Drug Repositioning for Coronavirus (COVID-19) using Different Deep Learning Architectures

Afaf A. Morad[1], Mohamed A. Aborizka[2], Fahima A. Maghraby[3]

Arab Academy for Science Technology & Maritime Transport-College of Computing and Information Technology,
Heliopolis, Cairo Governorate, Egypt[1, 2, 3]

Abstract—In December 2019, the COVID-19 epidemic was found in Wuhan, China, and soon hundreds of millions were infected. Therefore, several efforts were made to identify commercially available drugs to repurpose them against COVID-19. Inferring potential drug indications through computational drug repositioning is an efficient method. The drug repositioning problem is a top-K recommendation function that presents the most likely drugs for specific diseases based on drug and disease-related data. The accurate prediction of drug-target interactions (DTI) is very important for drug repositioning. Deep learning (DL) models were recently exploited for promising DTI prediction performance. To build deep learning models for DTI prediction, encoder-decoder architectures can be utilized. In this paper, a deep learning-based drug repositioning approach is proposed, which is composed of two experimental phases. Firstly, training and evaluating different deep learning encoder-decoder architecture models using the benchmark DAVIS Dataset. The trained deep learning models have been evaluated using two evaluation metrics; mean square error and the concordance index. Secondly, predicting antiviral drugs for Covid-19 using the trained deep learning models created during the first phase. In this phase, these models have been experimented to predict different antiviral drug lists, which then have been compared with a recently published antiviral drug list for Covid-19 using the concordance index metric. The overall experimental results of both phases showed that the most accurate three deep learning compound-encoder/protein-encoder architectures are Morgan/AAC, CNN/AAC, and CNN/CNN with best values for the mean square error, the first phase concordance index, and the second phase concordance index.

Keywords—Antiviral drugs; computational drug repositioning; coronavirus; deep learning; drug-target interactions

I. INTRODUCTION

Since December 2019, Coronavirus disease (COVID-19) has become a crucial public issue across the world. There is a real need to develop antiviral drugs for COVID-19 to stop viral infections. Recent efforts have been carried out to design novel inhibitors or utilize a drug repurposing strategy to determine anti-COVID-19 drugs that can serve as promising inhibitors versus coronavirus protease [1, 2].

Drug discovery and development is a time-consuming, complicated and costly task, including the identification of candidates, synthesis, characterization, screening, assays for therapeutic efficacy, and clinical trials [3, 4, 5]. However, drug development success rates are extremely low. In clinical phrases, numerous investigational drugs have failed due to insufficient achievement, safety concerns, or commercial purposes [6]. Alternative drug development strategy drug repositioning seeks to identify novel uses for present drugs and can decrease the risk and costs associated with the development of new drugs [7, 8]. Inferring potential drug indications via computational drug repositioning is an efficient method. The drug repositioning problem is a top-K recommendation function that presents the most probable drugs for certain diseases based on drug and disease-related data. For drug repositioning, it is crucial to accurately predict drug-target interactions (DTI), which define the binding of substances to protein targets [9]. The precise identification of molecular drug targets is essential for drug discovery and development [10, 11] and is particularly important for discovering effective and safe treatments for new pathogens, such as SARS-CoV-2 [12].

Diverse issues in bioinformatics and cheminformatics applications [13], and more specifically, drug development and discovery [14], have been successfully solved using deep learning (DL) techniques [15, 16]. By comparing DL techniques to traditional machine learning (ML) algorithms, DL algorithms designed to predict drug-target binding affinities (DTBA) occasionally do better [17]. These DL-based DTBA prediction algorithms differ in two key aspects. The representation of the input data is the first aspect. The input drug features can include, for instance, extended connectivity fingerprint (ECFP), ligand maximum common substructure (LMCS), simplified molecular input line entry system (SMILES), or a combination of these features. The second aspect relates to the DL system design created using various neural network (NN) types [18]. The construction of the many NN types varies and may include a number of layers, hidden units, filter sizes, or an integrated activation function. Each variety of NN has particular advantages that make them better suited for particular applications. These variant types of NN are Feedforward Neural Networks; FNN, Radial Basis Function Neural Networks; RBNN, Multilayer Perceptron; MLP, Recurrent Neural Networks; RNN, Convolutional Neural Networks; CNN, and Modular Neural Networks; MNN, etc. [19].

Deep learning (DL) models for predicting drug-target interactions (DTI) often use encoder-decoder architectures [20]. The encoder converts ligand or protein representations into numerical vectors for training or evaluating the DL model. There are many different encoder-decoder architectures available for DTI prediction, but only a few have been explored in previous research. In this study, we propose a DL-based method for drug repositioning, which involves two experimental phases. Firstly, training and evaluating different

deep network architectures of compound encoders (Morgan, CNN, CNN_RNN, MPNN, Transformer) and protein encoders (Transformer, AAC, CNN, CNN_RNN, Conjoint_triad) using the benchmark DAVIS Dataset [21]. The trained models have been evaluated using the mean square error and the concordance index metrics. Based on the experimental evaluation of this phase, the most superior five DL compound-encoder/protein-encoder architectures are Morgan/Transformer, Morgan/AAC, CNN/AAC, Morgan/CNN, and CNN/CNN. Secondly, predicting antiviral drugs for Covid-19 using the trained models. In this phase, the trained models created in the first phase have been experimented with to predict different antiviral drug lists for Covid-19. These lists have been evaluated by comparing them to a recently published antiviral drug list for Covid-19 [22] using the concordance index metric. The experimental evaluation of this phase showed that the most superior five DL compound-encoder/protein-encoder architectures are Morgan/Conjoint_triad, Morgan/AAC, CNN/AAC, CNN/CNN, and CNN/CNN_RNN. The overall experimental evaluation of both phases showed that the most accurate three DL compound-encoder/protein-encoder architectures are Morgan/AAC, CNN/AAC, and CNN/CNN with best values for the mean square error, the first phase concordance index, and the second phase concordance index. To summarize, there are two main contributions to this paper:

- Training and evaluating different twenty-one deep network architectures of compound encoders and protein encoders for drug repositioning using the benchmark DAVIS Dataset.

- Predicting antiviral drug lists for Covid-19 using the trained twenty-one models, and then comparing them to a recently published antiviral drug list for Covid-19.

The rest of this paper is organized in the following manner. While Section II provides an overview of the key scientific concepts, Section III reviews the related work. The conceptual model, system architecture, used benchmark DAVIS dataset, and reference Coronavirus antiviral medicine list are all described in Section IV along with the suggested technique. Section V describes the tools and libraries exploited in implementation and then presents the experimental evaluation using two main experiments: Evaluating the trained models using the DAVIS dataset and predicting antiviral drugs for Covid-19 using the trained models. Finally, the paper is concluded in Section VI.

II. BACKGROUND

DTIs have a very important function in the drug discovery procedure. DTIs recognize the interaction sites between protein targets and drug compounds and describe the attributes of the interactions sites. DTI aims in recognizing new ligands versus specified protein targets. A large number of researches have gotten advantages from recognizing DTIs containing drug repositioning [23, 24]. Costly and time-consuming laboratory tests are required to determine the affinity value for a sizable number of drug-target combinations. Therefore, computational approaches have gotten more attention in the recent years [25].

A crucial step in predicting drug-target interactions DTI is feature extraction (also called feature encoding) from the input data. Feature extraction obtains useful, discriminating, and non-trivial information from input data to facilitate subsequent learning phases. Fig. 1 illustrates the two types of feature encoding techniques: data-driven and non-data-driven. The main distinction between these two groupings is that data-driven approaches develop characteristics for each input automatically. In strategies that are not data-driven, features are calculated in a fixed way for each input. The data-driven approaches are mainly based on deep learning methods, which is a set of machine learning algorithms that uses a model of the human visual system to create new hierarchical feature representations [26].
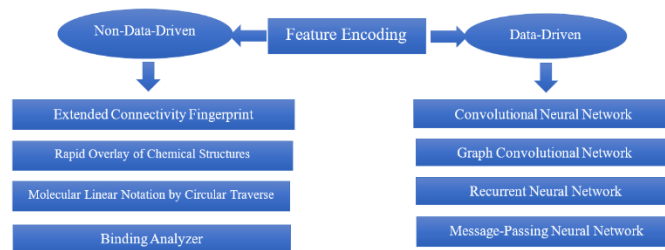

Fig. 1. Feature encoding methods [26].

A neural network with two or more hidden layers is deemed a deep learning. The input layer receives the input features directly, while the output layer generates predictions through a series of non-linear transformations utilizing hidden layers. Each output node corresponds to a class-based prediction task. If there is only one node in the output layer, then the network is considered a single-task deep learning. Otherwise, it is known as a multi-task deep learning [27].

A basic sequence, a derived atomic fingerprint, or a mixture of both can constitute the input feature space for a deep network. Numerous studies have used a network's input to be a raw molecular sequence. Other works convert a raw sequence to a more appropriate form, such as one-hot coding, to feed it to deep networks like CNN. In one-hot coding, each character in the sequence is represented by a binary vector with its matching bit set to one and the other bits set to zero.

There are many feature encoding architectures in DTI prediction such as Convolutional Neural Network CNN and Message-Passing Neural Network MPNN. For example, the Transformer encoders [28] are multi-layered bidirectional Transformer encoders following the initial Transformer model. The Transformer encoder is capable of modelling a sequence without the aid of a CNN or RNN. Transformer, in contrast to these earlier sequence processing layers, can efficiently encode the relationship between far-flung tokens (atoms) in a sequence. Various Transformer-based NLP models surpass earlier techniques in many benchmarks due to this effective context modelling.

Molecular descriptors must be created from symbolic representations of molecules, such as the SMILES (Simplified Molecular Input Line Entry System) format, in order to do deep learning. "Morgan Fingerprints," a vector representation of molecular attributes, is a commonly adopted way to describe

a molecule. Morgan fingerprint, also known as extended-connectivity fingerprint (ECFP) [29], is often used algorithm, or encoder specification. ECFPs are innovative category of topological fingerprints for molecular characterization. For the purpose of modelling structure-activity, ECFPs were created. ECFPs are circular fingerprints with a variety of useful properties: their ability to be calculated quickly; they are not predefined and can represent an essentially infinite number of different molecular features; and their features indicate the presence of specific substructures, which makes it simpler to interpret analysis results.
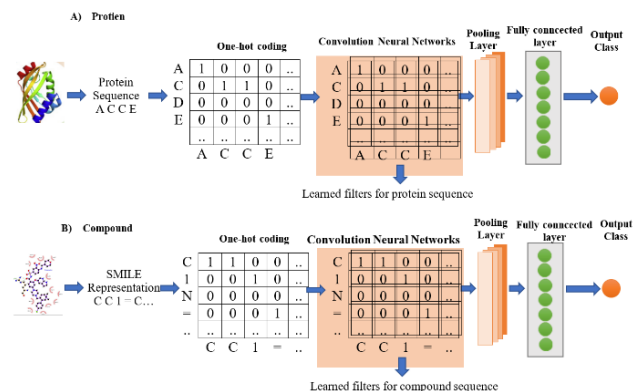

Fig. 2. One-dimension convolutional layer operation on a protein sequence and a molecule sequence [26].

The most popular deep neural network encoder is CNN, which operates on a grid data structure like digital images. As depicted in Fig. 2 [26], CNN consists of multiple convolutional and pooling layers arranged in an arbitrary order. Convolutional layers discover a series of filters that derive a group of local patterns from a specific receptive field of the layer input. In subsequent convolutional layers, the receptive field also expands. By down-sampling the layer's input, the pooling layer expands the receptive field. Moreover, the pooling layer does not define any additional parameters. The CNN input might consist of one-dimensional or two-dimensional matrices that are scanned along the sequence in only one or two directions, respectively. Until now, 1D CNNs have been used for DNA sequences for categorization, DNA-protein binding, and motif extraction, among other tasks [30, 31, 32].

In CNN, Fig. 2 depicts how to apply the one-dimension convolutional layer on a small molecule sequence or a protein sequence. In Fig. 2A, a protein sequence is depicted as an amino acids' sequence contained in a matrix where every amino acid is encoded by a single-hot code (all bits are zeros except the corresponding bit of the symbol is one). As seen, by reducing the loss function on both positive and negative task samples, the filter is moved along the sequence. Fig. 2B depicts the SMILE sequence as a string of characters. Every character represents an atom or molecule structural indicator. Every character is then encoded with a one-hot code and inserted in each matrix column. In both instances, the learned filter is displayed.

III. RELATED WORK

Deep learning methods are now widely used in many research fields like; speech recognition [33, 34, 35] and image processing [36, 37, 38], involving bioinformatics such as genomics works [39,40] and quantitative-structure activity relationship (QSAR) researches in drug discovery [41]. The primary benefit of deep learning architectures aims to offer improved raw data representations through non-linear modifications in every layer [42]; hence facilitates the learning of the data's hidden patterns.

Several studies utilizing Deep Neural Networks (DNN) for the prediction of DTI binary class with various input models for drugs and proteins have been conducted previously [43,44,45], as well as a few studies utilizing stacked auto-encoders [46] and deep belief networks [47]. Likewise, stacked auto-encoder-based models using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were used to describe genomic and chemical structures as real-valued vectors [48, 49]. The protein-ligand interaction scoring is performed by using deep learning methods. The protein-ligand interaction scoring commonly utilizes CNNs, which learn from the three-dimensional compositions of the protein-ligand complexes [50, 51, 52, 53].

Pahikkala et al. [54] used a method called KronRLS (Kronecker Regularized Least Squares), which only requires two dimension-based chemical similarity-based representations of the medications and the Smith-Waterman similarity representation of the targets. Latterly, the SimBoost approach was suggested to forecast the scores of binding affinity with a gradient boosting machine by utilizing feature engineering to provide DTI [55]. They exploited similarity-based information of DT pairings and attributes acquired from the pairs' network-based interactions. Both kinds of research obtained similarity-based information using 2D representations of the substances and typical machine learning algorithms.

The ensemble of deep learning models (EnsembleDLM) for DTI prediction was presented [56]. EnsembleDLM utilizes a set of chemical compounds and proteins and assembles the predictions from numerous deep neural networks. It provides good achievements in cross-domain applications spanning various bio-activity types and protein classes. By using transfer learning, the EnsembleDLM obtained a good performance (Pearson correlation coefficient and concordance index).

To construct negative DTIs, a new similarity-based strategy was presented [57]. Multiple least absolute shrinkage and selection operator (LASSO) models were presented to incorporate various collections of feature sets in order to investigate the strength of prediction and forecast DTIs. In addition, LASSO Deep Neural Network (LASSO-DNN) model was developed to predict DTIs based on the features retrieved from the LASSO models with the highest achievement. LASSO-DNN was compared to LASSO, standard logistic (SLG) regression, support vector machine (SVM), and conventional DNN models. The LASSO-DNN outperformed the SLG, LASSO, SVM, and regular DNN models, as demonstrated by the results of the experiments.

Wang et al. [58] provided the three-step strategy for identifying obscure DTIs using deep learning. The first step is an illustration of drug-target pairings, where the drug compounds are encoded as fingerprint characteristics. In contrast, the protein sequence features are produced by applying Legendre Moments (LM) to a position-specific scoring matrix (PSSM), including evolutionary protein information. The second step involves compression and fusion of features. The sparse principal component analysis (SPCA) was utilized to reduce the dimension and redundancy of the features.

Eventually, the prediction task that used the deep long short-term memory (DeepLSTM) model was exploited. The experimental results proved that the suggested technique outperforms other DTI prediction methods.

DTI prediction model with 2D paired distance maps of proteins and molecular graphs as inputs for targets and medicines was presented [59]. To retrieve the interactive effects of targets and drugs, the mutual interaction neural network (MINN) by integrating two interacting transformers (Interformer, for short) with an enhanced communicative message passing neural network (CMPNN) (titled Inter-CMPNN) was proposed.

In conclusion, the majority of computational algorithms suggested forecasting DT interactions concentrated on binary classification, wheresoever the primary objective is to assess whether or not a drug-target pair interacts. Yet, the interactions of protein-ligand suppose a continuum of binding affinity values, often known as binding strength. Increased availability of affinity information in drug-target KBs (knowledge bases) enables the application of sophisticated learning approaches like deep learning architectures to predict binding affinities. The main contribution of this research is to train and evaluate twenty-one different deep network architectures of compound encoders and protein encoders for drug repositioning.

IV. RESEARCH METHODOLOGY

A. Proposed Approach

Fig. 3 shows the proposed approach conceptual model, which exploits a benchmark dataset called DAVIS [21] and a reference Coronavirus antiviral drug list [22], and then outputs a ranked list of candidate drugs for Coronavirus for each trained deep learning model. According to various compound-decoders/protein-decoders, twenty-one DL models have been trained.
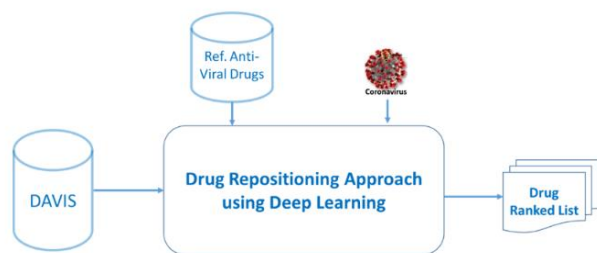

Fig. 3. The proposed approach conceptual model.

This research work has two main objectives. Firstly, training and evaluating models of different deep learning compound-encoders and protein-encoders using the benchmark DAVIS Dataset. Secondly, predicting antiviral drugs for Covid-19 using the trained models. To achieve these objectives, Fig. 4 shows the system architecture, which includes two phases: the training phase and the prediction phase. In the training phase, both drug compounds and disease target proteins of the DAVIS dataset are encoded using different neural network architectures and then their embeddings are decoded to generate the trained model. For example, as shown in Fig. 4, the drug compound is encoded using the Transformer architecture, and the disease target protein is encoded using the CNN architecture. The procedure of converting ligand or protein representations into numerical vectors utilized to train or assess a model of machine learning is known as encoding. Table I and Table II show the compound-encoders and protein-encoders exploited in this research, respectively. The cross-validation technique was applied on all DAVIS dataset instances.
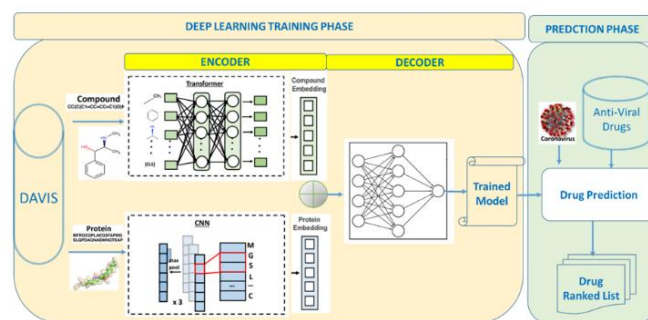

Fig. 4. The system architecture.

TABLE I. THE COMPOUND-ENCODERS AND THEIR DESCRIPTIONS

Compound Encoder	Description
Morgan	Extended-Connectivity Fingerprints
CNN	Convolutional Neural Network on SMILES
CNN_RNN	A GRU/LSTM on top of a CNN on SMILES
MPNN	Message-passing neural network
Transformer	Transformer Encoder on ESPF

TABLE II. THE PROTEIN-ENCODERS AND THEIR DESCRIPTIONS

Protein Encoder	Description
Morgan	Transformer Encoder on ESPF
CNN	Amino acid composition up to 3-mers
CNN_RNN	Convolutional Neural Network on target seq
MPNN	A GRU/LSTM on top of a CNN on target seq
Transformer	Conjoint triad features

In the prediction phase, the proposed solution exploits the trained model to predict a ranked list of drugs candidate for Coronavirus. The predicted ranked list of drugs is evaluated versus a reference list of antiviral drugs published recently for

Coronavirus. Fig. 5 shows how different trained models are generated and how different deep learning architectures are exploited. For example, the trained model 1 was generated by encoding drug compounds and disease target proteins using DNN and CNN, respectively. In the same way, the trained model 2 was generated by encoding drug compounds and disease target proteins using DNN and Transformer, respectively. In this research, twenty-one combinations of neural network architectures were experimented.
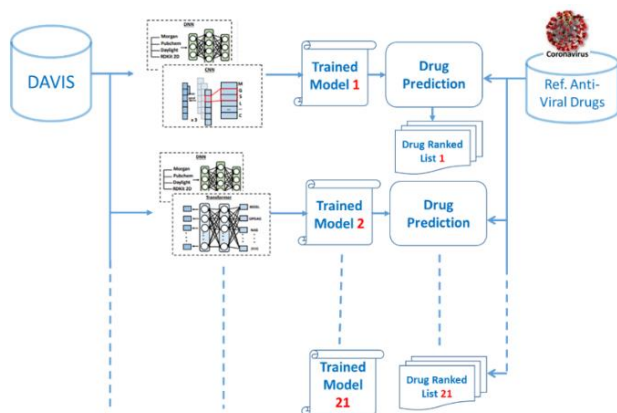

Fig. 5. Predicting twenty-one different drug-ranked lists for coronavirus.

B. The Benchmark Dataset

The selectivity assays of the set of kinase proteins and the pertinent inhibitors with their definite dissociation constant (Kd) values are included in the DAVIS dataset [21]. The DAVIS dataset compound SMILES strings were retrieved from the Pubchem compound database using their Pubchem CIDs [60]. The DAVIS dataset's protein sequences were retrieved from the UniProt protein repository using gene names and RefSeq accession codes [61]. Table III shows the number of Drug-Target Interaction Instances.

TABLE III. DAVIS DATASET SUMMARY

Proteins	Compounds	Interactions
424	68	30056

C. The Reference Coronavirus Antiviral Drug List

Recently, [22] used virtual screening and molecular docking methods to locate prospective inhibitors from existing drugs that can respond to COVID-19. Based on their binding energy (kcal/mol), the authors ranked 121 drugs as potent drugs against SARS-CoV-2 since they tightly bind to their main protease. Table IV presents a list of ranked Drugs based on their docking scores Binding energy (BE).

V. RESULTS AND DISCUSSION

A. Implementation Setup

In this research, the DeepPurpose DTI prediction tool was exploited. This tool provides a set of eight ligand encoders that can be combined with other protein representations and architectures to generate new models [62]. The DeepPurpose tool contains seven distinct protein encoders, which can be divided into two distinct categories: expert-designed

algorithms and text-processing based on neural networks. Among algorithmic encoders, the ACC encoder produces a vector of 8420 elements describing the frequency of all amino acid k-mers for k values up to three [63]. In contrast, the conjoint triad encoder offers a 3-mer frequency count utilizing a restricted amino acid alphabet [64]. Neural network encoders, in contrast, apply directly on the sequence.

TABLE IV. A LIST OF RANKED DRUGS BASED ON THEIR DOCKING SCORES BINDING ENERGY (BE) [22]

Drug	BE	Drug	BE
Beclabuvir	-10.4	Pibrentasvir	-6.9
Nilotinib	-9.9	Valaciclovir	-6.9
Tirilazad	-9.6	Valganciclovir	-6.9
Paritaprevir	-9.2	Pirodavir	-6.8
Raltegravir	-9.1	Vidarabine	-6.8
Venetoclax	-9	Zanamivir	-6.8
Bictegravir	-8.9	Daclatasvir	-6.7
Danoprevir	-8.9	Alovudine	-6.6
Pimodivir	-8.8	Entecavir	-6.6
Voxilaprevir	-8.7	Famciclovir	-6.6
Faldaprevir	-8.6	Idoxuridine	-6.6
Setrobuvir	-8.6	Laninamivir	-6.6
Letermovir	-8.5	Tenofovir	-6.6
Bisantrene	-8.4	Cidofovir	-6.5
Rilpivirine	-8.4	Dasabuvir	-6.5
Indinavir	-8.2	Didanosine	-6.5
Ombitasvir	-8.2	Efavirenz	-6.5
Saquinavir	-8.2	Fiacitabine	-6.5
Simeprevir	-8.2	Lobucavir	-6.5
Digitoxin	-8.1	Penciclovir	-6.5
Dolutegravir	-8.1	Telbivudine	-6.5
Remdesivir	-8.1	Tenofovir disoproxil	-6.5
Lopinavir	-8	Umifenovir	-6.5
Maraviroc	-8	Capravirine	-6.4
Calanolide A	-7.9	Grazoprevir	-6.4
Darunavir	-7.9	Interferon alfa-2b	-6.4
Elbasvir	-7.9	Peramivir	-6.4
Etravirine	-7.9	Sorivudine	-6.4
Velpatasvir	-7.9	Zidovudine	-6.4
Vesatolimod	-7.9	Boceprevir	-6.3
Digoxin	-7.8	Emivirine	-6.3
Tipranavir	-7.8	Gancicolovir	-6.3
Amprenavir	-7.7	Ribavirin	-6.3
Maribavir	-7.7	Stavudine	-6.3

Nelfinavir	-7.7	Cenicriviroc	-6.2
Ruzasvir	-7.7	Dexelvucitabine	-6.2
Sofosbuvir	-7.7	MK-0608	-6.2
Doravirine	-7.6	Oseltamivir	-6.2
Mericitabine	-7.6	R1479	-6.2
Ritonavir	-7.6	Taribavirin	-6.1
Atazanavir	-7.5	Chloroquine	-6
Dapivirine	-7.5	Hydroxychloroquine	-6
Delavirdine	-7.5	Ledipasvir	-6
Fosamprenavir	-7.5	Elvucitabine	-5.9
Asunaprevir	-7.4	Emtricitabine	-5.9
Baloxavir marboxil	-7.4	Tromantadine	-5.9
Elvitegravir	-7.3	Acyclovir	-5.8
Sovaldi	-7.3	PSI-6130	-5.8
Tenofovir alafenamide	-7.3	Triazavirin	-5.8
Vedroprevir	-7.3	Aciclovir	-5.7
Galidesivir	-7.2	Fingolimod	-5.7
Methylprednisolone	-7.2	Lamivudine	-5.7
Clevudine	-7.1	Zalcitabine	-5.7
Telaprevir	-7.1	Ingavirin	-5.6
Thalidomide	-7.1	Rimantadine	-4.6
Abacavir	-7	Favipiravir	-4.4
Adefovir dipivoxil	-7	Amantadine	-4.3
Cobicistat	-7	Docosanol	-4.3
Nevirapine	-7	Foscarnet	-4.3
Tenofovie alafenamide	-7		

Every amino acid is transformed into a numerical value as a fixed-length one-dimension array using the DeepPurpose CNN encoder. Subsequently, it employs a convolutional neural network [65] to learn spatial information from the sequence (local amino acid neighborhoods) that may be pertinent to the DTI binding model.

Fig. 6 shows different compound-encoders and different protein-encoders developed by the DeepPurpose DTI prediction tool [60].
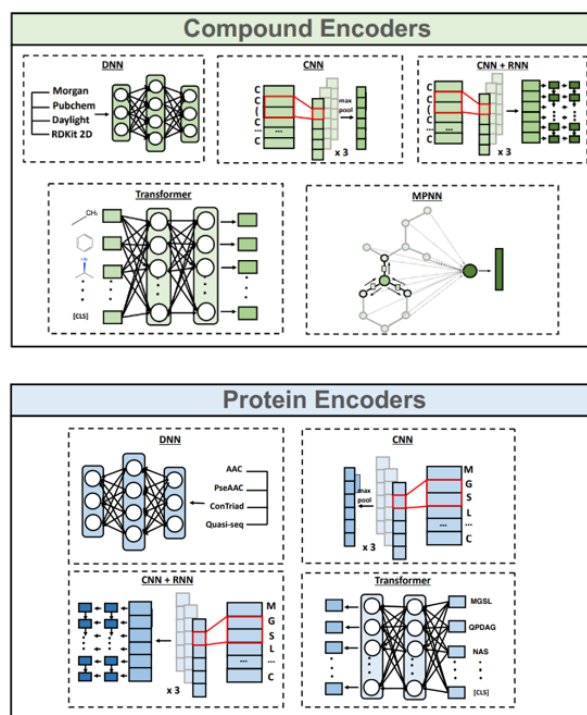

Fig. 6. Compound-encoders and protein-encoders developed by the DeepPurpose tool [62].

The DeepPurpose DTI prediction tool enables developers to set some pamperers for the model training phase. Table V shows some parameter settings for NN architectures, which had been used during the model training phase.

TABLE V. PARAMETER SETTINGS FOR NN-BASED TRAINED MODEL

Parameter	Range
Filter length (compounds)	[32,64,96]
Filter length (proteins)	[32,64,96]
Hidden neurons	1024; 1024; 512
Batch size	256
Epoch	100
Learning rate	0.001
CNN Drug Kernels	[4,8,12]
CNN Target Kernels	[4,8,12]

B. Evaluation Metrics

1) Mean squared error (MSE): During training, a learning model attempts to reduce the gap between the actual (real) value and the prediction. The mean squared error (MSE) is picked as the loss function because a regression problem is used. Equation 1 shows how the MSE is calculated.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (P_i - Y_i)^2 \quad (1)$$

Where P is the vector of predictions, Y is the vector of actual outputs, and n represents the total number of samples.

2) Concordance index (CI): The Concordance Index (CI) was used to evaluate the effectiveness of a model that outputs continuous values [17]. CI determines whether the anticipated binding affinity values of two random drug–target pairs were predicted in the same sequence as their true values were. Equation 2 shows how the CI is calculated.

$$CI = \frac{1}{Z} \sum_{\delta_i > \delta_j} h(b_i - b_j)$$

$$h(x) = \begin{cases} 1, & if\ x > 0 \\ 0.5, & if\ x = 0 \\ 0, & if\ x < 0 \end{cases} \quad (2)$$

Where b_i represents the prediction value for the greater affinity δ_i, b_j represents the prediction value for the lesser affinity δ_j, Z is a normalization constant, and h(x) is the step function.

C. Phase 1: Evaluating the Trained Models using the Davis Dataset

Based on exploiting different deep learning compound-encoders/protein-encoders, twenty-one trained models were evaluated using the benchmark DAVIS Dataset. The cross-validation technique was applied. Both Fig. 7 and Fig. 8 show the evaluation results for the different twenty-one trained models in terms of MSE (Mean Square Error) and CI (Concordance Index) evaluation metrics. It is noted that the most superior five DL compound-encoder/protein-encoder architectures are Morgan/Transformer, Morgan/AAC, CNN/AAC, Morgan/CNN, and CNN/CNN.
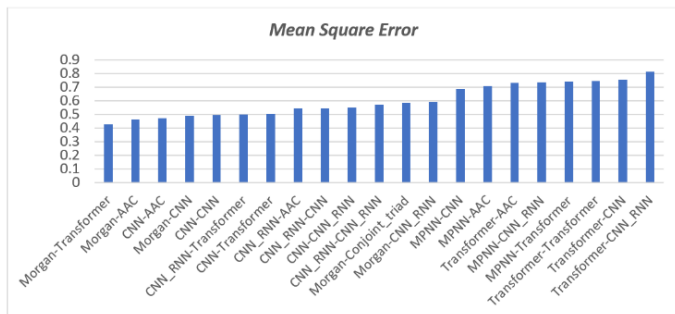

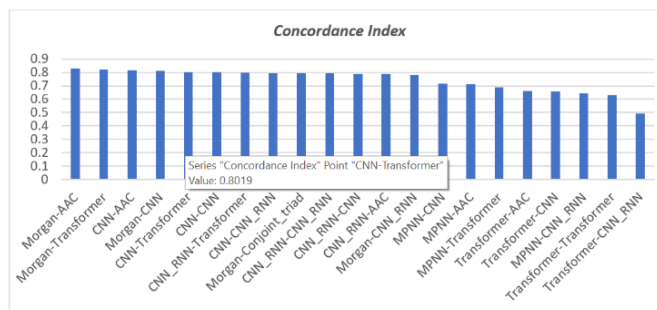Fig. 7. The MSE evaluation results for the different twenty-one trained models.

Fig. 8. The CI evaluation results for the different 21 trained models.

D. Phase 2: Predicting Antiviral Drugs for Covid-19 using the Trained Models

To evaluate the predicting accuracy for the trained models versus the 121-reference antiviral ranked drugs, the Concordance Index was used to compare each predicted ranked list versus the reference antiviral ranked drugs. Fig. 9 shows the Concordance Index calculating algorithm for a Predicted Drug List versus the Covid-19 Drug Reference List. Fig. 10 shows the Concordance Index for the twenty-one Predicted Drug Lists versus the Covid-19 Drug Reference List. As shown in Fig. 10, the most superior five DL compound-encoder/protein-encoder architectures are Morgan/Conjoint_triad, Morgan/AAC, CNN/AAC, CNN/CNN, and CNN/CNN_RNN.

The overall experimental evaluation for the two experimental phases can be summarized in Fig. 11, which shows the evaluation results using the CI metric for the different twenty-one trained models in the model testing phase using the DAVIS dataset and in the predicting phase versus the coronavirus drug reference list. As shown in Fig. 12, it is worth noting that the most accurate three deep learning compound-encoder/protein-encoder architectures are Morgan/AAC, CNN/AAC, and CNN/CNN with best values for the mean square error, the first phase concordance index, and the second phase concordance index.

Input: Drug Reference List (DR)
        Predicted Drug List (P)
Output: CI
Select all pairs from DR -> D, where DRi > DRj
Let S = 0
For each pair in D (Di , Dj) Do
    Begin
            If (Pi > Pj)
                    Add 1 to S
            Else If (Pi = Pj)
                    Add 0.5 to S
    End
Let CI = S / No. of D pairs

Fig. 9. The concordance index calculating algorithm for a predicted drug list versus the Covid-19 drug reference list.
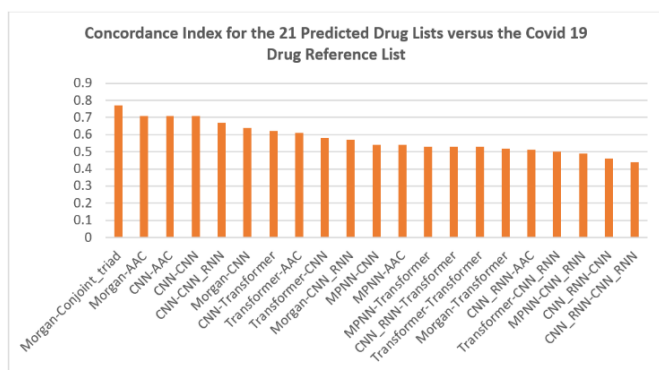
Fig. 10. The concordance index for the twenty-one predicted drug lists versus the Covid-19 drug reference list.
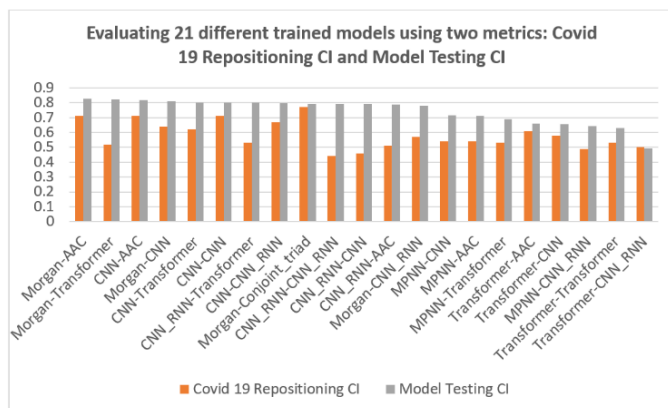

Fig. 11. Evaluating twenty-one different trained models using two metrics: Covid-19 repositioning CI and model testing CI.
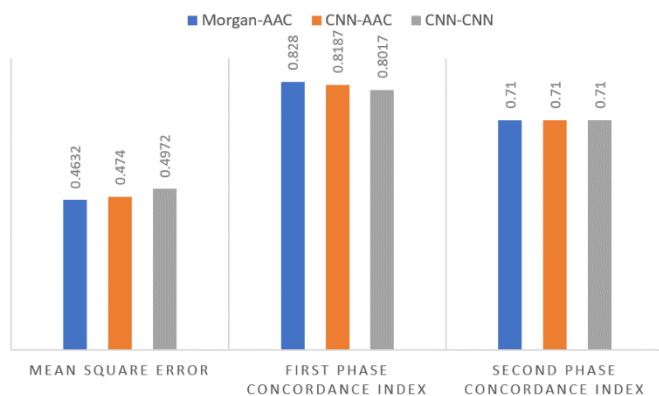

Fig. 12. The most accurate three deep learning compound-encoder / protein-encoder architectures: Morgan / AAC, CNN / AAC, and CNN / CNN.

VI. CONCLUSION

This research proposed a deep learning-based drug repositioning method to train and evaluate twenty-one models based on deep learning compound-encoders and protein-encoders. The trained models have been evaluated using two experiments. Firstly, testing the trained models by applying the cross-validation technique on the benchmark DAVIS Dataset. Secondly, comparing the predicted antiviral drug lists by the trained models to a recently published antiviral drug list for Covid-19. The experimental evaluation showed the most accurate three deep learning compound-encoder/protein-

encoder architectures are Morgan/AAC, CNN/AAC, and CNN/CNN with best values for the mean square error, the first phase concordance index, and the second phase concordance index. As a future work, the same 21 different deep network architectures of protein and compound encoders are suggested to be trained and to be evaluated using other datasets and other viral diseases.

REFERENCES

[1] R. J. Khan, R. Jha, G. M. Amera, M. Jain, E. Singh, A. Pathak, and A. Singh, "Targeting novel coronavirus 2019: A systematic drug repurposing approach to identify promising inhibitors against 3C-like proteinase and 2'-O-ribose methyltransferase", Journal of Biomolecular Structure and Dynamics, vol. 2, pp. 1–40, 2020.

[2] P. Sarma, N. Sekhar, M. Prajapat, P. Avti, H. Kaur, S. Kumar, and B. Medhi, "In-silico homology assisted identification of inhibitor of RNA binding against 2019-nCoV N-protein (N terminal domain)", Journal of Biomolecular Structure & Dynamics, vol. 39, no. 8, 2021.

[3] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C.C. Persinger, B.H. Munos, S.R. Lindborg, and A.L. Schacht, "How to improve R&D productivity: the pharmaceutical industry's grand challenge", Nature reviews Drug discovery, vol. 9, no. 3, pp. 203-214, 2010.

[4] N. Prakash and P. Devangi, "Drug Discovery", Journal of Antivirals & Antiretrovirals, vol. 2, no. 4, pp. 63-68, 2010.

[5] H. Xue, J. Li, H. Xie, and Y. Wang, "Review of drug repositioning approaches and resources", International journal of biological sciences, vol. 14, no. 10, pp. 1232-1244, 2018.

[6] T. J. Hwang, D. Carpenter, J. C. Lauffenburger, B. Wang, , J. M. Franklin, and A. S. Kesselheim, "Failure of investigational drugs in late-stage clinical development and publication of trial results", JAMA internal medicine, vol. 176, no. 12, pp. 1826- 1833, 2016.

[7] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, and A. Norris, "Drug repurposing: progress, challenges and recommendations", Nature reviews Drug discovery, vol. 18, no. 1, pp.41-58, 2019.

[8] G. Jin and S. T. C. Wong, "Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines", Drug discovery today, vol. 19, no. 5, pp. 637-644, 2014.

[9] R. Santos, O. Ursu, A. Gaulton et al., "A comprehensive map of molecular drug targets", Nature reviews Drug discovery, vol. 16, no. 1, pp. 19–34, 2017.

[10] A. Rutkowska, D. W. Thomson, J. Vappiani et al., "A Modular Probe Strategy for Drug Localization, Target Identification and Target Occupancy Measurement on Single Cell Level", ACS Chemical Biology, vol. 11, no. 9, pp. 2541–2550, 2016.

[11] M. Zitnik, F. Nguyen, B. Wang et al., "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities", Information Fusion, vol. 50, pp. 71–91, 2019.

[12] T. P. Velavan and C. G. Meyer, "The COVID19 epidemic", Tropical Medicine & International Health, vol. 25 , no. 3, pp. 278–280, 2020.

[13] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, and X. Gao, "Deep learning in bioinformatics: introduction, application, and perspective in the big data era", Methods, vol. 166, pp. 4–21, 2019.

[14] M. Karimi, D. Wu, Z. Wang, and Y. Shen, "DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks", Bioinformatics, vol. 35, no. 18, pp. 3329–3338, 2019.

[15] S.W. Naidoo, N. Naicker, S.S. Patel, and P. Govender, "Computer vision: the effectiveness of deep learning for emotion detection in marketing campaigns", International Journal of Advanced Computer Science and Applications, Vol. 13, no. 5, 2022.

[16] J. Kalezhi, M. Chibuluma, C. Chembe, V. Chama, F. Lungo, and D. Kunda, "A Cross Platform Contact Tracing Mobile Application for COVID-19 Infections using Deep Learning", International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, 2022.

[17] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction", Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.

[18] S. Krig, "Feature learning and deep learning architecture survey," in Computer Vision Metrics (Cham: Springer), 375–514, 2016.

[19] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications", Neurocomputing, vol. 234, pp. 11–26, 2017.

[20] K. Cho, B. v. Merrienboer, D. Bahdanau et al., "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches", In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation at ACL, pages 103–111, Doha, Qatar, 2014.

[21] M. I. Davis, J. P. Hunt, S. Herrgard et al., "Comprehensive analysis of kinase inhibitor selectivity", Nature Biotechnology, vol. 29, no. 11, 1046–1051, 2011.

[22] C. Samir, B. Assia, A. Adnane et al., "Discovery of potent SARS-CoV-2 inhibitors from approved antiviral drugs via docking and virtual screening", Combinatorial chemistry & high throughput screening, vol. 24, no. 3, pp.441-454, 2021.

[23] A. Masoudi, Z. Mousavian, and J. H. Bozorgmehr, "Drugtarget and disease networks: polypharmacology in the postgenomic era", vol. 17, 2013.

[24] Y. M. Sobhanzadeh, Y. Omidi, M. Amanlou et al., "Drug databases and their contributions to drug repurposing". Genomics, vol. 112, no. 2, pp. 1087-1095, 2020.

[25] A. Ezzat, M. Wu, X.-L. Li et al., "Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey", Briefings in Bioinformatics, vol. 20, no. 4, pp. 1337–1357, 2019.

[26] A. Karim, R. Parvin, P. Antti et al., "Deep learning in drug target interaction prediction: current and future perspectives". Current Medicinal Chemistry, vol. 28, no. 11, pp. 2100-2113, 2021.

[27] A. S. Rifaioglu, E. Nalbat, V. Atalay et al., "DEEPScreen: high performance drug–target interaction prediction with convolutional neural networks using 2-D structural compound representations", Chemical Science, vol. 11, no. 9, pp. 2531-2557, 2020.

[28] B. Shin, S. Park, K. Kang et al., "Self-attention-based molecule representation for predicting drug-target interaction", 4th Machine Learning for Healthcare Conference, PMLR 106:230-248, 2019.

[29] D. Rogers and M. Hahn, "Extended-connectivity fingerprints. Journal of chemical information and modeling", vol. 50, no. 5, pp.742-754, 2010.

[30] H. Zeng, M. D. Edwards, G. Liu et al., "Convolutional neural network architectures for predicting DNA protein binding. Bioinformatics", vol. 32, no. 12, i121-i127, 2016.

[31] J. Lanchantin, R. Singh, B. Wang et al., "Deep motif dashboard: Visualizing and understanding genomic sequences using deep neural networks", Biocomputing, pp. 254-265, 2017.

[32] B. Alipanahi, A. Delong, M. T. Weirauch et al., "Predicting the sequence specificities of DNA- and RNAbinding proteins by deep learning", Nature Biotechnology, vol. 33, no. 8, pp. 831-838, 2015.

[33] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition". IEEE Transactions on Audio, Speech, and Langua.ge Processing, vol. 20, no. 1, pp. 30-42, Jan. 2012.

[34] A. Graves; A. Mohamed; G. Hinton, "Speech recognition with deep recurrent neural networks", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 6645–6649, 2013

[35] G. Hinton, L. Deng, D. Yu et al., "Deep neural networks for acoustic modeling in speech recognition" IEEE Signal Processing Magazine, vol. 29, no. 6, 82–97, 2012.

[36] D. Ciregan, U. Meier, J. Schmidhuber, "Multi-column deep neural networks for image classification". Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3642–3649, 2012.

[37] J. Donahue, Y. Jia, O. Vinyals et al., "Decaf: a deep convolutional activation feature for generic visual recognition", 31st International Conference on Machine Learning, PMLR, vol. 32, no. 1, pp. 647-655, 2014.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", 3rd International Conference on Learning Representations (ICLR), Hilton San Diego Resort & Spa, May 7–9, 2015.

[39] M. K. Leung, H. Y. Xiong, L. J. Lee et al., "Deep learning of the tissue-regulated splicing code", Bioinformatics, vol. 30, no. 12, pp. i121–i129, 2014.

[40] H. Y. Xiong, B. Alipanahi, L. J. Lee et al., "The human splicing code reveals new insights into the genetic determinants of disease", Science, vol. 347, no. 6218, 2014.

[41] J. Ma, R. P. Sheridan, A. Liaw et al., "Deep neural nets as a method for quantitative structure–activity relationships", Journal of Chemical Information and Modeling, vol. 55, no. 2, pp. 263–274, 2015.

[42] M. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey, "Deep learning of the tissue-regulated splicing code", Bioinformatics, vol. 30, no. 12, pp. i121-i129, 2014.

[43] P. W. Hu, K. C. Chan, and Z. H. You, "Large-scale prediction of drug-target interactions from eep representations", In 2016 international joint conference on neural networks (IJCNN), IEEE, pp. 1236-1243, 2016.

[44] K. Tian, M. Shao, Y. Wang et al., "Boosting compound-protein interaction prediction by deep learning", Methods, vol. 110, pp. 64-72. 2016.

[45] M. Hamanaka, K. Taneishi, H. Iwata et al., "CGBVS - DNN: Prediction of Compound - protein Interactions Based on Deep Learning", Molecular informatics, vol. 36, no. 1-2, pp.1600045, 2017.

[46] L. Wang, Z.H. You, X. Chen et al., "A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network", Journal of Computational Biology, vol. 25, no. 3, pp.361-373, 2018.

[47] M. Wen, Z. Zhang, S. Niu et al., "Deep-learning-based drug–target interaction prediction", Journal of proteome research, vol. 16 no. 4, pp.1401-1409, 2017.

[48] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud et al., "Automatic chemical design using a data-driven continuous representation of molecules", ACS central science, vol. 4, no. 2, pp. 268-276, 2018.

[49] S. Jastrzębski, D. Leśniak, and W. M. Czarnecki, "Learning to smile (s)", arXiv preprint arXiv: 1602.06289, 2016.

[50] Gomes, J., Ramsundar, B., Feinberg, E.N., and Pande, V.S., "Atomic convolutional networks for predicting protein–ligand binding affinity", arXiv preprint arXiv: 1703.10603, 2017.

[51] M. Ragoza, J. Hochuli, E. Idrobo et al., "Protein–ligand scoring with convolutional neural networks", Journal of chemical information and modeling, vol. 57, no. 4, pp. 942-957, 2017.

[52] I. Wallach, M. Dzamba, and A. Heifets, "AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery", arXiv preprint arXiv:1510.02855, 215.

[53] P. W. Rose, A. Prlić, A. Altunkaya et al., "The RCSB protein data bank: integrative view of protein, gene and 3D structural information", Nucleic acids research, p.gkw1000, 216.

[54] T. Pahikkala, A. Airola, S. Pietilä et al., "Toward more realistic drug–target interaction predictions", Briefings in bioinformatics, vol. 16, no. 2, pp. 325-337, 2014.

[55] T. He, M. Heidemeyer, F. Ban et al., "SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines", Journal of cheminformatics, vol. 9, no. 1, pp. 1-14, 2017.

[56] P. Kao, S. Kao, N. Huang, and Y. Lin, "Toward Drug-Target Interaction Prediction via Ensemble Modeling and Transfer Learning", IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Houston, TX, USA, pp. 2384-2391, 2021.

[57] Y. Jiaying, D. Robert, and P. Hu, "Predicting drug-target interaction network using deep learning model", Computational Biology and Chemistry, Vol. 80, PP. 90-101, 2019.

[58] Y. Wang, Z. You, S. Yang et al., "A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network", BMC medical informatics and decision making, vol. 20 no. 2, pp.1-9, 2020.

[59] F. Li, Z. Zhang, J. Guan, and S. Zhou, "Effective drug–target interaction prediction with mutual interaction neural network", Bioinformatics, Vol. 38, no. 14, pp. 3582–3589, 2022.

[60] E.E. Bolton, Y. Wang, P.A. Thiessen, and S.H. Bryant, "PubChem: integrated platform of small molecules and biological activities", In Annual reports in computational chemistry, vol. 4, pp. 217-241, Elsevier, 2008.

[61] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, and F. Cheng, "Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2", Cell discovery, vol. 6, no. 1, p.14, 2020.

[62] K. Huang, T. Fu, L. M. Glass et al., "DeepPurpose: A deep learning library for drug-target interaction prediction", Bioinformatics, vol. 36, no. 22-23, pp. 5545–5547, 2020.

[63] H. Huang, and X. Gong, "A review of protein inter-residue distance prediction", Current Bioinformatics, vol. 15, no. 8, pp.821-830, 2020.

[64] M. Bagherian, E. Sabeti, K. Wang, M.A. Sartor, Z. Nikolovska-Coleska, and K. Najarian, "Machine learning approaches and databases for prediction of drug–target interaction: a survey paper", Briefings in bioinformatics, vol. 22, no. 1, pp.247-269, 2021.

[65] Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks", Communications of the ACM, vol. 60, no. 6, pp. 84–90, 2017.