Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CSDVP: Compressed Sensing 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, Matrix factorization, Virus

Posted Date: September 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-910042/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

CSDVP: Compressed Sensing 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. Although no drug has been definitively approved for the treatment of this disease, the effectiveness of a few drugs for the treatment of the disease has been observed. In this study, with the help of computational matrix factorization methods, the associations between drugs and viruses have been predicted. By combining the similarities between the drugs and the similarities between the viruses and using the compressed sensing technique, we investigated the association between the drug and the virus. The Compressed Sensing approach to Drug-Virus Prediction (CSDVP) can work well. We compared the proposed method with other methods in this field and found its accuracy is more desirable than other methods. In fact, the CSDVP approach with the HDVD dataset and evaluation through 5-fold CV, with AUC = 0.96 and AUPR = 0.85, can identify the relationship between drugs and viruses. We also investigated the effect of drug properties on model performance improvement using autoencoder. Thus, with each decrease in the size of the characteristics in different sizes, we examined the performance of the CSDVP model in predicting the drug-virus relationship. The relationship between drugs and coronavirus infection is also analyzed, and the results are presented.

Keywords: Autoencoder, Compressed sensing, Drug, Matrix factorization, Virus

[1] e-mail address: milad1besharai@aut.ac.ir
[2] e-mail address: arg135@pitt.edu

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 [8]. 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) [19]. This virus is different from SARS-CoV and MERS-CoV, SARS-CoV-2. Covid-19 is also the most pathogenic human coronavirus ever detected [18]. 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 [26].

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 [15, 26]. 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 [22]. 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 [18]. 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 [26]. 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 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 [14, 21, 23], we predicted the relationship between drugs and viruses. Because studies based on compressed sensing have been successful, we also used this method to predict unknown drug-virus relationships. This method, which has been very effective in recovering signals, also seems to be effective in finding drug-virus associations that are not known [23]. In this paper, we used human drug virus database (HDVD), according to Study [26]. In the problem of drug-virus prediction, we refer to ‘signals’ as drug-virus associations, some of which are known and some of which are unknown. ‘Samples’ are also considered drug-virus pairs whose relationship is known. This method seems to work well for predicting drug-virus association, similar to the drug-ADR prediction problem [23]. Because in this problem, the data has noise, and the number of positive data (known drug-virus relation) is low. Based on this framework, we used the CS method to find the relationship between the drug and the virus. First, to calculate the similarity between each drug pair, we extracted the different properties of the drugs from different databases. We extracted various drug-related features such as structural properties (fingerprint), phenotype, genes, side effect and indication from PubChem [10], CTD [17] and SIDER [12] databases. Then, with different measures, we calculated the similarity between the drugs based on the extracted features. For viruses, in addition to their sequence information, we used pre-trained model Biobert (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [13]. The name of each virus is encoded in vectors of size 768. Similarities between viruses are also obtained using these two features. We then recover drug-virus relationships based on drug-virus relationships, drug similarity, and virus similarity. Finally, we identify the drugs that can be beneficial for the coronavirus.

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 [28]. 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

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 $R_{ij} = 0$. Finally, our goal is to find anonymous drug-virus relationships using the compressed sensing technique. Figure 1 provides an overview of the proposed model (CSDVP) and problem-solving process.
As shown in Fig. 1, problem compressed sensing drug-virus prediction (CSDVP) 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) [9], 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)).

Human drug virus database (HDVD). In this study, we used the data set used in [18]. The details of this dataset are described in Table 1.

Table 1: HDVD dataset.

Dataset	drugs	viruses	drug–virus associations
HDVD	219	34	455

Similarities between drugs. To predict the relationship 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 [9, 23]. 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 [9, 27]. 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 [17].
- **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 [17].
- **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 [12].
- **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 [12].

For viruses, in addition to the similarities obtained through their genome sequences [18], 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 [9,23]:

- **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 [9].

- **Cosine:** The criterion of cosine similarity is defined as follows [9]:

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

- **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 [9].

- **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 [11, 23].

- **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 [9].

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 [9]. 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, $A(K_{i,d}, RT_d)$ and $A(K_{i,v}, RT_v)$ 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).

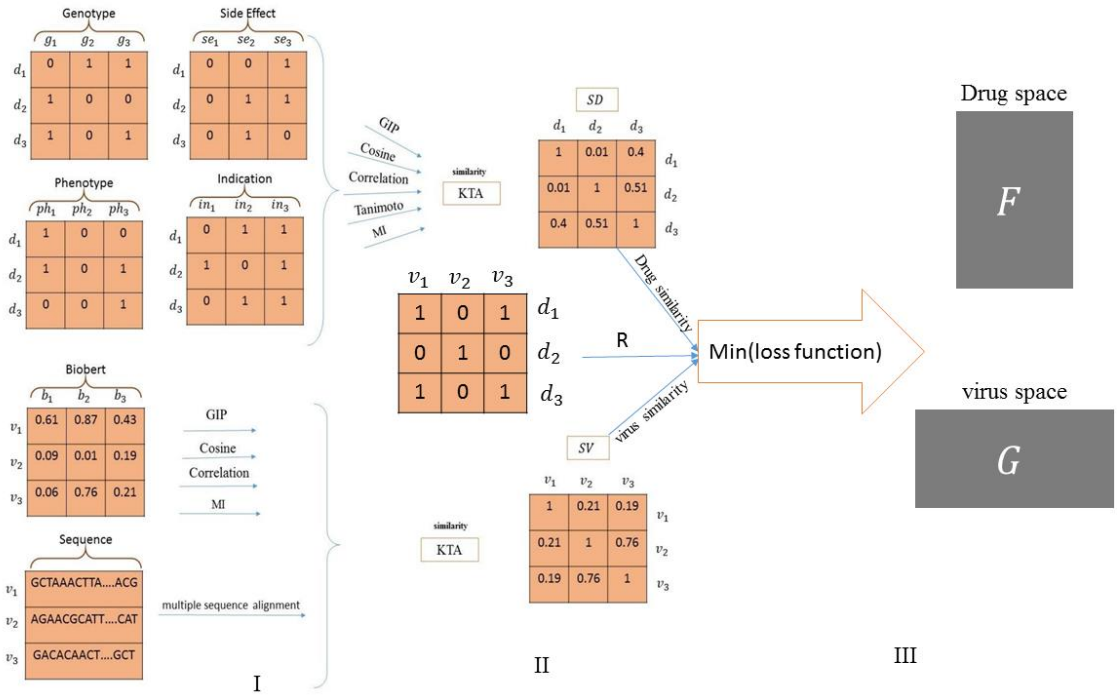

Figure 1: Overview of the work flow of this study. I. In this part, we first extract the characteristics of each drug (d_i) and also for each virus (v_i), in addition to their genomic sequence, using the Biobert model, we extract its specificity vector for each virus name. II. Then we calculate the similarity 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.

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

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

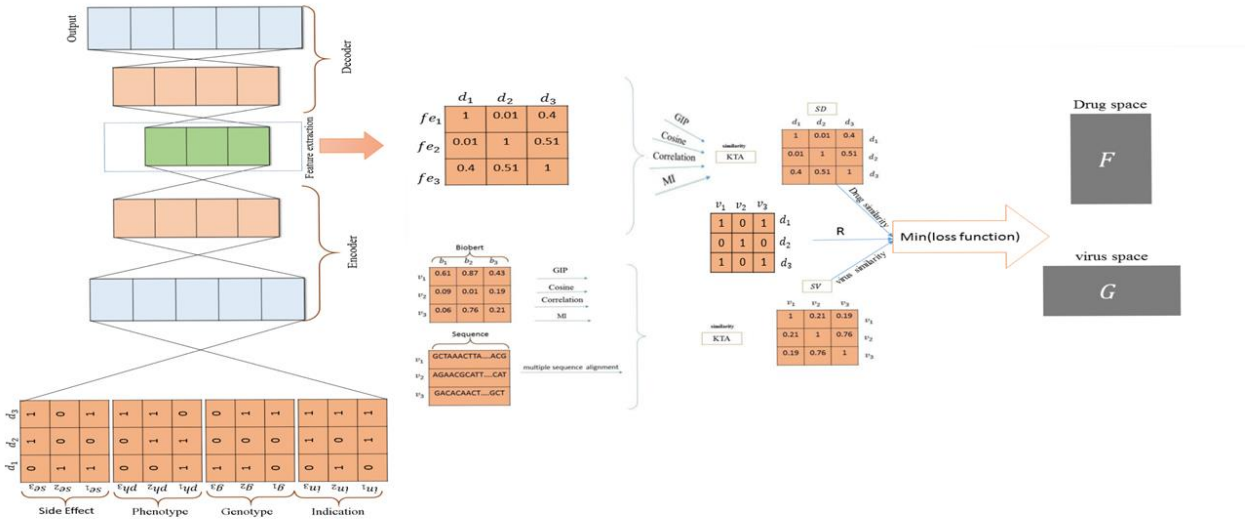

Figure 2: Overview of work using Autoencoder. I. In this part, we have formed a vector for each drug during 19821, which is obtained by combining all the drug properties. For each virus (v_i), in addition to their genomic sequence, using the Biobert model, we extract its specificity vector for each virus 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.

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

According to Figure 2, 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 proposed approaches (CSDVP) 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 approach (CSDVP) is to predict the drug-virus association based on the CS technique. In addition to integrating the features and comparing our proposed method (CSDVP) with other approaches, we also examined the importance of drug specificity by reducing dimensions using an automated encoder.

In this section, we evaluate two frameworks (with reduce dimension and without reduce dimension). We also compared our proposed approach (CSDVP) with the Similarity Constrained Probabilistic Matrix Factorization (SCPMF) [18], 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 (CSDVP) 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".

Model evaluation

We evaluated our model based on its performance in predicting drug-virus association. To evaluate the proposed approach (CSDVP), 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 5-fold cross validation, and the size of the latent factor reduction is equal to 18.

Performance of CSDVP in the 5-fold CV

We evaluate the performance of several features and a single feature of drugs and viruses in the HDVD dataset. The prediction results are shown in Tables 3 and 4. 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.

Table 3: 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
CSDVP	**0.96**	**0.85**	**0.96**	**0.85**	**0.96**	**0.85**	**0.96**	**0.85**	**0.96**	**0.85**
RLS	0.55	0.12	0.55	0.12	0.55	0.12	0.55	0.12	0.55	0.11
BNNR	0.51	0.6	0.50	0.6	0.52	0.7	0.50	0.6	0.52	0.7
IMC	0.54	0.1	0.55	0.11	0.55	0.11	0.55	0.11	0.54	0.1
NCP	0.42	0.12	0.41	0.11	0.41	0.1	0.41	0.1	0.41	0.1
SCPMF	0.64	0.18	0.64	0.24	0.62	0.2	0.62	0.19	0.62	0.19

Table 4: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
CSDVP	**0.95**	**0.76**	**0.95**	**0.78**	**0.96**	**0.78**	**0.96**	**0.78**	**0.95**	**0.78**

RLS	0.55	0.13	0.55	0.11	0.55	0.11	0.55	0.11	0.55	0.10
BNNR	0.51	0.6	0.50	0.6	0.52	0.7	0.50	0.6	0.52	0.7
IMC	0.58	0.12	0.58	0.13	0.58	0.13	0.59	0.13	0.59	0.12
NCP	0.42	0.12	0.42	0.11	0.41	0.1	0.41	0.1	0.41	0.1
SCPMF	0.64	0.17	0.64	0.24	0.61	0.2	0.61	0.18	0.61	0.2

As we can see in Tables 3 and 4, the CSDVP model performs better than other models. Feature sequences for viruses are more effective in predicting the relationship between drugs-viruses.

Table 5: 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, and for the viruses, the similarities are based on the sequences and the similarities are based on the encoded Biobert vectors.

	Model					
	CSDVP	RLS	BNNR	IMC	NCP	SCPMF
AUC	**0.96**	0.55	0.51	0.58	0.42	0.64
AUPR	**0.8**	0.11	0.7	13	0.11	0.2

After executing the models based on 5-fold CV, the results in Table 5 were obtained. These results, the mean values of AUC and AUPR after five runs, indicate that the proposed CSDVP 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 6. We also specified the value of the autoencoder mean squared error (MSE) error in Table 6 for each dimension reduction value.

Table 6: Investigating the importance of drug features in predicting drug-virus relationship in the CSDVP model.

dimension	CSDVP		
	AUC	AUPR	MSE
19821	0.96	0.8	-
15000	0.96	0.79	0.01
10000	0.96	0.79	0.01
5000	0.96	0.79	0.01
1000	0.96	0.79	0.01
500	0.96	0.79	0.01

As you can see in Table 6, we reduced the size of the medicinal properties from 1982 to 500, and as a result, the amount of AUPR decreased by only one percent. This suggests that only a limited number of drug features combined with virus features can predict drug-virus association.

CSDVP 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 CSDVP model, among the drugs in the HDVD dataset.

Table 7: We have shown the possible values predicted by Methods CSDVP 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 CSDVP	Probability SCPMF [3]	Study
Mycophenolate Mofetil (DB00688)	0.73	0.23	[3, 20]
Censavudine (DB12074)	0.77	0.3	-
Bortezomib (DB00188)	0.7	0.3	[4, 16]

Mesalazine (DB00244)	0.66	0.27	[24]
Ramipril (DB00178)	0.67	0.23	[1]

According to Table 7, 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 use the compressed sensing technique, one of the matrix factorization methods, to predict the relationship between drugs and viruses. We also use the automatic encoder to find the number of the best properties of the drug, which is achieved by reducing the dimension of the drug properties and finally using these properties to predict whether there is a causal relationship

between the drug or not. After comparing our proposed approach with other matrix factorization methods, we identified drugs with a high potential for association with the Covid-19 virus. Because the proposed CSDVP model was able to predict the drug-virus relationship more accurately than other models, the identification of these drugs could help find drugs that are effective in treating Covid-19.

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] 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

[9] 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

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

[11] 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

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

[13] 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

[14] 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

[15] 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

[16] 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

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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

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

[24] 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

[25] 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

[26] 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

[27] Wang, Y. a. (2009). PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research*, W623--W633; doi.org/10.1093/nar/gkp456

[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