AntiViralDL: Computational Antiviral Drug Repurposing Using Graph Neural Network and Self-Supervised Learning

Pan Zhang, Xiaowen Hu, Guangdi Li, and Lei Deng

***Abstract*—Viral infections have emerged as significant public health concerns for decades. Antiviral drugs, specifically designed to combat these infections, have the potential to reduce the disease burden substantially. However, traditional drug development methods, based on biological experiments, are resource-intensive, time-consuming, and low efficiency. Therefore, computational approaches for identifying antiviral drugs can enhance drug development efficiency. In this study, we introduce AntiViralDL, a computational framework for predicting virus-drug associations using self-supervised learning. Initially, we construct a reliable virus-drug association dataset by integrating the existing Drugvirus2 database and FDA-approved virus-drug associations. Utilizing these two datasets, we create a virus-drug association bipartite graph and employ the Light Graph Convolutional Network (LightGCN) to learn embedding representations of viruses and drugs. To address the sparsity of virus-drug association pairs, AntiViralDL incorporates contrastive learning to improve prediction accuracy. We implement data augmentation by adding random noise to the embedding representation space of virus and drug nodes, as opposed to traditional edge and node dropout. Finally, we calculate an inner product to predict virus-drug association relationships. Experimental results reveal that AntiViralDL achieves AUC and AUPR values of 0.8450 and 0.8494, respectively, outperforming four benchmarked virus-drug association prediction models. The case study further highlights the efficacy of AntiViralDL in predicting anti-COVID-19 drug candidates.**

***Index Terms*—Bipartite graph, contrastive learning, graph convolutional network, self-supervised learning, virus-drug associations.**

I. Introduction

Viral infections are causing an increasing global disease burden, especially acute viral infections caused by certain viruses with strong transmissibility, which much more likely lead to sudden outbreaks [1]. For instance, Coronaviruses (CoVs), members of the Coronaviridae family, are recognized as potentially lethal viruses due to their occasional causation of severe respiratory tract infections in humans and other mammals [2], [3]. Over the past two decades, the highly contagious and rapidly spreading nature of coronaviruses has incited public panic on three separate occasions [4], [5], [6]. The recent COVID-19 outbreak, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus, has led to millions of deaths and substantial economic impact worldwide [7]. This ongoing crisis, which has persisted for over three years, continues to attract widespread attention [8], [9], [10]. Although several drugs have been approved or granted emergency use authorizations (EUAs) by the US Food and Drug Administration (FDA), their effectiveness remains uncertain due to the emergence of new SARS-CoV-2 variants, such as Delta and Omicron [11], [12]. Consequently, the development of effective antiviral drugs via drug repurposing against the constantly evolving viruses or newly emerging pathogenic viruses remains a critical task.

The development of clinical drugs encompasses three stages: drug discovery, preclinical, and clinical research, causing the entire process from drug discovery to approval to be quite lengthy [13]. In the face of pandemics like COVID-19, it is crucial to identify available drugs for emergency use as swiftly as possible [14]. However, traditional new drug discovery, which relies on wet experiments, can be costly, as extensive testing is required to determine the effectiveness of antiviral drugs [15]. Moreover, antiviral drug screening experiments necessitate strict experimental conditions, such as Biosafety Level 3 (BSL-3) laboratories, and each step of the process–including cell culture, virus infection, or drug treatment–must be rigorously controlled and monitored [16], [17]. Given these factors, drug repurposing

Manuscript received 9 May 2023; revised 27 August 2023, 20 September 2023, and 7 October 2023; accepted 21 October 2023. Date of publication 3 November 2023; date of current version 5 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 92269112, 62272490, and 61972422 and in part by the Scientific Research Program of FuRong Laboratory under Grant 2023SK2090. *(Corresponding authors: Guangdi Li; Lei Deng.)*

This work did not involve human subjects or animals in its research.

Pan Zhang is with the Hunan Provincial Key Laboratory of Clinical Epidemiology, Xiangya School of Public Health, Central South University, Changsha 410083, China (e-mail: paulazhang@csu.edu.cn).

Xiaowen Hu is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: altriavin@163.com).

Guangdi Li is with the Hunan Provincial Key Laboratory of Clinical Epidemiology, Xiangya School of Public Health, Central South University, Changsha 410083, China, and also with the Furong Laboratory, Central South University, Changsha 410083, China (e-mail: liguangdi@csu.edu.cn).

Lei Deng is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: leideng@csu.edu.cn).

The programming code and datasets of AntiViralDL are available at: https://github.com/paulacode/AntiViralDL.

Digital Object Identifier 10.1109/JBHI.2023.3328337

through the identification of virus-drug associations is undoubtedly a practical option for screening candidate drugs for further validation, ultimately accelerating drug development and application [18].

Recently, with the widespread application of recommendation algorithms in the field of biomedical research [19], [20],numerous computational drug repurposing methods have been developed for predicting associations between drugs and diseases [21], [22], [23]. There are also certain approaches that mainly focus on the association between antiviral drugs and diseases caused by viral infections [24], [25], [26], [27]. These methods can be broadly categorized into two groups [28]. The first group consists of network-based approaches that aim to discover new drug-disease relationships using graphs and integrated networks. Luo et al. [29] proposed a drug repositioning recommendation system (DRRS) that integrates drug and disease similarities, as well as known drug-disease associations, based on the Singular Value Thresholding (SVT) algorithm to explore novel clinical indications for existing drugs. Su et al. [24] presented a deep learning method utilizing a novel constrained multi-view nonnegative matrix factorization (CMNMF) model, which preserves intrinsic similarity information of drugs and viruses in a low-rank latent feature space (LFS) to identify potential drugs for new viral infections. Zhou et al. [25] proposed a virus-drug association (VDA) prediction using a heterogeneous network based on the KATZ model, which transforms the prediction of novel VDAs into measuring the number of direct neighbor nodes and all other nodes connected to those direct neighbors in the network. The second group comprises learning-based approaches that utilize learning models with feature vectors and extract patterns from existing databases or natural language text. Deepthi et al. [26] developed a deep learning ensemble approach, DLEVDA, by integrating drug chemical structure similarities and virus genome sequence information into a convolutional neural network, and then using an extreme gradient boosting classifier to recommend potential anti-SARS-CoV-2 drugs. Wang et al. [30] constructed a deep learning method employing a directed message-passing neural network to screen broad-spectrum antiviral drugs against coronaviruses and then used transfer learning to fine-tune the initial model for identifying specific drugs targeting SARS-CoV-2.

Although deep learning methods have been widely applied in drug repurposing, and the COVID-19 outbreak has prompted researchers to recognize the importance of drug repurposing for antiviral drugs, both network-based and structure-based computational methods currently exhibit certain limitations in the field of virus-drug association prediction. First, they overlook crucial datasets of approved virus-drug associations, which may result in biased prediction outcomes. Second, most models tackle supervised learning-based association prediction tasks, with supervised signals originating from observed virus-drug associations. However, these observed associations are often sparse, potentially increasing the risk of overfitting in the model. Consequently, there is a need to develop a more concise and high-performance virus-drug association prediction model.

In this study, we propose a novel graph contrastive learning method, AntiViralDL, to predict potential associations between viruses and antiviral drugs, with the goal of enhancing the efficiency and speed of discovering and developing new antiviral drugs while providing guidance for their development. Initially, we construct a virus-drug bipartite graph based on the associations between viruses and drugs and employ a lightweight graph convolutional network to learn the embedding representations of viruses and drugs through neighbor node aggregation. Given the sparsity of virus-drug associations, we incorporate contrastive learning to improve the model's prediction accuracy. Traditional graph contrastive learning methods perform data augmentation on the graph structure, such as node random dropout, edge random dropout, and random walk. However, these data augmentation methods may result in the loss of critical information, leading to unreliable prediction structures. Furthermore, data augmentation on the graph structure can be complex. Therefore, we innovatively add random noise to the embedding representation space of virus and drug nodes to complete the contrastive learning task. Ultimately, the inner product is used to predict potential associations between viruses and drugs. To ensure the reliability of our experimental results, we use evidence from in vitro experiments and clinical trials to identify virus-drug associations. To evaluate the performance of our proposed model, we conduct five-fold cross-validation. The comparative experimental results indicate that AntiViralDL outperforms four association prediction benchmarked models. Case analysis demonstrates the effectiveness of AntiViralDL, which has the potential to be a powerful tool for clinical and biological research.

TABLE I
STATISTICS FOR EACH VIRUS-DRUG ASSOCIATION DATASET

Data set	Viruses	Drugs	Associations
DrugVirus2	153	231	1519
US FDA	16	111	142
Merged data	158	336	1648

II. MATERIALS AND METHODS

A. Data Set

We extracted two datasets for known virus-drug associations from different sources: the DrugVirus.info 2 database [31] and the US FDA (https://www.fda.gov/). The DrugVirus.info 2 database was considered the primary dataset, containing 1,519 virus-drug associations between 231 drugs and 153 viruses. We manually collected current FDA-approved drugs for the treatment of viral infectious diseases up until October 2022, constituting a new dataset with 142 virus-drug associations between 111 drugs and 16 viruses. After removing all redundant records with the same virus and drug, we obtained a merged dataset comprising 1,648 virus-drug associations between 158 viruses and 336 drugs. The detailed statistics for the three virus-drug association datasets mentioned above are listed in Table I.

B. Virus-Drug Bipartite Graph

After obtaining a virus-drug association database, we needed to model the associations between viruses and drugs by constructing a virus-drug bipartite graph. First, based on the virus-drug association pairs, we constructed an association matrix $A \in$

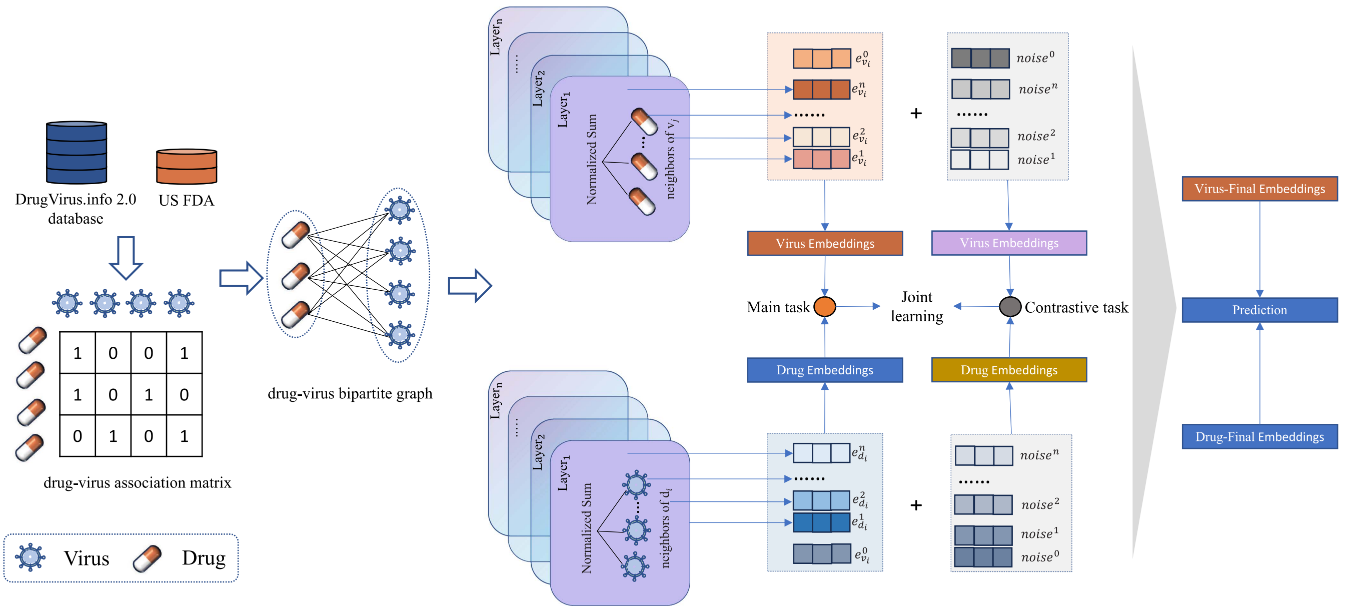

Fig. 1. Workflow of AntiViralDL. We first collect known associations between viruses and drugs from two different databases, namely the DrugVirus.info 2 database and the FDA, and then construct a bipartite graph of viruses and drugs based on viruses-drug association pair. AntiViralDL consists of two tasks: the main task and the contrastive task. In the main task, we input the bipartite graph into the LightGCN framework to learn representations of virus and drug nodes. Subsequently, we calculate the inner product to predict virus-drug associations. Due to the sparse supervised signal in the dataset, we further enhance the efficiency of learned virus and drug representations by designing a secondary contrastive task for self-supervised learning. In the contrastive task, we augment the data by adding random noise to the embedding representations of virus and drug nodes, and build the contrastive learning paradigm. Finally, we integrate the main and contrastive tasks to predict associations between viruses and drugs better.

$R^{|V|\times|D|}$, where V and D represent the sets of viruses and drugs, respectively. If there was an association between virus V_i and drug D_j, then $A_{i,j} = 1$; otherwise, $A_{i,j} = 0$. After obtaining the virus-drug association matrix, the virus-drug bipartite graph can be represented as $G(V, D, E)$, where $E = \{y_{vd} | v \in V, d \in D\}$ represents the verified associations between viruses and drugs.

C. Antiviraldl

In this work, we propose a new method based on self-supervised learning and graph convolutional networks called AntiViralDL to predict potential associations between viruses and drugs. AntiViralDL takes the virus-drug bipartite graph as input and outputs the association scores between specific viruses and drugs. First, we randomly initialize the embeddings of viruses and drugs. Subsequently, LightGCN was employed to aggregate feature representations of neighboring nodes on the virus-drug bipartite graph. LightGCN is a graph neural network algorithm designed for recommendation systems, with a focus on handling virus-drug associations data [32]. The fundamental idea of LightGCN involves simplifying graph convolution operations into a lightweight form of neighborhood propagation.Afterwards, we introduced random noise into the embeddings of virus and drug nodes at each layer of LightGCN to achieve efficient data augmentation. Following this, we constructed a contrastive learning paradigm to enhance the model's representational capacity. Finally, the inner product is used to calculate the association scores between specific viruses and drugs. The entire workflow of AntiViralDL is shown in Fig. 1.

1) Initialize Embedding: In this section, we need to initialize the ID embedding matrix corresponding to viruses and drugs, respectively, and use embedding lookup table to map them to their corresponding feature space through the IDs of viruses and drugs. In mathematical terms, the ID embedding matrix for viruses and drugs can be abstracted as:

$$\begin{cases} E_v = [e_{v_1}, e_{v_2}, \ldots, e_{v_m}] \\ E_d = [e_{d_1}, e_{d_2}, \ldots, e_{d_n}] \end{cases} \tag{1}$$

where E_v and E_d were the feature matrices for virus and drug respectively, $e_{v_i} \in R^T$ and $e_{d_j} \in R^T$ were the feature vectors for the i-th virus and the j-th drug respectively, and T represented the dimensionality of the feature vectors.

2) Graph Convolutional Network for AntiViralDL: Graph convolutional networks(GCN) have gained widespread application in the field of association prediction [33], [34], [35]. Traditional GCN methods consist of three crucial components: feature transformation, non-linear activation, and neighbor node aggregation. He et al. [32] experimentally verified that neighbor node aggregation plays a primary role among these three components. Building upon this insight, we solely utilize neighbor node aggregation to update the information of nodes in the graph as follows:

$$\begin{cases} e_v^{(l+1)} = \sum_{d \in N_v} \frac{1}{\sqrt{|N_v|}\sqrt{|N_d|}} e_v^{(l)} \\ e_d^{(l+1)} = \sum_{v \in N_d} \frac{1}{\sqrt{|N_d|}\sqrt{|N_v|}} e_d^{(l)} \end{cases} \tag{2}$$

where e_v^l and e_d^l denote the embeddings of virus v and drug d at layer l, respectively. N_v and N_d represent the sets of first-hop neighboring nodes for virus v and drug d, and $|N_v|$ and $|N_d|$ indicate the number of neighboring nodes for virus v and drug d, respectively. The term $\frac{1}{\sqrt{|N_v||N_d|}}$ serves as the graph laplacian norm.

Finally, the feature representation obtained by each layer of LightGCN is summed to obtain the embedding of the final node as follows:

$$\begin{cases} e_v = \sum_{l=0}^{L} e_v^{(l)} \\ e_d = \sum_{l=0}^{L} e_d^{(l)} \end{cases} \tag{3}$$

where L represents the number of layers of LightGCN.

3) Contrastive Learning:

Traditional graph-based data augmentation methods mainly include methods based on node dropout and edge dropout [36]. However, these methods might lose critical node or edge information due to changes in the graph structure, resulting in poor robustness of the model. Moreover, learning uniform feature representations for nodes by directly manipulating the graph structure often proves to be tricky and requires significant computational costs. Inspired by [37], we focus on data augmentation in the embedding space of viruses and drugs. Formally, for each given feature representation of a node, we achieve efficient embedding-level data augmentation by directly adding different noise to the feature representation as follows:

$$\begin{cases} e_i' = e_i + \Delta_i' \\ e_i'' = e_i + \Delta_i'' \end{cases} \tag{4}$$

where these noise vectors Δ_i' and Δ_i'' were the subject to $\|\Delta\|_2 = \varepsilon$ and $\Delta = \bar{\Delta} \odot sign(e_i)$, $\Delta' \in R^T \sim U(0,1)$. $\|\Delta\|_2 = \varepsilon$ represent that the granularity of the noise is equivalent to the vector on the hypersphere with ε as the radius. $\Delta = \bar{\Delta} \odot sign(e_i)$, $\Delta' \in R^T \sim U(0,1)$ indicates that the noise vector and the original representation are located in the same super quadrant to avoid excessive semantic deviation caused by adding noise, $sign(\cdot)$ and $U(0,1)$ represent transition function and uniform distribution respectively.

In the contrastive learning task, we need to minimize the feature representation obtained by the same embedding augmentation, and maximize the feature representation of different embeddings and their added noise. We use the InfoNCE loss [38] for the contrastive learning auxiliary task:

$$L_{cl} = \sum_{i \in B} -log \frac{\exp\left({e_i'}^T e_i''/\tau\right)}{\sum_{j \in B} \exp\left({e_i'}^T e_j''/\tau\right)} \tag{5}$$

Among them, i and j are the viruses and drugs sampled from the training batch, and e_i', e_i'' and e_j'' are the feature representations of virus i and drug j after adding random noise, respectively. τ is the temperature. It can be observed that InfoNCE is mainly designed to reduce the distance between e_i' and e_i'', as they are generated from data augmentation of a feature vector, which belongs to positive samples, while increasing the distance between e_i' and e_j'', as they come from data augmentation between different feature vectors, which belong to negative samples.

D. Prediction

Finally, when we obtain the final embeddings of the virus and the drug, the inner product is used to calculate the preference score y_{ij} between virus v_i and drug d_j as follows:

$$\hat{y}_{ij} = e_{v_i}^T e_{d_j} \tag{6}$$

E. Optimization

In the optimization part, we used BPR loss [39] for the main task and InfoNCE loss [38] for the contrastive learning auxiliary task. Formally, the BPR loss can be abstracted as:

$$L_{BPR} = \sum_{(v,d_+,d_-) \in B} -log\sigma(\hat{y}_{vd_+} - \hat{y}_{vd_-}) \tag{7}$$

where (v, d_+) represented a positive sample, (v, d_-) represented a negative sample and σ represented a nonlinear activation function, B represented the set of associated pairs. For each virus v, we randomly sample an equal number of negative samples related to it as there are positive samples. As a result, the ratio of positive to negative samples in the associated set B is 1:1. Finally, we can use join learning to obtain the total loss as follows:

$$L = L_{BPR} + \lambda L_{cl} \tag{8}$$

where λ was a hyperparameter for balancing L_{BPR} and L_{cl}.

III. Results

A. Experimental Setup

To effectively reduce the possibility of accidental prediction results, in our experiment, we used five-fold cross-validation to evaluate the performance of the AntiViralDL model, in which the whole dataset of virus-drug associations was randomly divided into five subsets. Each time, four subsets among them were taken as the training sets to train the model while the other one subset was used as the test set to make the prediction. This cross-validation process did not finish until each subset had been used as the test set, and then the average value of five-fold cross-validation was regarded as the final result of the AntiViralDL model. We computed the AUC (area under the receiver operating characteristic curve) and the AUPR (area under the accurate recall curve) as primary evaluation indicators as both AUC and AUPR were based on the comparison between model prediction results and real labels, which could reflect the model's capacity to distinguish between positive and negative samples.

Several hyperparameters were set in AntiViralDL: the number of training epochs was 1000, and the learning rate was 0.01. Moreover, the number of graph convolution layers was chosen from the ranges [1, 2, 3, 4], the embedding dimension was chosen from the ranges [16, 32, 64, 128, 256], and the values of lambda was chosen from the ranges [0.1, 0.3, 0.5, 0.7, 0.9]. In AntiViralDL, we implemented all the experiments based on the open-source deep learning framework TensorFlow 1.14.0 and TensorFlow 2.4.

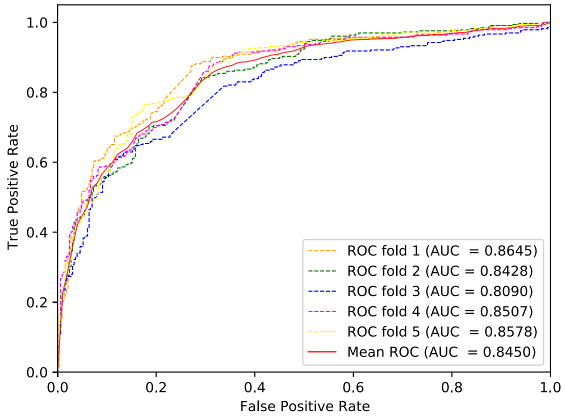

Fig. 2. Five-fold cross-validation ROC curves of AntiViralDL on the Merged Dataset.

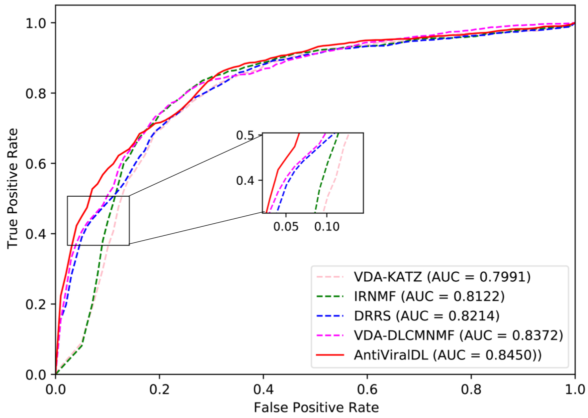

Fig. 3. ROC curves of AntiViralDL compared with four related models for virus-drug association.

TABLE II
FIVE-FOLD CROSS-VALIDATION RESULTS GENERATED BY ANTIVIRALDL ON THE MERGED DATASET

Validation set	AUC	AUPR
1	0.8645	0.8595
2	0.8428	0.8469
3	0.8090	0.8237
4	0.8507	0.8602
5	0.8578	0.8570
Average	0.8450	0.8494

TABLE III
PERFORMANCE COMPARISON OF ANTIVIRALDL AND FOUR BENCHMARKED METHODS ON THE MERGED DATASET

Methods	AUC	AUPR
VDA-KATZ	0.7991	0.7496
IRNMF	0.8122	0.7610
DRRS	0.8214	0.8172
VDA-DLCMNMF	0.8372	0.8318
AntiViralDL	0.8450	0.8494

B. Five-Fold Cross-Validation Performance

In this study, to scientifically evaluate the performance of the proposed model, we conducted five-fold cross-validation as mentioned earlier. We plotted the ROC curve of the model based on the five-fold cross-validation and calculated their AUC values separately. The ROC curves generated from the five experiments are shown in Fig. 2. The detailed results of the five-fold cross-validation are shown in Table II. From Table II, the results show that the AUC values of the five validations are 0.8645, 0.8428, 0.8090, 0.8507, and 0.8578, with an average of 0.8450. The AUPR values are 0.8595, 0.8469, 0.8237, 0.8602, and 0.8570, with an average of 0.8494. Based on the experimental results, it can be concluded that this model is able to effectively predict virus-related drugs with good overall performance. This also provides evidence that this model could be a promising tool for exploring virus-drug associations.

C. Comparison of Models

In order to further identify the performance of the AntiViralDL model, we compared it with four state-of-the-art methods to predict drug-disease associations, including VDA-KATZ [25], IRNMF [40], DRRS [29], and VDA-DLCMNMF [24].

- VDA-KATZ [25] model was a deep learning model for binary classification, which was based on KATZ algorithm to calculate the connection paths between different viruses and drugs in a large-scale heterogeneous network, so that it contributed to recommend potential drugs against SARS-CoV-2.
- IRNMF [40] was a virus-drug association network based on the indicator regularized non-negative matrix factorization (IRNMF) method through the combination of the indicator matrix and the Karush-Kuhn-Tucker condition to predict the potential drugs for the treatment of COVID-19.
- DRRS [29] used the Singular Value Thresholding (SVT) algorithm to construct a recommendation system that could handle large heterogeneous network based on drug similarity, disease similarity and drug-disease association.
- VDA-DLCMNMF [24] constructed a constrained multi-view nonnegative matrix factorization (CMNMF) model, which was used to process multi-modal data. The model combined the advantages of deep learning and non-negative matrix decomposition, and integrated the drug chemical structure and viral genome sequence into the VDA matrix to obtain an enhanced association matrix, which could effectively extract the feature information from the multi-modal data.

Similarly, for each method, we conduct five-fold cross-validation and take the average value as the final result. For an intuitive comparison on the performance of all five models, we respectively drawn the ROC curves in Fig. 3. We noted that our method outperformed above four approaches related to drug repositioning on both metrics. As shown in Table VI, The AUC and AUPR of the AntiViralDL model on the merged dataset were 0.8450 and 0.8494, respectively. VDA-DLCMNMF was ranked second in performance among the four benchmark methods, with AUC and AUPRC values of 0.8372 and 0.8318, both of which were lower than AntiViralDL. In addition, the other three models, VDA-KATZ, IRNMF, and DRRS, all exhibited

TABLE IV
EXPERIMENTAL RESULTS AT DIFFERENT POSITIVE AND NEGATIVE SAMPLE RATIOS

Methods	AUC		
	1:1	**1:5**	**1:10**
VDA-KATZ	0.7991	0.7954	0.7911
IRNMF	0.8122	0.8110	0.8103
DRRS	0.8214	0.8207	0.8189
VDA-DLCMNMF	0.8372	0.8354	0.8321
AntiViralDL	0.8450	0.8418	0.8409

TABLE V
EXPERIMENTAL RESULTS OF ABLATION STUDY

		AUC	**AUPR**
Data Augmentation	**SGL-None**	0.8110	0.8048
	SGL-ED	0.8272	0.8253
	SGL-ND	0.8204	0.8116
	SGL-RW	0.8126	0.8163
	AntiViralDL	0.8450	0.8494
Feature Composition	**concatenation**	0.5920	0.6863
	element-wise max	0.8011	0.7748
	element-wise min	0.8218	0.8040
	sum(AntiViralDL)	0.8450	0.8494
difficult noise	**random noise**	0.8158	0.8064
	similar noise(AntiViralDL)	0.8450	0.8494

lower AUC values 0.7991, 0.8122, 0.8214 and AUPR values 0.7496, 0.7610, 0.8172. Overall, the AntiViralDL model performed better than the other four models on the merged dataset.

In addition, we verified the performance of AntiViralDL with different positive-to-negative sample ratios of 1:1, 1:5, and 1:10, as shown in Table IV. From Table IV, it can be observed that AntiViralDL achieves optimal performance even with varying positive-to-negative sample ratios. The AUC values for the 1:1, 1:5, and 1:10 ratios are 0.8450, 0.8418, and 0.8409, respectively, which are higher than the second-best methods by 7.8%, 6.4%, and 8.8%, respectively.

The excellent predictive performance of AntiViralDL may be attributed to the following reasons. Firstly, AntiViralDL models virus-drug associations as a graph and uses graph neural networks to explore the high-order connectivity of the graph. Secondly, existing methods model virus-drug association prediction as supervised learning, with the supervision signal coming from known virus-drug associations. However, the limited number of known virus-drug associations leads to insufficient learning of the model. AntiViralDL uses self-supervised learning techniques to alleviate the impact of sparse supervision signals. In addition, most existing methods are based on virus similarity and drug similarity, and the computed similarity often contains inevitable noise due to the lack of additional biological data, resulting in inaccurate representations of virus and drug. AntiViralDL does not require similarity information and could predict unknown virus-drug associations solely based on the virus-drug association matrix.

D. Ablation Study

In order to investigate the respective contributions of different components within the model, we conducted three types of ablation experiments, encompassing data augmentation, feature combination, and noise variation, respectively.

1) Analysis of Different Data Augmentation Methods: To further provide evidence for the advantages of AntiViralDL based on self-supervised learning in antiviral drug repurposing, we conduct an ablation experiment. We first design four variants of AntiViralDL and evaluate their performance on the merged dataset. The detailed information about these four variant models are presented below.

- SGL-None [32] is a simplified GCN model with no data augmentation.
- SGL-ND [41] denotes Node Dropout as the data augmentation technique for contrastive learning.
- SGL-ED [41] denotes Edge Dropout as the data augmentation technique for contrastive learning.
- SGL-RW [41] denotes Random Walk as the data augmentation technique for contrastive learning.

AntiViralDL and the other four variant models predict virus-drug associations based on the same dataset, the average values under 5-fold cross-validation of the three models are listed in Table V. Overall, AntiViralDL outperforms SGL-None, SGL-ED, SGL-ND and SGL-RW in both evaluation metrics AUC and AUPR. Compared with SGL-None, the satisfactory performance of AntiViralDL is attributed to the advantages of self-supervised learning in the alleviation of sparse supervision signals. The performance of SGL based on traditional graph data augmentation is lower than AntiViralDL, indicating that adding random noise to the representation is a simpler and more efficient way to extract the essential information of the original graph.

2) Analysis of Feature Composition: The ultimate embedding of viruses and diseases is obtained by composing the features from each layer of graph convolution. In order to investigate the impact of different methods of feature composition on the performance of the model, we validated the effects of feature summation, feature concatenation, element-wise maximum, and element-wise minimum on the model performance, as shown in Table V. From Table V, it can be observed that summing the features from each layer of the graph convolution network achieves the best model performance, with AUC of 0.8450 and AUPR of 0.8494, both higher than the second-best method by 2.32% and 4.54%, respectively.

3) The Influence of Different Noises: Compared to traditional graph-based contrastive learning methods, CL innovatively introduces noise similar to node embeddings into the node embeddings to achieve data augmentation. In order to explore the impact of different noises (similar noises to node embeddings and completely random noises) on model performance, corresponding experiments were conducted, and the results are shown in Table V. From Table V, it can be observed that adding noise similar to node embeddings plays a significant role in improving model performance. We speculate that this is because similar noise can achieve positive sample alignment and make the learned embeddings uniformly distributed on the same hypersphere, which are the two main reasons why contrastive learning can learn good node feature representations from limited samples [42].

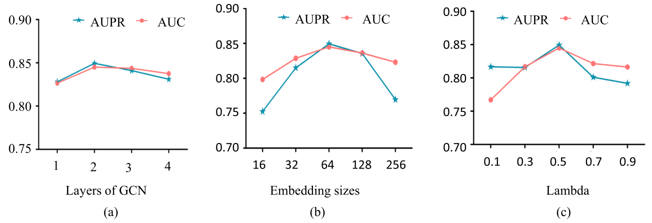

Fig. 4. Influence of different parameters (a) effect of different GCN layers,(b) effect of different Embedding Sizes,(c) effect of lambda.

TABLE VI
EXPERIMENTAL RESULTS OF PARAMETER SENSITIVITY ANALYSIS

		AUC	AUPR
layer	**1**	0.8265	0.8280
	2	0.8450	0.8494
	3	0.8437	0.8410
	4	0.8376	0.8311
embedding sizes	**16**	0.7984	0.7525
	32	0.8288	0.8151
	64	0.8450	0.8494
	128	0.8363	0.8355
	256	0.8231	0.7696
lambda	**0.1**	0.7670	0.8168
	0.3	0.8168	0.8158
	0.5	0.8450	0.8494
	0.7	0.8216	0.8010
	0.9	0.8164	0.7919

E. Parameter Sensitivity Analysis

To ensure the scientificity and accuracy of the AntiViralDL model in this study, we evaluated the impact of two parameters used to construct the model, including GCN layers and embedding size, on its performance. Alternatively, we calculated the AUC and AUPR values separately for different GCN layers or embedding sizes while keeping other conditions constant.

Firstly, the influence of GCN layers was evaluated. In this study, with the remaining parameters unchanged, GCN layers were replaced sequentially from 1 to 4, and the average results of 5-fold cross-validation were used for evaluation. The AUC and AUPR values are shown in Table VI. It was found that when the GCN layer was 2, the model had the highest average values of AUC and AUPR, indicating the best performance in Fig. 4(a). The optimal number of GCN layers should be related to the sparsity of the neighbor matrix used to construct the graph. Too few layers may not fully explore the high-order connectivity from the dataset, while too many layers may lead to over-smoothing, resulting in the representation vectors becoming indistinguishable.

Secondly, the effect of embedding size was evaluated. Similarly, keeping other parameters unaltered, embedding sizes from [16, 32, 64, 128, 256] were selected orderly, and the average values of AUC and AUPR from 5-fold cross-validation were shown in Table VI. The experimental results showed that the model performed best when the embedding size was 64 in Fig. 4(b), and both too high or too low embedding sizes led to a decrease in model performance. An excessively large embedding size could lead to overfitting as the model may overfit the noise in the training data.

TABLE VII
TOP 10 DRUGS ASSOCIATED WITH COVID-19 PREDICTED BY ANTIVIRALDL

Rank	Drug name	Status	Score	Evidence (PubMed ID or number of clinical trial)
1	Baloxavir marboxil	Clinical trial	0.99	33115675
2	Laninamivir octanoate	-	0.96	unconfrmed
3	Arbidol	Phase 4	0.96	32373347 (NCT04260594)
4	Peramivir	-	0.96	unconfrmed
5	Zanamivir	-	0.94	unconfrmed
6	Oseltamivir	Phase 3	0.93	35531426 (NCT04338698)
7	Rimantadine	In vitro	0.91	34696509
8	Saliphenylhalamide	-	0.90	unconfrmed
9	Idoxuridine	-	0.82	unconfrmed
10	Regorafenib	Clinical trial	0.72	NCT05594147

Third, the effect of lambda was evaluated. Similarly, keeping other parameters unaltered, lambda from [0.1, 0.3, 0.5, 0.7, 0.9] were selected orderly, and the average values of AUC and AUPR from 5-fold cross-validation were shown in Table VI. The experimental results showed that the model performed best when the lambda was 0.5 in Fig. 4(c), The selection of an appropriate value for lambda is crucial for enhancing model performance and ensuring model balance. Both excessively high and low values of lambda are not conducive to model improvement. In the case of a low lambda, it can result in a decrease in AUC, while a high lambda can lead to a decrease in AUPR. It is important to find a balanced lambda that promotes model performance and stability by considering the contrast loss between the main task and the auxiliary task in comparison learning. Such a balance is beneficial for the overall model performance.

F. Case Study

Repurposing existing antiviral drugs to treat viral infections beyond their approved indications is a highly promising trend in drug development, as it can significantly reduce the time and cost of developing new drugs [43]. The COVID-19 pandemic, a global concern for nearly three years, has seen new variants with reduced virulence and self-limiting infections. However, approved drugs still exhibit limitations for vulnerable populations, such as the elderly and children. Consequently, it remains crucial to pursue drug repurposing for COVID-19 based on extensive known virus-antiviral drug associations. In this article, we use COVID-19 as a case study.

To thoroughly assess AntiViralDL's applicability, we trained the model to predict potential drugs related to COVID-19. We first removed known COVID-19 drugs and input the remaining ones into the trained AntiViralDL for prediction. Then, we ranked the predicted drugs in descending order based on the calculated correlation score, selecting the top 10 candidate drugs as our predicted COVID-19-related drugs. We validated our predictions through existing literature mining (PubMed ID) and querying clinical trials of the drugs (https://clinicaltrials.gov/).

In Table VII, we list several antivirals with detailed descriptions. Baloxavir marboxil, a novel anti-influenza drug approved by the FDA in 2018, inhibits the influenza virus polymerase complex, preventing virus replication and transmission. It acts more quickly and lasts longer than traditional anti-influenza drugs such as oseltamivir and zanamivir [44]. Lou et al. [45] demonstrated Baloxavir marboxil's significant in vitro antiviral activity against SARS-CoV-2, although clinical benefits remain unproven. Laninamivir octanoate and Peramivir, anti-influenza drugs, inhibit influenza neuraminidase to prevent virus replication and spread. Arbidol, primarily used to treat influenza and other respiratory viral infections, binds to viral membrane fusion proteins, inhibiting virus entry into host cells. Proven to be a SARS-CoV-2 inhibitor, Arbidol has demonstrated efficacy in vitro and clinical improvements in vivo [46]. Oseltamivir, a classic antiviral drug for influenza, inhibits the neuraminidase of influenza viruses, affecting both type A and type B strains. While there is limited evidence suggesting a direct therapeutic effect on SARS-CoV-2, Zendehdel et al. [47] proposed that Oseltamivir may have an adjuvant effect on COVID-19 treatment. The study found that patients using Oseltamivir experienced shorter illness durations and lower mortality rates compared to those who did not. Interestingly, some top-ranked drugs have not yet been confirmed to be connected with COVID-19, warranting further investigation.

IV. Conclusion

In this study, we developed a computational method called AntiViralDL, based on self-supervised learning, to predict virus-drug associations. AntiViralDL effectively mitigates the impact of sparse supervised signals in such data and is suitable for large-scale virus-drug association predictions. We constructed a virus-drug association bipartite graph by integrating the existing Drugvirus 2 database and FDA-approved virus-drug associations dataset, ensuring the validity of virus-drug associations. Additionally, we employed LightGCN to learn the embedding representations of viruses and drugs. Moreover, AntiViralDL incorporated a contrastive learning approach to enhance the model's predictive accuracy due to the sparsity of virus-drug associations. Traditional data augmentation methods, such as node dropping, edge perturbation, and random walk, can result in the loss of vital information, leading to unreliable prediction outcomes. To address this issue, we performed data augmentation by adding random noise to the embedding representations of viruses and drugs. We then predicted virus-drug associations using the inner product method. Comparative experimental results demonstrated that AntiViralDL achieved an AUC of 0.8450, outperforming four other benchmarked virus-drug association prediction models, which indicates its exceptional performance in this domain.

AntiViralDL analyzes large-scale virus and drug data validated through biological experiments. By mining the potential associations between viruses and drugs, it aids in identifying drugs originally intended for a specific virus that could potentially exhibit therapeutic effects in other viral diseases, facilitating drug repurposing. The most promising drug candidates for biological experiments or clinical trials are recommended, promoting the rapid screening of targeted antiviral drugs and facilitating enhanced clinical decision-making.

In the future work, we plan to extract multi-source heterogeneous features of viruses and antiviral drugs to enhance the model's generalization capability and improve predictive accuracy for the discovery of important virus-drug associations. Moreover, we plan to employ transfer learning methods for pretraining on large-scale datasets, aiming to enhance the model's generalization and improve the performance of the virus-disease association model.

Acknowledgment

We are grateful for resources from the High Performance Computing Center of Central South University.

References

[1] R. M. Meganck and R. S. Baric, "Developing therapeutic approaches for twenty-first-century emerging infectious viral diseases," *Nature Med.*, vol. 27, no. 3, pp. 401–410, 2021.

[2] S. C. Baker, "Coronaviruses: From common colds to severe acute respiratory syndrome," *Pediatr. Infect. Dis. J.*, vol. 23, no. 11, pp. 1049–1050, Nov. 2004.

[3] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," *J. Med. Virology*, vol. 92, no. 10, Oct. 2020, Art. no. 2249.

[4] N. Zhong et al., "Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, people's Republic of China, in February, 2003," *Lancet*, vol. 362, no. 9393, pp. 1353–1358, 2003.

[5] A. Zumla, D. S. Hui, and S. Perlman, "Middle east respiratory syndrome," *Lancet*, vol. 386, no. 9997, pp. 995–1007, 2015.

[6] S. Perlman, "Another decade, another coronavirus," *New England J. Med.*, vol. 382, pp. 760–762, 2020.

[7] P. Zhou et al., "A pneumonia outbreak associated with a new coronavirus of probable bat origin," *Nature*, vol. 579, no. 7798, pp. 270–273, 2020.

[8] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *Lancet*, vol. 395, no. 10223, pp. 470–473, 2020.

[9] Y. Pan et al., "Characterisation of SARS-CoV-2 variants in Beijing during 2022: An epidemiological and phylogenetic analysis," *Lancet*, vol. 401, no. 10377, pp. 664–672, 2023.

[10] C. Jia et al., "Immune repertoire sequencing reveals an abnormal adaptive immune system in COVID-19 survivors," *J. Med. Virol.*, vol. 95, no. 1, 2023, Art. no. e28340.

[11] C. K. Wong, I. C. Au, K. T. Lau, E. H. Lau, B. J. Cowling, and G. M. Leung, "Real-world effectiveness of molnupiravir and nirmatrelvir plus ritonavir against mortality, hospitalisation, and in-hospital outcomes among community-dwelling, ambulatory patients with confirmed SARS-CoV-2 infection during the omicron wave in Hong Kong: An observational study," *Lancet*, vol. 400, no. 10359, pp. 1213–1222, 2022.

[12] G. Li, R. Hilgenfeld, R. Whitley, and E. D. Clercq, "Therapeutic strategies for COVID-19: Progress and lessons learned," *Nature Rev. Drug Discov.*, vol. 22, pp. 449–475, 2023.

[13] C. Dietz and B. Maasoumy, "Direct-acting antiviral agents for hepatitis C virus infection–from drug discovery to successful implementation in clinical practice," *Viruses*, vol. 14, no. 6, 2022, Art. no. 1325.

[14] J. L. Goodman and L. Borio, "Finding effective treatments for COVID-19: Scientific integrity and public confidence in a time of crisis," *Jama*, vol. 323, no. 19, pp. 1899–1900, 2020.

[15] N. Berdigaliyev and M. Aljofan, "An overview of drug discovery and development," *Future Med. Chem.*, vol. 12, no. 10, pp. 939–947, 2020.

[16] D. Li et al., "In vitro and in vivo functions of SARS-CoV-2 infection-enhancing and neutralizing antibodies," *Cell*, vol. 184, no. 16, pp. 4203–4219, 2021.

[17] Y. J. Hou et al., "SARS-CoV-2 reverse genetics reveals a variable infection gradient in the respiratory tract," *Cell*, vol. 182, no. 2, pp. 429–446, 2020.

[18] H. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, "Advancing drug discovery via artificial intelligence," *Trends Pharmacological Sci.*, vol. 40, no. 8, pp. 592–604, 2019.

[19] K. Zheng et al., "SPRDA: A link prediction approach based on the structural perturbation to infer disease-associated piwi-interacting RNAs," *Brief. Bioinf.*, vol. 24, no. 1, 2023, Art. no. bbac498.

[20] L. Wong, L. Wang, Z.-H. You, C.-A. Yuan, Y.-A. Huang, and M.-Y. Cao, "GKLOMLI: A link prediction model for inferring miRNA–lncRNA interactions by using Gaussian kernel-based method on network profile and linear optimization algorithm," *BMC Bioinf.*, vol. 24, no. 1, 2023, Art. no. 188.

[21] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Brief. Bioinf.*, vol. 22, no. 4, 2021, Art. no. bbaa243.

[22] Y. Gu, S. Zheng, Q. Yin, R. Jiang, and J. Li, "REDDA: Integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction," *Comput. Biol. Med.*, vol. 150, 2022, Art. no. 106127.

[23] H.-J. Jiang, Z.-H. You, and Y.-A. Huang, "Predicting drug- disease associations via sigmoid kernel-based convolutional neural networks," *J. Transl. Med.*, vol. 17, no. 1, pp. 1–11, 2019.

[24] X. Su, L. Hu, Z. You, P. Hu, L. Wang, and B. Zhao, "A deep learning method for repurposing antiviral drugs against new viruses via multi-view nonnegative matrix factorization and its application to SARS-CoV-2," *Brief. Bioinf.*, vol. 23, no. 1, 2022, Art. no. bbab526.

[25] L. Zhou et al., "Probing antiviral drugs against SARS-CoV-2 through virus-drug association prediction based on the KATZ method," *Genomics*, vol. 112, no. 6, pp. 4427–4434, 2020.

[26] K. Deepthi, A. Jereesh, and Y. Liu, "A deep learning ensemble approach to prioritize antiviral drugs against novel coronavirus SARS-CoV-2 for COVID-19 drug repurposing," *Appl. Soft Comput.*, vol. 113, 2021, Art. no. 107945.

[27] R. Sharma, S. Shrivastava, S. K. Singh, A. Kumar, A. K. Singh, and S. Saxena, "Deep-AVPpred: Artificial intelligence driven discovery of peptide drugs for viral infections," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 5067–5074, Oct. 2022.

[28] S. S. Sadeghi and M. R. Keyvanpour, "An analytical review of computational drug repurposing," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 2, pp. 472–488, Mar./Apr. 2021.

[29] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, 2018.

[30] S. Wang, Q. Sun, Y. Xu, J. Pei, and L. Lai, "A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2," *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab211.

[31] A. Ianevski et al., "DrugVirus. info 2.0: An integrative data portal for broad-spectrum antivirals (BSA) and BSA-containing drug combinations (BCCs)," *Nucleic Acids Res.*, vol. 50, no. W1, pp. W272–W275, 2022.

[32] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[33] L. Wang and C. Zhong, "gGATLDA: LncRNA-disease association prediction based on graph-level graph attention network," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–24, 2022.

[34] J. Zheng, Y. Qian, J. He, Z. Kang, and L. Deng, "Graph neural network with self-supervised learning for noncoding RNA–drug resistance association prediction," *J. Chem. Inf. Model.*, vol. 62, no. 15, pp. 3676–3684, 2022.

[35] X. Lei, J. Tie, and Y. Pan, "Inferring metabolite-disease association using graph convolutional networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 2, pp. 688–698, Mar./Apr. 2022.

[36] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Proc. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 5812–5823.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7–9, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

[38] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[39] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. 25th Conf. Uncertainty Artif. Intell., ser. UAI '09*, Arlington, Virginia, USA, 2009, pp. 452–461. [Online]. Available: https://arxiv.org/pdf/1205.2618

[40] X. Tang, L. Cai, Y. Meng, J. Xu, C. Lu, and J. Yang, "Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19," *Front. Immunol.*, vol. 11, 2021, Art. no. 603615.

[41] J. Wu et al., "Self-supervised graph learning for recommendation," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, ser. SIGIR '21*, New York, NY, USA, 2021, pp. 726–735. [Online]. Available: https://doi.org/10.1145/3404835.3462862

[42] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9929–9939.

[43] L. Lu, S. Su, H. Yang, and S. Jiang, "Antivirals with common targets against highly pathogenic viruses," *Cell*, vol. 184, no. 6, pp. 1604–1620, 2021.

[44] T. Komeda et al., "Comparison of hospitalization incidence in influenza outpatients treated with baloxavir marboxil or neuraminidase inhibitors: A health insurance claims database study," *Clin. Infect. Dis.*, vol. 73, no. 5, pp. e1181–e1190, 2021.

[45] Y. Lou et al., "Clinical outcomes and plasma concentrations of baloxavir marboxil and favipiravir in COVID-19 patients: An exploratory randomized controlled trial," *Eur. J. Pharmaceut. Sci.*, vol. 157, 2021, Art. no. 105631.

[46] X. Wang et al., "The anti-influenza virus drug, arbidol is an efficient inhibitor of SARS-CoV-2 in vitro," *Cell Discov.*, vol. 6, no. 1, 2020, Art. no. 28.

[47] A. Zendehdel, M. Bidkhori, M. Ansari, S. Jamalimoghaddamsiyahkali, and A. Asoodeh, "Efficacy of oseltamivir in the treatment of patients infected with COVID-19," *Ann. Med. Surg.*, vol. 77, 2022, Art. no. 103679.