Computers and Chemical Engineering 166 (2022) 107947

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

DRPADC: A novel drug repositioning algorithm predicting adaptive drugs for COVID-19

Guobo Xie[a], Haojie Xu[a], Jianming Li[a], Guosheng Gu[a,*], Yuping Sun[a], Zhiyi Lin[a], Yinting Zhu[a], Weiming Wang[a], Youfu Wang[b], Jiang Shao[c]

[a] *School of Computer Science, Guangdong University of Technology, Guangzhou 510006, China*
[b] *Huaneng Qinghai Power Generation Co., Ltd. New Energy Branch, Xining 810000, China*
[c] *School of Architecture & Design, China University of Mining and Technology, Xuzhou 221116, China*

ARTICLE INFO

Keywords:
Weight *K* nearest known neighbors
Central kernel alignment multiple kernel learning
Compressed sensing
Drug repositioning
COVID-19

ABSTRACT

Given that the usual process of developing a new vaccine or drug for COVID-19 demands significant time and funds, drug repositioning has emerged as a promising therapeutic strategy. We propose a method named DRPADC to predict novel drug-disease associations effectively from the original sparse drug-disease association adjacency matrix. Specifically, DRPADC processes the original association matrix with the WKNKN algorithm to reduce its sparsity. Furthermore, multiple types of similarity information are fused by a CKA-MKL algorithm. Finally, a compressed sensing algorithm is used to predict the potential drug-disease (virus) association scores. Experimental results show that DRPADC has superior performance than several competitive methods in terms of AUC values and case studies. DRPADC achieved the AUC value of 0.941, 0.955 and 0.876 in Fdataset, Cdataset and HDVD dataset, respectively. In addition, the conducted case studies of COVID-19 show that DRPADC can predict drug candidates accurately.

1. Introduction

COVID-19, a type of pneumonia discovered in 2019, is caused by a novel coronavirus. As shown in Fig. 1, the pathogenic virus causing COVID-19 was named "novel coronavirus SARS-CoV-2" because of its resemblance to previously identified coronaviruses (Abd El-Aziz Tarek and Stockand James, 2020; Catrin et al., 2020; Elisabetta et al., 2004). This is the seventh coronavirus to be discovered and studied in humans; the previous six were HCoV-OC43, 229E (HCoV-229E), HCoV-HKU1, HCoVNL63, Middle East respiratory syndrome-associated coronavirus (MERS-Cov), and severe acute respiratory syndrome-associated coronavirus (SARS-Cov) (Cha et al., 2018). There is a dearth of commercially available drugs for the treatment of COVID-19 (Wei-jie et al., 2020), and more therapeutic drugs are still in clinical trials stage. More seriously, the genetic material of the SARS-CoV-2 is single-stranded RNA (Imran and Gwanggil, 2021), which is susceptible to mutation during the transcriptional replication phase. Such mutation may alter the *in vitro* characteristics of SARS-CoV-2, potentially threatening the effectiveness of existing COVID-19 vaccines. Researchers have now identified various mutated strains of SARS-CoV-2 (including the Delta strain), so developing effective drugs to treat COVID-19 is an urgent task.

Drug discovery and development present several challenges, including high attrition rates, long development times, and substantial costs (Yosef et al., 2020; Han et al., 2019; Sudeep et al., 2019; Huimin et al., 2020; Muhammad et al., 2014). Now there is an increasing interest in drug-repositioning techniques using computer algorithmic models (Hurle et al., 2013). Compared with traditional drug clinical trials, drug-repositioning techniques can effectively reduce the financial and time costs of the drug development process. In recent years, various drug-repositioning methods using computer algorithm models have been proposed as auxiliary tools (Qi et al., 2019) and validated by biological data. Because most diseases are caused by corresponding pathogens, drugs that inhibit or kill that pathogen in the organism are generally the effective ones for that disease. Therefore, it is feasible to use drug repositioning techniques to screen for therapeutic agents adapted to SARS-CoV-2.

Current drug-repositioning methods can be divided into three categories: machine learning methods, network propagation-based methods and matrix completion or factorization-based methods (Xing et al., 2016; Jawad et al., 2021). These methods are based on assumption that

[*] Corresponding author.
E-mail address: gsgu@gdut.edu.cn (G. Gu).

https://doi.org/10.1016/j.compchemeng.2022.107947
Received 27 September 2021; Received in revised form 13 April 2022; Accepted 27 July 2022
Available online 4 August 2022

similar drugs are associated with similar diseases and vice versa. Machine learning methods include a method proposed by Yongcui et al. (2013) to determine the similarity between drug-disease pairs using a kernel function and train the classification kernel of a support vector machine to discover new drug-disease interactions. Machine learning methods rely heavily on known sample labels in the dataset. However, stable negative sample data are difficult to obtain in practice, which limits the generalizability of these prediction methods. Xing et al. (2016) developed a network-based plastic regularized least squares collaborative drug combination prediction method, which can overcome the problem of lack of stable negative samples data. However, their performance is not stable because of the lack of negative samples in training, which limits their applications.

Many network-based drug-repositioning methods use an association matrix of drugs and indications to build a network, and use wandering or diffusion to propagate the association resources from the drug side to the indication side to infer the missing edges within the network. Wenhui et al. (2013) proposed a heterogeneous graph-based inference (HGBI) technique for drug-target association prediction by combining known drug-target interactions and drug-target similarity information; this model was successfully applied to predict drug-disease associations (Wenhui et al., 2014; Mengyun et al., 2020). Victor et al. (2015) established a three-layer heterogeneous network containing different types of elements and interactions, called DrugNet. Their study showed that DrugNet was effective for discovering new drug candidates. Huimin et al. (2016) developed a comprehensive similarity measures and Bi-Random walk (MBIRW) algorithm, which used available drug-disease associations information to enhance drug similarity and disease similarity. However, in the process of network propagation, information resources are biased to select the edges with larger weights, meaning that nodes lacking associated information are not allocated resources for propagation for a long time (i.e., the cold start problem); this can affect the accuracy of the prediction results.

When a matrix contains partially known information, matrix completion is required. A number of methodological models for matrix factorization and complementation have been proposed in recent years. For instance, Huimin et al. (2018) constructed a drug-disease heterogeneous network based on the similarity information of drugs and diseases and designed a Drug Repositioning Recommendation System (DRRS), which achieved the prediction of unknown drug-disease effects. Mengyun et al. (2019) proposed a scheme for preprocessing the drug-disease association matrix using a bounded kernel norm regularization (BNNR) model. Under the assumption of low rank, the model eliminated the effect of noisy data on the accuracy of model prediction by introducing a regularization term. Mengyun et al. (2020) proposed a method based on matrix-completion heterogeneous graphical inference (HGIMC) that could predict the potential adaptation of a drug to a disease. Yajie et al. (2021) proposed a similarity-constrained probability matrix factorization (SCPMF) model based on the similarity information of drugs and diseases to identify unknown association targets of drugs. Poleksic (2021) augmented the set of relationships between different biomedical entities based on the concept of meta-paths to enrich the associated information of paths, and predicted the association between different biomedical entities using a compressed sensing matrix completion algorithm.

Compressed sensing, a class of matrix-completion methods, has an inherent advantage over supervised methods of machine learning for drug repositioning because it uses a "submatrix simulation" technique to predict potential drug-disease interactions without treating all missing data as negative data. However, networks or graphs based methods rely on association data as a resource propagation path for bipartite graphs and tend to suffer from cold-start problems that affect the prediction performance because of sparse association information. By contrast, compressed sensing starts with the existing matrix information, uses submatrices to capture the original matrix information, and generates a low-rank simulation matrix to complete the missing parts of the original association matrix. This reduces the redundancy of the model to a certain extent compared with that of network- or graph-based methods.

Inspired by the compressed-sensing-based matrix completion model proposed by Poleksic (2021), we use a compressed-sensing-based method in this work to find potential adaptive drugs for COVID-19. We introduce the Weight K nearest known neighbors (WKNKN) algorithm (Ali et al., 2016) to preprocess the original matrix and increase its rank, which can complete more information. In addition, compressed sensing allows us to affiliate drug and disease similarity features to aid in the correction of the decomposed submatrices. To better integrate different types of similarity information, we fuse multiple drug and disease similarity matrices using a central kernel alignment multiple kernel learning model (CKA-MKL) based on the study of Yijie et al. (2019). After the processing of CKA-MKL, we obtain the weights of each similarity matrix (the kernel) and choose the best matrix combination, thereby improving the complementarity of various types of similarity information. In summary, we propose a compressed-sensing-based prediction method combining a multiple kernel learning model with central kernel alignment to predict the probability of potential drug-disease associations, and we validate the feasibility of this method as a drug-discovery aid for finding drugs that can interact with SARS-CoV-2.

2. Materials and methods

2.1. Materials

2.1.1. Drug-disease datasets

In this work, we used Fdataset and Cdataset as gold-standard datasets to test the performance of our proposed model. Fdataset is a collection of 593 drugs, 313 diseases, and 1933 validated drug-disease associations from different data sources compiled by Assaf et al. (2011). Cdataset was compiled by Huimin et al. (2016) and includes 409 diseases, 663 drugs and 2532 disease and drug associations. Each dataset contains three information matrices, as follows:

Drug-disease association adjacency matrix $Y \in R^{n_D \times n_d}$ (n_D denotes the number of drugs and n_d denotes the number of diseases). The value of element $Y(D_i, d_j)$ is 1 if the drug D_i is associated with the disease d_j; otherwise, it is 0;

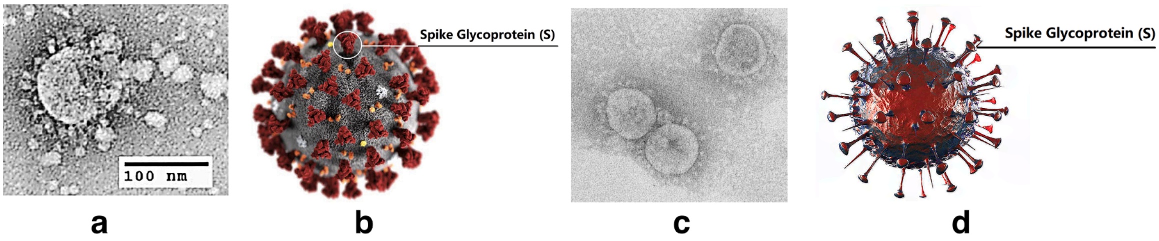


Fig. 1. Viral structure of SARS-CoV and SARS-CoV-2. (a) Electron microscopy image of a thin section of SARS-CoV-2 found in the cytoplasm of an infected cell by Catrin et al. (2020), showing the spherical particles and cross-sections of the nucleocapsid of the virus. (b) Schematic diagram of SARS-CoV-2 virus particles. (c) Electron microscopy image of SARS-CoV virus cultured in Vero cells by Elisabetta et al. (2004). (d) Model diagram of SARS-CoV virus particles.

Drug chemical structure similarity matrix $DS_{chem} \in R^{n_D \times n_D}$. The matrix consists of the chemical structure information of the corresponding drug derived from a chemical development kit (Christoph et al., 2003), where the drug-drug pair $DS_{chem}(i,j)$ is represented by the two-dimensional chemical fingerprint scores of drugs D_i and D_j;

Diseases semantic similarity matrix $dS_{sem} \in R^{n_d \times n_d}$. We obtain the similarity information for corresponding disease pairs using MimMiner (Van Driel et al., 2006). The semantic similarity matrix dS_{sem} of diseases is obtained by computing the semantic similarity between diseases through text mining (Fleuren Wilco and Wynand, 2015).

2.1.2. Drug-virus dataset

Yajie et al. (2021) used text mining techniques to collect a large number of experimentally validated drug-virus interaction terms to construct an experimentally supported human drug-virus association database, HDVD. We used the HDVD as a dataset for finding therapeutic drugs for COVID-19. The HDVD contains 219 drugs, 34 viruses, and 455 confirmed human drug-virus interactions. Analogous to the gold-standard datasets described above, HDVD contains the following three information matrices:

Drug-virus association adjacency matrix $A \in R^{m_D \times m_v}$ (m_D denotes the number of drugs and m_v denotes the number of diseases). The value of element $A(D_i, V_j)$ is 1 if drug D_i is associated with the virus V_j; otherwise, it is 0;

Drug chemical structure similarity matrix $DS_{chem} \in R^{m_D \times m_D}$. SMILES (simplified molecular input line entry system) is an information format that describes molecular structures as a one-dimensional representation (Hakime et al., 2016). We used the SMILES format to download the chemical structure information of the corresponding drugs from the DrugBank database, calculating the Molecular Access System (MACCS) fingerprint of each drug with Open Babel v2.3.1 (O'Boyle et al., 2011), and use the Tanimoto coefficient to measure the absolute similarity (Bajusz et al., 2015) to construct a drug chemical structure similarity matrix DS_{chem};

Viral gene sequence similarity matrix $VS_{gen} \in R^{m_v \times m_v}$. MAFFT is a similarity-based multiple sequence comparison method (Katoh and Standley, 2013). We downloaded the genomic nucleotide sequences of viruses in Homo sapiens from the National Center for Biotechnology Information, and use MAFFT version 7 to calculate sequence similarity between viruses to construct a viral gene sequence similarity matrix VS_{gen}.

2.2. Method

Suppose the number of drugs and diseases are n_D and n_d, respectively, and $Y_{n_D \times n_d}$ denotes the association matrix. $Y_{ij} = 1$ if the association between drug i and disease j is known, otherwise $Y_{ij} = 0$. An algorithm predicting Drug-disease associations requires Y and corresponding feature matrix X as input, then outputs a score for each pair of drug and disease. p^{n+1} denotes the score matrix, $p_{ij}^{n+1} \in [0, 1]$, i.e the prediction result.

2.2.1. Calculating Jaccard similarity of drugs and diseases

Jaccard similarity is a normalized form of the common neighborhood measure between vectors (Pang-Ning et al., 2016). For drugs D_x and D_y, Jaccard similarity gives the probability that a random node pair from the set of disease associations between the D_x and D_y is one of the co-acting diseases of D_x and D_y. If the number of mutual diseases of D_x and D_y nodes is higher, this metric is larger, which means that D_x and D_y are more similar. The formula is as follows:

$$DS_{jac}(D_x, D_y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{1}$$

where, $\Gamma(x)$ is the set of diseases associated with the drug D_x.

We can calculate the Jaccard similarity matrix dS_{jac} of diseases in a similar manner.

2.2.2. Calculating Gaussian kernel similarity matrix of drugs and diseases

Gaussian kernel similarity is a method commonly used to calculate the similarity between different types of nodes. By projecting the data in a high dimension through a radial basis, the distance between different node vectors can be calculated to obtain the similarity weights between nodes; therefore, Gaussian kernel similarity is also called radial basis function kernel similarity. In the adjacency matrix Y_F, row i indicates whether drug D_i is associated with each disease, and column j indicates whether disease d_j is associated with each drug. The vectors $Q(D_i)$ and $Q(d_j)$ represent the ith row vector and the jth column vector as the eigenvectors of the Gaussian kernel, respectively. Therefore, we denote the Gaussian kernel similarity between diseases d_i and d_j as dS_{Gaus}, and the Gaussian kernel similarity between drugs D_i and D_j as DS_{Gaus}, calculated as follows:

$$dS_{Gaus}(d_i, d_j) = exp\left(- \alpha_d \| Q(d_i) - Q(d_j) \|^2 \right) \tag{2}$$

$$DS_{Gaus}(D_i, D_j) = exp\left(- \alpha_D \| Q(D_i) - Q(D_j) \|^2 \right) \tag{3}$$

where, the kernel bandwidths α_d and α_D are defined as:

$$\alpha_d = \alpha_d^{'}(\frac{1}{n_d}\sum_{i=1}^{n_d} \| Q(d_i) \|^2) \tag{4}$$

$$\alpha_D = \alpha_D^{'}(\frac{1}{n_D}\sum_{i=1}^{n_D} \| Q(D_i) \|^2) \tag{5}$$

where, the initial kernel bandwidth coefficients $\alpha_d^{'}$ and $\alpha_D^{'}$ are both set to 1 (Xiujuan and Cheng, 2020).

2.2.3. DRPADC

In this study, we propose a compressed sensing prediction method incorporating a multicore learning model with central kernel symmetry. First, the WKNKN algorithm (Ali et al., 2016) is introduced to reduce the sparsity of the drug-disease association adjacency matrix. According to the work of Yijie et al. (2019), processing multiple similarity matrices using CKA-MKL allows us to obtain the weights of each similarity matrix (kernel) and thus select the best matrix combination. We use the CKA-MKL to fuse the different similarity information matrices to obtain the integrated similarity matrix. Finally, the drug-disease association prediction scores are calculated using the compressed sensing algorithm. A general flow chart for the algorithm is shown in Fig. 2. The details of the principle and flow of each module in the algorithm are described below.

2.2.3.1. Reducing sparseness by WKNKN. The compressed sensing algorithm uses the decomposed submatrix to sample the information of the original association matrix and generate a simulated approximation matrix. The sparsity of the target matrix affects the descriptiveness of the elements in the submatrices, which in turn affects the recovery accuracy of the approximation matrix. Previous studies (Huimin et al., 2018) have shown that the drug-disease association matrix Y in both Cdataset and Fdataset is sparse (the same is true for HDVD); this is because there are still many drug-disease associations that have not been confirmed. Therefore, we reduce the sparsity of Y by using WKNKN (Ali et al., 2016) to mine the likelihood scores of potential interactions. That is, the WKNKN algorithm is used to estimate the likelihood value of the interactions between drug-disease pairs in the adjacency matrix Y. This is performed using the following three steps:

Step1. The interaction likelihood score for each drug D_i is estimated using the chemical similarity matrix DS_{chem} of the K known drugs closest to it and their corresponding association effects. The derived equation is

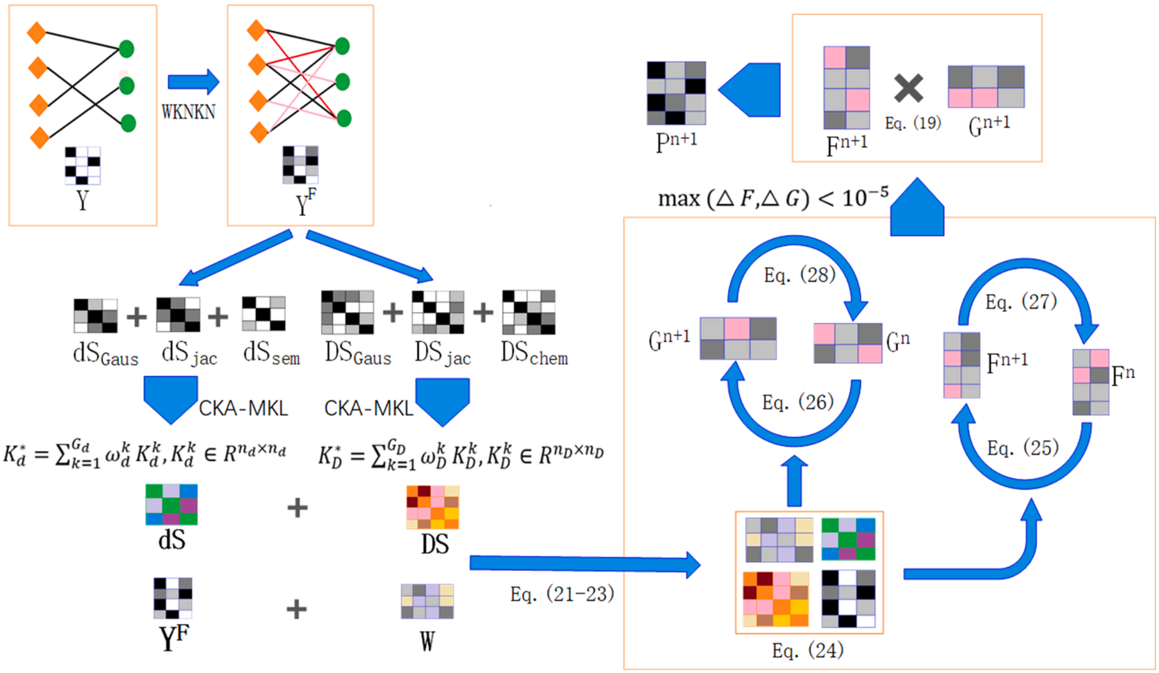

Y
WKNKN
Y^F
dS_Gaus + dS_jac + dS_sem
DS_Gaus + DS_jac + DS_chem
CKA-MKL
CKA-MKL
$K_d^* = \sum_{k=1}^{G_d} \omega_d^k K_d^k, K_d^k \in R^{n_d \times n_d}$
$K_D^* = \sum_{k=1}^{G_D} \omega_D^k K_D^k, K_D^k \in R^{n_D \times n_D}$
dS + DS
Y^F + W
Eq. (21-23)
Eq. (24)
Eq. (25)
Eq. (26)
Eq. (27)
Eq. (28)
G^n+1
G^n
F^n+1
F^n
max (△F,△G) < 10^−5
P^n+1
F^n+1
Eq. (19)
G^n+1

Fig. 2. Flowchart of DRPADC. The red box shows the iterative process of the compressed sensing model (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

as follows:

$$Y_D(D_i, :) = \frac{1}{Q_D}\sum_{k_D=1}^{K}\omega_{k_D} Y_D(D_{k_D}, :) \tag{6}$$

where, D_1 to D_{k_D} denote the K-nearest drugs of D_i in descending order; $Q_D = \sum_{k_D=1}^{K} DS_{chem}(D_i, D_{k_D})$ is the regularization term; $\omega_{k_D} = T^{k_D - 1} DS_{chem}(D_i, D_{k_D})$ is the weighting factor, where T is the decay term and satisfies $T \leq 1$.

Step2. Similarly, for each disease d_j, the semantic similarity of its closest K known diseases and their corresponding interactions are used to estimate the interaction likelihood score of d_j. The derived formula is as follows:

$$Y_d(:, d_j) = \frac{1}{Q_d}\sum_{k_d=1}^{K}\omega_{k_d} Y_d(:, d_{k_d}) \tag{7}$$

where, d_1 to d_{k_d} denote the K-nearest neighbor diseases of disease d_i in descending order; $Q_d = \sum_{k_d=1}^{K} dS_{sem}(d_{k_d}, d_j)$ is the regularization term; $\omega_{k_d} = T^{k_d - 1} dS_{sem}(d_{k_d}, d_j)$is the weighting factor, where T is the decay term and satisfies $T \leq 1$.

Step3. Finally, if $Y_{i,j} = 0$, we replace it by taking the average of Y_D and Y_d, as follows:

$$Y_F(i,j) = \max\left(\frac{Y_D + Y_d}{2}, Y(i,j)\right). \tag{8}$$

According to the introduction of the above model, the procedure of WKNKN is summarized in Algorithm 1. There are 99%, 99.1% and 93.9% of 0 elements in the association matrix of Fdataset, Cdataset and HDVD respectively, which indicates that the elements in the original matrix are too sparse. Therefore, we use WKNKN algorithm to preprocess the original matrix, so that the 0 elements in the matrix can be replaced by likelihood scores between 0 and 1 to mine more potential interactions between drugs and diseases or viruses. The proportion of element 0 in the association matrix of Fdataset, Cdataset and HDVD before and after performing WKNKN algorithm are shown in Table 1.

Table 1
The proportion of element 0 in the association matrix.

Datasets	Before WKNKN	After WKNKN
Fdataset	99.0%	89.2%
Cdataset	99.1%	91.2%
HDVD	93.9%	64.1%

Algorithm 1. WKNKN

```
Inputs: Adjacent matrix $Y \in R^{n_D \times n_d}$, chemical similarity matrix $DS_{chem} \in R^{n_D \times n_D}$, and
  disease semantic similarity matrix $dS_{sem} \in R^{n_d \times n_d}$,neighborhood sizes $K$, decay term
  $T$.
  Output: Association probability matrix $Y_F$.
  Algorithm
  Compute association probability matrix $Y_F$ by WKNKN:
  for i = 1 →$n_D$ do
  calculate $Y_D$ by Eq. (6);
  end for
  for j = 1 →$n_d$ do
  calculate $Y_d$ by Eq. (7);
  end for
  $Y_F = \max(\frac{Y_D + Y_d}{2}, Y)$.
  Output: $Y_F$.
```

2.2.3.2. Central kernel alignment algorithm multiple kernel learning alignment. A number of methods have been proposed to mine the similarity of drugs and diseases (Esra and Buket, 2017) and excavate their similar characteristics from different perspectives. In drug-repositioning methods, the use of different similarity information has different effects on model prediction performance; therefore, most current methods use a combination of multiple similarity information. However, they generally lack a systematic method of combining this information and generally only consider the mean or complementarity, etc. In order to integrate multiple types of similarity information, based on previous work by Yijie et al. (2019), we process multiple similarity matrices and obtain the weights of each similarity matrix (kernel) by CKA-MKL, so as to select the best combination of similarity matrices and improve the complementarity of similarity information. Specifically, we obtain three similar kernel matrices for each of the similar kernel sets of drugs and diseases ($DS_{chem}, DS_{jac}, DS_{Gaus} \in R^{n_D \times n_D}$; $dS_{sem}, dS_{jac}, dS_{Gaus} \in R^{n_d \times n_d}$). Then, we combine the three kernel matrices in two spaces separately using CKA-MKL. The optimal kernel is calculated as follows:

$$K^* = \sum_{i=1}^{k} \omega_i K_i, K_i \in R^{N \times N}, \sum_{i=1}^{k} \omega_i = 1 \tag{9}$$

where, k is the number of kernels, $K_i$is the value in the drug similar kernel set ($DS_{chem}, DS_{jac}, DS_{Gaus}$) and disease similar kernel set ($dS_{sem}, dS_{jac}, dS_{Gaus}$), and ω_i is the weight of kernel K_i. N is the number of nodes.

According to Yijie et al. (2019), the kernel alignment score can be described by calculating the cosine correlation between two kernels. The larger the correlation between the kernels, the higher the alignment between the kernels. Therefore the value of kernel alignment is defined as follows:

$$A(P, Q) = \frac{\langle P, Q \rangle_F}{\| P \|_F \| Q \|_F} \tag{10}$$

where, $P, Q \in R^{N \times N}$, $N \in \{n_D, n_d\}$, $\langle P, Q \rangle_F = Trace(P^T Q)$ is the Frobenius inner product, and $\| P \|_F = \sqrt{\langle P, P \rangle_F}$ is the Frobenius parametrization. In fact, the kernel alignment score can be viewed as a correlation between two kernels (the characteristic kernel K^* and the ideal kernel matrix $Y_F Y_F^T$). To obtain the optimal weights of the kernel, we should maximize the alignment score between K^* and $Y_F Y_F^T$. Therefore, the objective function after central kernel alignment is as follow:

$$\begin{gathered} \underset{\omega \geq 0}{max} CA\left(K^*, y_F y_F^T\right) = \underset{\omega \geq 0}{max} \frac{\left\langle U_N K^* U_N, y_F y_F^T \right\rangle_F}{\| U_N K^* U_N \|_F \| y_F y_F^T \|_F} \\ subjectto K^* = \sum_{i=1}^{k} \omega_i K_i, \omega_i \geq 0, i = 1, 2, ..., k, \sum_{i=1}^{k} \omega_i = 1 \end{gathered} \tag{11}$$

where, $U_N = I_N - (1/N) l_N l_N^T (U_N \in R^{N \times N})$ denotes a central kernel matrix. $I_N \in R^{N \times N}$ is a unitary matrix. I_N is a unit vector. Eq. (11) can also be transformed into:

$$\begin{gathered} \underset{\omega \geq 0}{max} \frac{\omega^T a}{\sqrt{\omega^T M \omega}} \\ subjectto K^* = \sum_{i=1}^{k} \omega_i K_i, \omega_i \geq 0, i = 1, 2, ..., k, \sum_{i=1}^{k} \omega_i = 1 \end{gathered} \tag{12}$$

where, $a \in R^{k \times 1}$ and $M \in R^{k \times k}$ are calculated from Eqs. (13) and (14), respectively:

$$a = \left(\left\langle U_N K_1 U_N, y_F y_F^T \right\rangle_F, ..., \left\langle U_N K_k U_N, y_F y_F^T \right\rangle_F\right)^T \in R^{k \times 1} \tag{13}$$

$$\begin{gathered} M = \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,k} \\ M_{2,1} & M_{2,2} & \cdots & M_{2,k} \\ \vdots & \vdots & M_{e,f} & \vdots \\ M_{k,1} & M_{k,2} & \cdots & M_{k,k} \end{bmatrix}_{k \times k}. \\ M_{e,f} = \left\langle U_N K_e U_N, U_N K_f U_N \right\rangle_F, e, f = 1, 2, ..., k \end{gathered} \tag{14}$$

The final objective function can also be expressed as:

$$\begin{gathered} \underset{\omega \geq 0}{max} \omega^T M \omega - 2\omega^T a \\ subjectto K^* = \sum_{i=1}^{k} \omega_i K_i, \omega_i \geq 0, i = 1, 2, ..., k, \sum_{i=1}^{k} \omega_i = 1 \end{gathered}. \tag{15}$$

To obtain the reconfiguration weights for each similar kernel, we use standard quadratic programming to solve Eq. (15). Therefore, we obtain the weights of drugs and diseases ($\omega_D, \omega_d \in R^{3 \times 1}$) and combine them with the drug ($DS_{chem}, DS_{jac}, DS_{Gaus} \in R^{n_D \times n_D}$) and disease ($dS_{sem}, dS_{jac}, dS_{Gaus} \in R^{n_d \times n_d}$) similarity kernels, respectively, according to the following equations:

$$K_d^* = \sum_{i=1}^{k} \omega_d^i K_d^i, K_d^i \in R^{n_d \times n_d} \tag{16}$$

$$K_D^* = \sum_{j=1}^{k} \omega_D^j K_D^j, K_D^j \in R^{n_D \times n_D}. \tag{17}$$

where, $\omega_d^i = \{\omega_d^1, \omega_d^2, ..., \omega_d^k\}$ and $\omega_D^i = \{\omega_D^1, \omega_D^2, ...\omega_D^k\}$ are the optimal weights of the disease kernel and the drug kernel, respectively.

According to the above alignment model, the procedure of CKA-MKL is summarized in Algorithm 2. CKA-MKL combines three similar kernel matrices of drugs and diseases in two spaces respectively, and obtains optimal combination similarity matrices of drugs and diseases, thereby improving the complementarity of various types of similarity information.

```
Algorithm 2. CKA-MKL
Inputs: Association probability matrix $Y_F$, chemical similarity matrix $DS_{chem} \in R^{n_D \times n_D}$,
  disease semantic similarity matrix $dS_{sem} \in R^{n_d \times n_d}$, Jaccard similarity matrices $DS_{jac}$
  and $dS_{jac}$of drugs and diseases, and Gaussian-based similarity matrix $DS_{Gaus}$ and
  $dS_{Gaus}$ of drugs and diseases.
  Output: Optimal combination similarity matrices $K_D^*$ and $K_d^*$ of drugs and diseases.
  Algorithm
  Define the objective function Eqs. (11) and (12) of CKD-MKL according to Eqs. (9)
  and (10);
  Change the objective function Eq. (12) to Eq. (15) according to Eqs. (13) and (14);
  Calculate the optimal combination similarity matrix weights $\omega_D$ and $\omega_d$ for drugs
  and diseases by minimizing Eq. (15);
  Calculate the optimal combination similarity matrices of drugs and diseases
  according to Eqs. (16) and (17), respectively.
  Output: $K_D^*$, $K_d^*$.
```

2.2.3.3. Compressed sensing model. Compressed sensing is a matrix completion class method based on the principle that the simulation matrix obtained by the inner product of sub-matrices can approximate the target matrix. The elements in the simulation matrix are used as estimates of the elements of the unobservable part of the target matrix. In other words, given a matrix $Y \in R^{n_D \times n_d}$ with missing interaction information, where n_D represents the number of drugs and n_d represents the number of diseases. Y_F is obtained based on the WKNKN algorithm by simulating the inner product of two low-dimensional sub-matrices $F \in R^{n_D \times r}$ and $G \in R^{n_d \times r}$ $\left(F, G \sim N\left(0, \frac{1}{\sqrt{r}}\right)\right)$, which correspond to the potential features of drugs and diseases, respectively. The probabilities of drug-disease interactions are approximated by mapping the association information of drugs and diseases to a low-dimensional common potential space. Optimization of the sub-matrices is achieved by reducing the gap between the simulation matrix and the target matrix during the complementation process. To avoid over-optimization of the model, the complexity of the sub-matrix is calculated as a penalty term for sub-matrix optimization when calculating the difference between the simulation matrix and the target matrix. In addition, during optimization, the compressed sensing model allows the introduction of drug and disease similarity information in order to continuously adjust the sub-matrix model.

Therefore, given a drug D_i and disease d_j, the probability of their interaction events can be calculated by the following equation:

$$p_{i,j} = p\left(y_{i,j}^F = 1 \middle| f_i, g_j\right) = 1 \bigg/ \left(1 + exp\left(f_i g_j^T\right)^{-1}\right) \tag{18}$$

or in matrix form:

$$p(Y_F | F, G) = \prod_{i,j} \left(p_{i,j}^{y_{i,j}^F} (1 - p_{i,j})^{1 - y_{i,j}^F}\right)^{w_{i,j}} \tag{19}$$

where, $p_{i,j}$ is the interaction probability of drug D_i and disease d_j, f_i is the ith row of the drug sub-matrix F, g_j is the jth row of the disease sub-matrix G, and $w_{i,j}$ is the initial weight of drug D_i and disease d_j

(Hansaim et al., 2016). By Bayesian inference, we can derive the probability of $p(Y_F|F,G)$ as follow:

$$p(F,G|Y_F)\propto p(Y_F|F,G)p(F)p(G). \tag{20}$$

As described by Steck (2010), we obtain the loss function of the model:

$$LOSS=\sum_{i,j}w_{i,j}\left\{\ln\left(1+exp\left(f_ig_j^T\right)\right)-y_{i,j}^Ff_ig_j^T\right\}+\lambda_F\|\ F\ \|_2^2+\lambda_G\|\ G\ \|_2^2. \tag{21}$$

To improve the accuracy of the method's predictions, we further extend the loss function based on the assumption that similar drugs may interact with similar diseases. Specifically, let $DS\in R^{n_D\times n_D}$ be the drug similarity matrix, where each entry $DS_{i,j}$ denotes the similarity between drugs D_i and D_j, and let $dS\in R^{n_d\times n_d}$ be the disease similarity matrix. The combination of similar diseases with similar drugs is explained by minimizing the distance between the properties of the drugs:

$$tr\left(F^T(D_{DS}-DS)F\right)=\frac{1}{2}\sum_{i=1}^{n_D}\sum_{j=1}^{n_D}DS(i,j)\|\ F(i,:)-F(j,:)\ \|_2^2. \tag{22}$$

Similarly, similarities between diseases are minimized:

$$tr\left(G^T(d_{dS}-dS)G\right)=\frac{1}{2}\sum_{i=1}^{n_d}\sum_{j=1}^{n_d}dS(i,j)\|\ G(i,:)-G(j,:)\ \|_2^2 \tag{23}$$

Combining the regularization terms (22) and (23) into (21) and introducing two additional adjustable parameters λ_M and λ_N, our loss function can be transformed into:

$$\begin{aligned}LOSS=&\sum_{i,j}w_{i,j}\left\{\ln\left(1+exp\left(f_ig_j^T\right)\right)-y_{i,j}^Ff_ig_j^T\right\}+\lambda_F\|\ F\ \|_2^2+\lambda_G\|\ G\ \|_2^2\\&+\lambda_Mtr\left(F^T(D_{DS}-DS)F\right)+\lambda_Ntr\left(G^T(d_{dS}-dS)G\right)\end{aligned} \tag{24}$$

According to Liu et al. (2016), compressed sensing can be used to optimize the model using an iterative gradient descent method (AdaGrad) (John et al., 2011). During the iterative process, the partial derivative of the loss function can be written as:

$$BF=\frac{\partial LOSS}{\partial F}=\{W\circ[P-Y_F]\}G+2\lambda_rF+2\lambda_M(D_{DS}-DS)F \tag{25}$$

$$BG=\frac{\partial LOSS}{\partial G}=\left\{W^T\circ\left[P-Y_F^T\right]\right\}F+2\lambda_rG+2\lambda_N(d_{dS}-dS)G \tag{26}$$

where, $\circ$ represents the Hadamard product. The submatrices F and G are updated according to the following equations:

$$F^{n+1}=F^n-k\left(\frac{BF^n}{\|\ BF^n\ \|_F}\right) \tag{27}$$

$$G^{n+1}=G^n-k\left(\frac{BG^n}{\|\ BG^n\ \|_F}\right) \tag{28}$$

$$\Delta F=\|\ F^{n+1}\ \|_F-\|\ F^n\ \|_F \tag{29}$$

$$\Delta G=\|\ G^{n+1}\ \|_F-\|\ G^n\ \|_F \tag{30}$$

where, k is the learning rate of the iterative process; the superscript n is the current number of iterations; the end condition for the update of submatrices F and G is $\max(\Delta F,\Delta G)<10^{-5}$; $\|\ BF\ \|_F$ is the Frobenius paradigm of BF, and $\|\ BG\ \|_F$ is defined similarly as:

$$\|\ BF\ \|_F=\left(\sum_{i=1}^{n_D}\sum_{j=1}^{r}\left|BF_{i,j}\right|^2\right)^{\frac{1}{2}} \tag{31}$$

$$\|\ BG\ \|_F=\left(\sum_{i=1}^{n_d}\sum_{j=1}^{r}\left|BG_{i,j}\right|^2\right)^{\frac{1}{2}} \tag{32}$$

According to the above description, the procedure of compressed sensing model is summarized in Algorithm 3. By combining two optimal combination similarity matrices derived from CKA-MKL algorithm, we use compressed sensing model to predict the potential drug-disease (virus) association scores.

Algorithm 3. Compressed sensing model

```
Inputs: Association probability matrix $Y_F$, optimal combination similarity matrices $K_D^*$
  and $K_d^*$ of drugs and diseases, learning rate $k$, Frobenius parametric term coefficients
  $\lambda_r$ of the submatrices $F$ and $G$, and the regularized term coefficients $\lambda_M$ and $\lambda_N$ of the
  submatrices $F$ and $G$.
  Output: Association prediction matrix $P^{n+1}$.
  Algorithm
Construct ~~the~~ initial component matrices $F^0\in R^{n_D\times r}$ and $G^0\in R^{n_d\times r}$;
  Calculate the initial probability matrix $P^0$ according to Eq. (19);
  Construct loss function Eq. (21);
  Rewrite Eq. (21) as Eq. (24) according to Eqs. (22) and (23);
  Do
  Calculate the partial derivatives of $F$ and $G$ from Eqs. (25) and (26), respectively by
  solving Eq. (24);
  Update the matrices $F$ and $G$ according to Eqs. (27) and (28), respectively;
  Until $\max(\Delta F,\Delta G)<10^{-5}$.
  Calculate the probability matrix $P^{n+1}$ with matrices $F^{n+1}$ and $G^{n+1}$ according to Eq.
  (19).
  Output: $P^{n+1}$.
```

3. Result

Under the running environment of win10 Professional Edition and i5-9300H CPU, we used MATLAB 2018b software to run the code which can be downloaded from https://github.com/kk-2010000/drpadc1. The purpose of this study was to construct a drug-repositioning computational method with excellent performance and reliable results for drug development against COVID-19. Therefore, we used Fdataset and Cdataset, which are used in drug-repositioning research and available from DrugBank, OMIM and Mesh public database, as gold-standard datasets and compared DRPADC with those proposed by other researchers in recent years to validate the performance of DRPADC. Then, we compared DRPADC with methods of the same type on the HDVD dataset to further validate its performance.

3.1. Performance comparison of methods on the gold standard datasets

To test the performance of DRPADC, we perform 10 times 10-fold cross-validation (CV) on two gold-standard datasets, Fdataset and Cdataset, for comparison with other classical drug-repositioning methods. Here, we randomly divide all the known drug-disease associations into 10 equal parts. We take turns to select one part to form the test sample set and the rest the training sample set. The 10-fold cross-validation is repeated 10 times for each method, and the average of all the 10 results is regarded as the final score. After testing all interactions, we calculate the true positive rate (*TPR,* the proportion of positive samples that are correctly identified among positive samples) and the false positive rate (*FPR,* the proportion of incorrectly identified negative samples to all negative samples) using the following equations:

$$TPR(or\mathrm{Recall})=\frac{TP}{TP+FN} \tag{33}$$

where, *TP* indicates the number of correctly identified positive samples and *FN* indicates the number of incorrectly identified negative samples; and

$$FPR=\frac{FP}{TN+FP} \tag{34}$$

where, *FP* is the number of correctly identified positive samples and *TN* is the number of correctly identified negative samples.

Since the precision measure is the percentage of correctly identified positive samples out of all retrieved samples, the higher the precision value, the better the prediction of the model. Therefore, we calculate the precision of each method on the various datasets. Precision is calculated as follows:

$$Precision = \frac{TP}{TP + FN} \tag{35}$$

We constructed receiver operating characteristic (ROC) curves and calculated the area under the curve (AUC), which is widely used to describe overall predictive performance (Zou et al., 2015). An AUC value closer to 1 indicates better performance, while an AUC value closer to 0.5 indicates performance closer to random (Peng et al., 2020). Since the area under the Precision-recall curve (AUPR) is more suitable than AUC for evaluating highly unbalanced or skewed datasets, we also added AUPR to assess the performance of various methods. We calculated the *TPR, FPR* and *Precision* by changing the preset ranking thresholds for plotting the respective ROC and PR (Precision-recall) curves. As shown in Fig. 3 and Table 2, we obtained better results with DRPADC compared with the other methods. DRPADC achieved an AUC value of 0.955 in Cdataset, which is 0.5% higher than that of the second highest-achieving method, BNNR (AUC = 0.950). DRPADC had the highest AUC value of 0.941 in Fdataset, 0.8% higher than that obtained with BNNR (AUC = 0.933). DrugNet achieved the lowest AUC values in both datasets, 16.3% (Fdataset) and 15.1% (Cdataset) lower than those of DRPADC, respectively. In addition, as the AUPR row shown in Table 2, DRPADC achieves the highest AUPR value in different datasets. In summary, DRPADC outperformed the other methods shown in Fig. 3 on the gold-standard datasets.

3.2. Performance comparison of methods in HDVD dataset

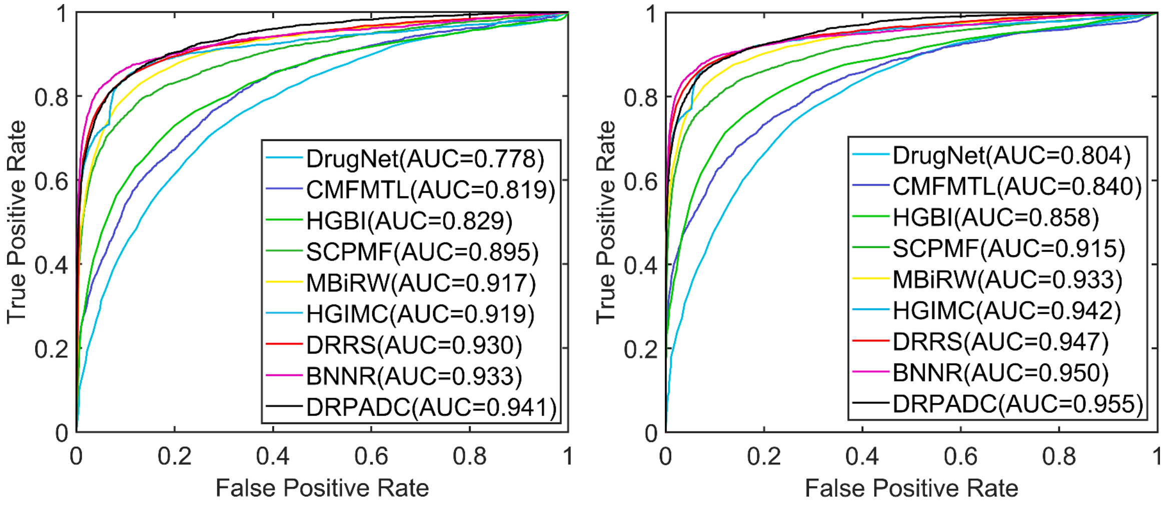
To further verify the performance of DRPADC, we validated the proposed method and other recent prediction methods of the same type (Mengyun et al., 2020, 2019; Yajie et al., 2021; Feng et al., 2020; Chengqian et al., 2019, 2018) on HDVD with 10 times 10-fold-CV, constructed ROC plots for each comparison method, and calculated the corresponding AUC values. As shown in Fig. 4a and the AUC column of Table 3, DRPADC achieved AUC = 0.876 in the 10 times 10-fold-CV. The other methods tested were BNNR (AUC = 0.876), SCPMF (AUC = 0.860), CMFMTL (AUC = 0.850), HGIMC (AUC = 0.785), GMCLDA (AUC = 0.735) and SIMCLDA (AUC = 0.705). The AUC value of DRPADC was the same as that of BNNR and better than those of the other methods.

To further evaluate the performance of DRPADC compared with other matrix completion methods on the HDVD dataset, we computed precision metrics using the same validation framework, plotted PR curves, and calculated the corresponding AUPR values. As shown in Fig. 4b and the AUPR column of Table 3, DRPADC had the highest AUPR value of 0.507, which is higher than the other methods, namely BNNR (AUPR = 0.489), SCPMF (AUPR = 0.486), CMFMTL (AUPR = 0.440), HGIMC (AUPR = 0.417), GMCLDA (AUPR = 0.177) and SIMCLDA (AUPR = 0.152). In summary, DRPADC outperforms the other methods in Fig. 4 on the HDVD dataset.

4. Discussion

4.1. Ablation experiments

To verify the validity of each part of the method proposed in this paper, ablation experiments were conducted in two gold-standard datasets, Fdataset and Cdataset. The comparisons made are shown in Table 4, where DRPADC is the one proposed in this paper (i.e., WKNKN + CKA-MKL + compressed sensing); model 1 is a combination of the WKNKN algorithm, similarity mean fusion, and compressed sensing; model 2 is a combination of CKA-MKL and compressed sensing; model 3 is a combination of similarity mean fusion and compressed sensing; model 4 is a combination of WKNKN, single similarity (with no introduction of Jaccard similarity or Gaussian kernel similarity for drugs and diseases), and compressed sensing; and model 5 uses single similarity (with no introduction of Jaccard similarity or Gaussian kernel similarity for drugs and diseases) and compressed sensing. As shown in Fig. 5, DRPADC showed a slight improvement in performance compared with model 1; this was because processing multiple similarity matrices using CKA-MKL enables the weights of each similarity matrix (kernel) to be obtained and the best matrix combination to be selected to improve the complementarity of the similarity information. All three sets of comparison models, DRPADC and model 2, model 1 and model 3, and model 4 and model 5, illustrate that introducing WKNKN to preprocess the drug-disease association matrix can improve the representativeness of the sampled features of the sub-matrices with respect to the original matrix data and improve the accuracy of association prediction. The comparison between model 2 and model 5 illustrates that the effectiveness of drug repositioning can be improved to some extent by enrichment with multiple similarity information.



（a）Fdataset （b）Cdataset

Fig. 3. ROC plots of the proposed method and other methods on the Fdataset (a) and Cdataset (b).

Table 2
AUC values for 10 times 10-fold-CV of various methods on Fdataset and Cdataset.

test	Metrics	DRPADC	BNNR	DRRS	HGIMC	MBiRW	SCPMF	HGBI	CMFMTL	DrugNet
10-fold-CV of Fdataset	AUC	**0.941** **(0.001)**	0.933 (0.002)	0.930 (0.001)	0.919 (0.001)	0.917 (0.001)	0.895 (0.001)	0.829 (0.012)	0.819 (0.001)	0.778 (0.001)
	AUPR	**0.521** **(0.001)**	0.402 (0.004)	0.341 (0.001)	0.394 (0.001)	0.264 (0.002)	0.357 (0.001)	0.102 (0.010)	0.195 (0.001)	0.155 (0.001)
10-fold-CV of Cdataset	AUC	**0.955** **(0.001)**	0.950 (0.001)	0.947 (0.002)	0.942 (0.002)	0.933 (0.003)	0.915 (0.001)	0.858 (0.014)	0.840 (0.001)	0.804 (0.001)
	AUPR	**0.607** **(0.001)**	0.441 (0.003)	0.378 (0.001)	0.428 (0.003)	0.310 (0.002)	0.423 (0.001)	0.129 (0.011)	0.305 (0.001)	0.201 (0.001)

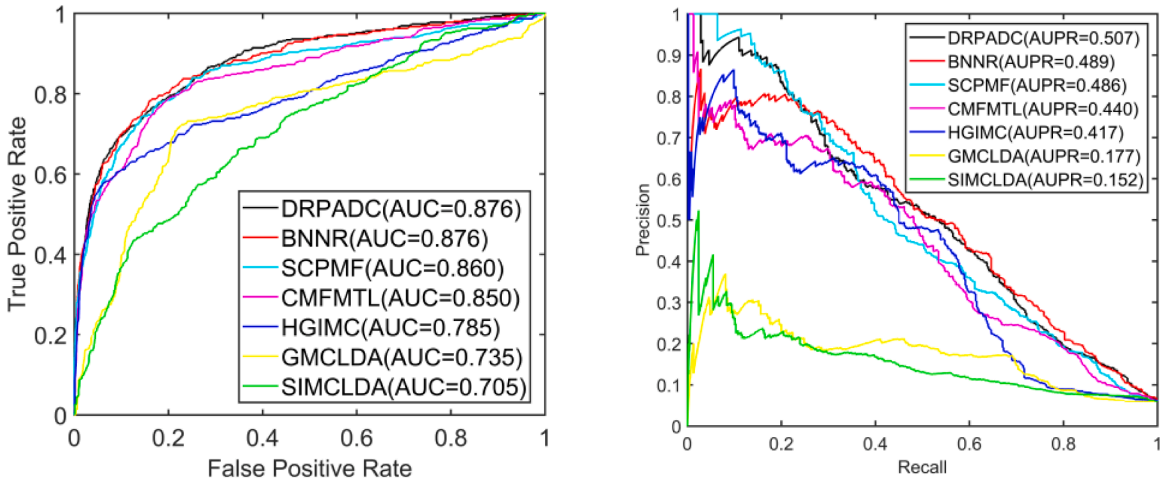


Fig. 4. ROC comparison graph (a) and PR comparison graph (b) for the proposed model and other matrix completion methods on the drug-virus dataset HDVD.

Table 3
AUC values and AUPR values for various matrix completion methods on HDVD.

Methods	AUC	AUPR
DRPADC	**0.876 (0.001)**	**0.507 (0.001)**
BNNR	0.876 (0.005)	0.489(0.013)
SCPMF	0.860 (0.001)	0.486 (0.001)
CMFMTL	0.850 (0.001)	0.440 (0.001)
HGIMC	0.785 (0.005)	0.417 (0.009)
GMCLDA	0.735 (0.005)	0.177 (0.003)
SIMCLDA	0.705 (0.007)	0.152 (0.008)

4.2. *Parameter analysis*

To verify the sensitivity of the performance of DRPADC to its parameters, we perform a parameter analysis on the Fdataset, Cdataset, and HDVD datasets, respectively. The parameters not involved in the conditioning experiments were set to fixed values for this analysis. When evaluating parameters $K$and T, the other parameters were fixed, i.e., λ_r, λ_M and λ_N were fixed as 0.1, 1 and 0.1, respectively. In the subsequent parameter evaluation experiments, the optimal values of the corresponding parameters were replaced in turn. As shown in Fig. 6, the values of the neighborhood parameter K and the attenuation parameter T in the WKNKN algorithm ranged from 1 to 10 and from 0.1 to 1, respectively. For Fdataset, the best performance was achieved with $K = 7$ and $T = 0.6$ (Fig. 6a). For Cdataset, the best performance was achieved with $K = 6$ and $T = 0.7$ (Fig. 6b). For HDVD, the best performance was located at $K = 5$ and $T = 0.3$ (Fig. 6c).

Next, we parametrically analyzed the parameters λ_r, λ_M and λ_N in the loss function of the compressed sensing algorithm in different datasets. All three parameters were set in the range of $[0.1, 1]$, incremented by 0.1 each time. As shown in Fig. 7a, the performance of DRPADC in all three datasets tended to decrease as the value of λ_r increased, and DRPADC performed best when λ_r was 0.1. As shown in Fig. 7b, the value of λ_M tended to level off after 0.4 and the performance reached stability, so we took the optimal value of λ_M in each of the three datasets: HDVD ($\lambda_M = 1$), Fdataset ($\lambda_M = 0.7$) and Cdataset ($\lambda_M = 0.5$). As shown in Fig.7c, the performance of DRPADC in the Fdataset and HDVD datasets showed a decreasing trend with increasing λ_N values, whereas in Cdataset it showed a decreasing trend from a λ_N value of 0.2. Therefore, we took the optimal values of λ_N in each of the three datasets: HDVD ($\lambda_N = 0.1$),

Table 4
Comparison of the combination used in various modules.

Model\module	WKNKN	CKA-MKL similarity	Mean similarity	Single similarity	Compressed sensing	AUC of Fdataset	AUC of Cdataset
DRPADC	√	√			√	0.941	0.955
Contrast Model 1	√		√		√	0.937	0.952
Contrast Model 2		√			√	0.933	0.946
Contrast Model 3			√		√	0.929	0.945
Contrast Model 4	√			√	√	0.930	0.946
Contrast Model 5				√	√	0.925	0.940

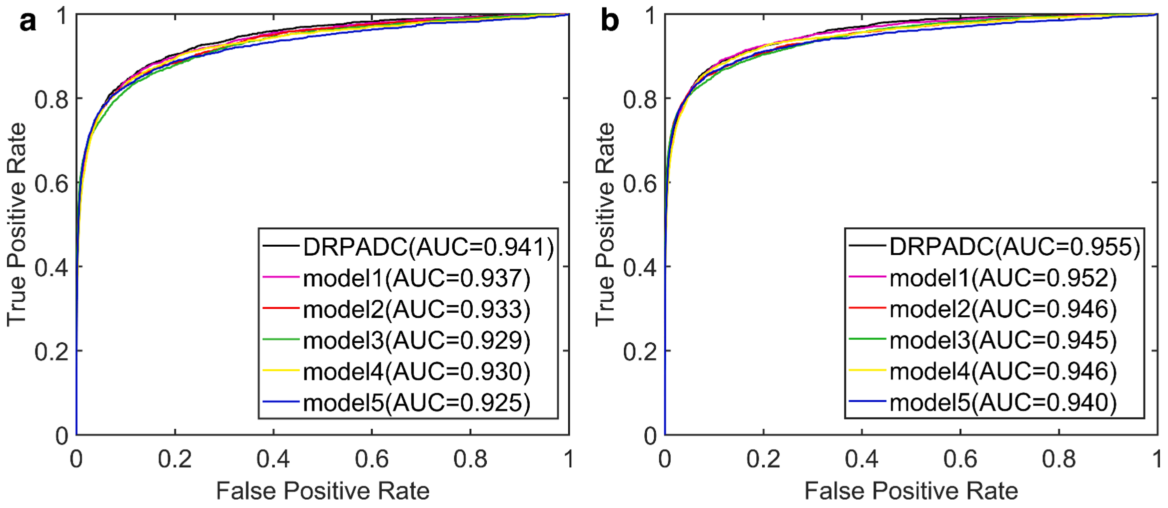
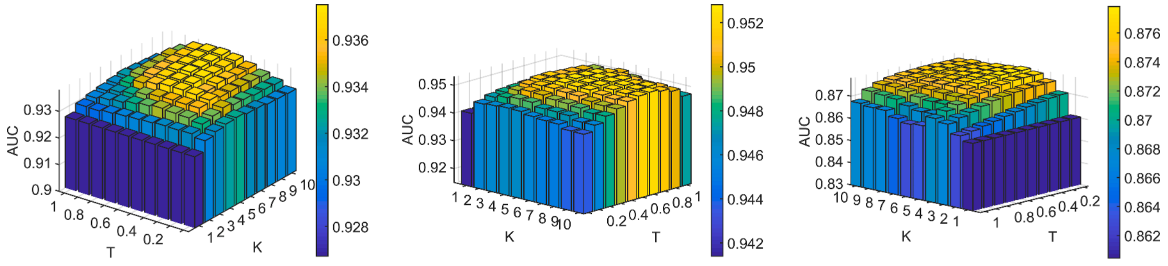


Fig. 5. ROC plots for the modules in the ablation experiments for the model proposed in this paper. (a) ROC plot for the experiment based on Fdataset; (b) ROC plot for the experiment based on Cdataset.



（a）Fdataset

（b）Cdataset

（c）HDVD

Fig. 6. Parameter analysis of the proposed model by 10 times 10-fold-CV in various datasets. (a) Sensitivity analysis of parameters K and T of WKNKN algorithm on Fdataset; (b) sensitivity analysis of parameters K and T of WKNKN algorithm on Cdataset; (c) sensitivity analysis of parameters K and T of WKNKN algorithm on HDVD.

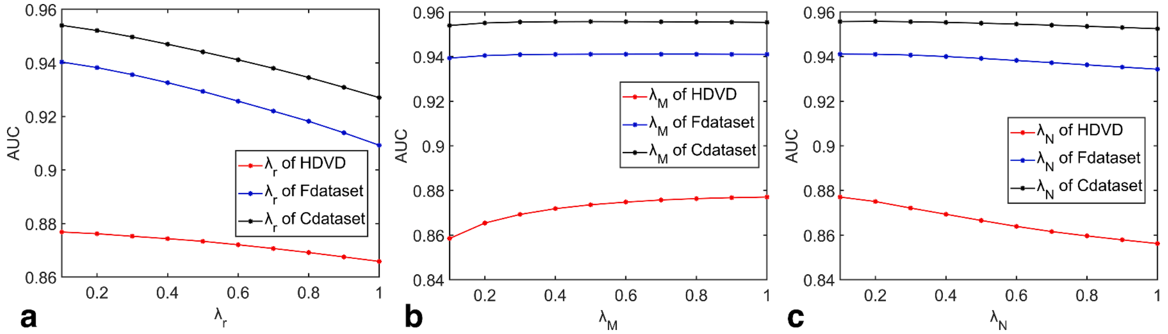


Fig. 7. Effect of parameters λ_r, λ_M and λ_N on model performance on the HDVD, Fdataset and Cdataset datasets, respectively.

Cdataset (λ_N = 0.2), Fdataset (λ_N = 0.1).

4.3. Case study

To validate the performance of DRPADC in practical applications, we performed a case study of COVID-19. The top 15 predicted drug candidates for COVID-19, as determined by their final scores of predicted association, are listed in Table 5. For each drug, the rank (predicted scores ranked in descending order), the registration number in the Drugbank library, the canonical name, and any relevant evidence reported in the literature are shown. Of these top 15 drug candidates, 12 drugs (80% success rate) have been validated by a variety of evidence. Chloroquine is an inexpensive, safe, and widely used anti-malarial drug which has been in use for over 70 years; it is highly effective *in vitro* control of SARS-CoV-2 infection and therefore may be clinically indicated for COVID-19 (Choy et al., 2020). Touret and Xavier de (2020) suggest that combined treatment with Remdesivir and Emetine may provide better clinical benefits. Wang et al. (2020) found that Chloroquine and Remdesivir were very effective in controlling SARS-CoV-2 infection *in vitro*. Ribavirin was originally suggested in clinical practice for use in the COVID-19 Pneumonia Diagnostic and Treatment Program (Revised 5th Edition) (Khalili et al., 2020). Therefore, Ribavirin is the first drug candidate predicted for possible use in the treatment of COVID-19. Camostat mesylate, one of the components of Camostat, blocks SARS-CoV-2 interaction in pulmonary cells and may be regarded for extra-label therapy of COVID-19 infection (Hoffmann et al.,

Table 5
Top 15 drug candidates for the therapy of COVID-19 identified by our approach.

Rank	Drugbank ID	Drug name	Evidences
1	DB00608	Chloroquine	(Choy et al., 2020)
2	DB00811	Ribavirin	(Khalili et al., 2020)
3	DB15660	N4-Hydroxycytidine	(Yajie et al., 2021)
4	DB13729	Camostat	(Hoffmann et al., 2020; Zhou et al., 2015)
5	DB00507	Nitazoxanide	(Khatri and Mago, 2020)
6	DB01024	Mycophenolic Acid	(Yajie et al., 2021)
7	DB12617	Mizoribine	Unknow
8	DB06803	Niclosamide	(Xu et al., 2020)
9	DB01043	Memantine	(Singh and Arkin., 2020)
10	DB00613	Amodiaquine	Unknow
11	DB00441	Gemcitabine	(Ya-Nan et al., 2020)
12	DB03718	6-Azauridine	Unknow
13	DB12139	Alisporivir	(Almasi and Mohammadipanah, 2020)
14	DB14761	Remdesivir	(Yajie et al., 2021; Al-Tawfiq et al., 2020; Grein et al., 2020)
15	DB13393	Emetine	(Touret and Xavier de, 2020)

2020; Zhou et al., 2015). A combination of Nitazoxanide and Camostat may be proposed as early clinical treatment and evaluation of COVID-19 based on its pathophysiological and pharmacological potential (Khatri and Mago, 2020). Niclosamide is an anthelmintic approved by the Food and Drug Administration that modulates a variety of functional signaling pathways and bioprocesses, and has been certified as a multi-purpose drug (Li et al., 2014, 2017). For instance, Niclosamide is effective against a diversity of viral infections, including ZIKV virus, MERS-CoV, SARS-CoV, hepatitis C virus, and human adenovirus (Andrews et al., 1982; Organization, 2019). Xu et al. (2020) envisage that Niclosamide may offer therapeutic potential against SARS-CoV-2. In addition, the E protein channel activity of SARS-CoV-2 was shown to be inhibited by memantine (Singh and Arkin., 2020). Remdesivir is a nucleotide analogue precursor drug with a broadly based spectrum of anti-viral activity, including activity against paramyxoviruses, pulmonary viruses, filoviruses and coronaviruses (Lo et al., 2017; Sheahan et al., 2017). Remdesivir inhibits viral RNA polymerase and shows anti-SARS-CoV-2 activity *in vitro* (Al-Tawfiq et al., 2020; De Wit et al., 2020; Grein et al., 2020). Gemcitabine is an effective broad-spectrum antiviral agent against a variety of RNA viruses, including MERS-CoV and SARS-CoV (Dyall et al., 2014), and a recent bioassay study demonstrated that it inhibited the multiplication of SARS-CoV-2 (Ya-Nan et al., 2020). The 18-kDa cytoplasmic cyclophilin A is an important cellular molecule required for the replication of RNA viruses, including HIV (Jeremy et al., 1993), HCV (Koichi et al., 2005) and coronaviruses (Almasi and Mohammadipanah, 2020). Recent studies have demonstrated that non-immunosuppressive analogs such as Alisporivir inhibit the activity of procyclins (Almasi and Mohammadipanah, 2020). Yajie et al. (2021) analyzed the effect of N4-Hydroxycytidine and Mycophenolic Acid on the functional receptor of SARS-CoV-2, ACE2 (cellular receptor for angiotensin-converting enzyme 2), using a molecular docking approach. Both drugs were shown to have several important binding sites for ACE2, suggesting a therapeutic effect of these drugs on COVID-19. A clinical trial of umifenovir alone was recently initiated in China (McKee et al., 2020); most of the 15 drug candidates for the treatment of COVID-19 predicted by our approach can be found in McKee et al. (2020).

In addition, DRPADC predicted the drug candidates Mizoribine, Amodiaquine, and 6-Azauridine. The antimalarial Chloroquine analogue Amodiaquine was shown to have inhibitory effects against MERS-CoV in a previous study (Dyall et al., 2014), and 6-Azauridine is a pyrimidine analogue that can inhibit a variety of viruses by inhibiting viral RNA synthesis (Cao et al., 2015). Viruses inhibited by 6-Azauridine include another human coronavirus, HCoV-NL63 (Krzysztof et al., 2006). Regarding Mizoribine, although there are no reports in the literature about its association with COVID-19, it has the potential to be a treatment for COVID-19. These predicted drug candidates provide promising directions for drug developers and will help to advance the drug development process.

5. Conclusion

COVID-19 is still spreading worldwide, and research in medicine and pharmacology is ongoing in the effort to develop therapeutic drugs and vaccines; however, the development progress remains slow owing to the various limitations of medical trials. In this study, to help advance the drug research process, we proposed a new compressed sensing algorithm combining a central kernel symmetry with a multicore learning algorithm called DRPADC to pinpoint drug candidates with high confidence intervals for the potential treatment of COVID-19. In this model, we used the WKNKN algorithm to reduce the sparsity of the drug-disease association matrix, thereby expanding the information complement of the model and improving its predictive performance. In addition, we improved the complementarity of various types of similarity information by fusing multiple drug and disease similarity matrices through CKA-MKL. Finally, we used a compressed sensing algorithm to calculate drug-disease association prediction scores. After validation of the experimental results, DRPADC achieved satisfactory results in the rapid discovery of COVID-19 drug candidates, outperforming drug-repositioning methods of similar type proposed in recent years.

However, there are still some points that should be considered regarding our approach in the future. Although we introduce multiple similarity data for drugs and diseases (or viruses), there are many types of similarity data that are not currently considered. In addition, DRPADC does not completely circumvent the effects of association sparsity. For new diseases and drugs with sparse association data or feature data, existing methods, including DRPADC, have not been able to achieve good prediction results. In future work, we will further investigate drug-repositioning methods and continue to develop new algorithms to cope with these current problems. In addition, owing to ongoing research, more correlation data will be added to the existing datasets in the future, providing quality data resources for drug-repositioning methods. We expect that drug-repositioning technology will help to advance the drug development process.

CRediT authorship contribution statement

Guobo Xie: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Haojie Xu:** Formal analysis, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Jianming Li:** Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Guosheng Gu:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – review & editing. **Yuping Sun:** Formal analysis, Funding acquisition, Supervision, Writing – review & editing. **Zhiyi Lin:** Formal analysis, Funding acquisition, Supervision, Writing – review & editing. **Yinting Zhu:** Formal analysis, Writing – review & editing. **Weiming Wang:** Formal analysis, Funding acquisition, Writing – review & editing. **Youfu Wang:** Project administration. **Jiang Shao:** Funding acquisition.

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.

Acknowledgment

We thank anonymous reviewers for valuable suggestions.

Fundings

This work was supported by the National Natural Science Foundation of China (72001202, 62002070, 61802072); the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University (2021013); the Science and Technology Plan Project of Guangzhou City (202102021236, 201902020012, 201902020006).

Availability of data and material

The available code and dataset of DRPADC can be downloaded from here https://github.com/kk-2010000/drpadc1.

Ethics approval

This paper does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate

All listed authors have read the submitted manuscript and agree to its submission.

Consent for publication

All listed authors consent for this publication.

References

Abd El-Aziz Tarek, M., Stockand James, D., 2020. Recent progress and challenges in drug development against COVID-19 coronavirus (SARS-CoV-2)-an update on the status. Infect. Genet. Evol. 83, 104327 https://doi.org/10.1016/j.meegid.2020.104327.

Al-Tawfiq, J.A., Al-Homoud, A.H., Memish, Z.A., 2020. Remdesivir as a possible therapeutic option for the COVID-19. Travel Med. Infect. Dis. 34, 101615 https://doi.org/10.1016/j.tmaid.2020.101615.

Ali, E., et al., 2016. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans. Comput. Biol. Bioinform. 14 (3), 646–656. https://doi.org/10.1109/TCBB.2016.2530062.

Almasi F., Mohammadipanah F., 2020. Potential targets and plausible drugs of coronavirus infection caused by 2019-nCoV. Authorea Preprints. 10.22541/au.158766083.33108969.

Andrews, P., Thyssen, J., Lorke, D., 1982. The biology and toxicology of molluscicides, bayluscide. Pharmacol. Ther. 19, 245–295. https://doi.org/10.1016/0163-7258(82)90064-X.

Assaf, G., et al., 2011. Predict: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. 7 (1), 496. https://doi.org/10.1038/msb.2011.26.

Bajusz, D., Racz, A., H́eberger, K., 2015. Why is tanimoto ́index an appropriate choice for fingerprint-based similarity calculations. J. Cheminform. 720 https://doi.org/10.1186/s13321-015-0069-3.

Cao, J., Forrest, J.C., Zhang, X., 2015. A screen of the NIH clinical collection small molecule library identifies potential anti-coronavirus drugs. Antivir. Res. 114, 1–10. https://doi.org/10.1016/j.antiviral.2014.11.010.

Catrin, S., et al., 2020. World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID19). Int. J. Surg. 76, 71–76. https://doi.org/10.1016/j.ijsu.2020.02.034.

Cha, Y., et al., 2018. Drug repurposing from the perspective of pharmaceutical companies. Br. J. Pharmacol. 175 (2), 168–180. https://doi.org/10.1111/bph.13798.

Chengqian, L., et al., 2018. Prediction of lncRNA–disease associations based on inductive matrix completion. Bioinformatics 34 (19), 3357–3364. https://doi.org/10.1093/bioinformatics/bty327.

Chengqian, L., et al., 2019. Predicting human lncrna-disease associations based on geometric matrix completion. IEEE J. Biomed. Heal. Inform. 24 (8), 2420–2429. https://doi.org/10.1109/JBHI.2019.2958389.

Choy, K.T., et al., 2020. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication *in vitro*. Antivir. Res. 104786 https://doi.org/10.1016/j.antiviral.2020.104786.

Christoph, S., et al., 2003. The chemistry development kit (cdk): an open-source java library for chemo-and bioinformatics. J. Chem. Inform. Comput. Sci. 43 (2), 493–500. https://doi.org/10.1021/ci025584y.

De Wit, E., et al., 2020. Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection. Proc. Nat. Acad. Sci. 117, 6771–6776. https://doi.org/10.1073/pnas.1922083117.

Dyall, J., et al., 2014. Repurposing of clinically developed drugs for treatment of middle east respiratory syndrome coronavirus infection. Antimicrob. Agents Chemother. 58, 4885–4893. https://doi.org/10.1128/AAC.03036-14.

Elisabetta, T., et al., 2004. An efficient method to make human monoclonal antibodies from memory B cells: potent neutralization of SARS coronavirus. Nat. Med. 10 (8), 871–875. https://doi.org/10.1038/nm1080.

Esra, G., Buket, K., 2017. A link prediction approach for drug recommendation in disease-drug bipartite network. In: Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE, pp. 1–4. https://doi.org/10.1109/IDAP.2017.8090219.

Feng, H., et al., 2020. Predicting drug-disease associations *via* multi-task learning based on collective matrix factorization. Front. Bioeng. Biotechnol. 8218 https://doi.org/10.3389/fbioe.2020.00218.

Fleuren Wilco, W.M., Wynand, A., 2015. Application of text mining in the biomedical domain. Methods 74, 97–106. https://doi.org/10.1016/j.ymeth.2015.01.015.

Grein, J., et al., 2020. Compassionate use of remdesivir for patients with severe Covid-19. N. Engl. J. Med. 382, 2327–2336. https://doi.org/10.1056/NEJMoa2007016.

Hakime, O., Elif, O., Arzucan̈, O., 2016. A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction. BMC Bioinform. 17 (1), 1–11. https://doi.org/10.1186/s12859-016-0977-x.

Han, S., et al., 2019. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. Genomics 111 (6), 1839–1852. https://doi.org/10.1016/j.ygeno.2018.12.007.

Hansaim, L., et al., 2016. Improved genome-scale multitarget virtual screening *via* a novel collaborative filtering approach to cold-start problem. Sci. Rep. 6 (1), 1–11. https://doi.org/10.1038/srep38860.

Hoffmann, M., et al., 2020. SARS-CoV-2 Cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell 181 (2), 271–280. https://doi.org/10.1016/j.cell.2020.02.052.

Huimin, L., et al., 2016. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics 32 (17), 2664–2671. https://doi.org/10.1093/bioinformatics/btw228.

Huimin, L., et al., 2016. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics 32 (17), 2664–2671. https://doi.org/10.1093/bioinformatics/btw228.

Huimin, L., et al., 2018. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics 34 (11), 1904–1912. https://doi.org/10.1093/bioinformatics/bty013.

Huimin, L., et al., 2020. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief. Bioinform. 22, 1604–1619. https://doi.org/10.1093/bib/bbz176.

Hurle, M.R., et al., 2013. Computational drug repositioning: from data to therapeutics. Clin. Pharmacol. Ther. 93 (4), 335–341. https://doi.org/10.1038/clpt.2013.1.

Imran, A., Gwanggil, J., 2021. Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. Interdiscipl. Sci. Comput. Life Sci. 14, 1–16. https://doi.org/10.1007/s12539-021-00465-0.

Jawad, R., et al., 2021. COVID-19 in the age of artificial intelligence: a comprehensive review. Interdiscipl. Sci. Comput. Life Sci. 13, 1–23. https://doi.org/10.1007/s12539-021-00431-w.

Jeremy, L., et al., 1993. Human immunodeficiency virus type 1 gag protein binds to cyclophilins A and B. Cell 73 (6), 1067–1078. https://doi.org/10.1016/0092-8674(93)90637-6.

John, D., Elad, H., Yoram, S., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (7), 2122–2125. https://www.jmlr.org/papers/volume12/duchi11a/duchi11a.pdf.

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30772–30780. https://doi.org/10.1093/molbev/mst010.

Khalili, J.S., et al., 2020. Novel coronavirus treatment with ribavirin: groundwork for an evaluation concerning COVID-19. J. Med. Virol. 92 (7), 740–746. https://doi.org/10.1002/jmv.25798.

Khatri, M., Mago, P., 2020. Nitazoxanide/Camostat combination for COVID-19: an unexplored potential therapy. Chem. Biol. Lett. 7, 192–196. http://pubs.iscience.in/journal/index.php/cbl/article/view/1085.

Koichi, W., et al., 2005. Cyclophilin B is a functional regulator of hepatitis C virus RNA polymerase. Mol. Cell 19 (1), 111–122. https://doi.org/10.1016/j.molcel.2005.05.014.

Krzysztof, P., et al., 2006. Inhibition of human coronavirus NL63 infection at early stages of the replication cycle. Antimicrob. Agents Chemother. 50 (6), 2000–2008. https://doi.org/10.1128/AAC.01598-0.

Li, Y., et al., 2014. Multi-targeted therapy of cancer by niclosamide: a new application for an old drug. Cancer Lett. 349, 8–14. https://doi.org/10.1016/j.canlet.2014.04.003.

Li, Z., et al., 2017. Existing drugs as broad-spectrum and potent inhibitors for zika virus by targeting NS2BNS3 interaction. Cell Res. 27, 1046–1064. https://doi.org/10.1038/cr.2017.88.

Liu, Y., et al., 2016. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput. Biol. 12, e1004760 https://doi.org/10.1371/journal.pcbi.1004760.

Lo, M.K., et al., 2017. GS-5734 and its parent nucleoside analog inhibit filo-, Pneumo-, and paramyxoviruses. Sci. Rep. 7, 43395. https://doi.org/10.1038/srep43395.

McKee, D.L., et al., 2020. Candidate drugs against SARS-CoV-2 and COVID-19. Pharmacol. Res. 157, 104859 https://doi.org/10.1016/j.phrs.2020.104859.

Mengyun, Y., et al., 2019. Drug repositioning based on bounded nuclearnorm regularization. Bioinformatics 35 (14), i455–i463. https://doi.org/10.1093/bioinformatics/btz331.

Mengyun, Y., et al., 2020. Heterogeneous graph inference with matrix completion for computational drug repositioning. Bioinformatics 36, 22–23. https://doi.org/10.1093/bioinformatics/btaa1024.

Muhammad, S.A., et al., 2014. Prioritizing drug targets in clostridium botulinum with a computational systems biology approach. Genomics 104 (1), 24–35. https://doi.org/10.1016/j.ygeno.2014.05.002.

O'Boyle, N.M., et al., 2011. Open babel: an open chemical toolbox. J. Cheminform. 333 https://doi.org/10.1186/1758-2946-3-33.

Organization W.H., 2019. World Health Organization model list of essential medicines: 21st List 2019. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/325771/WHO-MVP-EMP-IAU-2019.06-eng.pdf.

Pang-Ning T., Michael S., and Vipin K., 2016. Introduction to data mining. Pearson Education India. https://paulallen.ca/documents/2015/01/kumar-v-introduction-to-data-mining-instructors-solution-manual.pdf/.

Peng, L.H., et al., 2020. A computational study of potential mirna-disease association inference based on ensemble learning and kernel ridge regression. Front. Bioeng. Biotechnol. 8, 40. https://doi.org/10.3389/fbioe.2020.00040.

Poleksic, A., 2021. Overcoming sparseness of biomedical networks to identify drug repositioning candidates. IEEE/ACM Trans. Comput. Biol. Bioinform. 1. https://doi.org/10.1109/TCBB.2021.3059807.

Qi, Z., et al., 2019. Computational model development of drugtarget interaction prediction: a review. Curr. Protein Pept. Sci. 20 (6), 492–494. https://doi.org/10.2174/1389203720666190123164310.

Sheahan, T.P., et al., 2017. Broad-spectrum antiviral GS-5734 inhibits both epidemic and zoonotic coronaviruses. Sci. Transl. Med. 9 (396) https://doi.org/10.1126/scitranslmed.aal3653.

Singh, T.P.P., Arkin, I.T., 2020. SARS-CoV-2 E protein is a potential ion channel that can be inhibited by gliclazide and memantine. Biochem. Biophys. Res. Commun. 530 (1), 10–14. https://doi.org/10.1016/j.bbrc.2020.05.206.

Steck, H., 2010. Training and testing of recommender systems on data missing not at random. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 713–722. https://doi.org/10.1145/1835804.1835895.

Sudeep, P., et al., 2019. Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discov. 18 (1), 41–58. https://doi.org/10.1038/nrd.2018.168.

Touret, F., Xavier de, L., 2020. Of chloroquine and COVID-19. Antivir. Res. 177, 104762 https://doi.org/10.1016/j.antiviral.2020.104762.

Van Driel, M.A., et al., 2006. A textmining analysis of the human phenome. Eur. J. Hum. Genet. 14 (5), 535–542. https://doi.org/10.1038/sj.ejhg.5201585.

Victor, M., et al., 2015. Drug-net: network-based drug-disease prioritization by integrating heterogeneous data. Artif. Intell. Med. 63 (1), 41–49. https://doi.org/10.1016/j.artmed.2014.11.003.

Wang, M., et al., 2020. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. Cell Res. 30, 269–271. https://doi.org/10.1038/s41422-020- 0282-0.

Wei-jie, G., et al., 2020. Clinical characteristics of coronavirus disease 2019 in China. N. Engl. J. Med. 382 (18), 1708–1720. https://doi.org/10.1056/NEJMoa2002032.

Wenhui, W., Yang, S., Li, J., 2013. Drug target predictions based on heterogeneous graph inference. Biocomputing 2013, 53–64. https://doi.org/10.1142/97898144479730006.

Wenhui, W., et al., 2014. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30 (20), 2923–2930. https://doi.org/10.1093/bioinformatics/btu403.

Xing, C., et al., 2016. Drug–target interaction prediction: databases, web servers and computational models. Brief. Bioinform. 17 (4), 696–712. https://doi.org/10.1093/bib/bbv066.

Xing, C., et al., 2016. Nllss: predicting synergistic drug combinations based on semi-supervised learning. PLoS Comput. Biol. 12 (7), e1004975 https://doi.org/10.1371/journal.pcbi.1004975.

Xiujuan, L., Cheng, Z., 2020. Predicting metabolite disease associations based on linear neighborhood similarity with improved bipartite network projection algorithm. Complexity. https://doi.org/10.1155/2020/9342640.

Xu, J., et al., 2020. Broad spectrum antiviral agent niclosamide and its therapeutic potential. ACS Infect. Dis. 6, 909–915. https://doi.org/10.1021/acsinfecdis.0c00052.

Ya-Nan, Z., et al., 2020. Gemcitabine, lycorine and oxysophoridine inhibit novel coronavirus (SARS-CoV-2) in cell culture. Emerg. Microbes Infect. 9 (1), 1170–1173. https://doi.org/10.1080/22221751.2020.1772676.

Yajie, M., et al., 2021. Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. Appl. Soft Comput. 103, 107135 https://doi.org/10.1016/j.asoc.2021.107135.

Yijie, D., Jijun, T., Fei, G., 2019. Identification of drug-side effect association *via* multiple information integration with centered kernel alignment. Neurocomputing 325, 211–224. https://doi.org/10.1016/j.neucom.2018.10.028.

Yongcui, W., et al., 2013. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PloS One 8 (11), e78518. https://doi.org/10.1371/journal.pone.0078518.

Yosef, M.S., et al., 2020. Drug databases and their contributions to drug repurposing. Genomics 112 (2), 1087–1095. https://doi.org/10.1016/j.ygeno.2019.06.021.

Zhou, Y., et al., 2015. Protease inhibitors targeting coronavirus and filovirus entry. Antivir. Res. 116, 76–84. https://doi.org/10.1016/j.antiviral.2015.01.011.

Zou, Q., et al., 2015. Prediction of microRNA-disease associations based on social network analysis methods. Bio. Med. Res. Int. https://doi.org/10.1155/2015/810514.