Machine learning for Drug-Virus Prediction

Milad Besharatifard (milad1besharati@aut.ac.ir)
Amirkabir University of Technology
Arshia Gharagozlou
pittsburgh university

Research Article

Keywords: Autoencoder, Compressed sensing, Drug, Neural Network, Random forest, Virus

Posted Date: September 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-910042/v3

Machine learning for Drug-Virus Prediction

Milad Besharatifard[1], Arshia Gharagozlou[2]

Abstract

The 2019 Coronavirus (COVID-19) epidemic has recently hit most countries hard. Therefore, many researchers around the world are looking for a way to control this virus. Examining existing medications and using them to prevent this epidemic can be helpful. Drug repositioning solutions can be effective because designing and discovering a drug can be very time-consuming. In this study, we used a binary classifier learning method to predict the drug-virus relationship. The feature vector for each drug-virus pair is based on the similarity between drugs and the similarity between viruses. We calculated the similarities between the drugs using their structural properties (fingerprint) and their phenotype. We also calculated the similarities between viruses based on their genome sequence and the vector encoded by the Biobert model. Finally, using the HDVD dataset, we formed the similarity vectors of each drug-virus pair and considered it as input to neural network and random forest models. In these models, we randomly selected 20% of the positive data and the same amount of negative data. Finally, the performance of the proposed approach for this test data is considered, after five tests, as AUC=0.97 and AUPR = 0.96. We also used the Compressed Sensing (CS) matrix factorization model to predict the drug-virus association. We also investigated the importance of drug features in predicting drug-virus association by using Autoencoder and reducing the dimension of drug properties.

Keywords: Autoencoder, Compressed sensing, Drug, Neural Network, Random forest, Virus

[1] e-mail address: milad1besharai@aut.ac.ir, Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran.
[2] e-mail address: arg135@pitt.edu, Department of Mathematics, University of Pittsburgh, Pennsylvania, USA.

Introduction

Acute Respiratory Syndrome (SARS) of Corona-2 (SARS-CoV-2) virus has caused widespread disruption in most economic and social fields, and its reckless spread has forced many countries to become infected with the virus [9]. From the first case in Wuhan, China, in December 2019 until today, despite the vaccination of many people, there are still deaths from COVID-19 (VIrus-2019 coronary heart disease) [20]. This virus is different from SARS-CoV and MERS-CoV, SARS-CoV-2. Covid-19 is also the most pathogenic human coronavirus ever detected [19]. Meanwhile, much research has focused on finding a solution to treat people with COVID-19. Various laboratory and computational studies are underway in multiple fields, and to date, several vaccines have been approved to control the virus. On October 22, 2020, Remdesivir was approved by the US Food and Drug Administration (FDA) as the first official treatment for COVID-19 [27].

The purpose of drug repositioning is to find a new therapeutic target in drugs. With the spread of the coronavirus, the importance of using this method to find effective drugs for this new and dangerous virus has doubled. For example, in 2020, based on a study by Lim et al., It was found that ribavirin, previously used to treat infectious diseases such as hepatitis, would also be effective in treating Covid-19 [16, 27]. The usefulness of drug repositioning compared to traditional drug discovery methods is to optimize the time and cost of drug production and reduce the potential risks associated with drug toxicity.

In recent years, many studies have been conducted to find effective drugs in the treatment of COVID-19 using drug repositioning. In 2020, Peng et al. clinically reviewed about 20 drugs and identified which drugs could effectively treat Covid-19 [23]. Che et al. were also able to predict useful drugs in the treatment of Covid-19 by embedding a knowledge chart [5]. This study formed a relationship between drugs, genes, diseases, side effects, and pathways. They used the Graph Convolutional Network with Attention to identify potential relationships between drugs and diseases. Another model was proposed in 2021 by Meng et al. They predicted the drug-virus relationship based on the matrix factorization model [19]. This method uses chemical structures of drugs and virus genomic sequences to calculate the similarities between drugs and viruses, respectively. Finally, using the matrix factorization approach predicts the relationship between

each drug-virus pair. Tang et al. Also identified the drug-virus relationship in 2021 using matrix factorization [27]. In this method, using similarity matrices of drugs based on their structure and also similarity matrices of viruses based on their sequence, they predicted the drug-virus relationship.

In this study, we presented a similarity-based approach to predicting the association between drugs and viruses. In this way, we calculate the similarity between drugs with the help features of fingerprint and phenotype for each drug. For viruses, we also calculated the similarities between them using their sequence information and the pre-trained model Biobert (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [14]. The name of each virus is encoded in vectors of size 768. Similarities between viruses are also obtained using these two features. Finally, by combining these similarities and using a neural network model [8], we predicted the drug-virus association. In this paper, we used human drug virus database (HDVD), according to Study [27]. In addition, we used one of the "compressed sensing" (CS) techniques [7], which is based on reducing the dimension of the matrices. Using this method, which has been used in various bioinformatics issues [15, 22, 24], we predicted the relationship between drugs and viruses. We examined the effect of drug features on this model with the help of Autoencoder. In the following, we will describe our proposed approach in more detail.

In addition, we explored another approach that examined the importance of drug properties. We reduced the properties of the drugs by concatenating the properties of the drugs with each other and using autoencoder [29]. In other words, we once predicted the drug-virus relationship based on all the properties we extracted for drugs in this study. We also re-predicted the relationship between the drug and the virus by reducing the dimension of drug feature vectors.

Methods

After describing the data set used in this section, we first express our proposed approach based on similarity vectors and classifier learning models to predict drug-virus interaction. In the following, using the CS model and the different properties of drugs and viruses, we will predict the drug-virus relationship and examine the importance of each drug features using Autoencoder.

Human drug virus database (HDVD). In this study, we used the dataset used in [19]. The details of this dataset are described in Table 1.

Table 1: Human drug virus database (HDVD).

Dataset	drugs	viruses	drug–virus associations
HDVD	219	34	455

Problem Description. The problem of predicting the drug-virus association can be considered as a bipartite network. Which has n drugs and m viruses. Now the matrix adjacent to the intended network has a dimension of $m \times n$ ($R_{m\times n}$). We denote the set of drugs by $D = \{d_1, d_2, \dots, d_m\}$ and the set of viruses by $V = \{v_1, v_2, \dots, v_n\}$. In the network adjacency matrix, $R_{ij} = 1$ means that drug d_i is associated with virus v_j; otherwise, it is $R_{ij} = 0$. In this problem, we attempt to identify the unknown drug-virus association ($R_{ij} = 0$).

Similarities between drugs. To predict the association between drugs and viruses, we first obtain different characteristics for each drug, such as side effects, indications, genes, phenotypes, and fingerprints (structural features). Then, based on various computational criteria such as Gaussian Interaction Profile (GIP), Cosine, correlation, Tanimoto, and Mutual Information (MI), we obtained similarities between drugs [10, 24]. For each drug d_i, we create fingerprint profiles (FP), genotype (G), phenotype (PH), side effect (SE), and indication (IN). In general, the profile of each d_i drug can be displayed as follows:

$$d_i^U = \{(u_1, u_2, \dots, u_l) \mid U \in \{FP, G, PH, SE, IN\}, u_i \in \{0,1\}\}.$$

For example d_i^{FP} is a binary vector, each component of which represents a property in the structure of the drug; if there is that property in the drug d_i, it is equal to 1 and otherwise 0. The amount of l in the d_i^U vector can vary depending on the length of each profile. The following is a brief description of each of the features used.

- **Fingerprint:** We can encode any drug into a binary line vector 881-dimensional using chemistry development kit (CDK) service and Pubchem database [10, 11, 26]. In this binary vector, each bit represents the presence of a predefined piece of chemical structure. If this property exists, we set that bit to 1 and otherwise to 0.
- **Genotype:** We consider the set of all genes that change due to drug use, function, and regulation to be drug-dependent genes. The set of these genes can be extracted from CTD database [18].
- **Phenotype:** Phenotype refers to a non-disease biological event. For example, cell cycle reduction is a phenotype. Under the influence of drug use, cellular, molecular, and physiological phenotypes are formed. All chemical-phenotypic interactions are available under the CTD database [18].
- **Side effect:** A side effect is an effect of a drug that is separate from the main therapeutic effect of the drug. These side effects are available from the Sider database [13].
- **Indication:** The set of disorders for which a drug is prescribed or used for treatment is called the "indication" of that drug. Indications of a drug can be extracted from the Sider database [13].

For viruses, in addition to the similarities obtained through their genome sequences [19], we also received specific vectors for each virus using the Biobert model. Biobert is a pre-trained model on a variety of biomedical texts (such as PubMed publications) that can give us a good representation of the viruses used in the dataset. The name of each input virus of the Biobert model and its output are vectors of size 768.

In the following, we will review the meters we used to calculate the similarity in this study [10,24]:

- **Gaussian Interaction Profile (GIP):** The Gaussian similarity criterion based on the exponential function of EXP is defined as follows:

$$S_{GIP}(f_i, f_j) = \exp(-\gamma \, ||f_i - f_j||^2),$$

  In the Gaussian kernel, γ is the bandwidth controlling parameter. f_i and f_j are also input vectors (such as drug property vectors) calculated using the Gaussian criterion [10].

- **Cosine:** The criterion of cosine similarity is defined as follows [10]:

$$S_{Cos}(f_i, f_j) = \frac{f_i . f_j^T}{|f_j||f_j|}.$$

- **Correlation:** We also used the correlation criterion to calculate the similarity, which is defined as follows:

$$S_{Corr}(f_i, f_j) = \frac{COV\ (f_i, f_j)}{\sqrt{Var(f_i).Var(f_j)}},$$

In this relation, COV means covariance, and Var means variance [10].

- **Tanimoto:** Another similarity criterion is based on the Tanimoto coefficient, which is expressed as follows:

$$S_{TAN}(f_i, f_j) = \frac{|f_i \wedge f_j|}{|f_i \vee f_j|}.$$

The notation $|f_i \wedge f_j|$ indicates that in several components of the f_i and f_j feature vectors, both have the same value 1. $|f_i \vee f_j|$ also represents the number of f_i and f_j vectors where at least one of the components is equal to 1 [12, 24].

- **Mutual Information (MI):** We also used the mutual information relationship to calculate similarity. This relationship is defined as follows:

$$S_{MI}(f_i, f_j) = \sum_{u=0}^{1}\sum_{v=0}^{1} fr(u, v) \log\left(\frac{fr(u, v)}{fr(u)fr(v)}\right),$$

in this relation, $fr(u)\ (fr(v))$ refers to the frequency of the u (v) in the f_i (f_j) vector. The $fr(u, v)$ frequency is relative [10].

For viruses, as mentioned, we calculated their similarity once based on the sequence of their genomes (S_{Seq}) and once based on the vectors obtained by Biobert ($S_{Biobert}$).

After finding the similarities between the matrices based on the different properties and criteria, we integrate them according to the kernel target alignment (KTA) method [10]. The weight of each similarity matrix is obtained according to the KTA method as follows:

$$\beta_{i,d} = \frac{A(K_{i,d}, RT_d)}{\sum_{i=1}^{l} A(K_{i,d}, RT_d)}, \qquad (i = 1,2,3, \dots, l), \qquad (1)$$

$$\beta_{i,v} = \frac{A(K_{i,v}, RT_v)}{\sum_{i=1}^{l} A(K_{i,v}, RT_v)}, \qquad (i = 1,2,3, \dots, l'), \qquad (2)$$

In the above relation, $A\left(K_{i,d}, RT_d\right)$ and $A\left(K_{i,v}, RT_v\right)$ means the similarity of cosine between matrices, which is defined as follows:

$$A(P', P) = \frac{< P', P >_F}{\|P'\|_F \|P\|_F},$$

Which $\|P'\|_F$ and$< P', P >_F$ is obtained as follows:

$$\|P\|_F = \sqrt{< P, P >_F},$$

$$< P, Q >_F = Trace(P^T Q).$$

In Eq.(1) and Eq.(2), RT_d (RT_v) means $RT_d = RR^T$ $(RT_v = R^T R)$ and also l (l') means the number of drug similarity matrices (viruses similarity matrices).

Finally, by minimizing the loss function (3), we obtain the latent factor of drug space $(F = (f_{ij}))$ and latent factor of virus space $(G = (g_{ij}))$, and from their combination, we obtain the probability of any drug-virus associations (see Figure 1 (part (III))).

$$\sum_{i,j} W_{i,j} \left\{ ln\left(1 + e^{f_i g_j^T}\right) - \left(r_{i,j}\right) f_i g_j^T \right\} + \lambda_r \|F\|_F^2 + \lambda_r \|G\|_F^2 + \lambda_{SD}\, tr(F^T\, (D_{SD} - SD)F) + \lambda_{SV}\, tr(G^T\, (D_{SV} - SV)G), \quad (3)$$

In Eq.(3) SD (SV) means a matrix of similarities between drugs (viruses) that are obtained after combination with the KTA method. $W_{i,j}$ is a drug-virus frequency matrix derived from the drug-virus association matrix[3]. F^T is the transpose of F and $\|\quad\|_F$ means the Frobenius norm. The tr also represents the trace of the matrix and D_{SD} means 'degree matrix' of SD [24]. After finding

[3] Unfortunately, our current experiments do not use weights due to the unavailability of virus-drug frequency data.

the matrices of F and G using the following equation, the predicted values for each drug-virus relationship are calculated.

$$P = \frac{\exp(FG^T)}{(1 + \exp\,(FG^T))}.$$

Finally, our goal is to find the drug-virus association using a neural network classifier learning model. As shown in Fig 1, we used similarity-based vectors to predict the drug-virus association. We first put the set of drugs associated with each virus (v) in the S_v set. We also put a set of related viruses in the S_d for each drug (d). The feature vector for each drug-virus pair is then formed as follows:

$$fe_1(d,v) = \max_i\{\, S_{Cos}(d^{FP}, d_i^{FP}) \mid d_i \in S_v \backslash \{d\}\,\},$$

$$fe_2(d,v) = \max_i\{\, S_{Cos}(d^{PH}, d_i^{PH}) \mid d_i \in S_v \backslash \{d\}\,\},$$

$$fe_3(d,v) = \max_i\{\, S_{Cos}(v^{Bert}, v_i^{Bert}) \mid v_i \in S_d \backslash \{v\}\,\},$$

$$fe_4(d,v) = \max_i\{\, S_{Seq}(v^{seq}, v_i^{seq}) \mid v_i \in S_d \backslash \{v\}\,\}.$$

So for each drug-virus (d, v) pair, we have a vector in dimension 4 (FE), each of which consists of fe_1, fe_2, fe_3, and fe_4 (See Fig 1). In other words, for drug d, we calculated its similarity to other drugs that interact with virus v, and maximum of similarity, we considered it as one of the components of the feature vector FE. Finally, we predicted the drug-virus association with the help of the obtained feature vectors and the neural network model.

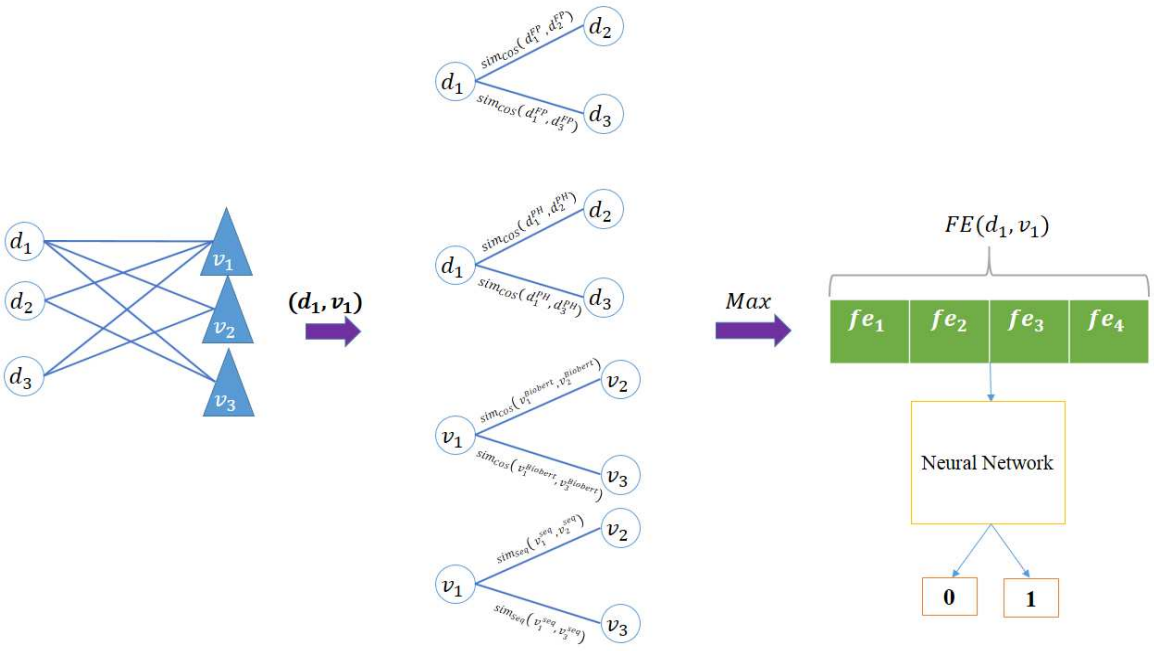

Figure 1: The structure of the proposed model. For all drugs and viruses in the data set, we have shown the relationship between them with the help of nodes and edges. For each specific drug-virus pair (for example, we considered d_1-v_1), we calculate the similarity between d_1 and other drugs, based on their association with the v_1 and the fingerprint and phenotype features.

In Figure 1, we show the process of constructing feature vectors for each drug-virus pair for a sample (d_1, v_1).

We also used the Compressed Sensing (CS) matrix factorization model to predict the drug-virus association and examined the different features of drugs and viruses. As shown in Fig. 2, problem drug-virus prediction is divided into three parts. In Part (I), model inputs are made. The inputs to the problem are drug-virus adjacency network matrix (R), drug features matrix (F_D), and virus features matrix (F_V). In the next part (II), we calculated the similarity between drugs and the similarity between viruses using different computational criteria. Then, using method Kernel Target Alignment-based Multiple Kernel Learning (KTA-MKL) [10], we combined the similarity matrix of drugs. We similarly combined the similarities between the viruses. Finally, using these input matrices and compressed sensing technique, we predicted the drug-virus relationship (part (III)).

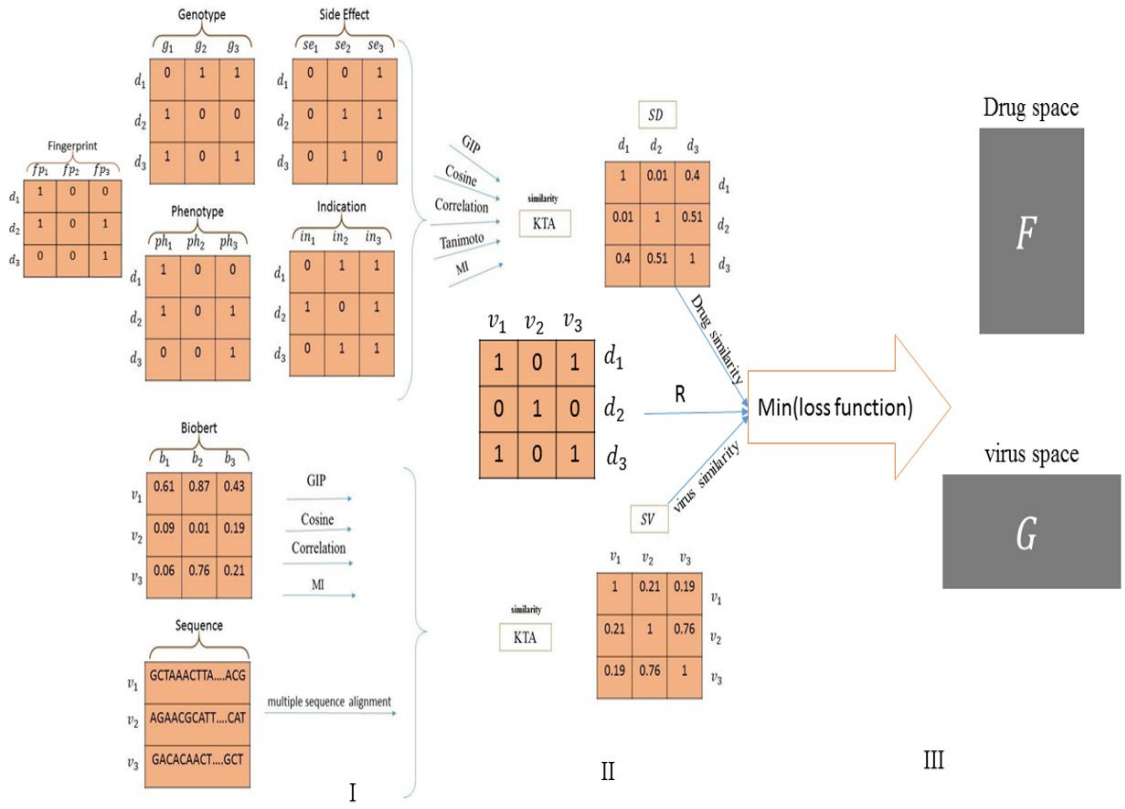

Figure 2: Overview of the work flow of this study. I. In this part, we first extract the characteristics of each drug (d_i) and also for each virus (v_i), in addition to their genomic sequence, using the Biobert model, we extract its specificity vector for each virus name. II. Then we calculate the similarity between drugs and viruses with different computational criteria (GIP,MI, Cosine, Correlation). III. Finally, by minimizing the loss function and finding the hidden factors, we predict the drug-virus association.

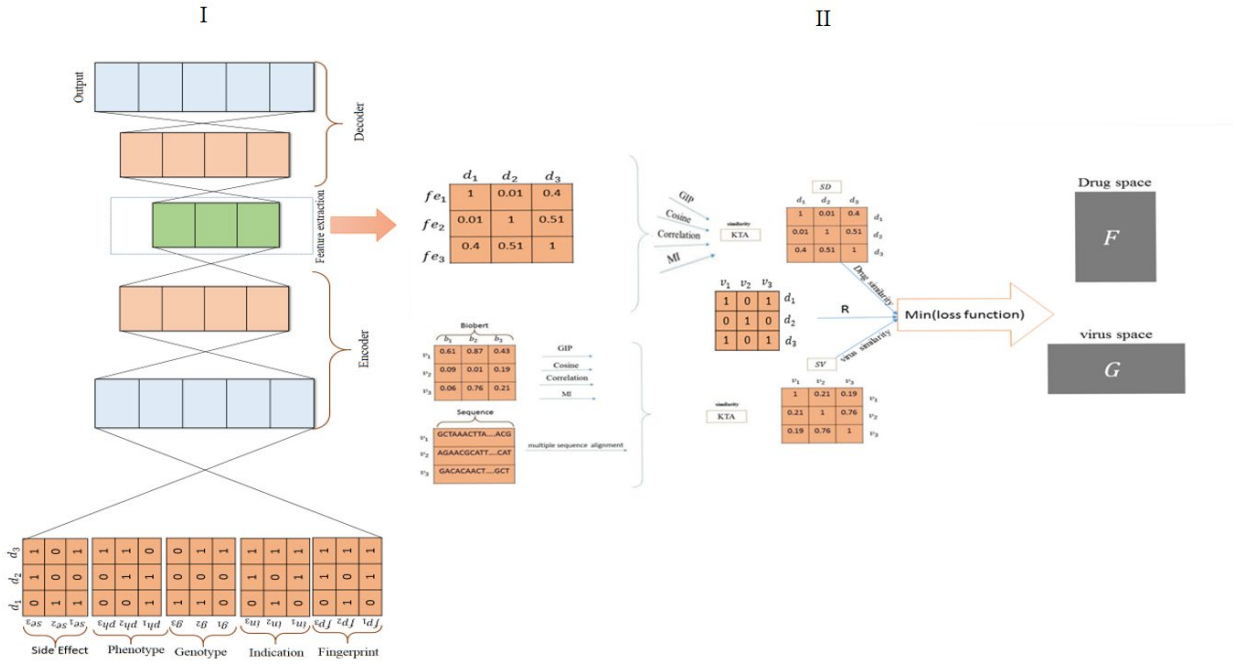

Figure 3: Overview of work using Autoencoder. I. In this part, we have formed a vector for each drug during 19821, which is obtained by combining all the drug properties. For each virus (v_i), in addition to their genomic sequence, using the Biobert model, we extract its specificity vector for each virus name. II. We calculate the similarity of drugs using the obtained features by reducing the specificity of the properties (We showed each of the features obtained from the hidden layer of autoencoder with fe_l in the figure.) with the autoencoder and the stated computational criteria (MI, GIP, Cosine, Correlation). For viruses, in addition to genomic sequences, we calculate the similarity between each pair of viruses with Biobert derived vectors and various computational criteria. Finally, by minimizing the loss function of the CS technique, we find the latent factors in the space of drugs (F) and viruses (G) and use them to calculate the drug-virus relationship.

According to Figure 3, this time, we concatenated all the properties of the drug (Part (I)). After concatenating feature vectors for each drug, the feature vector of dimension 19821 for each drug was obtained. Then, using an autoencoder, we reduce the feature dimension from 19821, and each time by calculating the similarity of the drugs based on the obtained features (part (II)), we reviewed the CS model in predicting the relationship between drugs and viruses.

Table 2 also shows the size of each of the drug and virus characteristics we used in this study.

Table 2: The dimension of each feature used is used.

input	Drug					virus
features	fingerprint	genotype	phenotype	side effect	indication	Biobert
dimension	881	14361	1495	2396	688	768

Results

Our proposed approaches are based on similarity vector and classification (random forest and neural network) models. In these models, we randomly select as much as 20% of the positive data from the positive data and randomly select the same amount from the negative data. First, we form the similarity vectors for each pair (d, v). Then, each pair is wholly removed from the data set (training set). We also compared our proposed approach (Neural Network model) with the CS, Similarity Constrained Probabilistic Matrix Factorization (SCPMF) [19] and random forest models. Similarly, in the matrix factorization model, we masked the data (drug-virus pair) that we set aside to test the neural network model and, therefore, compared the models to five experiments. We also examined the importance of drug features by reducing dimensions using an automated encoder. In this section, we evaluate two other frameworks (with dimension reduction and without dimension reduction). We also compared the CS approach to other models such as the Similarity Constrained Probabilistic Matrix Factorization (SCPMF) [19], Inductive Matrix Completion (IMC) [6], Regularized Least Squares (RLS) [30], Network Consistency Projection (NCP) [29] and Bounded Nuclear Norm Regularization (BNNR) [31] models.

All codes and tests on Matlab 2018b run on Windows and Intel Core i5-2430M processors and 4 GB of memory. In the following, we first state the values that we considered for the parameters of the proposed approach (random forest model) and other methods and then assert the criteria we used to evaluate our model.

Parameters setting

In the CS approach, we set the value of the parameters to $\lambda_r = 0.5$, $\lambda_{SD} = 0.01$, and $\lambda_{SV} = 10$. The reduction value of the given dimension is equal to 18 and also the number of repetitions to minimize the loss function is equal to 100. The autoencoder used has one hidden layer, and its activating function is for the hidden layer “Sigmoid”.

We also configured hyper parameter of random forest (RF) and neural network (NN) is shown in table 3. In fact, "MinLeafSize" refers to the minimum observations per leaf, which is important in dividing the nodes in the decision trees. Moreover "NumPredictorsToSample" means the number of variables that are randomly selected at each decision.

Table 3: Hyper parameters used in the RF (Random Forest) and NN (Neural Network) models.

Parameter	Value
MinLeafSize (RF)	3
NumPredictorsToSample (RF)	1
Hidden Sizes layers (NN)	100
Learning rate (NN)	0.05

Model evaluation

We evaluated our model based on its performance in predicting drug-virus association. To evaluate the proposed approach (neural network model), we used the measurement criterion of the area under the receiver operating characteristic curve (AUC). This curve is obtained based on the false positive rate (FPR) and the classifier model's real positive rate (TPR) under different classification thresholds. The TPR and FPR values are obtained as follows:

$$FPR = \frac{FP}{FP + TN}. \qquad TPR = \frac{TP}{TP + FN}.$$

FP means the number of incorrect predictions in the positive samples, TN implies the number of correct identifications in the negative samples, TP means the number of correct predictions in the positive samples. Finally, FN, the number of incorrect labels in the sample Shows negatives. Since the AUC is not the only suitable metric for the problem, we also used the area under the Precision-Recall curve (AUPRC) measure for evaluation. This measure measures the area under the call accuracy curve (PR). In other words, the relationship between sensitivity (recall) and positive predictive value (precision) is shown. These concepts are defined as follows:

$$precision = \frac{TP}{TP + FP}. \quad recall = \frac{TP}{TP + FN}.$$

All comparisons are based on 10-fold Cross-Validation for training models with training data sets and testing models based on a selection of 20% of positive data and the same amount of negative data, and the size of the latent factor reduction is equal to 18.

Performance of proposed model

In this section, we first examine neural network (NN) and random forest (RF) models based on similarity vectors to predict the drug-virus association. We also compare these proposed approaches with the CS model and SCPMF model. The results, after five experiments on different test data, are given in Table 4.

Table 4: Test data in these models were randomly selected from 20% of positive and negative data. The results are shown after five tests in each model.

Model	AUC	AUPR
NN	**0.97**	**0.96**
	0.92	**0.94**
	0.97	**0.96**
	0.94	**0.92**
	0.97	**0.96**
RF	0.84	0.91
	0.77	0.87
	0.76	0.86
	0.84	0.90
	0.88	0.92
CS	0.88	0.90
	0.88	0.88
	0.87	0.89
	0.90	0.91
	0.80	0.77
SCPMF	0.94	0.83
	0.81	0.73
	0.96	0.85
	0.97	0.86
	0.98	0.88

We evaluate the performance of several features and a single feature of drugs and viruses in the HDVD. The prediction results are shown in Tables 5 and 6. It is necessary to note that in all models, we integrated the similarities obtained based on different computational criteria or different properties of drugs and viruses with the help of the KTA method. Among the features of the drug, phenotype and fingerprint were able to have a more significant impact than other features. It should be noted that the comparison between the matrix analysis models (shown in Tables 5, 6 and 7) is based on 5-fold cross-validation. This is why the AUC of predicting models by these two features is better than other features. (see Table 5 and Table 6)

Table 5: The evaluation of the models is based on the different properties of the drugs. The similarity of viruses is calculated based on the sequence of their genomes.

Model	Drug									
	Fingerprint		Side effect		Phenotype		Indication		Gene	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
CS	**0.87**	**0.37**	**0.85**	**0.4**	**0.84**	**0.37**	**0.85**	**0.37**	**0.84**	**0.36**
RLS	0.84	0.35	0.83	0.36	0.83	0.32	0.83	0.35	0.82	0.32
BNNR	0.53	0.07	0.5	0.07	0.54	0.07	0.51	0.06	0.55	0.08
IMC	0.61	0.14	0.62	0.16	0.63	0.17	0.63	0.17	0.61	0.17
NCP	0.38	0.15	0.40	0.15	0.38	0.13	0.37	0.13	0.38	0.14
SCPMF	0.86	0.29	0.83	0.35	0.78	0.32	0.79	0.27	0.78	0.3

Table 6: The evaluation of the models is based on the different properties of the drugs. The similarity between the viruses was calculated based on 768-dimensional vectors encoded by the Biobert model, with different similarity criteria.

Model	Drug									
	Fingerprint		Side effect		Phenotype		Indication		Gene	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
CS	**0.83**	0.27	**0.83**	0.31	**0.83**	**0.25**	**0.83**	**0.31**	**0.82**	0.25
RLS	0.81	**0.28**	0.82	0.24	0.82	0.24	**0.83**	0.3	0.81	0.24
BNNR	0.52	0.07	0.50	0.06	0.54	0.07	0.51	0.06	0.55	0.08
IMC	0.7	0.19	0.71	0.2	0.7	0.2	0.71	0.19	0.71	0.18
NCP	0.39	0.17	0.39	0.15	0.38	0.13	0.37	0.12	0.37	0.12
SCPMF	0.82	0.27	0.82	**0.35**	0.76	**0.31**	0.76	0.27	0.75	**0.29**

As we can see in Tables 5 and 6, the CS model performs better than other models. Feature sequences for viruses are more effective in predicting the relationship between drugs-viruses.

Table 7: The results based on 5-fold CV are shown in different models. These results are based on the fact that the similarities between the drugs are calculated and integrated based on all the characteristics (PH, FP, G, SE and IN), and for the viruses, the similarities are based on the sequences and the similarities are based on the encoded Biobert vectors.

	Model					
	CS	RLS	BNNR	IMC	NCP	SCPMF
AUC	**0.86**	0.83	0.53	0.68	0.4	0.82
AUPR	**0.34**	0.31	0.07	0.2	0.15	0.31

After executing the models, the results in Table 7 were obtained. These results, the mean values of AUC and AUPR after five runs, indicate that the proposed CS model performs better in predicting the drug-virus relationship. Another point is that using only similarities based on the genome sequences of viruses can be more AUPR in models (see Tables 3 and 5).

We also measured the performance of the proposed model in predicting drug-virus association, based on the obtained characteristics, after concatenating the features of the drugs and using an autoencoder to reduce the dimension. The results can be seen in Table 8. We also specified the value of the autoencoder mean squared error (MSE) error in Table 8 for each dimension reduction value.

Table 8: Investigating the importance of drug properties in predicting the relationship between drugs and viruses using reducing the properties dimension in the CS model.

dimension	CS		
	AUC	AUPR	MSE
19821	0.85	0.33	-
15000	0.83	0.27	0.01
10000	0.83	0.27	0.01
5000	0.83	0.27	0.01
1000	0.83	0.26	0.01
500	0.83	0.25	0.01

As you can see in Table 8, we reduced the size of the medicinal properties from 19821 to 500, and as a result, the amount of AUPR decreased by only eight percent. It can be concluded that only a limited number of medicinal properties together with the properties of the virus can optimally predict the relationship between the drug and the virus. As shown in Figure 4, we see that each virus is associated with about 13 drugs on average. As a result, the sparsity of the dataset used is high, making it difficult to accurately detect the drug-virus relationship by computational models.

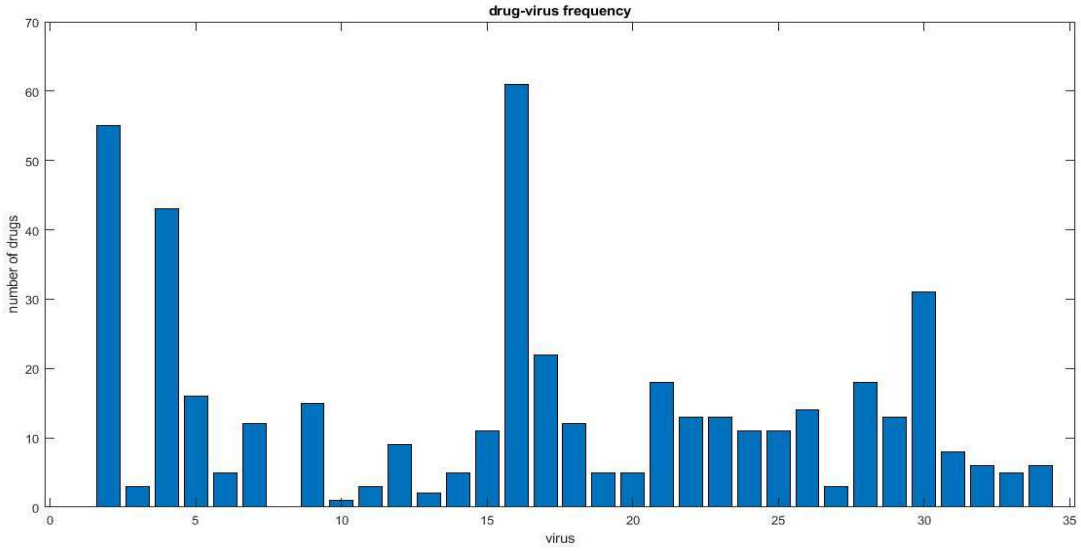

Figure 4: Drug-virus frequency. It has been shown how many drugs each virus is associated with within the HDVD.

CS for COVID-19

Coronavirus 2 (SARS-CoV-2), an infectious disease caused by acute respiratory syndrome, was reported in China in December 2019 [2]. Covid-19 has caused many challenges and problems to many countries around the world to date. In this study, we identify drugs with a high potential for association with this type of virus with the help of the CS model, among the drugs in the HDVD.

Table 9: We have shown the possible values predicted by Methods CS and SCPMF for the association between drugs and Covid virus 19. In the last column, we present the studies that have shown this relationship.

Drug	Probability CS	Probability SCPMF [3]	Study
Mycophenolate Mofetil (DB00688)	0.73	0.23	[3, 21]
Censavudine (DB12074)	0.77	0.3	-
Bortezomib (DB00188)	0.7	0.3	[4, 17]
Mesalazine (DB00244)	0.66	0.27	[25]
Ramipril (DB00178)	0.67	0.23	[1]

According to Table 9, we have identified the most likely drugs associated with the coronavirus. We also examined these relationships in SCPMF and showed the possible values that that model predicts. Identifying these connections with computational methods and examining them more closely can be effective in finding more effective drugs for the coronavirus.

Discussion

The coronavirus, which has progressed uncontrollably in many countries, has caused many problems. In addition to vaccine production, the use of available drugs effective in controlling mortality from Covid-19 can also be a promising path to safety and health. Machine learning and data mining models have been able to help laboratory methods to find the drug-virus relationship to a great extent. These methods can predict drug-virus relationships at a better cost and time. In addition to the biological characteristics of drugs and viruses, the use of clinical data can also be effective. For example, extracting relevant information from social networks such as Twitter and electronic databases of medical records and teaching this information along with biological data from drugs and viruses to learning models can improve prediction efficiency.

Conclusion

In this paper, we present an approach based on similarity vectors. For each drug-virus pair given as input, we assume the maximum similarity of the drug to other virus-related drugs. Drugs similarity criteria are based on two structural features (fingerprint) and phenotype of drugs. We do the same for viruses and use similarities between them based on the sequence of each virus and Biobert model. Finally, we classify each drug-virus pair using vectors obtained with neural network and random forest binary classification models based on drug-virus interaction. In addition, we used CS matrix factorization models and examined different drug features to see how effective they were in predicting drug-virus association. We also compared it with other models and finally evaluated the importance of each of the drug features by reducing the dimension of features using Autoencoder. Finally, it can be said that learning models work better in predicting the relationship between drugs and viruses. Also, by considering more features of viruses and drugs, the performance of these models can be improved.

Acknowledgements

The authors received no financial support for the research, authorship, and/or publication of this article.

References

[1] Ajmera, V. a. (2021). RAMIC: Design of a randomized, double-blind, placebo-controlled trial to evaluate the efficacy of ramipril in patients with COVID-19. *Contemporary Clinical Trials*, 106330; doi.org/10.1016/j.cct.2021.106330

[2] Asai, A. a. (2020). COVID-19 drug discovery using intensive approaches. *International journal of molecular sciences*, 2839; doi.org/10.3390/ijms21082839

[3] Balestri, R. a. (2020). Occurrence of SARS-CoV-2 during mycophenolate mofetil treatment for pemphigus. *J Eur Acad Dermatol Venereol*, e435--e436.

[4] Bellesso, M. a. (2021). Second COVID-19 infection in a patient with multiple myeloma in Brazil-reinfection or reactivation? *Hematology, Transfusion and Cell Therapy*, 109--111; doi.org/10.1016/j.htct.2020.12.002

[5] Che, M. a. (2021). Knowledge-Graph-Based Drug Repositioning against COVID-19 by graph convolutional network with attention mechanism. *Future Internet*, 13; doi.org/10.3390/fi13010013

[6] Chen, X. a.-N.-Q. (2018). Predicting miRNA--disease association based on inductive matrix completion. *Bioinformatics*, 4256--4265; doi.org/10.1093/bioinformatics/bty503

[7] Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 1289--1306; doi.org/10.1109/TIT.2006.871582

[8] Dreiseitl, S. a.-M. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 352-359.

[9] Ghosh, K. a. (2021). Chemical-informatics approach to COVID-19 drug discovery: Exploration of important fragments and data mining based prediction of some hits from natural origins as main protease (Mpro) inhibitors. *Journal of molecular structure*, 129026; doi.org/10.1016/j.molstruc.2020.129026

[10] Guo, X. a. (2020). A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment. *BioMed Research International*; doi.org/10.1155/2020/4675395

[11] Kim, S. a. (2016). PubChem substance and compound databases. *Nucleic acids research*, D1202--D1213; doi.org/10.1093/nar/gkv951

[12] Kristensen, T. G. (2010). A tree-based method for the rapid screening of chemical fingerprints. *Algorithms for Molecular Biology*, 1--10; doi.org/10.1186/1748-7188-5-9

[13] Kuhn, M. a. (2016). The SIDER database of drugs and side effects. *Nucleic acids research*, D1075--D1079; doi.org/10.1093/nar/gkv1075

[14] Lee, J. a. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 1234--1240; doi.org/10.1093/bioinformatics/btz682

[15] Lim, H. a. (2016). Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Scientific reports*, 1--11; doi.org/10.1038/srep38860

[16] Lim, J. a. (2020). Case of the index patient who caused tertiary transmission of COVID-19 infection in Korea: The application of lopinavir/ritonavir for the treatment of COVID-19 infected pneumonia monitored by quantitative RT-PCR. *Journal of Korean medical science*, e79--e79; doi.org/10.3346/jkms.2020.35.e79

[17] Longhitano, L. a. (2020). Proteasome inhibitors as a possible therapy for SARS-CoV-2. *International journal of molecular sciences*, 3622; doi.org/10.3390/ijms21103622

[18] Mattingly, C. J. (2003). The comparative toxicogenomics database (CTD). *Environmental health perspectives*, 793--795; doi.org/10.1289/ehp.6028

[19] Meng, Y. a. (2021). Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Applied soft computing*, 107-135; doi.org/10.1016/j.asoc.2021.107135

[20] Mongia, A. a. (2021). A computational approach to aid clinicians in selecting anti-viral drugs for COVID-19 trials. *Scientific reports*, 1-12; doi.org/10.1038/s41598-021-88153-3

[21] Neurath, M. F. (2021). COVID-19: biologic and immunosuppressive therapy in gastroenterology and hepatology. *Nature reviews Gastroenterology \& hepatology*, 1--11; doi.org/10.1038/s41575-021-00480-y

[22] Parvaresh, F. a. (2008). Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 275--285; doi.org/10.1109/JSTSP.2008.924384

[23] Peng, Y. a. (2021). A comprehensive summary of the knowledge on COVID-19 treatment. *Aging and disease*, 155; doi.org/10.14336/AD.2020.1124

[24] Poleksic, A. a. (2018). Predicting serious rare adverse reactions of novel chemicals. *Bioinformatics*, 2835--2842; doi.org/10.1093/bioinformatics/bty193

[25] Rizzello, F. a. (2021). COVID-19 in IBD: The experience of a single tertiary IBD center. *Digestive and Liver Disease*, 271--276; doi.org/10.1016/j.dld.2020.12.012

[26] Steinbeck, C. a. (2003). The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences*, 493--500; doi.org/10.1021/ci025584y

[27] Tang, X. a. (2021). Indicator Regularized Non-Negative Matrix Factorization Method-Based Drug Repurposing for COVID-19. *Frontiers in Immunology*, 3824; doi.org/10.3389/fimmu.2020.603615

[28] Wang, Y. a. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, 232--242; doi.org/10.1016/j.neucom.2015.08.104

[29] Xie, G. a. (2019). NCPHLDA: a novel method for human lncRNA--disease association prediction based on network consistency projection. *Molecular omics*, 442--450; doi.org/10.1039/C9MO00092E

[30] Yang, H. a. (2021). Drug-disease associations prediction via Multiple Kernel-based Dual Graph Regularized Least Squares. *Applied Soft Computing*, 107811; doi.org/10.1016/j.asoc.2021.107811

[31] Yang, M. a. (2019). Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics*, i455--i463; doi.org/10.1093/bioinformatics/btz331