Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Methods 198 (2022) 11–18

Contents lists available at ScienceDirect

Methods

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

Heterogeneous graph attention networks for drug virus association prediction

Yahui Long [a,b], Yu Zhang [b], Min Wu [c], Shaoliang Peng [a], Chee Keong Kwoh [b], Jiawei Luo [a,*], Xiaoli Li [c,*]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410000, China
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore
[c] Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 138632, Singapore

ARTICLE INFO

Keywords:
COVID-19
SARS-CoV-2
Drug
Heterogeneous graph attention networks
Association prediction

ABSTRACT

Coronavirus Disease-19 (COVID-19) has lead global epidemics with high morbidity and mortality. However, there are currently no proven effective drugs targeting COVID-19. Identifying drug-virus associations can not only provide insights into the understanding of drug-virus interaction mechanism, but also guide and facilitate the screening of compound candidates for antiviral drug discovery. Since conventional experiment methods are time-consuming, laborious and expensive, computational methods to identify potential drug candidates for viruses (e.g., COVID-19) provide an alternative strategy. In this work, we propose a novel framework of Heterogeneous Graph Attention Networks for Drug-Virus Association predictions, named HGATDVA. First, we fully incorporate multiple sources of biomedical data, e.g., drug chemical information, virus genome sequences and viral protein sequences, to construct abundant features for drugs and viruses. Second, we construct two drug-virus heterogeneous graphs. For each graph, we design a self-enhanced graph attention network (SGAT) to explicitly model the dependency between a node and its local neighbors and derive the graph-specific representations for nodes. Third, we further develop a neural network architecture with tri-aggregator to aggregate the graph-specific representations to generate the final node representations. Extensive experiments were conducted on two datasets, i.e., DrugVirus and MDAD, and the results demonstrated that our model outperformed 7 state-of-the-art methods. Case study on SARS-CoV-2 validated the effectiveness of our model in identifying potential drugs for viruses.

1. Introduction

Coronavirus Disease-19 (COVID-19) is an infectious disease caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) [1]. SARS-CoV-2 is an enveloped, positive-sense, single-stranded RNA beta-coronavirus of the family Coronaviridae [2,3]. Coronaviruses (CoVs) typically affect the respiratory tract of mammals, including humans, and lead to mild to severe respiratory tract infections [4]. COVID-19 has lead to global epidemics with high morbidity and mortality. However, there are currently no antiviral drugs with proven clinical efficacy for the treatment of COVID-19. As COVID-19 is a new disease and our knowledge about SARS-CoV-2 is limited, it thus brings great challenge to develop new antiviral drugs against COVID-19 in a short time.

Recently, numerous research scientists around the world have focused on identifying potential drugs that can be repurposed to effectively treat COVID-19. Many common drugs approved for treating other human diseases are discovered to be effective for COVID-19 and are undergoing clinical trials. For example, Choy et al. [5] demonstrated remdesivir and lopinavir could inhibit SARS-CoV-2 replication in vitro. After that, Zhu et al. [6] indicated that Arbidol had superior effectivity to lopinavir/ritonavir in treating COVID-19. In the past decades, numerous drug-virus associations have been experimentally or clinically confirmed. For example, it was demonstrated that Azacytidine could generate activity against adenoviruses types 1, 2, 5 by inhibiting synthesis of viral DNA and protein [7]. Stadler et al. [8] showed that Amiodarone could alter late compartments of the endocytic pathway and inhibits SARS coronavirus infection. Hence, identifying drug-virus associations is very useful for disease prevention and treatment, as well as drug development. Considering that conventional experiment methods are time-consuming, laborious and expensive, computational

* Corresponding authors.
E-mail addresses: luojiawei@hnu.edu.cn (J. Luo), xlli@i2r.a-star.edu.sg (X. Li).

https://doi.org/10.1016/j.ymeth.2021.08.003
Received 27 May 2021; Accepted 11 August 2021
Available online 20 August 2021

methods provide a low cost complementary and can guide the screening of candidate compounds for drug discovery.

More recently, several computational methods have been proposed for drug-microbe association prediction. For example, Zhu et al. [9] presented a KATZ-based method named HMDAKATZ for drug-microbe predictions using drug chemical similarity and Gaussian kernel similarity. Long et al. [10] proposed a novel computational model named HNERMDA to predict drug-microbe associations based on a heterogeneous network. Following that, Long et al. [11] developed another prediction model called GCNMDA to infer latent drug-microbe associations combining with microbial protein interactions and drug chemical information. GCNMDA first encoded node representations using graph convolutional network (GCN), and then used the learned representations to identify potential associations between drugs and microbes. However, all the above existing methods do not fully consider the biological knowledge associated with viruses. Very recently, Andersen et al. [12] released a comprehensive database called DrugVirus that records experimentally and clinically validated drug-virus associations. In addition, there are many other available knowledge databases for drugs and viruses, such as Drugbank [13], Uniprot [14] and Virhostnet [15]. The availability of these rich data provides a golden opportunity for us to develop deep learning methods for drug-virus association predictions. In particular, graph neural network, e.g., graph attention network (GAT) proposed by Veličković et al. [16], is a promising deep learning technique due to its powerful capability for graph-structured data. For example, GAT has been successfully applied in various bioinformatics tasks, such as microbe-disease prediction [11] and enhancer-promoter prediction [17]. As such, we are motivated to generalize GAT for novel drug-virus predictions.

However, there exist several challenges when using GAT for drug-virus predictions. First, biological data related to drugs and viruses are often heterogeneous and different source data represent distinct biological meanings. Thus it is a challenge to integrate them as effective input features in a GAT framework for drug-virus predictions. Second, the observed/known drug-virus associations are limited and sparse so that it brings great challenges for using GAT to model drug-virus associations. To deal with above issues, we developed a novel Heterogeneous Graph Attention Network (HGAT) based framework for Drug-Virus Association prediction (HGATDVA). In particular, we first built two networks/graphs for drug-virus prediction, i.e., a drug-virus heterogeneous network with known drug-virus associations and a drug-host-virus heterogeneous network by integrating drug-target interactions with virus-host (human) protein interactions. Then we exploited multiple biomedical data, e.g., virus genome sequences, drug chemical structure information, viral protein sequences, drug-drug interactions, etc., to derive input features for drugs, viruses and proteins. For each graph, we designed a self-enhanced attention mechanism to learn graph-specific representation for each node. We further developed a Multilayer perceptron (MLP) based tri-aggregator to combine graph-specific representations and thus generated the final representations for nodes. Comprehensive experiments on two datasets (i.e., DrugVirus and MDAD) showed that our proposed HGATDVA model consistently outperformed seven state-of-the-art methods. Case study for SARS-CoV-2 further confirmed the effectiveness of our proposed model in identifying potential related drugs for viruses.

Overall, our contributions are summarized as follows.

- We integrated various data sources and constructed two heterogeneous networks, namely a drug-virus heterogeneous network and a drug-host-virus heterogeneous network, for drug-virus association prediction.
- We proposed a novel GAT-based framework for novel drug-virus prediction on two heterogeneous networks.
- We designed an self-enhanced attention mechanism to learn node representation, which explicitly models the dependency between nodes and their local neighbors in each heterogeneous network. We further developed a tri-aggregator to combine graph-specific representations as final representations.
- Comprehensive experiments on two datasets and case study for SARS-CoV-2 demonstrated that our model was a promising tool to identify potential drugs for viruses.

Table 1
The statistics for each drug-virus/microbe association dataset.

	DrugVirus	MDAD
# Drugs	202	1373
# Viruses/ microbes	104	173
# Associations	1016	2470
Density	4.836%	1.040%

2. Materials

2.1. Networks constructions for drugs and viruses

We use two different datasets for known drug-virus/microbe associations, i.e., DrugVirus [12] and MDAD [18]. DrugVirus dataset records activities and development statuses of 118 compounds/drugs which altogether target 83 human viruses, including recently occurred novel coronavirus named SARS-CoV-2. Besides, we manually curate 140 clinically or experimentally validated drug-virus associations between 84 drugs and 21 viruses from existing literature. Overall, we obtain 1016 observed drug-virus associations involving 202 drugs and 104 viruses. MDAD includes 5505 clinically or experimentally verified microbe-drug associations between 1388 drugs and 174 microbes. After removing the repeated data, we finally attain 2470 associations between 1373 drugs and 173 microbes. The statistics for each database are shown in Table 1. We construct a drug-virus/microbe heterogeneous network named *Net*1 by connecting Gaussian kernel drug similarity network to Gaussian kernel virus/microbe similarity network, via drug-virus/microbe bipartite network. Following the method [11], we calculate Gaussian kernel similarity for drugs and viruses/microbes based on known drug-virus/microbe associations.

We further construct a drug-host-virus heterogeneous network named *Net2* by integrating drug-target interactions (DTIs) with virus-host (human) protein–protein interactions (PPIs). In particular, we download DTIs from the latest version of Drugbank [13] and PharmGKB [19] databases. PPIs are derived from Virhostnet [15] and mentha [20] databases. After mapping the shared proteins (i.e., targets) between DTIs and PPIs, we finally obtained 180 DTIs between 202 drugs and 119 host proteins, and 256 PPIs between 83 viral proteins and 119 host proteins. Note that we only select viral proteins that are associated with more than one out of 104 viruses.

For each graph, we define an adjacent matrix as inputs of the model. For *Net*1, taking drug-virus pairs as example, we first use an binary matrix $I_1 \in \mathbb{R}^{nd\times nv}$ to represent drug-virus associations, with nd and nv denoting the numbers of drugs and viruses respectively. If the association between drug d_i and virus v_j is clinically or experimentally confirmed, $(I_1)_{ij}$ is equal to 1, otherwise 0. Then we represent its adjacent matrix $A^1 \in \mathbb{R}^{(nd+nv)\times(nd+nv)}$ as follows,

$$A^1 = \begin{bmatrix} S_d & I_1 \\ I_1^T & S_v \end{bmatrix}, \tag{1}$$

where $S_d \in \mathbb{R}^{nd\times nd}$ and $S_v \in \mathbb{R}^{nv\times nv}$ represent Gaussian kernel similarity matrices for drugs and viruses respectively. Similarly, for *Net*2, we denote drug-target interactions and virus-host protein interactions as $I_2 \in \mathbb{R}^{nd\times nm}$ and $I_3 \in \mathbb{R}^{np\times nm}$, respectively. nm and np represents the numbers of host proteins and viral proteins respectively. Hence, the adjacent matrix $A^2 \in \mathbb{R}^{(nd+nm+np)\times(nd+nm+np)}$ for *Net*2 is formulated as follows:

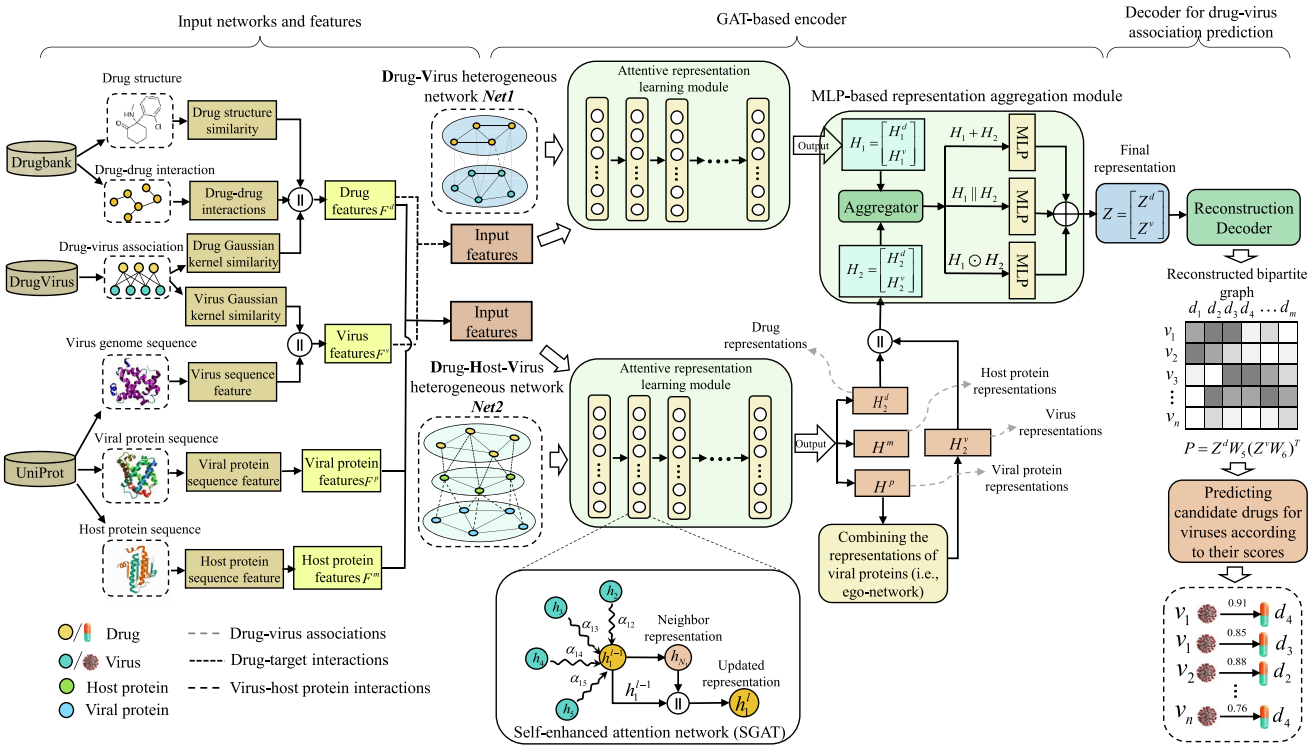


Fig. 1. The overall architecture of HGATDVA for drug-virus predictions.

$$A^2 = \begin{bmatrix} 0 & I_2 & 0 \\ I_2^T & 0 & I_3^T \\ 0 & I_3 & 0 \end{bmatrix}. \tag{2}$$

$$F^2 = \begin{bmatrix} F^d & 0 & 0 \\ 0 & F^m & 0 \\ 0 & 0 & F^p \end{bmatrix}. \tag{4}$$

2.2. Features for drugs and viruses

In this work, we construct rich features for drugs, viruses, host proteins and viral proteins from different biological databases, including Drugbank [13], DrugVirus [12] and UniProt [14]. Specifically, we first download drug chemical structure information and drug-drug interactions from Drugbank database. Then we measure drug structure similarity using the method proposed by Hattori et al. [21]. We finally generate drug feature $F^d \in \mathbb{R}^{nd \times r_1}$ (r_1 is the dimension of drug feature) by concatenating drug structure similarity matrix, drug-drug interaction matrix with Gaussian kernel drug similarity matrix.

Furthermore, we collect genome sequences for viruses, and protein sequences for host proteins and viral proteins from UniProt database. Here we use k-mer feature representation method [22] to extract sequence features for viruses, host proteins and viral proteins from the collected sequences. For viruses, we concatenate the virus sequence features with Gaussian kernel virus similarity as their final features, denoted as $F^v \in \mathbb{R}^{nv \times r_2}$ (r_2 represents feature dimension). For host proteins and viral proteins, we utilize the extracted sequence features $F^m \in \mathbb{R}^{nm \times r_3}$ and $F^p \in \mathbb{R}^{np \times r_4}$ as their features respectively. r_3 and r_4 denote the feature dimensions of host and viral proteins respectively. The whole process to generate features for drugs, viruses, host and viral proteins is shown at the left part of Fig. 1. In consistent with Eq. (1) and Eq. (2), the feature matrices $F^1 \in \mathbb{R}^{(nd+nv) \times (r_1+r_2)}$ and $F^2 \in \mathbb{R}^{(nd+nm+np) \times (r_1+r_3+r_4)}$ for *Net*1 and *Net*2 are constructed as follows:

$$F^1 = \begin{bmatrix} F^d & 0 \\ 0 & F^v \end{bmatrix}, \tag{3}$$

3. Methods

In this work, we propose a novel heterogeneous graph attention network (HGAT) based framework named HGATDVA to predict novel drug-virus associations. As shown in Fig. 1, HGATDVA consists of three main steps. First, we design an attentive representation learning module with self-enhanced attention mechanism to learn two graph-specific representations for each node from the constructed two graphs respectively. Second, we further develop a neural network architecture to aggregate graph-specific representations for nodes. Third, we reconstruct the drug-virus bipartite graph based on the learned representations. Next, we introduce the above three steps in details.

3.1. Self-enhanced GAT for representation learning

Graph attention network (GAT) [16], which aims to preserve the importance of different neighbors, possesses excellent performance in addressing graph-structured data. However, while standard GAT considers the importance of neighbors, it simultaneously weakens the importance of centre node itself. In fact, node itself plays more important role than neighbors during the representation learning process. In this work, we adopt **S**elf-enhanced **G**raph **A**ttention **N**etworks (SGATs) to learn node representations for drug-virus predictions. The key idea behind SGATs is to retrieve the importance of node itself to strengthen node representation learning.

3.1.1. Preliminary representation learning

Recall that we have derived adjacent matrices (i.e., A^1 and A^2) and feature matrices (i.e., F^1 and F^2) for *Net*1 and *Net*2 respectively. As such,

we can use them to learn node representations. Specifically, we implement multi-layer SGATs on each graph and thus can obtain a graph-specific representation for each drug and virus. More specifically, given a node, we first learn the importance of its neighbors, then derive its neighbor representation by aggregating the representations of all local neighbors according to their attention coefficients, and finally generate its self-enhanced representation by concatenating current representation with the aggregated neighbor representation. Mathematically, the attention score of a pair between drug d_i and virus v_j is formulated as follows:

$$e_{ij}^{l} = f\left(W_1 h_i^{(l-1)}, W_1 h_j^{(l-1)}\right), \tag{5}$$

where $f(\cdot)$, parameterized by a weight matrix $W_1 \in \mathbb{R}^{d_1 \times d_2}$ (d_1 and d_2 are the dimensions of $h^{(l-1)}$ before and after transformation respectively), represents a feed-forward neural network, which transforms linearly input features into high-level features. e_{ij}^{l} is attention score that represents the importance of neighbor v_j to d_i in the l-th layer. $h_i^{(l-1)}$ denotes the output representation of node d_i in the $(l-1)$-th layer. Note that h_i^0 is defined as the raw input features F_i of node d_i. To make attention coefficient across different nodes easily comparable, we further normalize attention scores across all neighbors using the softmax function:

$$\alpha_{ij} = softmax(e_{ij}^{l}) = \frac{\exp(e_{ij}^{l})}{\sum_{t \in \mathcal{N}_i} \exp(e_{it}^{l})}, \tag{6}$$

where α_{ij} is attention coefficient and $\mathcal{N}_i$ represents the set of local neighbors of node d_i.

After that, we can obtain the neighbor representation $h_{\mathcal{N}_i}$ for d_i by aggregating the representations of all its neighbors $\mathcal{N}_i$ according to their attention coefficients as follows.

$$h_{\mathcal{N}_i} = ReLU\left(\sum_{t \in \mathcal{N}_i} \alpha_{it} \cdot W_1 h_t^{(l-1)}\right). \tag{7}$$

where *ReLU* denotes activation function.

As mentioned above, the representation learned from standard GAT may be insufficiently informative since the parts of weight values are assigned to neighbors and thus lead to the importance of node itself reduced. Motivated by that, we explicitly model the dependencies between nodes and neighbors to enrich node representation. Formally, we yield a self-enhanced representation for d_i by concatenating its neighbor representation $h_{\mathcal{N}_i}$ with current representation $h_i^{(l-1)}$ as follows:

$$h_i^{l} = h_{\mathcal{N}_i} \| W_1 h_i^{(l-1)}. \tag{8}$$

Finally, we adopt multi-head attention to stabilize the learning process of attention coefficients.

$$h_i^{l} = \Big\|_{k=1}^{K} ReLU\left(\left(\sum_{t \in \mathcal{N}_i} \alpha_{it}^{k} \cdot W_1^{k} h_t^{(l-1)}\right) \middle\| W_1^{k} h_i^{(l-1)}\right), \tag{9}$$

K denotes the number of attentional heads, α_{ij}^{k} represents the k-th attention coefficient between d_i and v_j. Here we adopt multi-layer SGATs to learn node representations. As the layer iterates, nodes incrementally gain more and more information from global neighbors. Empirically, we set l as 2.

3.1.2. Modeling virus-protein interactions

We carry out SGATs on *Net*2 and can then derive the second representations $H_2^d \in \mathbb{R}^{nd \times Kd_2}$ for drugs, as well as viral protein representation $H^p \in \mathbb{R}^{np \times Kd_2}$, as shown in the middle of Fig. 1. Considering a virus v_i, we use $\mathcal{N}_{v_i}$ to denote the set of its proteins, termed ego-network. To characterize the first-order connectivity structure of virus v_i, we generate the second representation $(H_2^v)_i$ for v_i through the following linear combination of its ego-network.

$$(H_2^v)_i = \sum_{t \in \mathcal{N}_{v_i}} H_t^p. \tag{10}$$

3.2. Multi-Layer Perceptron-based representation aggregation

After implementing SGATs on two graphs (i.e., *Net*1 and *Net*2), we can derive two graph-specific representations named H_1 and H_2 for nodes respectively. In fact, different graphs include distinct semantic information between nodes. To more accurately capture this valuable information, we further design a Multi-Player Perceptron (MLP) based aggregation architecture with tri-aggregator to integrate graph-specific representations. Specifically, the tri-aggregator is defined as follows:

$$Z = \begin{bmatrix} Z^d \\ Z^v \end{bmatrix} = LeakyReLU(W_2(H_1 + H_2) + b_2) + LeakyReLU(W_3(H_1 \| H_2) + b_3) + LeakyReLU(W_4(H_1 \odot H_2) + b_4), \tag{11}$$

where *LeakyReLU* denotes activation function, $\|$ and $\odot$ denote concatenation and element-wise product operations respectively. $W_2 \in \mathbb{R}^{Kd_2 \times d_3}$, $W_3 \in \mathbb{R}^{2Kd_2 \times d_3}$, $W_4 \in \mathbb{R}^{Kd_2 \times d_3}$ represent learnable weight matrices with d_3 denoting the number of neurons in the MLP. $b_2 \in \mathbb{R}^{d_3}$, $b_3 \in \mathbb{R}^{d_3}$, $b_4 \in \mathbb{R}^{d_3}$ represent learnable bias matrices. Here we introduce three types of aggregators, i.e., sum, concatenation and element-wise product, to aggregate graph-specific representations, which enables our model to fully capture rich semantic information hidden in different graphs.

3.3. Decoder for drug-virus associations re-construction

We have derived feature representations $Z^d \in \mathbb{R}^{nd \times d_3}$ for drugs and feature representations $Z^v \in \mathbb{R}^{nv \times d_3}$ for viruses. Then we can utilize them to reconstruct drug-virus associations in Eq. (12) and define the loss function in Eq. (13).

$$P = sigmoid\left(Z^d W_5 (Z^v W_6)^T\right), \tag{12}$$

$$\mathscr{L}_{REC} = \sum_{(i,j) \in \mathscr{O}^+ \cup \mathscr{O}^-} \Theta(P_{ij}, A_{ij}), \tag{13}$$

where $W_5 \in \mathbb{R}^{d_3 \times d_4}$, $W_6 \in \mathbb{R}^{d_3 \times d_4}$ are trainable weight matrices that project representations back into original features. d_4 denotes the dimension of weight matrix. *sigmoid* means activation function and Θ is MSE (i.e., mean square error) loss function. For better training our model, here we adopt negative sampling strategy to train the model. $\mathscr{O}^+$ and $\mathscr{O}^-$ represent the sets of positive and negative samples respectively.

3.4. Model training

In the decoder, there are two trainable weight matrices W_5 and W_6. We add a regularization term in Eq. 14 to limit their influences on our model. Thus the overall loss function can be defined as follows.

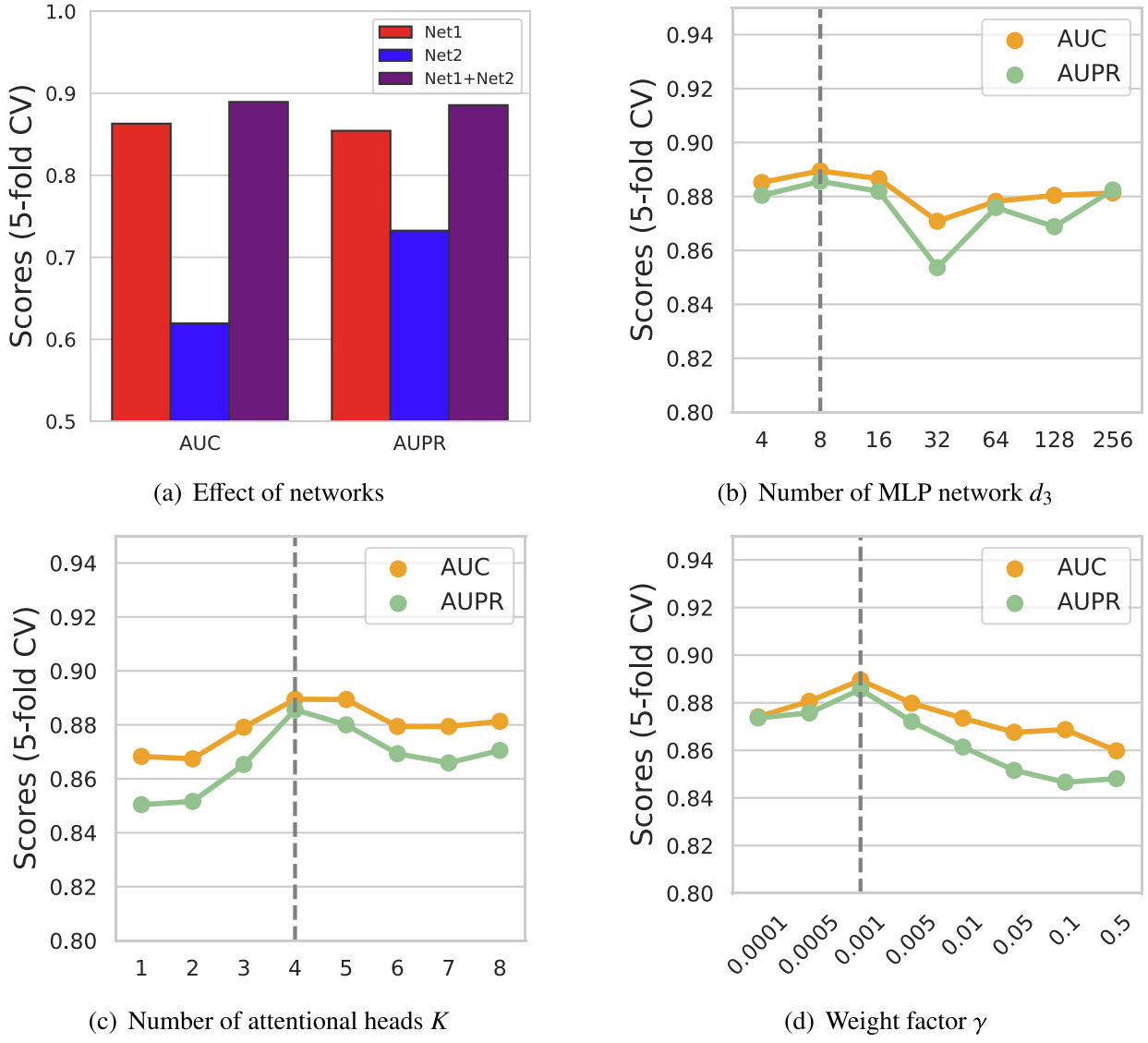
Table 2
The AUC and AUPR for various methods on two datasets. The best results are marked in bold and the second best is underlined.

Methods	DrugVirus		MDAD	
	AUC	AUPR	AUC	AUPR
HMDAKATZ	0.7750 ± 0.0038	0.7525 ± 0.0031	0.9015 ± 0.0007	0.9053 ± 0.0006
WMGHMDA	0.7337 ± 0.0013	0.7693 ± 0.0025	0.8097 ± 0.0012	0.8657 ± 0.0016
NTSHMDA	0.7680 ± 0.0028	0.7268 ± 0.0030	0.8325 ± 0.0033	0.8028 ± 0.0026
WNN-GIP	0.8002 ± 0.0193	0.8436 ± 0.0183	0.8721 ± 0.0162	0.8922 ± 0.0137
IMCMDA	0.6235 ± 0.0245	0.6962 ± 0.0302	0.7466 ± 0.0102	0.7773 ± 0.0113
GCNMDA	0.8685 ± 0.0125	0.8567 ± 0.0132	0.9122 ± 0.0112	0.9169 ± 0.0087
EGATMDA	0.8405 ± 0.0123	0.8264 ± 0.0112	0.8517 ± 0.0088	0.8311 ± 0.0110
GCMDR	0.8485 ± 0.0062	0.8509 ± 0.0040	0.8243 ± 0.0168	0.8206 ± 0.0141
GCN	0.8182 ± 0.0122	0.8093 ± 0.0290	0.8666 ± 0.0164	0.8778 ± 0.0164
GAT	0.7402 ± 0.0212	0.6942 ± 0.0196	0.8213 ± 0.0206	0.8371 ± 0.0286
HGATDVA-GAT	0.8701 ± 0.0168	0.8542 ± 0.0152	0.8981 ± 0.0140	0.9142 ± 0.0086
HGATDVA	**0.8895 ± 0.0171**	**0.8856 ± 0.0103**	**0.9254 ± 0.0092**	**0.9246 ± 0.0059**

$$\mathscr{L}_{Overall} = \mathscr{L}_{REC} + \gamma(\|W_5\|_F^2 + \|W_6\|_F^2) \tag{14}$$

where γ denotes weight factor that regularizes the influences of parameters W_5 and W_6.

Following Long et al. [11], we adapt Adam optimizer [23] to train our model. After that, we obtain the predicted score matrix P and prioritize candidate drugs for viruses (e.g., SARS-CoV-2) according to their probability scores to screen the most possible antiviral drugs.



(a) Effect of networks

(b) Number of MLP network d_3

(c) Number of attentional heads K

(d) Weight factor γ

Fig. 2. Network and parameter sensitivity analysis for HGATDVA on DrugVirus in 5-fold CV.

4. Results

In this section, we first briefly introduce experimental settings and then validate the effectiveness of our proposed model of HGATDVA through the comparison with seven state-of-the-art methods and case study.

4.1. Experimental settings

In this work, we implemented standard 5-fold cross-validation (5-fold CV) on two datasets, i.e., DrugVirus and MDAD, to validate the effectiveness of our proposed HGATDVA model. Specifically, we randomly divided all known drug-virus association pairs into five groups. For each round, we selected in turn four groups of drug-virus pairs (i.e., positive samples) as training samples while the rest of one group of drug-virus pairs were employed as test samples. Here we adopted negative sampling strategy to better train our model. For each iteration, together with training positive samples, we randomly sampled an equal-size sets of pairs from unknown drug-virus pairs as negative samples to train the model. Similarly, we randomly selected the same number of negative samples as that of test positive samples for testing.

In our model, the training epoch was set to 600 and the learning rate in the optimization algorithm was set to 0.005. In the next section, we would discuss the influences of several important parameters on our model in detail, including the number of attentional heads K, the number of neurons of MLP network d_3 and weight factor γ in the overall loss. All the experiments in this work were implemented based on the open source machine learning framework Tensorflow (https://github.com/tensorflow/tensorflow).

4.2. Baseline methods

As mentioned above, identifying drug-virus associations is a new issue and few computation methods have been developed for this important task. Thus we compare our proposed HGATDVA model with 2 approaches developed for microbe-drug predictions, as well as 5 approaches proposed for addressing other biological link/association prediction problems. Baseline methods are introduced as follows:

- GCNMDA [11]: is a novel *graph convolutional network (GCN)*-based framework, designed for microbe-drug prediction.
- HMDAKATZ [9]: is a *KATZ*-based computational model for identifying microbe-drug associations.
- WMGHMDA [24]: is a *meta-graph* based computational model proposed for microbe-disease association prediction.
- WNN-GIP [25]: is a *weighted nearest neighbor-Gaussian interaction profile* model, developed for drug-target prediction.
- NTSHMDA [26]: is a *random walk with restart* based model, proposed to predict microbe-disease associations.
- GCMDAR [27]: is a *graph convolutional network* based approach for identifying miRNA-drug resistance associations.
- IMCMDA [28]: is a *inductive matrix completion (IMC)* based method, designed for miRNA-disease predictions.
- HGATDVA-GAT: is a variant of our proposed HGATMDA model, which uses standard GAT to learn node representation.

For a fair comparison, all existing six baseline methods adopted the default parameter values which were suggested in their original papers and were implemented on the same benchmark datasets, i.e., DrugVirus and MDAD. Note that for MDAD, all baseline methods used Gaussian kernel similarities for microbes and drugs as input features. Besides, machine learning-based baseline models (e.g., GCNMDA, GCMDAR and IMCMDA) utilized the same number of randomly sampled unknown pairs (i.e., negative samples) as that of positive samples for training.

Table 2 shows the results on two datasets, which indicate that our proposed HGATDVA model consistently outperforms 7 baseline methods in terms of AUC and AUPR. In particular, HGATDVA achieves an average AUC of 0.8895 and AUPR of 0.8856 on dataset DrugVirus, which are 2.42% and 3.37% higher than the second best method GCNMDA. For MDAD, our model also performs better than all baseline methods with average AUC of 0.9254 and average AUPR of 0.9246, which are 1.32% and 0.84% better than that of the second best method GCNMDA. From Table 2, we can also observe that HGATDVA-GAT achieves lower AUC and AUPR values than HGATDVA, which demonstrates that SGATs is useful for enriching node representation learning. This is one of main reasons why our model is superior to baseline methods. In addition, we design a tri-aggregator to aggregate representations learned from different graphs, which enables our model to more accurately capture semantic information between nodes and thus helps to enhance the prediction capability of our model.

From Table 2, it is found that compared with MDAD dataset, all methods achieve worse performance on DrugVirus in terms of AUC and AUPR. As shown in Table 1, DrugVirus is sparse and much smaller than MDAD, which results in less known pairs available for training for most methods and thus reduces the prediction accuracy. However, our model still achieves relatively satisfactory prediction performance.

4.3. Effect of different data source

Recall that we construct two heterogeneous network for drugs and viruses, i.e., *Net*1 and *Net*2, respectively. To assess their influences on HGATDVA, we implement our model on DrugVirus dataset with one of both networks as input and used 5-fold CV to evaluate its performance. The results are displayed in Fig. 2 (a), from which we can observe that both networks help to improve the prediction performance of our model. Besides, we can conclude that *Net*1 contributes much more than *Net*2. The main reason is because the second network *Net*2 is sparser compared to the first network *Net*1. Thus the node representations learned from *Net*2 may relatively weak.

4.4. Parameter analysis

In our work, there are several important parameters that can influence the performance of our proposed HGATDVA model, such as the number of neurons of MLP network d_3, the number of attentional heads K and the weight factor γ. Here we conduct parameter analysis for these parameters. All the experiments are implemented on DrugVirus dataset and evaluated by 5-fold CV.

In our proposed framework, the number of neurons of MLP network d_3 determines the dimension of node representation. To measure its impact on our model, we select its value from {4, 8, 16, 32, 64, 128, 256}. Fig. 2 (b) shows the performance first slightly increases and then decreases with $d_3 = 8$ achieving the best results. Our model adopts multi-attention heads to stabilize the process of attention coefficient learning. We evaluate our model by changing the number of attentional heads K from 1 to 8 with a step value of 1. From Fig. 2 (c), we observe a small or larger value is not good for the model performance. Our model will obtain more desirable performance when K is set to 4. In the decoder, we use a weight factor γ to control the influences of weight matrices W_5 and W_6. We choose its value from {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5} to assess its impact. The results from Fig. 2 (d) indicate that a small value of γ regularizes the impact of weight matrix well, and when γ is more than 0.001, the performance will gradually decrease and the best performance is reached when γ is set to 0.001.

4.5. Case study

For further validating the effectiveness of our proposed HGATDVA model, we carry out case study on dataset DrugVirus for SARS-CoV-2. We first remove all known entries, and then prioritize all candidate drugs according to their prediction scores. We finally evaluate the

Table 3
The top 20 predicted SARS-CoV-2-associated drugs. The first column records top 10 drugs, while the third column records top 11–20 drugs. "*" denotes the drugs are predicted by other in silico prediction approaches.

Drug	Evidence	Drug	Evidence
Itraconazole	Unconfirmed	Regorafenib	Stukalov et al. [34]
Mycophenolic acid	PMID:32579258	Vidarabine*	PMID:32488835
Favipiravir	PMID:32297571	Amiloride	PMID:32428379
Pleconaril	PMID:32295237	Trifluridine	PMID:32476594
Darunavir*	PMID:32306822	Ritonavir	PMID:32360480
Cidofovir	PMID:32562705	Cyclosporine	PMID:32529737
Nitazoxanide	PMID:32817953	Sorafenib	Unconfirmed
Indinavir*	PMID:32294562	Amodiaquine	PMID:32545799
Obatoclax	PMID:32545799	Niclosamide	PMID:32125140
Brequinar*	PMID:32426387	Saquinavir*	PMID:32294562

performance of our model by checking how many drugs could be confirmed by previously published literature among the top 10 and 20 ranking list.

As mentioned above, SARS-CoV-2, the causative agent of COVID-19, is an enveloped, positive-sense, single-stranded RNA betacoronavirus of the family Coronaviridae [2], which can affect the respiratory tract of humans and lead to mild to severe respiratory tract infections [4]. Recently some drugs, which are approved for treating other diseases, have been demonstrated to be promising candidate drugs against COVID-19. For example, it was demonstrated that Chloroquine and Hydroxychloroquine have in vitro activity against SARS-CoV-2 [29]. Wang et al. [30] showed that Remdesivir could inhibited virus infection efficiently in a human cell line, which was sensitive to COVID-19.

The results in Table 3 indicate that 9 and 19 out of the top 10 and 20 predicted drugs which are associated with SARS-CoV-2 can obtain validations from previous reports. It is found that the majority of drugs can be verified by wet-lab or clinic trials. For example, Risner et al. [31] conducted a screen of small molecules in cell culture and finally discovered that Nitazoxanide was able to inhibit SARS-CoV-2 infection. Lanevski et al. [32] identified that obatoclax was an potential antiviral drugs against SARS-CoV-2 by screening safe-in-man broad-spectrum antivirals against the SARS-CoV-2 infection in Vero-E6 cells. Dong et al. [33] found that favipiravir had potential antiviral action on SARS-CoV-2 by undergoing clinic trials. Also, some identified drugs are successfully predicted by previous in silico approaches, such as Darunavir, Indinavir and Brequinar. The high prediction accuracy, i.e., 90% and 95%, indicates our model has powerful capability to predict candidate drugs for a given virus, and thus is a promising tool to assist pharmacologists and biologists in screening potential compounds for drug discovery.

5. Discussion and conclusion

COVID-19 has lead global epidemics with high morbidity and mortality. Due to the lack of proven available drugs against COVID-19, there is an urgent need to develop effective approaches to accelerate the development of vaccines and drugs. Identifying drug-virus associations can not only provide great insight into the understanding of interaction mechanisms between drugs and viruses, but also assist to narrow the screening scopes of compound candidates for drug discovery. Considering that conventional experiment methods are time-consuming, laborious and expensive, computational methods are an alternative strategy. However, to the best our knowledge, few computational methods have been proposed for this critical task.

In this work, we propose a heterogeneous graph attention network framework named HGATDVA for novel drug-virus predictions. First, we take full advantage of multiple biomedical data, including virus genome sequences, drug chemical structure information, and Gaussian kernel similarity, to construct rich features for drugs and viruses. Besides, we build two heterogeneous networks for drugs and viruses by utilizing different genres of biological link data, such as drug-virus associations, drug-target interactions and virus-host protein interactions. Second, we introduce a self-enhanced graph attention network (SGAT) for node representation learning, which explicitly models the dependency between nodes and neighbors, leading to more informative representations. To capture rich semantic information from different graphs, we further design a tri-aggregator to aggregate graph-specific representations for nodes. Extensive experiments on two datasets (i.e., DrugVirus and MDAD) demonstrated that our proposed HGATDVA model outperformed 7 state-of-the-art methods. Case study on SARS-CoV-2 further confirmed the effectiveness of our model in identifying potential drugs for viruses.

Declaration of competing interest

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

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant No. 61873089), the National Key R&D Program of China (Grant No. 2017YFB0202602 and 2017YFC1311003), the Chinese Scholarship Council (CSC) (201906130027), A*STAR and NTU of Singapore.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.ymeth.2021.08.003.

References

[1] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (7798) (2020) 270–273.
[2] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, The Lancet 395 (10223) (2020) 470–473.
[3] Q. Wang, J. Wu, H. Wang, Y. Gao, Q. Liu, A. Mu, W. Ji, L. Yan, Y. Zhu, C. Zhu, et al., Structural basis for rna replication by the sars-cov-2 polymerase, Cell 182 (2) (2020) 417–428.
[4] C.I. Paules, H.D. Marston, A.S. Fauci, Coronavirus infections-more than just the common cold, Jama 323 (8) (2020) 707–708.
[5] K.-T. Choy, A.Y.-L. Wong, P. Kaewpreedee, S.-F. Sia, D. Chen, K.P.Y. Hui, D.K.W. Chu, M.C.W. Chan, P.P.-H. Cheung, X. Huang, et al., Remdesivir, lopinavir, emetine, and homoharringtonine inhibit sars-cov-2 replication in vitro, Antiviral Res., p. 104786, 2020.
[6] Z. Zhu, Z. Lu, T. Xu, C. Chen, G. Yang, T. Zha, J. Lu, Y. Xue, Arbidol monotherapy is superior to lopinavir/ritonavir in treating covid-19, J. Infect. 81 (1) (2020) e21–e23.
[7] I. Alexeeva, N. Dyachenko, L. Nosach, V. Zhovnovataya, S. Rybalko, R. Lozitskaya, A. Fedchuk, V. Lozitsky, T. Gridina, A. Shalamay, et al., 6-azacytidine-compound with wide spectrum of antiviral activity, Nucleosides Nucleotides Nucl. Acids 20 (4–7) (2001) 1147–1152.
[8] K. Stadler, H.R. Ha, V. Ciminale, C. Spirli, G. Saletti, M. Schiavon, D. Bruttomesso, L. Bigler, F. Follath, A. Pettenazzo, et al., Amiodarone alters late endosomes and inhibits sars coronavirus infection at a post-endosomal level, Am. J. Respirat. Cell Mol. Biol. 39 (2) (2008) 142–149.
[9] L. Zhu, G. Duan, C. Yan, and J. Wang, Prediction of microbe-drug associations based on katz measure, in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019, pp. 183–187.
[10] Y. Long and J. Luo, Association mining to identify microbe drug interactions based on heterogeneous network embedding representation, IEEE J. Biomed. Health Inf. 2020.
[11] Y. Long, M. Wu, C.K. Kwoh, J. Luo, X. Li, Predicting human microbe-drug associations via graph convolutional network with conditional random field, Bioinformatics (2020).
[12] P.I. Andersen, A. Ianevski, H. Lysvand, A. Vitkauskiene, V. Oksenych, M. Bjørås, K. Telling, I. Lutsar, U. Dampis, Y. Irie, et al., Discovery and development of safe-in-man broad-spectrum antiviral agents, Int. J. Infect. Diseases (2020).
[13] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al., Drugbank 5.0: a major update to the drugbank database for 2018, Nucl. Acids Res. 46 (D1) (2018) D1074–D1082.
[14] U. Consortium, Uniprot: a worldwide hub of protein knowledge, Nucl. Acids Res. 47 (D1) (2019) D506–D515.

[15] S.D. Thibaut Guirimand, V. Navratil, Virhostnet 2.0: surfing on the web of virus/host molecular interactions data, Nucleic Acids Res. 43 (D1) (2015) D583–587.
[16] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph Attention Networks, in: International Conference on Learning Representations, 2018.
[17] Z. Hong, X. Zeng, L. Wei, X. Liu, Identifying enhancer–promoter interactions with neural network based on pre-trained dna vectors and attention mechanism, Bioinformatics 36 (4) (2020) 1037–1043.
[18] Y.-Z. Sun, D.-H. Zhang, S.-B. Cai, Z. Ming, J.-Q. Li, X. Chen, Mdad: a special resource for microbe-drug associations, Front. Cell. Infect. Microbiol. 8 (2018) 424.
[19] J.M. Barbarino, M. Whirl-Carrillo, R.B. Altman, T.E. Klein, Pharmgkb: A worldwide resource for pharmacogenomic information, Wiley Interdisc. Rev.: Syst. Biol. Med. 10 (4) (2018), e1417.
[20] A. Calderone, L. Castagnoli, G. Cesareni, Mentha: a resource for browsing integrated protein-interaction networks, Nat. Methods 10 (8) (2013) 690–691.
[21] M. Hattori, N. Tanaka, M. Kanehisa, S. Goto, Simcomp/subcomp: chemical structure search servers for network analyses, Nucl. Acids Res. 38 (suppl_2) (2010) W652–W656.
[22] Y. Zhang, C. Jia, M.J. Fullwood, C.K. Kwoh, Deepcpp: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for rna coding potential prediction, Briefings Bioinf. (2020).
[23] D.P. Kingma, J.A. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2019.
[24] Y. Long, J. Luo, Wmghmda: a novel weighted meta-graph-based model for predicting human microbe-disease association on heterogeneous information network, BMC Bioinf. 20 (1) (2019) 541.
[25] T. Van Laarhoven, E. Marchiori, Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile, PloS one 8 (6) (2013), e66952.
[26] J. Luo, Y. Long, Ntshmda: Prediction of human microbe-disease association based on random walk by integrating network topological similarity, IEEE/ACM Trans. Comput. Biol. Bioinf. 17 (4) (2018) 1341–1351.
[27] Y.-A. Huang, P. Hu, K.C. Chan, Z.-H. You, Graph convolution for predicting associations between mirna and drug resistance, Bioinformatics 36 (3) (2020) 851–858.
[28] X. Chen, L. Wang, J. Qu, N.-N. Guan, J.-Q. Li, Predicting mirna–disease association based on inductive matrix completion, Bioinformatics 34 (24) (2018) 4256–4265.
[29] X. Yao, F. Ye, M. Zhang, C. Cui, B. Huang, P. Niu, X. Liu, L. Zhao, E. Dong, C. Song, et al., In vitro antiviral activity and projection of optimized dosing design of hydroxychloroquine for the treatment of severe acute respiratory syndrome coronavirus 2 (sars-cov-2), Clin. Infect. Dis. (2020).
[30] M. Wang, R. Cao, L. Zhang, X. Yang, J. Liu, M. Xu, Z. Shi, Z. Hu, W. Zhong, G. Xiao, Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-ncov) in vitro, Cell Res. 30 (3) (2020) 269–271.
[31] S. Pant, M. Singh, V. Ravichandiran, U. Murty, H.K. Srivastava, Peptide-like and small-molecule inhibitors against covid-19, J. Biomol. Struct. Dyn. (2020) 1–10.
[32] A. Ianevski, R. Yao, M.H. Fenstad, S. Biza, E. Zusinaite, T. Reisberg, H. Lysvand, K. Løseth, V.M. Landsem, J.F. Malmring, et al., Potential antiviral options against sars-cov-2 infection, Viruses 12 (6) (2020) 642.
[33] L. Dong, S. Hu, J. Gao, Discovering drugs to treat coronavirus disease 2019 (covid-19), Drug Discov. Therapeut. 14 (1) (2020) 58–60.
[34] A. Stukalov, V. Girault, V. Grass, V. Bergant, O. Karayel, C. Urban, D.A. Haas, Y. Huang, L. Oubraham, A. Wang, et al., Multi-level proteomics reveals host-perturbation strategies of sars-cov-2 and sars-cov, Biorxiv, 2020.