OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–16

https://doi.org/10.1093/bib/bbab526

Problem Solving Protocol

A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2

Xiaorui Su, Lun Hu, Zhuhong You, Pengwei Hu, Lei Wang and Bowei Zhao

Corresponding author. L. Hu, Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China. Tel: +86 0991-3672967; E-mail: hulun@ms.xjb.ac.cn

Abstract

The outbreak of COVID-19 caused by SARS-coronavirus (CoV)-2 has made millions of deaths since 2019. Although a variety of computational methods have been proposed to repurpose drugs for treating SARS-CoV-2 infections, it is still a challenging task for new viruses, as there are no verified virus-drug associations (VDAs) between them and existing drugs. To efficiently solve the cold-start problem posed by new viruses, a novel constrained multi-view nonnegative matrix factorization (CMNMF) model is designed by jointly utilizing multiple sources of biological information. With the CMNMF model, the similarities of drugs and viruses can be preserved from their own perspectives when they are projected onto a unified latent feature space. Based on the CMNMF model, we propose a deep learning method, namely VDA-DLCMNMF, for repurposing drugs against new viruses. VDA-DLCMNMF first initializes the node representations of drugs and viruses with their corresponding latent feature vectors to avoid a random initialization and then applies graph convolutional network to optimize their representations. Given an arbitrary drug, its probability of being associated with a new virus is computed according to their representations. To evaluate the performance of VDA-DLCMNMF, we have conducted a series of experiments on three VDA datasets created for SARS-CoV-2. Experimental results demonstrate that the promising prediction accuracy of VDA-DLCMNMF. Moreover, incorporating the CMNMF model into deep learning gains new insight into the drug repurposing for SARS-CoV-2, as the results of molecular docking experiments reveal that four antiviral drugs identified by VDA-DLCMNMF have the potential ability to treat SARS-CoV-2 infections.

Keywords: SARS-CoV-2, drug repositioning, constrained multi-view nonnegative matrix factorization, deep learning, graph convolutional network

Introduction

Coronaviruses (CoVs) have become a major public health concern due to two severe CoV outbreaks at the beginning of 21st century, including severe acute respiratory syndrome-associated coronavirus (SARS-CoV) in 2002 [4] and Middle East respiratory syndrome-associated coronavirus (MERS-CoV) in 2012 [8]. Unfortunately, the coronavirus disease 2019 (COVID-19) caused by a new enveloped RNA β-CoV [55], named SARS-CoV-2, has produced a global pandemic since December 2019 [72] with tens of millions of infected people and millions of death. At present, there are still no proven effective drugs for SARS-CoV-2, despite increasing efforts made by pharmaceutical companies in drug development [15].

Due to the ability of significantly accelerating the drug development process, reducing overall costs and avoiding risks [31], drug repositioning is believed as a promising and efficient computational way to discover new indications of approved drugs. Hence, in order to response the urgent demand for effectively treating SARS-CoV-2 infections, a variety of drug repositioning methods have been recently proposed and they are broadly classified into two categories including structure-based and

Xiaorui Su is a doctoral student in Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi, China. Her research interests include machine learning, network representation learning, computational biology and bioinformatics.

Lun Hu, PhD, is a professor in Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include machine learning, big data analysis and its applications in bioinformatics.

Zhuhong You, PhD, is a professor in School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His research interests include neural networks, intelligent information processing, sparse representation and its applications in bioinformatics.

Pengwei Hu, PhD, is a professor in Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include machine learning, big data analysis and its applications in bioinformatics.

Lei Wang, PhD, is a professor in Big Data and Intelligent Computing Research Center, Guangxi Academy of Science, Nanning, China. His research interests include data mining, pattern recognition, machine learning, deep learning, computational biology and bioinformatics.

Bowei Zhao now is a doctoral student in Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Science, Urumqi, China. His research interests include machine learning, complex networks analysis, graph neural network and their applications in bioinformatics.

Received: September 16, 2021. **Revised:** October 20, 2021. **Accepted:** November 14, 2021

network-based methods [9]. Among them, structure-based methods target to identify chemical compounds that may act on SARS-CoV-2 by using molecular docking [49], whereas network-based ones predict novel virus-drug associations (VDAs), which associate SARS-CoV-2 to approved drugs, from a given VDA network. Though promising, structure-based methods suffer from the disadvantage of being time-consuming when searching for an effective compound against SARS-CoV-2 [1]. It is for this reason that we focus our study on network-based methods. However, for new viruses, such as SARS-CoV-2, without any known VDAs, predicting their potential drugs would certainly result in a well-known cold-start problem, which inhibits the development of accurate network-based methods to repurpose approved drugs [46].

To address this problem, network-based methods often integrate the biological knowledge of viruses into VDA networks, thus alleviating the influence exerted by the lack of VDAs involving new viruses [18]. In particular, IRNMFVDA [58] first constructs a VDA matrix based on VDAs, a drug similarity matrix and a virus similarity matrix. An indicator matrix is then used to determine the most likely drugs for SARS-CoV-2 with nonnegative matrix factorization (NMF). Similar to IRNMFVDA, SCPMF [40] also combines VDAs and the similarity information of drugs and viruses to generate a heterogeneous network and then identifies novel VDAs related to SARS-CoV-2 by utilizing similarity constrained probabilistic NMF. In addition to these NMF-based models, there are also several attempts made from an alternative view. For example, VDA-KATZ [70] considers the identification of novel VDAs as a problem of counting the number of connection paths between viruses and drugs in a heterogeneous network and applies a network-based association prediction model to infer possible drugs associated with SARS-CoV-2. Inspired by the link prediction works of layer attention graph convolution network (GCN)[66] and multi-view GCN [11], SANE [56] designs an attentive network embedding model by considering the sequence information of drugs and viruses as node attributes and potential drugs against COVID-19 can be identified with an attention-based pre-depth-first-search strategy. VDA-RWR [47] incorporates VDAs with the similarity information of drugs and viruses and then applies a random walk with restart method to estimate the probability of antiviral drugs against SARS-CoV-2.

Hence, the development of network-based methods has been inspired by the increasing coverage of genomic data [33], which gains new insights into the patterns characterizing already known VDAs to identify the missing ones. As mentioned above, state-of-the-art network-based methods rely on the perception that similar viruses are possibly be treated by the same drug. To this end, they construct a VDA matrix for a given VDA network and refine it by additionally considering the similarity information of drugs and viruses. However, such refinement fails to capture the structural and genomic forces that govern VDA networks if without placing further constraints to strengthen the perception of interest [17]. Taking NMF-based methods as an example, they adopt traditional NMF models to project viruses and drugs into a low-rank latent feature space (LFS), where not all similarity information of viruses and drugs are consistently preserved when compared with those in their original feature space [32]. Moreover, according to the results presented in Section of Ablation study, we note that the similarity between viruses and drugs, which is impossible to exist in reality, is observed from the visualization of their latent features obtained with NMF. With such noisy information, NMF-based methods assign a larger prediction score to virus-drug pairs that share more similar partners including not only drugs but also viruses and are prone to identify more false-positive VDAs by confusing the perception about how new viruses are associated with drugs. Our results suggest that the fundamental reason for the failure of network-based methods is the lack of such a constraint that accounts for the enhancement and consistency of our perception during refinement.

In this work, we propose a novel constrained multi-view nonnegative matrix factorization (CMNMF) model to ensure that, for drugs and viruses, their respective similarity information in a low-rank LFS are consistent with those in their original feature space by generating few noisy information. To do so, two similarity matrices are first constructed for drugs and viruses by using chemical structures and genome sequences respectively. Combining them with the VDA matrix, we further obtain an enhanced association matrix, each element of which indicates the association strength between corresponding drug and virus from a more comprehensive perspective. Modified from traditional NMF, CMNMF formulates additional constraints on these three matrices. In doing so, the optimization procedure of CMNMF is targeted toward to preserving the similarity information of drugs and viruses in LFS as much as possible. By taking the latent feature vectors of drugs and viruses as their initial node representations, we apply a graph convolutional network with attention-based neighbor sampling to optimize the representation of drugs and viruses in a given VDA network and then develop a deep learning method, namely VDA-DLCMNMF, for predicting potential drugs that can be used to treat infections caused by new viruses. The major contributions of this work are summarized as follows.

- Regarding network-based drug repurposing methods, our experimental results suggest that the fundamental reason for their failure is the lack of such a constraint that accounts for the enhancement and consistency of the intuitive perception about the potential drugs that new viruses are more likely to associate with.
- A novel CMNMF model is proposed to ensure that the respective similarity information of viruses and drugs are preserved when projected from their own

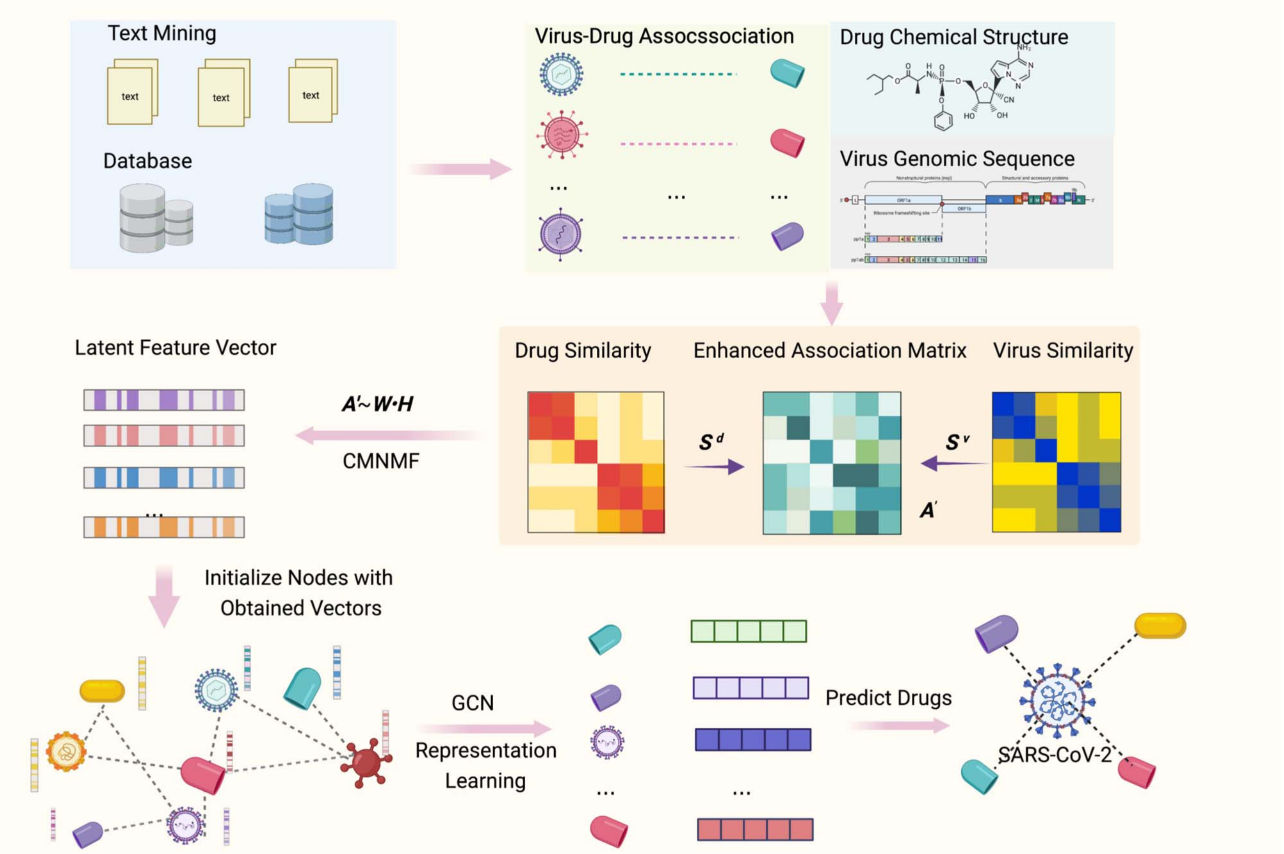

Figure 1. An illustration of the complete procedure of VDA-DLCMNMF.

feature spaces onto a unified LFS by generating few noisy information. To avoid random initialization, we take the latent feature vectors of viruses and drugs as their initial node representations and then develop a deep learning method, namely VDA-DLCMNMF, to precisely prioritize known drugs for new viruses.

- Experimental results on three VDA datasets with different size demonstrate the promising performance of VDA-DLCMNMF in repurposing antiviral drugs against SARS-CoV-2 in terms of several metrics. Besides, the results of molecular docking experiments reveal that incorporating CMNMF into deep learning gains new insight into the drug repurposing for SARS-CoV-2, as four novel drugs identified by our method are proved to have the potential ability to bind with important functional receptors of SARS-CoV-2.

The rest of this article is organized as follows. The section of Materials and Methods first describes the VDA datasets used in the experiments and then presents the details of VDA-DLCMNMF. Comparing VDA-DLCMNMF with several state-of-the-art models, we give the experimental results in the section of Experiments, following which we end the article with a in-depth discussion and a conclusion.

Table 1. The statistics of three datasets used in the experiments

Datasets	Viruses	Drugs	VDAs
HDVD	34	219	455
VDA1	11	78	96
VDA2	69	128	770

Materials and Methods

As shown in Figure 1, VDA-DLCMNMF is composed of three steps, including enhanced VDA matrix construction, LFS extraction with CMNMF and GCN-based drug repurposing. Before presenting the details of VDA-DLCMNMF, we describe the VDA datasets used in our experiments.

VDA datasets

To evaluate the performance of VDA-DLCMNMF, three datasets with different sizes are collected for discovering potential drugs against SARS-CoV-2, and they are denoted as HDVD, VDA1 and VDA2, respectively. The statistics of these three datasets are shown in Table 1.

HDVD [40] is a database for experimentally supported human drug-virus associations, built by assembling a significant number of experimentally validated drug–virus interaction entries from relevant literatures with

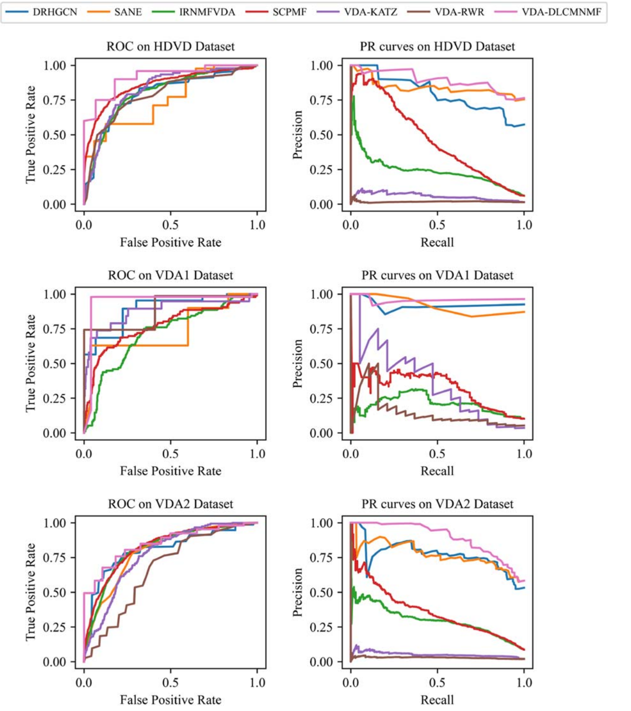

Figure 2. ROC and AUPR curves obtained by VDA-DLCMNMF and other competing methods on three datasets based on 5-fold CV.

text mining technology. HDVD includes 34 viruses, 219 drugs and 455 confirmed human drug–virus interactions.

VDA1 dataset is constructed based on the 96 known VDAs between 11 viruses similar to SARS-CoV-2, such as SARS-CoV [4], MERS-CoV [8] and influenza A viruses [10] and 78 small molecular drugs. These interactions are collected from the DrugBank [63], NCBI [50] and PubMed [5] datasets.

VDA2 dataset is collected from the DrugVirus.info database [2], which provides various experimentally validated VDA-related resources. After removing the viruses with incomplete genome sequences, VDAs contain totally 770 VDAs between 69 viruses and 128 drugs.

Enhanced VDA matrix construction

As mentioned before, the main idea of drug repurposing for new viruses follows the perception that similar viruses are more likely to be treated by the same drugs. In this step of VDA-DLCMNMF, we target to construct an enhanced VDA matrix by seamlessly considering both VDAs and the biological knowledge of drugs and viruses. To this end, we first construct two similarity matrices based on chemical structures of drugs and genomic sequences of viruses and then design a new association measure to obtain the enhanced VDA matrix.

In addition, it is supposed to illustrate the notations used throughout this paper before introducing VDA-DLCMNMF in detail. The vector is denoted by lowercase boldface letters (e.g. $\mathbf{v} \in \mathbb{R}^{d}$), matrix is denoted by uppercase boldface letters (e.g. $\mathbf{X} \in \mathbb{R}^{m\times n}$). Then, we denote the set by $\mathcal{D}(\cdot)$, hyperparameter by uppercase letters (e.g. T) and scalars by lowercase letters (e.g. d, k).

The similarity matrix of drugs is denoted as $\mathbf{S^d} \in \mathbb{R}^{n_d\times n_d}$, where n_d is the number of drugs, and it is constructed on the chemical structures, including atom, bond, branch, closed-loop and break specifications, obtained from simplified molecular input line entry system (SMILES) [62]. As one of the most commonly used database for molecular structures [60], SMILES has been widely used in calculating the similarity of drugs from the structural perspective. We download the SMILES database from DrugBank [63] with version 5.1.8 and use Babel chemistry toolbox [45] with version 2.3.1 to convert SMILES strings into molecular access system (MACCS) chemical fingerprints, each bit of which indicates the existence of a particular substructure in the compound. Given the MACCS fingerprints of all drugs, the Tanimoto index [67] is adopted to the measure the fingerprint-based molecular similarity between pairwise drugs. Assuming that d_i and d_j are two drugs and their MACCS fragment bit-strings sets are $\mathcal{D}(\mathrm{i})$ and $\mathcal{D}(j)$, respectively, $\mathbf{S}^{\mathbf{d}}_{ij}$ is the similarity between d_i and d_j in $\mathbf{S^d}$ and its value can be computed with (1).

$$\mathbf{S}^{\mathbf{d}}_{ij} = \frac{\mathcal{D}(\mathrm{i}) \cap \mathcal{D}(j)}{\mathcal{D}(\mathrm{i}) \cup \mathcal{D}(j)} \tag{1}$$

Regarding the similarity of viruses, the genomic sequences of viruses are used and they can be downloaded from the NCBI [50]. Let $\mathbf{S^v} \in \mathbb{R}^{n_v\times n_v}$, where n_v denotes the number of viruses, be the similarity matrix of viruses. For each element of $\mathbf{S^v}$, its value can be obtained by using a sequence alignment software MAFFT [24].

A VDA network is a bipartite graph, where edges are only existed between viruses and their associated drugs. A $n_d \times n_v$ matrix, denoted as $\mathbf{A}$, is introduced to represent the topological structure of a given VDA network. For an arbitrary element, say $\mathbf{A}_{i,j}$, its value is 1 if d_i and v_j are associated and 0 otherwise. The purpose of this step is to construct an enhanced association matrix, i.e. $\mathbf{A}^{'}$, by integrating $\mathbf{S^d}$, $\mathbf{S^v}$ and $\mathbf{A}$. With $\mathbf{A}^{'}$, we are able to ensure that (i) similar viruses are more likely to be associated with the same drugs and (ii) similar drugs are more likely to be associated with the same viruses. Obviously, the former one is of particular significance to repurpose drugs for new viruses. $\mathbf{A}^{'}$ can be obtained with (2).

$$\mathbf{A}^{'} = \mathbf{S^d}\mathbf{A}\mathbf{S^v} \tag{2}$$

According to the above formula, it is possible that the values of elements in $\mathbf{A}^{'}$ are much larger than those in $\mathbf{S^v}$ and $\mathbf{S^d}$. Hence, we adopt a min-max normalization to each row of $\mathbf{A}^{'}$, thus constraining the values within a range $[0, 1]$. For each element of $\mathbf{A}^{'}$, its value is given by (3).

$$\mathbf{A}^{'}_{i,j} = \frac{\mathbf{A}^{'}_{i,j} - min(\mathbf{A}^{'}_{i,:})}{max(\mathbf{A}^{'}_{i,:}) - min(\mathbf{A}^{'}_{i,:})}, \tag{3}$$

where $\mathbf{A}'_{i,:}$ is the i-th row of $\mathbf{A}'$ and $max(\mathbf{A}'_{i,:})$ and $min(\mathbf{A}'_{i,:})$ return the maximum and minimum values respectively in $\mathbf{A}'_{i,:}$.

The CMNMF model

Regarding the task of drug repurposing for new viruses, the failure of existing network-based methods is due to their incapability of capturing the structural and genomic forces that play an important role in determining VDAs. To overcome this problem, we propose the CMNMF model and present its details as below.

Assuming that k is the dimension of LFS, we define two matrices, i.e. $\mathbf{W} \in \mathbb{R}^{k\times n_d}$ and $\mathbf{H} \in \mathbb{R}^{k\times n_v}$, to denote the respective projection results of drugs and viruses in the LFS. Regarding the product of $\mathbf{W}$ and $\mathbf{H}$, CMNMF uses it to approximate $\mathbf{A}'$ rather than $\mathbf{A}$, which is commonly used by existing network-based methods. Thus, $\mathbf{W}$ and $\mathbf{H}$ are derived by using (4). In doing so, our perception about new viruses can be further enhanced.

$$\mathbf{A}' \approx \mathbf{W^T H} \tag{4}$$

In order to preserve the similarity information of drugs and viruses presented in their respective feature spaces, additional constraints are introduced by CMNMF from different views as given by (5) and (6), where $\mathbf{w_i} = \mathbf{W_{i,:}}$ is the latent feature vector of drug d_i, $\mathbf{h_j} = \mathbf{H_{j,:}}$ is the latent feature vector of virus v_j, and $||\cdot||_F$ is the Frobenius norm.

$$\sum_{i=1}^{n_d}\sum_{m=1}^{n_d} \mathbf{S}^d_{i,m}||\mathbf{w_i} - \mathbf{w_m}||_F^2 \tag{5}$$

$$\sum_{j=1}^{n_v}\sum_{n=1}^{n_v} \mathbf{S}^v_{j,n}||\mathbf{h_j} - \mathbf{h_n}||_F^2 \tag{6}$$

With (5) and (6), the optimization of $\mathbf{W}$ and $\mathbf{H}$ is driven not only by $\mathbf{A}'$ but also by $\mathbf{S}^d$ and $\mathbf{S}^v$. A theoretical analysis is provided to verify the rationality behind these two constraints. Taking (5) as an example, since $\mathbf{S}^d_{i,m}$ is a constant, the minimization of (5) can be achieved if the value of $||\mathbf{w_i} - \mathbf{w_m}||_F^2$ is small enough. In this regard, the latent feature vectors of drugs tend to gather together in the LFS. Moreover, if two drugs, i.e. d_i and d_m, are similar, they certainly have a larger value of $\mathbf{S}^d_{i,m}$, and thus $||\mathbf{w_i} - \mathbf{w_m}||_F^2$ should be smaller than others in order to minimize (5). In other words, the latent feature vectors of two drugs are much closer in the LFS if they are more similar. Given the above analysis, it is believed that the introduction of (5) and (6) allows CMNMF to preserve the similarity information of drugs and viruses in the LFS with only few noisy information generated.

The complete objective function of CMNMF is defined as

$$\begin{aligned}\mathcal{J}(\mathbf{W},\mathbf{H}) =& \frac{1}{2}\sum_{i=1}^{n_d}\sum_{j=1}^{n_v}(\mathbf{A}'_{i,j} - \mathbf{w_i}^T\mathbf{h_j})^2 \\ &+ \frac{\alpha}{2}\sum_{i=1}^{n_d}\sum_{m=1}^{n_d}\mathbf{S}^d_{i,m}||\mathbf{w_i} - \mathbf{w_m}||_F^2 \\ &+ \frac{\beta}{2}\sum_{j=1}^{n_v}\sum_{n=1}^{n_v}\mathbf{S}^v_{j,n}||\mathbf{h_j} - \mathbf{h_n}||_F^2 \\ &+ \frac{\lambda_W}{2}||\mathbf{W}||_F^2 + \frac{\lambda_H}{2}||\mathbf{H}||_F^2 \\ & s.t. \mathbf{W} \geq 0, \mathbf{H} \geq 0,\end{aligned} \tag{7}$$

where α, β, λ_W and λ_H are the regularization coefficients. The purpose of CMNMF is to find $\mathbf{W}$ and $\mathbf{H}$ that minimize $\mathcal{J}(\mathbf{W},\mathbf{H})$. Here, a stochastic gradient decent method is adopted to solve this minimization problem. In particular, we first eliminate the inequality constraints by introducing Lagrange multipliers and obtain $\mathcal{L}(\mathbf{W},\mathbf{H})$ as follows.

$$\begin{aligned}\mathcal{L}(\mathbf{W},\mathbf{H}) =& \frac{1}{2}Tr(\mathbf{A}'\mathbf{A}'^T) - Tr(\mathbf{W}^T\mathbf{H}\mathbf{A}'^T) \\ &+ \frac{1}{2}Tr(\mathbf{W}^T\mathbf{H}\mathbf{H}^T\mathbf{W}) + \frac{\alpha}{2}Tr(\mathbf{W}\mathbf{Q}_d\mathbf{W}^T) \\ &+ \frac{\beta}{2}Tr(\mathbf{H}\mathbf{Q}_v\mathbf{H}^T) + \frac{\lambda_W}{2}Tr(\mathbf{W}\mathbf{W}^T) + \frac{\lambda_H}{2}Tr(\mathbf{H}\mathbf{H}^T) \\ &+ Tr(\Phi\mathbf{W}^T) + Tr(\Psi\mathbf{H}^T).\end{aligned} \tag{8}$$

In (8), $Tr(\cdot)$ represents the matrix trace, $\Phi = [\phi_{ik}]$ and $\Psi = [\psi_{jk}]$ are the Lagrange multipliers for the inequalities of (7), i.e. $\mathbf{w_{ik}} \geq 0$ and $\mathbf{h_{jk}} \geq 0$, respectively, $\mathbf{Q_d}$ and $\mathbf{Q_v}$ are the Laplacian similarity matrices of $\mathbf{S}^d$ and $\mathbf{S}^v$, respectively. After that, we solve the minimization problem of $\mathcal{L}(\mathbf{W},\mathbf{H})$ by using its the Karush–Kuhn–Tucker conditions and obtain the update rules of $\mathbf{W}$ and $\mathbf{H}$ as follows.

$$\mathbf{w}'_{\mathbf{ik}} \leftarrow \mathbf{w_{ik}}\frac{(\mathbf{H}\mathbf{A}')_{ik} - (\alpha\mathbf{W}\mathbf{Q_d})_{ik}}{(\mathbf{H}\mathbf{H}^T\mathbf{W})_{ik} + (\lambda_W\mathbf{W})_{ik}} \tag{9}$$

$$\mathbf{h}'_{\mathbf{jk}} \leftarrow \mathbf{h_{jk}}\frac{(\mathbf{W}\mathbf{A}')_{jk} - (\alpha\mathbf{H}\mathbf{Q_v})_{jk}}{(\mathbf{W}\mathbf{W}^T\mathbf{H})_{jk} + (\lambda_H\mathbf{H})_{jk}} \tag{10}$$

An iterative procedure is applied by VDA-DLCMNMF to obtain the optimum results of $\mathbf{W}$ and $\mathbf{H}$ until a convergence is reached.

GCN-based drug repurposing

Without loss of generality, we still use $\mathbf{W}$ and $\mathbf{H}$ to denote their optimum results in terms of minimizing (8) and give the details of how VDA-DLCMNMF repurposes drugs for new viruses.

Compared with traditional network representation learning algorithms, such as DeepWalk [48] and Node2Vec [14], GNN has the advantage of processing and integrating node features [51]. As a specific type of GNN, spatial-based GCN [28, 68] receives much attention due to its high efficiency and flexibility in dealing with the heterogeneous information of input. Thus, VDA-DLCMNMF adopts it to learn the node representations of viruses and drugs in a given VDA network and then uses them to predict potential associations for new viruses.

When incorporating GCN into VDA-DLCMNMF, we adopt the attention-based strategy in neighborhood sampling for giving an accurate learning. To do so, the attention coefficients for all VDAs are first calculated, and then they are used to weight VDAs with the softmax function. Given a drug d_i and a virus v_j, e_{ij} and α_{ij} are defined as its attention coefficient and weight, respectively, and their definitions are given as follows.

$$e_{ij} = \mathcal{A}(\mathbf{w_i}, \mathbf{h_j}) \tag{11}$$

$$\alpha_{ij} = softmax_j(e_{ij}) = \frac{exp(e_{ij})}{\sum_{p \in \mathcal{N}(i)} exp(e_{ip})} \tag{12}$$

In (11), $\mathcal{A} : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ is a shared attentional mechanism. $\mathcal{N}(i)$ in (12) denotes the set of all VDAs involving d_i. For the sake of convenience, the attention weights for all pairs of d_i and v_j can be represented with a matrix $\mathbf{W^a} \in \mathbb{R}^{n_d \times n_v}$, which can be explicitly calculated with (13).

$$\mathbf{W^a} = softmax((\mathbf{W^T H}) \times \mathbf{A}) \tag{13}$$

A heuristic neighborhood sampling strategy is adopted by VDA-DLCMNMF to improve the efficiency of training a GCN. Taking d_i, or v_j, as an example, we select the top T viruses, or drugs, to compose $\mathcal{T}(d_i)$, or $\mathcal{T}(v_j)$, in the descending order of attention weights in $\mathbf{W^a}$. Since $\mathbf{W^a}$ is already determined before training, this process is only applied once. The representations of drugs and viruses are then updated iteratively by using $\mathcal{T}(d_i)$ and $\mathcal{T}(v_j)$, respectively, rather than the whole VDA network, thus reducing the training time.

Here, d_i is taken as an example to illustrate the process of learning its representation. Let $\mathbf{r}^{(l)}_{\mathcal{T}(d_i)}$ and $\mathbf{r}^{(l)}_{d_i}$ denote the neighbor information of d_i and the representation of d_i respectively at the l-th layer in the GCN, their definitions are

$$\mathbf{r}^{(l)}_{\mathcal{T}(d_i)} = \sum_{v_j \in \mathcal{T}(d_i)} \mathbf{W}^{\mathbf{a}}_{i,j} \mathbf{r}^{(l-1)}_{v_j} \tag{14}$$

$$\mathbf{r}^{(l)}_{d_i} = \sigma(\mathbf{W}^{(l)} \cdot CONCAT(\mathbf{r}^{(l-1)}_{d_i}, \mathbf{r}^{(l)}_{\mathcal{T}(d_i)}), \tag{15}$$

where $\mathbf{W}^{(l)}$ is the weight matrix of the l-th layer, $CONCAT(\cdot)$ is the concatenation function applied to learn the neighbor information from $T(d_i)$, and $\sigma(\cdot)$ is the sigmoid activation function. Regarding the initialization of $\mathbf{r}_{d_i}$, we have $\mathbf{r}^{\mathbf{(0)}}_{\mathbf{d_i}} = \mathbf{w_i}$ to avoid a random initialization. Similarly, we can also obtain $\mathbf{r}_{v_j}$ for an arbitrary virus v_j.

Assuming that there are total L layers in the GCN, for a new virus v_j, its probability of being associated with a drug d_i is computed by multiplying their final representations as given by (16). The activation function allows the result of s_{ij} fall into the range $[0, 1]$.

$$s_{ij} = \sigma(\mathbf{r}^{(L)}_{d_i} \cdot \mathbf{r}^{(L)}_{v_j}) \tag{16}$$

Complexity analysis

At each epoch, $\mathbf{W}$, $\mathbf{H}$,$\mathbf{W}^a$ and $\mathbf{r}^{(L)}$ are updated according to Equations (9), (10) and (13)–(15). The update of $\mathbf{W}$ and $\mathbf{H}$ takes time $O(kn_dkn_v)$. The time required to calculate $\mathbf{W}^a$ is $O(n_dkn_v)$. After calculating the attention weight matrix, we select the top T neighborhoods for each node in VDA network, which takes time $O(2n_dTn_v)$. The computation of $\mathbf{r}^{(L)}$ has the time complexity $O(T^L)$. Hence, the time used for one iteration is $O((k + k^2 + 2T)n_dn_v + T^L)$. In this study, we normally have $T \ll n_d, n_v, k$ and $L = 2$. As a result, the time complexity can be further simplified to $O(k^2n_dn_v)$. Assuming that the number of epochs is E, the overall time complexity is $O(Ek^2n_dn_v)$.

Results

Evaluation metrics and experimental settings

Five-fold cross-validation (CV) is used to evaluate the performance of VDA-DLCMNMF. We perform the 5-fold CVs by alternatively selecting one fold as the test set and the rest as the training set. The negative samples are selected on each fold to ensure that no unseen node is generated in negative dataset. In other words, negative samples are selected by paring up drugs and viruses whose associations are not found in each fold. To do so, we first obtain the complementary set of VDAs in each fold, then randomly select the negative samples with the same size of positive samples from the complementary set so as to compose the negative dataset. As a result, positive and negative samples are balanced in each fold.

Additionally, five evaluation metrics, including accuracy (Acc.), sensitivity (Sen.), specificity (Spe.) AUC and AUPR, are adopted to measure the performance. Definitions of the first three metrics are given as

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

$$Sen. = \frac{TP}{TP + FN} \tag{18}$$

$$Spe. = \frac{TN}{TN + FP}, \tag{19}$$

Table 2. Parameter settings used by VDA-DLCMNMF and four competing state-of-the-art methods

Methods\ Datasets	HDVD	VDA1	VDA2
IRNMFVDA	$\alpha = \beta = 0.8, \lambda_1 = \lambda_2 = 0.1$	$\alpha = \beta = 0.8, \lambda_1 = \lambda_2 = 0.1$	$\alpha = \beta = 0.8, \lambda_1 = \lambda_2 = 0.1$
SCPMF	$\lambda_W = \lambda_H = 1, \lambda_1 = \lambda_2 = 0.1$	$\lambda_W = \lambda_H = 1, \lambda_1 = \lambda_2 = 0.1$	$\lambda_W = \lambda_H = 1, \lambda_1 = \lambda_2 = 0.1$
VDA-KATZ	$\beta = 0.04, w = 0.9, \gamma_v = \gamma_d = 2.5,$	$\beta = 0.04, w = 0.9, \gamma_v = \gamma_d = 2.5$	$\beta = 0.04, w = 0.9, \gamma_v = \gamma_d = 2.5$
VDA-RWR	$r = 0.5, \mu = 0.7, \alpha = 0.7$	$r = 0.7, \mu = 0.9, \alpha = 0.5$	$r = 0.5, \mu = 0.9, \alpha = 0.9$
VDA-DLCMNMF	$\alpha = \beta = 0.003, \lambda_W = \lambda_H = 0.005$	$\alpha = \beta = 0.005, \lambda_W = \lambda_H = 0.1$	$\alpha = \beta = 0.002, \lambda_W = \lambda_H = 0.1$

where *TP*, *FP*, *TN* and *FN* denote the numbers of true positive, false positive, true negative and false negative associations, respectively. AUC is the area under the receiver operating characteristic (ROC) curve, which can be plotted by true positive rate and false positive rate. AUPR is the area under the precision-recall curve, which can be plotted by precision and recall (Sen.).

In the experiments, VDA-DLCMNMF is compared with five state-of-the-art network-based drug repositioning methods, including DRHGCN [35], SANE [56], IRNMFVDA [58], SCPMF [40], VDA-RWR [47] and VDA-KATZ [70]. Among them, DRHGCN is a general drug repurposing method by introducing a layer attention mechanism to combine the embeddings of drugs and viruses from multiple graph convolution layers, SANE addresses the cold-start problem by introducing an LSTM unit to learn initial representations for drug and virus from drug chemical structure and virus genomic sequence, while the other three methods are specifically proposed for discovering potential antiviral drugs for SARS-CoV-2 as mentioned in the section of Introduction. All the experiments are performed on the working machine equipped with Intel Core I7 2.6GHz and 16GB RAM.

Regarding the parameters involved in VDA-DLCMNMF, an in-depth analysis of parameter sensitivity is performed to determine their optimal values with grid search. We determine their optimal values by using the method of control variates, which alternatively update each parameter. In specific, we vary the values of α, β, λ_W and λ_H from 0 to 1 at a step size of 0.001. Given the constraints that $\alpha = \beta$ and $\lambda_W = \lambda_H$, we conduct several trials with different combinations of parameter values and select the one with the best performance as the final values of these parameters. In addition, regarding the parameters used in GCN, we select the layer $L = 2$ and $T = 4$. For the other competing methods, their parameters are assigned with the values recommended in relevant literature. The parameter settings for all the five methods are presented in Table 2.

Performance comparison of different methods

The detailed results of 5-fold CV are shown in Table 3. To visually compare the performance of all five methods, we also plot their ROC curves and PR curves in Figure 2.

For the performance of two NMF-based methods, we note that SCPMF performs better than IRNMFVDA across all three datasets, as the scores of Acc. obtained by SCPMF are better by 38.05%, 9.23% and 23.43% than those of IRNMFVDA on HDVD, VDA1 and VDA2, respectively. Besides, for IRNMFVDA, its fluctuation in Acc. is more intensive, as the standard deviation of Acc. yielded by IRNMFVDA is 0.098 larger than the others. The reasons attributable for the unsatisfactory performance of IRNMFVDA are 2-fold. First, the indicator matrix used by IRNMFVDA is constructed based on a VDA matrix and the attribute matrices of drugs and viruses and a simple concatenation of these three matrices makes the indicator matrix sparser. Taking HDVD as an example, the VDA network of HDVD is the sparsest when compared with those of the other two datasets and the worst performance of IRNMFVDA is observed on it. Second, when applied to different datasets, the generalization ability of IRNMFVDA is poor due to the lack of a proper normalization process.

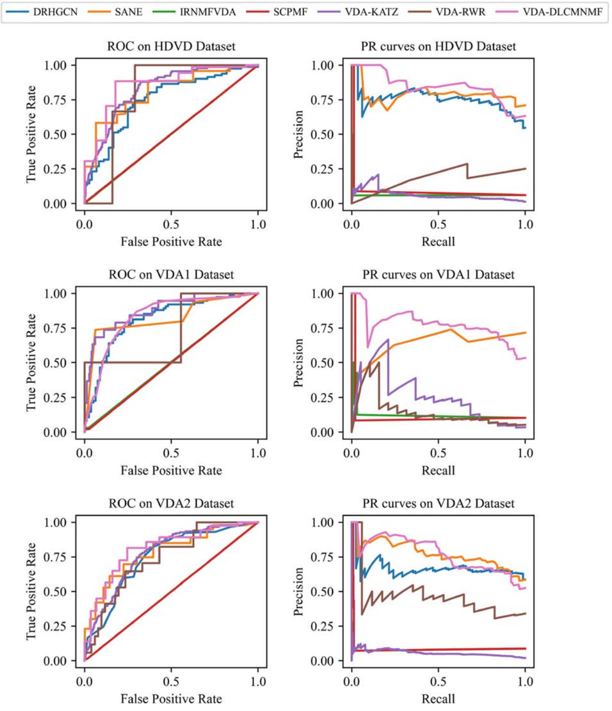

Figure 3. ROC and AUPR curves obtained by VDA-DLCMNMF and other competing methods on three datasets for de novo test.

Regarding VDA-KATZ and VDA-RWR, although both of them are network-based methods, they perform quite differently when predicting novel VDAs in the experiments. In particular, the best performance of VDA-KATZ is achieved in the 5-fold CV of VDA2, which could be an indicator that VDA-KATZ prefers dense VDA networks.

Table 3. Experimental results of 5-fold CV

Methods\ Datasets	**HDVD**					**VDA1**					**VDA2**				
	Acc.	**Sen.**	**Spe.**	**AUC**	**AUPR**	**Acc.**	**Sen.**	**Spe.**	**AUC**	**AUPR**	**Acc.**	**Sen.**	**Spe.**	**AUC**	**AUPR**
DRHGCN	0.7713	0.7689	0.7791	0.7713	0.7845	0.7298	0.7299	0.7841	0.8085	0.8271	0.7029	0.7030	0.7689	0.8177	0.7449
SANE	0.8352	0.8103	0.8580	0.8944	0.8598	0.7705	0.7489	0.7834	0.8080	0.8379	**0.8019**	0.7008	0.7307	0.8018	0.7553
IRNMFVDA	0.3856	0.7314	0.3631	0.8037	0.2156	0.6227	0.4852	0.6834	0.7102	0.2005	0.5317	0.6496	0.5205	0.8147	0.2971
SCPMF	0.7661	0.4693	0.7854	0.8549	0.4783	0.7150	0.4403	0.7464	0.7543	0.3684	0.7660	0.3969	0.8012	0.8293	0.3517
VDA-KATZ	0.5777	0.6995	0.5762	0.8253	0.0698	0.6691	0.6976	0.6684	0.8803	0.3380	0.7119	0.5441	0.7152	0.7743	0.0583
VDA-RWR	0.6550	0.7400	0.6700	0.7875	0.0218	0.8278	0.4824	0.7831	0.8582	0.1383	0.6613	0.5022	0.6643	0.6675	0.0322
VDA-DLCMNMF	**0.8649**	**0.8625**	**0.9118**	**0.9299**	**0.9097**	**0.9000**	**0.9667**	**0.9333**	**0.9250**	**0.9715**	0.7849	**0.7688**	**0.8361**	**0.8631**	**0.8770**

Best results are bolded.

The main reason for that phenomenon is ascribed to the motivation of VDA-KATZ, which is to utilize network paths to predict potential associations. In doing so, KDA-KATZ tends to yield a better performance for dense VDA networks with more network paths. Compared with VDA-KATZ, VDA-RWR is not always better than VDA-KATZ across all datasets. Among the results in Table 3, it is observed that for the VDA1 dataset, VDA-RWR performs the best among all methods except VDA-DLCMNMF in terms of Acc., whereas for the other two datasets, its performance in terms of Acc. is just fair. The main reason why VDA-RWR yields the second best score of Acc. on the VDA1 dataset is that VDA-RWR has a strong ability in learning the topological characteristics from small VDA networks. However, its worse performance obtained on HDVD and VDA2 demonstrates that its learning ability is constrained by the sparsity of VDA networks, which heavily affects the effectiveness of random walk from different perspectives. In particular, when dealing with sparse VDA networks, viruses and drugs with less degrees are difficult to be visited by the random walk of VDA-RWR, as their topological information is not sufficient. On the other hand, for dense VDA networks, more associations are involved during random walk, but the existence of false-positive associations could decrease the accuracy of VDA-RWR. Moreover, both VDA-KATZ and VDA-RWR achieve the low AUPR value with only 0.1554 and 0.0641 on average, which indicates that they are easily misled by false positive samples, especially when compared with those deep learning-based methods, including DRHGCN, SANE and VDA-CMNMF.

Regrading two deep learning-based methods DRHGCN and SANE, both of them have relatively stable performances among all metrics, which is mainly because of the attention mechanism adopted in them. Though both DRHGCN and SANE are constructed based on GCN, SANE performs better than DRHGCN across all three datasets, as the scores of Acc. obtained by SANE are better by 6.39%, 4.07% and 9.9% than those of DRHGCN on three datasets, respectively. The reason for this is that the information contained in SANE is not only the network topology but also the drug/virus attribute feature learned by LSTM from drug chemical structure/virus genomic sequence. On the other hand, though DRHGCN performs as not well as SANE, it still achieves the better performance than the other baselines, which further demonstrate that robustness of deep learning-based methods.

Among all drug repurposing methods compared in the experiments, VDA-DLCMNMF yields a bigger margin in terms of Acc., AUC and AUPR across all datasets, as it achieves the best performance on HDVD, VDA1 and VDA2. There are also several points worth further commentary. First, compared with NMF-based methods, VDA-DLCMNMF adopts the CMNMF model to preserve the similarity information of drugs and diseases when projecting them onto a unified LFS and thus yields a promising performance in drug repurposing. Second, the generalization ability of VDA-DLCMNMF is further improved by learning the network spatial structure with attention-based layer aggregation, and it is also for this reason that VDA-DLCMNMF yields the largest scores of Acc. and AUC in all cases. In summary, it is believed that VDA-DLCMNMF is a useful tool to discover novel VDAs.

De novo VDA prediction

To assess the capability of VDA-DLCMNMF in predicting potential indications for new drugs, we choose the drug pair with chemical similarity over 0.7 to conduct a de novo test. For each of this kind of drug pairs, we randomly select a drug and remove all VDAs related to it in turn as the test samples and other existing associations are used as training sample. The random selection procedure has been repeated for 50 times and the ROC and PR curves obtained by each prediction model are depicted in Figure 3.

First of all, according to the ROC and PR curves, we note that the overall performance of VDA-DLCMNMF is better than the other competing prediction models, as its average AUC and AUPR scores are the largest across all the three datasets. In this regard, we reason that VDA-DLCMNMF is more robust toward the bias resulted from the existence of redundancy drugs. Second, regarding the NMF-based models, i.e. IRNMFVDA and SCPMF, their ROC and PR curves indicate that their performances in the *de novo* test are much worse than those in the 5-fold CV. Hence, the performance of NMF-based models is heavily dependent on the similarity of drugs, which is an important information source for them to learn latent vectors. Lastly, the deep learning-based models,

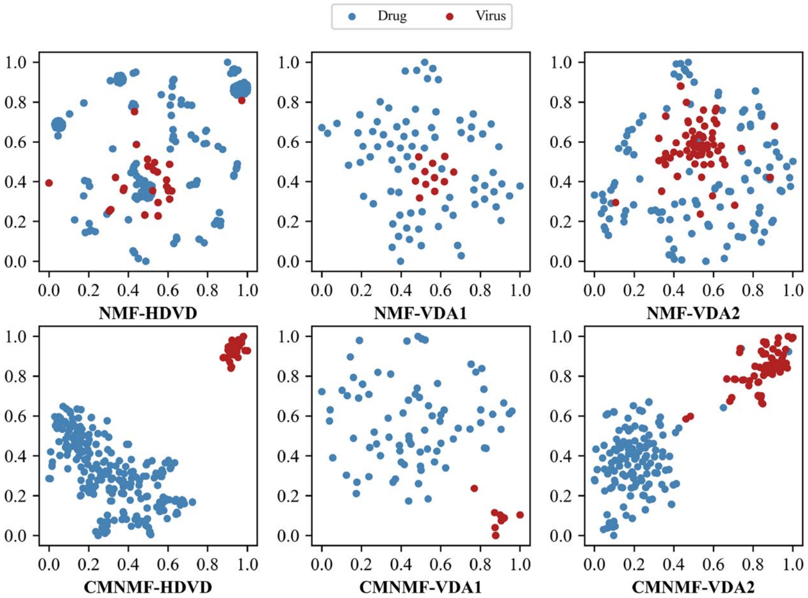

Figure 4. The latent feature vectors of drugs and viruses obtained by NMF and CMNMF are visualized in a 2D space by t-SNE.

including DRHGCN and SANE, are less affected by the removal of redundant drugs, as they do not need the similarity information of drugs to perform their tasks.

Ablation study

To better demonstrate the advantage of VDA-DLCMNMF in drug repurposing, an in-depth ablation study has been conducted with extensive experiments. To do so, we first design four variants of VDA-DLCMNMF and evaluate their performance on the datasets of HDVD, VDA1 and VDA2. A detailed description about these variants are listed as below, and the difference between them and VDA-DLCMNMF are also discussed.

- **NMF** is designed without considering any constraints. The input of NMF is only the VDA network, and the latent vectors of drugs and viruses are used as their final representations. Given a pair of drug and disease, their probability of being associated is computed with (16).
- **CMNMF** is used alone to discover novel VDAs. Similar to NMF, the latent vectors of drugs and viruses learned from CMNMF are explicitly used to compute the probability with (16).
- **GCN**$_{random}$ is a variant of the last part of VDA-DLCMNMF, which selects neighbor receptive field in a random manner.
- **GCN** is the last part of VDA-DLCMNMF. In the ablation study, the representation of drugs and viruses are randomly initialized for GCN.
- **NMF-GCN** is implemented by combining NMF and GCN. The main difference between NMF-GCN and VDA-DLCMNMF lies in the initial representation of drugs and viruses for GCN, as NMF-GCN and VDA-DLCMNMF use the latent feature vectors learned by NMF and CMNMF, respectively.

Experimental results of the variants of VDA-DLCMNMF are shown in Table 4. As mentioned before, the CMNMF model is adopted by VDA-DLCMNMF to obtain the reliable initial representations of drugs and viruses. Compared with NMF, CMNMF integrates S_d and S_v into A, thus obtaining an enhanced association matrix, i.e. $\mathbf{A}'$. To demonstrate the advantage and effectiveness of CMNMF, we first compare the performance of CMNMF with that of NMF on all the three datasets. As indicated by Table 4, it is seen that in terms of Acc., CMNMF perform better by 3.75%, 9.37% and 0.94% than NMF for the datasets of HDVD, VDA1 and VDA2, respectively. This could be a strong indicator that the latent vectors obtained by CMNMF are capable of retaining the characteristics of drugs and viruses from the perspectives of network topology and biological knowledge and hence they are able to improve the prediction accuracy of VDA-DLCMNMF.

In addition to the quantitative analysis, we visualize the latent vectors of drugs and viruses in a 2D space by t-SNE [59] and expect that the advantage of CMNMF can be better appreciated. According to Figure 4, it is observed that CMNMF is able to clearly distinguish drugs and viruses while introducing only few noisy information. In other words, the latent feature vectors obtained by CMNMF are more representative than those obtained by NMF. Moreover, we also note from Table 4 and Figure 4 that CMNMF contributes more in improving the accuracy of VDA-DLCMNMF on VDA1 than on the other two datasets, and the reasons are 2-fold. First, among all datasets, VDA1 is the smallest one with only 96 VDAs, and such a small VDA network is more sensitive to the quality of latent feature vectors of drugs and viruses. Second, comparing the distributions of latent vectors in 2D space, we note that the latent vectors obtained by NMF from VDA1 are uniformly distributed and they are not separated into different clusters as CMNMF do. In this

Table 4. Experimental results of ablation study

Methods\ Datasets	HDVD					VDA1					VDA2				
	Acc.	Sen.	Spe.	AUC	AUPR	Acc.	Sen.	Spe.	AUC	AUPR	Acc.	Sen.	Spe.	AUC	AUPR
NMF	0.7810	0.7802	0.7824	0.7811	0.8268	0.7395	0.6974	0.7672	0.7400	0.8522	0.7680	0.7558	0.7765	0.7680	0.8063
CMNMF	0.8185	0.8154	0.8220	0.8185	0.8694	0.8332	0.8442	0.8295	0.8334	0.9059	0.7774	0.7309	0.7955	0.7849	0.8388
GCN_{random}	0.7081	0.8201	0.6318	0.7965	0.7157	0.6250	0.6767	0.5833	0.5967	0.5670	0.7000	0.6423	0.7422	0.7769	0.7573
GCN	0.7516	0.6669	0.7761	0.8034	0.7761	0.6547	0.6137	0.7204	0.7606	0.7204	0.7558	0.7658	0.7656	0.8191	0.7657
NMF-GCN	0.8000	**0.8656**	0.7541	0.8979	0.8532	0.8500	0.8667	0.9095	0.9133	0.9333	0.7290	0.7343	0.7278	0.8018	0.7852
VDA-DLCMNMF	**0.8649**	0.8625	**0.9118**	**0.9299**	**0.9097**	**0.9000**	**0.9667**	**0.9333**	**0.9250**	**0.9715**	**0.7849**	**0.7688**	**0.8361**	**0.8631**	**0.8770**

Best results are bolded.

regard, the rationality behind the proposal of CMNMF can be verified.

Due to the advantage of CMNMF, VDA-DLCMNMF performs better than both NMF-GCN, GCN_{random} and GCN, which further indicates that initializing the representation of drugs and viruses with their latent feature vectors obtained by CMNMF provides a more effective way for VDA-DLCMNMF to discover novel VDAs. It is noted that regrading GCN_{random}, its performance in terms of Acc. is the worst among all competing models for each dataset. This could be a strong indicator that for GCN without attention mechanism, its performance is prone to be influenced by the noisy in VDA networks. However, the performance of GCN is still not as good as NMF-GCN, even if when it adopts attention mechanism, a conclusion thus be made that GCN is not applicable to predict isolated nodes in a given network, such as new viruses in VDA network, due to the random initialization of node representation. On the other hand, according to the performance of VDA-DLCMNMF and NMF-GCN, the accuracy of GCN can be improved by integrating with either CMNMF or NMF, which is able to solve the cold-start problem for isolated nodes. Additionally, the performance of CMNMF and NMF is also improved with the integration of GCN that strengthens the learning ability in terms of network representation. Therefore, concerning the respective advantages of CMNMF and GCN, it is the integration of them that leads to the promising performance of VDA-DLCMNMF in drug repurposing.

Application to drug repositioning of diseases

In order to prove the robustness of the DLCMNMF, we implement it on two golden standard datasets, including Fdataset [13] and Cdataset [37], for drug repurposing of diseases. We then compare the performance of CMNMF with two cutting-edge methods DRRS [36] and BNNR [65] on two datasets, respectively. In this section, 5-fold CV is used to evaluate the performance of DLCMNMF and two baselines, and all parameters are the same as their original works. The results are shown in Table 5.

Although both of DRRS and BNNR are trained on the same heterogeneous drug-disease networks, BNNR performs better than DRRS across all evaluation metrics expect Sen., and the main reason for that phenomenon is that BNNR specifically designs a relaxed penalty function to process noisy entries [65]. Regarding the performance of VDA-DLCMNMF, we note that VDA-DLCMNMF performs better by 0.17% and 35.46%, 0.94% and 27% than BNNR in terms of AUC and AUPR on Fdataset and Cdataset, respectively. Since BNNR demonstrates its ability in distinguishing negative samples as indicated by Spe., it also yields the best performance in terms of Acc.. However, for the task of drug repurposing, we are more concerned with the ability of discovering precise drug-disease associations. In this regard, VDA-DLCMNMF is preferred over BNNR, as the Sen. scores obtained by VDA -DLCMNMF are almost 60% and 48% larger than those of BNNR on FDataset and CDataset, respectively. In other words, the prediction accuracy of VDA-DLCMNMF in terms of Sen. AUC and AUPR could be a strong indicator that VDA-DLCMNMF is better in distinguishing between true positive samples and false negative samples when compared with DRRS and BNNR. Hence, we have reason to believe that VDA-DLCMNMF is also promising tool for the task of drug repurposing.

Identifying potential drugs for SARS-CoV-2

Taking SARS-CoV-2 as an example, we apply VDA-DLCMNMF to discover potential drugs that can be used to treat SARS-CoV-2 from the datasets of HDVD, VDA1 and VDA2. Specifically, for each VDA network of these datasets, we first apply VDA-DLCMNMF to obtain the probability of being associated with SARS-CoV-2 representation for each drug with (16) and then select the top 10 drugs in ascending order of probability for giving a detailed analysis. The results are shown in Table 6.

We note that 8, 7 and 9 out of the Top 10 drugs discovered in HDVD, VDA1 and VDA2, respectively, have been validated by recent publications. Among these validated drugs, Remdesivir obtains the largest probability scores in both HDVD and VDA1, and it has been recently recognized as a promising antiviral drug against a wide array of RNA viruses infection in cultured cells, mice and nonhuman primate models [39, 61]. As an adenosine analogue, Remdesivir incorporates into nascent viral RNA chains, thus resulting in the pre-mature termination. Moreover, Remdesivir is able to inhibit the viral infection of Vero-E6 cells by clinically isolating SARS-CoV-2 in an *in vitro* assay [61]. In the dataset of VDA2, chlorpromazine is predicted as the most likely drug for the

Table 5. Experimental results of case study on drug repositioning of diseases

Methods\ Datasets	**Fdataset**					**Cdataset**				
	Acc.	**Sen.**	**Spe.**	**AUC**	**AUPR**	**Acc.**	**Sen.**	**Spe.**	**AUC**	**AUPR**
DRRS	0.8314	0.5241	0.8374	0.9093	0.3512	0.8345	0.5249	0.8378	0.9093	0.3489
BNNR	**0.9576**	0.3637	**0.9638**	0.9280	0.5634	**0.9632**	0.4236	**0.9683**	0.9407	0.6566
DLCMNMF	0.8541	**0.9501**	0.7965	**0.9297**	**0.9090**	0.8432	**0.9058**	0.7737	**0.9501**	**0.9266**

Best results are bolded.

Table 6. The predicted top 10 drugs associated with SARS-CoV-2 on three datasets

Datasets	**Rank**	**Drug name**	**Evidence**	**Rank**	**Drug name**	**Evidence**
HDVD	1	Remdesivir	PMID:32020029	6	Chloroquine	PMID:32074550
	2	Tenofovir	–	7	Rimantadine	PMID:31133031; PMID:15288617
	3	EIDD-2801	[53]	8	Equilin	PMID:27169275; PMID:32194980
	4	Dactinomycin	PMID:1335030; PMID:32194980	9	Camostat	PMID:22496216
	5	Ribavirin	PMID:22555152	10	Berberine	–
VDA1	1	Remdesivir	PMID:32020029	6	Indinavir	PMID:15144898
	2	Cobicistat	–	7	Camostat	PMID:22496216
	3	Mycophenolic acid	PMID:5799033	8	Tenofovir	–
	4	Ribavirin	PMID:22555152	9	FK506	–
	5	Chloroquine	PMID:32074550	10	Zanamivir	PMID:15200845
VDA2	1	Chlorpromazine	PMID:8811199	6	Quinacrine	PMID:23301007
	2	Chloroquine	PMID:32074550	7	Tenofovir	–
	3	Gemcitabine	PMID:24841273	8	Indomethacin	PMID:5284360
	4	Ribavirin	PMID:22555152	9	Camostat	PMID:22496216
	5	Favipiravir	[44]	10	Zanamivir	PMID:15200845

treatment of SARS-CoV-2, and it is widely used to study virus entry by clathrin-mediated endocytosis of several viruses, including West Nile virus and influenza virus [6]. Since SARS-CoV also utilizes the clathrin-mediated endocytosis pathway for entry into host cell [6], it is possible for chlorpromazine to act similarly on MERS-CoV and SARS-CoV as a potential broad-spectrum CoV inhibitor.

It can also be observed from Table 6 that three drugs, including ribavirin, chloroquine and camostat, are found among the Top 10 drugs in all the three datasets. In particular, ribavirin is an approved antiviral drug to inhibit the production of Inosine-5′-monophosphate dehydrogenase, which interacts with the viral protein nsp14 [64]. It is for this reason that ribavirin has been recommended in the clinical practice for SARS-CoV-2 pneumonia diagnosis and Treatment Plan Edition 5-Revised [27]. As a traditional drug for the treatment of malaria, chloroquine phosphate is shown to have apparent efficacy and acceptable safety against COVID-19 associated pneumonia based on multi-center clinical trials conducted in China [12]. In addition to the ability of increasing the endosomal PH required for virus/cell fusion, Chloroquine is capable of interfering with the glycosylation of cellular receptors of SARS-CoV [29]. As a result, chloroquine is recommended in the next version of the Guidelines for the Prevention, Diagnosis, and Treatment of COVID-19 pneumonia, which is issued by the National Health Commission of the People's Republic of China for the treatment of COVID-19 infection for larger populations in future [12]. Known as one of commercial serine protease inhibitors, Camostat partially blocks the infection of SARS-CoV [20]. Moreover, when used together with cathepsin inhibitor EST, it can effectively prevent both cell entry and the multistep growth of SARS-CoV in human Calu-3 airway epithelial cells [25, 71].

In summary, the above analysis demonstrates the promising performance of VDA-DLCMNMF in discovering potential drugs for SARS-CoV-2, as most of the Top 10 drugs predicted by VDA-DLCMNMF are found to be effective when used to treat SARS-CoV-2 according to a careful literature review. Moreover, such a high accuracy could be also a strong indicator that VDA-DLCMNMF is able to precisely discover potential drugs for a new virus.

Molecular docking experiment

To further explain the reliability of VDA-DLCMNMF, we have conducted structure-based molecular docking experiments [42] to all the drugs listed in Table 6. For each drug, we compute its intermolecular binding ability with SARS-CoV-2 spike protein or human angiotensin-converting enzyme 2 (ACE2), which are important functional receptors for SARS and other CoVs [19, 34].

Specifically, we first download the structures of SARS-CoV-2 spike receptor-binding domain bound with ACES (PDB ID: 6M0J) from RCSB Protein Data Bank [3], and the chemical structures of drugs are obtained from the DrugBank in the PDB format. After that, the PDB data of

Table 7. Binding energies between predicted drugs and the SARS-CoV-2 spike protein/ACE2

Drug name	Binding energy (kcal/mol)	Drug name	Binding energy (kcal/mol)
Berberine	−7.39	Camostat	−7.43
Chloroquine	−6.40	Chlorpromazine	−6.82
Cobicistat	−7.93	Dactinomycin	−2.29
EIDD-2801	−5.45	Equilin	−7.68
Favipiravir	−4.24	FK506	−9.72
Gemcitabine	−4.89	Indinavir	−8.95
Indomethacin	−6.43	Mycophenolic acid	−5.60
Quinacrine	−6.50	Remdesivir	−7.25
Ribavirin	−6.87	Rlmantadine	−6.67
Tenofovir	−6.44	Zanamivir	−5.80

drugs are converted into pdbqt files by AutoDockTools [41]. For each drug, its pdbqt file is considered as the input of AutoDock software, with which we are able to complete the molecular docking experiment by taking the spike protein and ACE2 as receptors and each drug as a ligand of interest. The experimental results of molecular docking on all the 20 drugs in Table 6 are shown in Table 7, where the binding energies of these drugs are recorded. When using AutoDock to conduct molecular docking experiments, one should note that the binding energy is the binding free energy. For an arbitrary drug, the lower its binding energy is, the stronger its binding ability is.

We note that for Remdesivir, its binding energy with SARS-CoV-2 spike protein/ACE2 is −7.39 kcal/mol while that for Chlorpromazine is −6.82 kcal/mol. For ribavirin, chloroquine and camostat that are listed among the Top 10 drugs in all datasets, their binding energies are −6.87 kcal/mol, −6.40 kcal/mol and −7.43 kcal/mol, respectively. Overall, the binding energies of these five drugs are positioned at a relatively lower level as indicated by Table 7. This finding further validates the eligibility of these drugs in treating SARS-CoV-2.

According to the Table 6, there are a total of four drugs, including tenofovir [26], Berberine [16], cobicistat [30] and FK506 [52], yet to be validated, as there is no evidence to confirm their effort for the treatment of SARS-CoV-2. Hence, we have also conducted molecular docking experiments for these four drugs and presented their binding sites in Figure 5, where the green and cyan parts denote the structures of ACE2 and SARS-CoV-2 spike protein, respectively. It is observed from Table 7 that the binding synergies of these four drugs are even lower than several validated drugs, such as favipiravir, gemcitabine and dactinomycin, thus indicating a strong association they have with SARS-CoV-2. Moreover, particular attention is given to cobicistat and FK506, which obtain lower binding energies when compared with ribavirin and camostat. In particular, FK506 has the lowest binding energy with SARS-CoV-2 spike protein/ACE2 among all the 20 drugs. Overall, we reason that the associations of these four drugs are possibly existed, but missed by laboratory experiments, and thus they are likely to have therapeutic effects against SARS-CoV-2. It also should be noted that molecular docking do not necessarily prove that the drug can treat SARS-CoV-2, the results obtained by molecular docking just provide a therapeutic possibility. Accurate results require in-depth follow-up experimental verifications.

Discussion and Conclusion

To facilitate the development of antiviral drugs against new diseases, we propose a novel deep learning-based method, namely VDA-DLCMNMF, for drug repurposing with CMNMF and apply it to discover novel drugs that are more likely to treat SARS-CoV-2. Regarding the drug repurposing for new viruses, the major difficulty lying in here is that there are no known associations between new viruses and existing drugs; hence, it is of great significance to effectively solve the cold-start problem. To this end, we first construct an enhanced association matrix by integrating VDAs, chemical structures of drugs and genomic sequences of viruses. After that, the CMNMF model is designed to address the cold-start problem by precisely constructing the latent feature vectors of drugs and viruses in a unified LFS from the structural and genomic perspectives. VDA-DLCMNMF then adopts a GCN with attention-based neighbor sampling to learn the representations of drugs and viruses, which are initialized with their latent feature vectors at beginning. The probability of a drug being associated with a new virus can thus be computed based on their final representations. Extensive experiments have been conducted to evaluate the performance of VDA-DLCMNMF, and their results demonstrate the superior accuracy of VDA-DLCMNMF on three datasets created for SARS-CoV-2 when comparing it with several state-of-the-art drug repurposing methods. Moreover, for each dataset, most of the top 10 drugs predicted by VDA-DLCMNMF are validated by literature review. Even for those without any evidence, the results of molecular docking, to some extent, indicate their potential ability in treating SARS-CoV-2.

There are several reasons to explain the success of VDA-DLCMNMF in drug repurposing for SARS-CoV-2. First of all, the selection of proper biological knowledge of drugs and viruses, as well as how to process them, provides a solid basis for the following steps of VDA-DLCMNMF. Obviously, it is impossible for existing drug repurposing methods to discover novel VDAs of new viruses if without any other source of information about viruses, especially in an effort to discover potential drugs for new viruses. To this end, we make use of the chemical structures of drugs and the genomic sequences of viruses, and integrate them into a given VDA network for the purpose of constructing an enhanced association matrix. With such a matrix, we are able to strengthen our perception about the formation of VDAs involving new viruses, which is that similar viruses are more likely to be

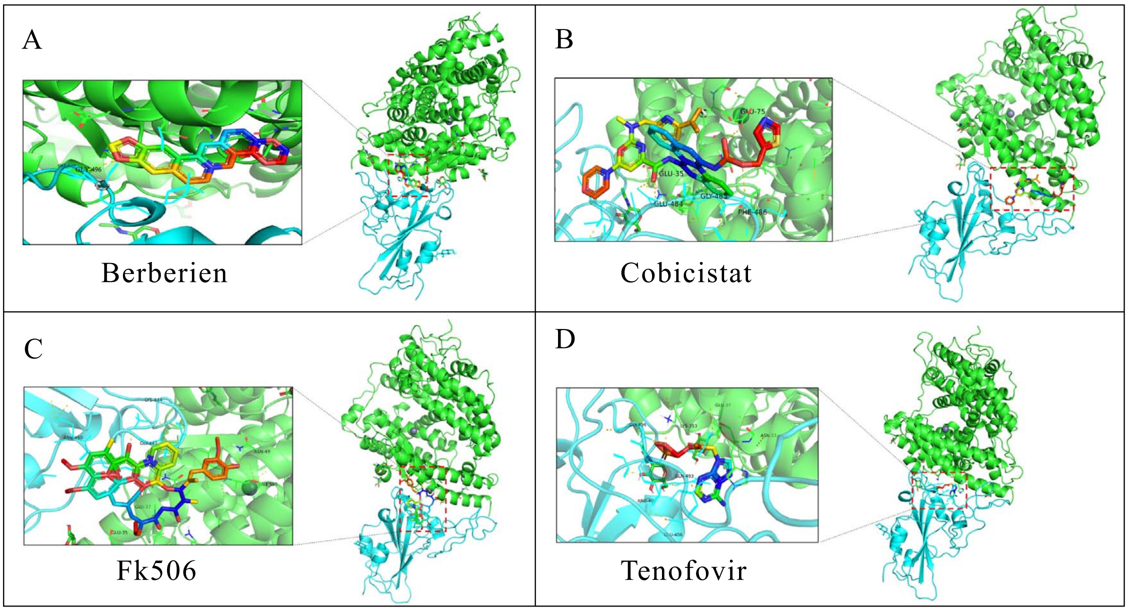

Figure 5. Molecular docking results for Berberine, cobicistat, FK506 and Tenofovir bound with SARS-CoV-2 spike protein/ACE2.

associated with the same drugs. Moreover, as indicated by [38], it is the integration with the biological knowledge of drugs and viruses that offers us an alternative view to improve the quality of VDA networks.

Second, our experimental results reveal that the fundamental reason accounting for the failure of existing network-based models is the lack of such an ability that precisely captures the characteristics of drugs and viruses, which are used to govern VDA networks respectively from the structural and genomic perspectives. Besides, the cold-start problem introduced by new viruses can also be addressed by properly capturing the characteristics of new viruses. In this regard, we develop the CMNMF model modified from traditional NMF. For the CMNMF model, its main purpose is to reconstruct the enhanced association matrix, rather than the original VDA matrix, by using the latent features of drugs and viruses. Moreover, additional constraints are defined from different views, thus ensuring that the similarity information of drugs and viruses are completely projected onto a unified LFS. An extra benefit of doing so is to avoid the noisy information generated after projection, such as the unexpected similarity between drugs and viruses. According to experimental results, the CMNMF model plays a critical role in contributing to the promising performance of CMNMF.

Last, but not least, VDA-DLCMNMF takes advantage of the powerful representation ability of GCN to learn the representations of drugs and viruses from a given VDA matrix. There are three points worth noting: (i) since GCN only accepts the adjacency matrix as input, we could not be able to apply the enhanced association matrix to GCN; (ii) instead of randomly initializing the representation of drugs and viruses, we use their latent feature vectors obtained with CMNMF to complete the initialization task; (iii) to accelerate the training of GCN, we adopt a heuristic neighborhood sampling strategy, which updates the representation of a virus, or a drug, by only using a part of viruses, or drugs, with high quality as indicated by their attention weights; and (iv) VDA-DLCMNMF integrates an attention mechanism into GCN, thus enhancing the information granularity by combining the latent features of drugs and viruses with the topological feature of VDA network.

Though experimental results indicate that VDA-DLCMNMF is a promising tool for repurposing drugs for new viruses, there is still room for further improvement. Specifically, we would like to adapt different solutions, such as variational inference [21, 54] and reparameterization techniques [43], to address the CMNMF model in a more efficient manner. Furthermore, as our future work, we are interested in exploring the possibility of using more biological knowledge, such as intracellular gene regulatory networks [7], drug–drug interactions [69] and drug–disease interactions [57], to construct complex heterogeneous networks and also using higher-order structures [22, 23] to enrich the representations of drugs and viruses.

Key Points

- An enhance association matrix is designed by integrating chemical structures of drugs and the genomic sequences of viruses into a given VDA network, thus strengthening our perception about the potential drugs that new viruses are more likely to associate with.
- We propose a novel CMNMF model to address the cold-start problem related to new viruses by reconstructing the enhance association matrix with the constraints from different views. The similarity information of drugs and viruses can

thus be completely projected onto a unified latent feature space.
- We develop a drug repositioning model, namely VDA-DLCMNMF, to identify potential drugs for new viruses with GCN. The latent feature vectors learned from CMNMF are used as the initial representations of drugs and viruses. VDA-DLCMNMF also adopts an attention-based neighbor sampling strategy to train GCN for drug repurposing.
- Experimental results on three VDA datasets demonstrate the promising performance of VDA-DLCMNMF in repurposing antiviral drugs against SARS-CoV-2. Four novel drugs identified by our method are proved to have the potential ability to bind with important functional receptors of SARS-CoV-2.

Data availability

The dataset and source code can be freely downloaded from https://github.com/Blair1213/DLMNMF.

Author contributions statement

X.S., L.H. and L.W. conceived the experiments; X.S. and L.H. conducted the experiments; Z.Y. and B.Z. analyzed the results.

Acknowledgments

The authors would like to thank colleagues and the anonymous reviewers who have provided valuable feedback to help improve the paper.

Funding

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (2021D01D05), in part by the Pioneer Hundred Talents Program of Chinese Academy of Sciences, in part by the National Natural Science Foundation of China (62172355), in part by the Awardee of the NSFC Excellent Young Scholars Program (61722212), in part by the Science and Technology Innovation 2030-New Generation Artificial Intelligence Major Project (2018AAA0100100) and the Tianshan youth-Excellent Youth (2019Q029).

References

1. AlonsoH, BliznyukAA, GreadyJE. Combining docking and molecular dynamic simulations in drug design. *Med Res Rev* 2006; **26**(5): 531–68.
2. PetterI, AndersenAI, LysvandH, *et al.* Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int J Infect Dis* 2020; **93**:268–76.
3. BermanHM, WestbrookJ, FengZ, *et al.* The protein data bank. *Nucleic Acids Res* 2000; **28**(1): 235–42.
4. BoschBJ, MartinaBEE, Van Der ZeeR, *et al.* (eds). Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides. *Proc Natl Acad Sci USA* 2004; **101**(22): 8455–60.
5. CaneseK, WeisS. PubMed: the bibliographic database. *NCBI Handbook* 2013; **2**:1.
6. Chamoun-EmanuelliAM, PecheurEI, SimeonRL, *et al.* Phenothiazines inhibit hepatitis c virus entry, likely by increasing the fluidity of cholesterol-rich membranes. *Antimicrob Agents Chemother* 2013; **57**(6): 2571–81.
7. ChengJ, ZhangJ, ZhongdaoW, *et al.* Inferring microenvironmental regulation of gene expression from single-cell RNA sequencing data using scMLnet with an application to COVID-19. *Brief Bioinform* 2021; **22**(2): 988–1005.
8. De GrootRJ, BakerSC, BaricRS, *et al.* Commentary: middle east respiratory syndrome coronavirus (MERS-CoV): announcement of the coronavirus study group. *J Virol* 2013; **87**(14): 7790–2.
9. DotoloS, MarabottiA, FacchianoA, *et al.* A review on drug repurposing applicable to COVID-19. *Brief Bioinform* 2020;1–16.
10. FraserC, DonnellyCA, CauchemezS, *et al.* Pandemic potential of a strain of influenza a (h1n1): early findings. *Science* 2009; **324**(5934):1557–61.
11. HaitaoF, HuangF, LiuX, *et al.* MVGCN: data integration through multi-view graph convolutional network for predicting links in biomedical bipartite networks. *Bioinformatics* 2021. https://doi.org/10.1093/bioinformatics/btab651.
12. GaoJ, TianZ, YangX. Breakthrough: chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Biosci Trends* 2020; **14**(1):1–11.
13. GottliebA, SteinGY, RuppinE, *et al.* PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011; **7**(1): 496.
14. GroverA, LeskovecJ. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM Digital Library, 2016, 855–64.
15. GuanW-J, Zheng-yi NiYH, LiangW-H, *et al.* Clinical characteristics of coronavirus disease 2019 in China. *N Engl J Med* 2020; **382**(18): 1708–20.
16. HahnFE, CiakJ. Berberine. In: *Mechanism of Action of Antimicrobial and Antitumor Agents*. Springer, 1975, 577–84.
17. HeT, BaiL, Y-S. Vicinal vertex allocation for matrix factorization in networks. *IEEE Trans Cybernet* 2021;1–14.
18. HeT, LiuY, KoTH, *et al.* Contextual correlation preserving multiview featured graph clustering. *IEEE Trans Cybernet* 2019; **50**(10): 4318–31.
19. HoffmannM, Kleine-WeberH, KrügerN, *et al.* The novel coronavirus 2019 (2019-nCoV) uses the SARS-coronavirus receptor ACE2 and the cellular protease TMPRSS2 for entry into target cells. *BioRxiv* 2020.
20. HoffmannM, Kleine-WeberH, SchroederS, *et al.* SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020; **181**(2): 271–80.
21. LunH, ChanKCC, YuanX, *et al.* A variational Bayesian framework for cluster analysis in a complex network. *IEEE Trans Knowl Data Eng* 2020; **32**(11): 2115–28.
22. HuL, PanX, YanH, *et al.* Exploiting higher-order patterns for community detection in attributed graphs. *Integr Comput-Aided Eng* 2021;**28**:1–12. Preprint.
23. LunH, ZhangJ, PanX, *et al.* HiSCF: leveraging higher-order structures for clustering analysis in biological networks. *Bioinformatics* 2021; **37**(4): 542–50.

24. KatohK, StandleyDM. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013; **30**(4): 772–80.
25. KawaseM, ShiratoK, van der HoekL, *et al.* Simultaneous treatment of human bronchial epithelial cells with serine and cysteine protease inhibitors prevents severe acute respiratory syndrome coronavirus entry. *J Virol* 2012; **86**(12): 6537–45.
26. KearneyBP, FlahertyJF, ShahJ. Tenofovir disoproxil fumarate. *Clin Pharmacokinet* 2004; **43**(9): 595–612.
27. KhaliliJS, ZhuH, MakNSA, *et al.* Novel coronavirus treatment with ribavirin: groundwork for an evaluation concerning COVID-19. *J Med Virol* 2020; **92**(7): 740–6.
28. KipfTN, WellingM. Semi-supervised classification with graph convolutional networks. 2016. Preprint arXiv:1609.02907.
29. LedfordH. Chloroquine hype is derailing the search for coronavirus treatments. *Nature* 2020; **580**(7805):573–4.
30. LepistE-I, PhanTK, RoyA, *et al.* Cobicistat boosts the intestinal absorption of transport substrates, including HIV protease inhibitors and GS-7340, *in vitro*. *Antimicrob Agents Chemother* 2012; **56**(10): 5409–13.
31. LiJ, ZhengS, ChenB, *et al.* A survey of current trends in computational drug repositioning. *Brief Bioinform* 2016; **17**(1): 2–12.
32. LiL, GaoZ, WangY-T, *et al.* SCMFMDA: predicting microRNA-disease associations based on similarity constrained matrix factorization. *PLoS Comput Biol* 2021; **17**(7): e1009165.
33. LiM, ZhangW. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Brief Bioinform* 2021. https://doi.org/10.1093/bib/bbab348.
34. LiW, MooreMJ, VasilievaN, *et al.* Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003; **426**(6965): 450–4.
35. LijunC, LuC, XuJ, *et al.* Drug repositioning based on the heterogeneous information fusion graph convolutional network. *Brief Bioinform* 2021;**22**(6):bbab319.
36. LuoH, LiM, WangS, *et al.* Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018; **34**(11): 1904–12.
37. LuoH, WangJ, LiM, *et al.* Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* 2016; **32**(17): 2664–71.
38. LvH, ShiL, BerkenpasJW, *et al.* Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Brief Bioinform* 2021;**22**(6):bbab320.
39. MalinJJ, SuárezI, PriesnerV, *et al.* Remdesivir against COVID-19 and other viral diseases. *Clin Microbiol Rev* 2020; **34**(1): e00162–20.
40. MengY, JinM, TangX, *et al.* Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. *Appl Soft Comput* 2021; **103**:107135.
41. MorrisGM, HueyR, LindstromW, *et al.* Autodock4 and autodocktools4: automated docking with selective receptor flexibility. *J Comput Chem* 2009; **30**(16): 2785–91.
42. MorrisGM, Lim-WilbyM. Molecular docking. In: *Molecular Modeling of Proteins*. Springer, 2008, 365–82.
43. MostafaH, WangX. Parameter efficient training of deep convolutional neural networks by dynamic sparse reparameterization. In: *International Conference on Machine Learning*. PMLR, 2019, 4646–55.
44. NaydenovaK, MuirKW, WuL-F, *et al.* Structure of the SARS-CoV-2 RNA-dependent RNA polymerase in the presence of favipiravir-RTP. *Proc Natl Acad Sci USA* 2021; **118**(7).
45. O'BoyleNM, BanckM, JamesCA, *et al.* Open Babel: an open chemical toolbox. *J Chem* 2011; **3**(1): 1–14.
46. PangJ, HuangY, XieZ, *et al.* Collaborative city digital twin for the COVID-19 pandemic: a federated learning solution. *Tsinghua Sci Technol* 2021; **26**(5): 759–71.
47. PengL, ShenL, JunlinX, *et al.* Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures. *Sci Rep* 2021; **11**(1): 1–11.
48. PerozziB, Al-RfouR, SkienaS. DeepWalk: online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York USA: ACM Digital Library, 2014, 701–10.
49. PinziL, RastelliG. Molecular docking: shifting paradigms in drug discovery. *Int J Mol Sci* 2019; **20**(18):4331.
50. SayersEW, BeckJ, BoltonEE, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2021; **49**(D1): D10.
51. ScarselliF, GoriM, TsoiAC, *et al.* The graph neural network model. *IEEE Trans Neural Netw* 2008; **20**(1): 61–80.
52. SchreiberSL, CrabtreeGR. The mechanism of action of cyclosporin A and FK506. *Immunol Today* 1992; **13**(4): 136–42.
53. SheahanTP, SimsAC, ZhouS, *et al.* An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat coronavirus. *Biorxiv* 2020.
54. ShenX, YiB, LiuH, *et al.* Deep variational matrix factorization with knowledge embedding for recommendation system. In: *IEEE Transactions on Knowledge and Data Engineering*. IEEE Xplore, 2019.
55. SohrabiC, AlsafiZ, O'neillN, *et al.* World health organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int J Surg* 2020; **76**:71–6.
56. XiaoruiS, YouZ, WangL, *et al.* SANE: a sequence combined attentive network embedding model for COVID-19 drug repositioning. *Appl Soft Comput* 2021;**111**:107831.
57. XiaoruiS, YouZ, YiH. Prediction of LncRNA-disease associations based on network representation learning. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, 1805–12.
58. TangX, CaiL, MengY, *et al.* Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. *Front Immunol* 2021; **11**:3824.
59. Van der MaatenL, HintonG. Visualizing data using t-SNE. *J Mach Learn Res* 2008; **9**(11):2579–605.
60. VidalD, ThormannM, PonsM. Lingo, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model* 2005; **45**(2): 386–93.
61. WangM, CaoR, ZhangL, *et al.* Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. *Cell Res* 2020; **30**(3): 3.
62. WeiningerD. Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988; **28**(1): 31–6.
63. WishartDS, FeunangYD, GuoAC, *et al.* Drugbank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018; **46**(D1): D1074–82.
64. WittineK, BabićMS, MakucD, *et al.* Novel 1, 2, 4-triazole and imidazole derivatives of l-ascorbic and imino-ascorbic acid: synthesis, anti-HCV and antitumor activity evaluations. *Bioorg Med Chem* 2012; **20**(11): 3675–85.

65. YangM, LuoH, LiY, *et al.* Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 2019; **35**(14): i455–63.
66. YuZ, HuangF, ZhaoX, *et al.* Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021; **22**(4): bbaa243.
67. ZengX, ZhuS, WeiqiangL, *et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chem Sci* 2020; **11**(7): 1775–97.
68. ZhangS, TongH, JiejunX, *et al.* Graph convolutional networks: a comprehensive review. *Comput Social Netw* 2019; **6**(1): 1–23.
69. ZhangT, LengJ, LiuY. Deep learning for drug–drug interaction extraction from the literature: a review. *Brief Bioinform* 2020; **21**(5): 1609–27.
70. ZhouL, WangJ, LiuG, *et al.* Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method. *Genomics* 2020; **112**(6): 4427–34.
71. ZhouY, VedanthamP, LuK, *et al.* Protease inhibitors targeting coronavirus and filovirus entry. *Antiviral Res* 2015; **116**:76–84.
72. ZhuN, ZhangD, WangW, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727–33.