Contents lists available at ScienceDirect

Applied Soft Computing Journal

journal homepage: www.elsevier.com/locate/asoc

Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study

Yajie Meng, Min Jin *, Xianfang Tang, Junlin Xu *

College of Computer Science and Electronic Engineering, Hunan University, Changsha, Hunan, 410082, China

ARTICLE INFO

Article history:
Received 14 September 2020
Received in revised form 10 January 2021
Accepted 20 January 2021
Available online 23 January 2021

Dataset link: https://github.com/luckymengmeng/SCPMF

Keywords:
Drug–virus interactions prediction
Similarity constrained probabilistic matrix factorization
Drug repositioning
Drug–disease associations prediction
COVID-19

ABSTRACT

The novel coronavirus disease 2019 (COVID-19) pandemic has caused a massive health crisis worldwide and upended the global economy. However, vaccines and traditional drug discovery for COVID-19 cost too much in terms of time, manpower, and money. Drug repurposing becomes one of the promising treatment strategies amid the COVID-19 crisis. At present, there are no publicly existing databases for experimentally supported human drug–virus interactions, and most existing drug repurposing methods require the rich information, which is not always available, especially for a new virus. In this study, on the one hand, we put size-able efforts to collect drug–virus interaction entries from literature and build the Human Drug Virus Database (HDVD). On the other hand, we propose a new approach, called SCPMF (similarity constrained probabilistic matrix factorization), to identify new drug–virus interactions for drug repurposing. SCPMF is implemented on an adjacency matrix of a heterogeneous drug–virus network, which integrates the known drug–virus interactions, drug chemical structures, and virus genomic sequences. SCPMF projects the drug–virus interactions matrix into two latent feature matrices for the drugs and viruses, which reconstruct the drug–virus interactions matrix when multiplied together, and then introduces the weighted similarity interaction matrix as constraints for drugs and viruses. Benchmarking comparisons on two different datasets demonstrate that SCPMF has reliable prediction performance and outperforms several recent approaches. Moreover, SCPMF-predicted drug candidates of COVID-19 also confirm the accuracy and reliability of SCPMF.

1. Introduction

Human coronavirus 229E (HCoV-229E), HCoV-OC43, HCoV-NL63, HCoV-HKU1, Severe Acute Respiratory Syndrome (SARS)-associated coronavirus (SARS-CoV), and Middle East Respiratory Syndrome (MERS)-associated coronavirus (MERS-CoV) are known as six human coronaviruses so far. Specifically, HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1with lower pathogenicity are prevalent and generally cause common cold symptoms, while MERS-CoV and SARS-CoV are zoonotic in origin reported in the 21st century [1]. In December 2019, patients with pneumonia of unknown cause emerged in Wuhan, Hubei Province, China. The pathogen has been identified as a new enveloped RNA betacoronavirus2, considered a relative of SARS and MERS, named SARS-CoV-2 [2]. Subsequently, the novel disease was declared coronavirus disease 2019 (COVID-19) by the World Health Organization (WHO) [3]. Different from SARS-CoV and MERS-CoV, SARS-CoV-2 is the seventh member of the coronaviruses, which is the most pathogenic human coronavirus identified so far [4]. The COVID-19 epidemic has spread very quickly and has swept more than 210 countries around the world. As of 10th July 2020, the COVID-19 infections continue to rise, with 12,102,328 cases and over 551,046 deaths worldwide. To date, no proven effective drugs or vaccines are available for COVID-19 [3].

Despite a substantial increase in pharmaceutical companies' investment, the approval rate of new drugs has remained stable [5]. A recent study points out that the development of a new proven drug basically takes billions of dollars and an average of about 9-12 years to successfully bring it to the market [6]. Traditional *de novo* drug discovery takes more than 10 years and poses considerable difficulties (e.g. time-consumption, substantial-cost and high-risk) [7]. Drug repurposing (also known as drug repositioning) has emerged as a feasible solution to improve the overall productivity of drug development, which uses existing drugs to find potential drugs for treating new indications. Compared with the traditional drug discovery, drug repurposing can significantly shorten the drug development timelines, reduce overall development costs, and avoid risks. Therefore, drug repurposing is a promising strategy for accelerating drug discovery of COVID-19 and minimizing the translational gap in drug development.

The wet-lab experiments for drug repurposing are typically expensive and time-consuming [5]. As a supplement of experimental approaches, Computational methods offer novel testable

* Corresponding authors.
E-mail addresses: jinmin@hnu.edu.cn (M. Jin), xjl@hnu.edu.cn (J. Xu).

https://doi.org/10.1016/j.asoc.2021.107135
1568-4946/© 2021 Elsevier B.V. All rights reserved.

hypotheses for drug discovery [8,9]. Computational methods obtain the potential interaction candidates with superior accuracy in a short time by narrowing down the search space for drug–virus interactions and significantly reduce the experimental workload by suggesting potential interaction candidates for validation [10]. However, without the complex networks connecting viruses, drugs, and diseases, the development of affordable computational drug repositioning technologies for screening potential anti-COVID-19 drugs is challenging. To date, there are very few drug–virus interactions databases, leading to many knowledge gaps about the newly emerged COVID-19. Due to little information about the new COVID-19, the computational approaches can hardly predict the interaction between the new virus and any drugs. Additionally, coronaviruses are features with mutating rapidly, altering tissue tropism, crossing the species barrier, and have a high tendency towards frequent genetic mutations and gene recombination [11]. There is an urgent need for comprehensive databases of the drug–virus interactions in biomedical researches amid the COVID-19 crisis.

In this paper, we first manually collect a high number of drug–virus interactions entries from literature to build a human drug virus database (HDVD), which is a manually curated database of experimentally supported drugs associated with various viruses, and supplies a foundation for drug repositioning to help screen anti-viral drugs. HDVD is freely accessible at (https://github.com/luckymengmeng/HDVD) for fellow researchers. To the best of our knowledge, there are very few publicly existing databases for drug–virus interactions. We then propose a novel computational drug repositioning approach, similarity constrained probabilistic matrix factorization (called SCPMF), to precisely identify potential indications for existing drugs. Specifically, SCPMF first integrates the observed drug–virus interactions (i.e., HDVD), drug-drug similarities (i.e., chemical structure similarities of drug pairs), virus-virus similarities (i.e., genomic sequence similarities of virus pairs) into a heterogeneous network. Next, SCPMF projects the observed drug–virus interactions matrix into two latent feature matrices for drugs and viruses, and reconstructs the interaction matrix as a product of two lower-rank drugs and virus matrices. Different from the conventional probabilistic matrix factorization method, SCPMF takes the biological information of the problem into account by introducing the similarity information as constraints for drugs and viruses. To evaluate the effectiveness of SCPMF, we applied SCPMF to two benchmark datasets by adopting AUC and AUPR metrics in the 5-fold Cross-Validation (CV) and local Leave-One-Out-Cross-Validation (LOOCV) experiment. The results showed that SCPMF achieved the highest AUCs and AUPRs, outperforming the state-of-the-art approaches. Finally, the analyses of the SCPMF-predicted drug candidates for COVID-19 demonstrates that SCPMF is a useful method to prioritize existing drugs for further investigation, which has the potential to accelerate drug discovery for COVID-19 and other emerging viral infections diseases.

Theoretically, the main contributions of this work as follows: (i) We develop a drug–virus interactions dataset named HDVD. To the best of our knowledge, there are very few publicly existing databases for drug–virus interactions. (ii) SCPMF respectively introduces similarity constraints for drugs and viruses into the probabilistic matrix factorization process, and hence leverages the biological information of the problem to boost the performance of SCPMF. (iii) We build a powerful computational drug repositioning methodology, which is complementary to existing experimental methods for rapidly and precisely discovering drug candidates for COVID-19 and other emerging viral infections diseases.

2. Material and methods

As shown in Fig. 1, SCPMF involves five steps: (i) construct the drug–virus network, (ii) calculate drug similarity scores by the similar chemical structure of drugs, (iii) calculate virus similarity scores by the genomic sequence of viruses, (iv) integrate the known drug–virus interactions, drug-drug similarities, virus-virus similarities networks to construct a heterogeneous network, and (v) use the proposed similarity constrained probabilistic matrix factorization approach to help suggest potential therapeutic drugs for COVID-19 and other emerging viral infections diseases based on the constructed heterogeneous network.

2.1. Human drug virus database (HDVD)

To construct the human virus-drug interactions network, we assembled a significant number of experimentally validated drug–virus interaction entries from literature by text mining technology and then built the HDVD, which is a database for experimentally supported human drug–virus associations. HDVD is freely accessible for the researcher. HDVD includes 34 viruses, 219 drugs, and 455 confirmed human drug–virus interactions.

2.2. Construction of the drug–virus interactions network

We use the known drug–virus interactions in HDVD to construct a drug–virus interactions network, where drugs and viruses are the nodes, and interactions between drugs and viruses are the edges. Let $G = (\mathrm{D},\ \mathrm{V},\ \mathrm{I})$ represent the drug–virus interactions network, where $D = \{d_1, d_{2\,\ldots,d_n}\}$ is the collection of drugs, $V = \{v_1, v_2, \ldots, v_m\}$ is the collection of viruses, and I is the collection of interactions between D and V. Let A_{n*m} represent the adjacency matrix of G. If d_i and v_j is associated, $A_{i,j} = 1$, otherwise $A_{i,j} = 0$. A^T represents the transpose of A_{n*m}.

2.3. Chemical structure similarity of drug pairs

A popular molecular structure 1D representation is SMILES (Simplified Molecular Input Line Entry System), which describes molecular structures in the form of special strings [12]. SMILES is the most compact text-based molecular representation and implicitly contains the information needed to compute all kinds of molecular structures, which has been used to obtain molecular similarity [13]. Therefore, we downloaded drug chemical structure information from the DrugBank database by adopting the SMILES format. We then calculated the Molecular Access System (MACCS) fingerprints of each drug via Open Babel v2.3.1 [14]. In this study, we used the Tanimoto index to measure the absolute similarity between two molecules. Tanimoto index is the most popular fingerprint-based molecular similarity metric in cheminformatics-related fields [15]. Let two drug molecules have a and b bits set in their MACCS fragment bit-strings, with c of these bits being set in the fingerprints of both drugs. Tanimoto index of a drug pair is given by [8]:

$$T = \frac{c}{a + b - c} \tag{1}$$

T is a value in the range of zero to one where zero means that no bits are common, and one means that all bits are the same.

2.4. Virus genomic sequence similarity

We downloaded the genome nucleotide sequence of viruses in *Homo sapiens* from the National Center for Biotechnology Information (NCBI). MAFFT is becoming popular in recent years because of its high performance. In this work, we calculated the sequence similarity between viruses by using MAFFT version

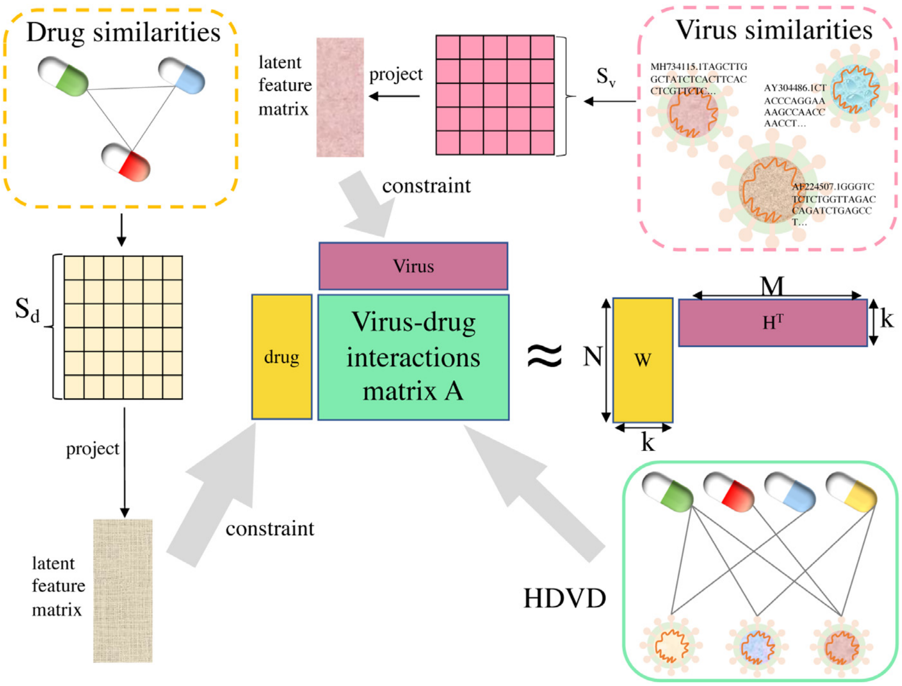
Drug similarities
latent feature matrix
project
Virus similarities
S_v
MH734115.1TAGCTTG GCTATCTCACTTCAC CTCGTTCTC...
AY304486.1CT ACCCAGGAA AAGCCAACC AACCT...
AF224507.1GGGTC TCTCTGGTTAGAC CAGATCTGAGCC T...
constraint
Virus
S_d
drug
Virus-drug interactions matrix A
≈
N
W
k
M
H^T
k
project
constraint
latent feature matrix
HDVD

Fig. 1. The workflow of SCPMF. We first create a confirmed drug–virus interactions network (denoted as the interaction matrix A with N drugs and M viruses) by developing HDVD. Due to the rapid mutation of SARS-CoV-2 so far, we focus on the chemical structure similarities of drug pairs (denoted as the similarity matrix S_d) and the genomic sequence similarities of virus pairs (denoted as the similarity matrix S_V). Afterwards, SCPMF integrates the drug-drug similarities, virus-virus similarities and confirmed drug–virus interactions networks to construct a heterogeneous network. Lastly, SCPMF projects the matrix A into two low-rank matrices (i.e., W and H) consisting of latent features for drugs and viruses, and then respectively introduces similarity as constraints for drugs and viruses in low-rank spaces. The SCPMF-predicted candidate drugs can be further analyzed and experimentally validated.

7 [16]. MAFFT is a multiple sequence alignment similarity-based method and offers various alignment strategies, such as progressive methods (e.g. PartTree recommended for a large-scale alignment), iterative refinement methods (e.g. FFT-NS-I suggested for a small-scale alignment), and structural alignment methods (e.g. Q-INS-I proposed for a small-scale RNA alignment).

2.5. SCPMF

In this study, we developed a human drug virus database (named HDVD) and a novel similarity constrained probabilistic matrix factorization methodology (called SCPMF), a practically useful framework, to effectively identify prospective drugs for the potential treatments of COVID-19 and other emerging viral infections diseases. Fig. 1 illustrates the basic idea of SCPMF.

In HDVD, the observed drug–virus interactions can be denoted as a binary matrix $A \in \{0,1\}^{N\times M}$ with N drugs and M viruses. A_{ij} is the (i,j)th entry of A. A_{ij} is equal to 1, if a drug d_i has interaction with a virus v_j; otherwise $A_{ij} = 0$. The pairwise chemical structure similarities between N drugs are denoted as a drug-drug similarity matrix S_d; the pairwise genomic nucleotide sequence similarities between M viruses are denoted as a virus-virus similarity matrix S_v. The value range of S_d and S_v is [0, 1]. Let $W \in R^{K\times N}$ and $H \in R^{K\times M}$ represent the latent drug and virus feature matrices, W_i and H_j represent drug-specific and virus-specific latent feature vectors, respectively. Then, our goal is to find drug and virus latent models ($W \in R^{K\times N}$ and $H \in R^{K\times M}$) whose product (W^TH) reconstructs the interaction matrix A. In probabilistic point of view, the conditional distribution on the observed interactions $A \in \{0,1\}^{N\times M}$ is given by:

$$P\left(A|W,H,\sigma^2\right) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}\left(A_{ij}|W_i^TH_j,\sigma^2\right)\right]^{I_{ij}} \tag{2}$$

where $\mathcal{N}\left(x|\mu,\sigma^2\right)$ is the probability density function of the Gaussian normal distribution, with mean μ and variance σ^2, and I_{ij} is the indicator function that is equal to 1 if the drug d_i and the virus v_j is connected and equal to 0 otherwise. Thus, $P\left(A|W,H,\sigma^2\right)$ provides us a probabilistic representation of the interaction matrix A. As a generative model for drug and virus latent models, we place zero-mean spherical Gaussian priors on drug and virus feature vectors as follows:

$$P\left(W|\sigma_W^2\right) = \prod_{i=1}^{N}\mathcal{N}\left(W_i|0,\sigma_W^2 I\right) \tag{3}$$

$$P\left(H|\sigma_H^2\right) = \prod_{j=1}^{M}\mathcal{N}\left(H_j|0,\sigma_H^2 I\right) \tag{4}$$

where $\boldsymbol{I}$ is a K-dimensional identity diagonal matrix. Thereafter, we take the log of the posterior distribution over the drug and virus features and transform it (see Supplementary File Equations 4 and 5). Maximizing the log-posterior over drug and virus features with hyperparameters kept fixed is equivalent to minimizing the sum-of-squared-errors objective function with quadratic regularization terms:

$$\min_{W_i,H_j}\sum_{i=1}^{N}\sum_{j=1}^{M}I_{ij}\left(A_{ij}-W_i^TH_j\right)^2+\frac{\lambda_W}{2}\sum_{i=1}^{N}\|W_i\|_{Fro}^2+\frac{\lambda_H}{2}\sum_{j=1}^{M}\left\|H_j\right\|_{Fro}^2 \tag{5}$$

where $\lambda_W = \sigma^2/\sigma_W^2$ and $\lambda_H = \sigma^2/\sigma_H^2$, and $\|\cdot\|_{Fro}^2$ denotes the Frobenius norm.

However, the conventional probabilistic matrix factorization model still has some room for improvement, which simply utilizes a probabilistic linear model with Gaussian noise to model the drug–virus interactions. Different from probabilistic matrix factorization, SCPMF takes the biological information of the problem (i.e., drug and virus similarity) into account. Therefore, we propose a new objective function of SCPMF as follows:

$$\min_{W_i,H_j} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}\left(A_{ij}-W_i^T H_j\right)^2+\frac{\lambda_W}{2}\sum_{i=1}^{N}\|W_i\|_{Fro}^2 + \frac{\lambda_H}{2}\sum_{j=1}^{M}\left\|H_j\right\|_{Fro}^2+\frac{\lambda_1}{2}\left\|W^TW-S_d\right\|_{Fro}^2+\frac{\lambda_2}{2}\left\|H^TH-S_v\right\|_{Fro}^2 \tag{6}$$

where W_i represents the K-dimensional latent feature vector for the drug, W^TW is the drugs weighted similarity matrix, and H^TH is the viruses weighted similarity matrix.

To improve the efficiency of SCPMF, we optimize the problem in Eq. (6) by utilizing the gradient descent algorithm. We define the corresponding Lagrange function $\mathcal{L}_f$ of Eq. (6), and obtain the partial derivatives equations of W and H. The detailed derivation process is shown in the Supplementary File Equations 8–12. Thus, we can obtain the updating rules as follows:

$$W_{ik}^{new} \leftarrow W_{ik}\frac{\left(I\cdot\left(HA^T\right)+2\lambda_1\left(W\left(S_d\right)\right)\right)_{ik}}{\left(I\cdot\left(HH^TW\right)\right)_{ik}+\left(\lambda_W W\right)_{ik}+\left(2\lambda_1\left(WW^TW\right)\right)_{ik}} \tag{7}$$

$$H_{jk}^{new} \leftarrow H_{jk}\frac{\left(I\cdot(WA)+2\lambda_1\left(W\left(S_v\right)\right)\right)_{jk}}{\left(I\cdot\left(WW^TH\right)\right)_{jk}+\left(\lambda_H H\right)_{jk}+\left(2\lambda_2\left(HH^TH\right)\right)_{jk}} \tag{8}$$

The matrices W and H are updated based on Eqs. (7) and (8) until the local minimum of the objective function. Finally, the predicted drug–virus interaction matrix is obtained as $A^* = W^TH$. Generally, the jth column of A^* indicates the interaction scores between virus v_j and drugs, and the larger the score, the more relevant it is.

3. Performance evaluation of SCPMF

3.1. Prediction of drug–virus interactions

We performed 5-fold CV and local LOOCV procedures based on 455 known drug–virus interactions between 219 drugs and 34 viruses from HDVD to test the performance of SCPMF. In the 5-fold CV experiment, the known interacting drug–virus pairs and the non-interacting drug–virus pairs were randomly divided into five parts, in each fold, the four parts of drug–virus pairs were selected as the training set, and the remaining one part of drug–virus pairs were held out as the testing part. We repeated the selection five times to ensure that each of the five parts was considered the testing part. In the local LOOCV experiment, each drug related to the jth virus was repeatedly left out in turn as the testing data, while $A_{n*(m-1)}$ is considered as the training data and the range of j is [1, m].

3.2. Evaluation metrics

In order to test the performance of SCPMF, we adopted both 5-fold CV and local LOOCV experiments. After all interactions have been tested, we calculate both True Positive Rate (TPR), False-Positive Rate (FPR), and Precision as follows:

$$TPR\,(or\ Recall)=\frac{TP}{TP+FN} \tag{9}$$

where TP is the number of positive samples identified correctly and FN is the number of negative samples identified incorrectly. TPR is the ratio of positive samples identified correctly among the total positive samples.

$$FPR=\frac{FP}{TN+FP} \tag{10}$$

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

where FP is the number of positive samples identified incorrectly and TN represents the number of negative samples identified correctly. FPR is the proportion of misidentified negative samples accounting for all the negative samples, and Precision is the percentage of the correctly identified positive samples among the retrieved samples. The larger Precision value means the better prediction performance.

For both 5-fold CV and local LOOCV, the testing data are ranked by SCPMF, a rank exceeding a preset threshold indicates a successful prediction and vice versa. By varying the rank threshold, we calculated TPR, FPR and Precision to construct the Receiver Operating Characteristic (ROC) curve and the Precision-Recall (PR) curve. The ROC curve is a probability curve where FPR is on the x-axis and TPR is on y-axis at various thresholds. AUC is the area under the ROC curve, which is widely used for describing the global prediction performance [17]. An AUC of one indicates an excellent performance whereas an AUC of 0.5 suggests a random performance [18]. Because the representation of the Precision-Recall curve (PR) is more effective than ROC on highly imbalanced or skewed datasets, we also utilize the PR curve and the area under the PR curve (AUPR) to comprehensively evaluate the performance of SCPMF. The larger the AUPR value, the better the prediction performance.

4. Results

Since the current researches on COVID-19 are mostly based on sequence information, there are very few drug–virus interactions prediction methods for repurposing existing drugs for COVID-19. However, the general problem of drug–virus interactions can actually be thought of as the network association prediction. In this work, we performed SCPMF and several association prediction approaches: IMCMDA [19], NCPMDA [20], RLSMDA [21], BNNR [22] on the HDVD. Specifically, Chen et al. proposed Inductive Matrix Completion for MiRNA-Disease Association prediction (IMCMDA) [19]. IMCMDA used the known miRNA-disease associations, the integrated miRNA similarity, and disease similarity to complete the missing miRNA-disease association by calculating the miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity. Gu et al. designed Network Consistency Projection for MiRNA-Disease Associations (NCPMDA) to identify the potential disease-related miRNAs [20]. NCPMDA is a non-parametric universal network-based method and does not require negative samples. It not only can identify the miRNA-disease associations in all diseases but also can predict the associations between miRNAs and isolated diseases (diseases without any known miRNA association is similar to COVID-19 without any known anti-viral drugs). Chen et al. presented a semi-supervised method, the Regularized Least Squares for MiRNA-Disease Association method (RLSMDA), to prioritize the miRNAs candidates for all the diseases simultaneously. RLSMDA also does not need negative samples [21]. Yang et al. developed a Bounded Nuclear Norm Regularization method (BNNR), which integrated drug-drug, Drug–disease and disease-disease networks and incorporated nuclear norm regularization and additional constraints to complete the Drug–disease matrix under the low-rank assumption [22]. For all parameters, we adopt the parameters of the previous published methods provided by the authors. Specifically, $\lambda_W = \lambda_H = 1, \lambda_1 = \lambda_2 = 0.1$ for SCPMF, NCPMDA is parameter-free, $W = 0.9$ for RLSMDA, and $\lambda_1 = \lambda_2 = 1$ for IMCMDA.

4.1. Performance of SCPMF in 5-fold CV

We compared the SCPMF with above-mentioned association prediction methods: Inductive Matrix Completion (IMC), Regularized Least Squares (RLS), Network Consistency Projection (NCP), Bounded Nuclear Norm Regularization (BNNR) for drug–virus interactions prediction. To compare the experimental results, we plotted ROC and PR curves and calculated AUC and AUPR based on the 5-fold CV (see Fig. 2). We found that SCPMF achieved superior performance (AUC = 0.8631 and AUPR = 0.509) in 5-fold CV experiment, outperforming that of the state-of-the-art approaches: IMC (AUC = 0.6423 and AUPR = 0.1649), RLS (AUC= 0.7341 and AUPR = 0.1792), NCP (AUC = 0.6711 and AUPR = 0.0908) and BNNR (AUC = 0.8537 and AUPR = 0.3947).

4.2. Performance of SCPMF in the local LOOCV

To fully evaluate the effectiveness of SCPMF, the local LOOCV experiment was carried on the HDVD (see Fig. 3). SCPMF showed a higher performance over other approaches in terms of both AUC and AUPR. Specifically, SCPMF obtained AUC value of 0.6936, outperforming that of IMC (0.5280), RLS (0.6480), NCP (0.6476), BNNR (0.6914). SCPMF achieved AUPR value of 0.1931, outperforming that of IMC (0.1092), RLS (0.1373), NCP (0.0831), BNNR (0.1657) as well.

4.3. SCPMF identifies the potential drugs for COVID-19

Table 1
The top 15 anti-COVID-19 drug candidates identified by SCPMF.

Rank	Accession number	Drug name	Evidence
1	DB00811	Ribavirin	[23]
2	DB00608	Chloroquine	[30]
3	DB00507	Nitazoxanide	[39]
4	DB13729	Camostat	[37,38]
5	DB13609	Umifenovir	[41]
6	DB15660	N4-Hydroxycytidine	Unconfirmed
7	DB12617	Mizoribine	Unconfirmed
8	DB06803	Niclosamide	[36]
9	DB04115	Berberine	Unconfirmed
10	DB14761	Remdesivir	[26,28]
11	DB12466	Favipiravir	[40]
12	DB01024	Mycophenolic Acid	Unconfirmed
13	DB00864	Tacrolimus	Unconfirmed
14	DB13393	Emetine	[29]
15	DB07715	Emodin	Unconfirmed

We listed top 15 SCPMF-predicted drugs for COVID-19 in Table 1. For each drug, we showcased the rank (predicted score sorted by descending order), Accession Number in DrugBank, canonical name, the literature-reported evidence. Among the top 15 drug candidates ranked according to the final predicted association scores, nine drugs (60% success rate) were verified by various evidences. Ribavirin, was initially recommended in clinical practice for the China 2019-nCoV pneumonia diagnosis and Treatment Plan Edition 5-Revised [23]. Herein, ribavirin is the top first predicted candidate for potentially treating COVID-19. Remdesivir, a nucleotide analogue prodrug, has a broad antiviral spectrum including filoviruses, pneumoviruses, paramyxoviruses, and coronaviruses [24,25]. Remdesivir inhibited viral RNA polymerases and had shown in vitro activity against COVID-19 [26–28]. Ref. [29] indicated the combination of remdesivir and emetine therapy may provide better clinical benefits. Chloroquine, a cheap, safe, and broadly used antimalarial drug that has been used for more than 70 years, was highly effective in controlling COVID-19 infection in vitro and thus may be clinically applicable against COVID-19 [30]. Ref. [31] again revealed that chloroquine and remdesivir are highly effective in controlling COVID-19 infection in vitro. Niclosamide was an FDA-approved anthelminthic drug regulating multiple signaling pathways and biological processes and identified as a multifunctional drug [32,33]. For example, niclosamide could effectively resist various viral infections such as SARS-CoV, MERS-CoV, ZIKV, HCV, and human adenovirus [34,35]. Ref. [36] envisioned that niclosamide might offer the therapeutic potential to battle COVID-19. Nitazoxanide, FDA approved drug potentiates host antiviral response, thereby reducing viral replication, titer, and ensuing immune dysregulation. Camostat mesylate, an ingredient of the camostat, can block SARS-CoV-2 infection of lung cells and could be considered for off-label treatment of COVID-19 infections [37,38]. Based on the combined pathophysiological and pharmacological potential, camostat and nitazoxanide combination potentially recommended for early evaluation and clinical trials of COVID-19 [39]. The results of Ref. [40] provided the preliminary evidence that Favipiravir can treat the SARS-CoV-2 infection. Umifenovir is a broad-spectrum antiviral drug. Clinical trials with umifenovir alone have been recently initiated in China [41]. Specially, almost of the top 15 SCPMF-predicted anti-COVID-19 drugs can be found in Ref. [41].

Furthermore, we analyzed unconfirmed drugs using the molecular docking approach, which characterizes the behavior of small molecules in the binding site of target proteins and models the interaction between a small molecule and a protein at the atomic level [42]. The cellular receptor angiotensin-converting enzyme 2 (ACE2) is considered to be an important functional receptor for SARS and other coronaviruses [43]. Like SARS-CoV, SARS-CoV-2 invades through the mediation of S-protein and ACE2 receptors on the human cell surface to infect human respiratory epithelial cells. With the disclosure of ACE2 as a target for SARS-CoV-2, blocking the combination of SARS-CoV-2 and ACE2 becomes one of the treatment options [44]. In this work, we examined the binding mode of SCPMF-predicted drugs to the cellular receptor ACE2 using molecular docking (see Fig. 4). Fig. 4 reveals that predicted drugs interact with multiple important binding sites on the cellular ACE2, especially the three unconfirmed drugs (N4-Hydroxycytidine, Berberine, Mycophenolic Acid). It again shows that the drug candidates identified by SCPMF have therapeutic effects on COVID-19.

4.4. Parameter sensitivity analysis

Parameter sensitivity analysis is significant for the performance of a model in different scenarios. Thus, we mainly focus on the λ_W, λ_H and λ_1, λ_2 and conducted 5-fold CV on the HDVD to tune the parameters of SCPMF. Specifically, the parameters λ_W, λ_H and λ_1, λ_2 are increased from 0.1 to 1 with a step of 0.1. Finally, we selected the parameters with the highest AUC and AUPR. The *x*-axis represents λ_1, λ_2 and *y*-axis represents λ_W, λ_H. Fig. 5A showed that the change of parameters has minimal effect on AUC. As seen in Fig. 5B, it is indicated that the influence of parameters is relatively significant for AUPR but within an acceptable range. The results further confirm the robustness of SCPMF to parameters.

4.5. Experiment on the other dataset

To test the adaptability of SCPMF for different datasets and make the prediction more convincing, we perform SCPMF on the other dataset named as Cdatasets [22] by performing 5-fold cross-validation. Cdatasets is generated by combining DNdatasets [45] and the gold standard dataset [46], which contains 663 drugs, 409 diseases, and 2352 known Drug–disease associations. It shows that SCPMF obtains an AUC value of 0.9216 in 5-fold CV, while

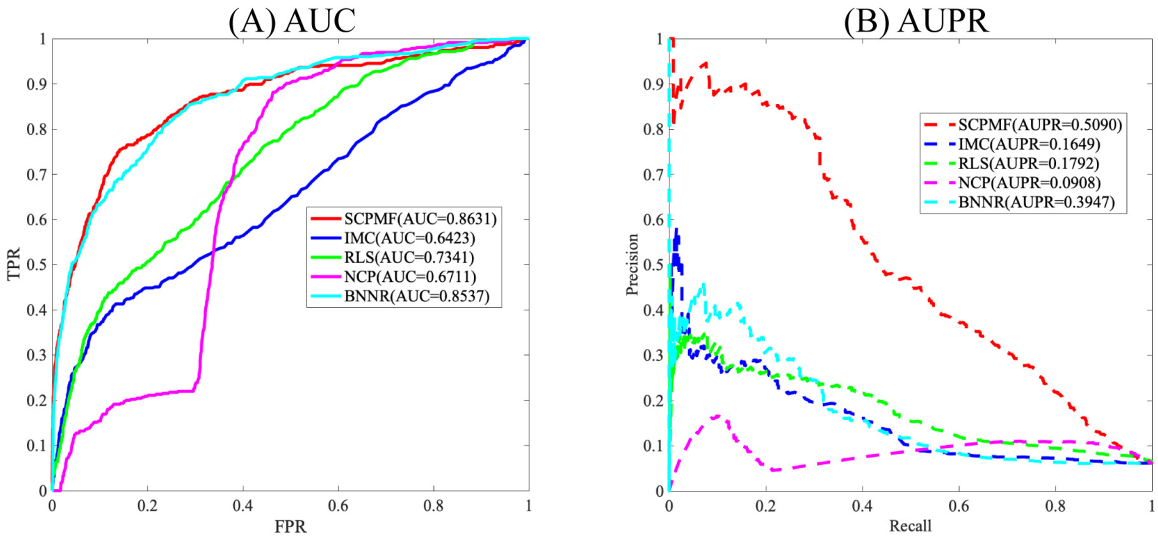

Fig. 2. Evaluation of SCPMF on HDVD in the 5-fold CV experiment. **(A)** The ROC curves of prediction results. **(B)** The PR curves of prediction results. As is shown, compared with other four methods, DRHGCN has achieved excellent performance (AUC=0.8631 and AUPR=0.509). The results indicate that SCPMF has a superior ability to accurately discover potential drug–virus interactions.

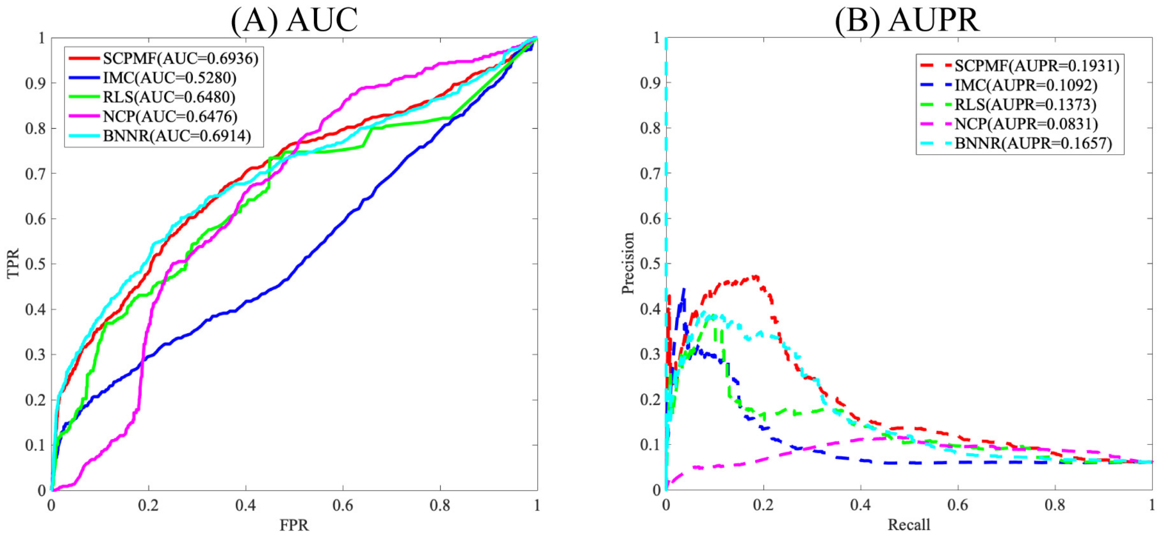

Fig. 3. Evaluation of SCPMF on HDVD in the local LOOCV experiment. **(A)** The ROC curves of prediction results. **(B)** The PR curves of prediction results. As is shown, SCPMF achieves an AUC of 0.6936 and an AUPR of 0.1931 in the local LOOCV, which is superior to the other four state-of-the-art methods. It again demonstrates that SCPMF achieves convincing performance for the discovery of potential drug–virus interactions, and thus can help identify high-confidence repurposed candidate drugs for COVID-19 and other emerging viral infections diseases.

IMC, RLS, NCP and BNNR have 0.6436, 0.8015, 0.8316, and 0.9103, respectively (see Fig. 6). The excellent performance evaluation results on Cdatasets further prove that SCPMF is reliable for drug repositioning.

5. Conclusion

As the COVID-19 pandemic is still rapidly spreading worldwide leading to a colossal toll in human suffering and lives, physicians are trying to search for effective antiviral therapies to save lives. Although multiple COVID-19 vaccine trials are underway, there is no enough vaccines for everyone in a short period of time or specific antiviral medication for COVID-19.

In this study, to fight the emerging COVID-19 pandemic, we put great efforts to create a human drug virus database (named HDVD). On the other hand, we proposed a novel similarity constrained probabilistic matrix factorization methodology, called SCPMF, to help identify high-confidence drug candidates for the potential treatment of COVID-19 and other emerging viral infections diseases. Specifically, due to the rapid mutation of SARS-CoV-2 so far, we focused on the chemical structure similarities of drug pairs and the genomic sequence similarities of virus pairs to obtain the drug-drug similarities network and the virus-virus similarities network. Then, we embedded them with the observed drug–virus interactions network to construct a heterogeneous network. Lastly, different from the classic probabilistic matrix factorization method, which adopts a probabilistic linear model with Gaussian noise to model the drug–virus interactions matrix as a product of two lower-rank drug and virus matrices. SCPMF introduces similarity constraints for drugs and viruses into the probabilistic matrix factorization process, hence leveraging the biological information of the problem boosts the performance of the model. We have validated the performance of SCPMF in terms of 5-fold CV, local LOOCV, and the other external dataset. Experimental results demonstrated that SCPMF

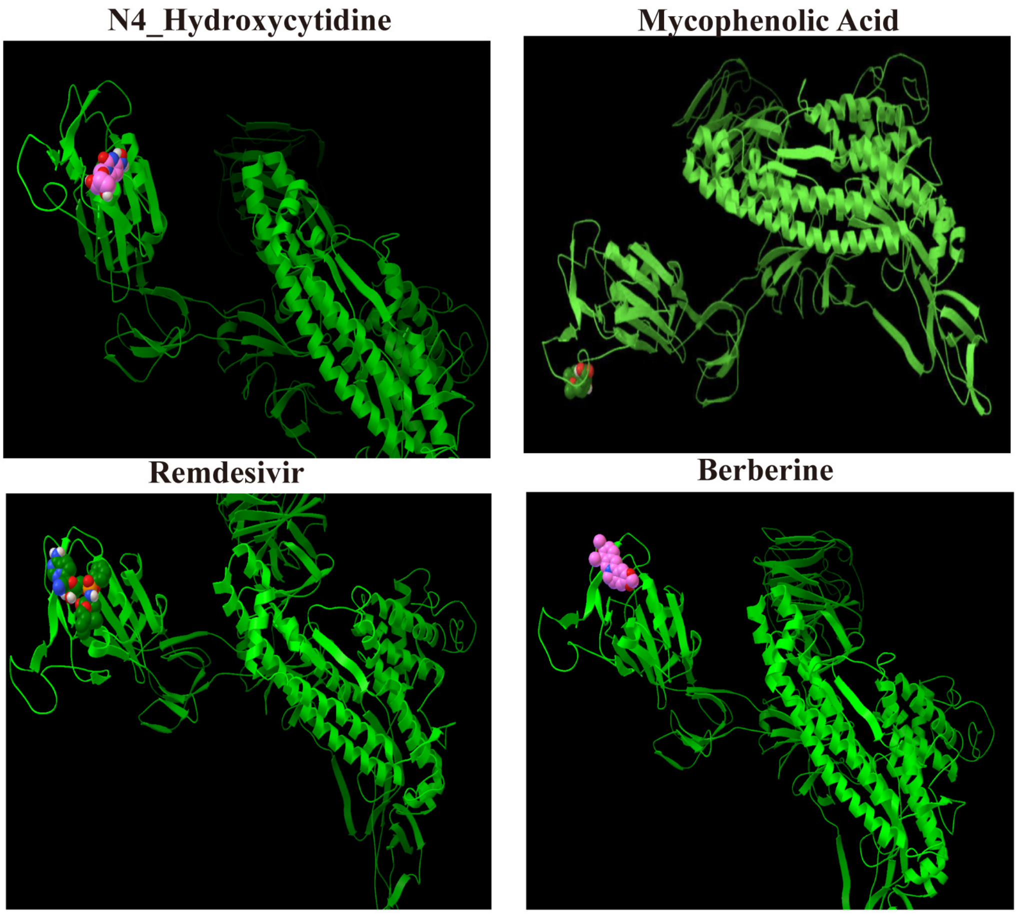

Fig. 4. The predicted ligand–protein binding mode between the drugs and the cellular receptor ACE2 using molecular docking. As is shown, in addition to Remdesivir, the other three SCPMF-predicted anti-COVID-19 drugs (i.e., N4-Hydroxycytidine, Berberine and Mycophenolic Acid, which have not been confirmed to be effective against COVID-19 so far) all interact with binding sites on ACE2. The results signify that SCPMF-predicted drug candidates have great potential efficacy against the COVID-19. We expect that the predicted candidate drugs targeting the emerging COVID-19 will provide a meaningful Ref. to assist clinicians.

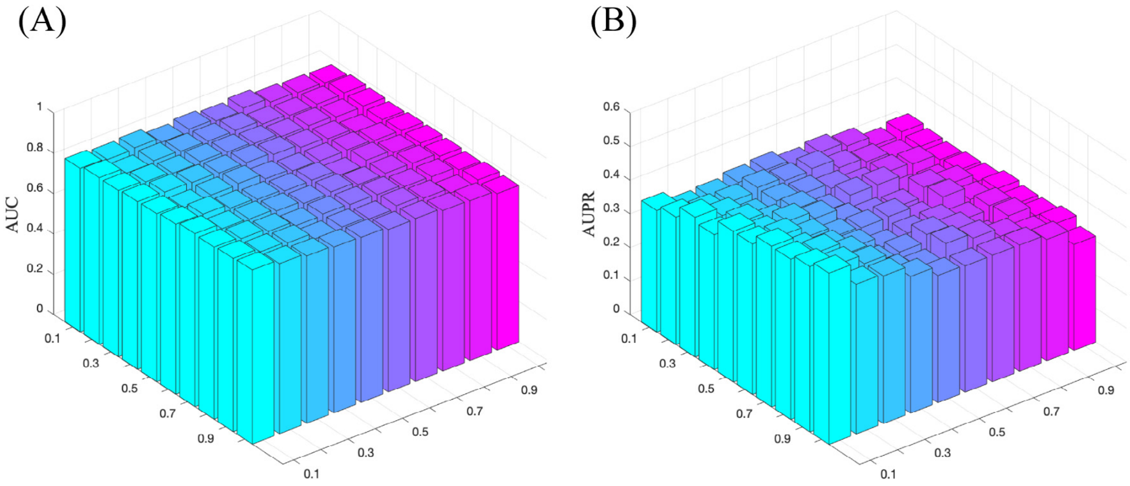

Fig. 5. The influence of parameters on the performance of SCPMF. **(A)** The influence of parameters on the AUC. **(B)** The influence of parameters on the AUPR. It shows that SCPMF is highly robust to parameter settings.

achieved convincing performance for the rapid discovery of candidate drugs for COVID-19 and other diseases and was superior to the state-of-the-art prediction methods.

Although we make the utmost efforts to collect the experimentally reported drug–virus interactions from clinical researches and published literature, the drug–virus interactions may be incomplete. In future work, we will capture comprehensive information to improve the HDVD. In summary, we presented HDVD and SCPMF, a practically useful framework which can help effectively identify prospective drugs for COVID-19 and other emerging viral infections diseases. Our proposed method can minimize the translational gap between preclinical testing

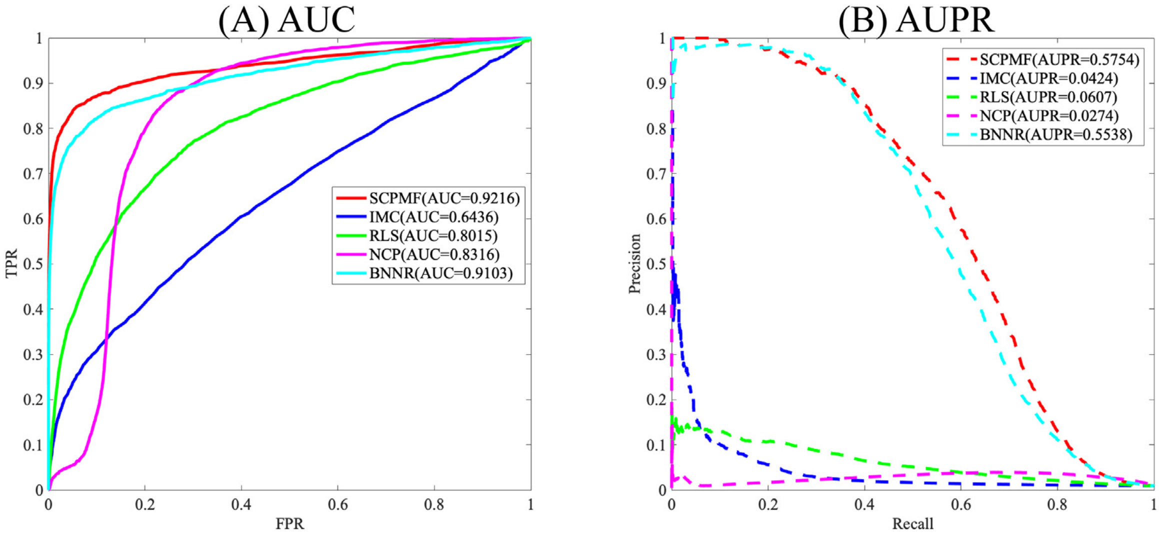

Fig. 6. Evaluation of SCPMF on the other Drug–disease associations dataset named Cdatasets in the 5-fold CV experiment. **(A)** The ROC curves of prediction results. **(B)** The PR curves of prediction results. As is shown, compared with other four methods, SCPMF had strong adaptability on different datasets, achieving the highest AUC of 0.9216 and AUPR of 0.5754 in the 5-fold CV. It signifies that SCPMF has a high generalization ability to assist the drug repositioning.

outcomes and clinical results, which is a significant problem in drug development.

CRediT authorship contribution statement

Yajie Meng: Entire processing of this study. **Min Jin:** Supervision. **Xianfang Tang:** Data curation. **Junlin Xu:** Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability

The source code and data are available at https://github.com/luckymengmeng/SCPMF.

Funding

The work was supported by the National Natural Science Foundation of China Grant 61773157 and the Hunan Provincial Innovation Foundation for Postgraduate, China under Grant No. CX20200434.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.asoc.2021.107135.

References

[1] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, A novel coronavirus from patients with pneumonia in China, 2019, New Engl. J. Med. (2020).

[2] C. Sohrabi, Z. Alsafi, N. O'Neill, M. Khan, A. Kerwan, A. Al-Jabir, C. Iosifidis, R. Agha, World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19), Int. J. Surg. (2020).

[3] W.-j. Guan, Z.-y. Ni, Y. Hu, W.-h. Liang, C.-q. Ou, J.-x. He, L. Liu, H. Shan, C.-l. Lei, D.S. Hui, Clinical characteristics of coronavirus disease 2019 in China, N. Engl. J. Med. 382 (2020) 1708–1720.

[4] Y. Cha, T. Erez, I. Reynolds, D. Kumar, J. Ross, G. Koytiger, R. Kusko, B. Zeskind, S. Risso, E. Kagan, Drug repurposing from the perspective of pharmaceutical companies, Br. J. Pharmacol. 175 (2018) 168–180.

[5] J. Avorn, The $2.6 billion pill–methodologic and policy considerations, N. Engl. J. Med. 372 (2015) 1877–1879.

[6] S. Liu, Q. Zheng, Z. Wang, Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus, Bioinformatics 36 (2020) 3295–3298.

[7] S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, Drug repurposing: progress, challenges and recommendations, Nature Rev. Drug Discov. 18 (2019) 41–58.

[8] X. Zeng, S. Zhu, W. Lu, Z. Liu, J. Huang, Y. Zhou, J. Fang, Y. Huang, H. Guo, L. Li, Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (2020) 1775–1797.

[9] Q. Zhao, H. Yu, M. Ji, Y. Zhao, X. Chen, Computational model development of drug-target interaction prediction: a review, Curr. Protein Pept. Sci. 20 (2019) 492–494.

[10] H. Lu, Drug treatment options for the 2019-new coronavirus (2019-nCoV), Biosci. Trends 14 (2020) 69–71.

[11] Y.A. Helmy, M. Fawzy, A. Elaswad, A. Sobieh, S.P. Kenney, A.A. Shehata, The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control, J. Clin. Med. 9 (2020) 1225.

[12] D. Vidal, M. Thormann, M. Pons, LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities, J. Chem. Inf. Model. 45 (2005) 386–393.

[13] H. Öztürk, E. Ozkirimli, A. Özgür, A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction, BMC Bioinformatics 17 (2016) 128.

[14] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminformatics 3 (2011) 33.

[15] D. Bajusz, A. Rácz, K. Héberger, Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J. Cheminformatics 7 (2015) 20.

[16] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.

[17] Q. Zou, J. Li, Q. Hong, Z. Lin, Y. Wu, H. Shi, Y. Ju, Prediction of microRNA-disease associations based on social network analysis methods, BioMed Res. Int. 2015 (2015).

[18] L.-H. Peng, L.-Q. Zhou, X. Chen, X. Piao, A computational study of potential mirna-disease association inference based on ensemble learning and kernel ridge regression, Front. Bioeng. Biotechnol. 8 (2020).

[19] X. Chen, L. Wang, J. Qu, N.-N. Guan, J.-Q. Li, Predicting miRNA–disease association based on inductive matrix completion, Bioinformatics 34 (2018) 4256–4265.

[20] C. Gu, B. Liao, X. Li, K. Li, Network consistency projection for human mirna-disease associations inference, Sci. Rep. 6 (2016) 36054.

[21] X. Chen, G.-Y. Yan, Semi-supervised learning for potential human microrna-disease associations inference, Sci. Rep. 4 (2014) 5501.

[22] M. Yang, H. Luo, Y. Li, J. Wang, Drug repositioning based on bounded nuclear norm regularization, Bioinformatics 35 (2019) i455–i463.

[23] J.S. Khalili, H. Zhu, N.S.A. Mak, Y. Yan, Y. Zhu, Novel coronavirus treatment with ribavirin: Groundwork for an evaluation concerning COVID-19, J. Med. Virol. (2020).

[24] M.K. Lo, R. Jordan, A. Arvey, J. Sudhamsu, P. Shrivastava-Ranjan, A.L. Hotard, M. Flint, L.K. McMullan, D. Siegel, M.O. Clarke, GS-5734 and its parent nucleoside analog inhibit Filo-, Pneumo-, and Paramyxoviruses, Sci. Rep. 7 (2017) 43395.

[25] T.P. Sheahan, A.C. Sims, R.L. Graham, V.D. Menachery, L.E. Gralinski, J.B. Case, S.R. Leist, K. Pyrc, J.Y. Feng, I. Trantcheva, Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses, Sci. Transl. Med. 9 (2017).

[26] J.A. Al-Tawfiq, A.H. Al-Homoud, Z.A. Memish, Remdesivir as a possible therapeutic option for the COVID-19, Travel Med. Infect. Dis. (2020).

[27] E. de Wit, F. Feldmann, J. Cronin, R. Jordan, A. Okumura, T. Thomas, D. Scott, T. Cihlar, H. Feldmann, Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection, Proc. Natl. Acad. Sci. 117 (2020) 6771–6776.

[28] J. Grein, N. Ohmagari, D. Shin, G. Diaz, E. Asperges, A. Castagna, T. Feldt, G. Green, M.L. Green, F.-X. Lescure, Compassionate use of remdesivir for patients with severe Covid-19, New Engl. J. Med. 382 (2020) 2327–2336.

[29] F. Touret, X. de Lamballerie, Of chloroquine and COVID-19, Antiviral Res. (2020) 104762.

[30] K.-T. Choy, A.Y.-L. Wong, P. Kaewpreedee, S.-F. Sia, D. Chen, K.P.Y. Hui, D.K.W. Chu, M.C.W. Chan, P.P.-H. Cheung, X. Huang, Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro, Antiviral Res. (2020) 104786.

[31] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong, G. Xiao, Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro, Cell Res. 30 (2020) 269–271.

[32] Y. Li, P.-K. Li, M.J. Roberts, R.C. Arend, R.S. Samant, D.J. Buchsbaum, Multi-targeted therapy of cancer by niclosamide: A new application for an old drug, Cancer Lett. 349 (2014) 8–14.

[33] Z. Li, M. Brecher, Y.-Q. Deng, J. Zhang, S. Sakamuru, B. Liu, R. Huang, C.A. Koetzner, C.A. Allen, S.A. Jones, Existing drugs as broad-spectrum and potent inhibitors for zika virus by targeting NS2B-NS3 interaction, Cell Res. 27 (2017) 1046–1064.

[34] P. Andrews, J. Thyssen, D. Lorke, The biology and toxicology of molluscicides, Bayluscide Pharmacol. Therapeut. 19 (1982) 245–295.

[35] W.H. Organization, World Health Organization Model List of Essential Medicines: 21st List 2019, World Health Organization, 2019.

[36] J. Xu, P.-Y. Shi, H. Li, J. Zhou, Broad spectrum antiviral agent niclosamide and its therapeutic potential, ACS Infect. Dis. 6 (2020) 909–915.

[37] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T.S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, SARS-CoV-2 Cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, Cell (2020).

[38] Y. Zhou, P. Vedantham, K. Lu, J. Agudelo, R. Carrion Jr, J.W. Nunneley, D. Barnard, S. Pöhlmann, J.H. McKerrow, A.R. Renslo, Protease inhibitors targeting coronavirus and filovirus entry, Antiviral Res. 116 (2015) 76–84.

[39] M. Khatri, P. Mago, Nitazoxanide/Camostat combination for COVID-19: An unexplored potential therapy, Chem. Biol. Lett. 7 (2020) 192–196.

[40] Q. Cai, M. Yang, D. Liu, J. Chen, D. Shu, J. Xia, X. Liao, Y. Gu, Q. Cai, Y. Yang, Experimental treatment with favipiravir for COVID-19: an open-label control study, Engineering (2020).

[41] D.L. McKee, A. Sternberg, U. Stange, S. Laufer, C. Naujokat, Candidate drugs against SARS-CoV-2 and COVID-19, Pharmacol. Res. (2020) 104859.

[42] X.-Y. Meng, H.-X. Zhang, M. Mezei, M. Cui, Molecular docking: a powerful approach for structure-based drug discovery, Curr. Comput.-Aided Drug Des. 7 (2011) 146–157.

[43] W. Li, M.J. Moore, N. Vasilieva, J. Sui, S.K. Wong, M.A. Berne, M. Somasundaran, J.L. Sullivan, K. Luzuriaga, T.C. Greenough, Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus, Nature 426 (2003) 450–454.

[44] M. Hoffmann, H. Kleine-Weber, N. Krüger, M.A. Mueller, C. Drosten, S. Pöhlmann, The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells, 2020, BioRxiv.

[45] V. Martinez, C. Navarro, C. Cano, W. Fajardo, A. Blanco, DrugNet: network-based drug–disease prioritization by integrating heterogeneous data, Artif. Intell. Med. 63 (2015) 41–49.

[46] A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, Mol. Syst. Biol. 7 (2011) 496.