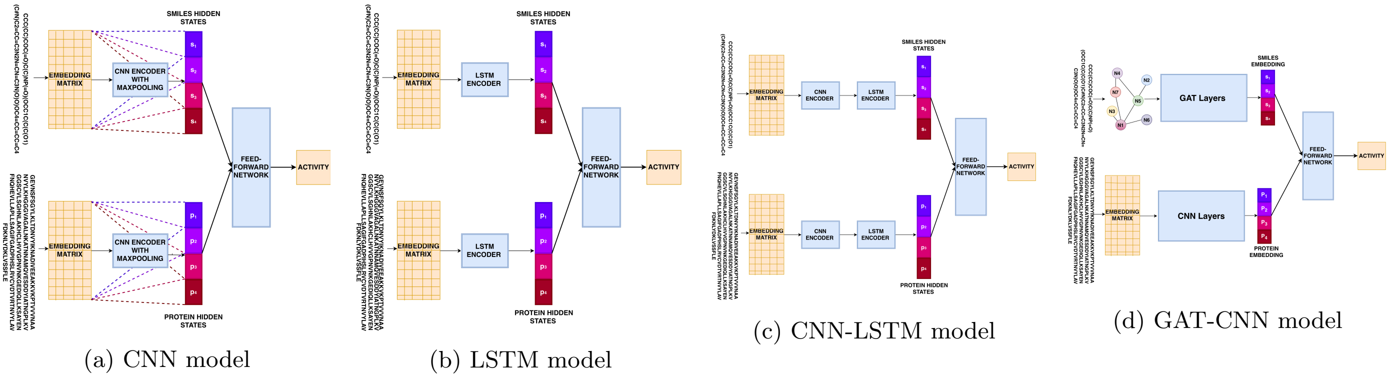
Random Forests (RF) belong to the class of ensemble supervised learning techniques. RF algorithm applies the technique of bagging or bootstrapped aggregating (Breiman, 2001) to decision tree learners. Our training dataset is depicted as $\mathcal{D}_{\text{train}} = \{(\mathrm{LS}_d^i, \mathrm{LS}_v^i), y_{dv}^i\}$, where $d \in \mathcal{S}, v \in \mathcal{V}$ and $i = 1, \ldots, N$. Here $y_{dv}^i \in \mathbb{R}$ and N is the total number of drug-viral interactions in the training set. Given $\mathcal{D}_{\text{train}}$, the bagging procedure repeatedly selects random samples with replacement and fits separate trees to these samples.

Gradient boosting machine (GBM) (Friedman, 2001) belongs to that family of predictive methods that uses an iterative strategy s.t. the learning framework will consecutively fit new models to have an accurate estimate of the response variable after each iteration. The notion behind this technique is to construct new tree-based learners to be as correlated as possible with the negative gradient of a given loss function (ℓ), calculated using all the training data $\mathcal{D}_{\text{train}}$. The advantage of the boosting procedure is that it works by decreasing the bias of the model, without increasing the variance. A more scalable and accurate version of GBM is XGBoost (Chen and Guestrin, 2016). XGBoost is based on the principle of tree boosting. It uses a scalable end-to-end tree boosting system with a weighted quantile sketch for approximate tree learning. More importantly, XGBoost can scale for a large number of samples using very little computational resources and achieve state-of-the-art predictive performance.

Support vector machines (SVM) were originally introduced in (Drucker *et al.*, 1997; Suykens and Vandewalle, 1999) and belong to the family of non-linear optimization technique used as a function estimator (regression) by constructing non-linear hyperplanes. A crucial step in building SVM models is the choice of the non-linear kernel function that encodes the similarity structure in the input data. In this work, we use the radial-basis function (RBF) kernel which is optimized using a standard cross-validation procedure. We used the 'sklearn' package (Pedregosa *et al.*, 2011) available in python (`https://www.python.org`) for building our optimal RF, XGBoost and SVM models after performing hyper-parameter optimization using 5-fold cross-validation.

3.5 End-to-End Deep Learning Models

We built 4 end-to-end deep learning models for our regression problem where the mapping functions g were CNN, LSTM, CNN-LSTM, and GAT-CNN. These models directly work on



(a) CNN model

(b) LSTM model

(c) CNN-LSTM model

(d) GAT-CNN model

Figure 4: Different end-to-end deep learning models used as data-driven predictive models for the task of estimating drug-viral protein activity.

the drug (x_d) and viral protein (x_v) representations, unlike traditional ML techniques. It has been shown previously (Khurana *et al.*, 2018; Elbasir *et al.*, 2019) that end-to-end deep learning models are useful for various bioinformatics applications.

3.5.1 CNN Model:

This deep learning architecture comprises two CNN encoders. For the drug and protein CNN encoders, each of the drug (x_d) and viral protein (x_v) representation is passed through an embedding layer ($e(\cdot)$) to generate drug embedding matrix and viral protein embedding matrix respectively. A single convolutional layer with multiple filter sizes, $k \in K = \{3, 6, 9, 12\}$, is applied on top of the embedding matrix followed by a max-pooling operation to generate hidden state vector for small molecules as well as viral protein sequences as depicted in Figure 4a. The hidden state vector h_d for drugs and h_v for viral protein sequences are then concatenated together (h) and are considered as the output of the CNN encoders.

We then have multiple feed-forward layers on top of h which are ultimately connected to the output unit corresponding to the activity score. The CNN encoders can capture contiguous sequences in the SMILES representations and k-mers in viral protein sequence, whereas the feed-forward layers capture the co-occurrence of such patterns that drive the activity score to be either high or low based on our training set $\mathcal{D}_{\text{train}}$. We use non-linear activations at every layer and optimize the model architecture w.r.t. hyper-parameters such as filter sizes, learning rate, etc.

3.5.2 LSTM Model:

The LSTM model consists of two LSTM encoders. We have an LSTM encoder based on the drug representation (x_d) and another one based on the viral protein representation (x_v). The drug LSTM encoder generates the hidden state vector (h_d) while the viral protein encoder generates the hidden state vector (h_v). The two hidden vectors are then concatenated together (h) as illustrated in Figure 4b.

We again have multiple feed-forward layers on top of h which is connected to the output unit representing the activity score. The LSTM encoders not only capture short but long term dependencies as well, due to the availability of memory units, based on SMILES strings and viral protein sequences and the feed-forward layers encapsulate the co-occurrence of such patterns driving the activity score to be high or low for a given drug-viral protein combination.

3.5.3 CNN-LSTM Model:

The CNN-LSTM model is a combination of CNN and the LSTM model. Similar to the previous two models, it consists of two encoders, one for the drug representation (x_d) while the other for the viral protein representation (x_v). For each encoder, the output of the convolutional layer instead of being passed to a max-pooling layer is directly passed to the recurrent layer (LSTM) which then generates the hidden state vector representations (h_d,h_v) for the compounds and the viral protein sequences respectively.

By combining the CNN and LSTM models, this model can better capture contiguous and well as long-term dependencies in the SMILES strings and viral protein sequences. The output of each encoder is concatenated together to generate hidden representation h which is passed to multiple feed-forward layers and is ultimately connected to the output layer consisting of one unit for the activity score.

3.5.4 Graph Attention Networks-Convolutional Neural Networks (GAT-CNN) Model:

This deep learning architecture is composed of two parts, graph attention networks (Veličković *et al.*, 2017) and convolutional neural networks. For a given drug, the compound structure can be presented as a graph consisting of the atoms in the compound (as nodes) and connected by edges if a bond exists between a pair of atoms. To convert a compound structure to the form of graph representations, we use the RDKit package which takes SMILES strings and converts them to a multi-dimensional binary feature vector. Furthermore, RDKit allows us to extract different atom features such as atom's degree, the total number of hydrogen, the number of hydrogen with the number of bonded neighbors, atom status as aromatic or not, the implicit value of atoms, and atom symbol. These features can be utilized as node properties for atoms. In total, we extract 78 such features from the SMILES strings. Given the graph-based representation of a drug molecule (x_d) along with the extracted node features, the GAT model learns an embedding representation for a drug encapsulating the topological information available in the graph of each drug.

The second component of this architecture is convolutional neural networks which take protein sequence as an input.

This component is composed of the embedding layer and multiple convolutional layers. At each convolutional layer, a non-linear activation function is applied and is followed by a max-pooling operator. This component of our GAT-CNN architecture learn protein embeddings (h_v) and then concatenates it with the SMILES embedding (h_d) generated by GAT to generate h which is then passed to feed-forward layers. The output layer outputs the score corresponding to the compound activity. The weights associated with GAT and CNN models are learned jointly through backpropagation procedure during the training phase while training on drug-viral protein samples.

The optimal model architecture hyper-parameters (like $h_d = 256$, $h_v = 64$) for each of the end-to-end deep learning models are provided in Supplementary Table 1.

4 RESULTS

4.1 Experimental Results on $\mathcal{D}_{\text{test}}$

We perform 10 randomizations for each of our predictive models by randomly splitting the full dataset $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ in proportions (0.9:0.1) for training and testing purposes respectively as mentioned earlier in the Materials section. Table 1 provides a comprehensive comparison of the mapping functions g utilized in our work including RF, SVM, XGBoost, CNN, LSTM, CNN-LSTM and GAT-CNN models These models are evaluated over 4 quality metrics, namely, mean absolute error (MAE), mean squared error (MSE), pearson correlation R (Pearson R) and the coefficient of determinination (R2), where each of these metrics are estimated using the predicted pChEMBL values vs the groundtruth pChEMBL values for drug-viral protein interactions ($\mathcal{D}_{\text{test}}$). For metrics, MAE and MSE, the lower the value and closer to 0, the better the predictive performance of the model, whereas for metrics, Pearson R and R2, the higher and closer the value to 1, the better the efficiency of the predictive model.

Model	**MAE**	**MSE**	**Pearson R**	**R2**
RF	0.630 ± 0.004	0.855 ± 0.015	0.739 ± 0.006	0.546 ± 0.003
SVM	0.596 ± 0.005	0.779 ± 0.016	0.767 ± 0.003	0.588 ± 0.005
XGBoost$^+$	0.567 ± 0.002	0.753 ± 0.007	0.775 ± 0.003	0.599 ± 0.005
CNN	0.587 ± 0.005	0.826 ± 0.017	0.758 ± 0.005	0.575 ± 0.008
LSTM	0.597 ± 0.003	0.809 ± 0.008	0.756 ± 0.001	0.571 ± 0.002
CNN-LSTM	0.646 ± 0.004	1.005 ± 0.013	0.700 ± 0.003	0.490 ± 0.004
GAT-CNN*	0.576 ± 0.005	0.761 ± 0.015	0.772 ± 0.004	0.597 ± 0.006

Table 1: Comparison of performance of devised ML techniques for our drug-viral activity prediction problem evaluated w.r.t. 4 metrics.

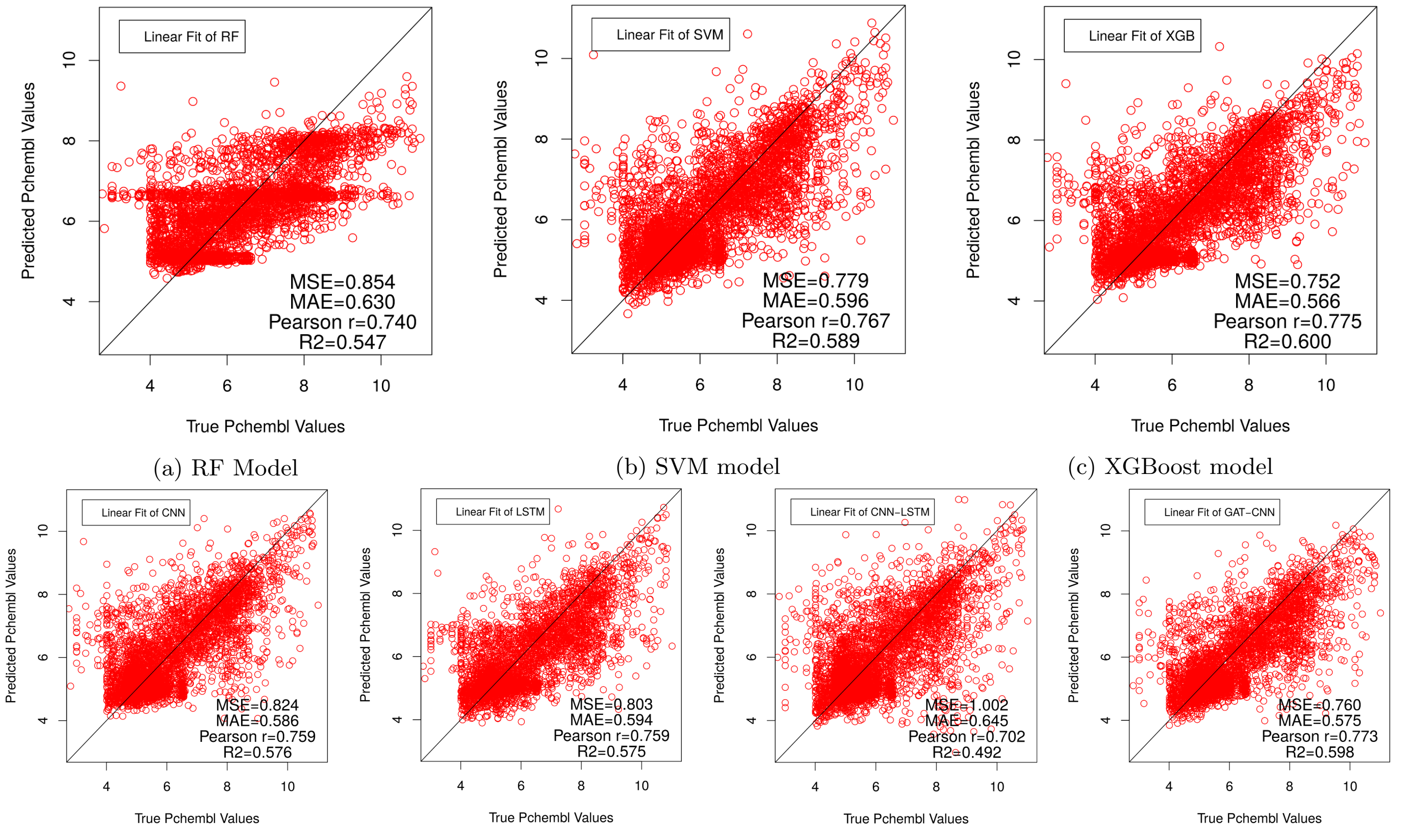
From Table 1, we observe that the best predictive model w.r.t. all quality metrics is the XGBoost model, highlighted in Table 1 by +, and is built on the numeric vector representation of drugs and viral protein sequences ($g(\text{LS}_d, \text{LS}_v; w)$) respectively. The XGBoost model achieves quality performance when compared to an ideal model for which the Pearson R and R2 would be 1. It is closely followed by the end-to-end deep learning model ($g(x_d, x_v; w)$) based on graph attention networks on the drugs and the convolutional neural networks on viral protein sequence model (GAT-CNN, depicted in Table 1 by $*$). Table 1 showcases that predictive performance of all our designed ML models are comparable, suggesting that these models can be used in an ensemble framework to rank drugs which can have the highest activity against a given viral protein of interest. It is noteworthy, that the standard deviations of each of our predictive model w.r.t. the 4 evaluation metrics are low, indicating low variance and high efficiency in the generalization performance of our proposed models.

Next, we evaluate the predictive performance of the best model obtained from the 10 randomizations for each mapping function g. The predictive capability of each of these models is highlighted in Figure 5. In Figure 5, the x-axis represents the true pChEMBL values for drug-viral protein activities available in $\mathcal{D}_{\text{test}}$ and the y-axis represents the predicted pChEMBL value by individual ML model. Ideally, we want to minimize the difference between predicted pChEMBL value and the true pChEMBL value i.e. the predictions should be aligned along the diagonal. Moreover, the smaller the scatter of the predictions along the diagonal, the lower is the variance of the predictive model and higher is the generalization performance as depicted in Figures 5c and 5g. Furthermore, we are interested in accurately estimating larger pChEMBL values as they suggest a higher activity for drug-viral protein combinations, thereby suggesting potential inhibition. We observe from Figure 5, that traditional ML models such SVM (see Figure 5b) and XGBoost (see 5c) as well as end-to-end deep learning models can efficiently estimate such pChEMBL values. For true pChEMBL values greater than 8 and predicted pChEMBL values also greater than 8, these models achieve MAE of 0.487, 0.545, 0.472, 0.562, 0.527, 0.645 and Pearson R of 0.647, 0.668, 0.682, 0.554, 0.598 and 0.637 respectively. However, the RF method fails to accurately estimate higher pChEMBL values (identify potential inhibitors efficiently), MAE of 0.945, and Pearson R of 0.525, as illustrated in Figure 5a.

We additionally compared the predictive performance of these models w.r.t. the ground-truth drug-viral protein interactions available in the test set $\mathcal{D}_{\text{test}}$ as illustrated in Figure 6. It can be observed from Figure 6 that the x-axis represents the sample id in $\mathcal{D}_{\text{test}}$, whereas, for each such sample, we have 8 values vertically spread along the y-axis. One of these values is the ground truth pChEMBL value, while the others are predicted interaction scores by our data-driven models. The closer the predicted scores are to the true pChEMBL value, the smaller is the error in our predictions. The spline fitted on the pChEMBL values for the ground truth (labeled as 'True' in Figure 6) activities correlate well with the splines fitted on the predicted pChEMBL values for each of our proposed models. The convergence of loss function for the best deep learning drug-viral protein activity prediction models are highlighted in Supplementary Figure 1c.

4.2 Experimental Results for COVID-19 Use Case

For COVID-19, we utilized proposed ML models in an ensemble framework to identify FDA approved drugs which can be most potent against its viral proteins. We focused on the 3 main viral proteases of SARS-COV-2 virus including the PL-pro, 3CL-pro and Spike protein whose primary structure is depicted in Table 2. We initally used a set of 117 FDA approved drugs which are in some stage of clinical trial for any known viral organism as indicated in (Andersen *et al.*, 2020). However, after filtering for large molecules ($l > 128$), we end up with a set $\mathcal{S}$ comprising 101 compounds (SMILES strings)



(a) RF Model (b) SVM model (c) XGBoost model

(d) CNN model (e) LSTM model (f) CNN-LSTM model (g) GAT-CNN model

Figure 5: Comparison of predictive performance of the optimal version of each ML model for 4 evaluation metrics on the test set $\mathcal{D}_{\text{test}}$. The best performance w.r.t. R2 is observed for the XGBoost model. However, the GAT-CNN deep learning model's performance is highly competitive and second best to the XGBoost model.

including known antivirals, antibiotics, antifungal and anticancer drugs (see Supplementary Table 2).

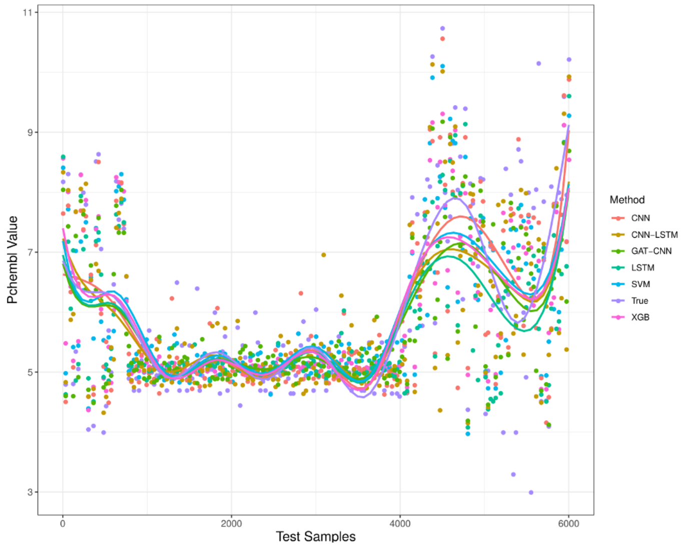


Figure 6: Comparison of predictions made by our proposed data-driven models w.r.t. the underlying groundtruth activity scores available in $\mathcal{D}_{\text{test}}$.

Uniprot Id	PDB Id	Protein Fragment	Sequence	L
P0DTD1	6W02	PL-PRO (NSP3)	GEVNS. . .SSFLE	170
P0DTD1	5R7Y	3CL-PRO	SGFRK. . .GVTFQ	306
P0DTC2	6MOJ	Spike Protein	TNLCPF. . .ATVCG	229

Table 2: Main proteases of SARS-COV-2 virus targeted for inhibition by our data-driven drug repurposing approach. Here . . . is used to save space.

We first obtain the predicted pChEMBL values for all our proposed ML methods corresponding to the set of drugs ($\mathcal{S}$) for each of the three main proteases of the SARS-COV-2 virus. After obtaining the predicted pChEMBL values, we average the predictions obtained from SVM, XGBoost, CNN, LSTM, CNN-LSTM and GAT-CNN models (top-performing models w.r.t. accurately estimating high pChEMBL values) to get a ranked list of drugs ordered by decreasing pChEMBL values for each of the 3 viral proteins as depicted in Table 3. By taking an average of the predictions, we allow our ranked list to be influenced by each of the ML models and the top-ranked predicted drugs should have high activity score for a majority of the 6 ML models used in the ensemble framework.

From Table 3, we observe that the list of top-ranked drugs includes 7 antivirals, 3 antibiotics, 6 anticancer, 2 antimalarial and 1 antifungal compound respectively. The power of drug repurposing is reflected in these results as several of the drugs identified are originally meant for a different disease or were designed for different functionality (chemotherapeutic agent, malarial drug, etc.) but can have potential antiviral capabilities. It is noteworthy, that majority (10 out of 19) of these drugs are commonly appearing in the top 15 ranked list of drugs and hence predicted to be effective against all the three main proteases of SARS-COV-2 virus. Similarly, 6 out of the remaining 9 drugs in Table 3, have strong activity scores for

Drug Name	PL-Pro	3CL-Pro	Spike Protein
Valaciclovir$^+$	6.342^1	6.305^1	6.328^3
Remdesivir$^+$	6.101^2	6.129^4	6.253^{11}
Nelfinavir$^+$	6.087^3	6.129^3	6.231^{13}
Regorafenib*	6.064^4	6.048^8	6.190^{14}
Trametinib*	6.058^5	6.153^2	6.473^1
Lopinavir$^+$	6.014^6	6.008^{10}	6.301^7
Mitoxantrone$^-$	6.009^7	6.100^5	6.405^2
Sorafenib*	5.989^8	5.994^{11}	6.100^{20}
Mefloquine$^\#$	5.979^9	6.021^9	6.099^{21}
Monensin$^-$	5.956^{10}	5.954^{15}	6.259^{10}
Topotecan*	5.929^{11}	6.066^6	6.302^6
Hydroxychloroquine$^\#$	5.928^{12}	6.057^7	6.098^{22}
Lobucavir$^+$	5.923^{13}	5.898^{18}	5.879^{41}
Bortezomib*	5.902^{14}	5.930^{16}	6.319^4
Posaconazole$^@$	5.893^{15}	5.962^{14}	6.307^5
Tilorone$^+$	5.851^{17}	5.964^{12}	6.250^{12}
Ritonavir$^+$	5.803^{23}	5.963^{13}	6.294^8
Salinomycin$^-$	5.830^{18}	5.853^{21}	6.270^9
Raloxifene*	5.819^{21}	5.929^{17}	6.182^{15}

Table 3: Top ranked 15 drugs for each of PL-Pro, 3CL-Pro and Spike proteins of SARS-COV-2 virus ordered by PL-Pro, 3CL-Pro and Spike protein respectively. The values represent the average predicted pChEMBL score by ensemble of our 6 proposed ML models. Here $+$, $-$, $*$, $\#$ and @ correspond to antiviral, antibiotics, anticancer, antimalarial and antifungal drugs respectively. The superscript for each predicted pChEMBL score reflects the ranking of the drug for that particular viral protein based on the list of ranked drugs.

2 out of the 3 main proteases of SARS-COV-2. This suggests that the drugs identified by the ensemble of our data-driven models can be universally effective against the SARS-COV-2 virus.

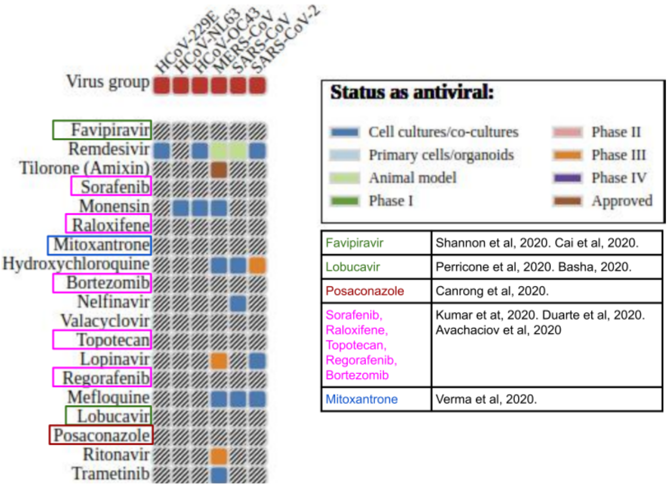


Figure 7: Validation of the top-ranked drugs identified by our data-driven approach. Several of the drugs identified by our framework are in some stage of clinical trials and several others (highlighted in boxes) have been identified as potential drugs for COVID-19 in recent literature.

From Figure 7, we observe that drugs such as Remdesivir, Monensin, Tilorone, Hydroxychloroquine, Nelfinavir, Lopinavir, Mefloquine, Ritonavir, Trametinib, identified as potential candidate drugs against SARS-COV-2 viral proteins by our data-driven framework, are in some stage of a clinical trial for one or more coronavirus based disease as reported in (Andersen *et al.*, 2020). Moreover, antivirals Favipiravir and Lobucavir have been reported in (Shannon *et al.*, 2020; Cai *et al.*, 2020) and (Perricone *et al.*, 2020; Basha, 2020) respectively as potential drugs for SARS-COV-2 virus as depicted in Figure 7. While the authors in (Shannon *et al.*, 2020), showcase that Favipiravir can be used to target the RNA polymerase process induced by the virus to produce proteins by functioning as polymerase inhibitor, the author in (Basha, 2020) identified Lobucavir as a potential drug using a computer-aided drug screening mechanism. Moreover, in (Wu *et al.*, 2020), a structure-based virtual ligand screening was performed i.e. molecular docking was utilized to determine that antifungal drug, Posaconazole, had a high binding affinity for the Spike protein (ranked 5th in our list for the Spike protein).

Additionally, we observe a myriad number of anticancer drugs such as Sorafenib, Regorafenib, Raloxifene, Topotecan have been identified as potential treatments for COVID-19 as indicated in (Duarte *et al.*, 2020; Arul *et al.*, 2020) and illustrated in Figure 7. By utilizing a data-driven approach, combining connectivity map and transcriptomic signature of lung carcinoma cells infected with the SARS-COV-2 virus, the authors in (Duarte *et al.*, 2020) identify anticancer drugs such as Raloxifene and Topotecan as potential therapeutic solutions. Similarly, in (Arul *et al.*, 2020), it was observed that anticancer drugs Regorafenib and Sorafenib have a high binding affinity score for the Spike protein through molecular docking experiments. In another experiment in (Verma *et al.*, 2020), Mitoxantrone was determined to be one of the most effective (highest binding affinity score) drugs against the 3CL-pro viral protein. A series of stabilizing interactions were observed between the drug and viral protease active sites in the case of Mitoxantrone, leading to a high binding affinity score.

Finally, in (Macchiagodena *et al.*, 2020), it was shown that the N3 ligand acts as a covalent inhibitor of 3CL-pro and has a strong binding affinity with the lowest energy minimum of -7.9 Kcal/mol, thereby inhibiting the 3CL-pro viral protease. We performed molecular docking for each of the 19 top-ranked drugs highlighted in Table 3 with the 3CL-pro protein utilizing the Rosalind online tool (`http://covid19.glamorous.ai/`). Rosalind uses Gypsum-dl (Ropp *et al.*, 2019) for compound preparation and Autodock vina (Trott and Olson, 2010) for docking. By default, it runs processes with the following parameters: exhaustiveness set to 8, the number of binding modes corresponds to 9, and a random seed.

From Table 4, we observe that majority of the 7 antiviral drugs estimated by our data-driven approach (including Nelfinavir, Remdesivir, and Lopinavir) have low binding energy which is comparable to the one obtained for N3 ligand and thus exhibit strong binding affinity against the 3CL-pro viral protease. Interestingly, all the 6 anticancer drugs and 3 antibiotics attain low binding energies, thereby exhibiting the potential to inhibit this viral protease. Drugs such as Trametinib and Monensin have shown to be effective inhibitors against multiple coronaviruses as illustrated in Figure 7 and in (Li and De Clercq, 2020) and (Pillaiyar *et al.*, 2020) respectively. It is also noteworthy, that even though our predictive framework identifies Hydroxychloroquine as a potential target (in the top 10 ranked drug list for 3CL-pro as indicated in Table 3) when performing molecular docking against 3CL-pro viral protein, it achieves a relatively high binding energy score (-6.472 Kcal/mol) suggesting weak binding affinity and hence lack of effectiveness as an inhibitor against the virus. This is complemented by the recent clinical trial (Geleris *et al.*,

Category	Antiviral	Anticancer	Antibiotics	Antimalarial	Antifungal
Drug Name	Nelfinavir, Remdesivir, Lopinavir, Valaciclovir, Lobucavir, Tilorone	Trametinib, Topotecan, Regorafenib, Raloxifene, Sorafenib, Bortezomib	Monensin, Mitoxantrone, Salinomycin	Mefloquine, Hydroxychloroquine	Posaconazole
Binding Energy	-8.386, -7.989, -7.872, -7.477, -7.389, -7.185, -6.291	-9.439, -8.753, -8.661, -8.480, -8.349, -8.170	-8.919, -8.349, -7.922	-8.274, -6.472	-8.308

Table 4: Comparison of the binding energy of the top-ranked drugs against 3CL-pro main viral protease estimated using Rosalind. The lower the binding energy the better is the binding affinity between the drug and the viral protein. Each group is ordered by increasing binding energy and a one-to-one correspondence exists between drug name and binding energy.

2020) showcasing that Hydroxychloroquine has no impact on the survival of the most severe outcomes from the COVID-19 disease.

5 DISCUSSION & CONCLUSION

In this work, we showcase that the problem of predicting activity score for drug-viral protein interactions can be formulated as a regression task. We illustrate that data-driven ML models ($g(\cdot)$) based on a simplistic representation of drugs (SMILES strings) and viral protein sequences (linear chain of amino acids) can be used efficiently for the aforementioned task. We demonstrated the effectiveness of 3 traditional ML methods: RF, SVM, and XGBoost and 4 end-to-end deep learning pipelines as mapping functions to accurately estimate these activity scores (all these techniques achieve Pearson R $>$ 0.7 on an independent test set $\mathcal{D}_{\text{test}}$). Moreover, the majority of the models (except RF) can accurately determine larger pChEMBL values ($>=$ 8.0) with a Pearson R $>$ 0.55. Since our models are based on representations of drugs (x_d) and viral proteins (x_v), we can enhance our models by using additional information such as physio-chemical properties as well as 2d images of drugs. Similarly, we can utilize supplementary information including physio-chemical and structural properties of proteins as showcased in (Khurana *et al.*, 2018; Elbasir *et al.*, 2019), to further strengthen our models in the future.

Since our predictive framework is built on $\mathcal{D}_{\text{train}}$ which contains information for over 97 different viral organisms along with their main proteases, our models are generalizable. This means that our models can efficiently produce a ranked list of potential inhibitors for the next big viral threat once the proteins associated with that viral organism are known. Moreover, it known that viruses frequently mutate (Fleischmann Jr, 1996). As a result, the viral protein will also have multiple point mutations i.e. few amino acids in the primary sequence of the viral protease can change. This can have an immense impact on the 3d structure as well as the functionality of the viral protein (Bhattacharya *et al.*, 2017). Thus, data-driven techniques identifying potential drugs based on virtual ligand screening using docking experiments (high-quality 3d structure of viral proteins) such as (Wu *et al.*, 2020; Basha, 2020; Verma *et al.*, 2020; Duarte *et al.*, 2020; Arul *et al.*, 2020) can suffer greatly in this situation. However, our models focus on the primary structure (linear chain of amino acids) and with point mutations, the vector representations LS_v and x_v will change. But since our mapping functions are generalizable (based on frequently cooccurring k-mers and subsequences in SMILES strings), we will end up with a revised ranked list of drugs as potential inhibitors for the mutated viral protein in a computationally efficient manner.

For the COVID-19 use-case, an ensemble of our data-driven models identifies a list of 19 drugs as potential inhibitors. These drugs include Remdesivir, Lopinavir, Nelfinavir, Ritonavir, Tilorone, Mefloquine, Hydroxychloroquine, Monensin, and Trametinib, which are in some stage of a clinical trial against one or more coronaviruses as depicted in Figure 7. Moreover, drugs such as Remdesivir, Lopinavir, Favipiravir, Hydroxychloroquine, and Trametinib were highlighted in a recent study (Sanders *et al.*, 2020) to potentially inhibit SARS-COV-2 virus by targeting different biological processes involved in the virus cycle. Furthermore, recent clinical trials have suggested the efficacy of Remdesivir (Beigel *et al.*, 2020) and the ineffectiveness of Hydroxychloroquine (Geleris *et al.*, 2020) for the most severe cases of COVID-19. We observe from our molecular docking experiments on the 3CL-pro viral protease of the SARS-COV-2 virus that a majority of the anticancer drugs in our ranked list exhibit high binding affinity, thereby demonstrating the true power of drug repurposing and suggesting a further investigation of the same.

Finally, a limitation of our work is that our mapping function g currently only considers the drug representation (x_d) and viral protein representation (x_v) and doesn't include any information about the host organism (x_h). Recently, in (Gordon *et al.*, 2020), the authors expressed 26 SARS-CoV-2 viral proteins in human cells and identified 332 high confidence human protein interactions with them. Based on this, they identified 69 compounds which can potentially target 66 human proteins interacting with SARS-COV-2 viral proteins. Similarly, in (Gysi *et al.*, 2020), a network-based approach is utilized to identify drug repurposing candidates. Their drug repurposing strategy relies on network proximity, diffusion, and AI-based metrics, allowing to rank all approved drugs based on their likely efficacy for COVID-19 disease leading to 81 promising candidates. Several drugs such as Ritonavir, Hydroxychloroquine, Bortezomib, Lopinavir, and Mitoxantrone appear in their set and are common to our ranked list of drugs. In future, we plan to extend our mapping function to become $g(x_d, x_v, x_h; w)$, by considering drug-viral protein interactions, drug-human protein target interactions, human protein-protein interactions, human protein-viral protein interactions in a knowledge graph representation and utilize a graph convolutional neural network (Kipf and Welling, 2016) based approach to identify potentially repurposable drugs for any viral disease. Another strand of work that we would like to explore is the use of Transformer Networks which use self-attention to capture long range dependency in sequence to sequence modeling. Recent work in natural language processing has convincingly demonstrated that Transfomer Networs are substantially more efficient than LSTMs with comparable level of accuracy Vaswani *et al.* (2017). In our particular instance, both the SMILES representation for drugs and linear chain of amino acid or primary structure of proteins can benefit from these newer approaches.

References

(2017). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, **45**(D1), D158–D169.

Andersen, P. I. *et al.* (2020). Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseases*.

Arul, M. N. *et al.* (2020). Searching for target-specific and multi-targeting organics for covid-19 in the drugbank database with a double scoring approach.

Ashburn, T. T. and Thor, K. B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, **3**(8), 673–683.

Bank, P. D. (1971). Protein data bank. *Nature New Biol*, **233**, 223.

Basha, S. H. (2020). Corona virus drugs–a brief overview of past, present and future. *Journal of PeerScientist*, **2**(2), e1000013.

Beck, B. *et al.* (2017). Assay operations for sar support. In *Assay Guidance Manual [Internet]*. Eli Lilly & Company and the National Center for Advancing Translational Sciences.

Beigel, J. H. *et al.* (2020). Remdesivir for the treatment of covid-19—preliminary report. *New England Journal of Medicine*.

Bhattacharya, R. *et al.* (2017). Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One*, **12**(3), e0171355.

Boeckmann, B. *et al.* (2003). The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, **31**(1), 365–370.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Cai, Q. *et al.* (2020). Experimental treatment with favipiravir for covid-19: an open-label control study. *Engineering*.

Chakraborti, S. and Srinivasan, N. (2020). Drug repurposing approach targeted against main protease of sars-cov-2 exploiting 'neighbourhood behaviour'in 3d protein structural space and 2d chemical space of small molecules. chemrxiv. *Preprint. https://doi. org/10.26434/chemrxiv*, **12057846**, v1.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cheng, F. (2019). In silico oncology drug repositioning and polypharmacology. In *Cancer Bioinformatics*, pages 243–261. Springer.

Cheng, F. *et al.* (2016). Drug repurposing: new treatments for zika virus infection? *Trends in molecular medicine*, **22**(11), 919–921.

Cheng, F. *et al.* (2017). Individualized network-based drug repositioning infrastructure for precision oncology in the panomics era. *Briefings in bioinformatics*, **18**(4), 682–697.

Cheng, F. *et al.* (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature communications*, **9**(1), 1–12.

Dong, E. *et al.* (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, **20**(5), 533–534.

Drucker, H. *et al.* (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Duarte, R. R. *et al.* (2020). Repurposing fda-approved drugs for covid-19 using a data-driven approach.

Dupond, S. (2019). A thorough review on the current advance of neural network structures. *Annual Reviews in Control*, **14**, 200–230.

Elbasir, A. *et al.* (2019). Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics*, **35**(13), 2216–2225.

Elbasir, A. *et al.* (2020). Bcrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics*, **36**(5), 1429–1438.

Fleischmann Jr, W. R. (1996). Viral genetics. In *Medical Microbiology. 4th edition*. University of Texas Medical Branch at Galveston.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Fu, L. *et al.* (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**(23), 3150–3152.

Gaulton, A. *et al.* (2017). The chembl database in 2017. *Nucleic acids research*, **45**(D1), D945–D954.

Geleris, J. *et al.* (2020). Observational study of hydroxychloroquine in hospitalized patients with covid-19. *New England Journal of Medicine*.

Gers, F. A. *et al.* (1999). Learning to forget: Continual prediction with lstm.

Goodfellow, I. *et al.* (2016). *Deep learning*. MIT press.

Gordon, D. E. *et al.* (2020). A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, pages 1–13.

Gupta, A. *et al.* (2018). Generative recurrent networks for de novo drug design. *Molecular informatics*, **37**(1-2), 1700111.

Gysi, D. M. *et al.* (2020). Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint arXiv:2004.07229*.

Haas, J. V. *et al.* (2017). Minimum significant ratio–a statistic to assess assay variability. In *Assay Guidance Manual [Internet]*. Eli Lilly & Company and the National Center for Advancing Translational Sciences.

Harris, D. and Harris, S. (2010). *Digital design and computer architecture*. Morgan Kaufmann.

Khurana, S. *et al.* (2018). Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, **34**(15), 2605–2613.

Kim, S. *et al.* (2016). Pubchem substance and compound databases. *Nucleic acids research*, **44**(D1), D1202–D1213.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kitchen, D. B. *et al.* (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, **3**(11), 935–949.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, **37**(2), 233–243.

Lamb, A. M. *et al.* (2016). Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609.

Lan, J. *et al.* (2020). Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, pages 1–6.

Landrum, G. (2013). Rdkit documentation. *Release*, **1**, 1–79.

LeCun, Y. *et al.* (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, **3361**(10), 1995.

Li, G. and De Clercq, E. (2020). Therapeutic options for the 2019 novel coronavirus (2019-ncov).

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Liu, T. *et al.* (2007). Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, **35**(suppl_1), D198–D201.

Macchiagodena, M. *et al.* (2020). Identification of potential binders of the main protease 3clpro of the covid-19 via structure-based ligand design and molecular modeling. *Chemical Physics Letters*, page 137489.

Mall, R. and Suykens, J. A. (2015). Very sparse lssvm reductions for large-scale data. *IEEE transactions on neural networks and learning systems*, **26**(5), 1086–1097.

Mall, R. *et al.* (2017). Detection of statistically significant network changes in complex biological networks. *BMC systems biology*, **11**(1), 32.

Mall, R. *et al.* (2018). Rgbm: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucleic acids research*, **46**(7), e39–e39.

Oprea, T. I. *et al.* (2011). Drug repurposing from an academic perspective. *Drug Discovery Today: Therapeutic Strategies*, **8**(3-4), 61–69.

Organization, W. H. (2020). Coronavirus disease (covid-2019) situation reports - 139. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200607-covid-19-sitrep-139.pdf.

Palotti, J. *et al.* (2019). Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ digital medicine*, **2**(1), 1–9.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Perricone, C. *et al.* (2020). The anti-viral facet of anti-rheumatic drugs: lessons from covid-19. *Journal of Autoimmunity*, page 102468.

Pillaiyar, T. *et al.* (2020). Recent discovery and development of inhibitors targeting coronaviruses. *Drug discovery today*, **25**(4), 668–688.

Polykovskiy, D. *et al.* (2018). Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv preprint arXiv:1811.12823*.

Pushpakom, S. *et al.* (2019). Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, **18**(1), 41–58.

Rawi, R. *et al.* (2018). Parsnip: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, **34**(7), 1092–1098.

Ropp, P. J. *et al.* (2019). Gypsum-dl: an open-source program for preparing small-molecule libraries for structure-based virtual screening. *Journal of cheminformatics*, **11**(1), 34.

Sanders, J. M. *et al.* (2020). Pharmacologic treatments for coronavirus disease 2019 (covid-19): a review. *Jama*, **323**(18), 1824–1836.

Shannon, A. *et al.* (2020). Favipiravir strikes the sars-cov-2 at its achilles heel, the rna polymerase. *bioRxiv*.

Smialowski, P. *et al.* (2007). Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, **23**(19), 2536–2542.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, **9**(3), 293–300.

Trott, O. and Olson, A. J. (2010). Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, **31**(2), 455–461.

Ullah, E. *et al.* (2017). Identification of cancer drug sensitivity biomarkers. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2322–2324. IEEE.

Ullah, E. *et al.* (2018). Harnessing qatar biobank to understand type 2 diabetes and obesity in adult qataris from the first qatar biobank project. *Journal of translational medicine*, **16**(1), 99.

Vaswani, A. *et al.* (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Veličković, P. *et al.* (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Verma, D. *et al.* (2020). Potential inhibitors of sars-cov-2 main protease (mpro) identified from the library of fda approved drugs using molecular docking studies. preprints 2020, 202004.

Wheeler, D. L. *et al.* (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, **36**(suppl_1), D13–D21.

Wishart, D. S. *et al.* (2018). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, **46**(D1), D1074–D1082.

Wu, C. *et al.* (2020). Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*.