Sequence Analysis

A Modelling Framework for Embedding-based Predictions for Compound-Viral Protein Activity

Raghvendra Mall [1*], Abdurrahman Elbasir [2,], Hossam Almeer [1], Zeyaul Islam [3], Prasanna R Kolatkar [3], Sanjay Chawla [1] & 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
[3]Qatar Biomedical Research Institute, Hamad Bin Khalifa Univeristy, Doha, 34110, Qatar

* corresponding authors.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A global effort is underway to identify compounds for the treatment of COVID-19. Since *de novo* compound design is an extremely long, time-consuming, and expensive process, efforts are underway to discover existing compounds that can be repurposed for COVID-19 and new viral diseases.
Model: We propose a machine learning representation framework that uses deep learning induced vector embeddings of compounds and viral proteins as features to predict compound-viral protein activity. The prediction model in-turn uses a consensus framework to rank approved compounds against viral proteins of interest.
Results: Our consensus framework achieves a high mean Pearson correlation of 0.916, mean R2 of 0.840 and a low mean squared error of 0.313 for the task of compound-viral protein activity prediction on an independent test set. As a use case, we identify a ranked list of 47 compounds common to three main proteins of SARS-COV-2 virus (PL-PRO, 3CL-PRO and Spike protein) as potential targets including 21 antivirals, 15 anticancer, 5 antibiotics and 6 other investigational human compounds. We perform additional molecular docking simulations to demonstrate that majority of these compounds have low binding energies and thus high binding affinity with the potential to be effective against the SARS-COV-2 virus.
Availability: All the source code and data is available at: `https://github.com/raghvendra5688/Drug-Repurposing` and `https://dx.doi.org/10.17632/8rrwnbcgmx.3`. We also implemented a web-server at: `https://machinelearning-protein.qcri.org/index.html`.
Contact: Raghvendra Mall and Ehsan Ullah
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The breakout of COVID-19 started in December 2019, in China's Hubei province [1], and to date, this pandemic has caused over 95 million infections and over 2 million deaths worldwide [2]. 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 compounds to treat COVID-19 and emerging viral diseases, it is not feasible to rely completely on the traditional process of compound discovery, which takes an average 15 years and costs $2-3 billion to bring a new compound to market [3]. A more pragmatic approach would be to perform drug repurposing, more specifically, accurately identify a set of candidate compounds which can exhibit high activity against viral proteins and potentially inhibit them using novel in-silico techniques.

In this paper, we present a consensus framework of in-silico embedding-based modeling techniques, which utilizes different combination of representations for compounds and viral proteins including:

- Morgan Fingerprints (MFP) [4] as chemoinformatic descriptors of compounds + a convolutional neural network (CNN) [5] autoencoder based vector representation for viral protein sequence.
- A teacher forcing - long short term memory neural network (TF-LSTM) [6] autoencoder based vector representation for compounds + CNN autoencoder based vector representation for viral proteins.
- Canonical SMILES based sequential representation of compounds + Primary structure (linear chain of amino acid) based sequential representation of viral proteins.

The goal of the consensus framework is to identify known and investigational compounds as candidates for viral diseases, using COVID-19

as a specific use case. The crux of our approach is that when new viruses emerge, already collected information on other viruses might be useful for inferring virus-specific compound activity. This is further suported by observations in quantitative structure-activity relationship (QSAR) models [7], where the intuition that compounds with similarities in structure and physio-chemical properties tend to have similar activities against given viral proteins is commonly utilized. For our use case, we focus on primary protein targets of severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2).

In the recent literature, a plethora of AI and network medicine based approaches have been applied for drug repurposing/repositioning [8, 9, 10, 11]. The most commonly solved problem is prediction of interaction/activity/binding affinity between compound and protein targets using variety of AI methods [12]. One of the limitations of these approaches is that they are often trained on human protein sequences (kinases, nuclear receptors, G-protein-coupled receptors) which are very different from viral protein sequences and hence they need not generalize well for in-silico compound-viral protein activity prediction. In another recent work, AtomNet [13], the authors predict the compounds-protein binding affinity using the 3d structural information extracted by convolutional neural network (CNN). However, high-quality 3d structure for novel viruses is seldom available. In [14], compound-target interactions were predicted using a hybrid approach of graph neural network [15] and recurrent neural network (RNN) [14] approach. Similarly, in [16], a hybrid CNN and RNN model called Molecule Transformer Drug-Target interaction predictor was proposed using known antiviral drugs for the potential treatment of SARS-CoV-2 infection [10]. One limitation of these approaches is that these models are trained on labelled compound-viral protein interactions in databases such as ChEMBL and do not benefit from viral protein sequences ($\sim$2.5 million) available in other databases such as Uniprot, as well as compounds ($\sim$2.5 million) in databases such as PubChem due to missing labelled interaction information. However, as shown in [17] that unsupervised or self-supervised learning on unlabelled data (i.e. learning a vector representation for a given data type) can greatly benefit the downstream supervised learning task.

Furthermore, there also exist network medicine based approaches which use knowledge graph representations [9, 8, 11] in combination with graph-theoretic (network-propagation, network proximity and diffusion) as well as graph neural network-based approaches to identify potential compounds targeting the COVID-19 as a disease. The knowledge graph is constructed using the interaction between multiple entities such as diseases, compounds, genes, human proteins and viral protein interactome. The goal is usually to identify links between existing approved compounds and new diseases such as COVID-19. Authors in [9] highlighted that the quality of the originally constructed knowledge graph from noisy sources was a potential limitation, which could impact their downstream Cov-KGE models. Additionally, these techniques can benefit from the vector representation learned for compounds and viral proteins using unsupervised learning framework as proposed in our work. The vector representation can be used in addition to the node representation learned through graph embedding procedure. In [11], the authors indicated that their deep graph neural network approach doesn't consider node features and are currently based only on the topology of the underlying graph.

In this work, we try to address several of these limitations following a data-driven perspective. We collect information about various viral organisms, their main proteins and their known compound interactions from plethora of resources including ChEMBL [18], PubChem [19], NCBI [20], UniProt [21], DrugBank [22] etc. **In this work, we use the term compounds for small molecules and compounds interchangabley**. The traditional approach for estimating compound (ligand) activity for a particular viral protein (enzyme) is through molecular docking [23]. 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 [24]. Moreover, it is computationally expensive to perform the docking simulations for a large number of compounds in combination with many viral proteins. However, it is relatively easy to collect information about the primary structure (linear chain of amino acids) for proteins associated with viruses from resources such as UniProt. Moreover, chemical information for compounds in the form of SMILES strings is readily available in resources such as DrugBank and ChEMBL. Finally, standardized activity (inhibition/potency/affinity) information for a plethora of compound-viral protein combinations is available in databases such as PubChem and ChEMBL.

These are essential resources required to build in-silico embedding-based compound-viral protein activity predictors using machine learning (ML) techniques. The primary notion is that by providing a large dataset of compound-viral protein activity, 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 compounds (or frequently occuring patterns in the MFP) that together drive the activity values to be high or low.

Our primary contributions are:

- Collection and curation of compound-viral protein activity from resources such as PubChem and ChEMBL leading to >60k interactions between >50k compounds and $\approx$ 100 viral organisms.
- Propose autoencoder frameworks (unsupervised) to obtain numeric vector representations for compounds ($\approx$ 2.5 million) and viral proteins ($\approx$ 2.5 million) respectively, which can be utilized for downstream compound-viral protein activity prediction task by traditional supervised ML techniques.
- Propose 4 different end-to-end deep learning techniques to predict compound-viral protein activity based on SMILES strings of compounds and primary structure of viral proteins.
- Showcase the effectiveness of the consensus framework as it outperforms all the individual modeling techniques on the test set.
- Identify a ranked list of 47 compounds as potential therapeutic agents for COVID-19 by targeting the three main proteins of the SARS-COV-2 virus using our consensus framework. These include 21 antivirals, 15 anticancer, 5 antibiotics, and 6 other investigational human compounds.
- Majority of the compounds in the top ranked list attain low binding energies (high binding affinity) in molecular docking experiments for each of the three viral proteins of SARS-COV-2 virus.
- Provide a general and extensible framework where individual components can be replaced to test if the respective change helps to improves the overall results. The entire source code is made publically available (`https://github.com/raghvendra5688/Drug-Repurposing`) and a web-server (`https://machinelearning-protein.qcri.org/index.html`) is also provided for the ease of non-experts.

Figure 1 illustrates our compound-viral activity prediction framework.

2 MATERIALS

In order to build our in-silico embedding-based compound-viral protein activity predictors, we collected information about compounds, viral protein sequences, and compound-viral protein interactions (activity values) from resources such as MOSES [25], ChEMBL, UniProt, PubChem and NCBI. Below we describe the details of data collection and curation steps required for the preparation of quality data, essential for accurate downstream predictive models.

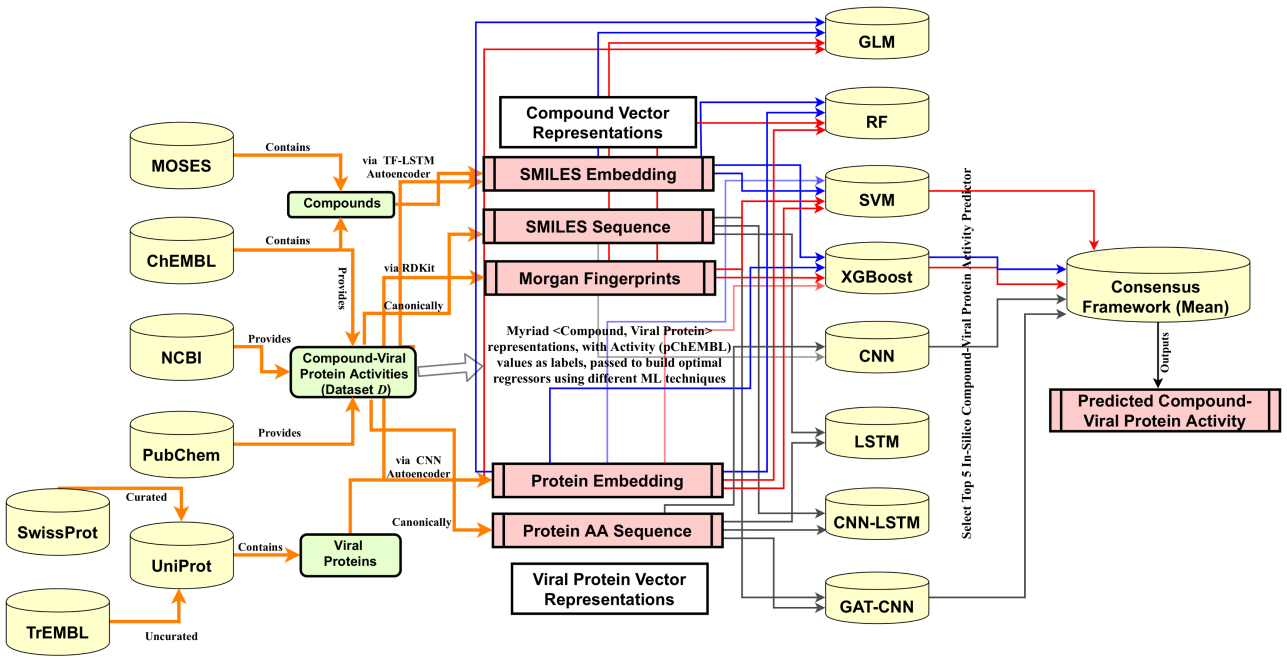


Figure 1: Flowchart of our proposed consensus framework. We collect ≈ 2.5 million SMILES representations of compounds from MOSES and ChEMBL databases. This is utilized to learn a SMILES embedding representation (numeric vector representation) via a TF-LSTM autoencoder model. We also collect ≈ 2.5 million viral protein amino acid (AA) sequences from Uniprot database. These are passed through a CNN autoencoder to learn viral protein embedding representation (numeric vector representation). We collect, curate and assimilate compound-viral protein activities from resources such as NCBI, PubChem and ChEMBL to build our dataset ($\mathcal{D}$). The corresponding bioactivities in these samples are transformed into a standardized pChEMBL value and are used to build downstream regression models. These regression models are various machine learning (ML) techniques which take advantage of different representations of compounds and viral proteins for in-silico compound-viral protein activity prediction. We then take a consensus of the top 5 predictors based on their performance w.r.t. 4 evaluation metrics on the test set. Here 'blue' color edges correspond to traditional ML models based on SMILES embedding + Protein embedding representations, 'red' color edges represent ML models based on Morgan Fingerprint (chemoinformatic descriptors) + Protein embedding representations and 'green' color edges correspond to end-to-end deep learning models based on canonical SMILES + Protein Amino Acid (AA) Sequence representations for predicting compound-viral protein activities.

2.1 Data Collection & Curation

2.1.1 Compounds:

We initially collected $556,134$ SMILES strings for compounds used in [26]. However, in order to have more robust and realistic set of molecules, the dataset was augmented with $1,936,962$ compounds available in the MOSES dataset [25]. Together these two datasets represented ≈ 2.5 million SMILES for compounds. We then filtered this dataset to remove salts and stereochemical information. In [26], the authors restricted their canonical SMILES sequence length to be in the range $[34, 74]$ for their LSTM based compound generation methodology. In [27], the authors highlighted that increasing the sequence length to 128 characters lead to better quality compound generation using an LSTM framework. In our work, we include compounds whose SMILES strings are in the range $[10, 128]$ to allow small sized compounds as well as large size ligands to be part of our chemical search space which is more inclusive and comprehensive than that used in [26]. As a result, our final compound set $\mathcal{S}$ consisted of $2,459,695$ canonical SMILES for small molecules.

To train the majority of traditional supervised ML algorithms, it is essential to have numeric vector representation for compounds. We used the set $\mathcal{S}$ to train a TF-LSTM [28, 6] based autoencoder [29] which generates a low dimensional vector representation (LS_c) for each compound. Furthermore, to have a comprehensive comparison, we also used traditional chemoinformatic descriptors such as Morgan Fingerprints (MFP) [4] dervied from compound structure as an alternative vector representation for each compound.

2.1.2 Viral Proteins:

We downloaded all the viral protein sequences available in UniProt [21] comprising a total of $2,684,774$ protein sequences. Among these $10,685$ are deposited in SwissProt [30] i.e. are manually curated and functionally annotated, whereas the remaining $2,674,089$ are obtained from TrEBML [30] and are not well-curated. These viral proteins span over $2,742$ viral organisms. A neccessary condition for training deep learning models with protein sequence is to have a fixed length L. In [31, 32], the authors used sequence legths of 800 and $1,200$ for training their deep learning models. In this work, we filter viral proteins to keep sequences with $L \leq 2000$ resulting in a set $\mathcal{V}$ of $2,658,225$ viral protein sequences, thereby, retaining $\approx 99\%$ of all viral proteins available in Uniprot.

In order to train traditional supervised methods, it is essential to have numeric vector representation for protein sequences. We utilized the set $\mathcal{V}$ to train a CNN [5] based autoencoder which then generates the required low dimensional representation (LS_v) for each viral protein sequence.

2.1.3 Compound-Viral Protein Activities:

The primary focus of our use case are the 3 main proteins 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 (PDB) [33] (PDB Ids: **6W02**,

5R7Y, **6M0J** respectively). This makes validation possible through molecular docking experiments. (b) For several other viral organisms, the PL-PRO and 3CL-PRO are the main proteins targeted by compounds [34]. (c) It has been shown [35], that Spike protein attaches the virion to the cell membrane by interacting with host receptor, initiating the infection.

However, our proposed framework can easily be extended to other viral proteins associated with the SARS-COV-2 virus as well as proteins associated with other viruses. As the SARS-COV-2 is a new virus, it is harder to get quality data about compound-viral protein activity. However, information about similar viruses, their main proteins and small molecules used to target these viral proteins are available in repositories such as PubChem, ChEMBL and BindingDB [36].

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 [20] which was used to download raw information from various NCBI Assay records. We 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 [37]. 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 [37].

We note these bioactivity measurements may vary across assays but to obtain a large set of compound-viral protein activities for the in-silico modeling techniques, it is essential to combine several of these bioactivites with certain restrictions. For example, we 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). The PubChem standard value for bioactivity is measured in micromolar ($\mu M = 10^{-6}$) concentration. We initially selected records containing IC_{50} value as done by [38], which is based on the concentration of a compound at which 50% inhibition of a viral protein is observed. Furthermore, it is known from enzyme kinetics (Cheng-Prusoff Equation [16]) that when a compound binds to a protein in an uncompetitive scenario i.e. an assay, the K_i value is equal to IC_{50} value. Similarly, it was shown in [12], that records containing K_d and Potency values as bioactivities (measured in PubChem standard value i.e. μM) can be combined with those holding IC_{50} and K_i values. Here combining corresponds to creating a dataset which includes all compound-viral protein bioactivity samples that either had K_i, K_d, IC_{50} or Potency as a label information for downstream supervised learning task. Thus, using these 4 measurements of compound-viral proteins activites and filtering records based on the aforementioned sequence lengths of compounds and viral proteins, we obtain an interaction set of $13,763$ compound-viral protein activities from PubChem.

We next downloaded all compounds and viral protein interactions available in ChEMBL [18] repository. As a part of internal quality checks provided by ChEMBL, we include only those compound-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 [18], 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 on a negative logarithmic scale. For instance, an IC_{50} value of 1 nanomolar ($nM = 10^{-9}$) would have a pChEMBL value of 9. The PubChem standard value is measured in micromolar (μM) concentration whereas the pChEMBL value is measurement in nanomolar (nM) concentration. Hence in order to have a change of unit and convert bioactivity measurements obtained from PubChem to standard pChEMBL value, we use the following formulae: $\text{pChEMBL} = -\log_{10}(\text{Activity}_{\text{PubChem}}) + 6$.

Here $\text{Activity}_{\text{PubChem}}$ corresponds to either IC_{50}, K_i, K_d or Potency. Hence 10^{-3} unit of PubChem standard value or 1 nM corresponds to a pChEMBL value of 9 ($= -\log_{10}(10^{-3})+6$). We initially obtain a set of $92,638$ such compound-viral protein activities and after filtering for only those records which contain IC_{50}, K_i, K_d 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 $54,756$ bioactivity samples obtained and curated via ChEMBL.

We take a union of the two data sources (ChEMBL and PubChem) resulting in the dataset $\mathcal{D}$ consisting of $60,195$ such interactions. These interactions comprise $54,617$ unique compounds, 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 $9:1$, which are then used as the training and independent test set respectively for the task of building in-silico embedding-based compound-viral protein activity predictors. The independent test set is pertinent to our framework as it enables us to take the consensus (mean) of the top k predictive models based on their performance on the test set.

All details of the steps followed to prepare, assimilate and curate compounds, viral proteins and compound-viral protein interactions is available in the 'README' file in the 'data' folder of the github repository (`https://github.com/raghvendra5688/Drug-Repurposing`) to enhance the reproducibility of our approach.

3 Methods

3.1 Overview

Compound-viral protein activity prediction can be modeled as a regression task. We learn a mapping function g that takes as input a joint compound and viral protein representation, (x_c, x_v) and outputs the activity value y_{cv}. In Figure 2, y_{cv} corresponds to the -$\log_{10}(IC_{50})$ and is used as standardized pChEMBL activity value. If ℓ is the model-specific loss function, then the regression task reduces to estimating the parameters w which minimizes $\min_w \sum_{c,v} \ell(y_{cv}, g(x_c, x_v; w))$

In this work, the mapping function g is a ML method including Generalized Linear Model [39], Random Forests [40], XGBoost [41], Support Vector Machines [42, 43] and ℓ is the squared loss function. For these techniques, x_c is either passed to a TF-LSTM [28] or Morgan Fingerprint generator [4] and x_v is passed to a CNN [5] to generate numeric vector representations LS_c (for compounds) and LS_v (for viral proteins) which are utilized by the aforementioned ML models to estimate activity values, such that $\hat{y}_{cv} = g(LS_c, LS_v; w)$.

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_c corresponds to canonical SMILES sequence for compounds and x_v reflects the primary structure or linear chain of amino acids (AA) for viral protein sequences and $\hat{y}_{cv} = g(x_c, x_v; w)$. The SMILES representation, is parameterized by a sequence of vectors, $x_c = \{x_{c,1}, x_{c,2}, \ldots, x_{c,l}\}$, where $x_{c,i}$ is a one-hot coded vector [44] i.e. a binary vector of length 72 (72 unique character combinations appearing in SMILES using the 'SmilesPE' package `https://github.com/XinhaoLi74/SmilesPE` in python) with 1 bit active for i^{th} character

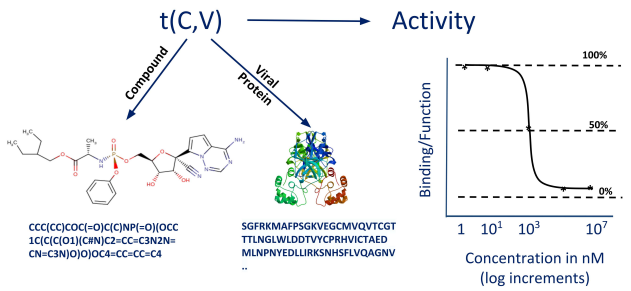


Figure 2: **Overview figure depicting our predictive modelling process.** For each compound c and each viral protein v, we use representations x_c and x_v based on SMILES strings and primary structure respectively. For each compound-viral protein interaction, the activity value used in the training set is obtained from myriad resources. Here -$\log_{10}(\text{IC}_{50})$ value measured in nM units i.e. -$\log_{10}(10^3 \times 10^{-9})$=6 is the standardized pChEMBL activity value (y_{cv}).

combination in the SMILES string and $l = 128$. Similarly, for each viral protein sequence (Protein AA 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 Compound Autoencoder: TF-LSTM

The goal of a compound autoencoder model [29] is to learn the innate low dimensional representation LS_c from SMILES strings of compounds (x_c) in an unsupervised setting such that compounds with similar patterns tend to be closer in the low dimensional space. Our compound 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. We are interested in the output of LSTM encoder that can be represented as $h = \text{EncoderLSTM}(e(x_c))$. Here $e(x_c)$ represents the SMILES embedding representation for compound, h correspond to hidden state representations encapsulating sequential information used as LS_c in our downstream predictive models. A detailed working mechanism of TF-LSTM is provided in Supplementary Material.

We trained this TF-LSTM model on $\approx$ 2.5 million SMILES strings for small molecules. Interestingly, 96.7% of the SMILES generated by our TF-LSTM model were valid small molecules (tested using RDKit [45] package) and had a mean categorical cross-entropy [46] error of 0.001. The convergence of the reconstruction error for our TF-LSTM model is depicted in Supplementary Figure 1a. Supplementary Figure 2a illustrates our TF-LSTM compound autoencoder model.

3.3 Protein Autoencoder: CNN

The goal of the viral protein autoencoder model is to learn a low dimensional Protein embedding representation LS_v from the AA 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 highlighted in Supplementary Figure 2b. The autoencoder was trained in an unsupervised fashion to learn a low dimensional space (LS_v). A detailed description of the protein autoencoder is provided in Supplementary Material.

We trained our autoencoder on $2,685,225$ viral proteins. The mean categorical cross-entropy [46] 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 four state-of-the-art ML models, namely, Generalized Linear Models (GLM) [39], Random Forests [40], XGBoost [41] and Support Vector Machines (SVM) [42, 43] as mapping function g. Thus, our predicted activity value can be represented as $\hat{y}_{cv} = g(\text{LS}_c, \text{LS}_v; w)$ for a given compound c and viral protein v. It has been shown that non-linear ML techniques such as Random Forests, XGBoost and SVMs can be used efficiently for a variety of bioinformatics problems [47, 48, 49, 50, 51, 52].

Generalized Linear Model (GLM) [39] is a flexible version of linear regression model which allows the errors or residuals of the response variable to follow a distribution other than the normal distribution. In our work, GLM serves as a baseline comparison technique. Random Forests (RF) belong to the class of ensemble supervised learning techniques. RF algorithm applies the technique of bagging or bootstrapped aggregating [40] to decision tree learners. Given $\mathcal{D}_{\text{train}}$, the bagging procedure repeatedly selects random samples with replacement and fits separate trees to these samples and aggreates them to build the final regressor.

Gradient boosting machine (GBM) [53] 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 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 [41]. It uses a scalable end-to-end tree boosting system with a weighted quantile sketch for approximate tree learning. XGBoost can scale for a large number of samples using very little computational resources.

Support vector machines (SVM) were originally introduced in [54, 42] and belong to the family of linear optimization techniques where regression task is considered as function estimation and achieved by constructing optimal hyperplanes. They only become suitable for non-linear regression task when a corresponding kernel is chosen [54, 42]. The choice of the kernel enables to encodes the similarity structure in the input data in high dimensional space. We use the radial-basis function (RBF) or universal kernel for our non-linear SVM model which is optimized using a standard cross-validation procedure.

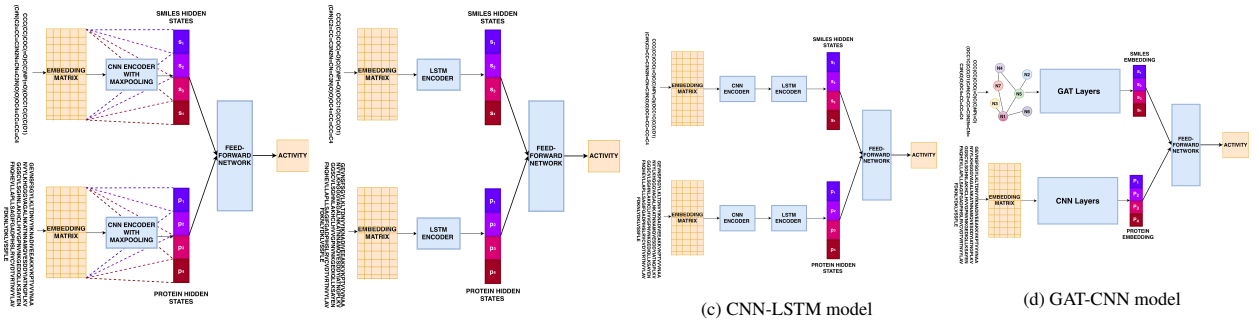
We used the 'sklearn' package [55] available in Python for building our optimal GLM, RF, XGBoost and SVM models after performing hyper-parameter optimization using 5-fold cross-validation. In order to do cross-validation, we shuffled the training dataset and then randomly split the data into 5 parts, using a combination of 4 parts as training set and 1 part as validation set to identify the optimal set of hyper-parameters. This process is repeated 5 times and the hyper-parameters with best average performance are then selected as optimal hyper-parameters. These hyper-parameters are used to build the final model on the entire training set.

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 the compound (x_c) and viral protein (x_v) representations, unlike traditional ML techniques.

3.5.1 CNN Model:

This deep learning architecture comprises two CNN encoders. For the compound and protein CNN encoders, each of the compound (x_c) and viral protein (x_v) representation is passed through an embedding layer ($e(\cdot)$) to generate compound 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 3a. The hidden state vector h_c for compounds and h_v for viral protein



(a) CNN model (b) LSTM model (c) CNN-LSTM model (d) GAT-CNN model

Figure 3: Different end-to-end deep learning models used as data-driven predictive models for the task of estimating compound-viral protein activity.

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 value. 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 value 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 compound representation (x_c) and another one based on the viral protein representation (x_v). The compound LSTM encoder generates the hidden state vector (h_c) 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 3b.

We again have multiple feed-forward layers on top of h which is connected to the output unit representing the activity value. 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 value to be high or low for a given compound-viral protein combination.

3.5.3 CNN-LSTM Model:

The CNN-LSTM model is a combination of CNN and the LSTM model. By combining the CNN and LSTM models, this model can capture spatially 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 value.

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

This deep learning architecture is composed of two parts, graph attention networks [56] and convolutional neural networks. For a given compound, the compound structure can be presented as a graph consisting of the atoms (nodes) in the compound 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. 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 compound molecule (x_c) along with the extracted node features, the GAT model learns an embedding representation for a compound encapsulating the topological information available in the graph of each compound.

The second component of this architecture is convolutional neural networks which take protein AA 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. It learns protein embedding (h_v) and concatenates it with the SMILES embedding (h_c) generated by GAT to produce h, which is then passed to feed-forward layers. The output layer provides the value corresponding to the compound activity.

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

3.6 Consensus Framework

In [11], the authors demonstrate that taking an aggregation of the results obtained from different methodologies can provide better performance than individual models while identifying suitable repurposable compounds for COVID-19. In a similar vein, we take a consensus i.e. the average of the pChEMBL values predicted by our top performing in-silico embedding-based compound-viral protein predictors on the independent test set. We argue that since our models are based on myriad representations of compounds (SMILES embedding or MFP or canonical SMILES) and viral proteins (Protein embedding or AA Sequence), it is imperative to take a consensus of the top predictive models as they learn different combinations of non-linear patterns from diverse representations of the data to attain optimal predictive performance as illustrated in Table 1. Figure 1 highlights the various combinations of input data representations and the top compound-viral protein predictors aggregated in the consensus framework based on the performance on the test set as illustrated in 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 (9:1) for training and testing purposes respectively as mentioned earlier in the Materials section. During each randomization, all the models are built using the same $\mathcal{D}_{\text{train}}$ and evaluated on the same test set $\mathcal{D}_{\text{test}}$ to avoid

Model	Representations	MAE	MSE	Pearson R	R2
Mean	Dummy regressor	1.100 ± 0.003	1.936 ± 0.006	-	-
Median	Dummy regressor	1.002 ± 0.002	2.310 ± 0.004	-	-
GLM	SMILES Emedding + Protein Embedding	0.662 ± 0.003	0.869 ± 0.009	0.740 ± 0.003	0.548 ± 0.005
RF	SMILES Emedding + Protein Embedding	0.557 ± 0.005	0.625 ± 0.010	0.826 ± 0.003	0.682 ± 0.005
SVM	SMILES Emedding + Protein Embedding	0.508 ± 0.004	0.478 ± 0.011	0.869 ± 0.002	0.755 ± 0.003
XGBoost*	SMILES Emedding + Protein Embedding	0.453 ± 0.003	0.423 ± 0.007	0.885 ± 0.002	0.783 ± 0.004
GLM	Morgan Fingerprint + Protein Embedding	0.647 ± 0.003	0.775 ± 0.008	0.774 ± 0.003	0.600 ± 0.005
RF	Morgan Fingerprint + Protein Embedding	0.529 ± 0.003	0.552 ± 0.004	0.849 ± 0.002	0.720 ± 0.003
SVM*	Morgan Fingerprint + Protein Embedding	0.439 ± 0.003	0.357 ± 0.005	0.905 ± 0.002	0.818 ± 0.003
XGBoost*	Morgan Fingerprint + Protein Embedding	0.404 ± 0.002	0.329 ± 0.003	0.911 ± 0.001	0.830 ± 0.002
CNN*	SMILES Sequence + Protein AA Sequence	0.451 ± 0.003	0.398 ± 0.006	0.892 ± 0.002	0.795 ± 0.004
LSTM	SMILES Sequence + Protein AA Sequence	0.500 ± 0.002	0.514 ± 0.006	0.863 ± 0.002	0.745 ± 0.003
CNN-LSTM	SMILES Sequence + Protein AA Sequence	0.516 ± 0.004	0.551 ± 0.009	0.852 ± 0.002	0.725 ± 0.004
GAT-CNN*	SMILES Sequence + Protein AA Sequence	0.478 ± 0.003	0.439 ± 0.007	0.880 ± 0.002	0.775 ± 0.003
μ_{Best} **(Top 10 Methods)**	All combination	0.423 ± 0.004	0.342 ± 0.009	0.911 ± 0.003	0.829 ± 0.005
μ_{Best} **(Top 5 Methods)**	All combinations	**0.403 ± 0.002**	**0.313 ± 0.006**	**0.917 ± 0.002**	**0.841 ± 0.003**

Table 1. Comparison of performance of devised ML techniques for our compound-viral activity prediction problem evaluated w.r.t. the 4 evaluation metrics on $\mathcal{D}_{\text{test}}$. Here we report the mean performance and ± corresponds to maximal standard deviation. Top 10 models are highlighted in bold and "*" superscript is added to top 5 models w.r.t. the 4 evaluation metrics. Last row corresponds to mean of top 5 methods.

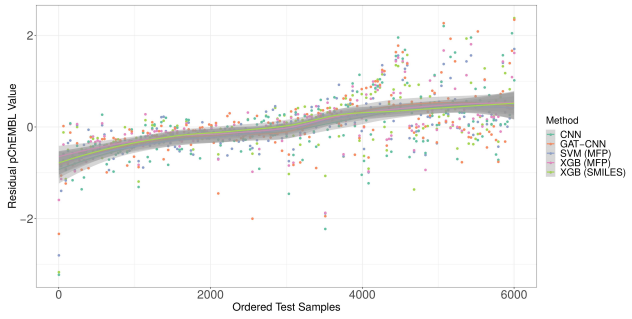
any unwanted bias in the downstream consensus framework. For the traditional machine learning techniques (GLM, RF, SVM and XGBoost), the optimal hyper-parameters are obtained using a 5-fold cross-validation technique on $\mathcal{D}_{\text{train}}$. However, in order to identify the optimal architecture for the end-to-end deep learning models, the training sets ($\mathcal{D}_{\text{train}}$) are divided on the fly into 80% for training and 20% for validation set owing to computational costs. The cross-validation performance of traditional supervised machine learning techniques (GLM, RF, SVM and XGB) using either SMILES embedding representation or Morgan Fingerprints representation for compounds and Protein embedding representation for viral proteins is depicted in Supplementary Figure 2c.

Table 1 provides a comprehensive comparison of the mapping functions g utilized in our work including baseline mean, median and optimal GLM regressors as well as optimal non-linear models such as RF, SVM, XGBoost, CNN, LSTM, CNN-LSTM and GAT-CNN. In Table 1, we report the mean and corresponding standard deviation ($\pm$) in performance for each of the 4 quality metric over the 10 randomizations. These 4 quality metrics are the mean absolute error (MAE), mean squared error (MSE), pearson correlation R (Pearson R) and the coefficient of determinination (R2). Each of these metrics are estimated using the predicted pChEMBL values vs the groundtruth pChEMBL values for compound-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 model's predictive capability.

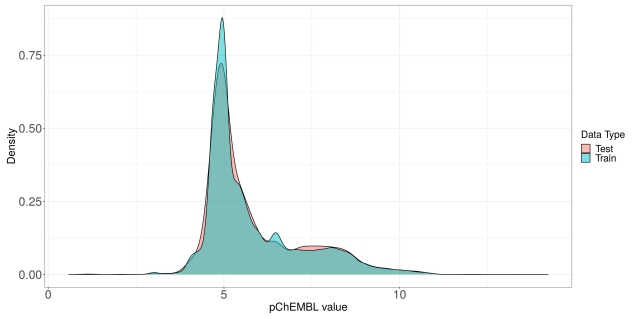
We highlight two baseline regressors i.e. the mean and the median regressor to showcase the effectiveness of our non-linear predictive models in Table 1. Here the mean regressor takes the mean value of all the compound-viral protein activities available in the training set and considers it as fixed output from the regressor. Similarly, the median regression outputs the median value of all the compound-viral protein activities available in the training set. The performance of these two baseline regressors are significantly lower than other machine learning techniques. Additionally, we demonstrate that the GLM models built on LS_c and LS_v for compounds and viral proteins respectively are two of the worst performing models w.r.t. 4 evaluation metrics. This necessitates the usage of non-linear machine learning techniques when using numeric vector representations for compounds (SMILES embedding/Morgan Fingerprint) and proteins (Protein embedding) for the task of accurate compound-viral protein activity prediction as illustrated in Table 1.

From Table 1, we observe that the best individual predictive model w.r.t. all quality metrics is the XGBoost model, highlighted in Table 1 by '$*$', and is built on the LS_c using the Morgan Fingerprint and LS_v obtained from protein autoencoder for the compounds and viral proteins respectively. It is closely followed by the SVM model on similar representations, the end-to-end CNN and GAT-CNN end-to-end deep learning models based on the sequence representations and the XGBoost model built on the SMILES embedding (LS_c) for compounds and protein embedding (LS_v) for viral proteins. These top 5 models each achieve Pearson R > 0.85 and R2 value in excess 0.75. Furthermore, we observe from Table 1, that when we take a consensus (average) of the top 10 predictive models, its performance is comparable to that of the best individual predictive (XGBoost) model. This can partly be reasoned due to the inclusion of models with much lower predictive capability, such as RF (SMILES Embedding/ Morgan Fingerprint + Protein Embedding) and CNN-LSTM deep learning models in the consensus, in comparison to the top performing predictive models. However, when we take a consensus of the top 5 predictors, we achieve the superior performance than the best individual predictor (XGBoost model) as depicted in Table 1. Its superior predictive capability can be attributed to the high Pearson R and R2 of the individual models included in the consensus and the ability to potentially capture different combinations of non-linear patterns from the diverse representations of the data. It is noteworthy, that the standard deviations of each predictive model obtained via 10 randomizations of $\mathcal{D}_{\text{test}}$ are low w.r.t. the 4 evaluation metrics as illustrated in Table 1, indicating low variance and high accuracy 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 Supplementary Figure 3. We additionally compare the predictive performance of the top 5 in-silico predictors w.r.t. the ground-truth compound-viral protein activites available for the same test set $\mathcal{D}_{\text{test}}$ as illustrated in Figure 4a. It can be observed from Figure 4a that the x-axis represents the sample id in $\mathcal{D}_{\text{test}}$, whereas for each such sample, we have 5 values vertically spread along the y-axis. Each of these values corresponds to the difference between the groundtruth and predicted interaction values by our in-silico embedding-based models. The closer the predicted score is to the true pChEMBL value, the smaller is the residual pChEMBL value (≈ 0). We observe more deviations from 0 in the residual pChEMBL values i.e. relatively larger errors in predictions, when the true pChEMBL value is either too small (close to sample id '0' on x-axis) or too large (close to sample



(a) Comparison of difference in groundtruth vs predicted pChEMBL values for top 5 in-silico embedding-based compound-viral activity predictors on the same $\mathcal{D}_{\text{test}}$.



(b) Comparison of density distribution of pChEMBL values in training and test set.

Figure 4: In Figure 4a, the x-axis represents compound-viral protein activity samples ordered by their groundtruth pChEMBL values (lowest to highest) and y-axis corresponds to the residual pChEMBL values. For a majority of the samples the residuals are close to zero for each of the top 5 predictors indicating the good predictive capability of these models. We use the 'loess' function with default parameters available in 'ggplot2' package in R to fit a smooth local regressor via a non-parametric approach for each in-silico predictor. Figure 4b shows the distribution as well as the density of the pChEMBL values available in $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$.

id '6000' on x-axis). This can partly be attributed to lack of availability of large number of compound-viral protein activity samples with small pChEMBL values (≤ 5) or large pChEMBL values (≥ 10) in the training set $\mathcal{D}_{\text{train}}$ as depcited in Figure 4b to train the in-silico embedding-based predictors. However, for a majority of the samples the residuals are close to 0 for each of the top 5 predictors showcasing their good predictive capability.

4.2 Experimental Results for COVID-19 Use Case

In a recent work [57], a library of compounds encompassing approximately $12,000$ clinical-stage or Food and Drug Administration (FDA)-approved small molecules were profiled by means of assay development and high throughput screening based on bioactivities against SARS-COV-2 virus in Vero E6 cells to identify candidate therapeutic drugs for COVID-19. The authors in [57] have deposited a total of 68 assays containing $2,483$ of these compounds in a publically available library named ReFRAME (`https://reframedb.org/`). As per the guidelines mentioned on their website, a good majority of these compounds are embargoed due to collaborations with pharmaceutical companies. Furthermore, we take a union of these compounds with 117 FDA approved drugs which are in some stage of clinical trial for any known viral organism as indicated in [58] and available at `http://drugvirus.info/`. After filtering these compounds based on the length of their SMILES sequences as per the criterion defined in the Materials section, we end up with a set of $\mathcal{S}$ comprising $1,482$ compounds including known antivirals, antibiotics, anticancer and other human investigational compounds (see details about preparing the set $\mathcal{S}$ in the Supplementary). We then make activity predictions (pChEMBL values) for each of these compounds on the 3 main proteins of SARS-COV-2 virus i.e. the PL-Pro (PDB ID: 6WO2), the 3CL-Pro (PDB ID: 5R7Y) and the Spike proteins (PDB ID: 6MOJ). Additional information about their primary AA sequence, Uniprot Ids, etc. are provided in Supplementary Table 2.

We first obtain the predicted pChEMBL values for the top 5 in-silico compound-viral activity predictors as indicated in Table 1 and take a consensus i.e. on average of these predictions for each of the three main proteins of the SARS-COV-2 virus. We then select the top 100 compounds with the highest predicted activity against each of these three viral proteins. By taking an intersection of the compounds in these lists, we obtain a set of 47 compounds which are consistently predicted to have high activity (high pChEMBL valuess) against all the three main viral proteins and thus can potentially be effective against the SARS-COV-2 virus. This candidate set includes 21 antivirals, 15 anticancer, 5 antibiotics and 6 other investigational human compounds as depicted in Table 2. Our candidate list includes antiviral therapies such as Lopinavir, Ritonavir and Filociclovir which have been undergoing clinical trials (`https://clinicaltrials.gov/`) for SARS-COV-2 as highlighted in [57]. Our consensus embedding-based in-silico framework also identifies Remdesivir, a viral RNA polymerase inhibitor [59], which has been granted emergency use authorization by the FDA for the treatment of COVID-19 on the basis of clinical trial data demonstrating a reduction in time to recovery [60].

In [8], the authors identified several compounds including Toremifene using a network-based drug repurposing approach for SARS-COV-2 which they further validated against the Spike viral protein using a comprehensive combination of homology modeling, molecular docking, molecular dynamics simulation, and binding affinity calculations in [61]. In a similar vein to showcase the accuracy of our consensus framework, we perform additional molecular docking experiments on the set of 47 compounds which consistently had high predicted activites against the three main viral proteins of SARS-COV-2 virus. All details related to molecular docking experiment setup are provided in the Supplementary. For each of the three main proteins, we highlight our predicted pChEMBL value and the corresponding binding energy score obtained via molecular docking for the 47 candidate compounds in Table 2. We observe that a good majority of the top ranked compounds consistently achieved low binding energy ($\leq$ -6 Kcal/mol) in the molecular docking experiments for all the considered viral proteins of SARS-COV-2 as illustrated in Table 2. It is noteworthy that among all the compounds in our final candidate list, LM 565 is the only compound which attains high binding energy score in the docking experiments for each of the three viral proteins and thus can potentially be a false positive. This illustrates that our consensus framework can serve as a data-driven screening tool which helps to reduce the list of candidate drugs from an initial set $\mathcal{S}$ ($1,482$ compounds) to the curated list of 47 potenital compounds ($\approx 3\%$ of original set $\mathcal{S}$) which can either be validated through molecular docking experiments (reducing computational costs) or through bioassays in absence of known 3d crystal structure of viral proteins.

Furthermore from our molecular docking experiments, we identified Rifabutin (in the set of 47 curated compounds), an antibiotic used to treat tuberculosis and Mycobacterium avium complex, to have the lowest binding energy scores for each of the three main viral proteins of SARS-COV-2.

Compound	PL-Pro		3CL-Pro		Spike Protein	
	Predicted pChEMBL	Binding Energy	Predicted pChEMBL	Binding Energy	Predicted pChEMBL	Binding Binding
Lopinavir^{+}	7.777	-6.3	7.851	-8.7	8.226	-5.0
Ritonavir^{+}	7.562	-6.4	7.777	-7.7	7.845	-5.5
Palinavir^{+}	7.416	-6.4	7.48	-7.2	7.699	-6
Simeprevir^{+}	7.646	-5.6	7.476	-6.1	8.206	-6.2
Cabotegravir^{+}	7.194	-7.1	6.951	-9.5	7.002	-6.8
L-870812^{+}	6.937	-7.1	6.895	-8.9	6.68	-7.2
MK-4965^{+}	7.319	-7.5	6.893	-9.6	7.302	-7.1
Tipranavir^{+}	6.634	-7.4	6.83	-8.3	6.794	-6.6
Zanamivir^{+}	6.798	-5.7	6.801	-5.9	6.748	-5.9
BMS-707035^{+}	6.938	-7.2	6.766	-8.8	6.511	-6.6
GSK-364735^{+}	7.086	-6.4	6.745	-9.6	6.552	-7
Paritaprevir^{+}	6.751	-6.8	6.571	5.6	7.443	-6.2
Filociclovir^{+}	6.542	-5.7	6.463	-7.1	6.647	-6.2
TMC-647055^{+}	6.717	-5.8	6.459	11.7	6.539	-5.5
Elvitegravir^{+}	6.462	-6.8	6.402	-8	6.236	-5.7
Dapivirine^{+}	6.584	-6.7	6.385	-8.7	6.32	-6.4
PLX-8394*	6.208	-9.1	6.358	-9.4	6.494	-7.2
Triciribine PO_3 *	6.385	-6.7	6.354	-8.1	6.314	-6.5
Zidovudine^{+}	5.966	-5.7	6.279	-7.4	6.264	-5.6
API-2$^{*/+}$	5.964	-6.7	6.175	-8.3	6.208	-5.7
Fluorouracil^{+}	5.965	-4.5	6.157	-5.2	6.353	-4.6
Gossypol*	6.029	-5.7	6.11	-4.2	6.069	-6
LM 565^{-}	6.137	8.7	6.094	72.4	6.461	-2.9
PF-03814735*	6.051	-7.6	6.091	-8.2	6.126	-7.1
Barasertib*	6.006	-8	6.087	-8.3	6.171	-6.8
Edoxudine^{+}	5.925	-5.9	6.075	-7.6	6.246	-5.6
Cefozopran^{-}	5.884	-7	6.049	-8.1	6.255	-6
Entrectinib*	6.231	-6.8	6.039	-9.3	6.023	-7
Clemizol*	6.085	-6.2	6.015	-8	6.105	-6
VBY-825^{+}	6.112	-6	6.006	-8	6.07	-4.7
R-763*	6.158	-6.6	6.002	-7.8	6.26	-6.7
Bietaserpine$^{@}$	6.054	-6.1	5.994	-2.1	6.323	-4.9
ACT-077825$^{@}$	5.916	-6.9	5.973	-6.7	6.223	-4.8
MP-412*	6.069	-6.6	5.971	-9	6.243	-5.6
Remdesivir^{+}	5.907	-6.2	5.964	-8	6.37	-6.4
ABT-263*	6.005	-4.2	5.925	1.9	6.211	-5.6
BMS-903452$^{@}$	5.929	-6.9	5.913	-7.8	6.174	-6.3
Brilacidin$^{@}$	6.016	-5.7	5.913	-2.2	6.266	-5.2
Taselisib*	5.934	-7	5.906	-8.6	6.142	-7.1
Goxalapladib$^{@}$	5.982	-6.9	5.905	-6.6	6.27	-5.1
HKI-357*	6.009	-6.8	5.884	-8.7	6.143	-6.2
Sitravatinib*	5.895	-6.3	5.879	-8.2	6.069	-7
Rifabutin^{-}	5.904	-9.4	5.878	-12.3	6.136	-12.1
Omadacycline^{-}	6.002	-6.1	5.865	-2.6	6.251	-5.3
Cefpiramide^{-}	5.883	-6.8	5.851	-8.3	6.179	-5.9
VCH-286$^{@}$	5.88	-6.6	5.847	-8.1	6.028	-4.6
BMS-754807*	5.915	-6.6	5.833	-8.3	6.095	-7.1

Table 2. Top ranked 47 compounds for each of PL-Pro, 3CL-Pro and Spike proteins of SARS-COV-2 virus consistently appearing in the ranked list of top 100 compounds against these viral proteins. The 'PPS' representS the predicted pChEMBL value by the consensus model whereas 'BE' corresponding to binding energy (units: Kcal/mol) obtained via molecular docking experiment. Here $+$, $-$, $*$, and @ correspond to antivirals, antibiotics, anticancer and other human compounds respectively. Here Rifabutin is highlighted in bold as it consistently achieves a low binding energy in the molecular docking experiments. Similarly, LM 565 is italizied as it constantly attains high binding energy score in the docking experiments and can potentially be a false positive.

In a recent review, the authors [62] highlighted that bacteriophages such as Rifabutin can be a potential game changer in the trajectory of COVID-19. Here we provide additional insights about the interaction of Rifabutin with SARS-COV-2 viral proteins. The PL-Pro viral protein has a right-hand thumb-palm-fingers architecture, contains a ubiquitin-like domain (UBL) at the N-terminal (see Figure 5A). Several Van der waal as well as hydrogen bond interactions stabilizes the PL-Pro-Rifabutin complex (see Figures 5B and 5C).

The 3CL-Pro viral protein regulates transcription and replication processes by cleaving the polyprotein chains into different non-structural proteins. It has 306 AA residues with three distinct domains (I-III). The domains I and II mainly have an antiparallel β-barrel structure, while domain III comprises five α-helices (see Figure 5D). Rifabutin docks at the interface between domain II and III of 3CL-Pro and the complex is stabilized by several interactions with AA residues from both domains (see Figures 5E and 5F). The core of RBD of Spike protein consists of antiparallel β-sheets (b1-4 and b7) with short interconnecting loops and helices (see Figure 5G). Rifabutin binds closer to the region of Spike protein-ACE-2 interaction site, and the complex is stabilized by hydrogen bonds and hydrophobic interactions (see Figures 5H and 5I).

5 DISCUSSION & CONCLUSION

In this work, we showcase that the problem of predicting activity value for compound-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 compounds (SMILES strings or Morgan Fingerprints) and viral protein sequences (AA sequence) can be used accurately for the aforementioned task. As our models are based on representations of compounds (x_c) and viral proteins (x_v), we can further enhance our models by using additional information such as 2d images of compounds. Similarly, we can utilize information such as physio-chemical and structural properties of proteins as showcased in [31, 32], to strengthen our models.

Our predictive framework is built on $\mathcal{D}_{\text{train}}$, which contains information for over 97 different viral organisms along with their main proteins, hence our models are generalizable. This means that our models can produce an accurate ranked list of potential inhibitors for the next big viral threat once its associated proteins are known and thus can be used as a data-driven screening tool. Moreover, it known that viruses frequently mutate [63]. As a result, the viral protein will also have multiple point mutations i.e. few AA in the viral protein sequence might change. This can have an immense impact on the 3d structure as well as the functionality of the viral protein [64]. Thus, techniques based on virtual ligand screening using docking experiments (high-quality 3d structure) such as [65, 66, 67] can suffer in this situation. However, our models focus on the primary structure and with point mutations, the vector representations LS_v and x_v will change. But since our mapping functions are generalizable (based on frequently co-occurring k-mers and subsequences in SMILES strings), we will end up with a revised ranked list of compounds for the mutated viral protein in a computationally efficient manner.

For the COVID-19 use case, our consensus framework identifies a list of 47 compounds as potential inhibitors. By further validating this curated list using molecular docking experiments, we identified Rifabutin as a potential inhibitor as it consistently achieved low binding energy score for all the three main proteins of SARS-COV-2 virus. This suggests that a hybrid drug-repurposing approach can be developed, where in-silico compound-viral protein activity predictors can be used initially to screen a large set of compounds to produce a much smaller list of compounds. This list can further be curated using molecular docking experiments (utilizing high quality 3d crystal structures) to prioritize the potential candidates for downstream in-vivo clinical trial stages.

Moreover, for the COVID-19 use case, our consensus framework recognized antivirals such as Remdesivir, Lopinavir, Ritonavir which have been identified by multiple in-silico and in-vitro studies [68, 69] to be potentially effective against the SARS-COV-2 virus. However, according to the recent results from the SOLIDARITY trial [70], the aforementioned antivirals appear to have little or no meaningful effect on overall

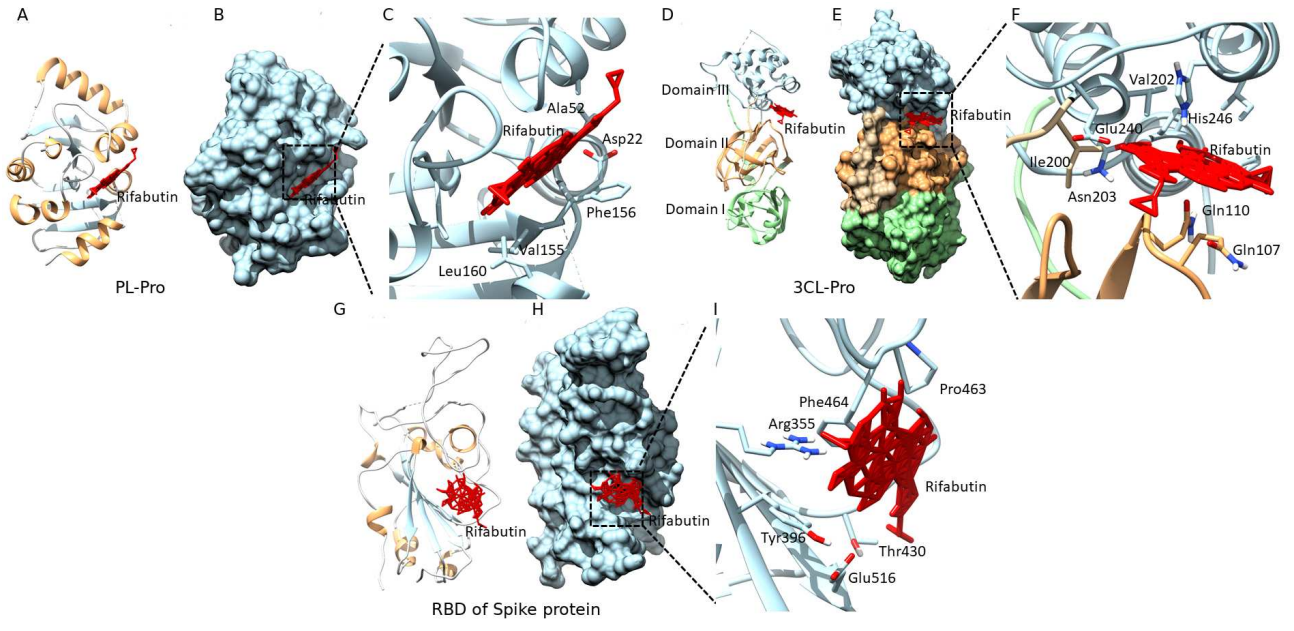


Figure 5: (A) Cartoon representation of the PL-Pro viral bound to Rifabutin (red). Protein was colored according to secondary structure: helices are brown and strands are blue. (B) Surface representation of complex structure highlighting the binding surface. (C) Rifabutin interactions with the AA residues of PL-Pro protease. (D) Cartoon representation of the 3CL-Pro viral protein bound to rifabutin (red). The three domains are shown in different color: domain I as light green, domain II as light brown and domain III as light blue. (E) Surface representation of complex structure highlighting the binding groove at the domain interface. (F) Rifabutin making significant interactions with the crucial AA residues of SARS-Cov-2 3CL-Pro protease. (G) Cartoon representation of the Spike protein bound to Rifabutin (red). Protein was rendered according to secondary structure elements. (H) Surface representation of complex structure highlighting the binding surface. (I) Rifabutin interacts with RBD of the Spike protein close to its binding to the receptor.

mortality rate in hospitals. This highlights a limitation of our work. Our current mapping function g only considers x_c and x_v and doesn't include any information about the host organism (x_h). Recently, in [71], 26 SARS-CoV-2 viral proteins were expressed in human cells and 332 high confidence human protein interactions were identified using a network-based drug-repurposing approach. Similarly, in [11], a consensus of network-based approaches was utilized to identify repurposing candidates. Their drug-repurposing strategy relied on network proximity, diffusion, and AI-based metrics, allowing to rank all approved compounds based on their likely efficacy for COVID-19 disease leading to 81 promising candidates. In [9], a network-based deep learning framework is used on top of a knowledge graph constructed on multiple entities such as diseases, drugs/compounds, genes and proteins (human and viral protein interactome) with the goal to identify links between existing approved compounds and COVID-19. Moreover, the tool CoVex [72] integrates the human protein-protein interaction and the host-interacting proteins to employ strategies such as trustrank or multi-steiner trees to identify repurposable drugs for COVID-19.

All the above mentioned approaches take into consideration the interaction with the human interactome, a key missing link in our current framework. In future, we plan to extend our mapping function to become $g(x_c, x_v, x_h; w)$, by considering compound-viral protein interactions, compound-human protein target interactions, human protein-protein interactions, human protein-viral protein interactions in a similar knowledge graph representation to identify potentially repurposable compounds for any viral disease. Another strand of work that we can be explored, is the use of Transformer Networks which use self-attention to capture long range dependency in sequence to sequence modeling for building the SMILES embedding representation. Recent work in natural language processing has convincingly demonstrated that Transfomer Networks are substantially more proficient than LSTMs with comparable level of accuracy [73]. In our particular instance, both the SMILES representation for compounds and linear chain of amino acids for proteins can benefit from these approaches.

References

[1]Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[2]World Health Organization. Coronavirus disease (covid-2019) situation reports - 139. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200607-covid-19-sitrep-139.pdf, June 2020.

[3]Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

[4]Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):1–15, 2020.

[5]Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[6]Alex M Lamb, Anirudh Goyal Alias Parth Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*, pages 4601–4609, 2016.

[7]Kunal Roy, Supratik Kar, and Rudra Narayan Das. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Academic press, 2015.

[8]Yadi Zhou, Yuan Hou, Jiayu Shen, Yin Huang, William Martin, and Feixiong Cheng. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell discovery*, 6(1):1–18, 2020.

[9]Xiangxiang Zeng, Xiang Song, Tengfei Ma, Xiaoqin Pan, Yadi Zhou, Yuan Hou, Zheng Zhang, Kenli Li, George Karypis, and Feixiong Cheng. Repurpose open data to discover therapeutics for covid-19 using deep learning. *Journal of proteome research*, 2020.

[10]Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. *Computational and structural biotechnology journal*, 2020.

[11]Deisy Morselli Gysi, Ítalo Do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Helia Sanchez, Rebecca Marlene Baron, Dina Ghiassian, Joseph Loscalzo, et al. Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint arXiv:2004.07229*, 2020.

[12]Maha Thafar, Arwa Bin Raies, Somayah Albaradei, Magbubah Essack, and Vladimir B Bajic. Comparison study of computational prediction tools for drug-target binding affinities. *Frontiers in Chemistry*, 7, 2019.

[13]Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

[14]Kyle Yingkai Gao, Achille Fokoue, Heng Luo, Arun Iyengar, Sanjoy Dey, and Ping Zhang. Interpretable drug target prediction using deep neural representation. In *IJCAI*, volume 2018, pages 3371–3377, 2018.

[15]Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[16]Benoit Beck, Yun-Fei Chen, Walthere Dere, Viswanath Devanarayan, Brian J Eastwood, Mark W Farmen, Stephen J Iturria, Phillip W Iversen, Steven D Kahl, Roger A Moore, et al. Assay operations for sar support. In *Assay Guidance Manual [Internet]*. Eli Lilly & Company and the National Center for Advancing Translational Sciences, 2017.

[17]Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. In *Advances in Neural Information Processing Systems*, pages 7647–7657, 2019.

[18]Anna Gaulton, Anne Hersey, Michal Nowotka, A Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

[19]Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[20]David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1):D13–D21, 2007.

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

[22]David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

[23]Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

[24]Sohini Chakraborti and Narayanaswamy Srinivasan. 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, 2020.

[25]Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, Artur Kadurin, Sergey Nikolenko, Alan Aspuru-Guzik, and Alex Zhavoronkov. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv preprint arXiv:1811.12823*, 2018.

[26]Anvita Gupta, Alex T Müller, Berend JH Huisman, Jens A Fuchs, Petra Schneider, and Gisbert Schneider. Generative recurrent networks for de novo drug design. *Molecular informatics*, 37(1-2):1700111, 2018.

[27]Matt O. Connor. *Deep Learning Coronavirus Cure*, 2020 (accessed June 25, 2020).

[28]Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *IET*, 1999.

[29]Mark A Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[30]Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31(1):365–370, 2003.

[31]Sameer Khurana, Reda Rawi, Khalid Kunji, Gwo-Yu Chuang, Halima Bensmail, and Raghvendra Mall. Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, 34(15):2605–2613, 2018.

[32]Abdurrahman Elbasir, Balasubramanian Moovarkumudalvan, Khalid Kunji, Prasanna R Kolatkar, Raghvendra Mall, and Halima Bensmail. Deepcrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics*, 35(13):2216–2225, 2019.

[33]Protein Data Bank. Protein data bank. *Nature New Biol*, 233:223, 1971.

[34]Georgie Fear, Slavko Komarnytsky, and Ilya Raskin. Protease inhibitors and their peptidomimetic derivatives as potential drugs. *Pharmacology & therapeutics*, 113(2):354–368, 2007.

[35]Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, et al. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, pages 1–6, 2020.

[36]Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[37]Joseph V Haas, Brian J Eastwood, Philip W Iversen, Viswanath Devanarayan, and Jeffrey R Weidner. 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, 2017.

[38]Ehsan Ullah, Raghvendra Mall, Halima Bensmail, Reda Rawi, Saila Shama, Nooral Al Muftah, and Ian Richard Thmpson. Identification of cancer drug sensitivity biomarkers. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2322–2324. IEEE, 2017.

[39]Alan Agresti. *Foundations of linear and generalized linear models*. John Wiley & Sons, 2015.

[40]Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[41]Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[42]Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[43]Raghvendra Mall and Johan AK Suykens. Very sparse lssvm reductions for large-scale data. *IEEE transactions on neural networks and learning systems*, 26(5):1086–1097, 2015.

[44]David Harris and Sarah Harris. *Digital design and computer architecture*. Morgan Kaufmann, 2010.

[45]Greg Landrum. Rdkit documentation. *Release*, 1:1–79, 2013.

[46]Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[47]Raghvendra Mall, Luigi Cerulo, Halima Bensmail, Antonio Iavarone, and Michele Ceccarelli. Detection of statistically significant network changes in complex biological networks. *BMC systems biology*, 11(1):32, 2017.

[48]Reda Rawi, Raghvendra Mall, Khalid Kunji, Chen-Hsiang Shen, Peter D Kwong, and Gwo-Yu Chuang. Parsnip: sequence-based protein solubility prediction using gradient boosting machine. *Bioinformatics*, 34(7):1092–1098, 2018.

[49]Raghvendra Mall, Luigi Cerulo, Luciano Garofano, Veronique Frattini, Khalid Kunji, Halima Bensmail, Thais S Sabedot, Houtan Noushmehr, Anna Lasorella, Antonio Iavarone, et al. Rgbm: regularized gradient boosting machines for identification of the transcriptional regulators of discrete glioma subtypes. *Nucleic acids research*, 46(7):e39–e39, 2018.

[50]Ehsan Ullah, Raghvendra Mall, Reda Rawi, Naima Moustaid-Moussa, Adeel A Butt, and Halima Bensmail. 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, 2018.

[51]Joao Palotti, Raghvendra Mall, Michael Aupetit, Michael Rueschman, Meghna Singh, Aarti Sathyanarayana, Shahrad Taheri, and Luis Fernandez-Luque. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ digital medicine*, 2(1):1–9, 2019.

[52]Abdurrahman Elbasir, Raghvendra Mall, Khalid Kunji, Reda Rawi, Zeyaul Islam, Gwo-Yu Chuang, Prasanna R Kolatkar, and Halima Bensmail. Bcrystal: an interpretable sequence-based protein crystallization predictor. *Bioinformatics*, 36(5):1429–1438, 2020.

[53]Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[54]Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.

[55]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[56]Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[57]Laura Riva, Shuofeng Yuan, Xin Yin, Laura Martin-Sancho, Naoko Matsunaga, Lars Pache, Sebastian Burgstaller-Muehlbacher, Paul D De Jesus, Peter Teriete, Mitchell V Hull, et al. Discovery of sars-cov-2 antiviral drugs through large-scale compound repurposing. *Nature*, 586(7827):113–119, 2020.

[58]Petter I Andersen, Aleksandr Ianevski, Hilde Lysvand, Astra Vitkauskiene, Valentyn Oksenych, Magnar Bjørås, Kaidi Telling, Irja Lutsar, Uga Dampis, Yasuhiko Irie, et al. Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseases*, 2020.

[59]Travis K Warren, Robert Jordan, Michael K Lo, Adrian S Ray, Richard L Mackman, Veronica Soloveva, Dustin Siegel, Michel Perron, Roy Bannister, Hon C Hui, et al. Therapeutic efficacy of the small molecule gs-5734 against ebola virus in rhesus monkeys. *Nature*, 531(7594):381–385, 2016.

[60]Food, Drug Administration, et al. Coronavirus (covid-19) update: Fda issues emergency use authorization for potential covid-19 treatment. *FDA News Release*, 1, 2020.

[61]William R Martin and Feixiong Cheng. Repurposing of fda-approved toremifene to treat covid-19 by blocking the spike glycoprotein and nsp14 of sars-cov-2. 2020.

[62]Marcin W Wojewodzic. Bacteriophages could be a potential game changer in the trajectory of coronavirus disease (covid-19). *PHAGE*, 1(2):60–65, 2020.

[63]W Robert Fleischmann Jr. Viral genetics. In *Medical Microbiology. 4th edition*. University of Texas Medical Branch at Galveston, 1996.

[64]Roshni Bhattacharya, Peter W Rose, Stephen K Burley, and Andreas Prlić. Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One*, 12(3):e0171355, 2017.

[65]D Verma, S Kapoor, S Das, and K Thakur. Potential inhibitors of sars-cov-2 main protease (mpro) identified from the library of fda approved drugs using molecular docking studies. preprints 2020, 202004, 2020.

[66]Rodrigo RR Duarte, Dennis C Copertino Jr, Luis P Iñiguez, Jez L Marston, Douglas F Nixon, and Timothy R Powell. Repurposing fda-approved drugs for covid-19 using a data-driven approach. *ChemRxiv*, 2020.

[67]Murugan Natarajan Arul, Sanjiv Kumar, Jeyaraman Jeyakanthan, and Vaibhav Srivastav. Searching for target-specific and multi-targeting organics for covid-19 in the drugbank database with a double scoring approach. *arXiv*, 2020.

[68]James M Sanders, Marguerite L Monogue, Tomasz Z Jodlowski, and James B Cutrell. Pharmacologic treatments for coronavirus disease 2019 (covid-19): a review. *Jama*, 323(18):1824–1836, 2020.

[69]John H Beigel, Kay M Tomashek, Lori E Dodd, Aneesh K Mehta, Barry S Zingman, Andre C Kalil, Elizabeth Hohmann, Helen Y Chu, Annie Luetkemeyer, Susan Kline, et al. Remdesivir for the treatment of covid-19—preliminary report. *New England Journal of Medicine*, 2020.

[70]Hongchao Pan, Richard Peto, Quarraisha Abdool Karim, Marissa Alejandria, Ana Maria Henao Restrepo, César Hernández García, Marie Paule Kieny, Reza Malekzadeh, Srinivas Murthy, Marie-Pierre Preziosi, et al. Repurposed antiviral drugs for covid-19; interim who solidarity trial results. *medRxiv*, 2020.

[71]David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, et al. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, pages 1–13, 2020.

[72]Sepideh Sadegh, Julian Matschinske, David B Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim Daniel Rose, et al. Exploring the sars-cov-2 virus-host-drug interactome for drug repurposing. *arXiv preprint arXiv:2004.12420*, 2020.

[73]Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 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., 2017.