



OPEN

## Prioritizing antiviral drugs against SARS-CoV-2 by integrating viral complete genome sequences and drug chemical structures

Lihong Peng<sup>1,4</sup>, Ling Shen<sup>1,4</sup>, Junlin Xu<sup>2</sup>, Xiongfei Tian<sup>1</sup>, Fuxing Liu<sup>1</sup>, Juanjuan Wang<sup>1</sup>, Geng Tian<sup>3</sup>, Jialiang Yang<sup>3</sup>✉ & Liqian Zhou<sup>1</sup>✉

The outbreak of a novel febrile respiratory disease called COVID-19, caused by a newfound coronavirus SARS-CoV-2, has brought a worldwide attention. Prioritizing approved drugs is critical for quick clinical trials against COVID-19. In this study, we first manually curated three Virus-Drug Association (VDA) datasets. By incorporating VDAs with the similarity between drugs and that between viruses, we constructed a heterogeneous Virus-Drug network. A novel Random Walk with Restart method (VDA-RWR) was then developed to identify possible VDAs related to SARS-CoV-2. We compared VDA-RWR with three state-of-the-art association prediction models based on fivefold cross-validations (CVs) on viruses, drugs and virus-drug associations on three datasets. VDA-RWR obtained the best AUCs for the three fivefold CVs, significantly outperforming other methods. We found two small molecules coming together on the three datasets, that is, remdesivir and ribavirin. These two chemical agents have higher molecular binding energies of  $-7.0$  kcal/mol and  $-6.59$  kcal/mol with the domain bound structure of the human receptor angiotensin converting enzyme 2 (ACE2) and the SARS-CoV-2 spike protein, respectively. Interestingly, for the first time, experimental results suggested that navitoclax could be potentially applied to stop SARS-CoV-2 and remains to further validation.

In late December, 2019, there was an outbreak of a novel febrile respiratory illness (COVID-19) in Wuhan, Hubei in China<sup>1,2</sup>. The illness was caused by a novel coronavirus named SARS-CoV-2 by the World Health Organization (WHO) and can transmit from human to human<sup>2</sup>. As of 10 a.m. Cest time on October, 18, 2020, 40,118,333 cases of SARS-CoV-2 infection and 1,114,749 cases of SARS-CoV-2-caused death have been confirmed around the world<sup>3</sup>. From February, 2020, WHO is seeking U.S. \$675 million for COVID-19 preparedness to prevent human to human transmission<sup>4</sup>.

SARS-CoV-2 is a new human-infecting single-stranded RNA virus<sup>2</sup>. It is very similar to two coronaviruses: severe acute respiratory syndrome (SARS) coronavirus (SARS-CoV) and Middle East respiratory syndrome (MERS) coronavirus (MERS-CoV). In November, 2002, SARS first emerged in Guangdong, China, and resulted in 8,098 infection cases with a fatality rate of 9.6%<sup>1</sup>. In September, 2012, MERS was first found in humans in the Middle East and resulted in 2,465 laboratory-confirmed cases of infection with a fatality rate 34.4%<sup>5</sup>.

As SARS-CoV-2 is an emerging virus, no specific antiviral treatment has been developed<sup>6</sup>. Therefore, finding effective drug treatment options is urgently needed for combating SARS-CoV-2<sup>7</sup>. However, it seems unrealistic to test new drugs targeting SARS-CoV-2 within such limited time<sup>8</sup>. An efficient method is to screen possible drugs from available public data repositories containing FDA-approved compounds<sup>7,9</sup>. Under such situation, computational methods could be chosen to identify special antiviral drug candidates<sup>10–12</sup>.

Although little is known about SARS-CoV-2, its complete genome sequence suggests strong homology with SARS-CoV<sup>13</sup>. To identify possible antiviral drugs, in this study, we investigated the relationship between the complete genome sequences of viruses similar to SARS-CoV-2, the chemical structures of drugs, and Virus-Drug Association (VDA) network topology. We then developed a novel Random Walk with Restart method (VDA-RWR) to find possible VDAs related to SARS-CoV-2 by integrating the genome sequences and the chemical structures into a unified framework. We compared VDA-RWR with NGRHMDA<sup>14</sup>, SMiR-NBI<sup>15</sup> and

<sup>1</sup>School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China. <sup>2</sup>College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China. <sup>3</sup>Geneis (Beijing) Co. Ltd., Beijing 100102, China. <sup>4</sup>These authors contributed equally: Lihong Peng and Ling Shen. ✉email: yangjl@geneis.cn; zhoulq11@163.com

Method	Dataset 1	Dataset 2	Dataset 3
NGRHMDA	$\alpha=0.4, \beta=0.8$	$\alpha=0.6, \beta=0.9$	$\alpha=0.9, \beta=0.9$
LRLSHMDA	$\eta M=0.9, \eta D=0.3$	$\eta M=0.8, \eta D=0.1$	$\eta M=0.6, \eta D=0.1$
VDA-RWR	$r=0.7, \mu=0.9, \alpha=0.5$	$r=0.5, \mu=0.9, \alpha=0.9$	$r=0.7, \mu=0.9, \alpha=0.9$

**Table 1.** The optimal values of parameters in VDA-RWR, NGRHMDA and LRLSHMDA.

LRLSHMDA<sup>16</sup>. These three methods were applied to biological association prediction in other application areas and obtained better prediction performance. We found that remdesivir and ribavirin come together on three datasets.

Molecular docking is a key bioinformatics modeling tool for drug discovery and used to predict the “best-fit” intermolecular binding between a small molecule and a target or two proteins at the atomic level. It characterizes the behavior of ligands in the binding sites of target proteins as well as elucidates fundamental biochemical processes<sup>17</sup>. The docking process comprises two basic steps: predicting conformation, position, and orientation of ligands within the binding sites and ranking these conformations based on the binding affinity<sup>18</sup>. We used AutoDock<sup>19</sup>, a molecular docking software, to measure the molecular activities of the predicted two compounds (remdesivir and ribavirin) binding to the SARS-CoV-2 spike protein/human receptor angiotensin converting enzyme 2 (ACE2). The docking showed that remdesivir and ribavirin have higher binding energies of  $-7.0$  kcal/mol and  $-6.59$  kcal/mol with the structure of the spike protein receptor-binding domain bound to the ACE2 receptor, respectively.

## Results

**Experimental settings.** In this section, we conducted extensive experiments to investigate the performance of our proposed VDA-RWR method. For the VDA matrix  $Y_{n \times m}$  from  $n$  viruses and  $m$  drugs, fivefold Cross-Validations (CVs) were performed under the following three different experimental settings.

- Fivefold Cross Validation 1 (CV1): CV on viruses, that is, random rows in  $Y$  (i.e., viruses) were selected for testing.
- Fivefold Cross Validation 2 (CV2): CV on drugs, that is, random columns in  $Y$  (i.e., drugs) were selected for testing.
- Fivefold Cross Validation 3 (CV3): CV on virus-drug pairs, that is, random entries in  $Y$  (i.e., virus-drug pairs) were selected for testing.

Under CV1, in each round, 80% of rows in  $Y$  were used as training set and the remaining 20% of rows were used as test set. Under CV2, in each round, 80% of columns in  $Y$  were used as training set and the remaining 20% of columns were used as test set. Under CV3, in each round, 80% of entries in  $Y$  were used as training set and the remaining 20% of entries were test set. These three settings CV1, CV2, and CV3 specially refer to potential VDA identification for (1) new viruses (especially for SARS-CoV-2), (2) new drugs, and (3) new virus-drug pairs, respectively.

Parameters  $r$ ,  $\mu$ , and  $\alpha$  denote the global restart rate, the transition probability, and the weight between the virus network and the drug network, respectively. For these three parameters, we performed cross validations on the training set to find the optimal values. In addition, the iteration stopped when  $\|p_{t+1} - p_t\|_2 \leq 1e - 11$ . SMiR-NBI need not set the parameters. For the parameters in NGRHMDA and LRLSHMDA, we conducted grid search to find the optimal values. The detailed settings are shown on Table 1.

**Evaluation metrics.** Sensitivity, specificity, F1 score, accuracy and AUC were widely applied to evaluate the proposed methods. Sensitivity denotes the ratio of correctly predicted positive VDAs to all positive VDAs. Specificity is the ratio of correctly predicted negative VDAs to all negative VDAs (all the unknown associations were labeled as negative). F1 score denotes the harmonic mean of recall and precision. Accuracy represents the ratio of correctly predicted positive and negative VDAs to all positive and negative VDAs. We used these five metrics to evaluate the performance of VDA-RWR. They were defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{F1score} = \frac{2TP}{2TP + FP + FN} \quad (4)$$

Datasets	Methods	Sensitivity	Specificity	F1 score	Accuracy	AUC
Dataset 1	NGRHMDA	0.7278	0.3991	0.0643	0.4092	0.7026
	SMiR-NBI	<b>0.8086</b>	0.2164	0.0366	0.2296	0.5806
	LRLSHMDA	0.1299	0.6171	0.0084	0.6057	0.1844
	VDA-RWR	0.4977	<b>0.7959</b>	<b>0.1055</b>	<b>0.7905</b>	<b>0.8157</b>
Dataset 2	NGRHMDA	0.3987	0.5521	0.0329	0.5495	0.4301
	SMiR-NBI	<b>0.8238</b>	0.0949	0.0332	0.1087	0.4003
	LRLSHMDA	0.3507	0.4667	0.0179	0.4643	0.3173
	VDA-RWR	0.5106	<b>0.6832</b>	<b>0.0844</b>	<b>0.6801</b>	<b>0.6932</b>
Dataset 3	NGRHMDA	0.4435	0.4560	0.0232	0.4563	0.4058
	SMiR-NBI	<b>0.9124</b>	0.0459	0.0227	0.0567	0.4092
	LRLSHMDA	0.1801	0.5817	0.0074	0.5766	0.2920
	VDA-RWR	0.5270	<b>0.7025</b>	<b>0.0812</b>	<b>0.7006</b>	<b>0.7276</b>

**Table 2.** The performance comparison of four methods on three datasets under CV1. Bold values indicates the best values for the different methods under the same evaluation.

Datasets	Methods	Sensitivity	Specificity	F1 score	Accuracy	AUC
Dataset 1	NGRHMDA	0.6435	0.6719	0.0850	0.6713	0.8329
	SMiR-NBI	<b>0.8510</b>	0.1917	0.0393	0.2064	0.6021
	LRLSHMDA	0.7938	0.5773	0.1122	0.5820	0.8249
	VDA-RWR	0.5070	<b>0.8932</b>	<b>0.1294</b>	<b>0.8846</b>	<b>0.9182</b>
Dataset 2	NGRHMDA	0.4867	<b>0.8027</b>	0.0719	<b>0.7967</b>	0.8017
	SMiR-NBI	<b>0.9971</b>	0.0929	0.0404	0.1098	0.7205
	LRLSHMDA	0.7720	0.4166	0.0639	0.4232	0.7334
	VDA-RWR	0.5045	0.7981	<b>0.0814</b>	0.7926	<b>0.8025</b>
Dataset 3	NGRHMDA	0.4579	0.6785	0.0279	0.6758	0.6772
	SMiR-NBI	<b>0.9751</b>	0.0434	0.0243	0.0549	0.5665
	LRLSHMDA	0.7420	0.5264	0.0493	0.5290	0.7468
	VDA-RWR	0.5054	<b>0.8098</b>	<b>0.0628</b>	<b>0.8061</b>	<b>0.8168</b>

**Table 3.** The performance comparison of four methods on three datasets under CV2. Bold values indicates the best values for the different methods under the same evaluation.

where  $TP$ ,  $FP$ ,  $TN$  and  $FN$  represent true positive, false positive, true negative and false negative, respectively.

AUC is the average area under the receiver operating characteristics (ROC) curve. The curve can be plotted by the ratio of True Positive Rate (TPR) to False Positive Rate (FPR) according to different thresholds. TPR and FPR are defined via Eqs. (4–5).

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{T} \quad (5)$$

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{F} \quad (6)$$

For these five evaluation metrics, higher values represent better performance.

**Performance evaluation under three fivefold CVs.** We compared VDA-RWR with NGRHMDA<sup>14</sup>, SMiR-NBI<sup>15</sup> and LRLSHMDA<sup>16</sup>. NGRHMDA was presented to find potential microbe-disease associations by integrating neighbor-based collaborative filtering and graph-based scoring<sup>14</sup>. SMiR-NBI can comprehensively identify new pharmacogenomic biomarkers by constructing a heterogeneous network connecting genes, drugs, and miRNAs<sup>15</sup>. LRLSHMDA was applied to predict human microbe-disease associations based on Laplacian regularized least squares<sup>16</sup>. These three state-of-the-art approaches obtained good performance in their corresponding applications. We performed these four methods for 100 times on three different fivefold CV settings on three datasets. The final performance was averaged over the five rounds for 100 times. The results are shown in Tables 2, 3, and 4. The best results were shown in bold in each column.

On dataset 1 and dataset 3, VDA-RWR outperformed other three methods in terms of specificity, accuracy, F1 score and AUC under three CVs. On dataset 2, although the sensitivity of VDA-RWR was slightly lower, VDA-RWR computed better specificity, accuracy, F1 score and AUC under majority of conditions. The slight difference can be produced by different data structures. AUC is one more important evaluation metric compared to other four measurements. AUC = 0.5 represents random performance and AUC = 1 shows perfect performance.

Datasets	Methods	Sensitivity	Specificity	F1 score	Accuracy	AUC
Dataset 1	NGRHMDA	0.5783	0.5567	0.0615	0.5572	0.6459
	SMiR-NBI	<b>0.8331</b>	0.1936	0.0385	0.2079	0.5723
	LRLSHMDA	0.8034	0.5813	0.1119	0.5863	0.8403
	VDA-RWR	0.4824	<b>0.7831</b>	<b>0.1153</b>	<b>0.8278</b>	<b>0.8582</b>
Dataset 2	NGRHMDA	0.4544	0.3562	0.0218	0.3581	0.3011
	SMiR-NBI	<b>0.8349</b>	0.0942	0.0336	0.1080	0.4156
	LRLSHMDA	0.7838	0.4840	<b>0.0733</b>	0.4896	<b>0.8248</b>
	VDA-RWR	0.5022	<b>0.6643</b>	0.0574	<b>0.6613</b>	0.6675
Dataset 3	NGRHMDA	0.3582	0.4081	0.0119	0.4074	0.2554
	SMiR-NBI	<b>0.9230</b>	0.0427	0.0230	0.0536	0.4365
	LRLSHMDA	0.8124	0.5239	0.0552	0.5275	<b>0.8169</b>
	VDA-RWR	0.5053	<b>0.7057</b>	<b>0.0556</b>	<b>0.7032</b>	0.7123

**Table 4.** The performance comparison of four methods on three datasets under CV3. Bold values indicates the best values for the different methods under the same evaluation.

VDA-RWR obtained the best AUCs under majority of conditions. In general, VDA-RWR is proper to discover potential VDAs.

In addition, under CV1, VDA-RWR computed better specificity, accuracy, F1 score and AUC on the three datasets. This result showed that VDA-RWR can effectively find possible antiviral drugs for new viruses (for example, SARS-CoV-2). Under CV2, VDA-RWR outperformed other three methods in terms of specificity, accuracy, F1 score and AUC on dataset 1 and dataset 3. Although the sensitivity, specificity and accuracy values of VDA-RWR were slightly lower than other individual methods on dataset 2, it obtained the best F1 score and AUC. Thus AUC can identify potential viruses associated with new drugs. Under CV3, VDA-RWR calculated the best specificity, F1 score and accuracy. Figure 1 showed the AUC values of four methods under CV1, CV2, and CV3. The results demonstrated that VDA-RWR obtained relatively higher AUCs under three different CVs. It suggested that VDA-RWR could be used to infer potential VDAs.

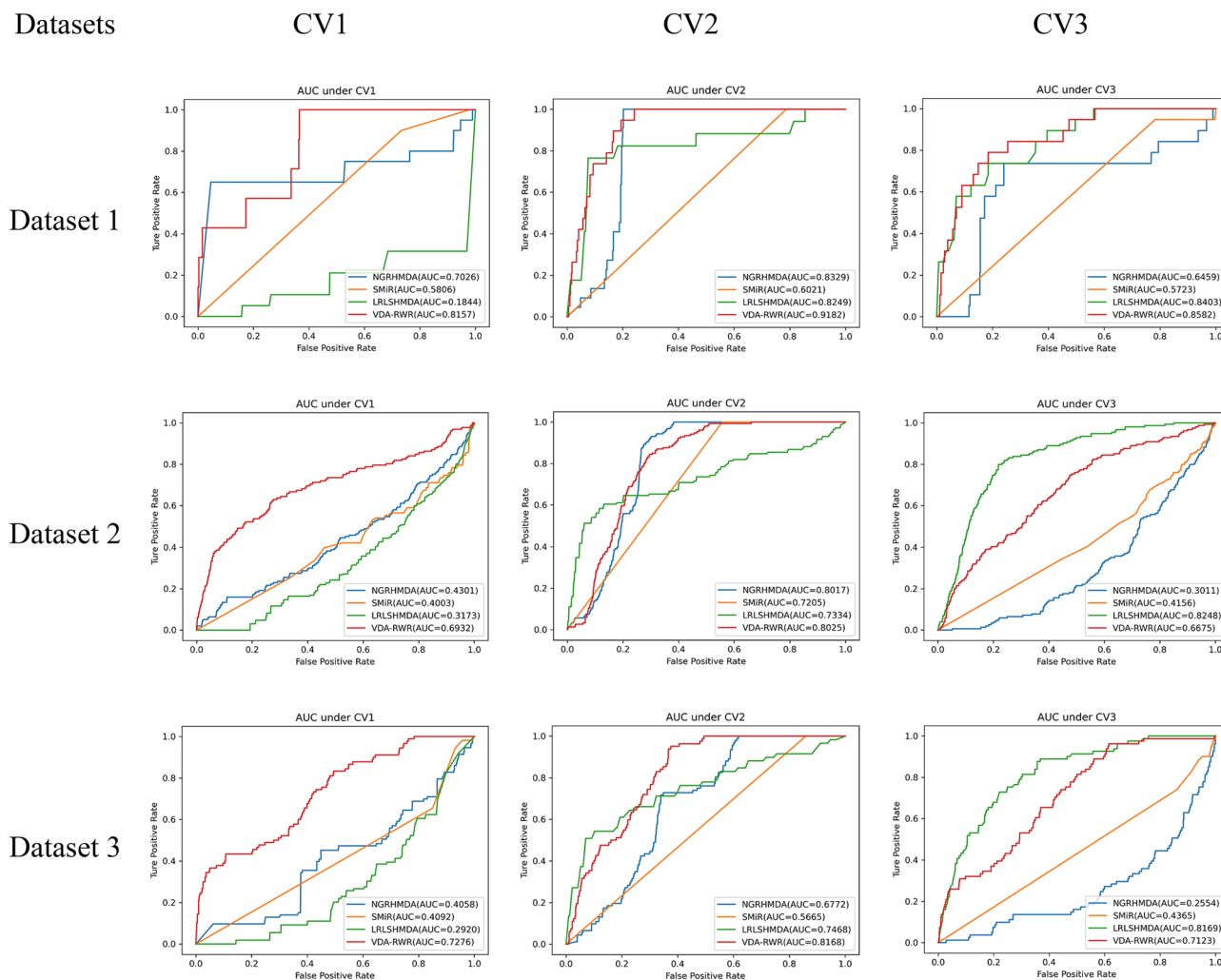
**Case study.** In this section, we want to find possible drugs for SARS-CoV-2 after verifying the performance of our proposed VDA-RWR method. We predicted the top 10 drugs with the highest association scores with SARS-CoV-2 on three datasets. The results were shown in Tables 5, 6, and 7, respectively. Among the predicted top 10 small molecules associated with SARS-CoV-2 on dataset 1, all drugs were supported by recent works. Among the predicted top 10 chemical agents related to SARS-CoV-2 on dataset 2, there were 9 VDAs validated by current literatures, that is, 90% chemical agents were reported. Among the predicted top 10 antiviral drugs against SARS-CoV-2 on dataset 3, all compounds were validated by recent publications.

The results on Tables 5, 6, and 7 showed that there were two FDA-approved drugs coming together on three datasets, that is, remdesivir and ribavirin. Remdesivir is a small molecular compound undergoing a clinical trial and shows superior antiviral activity against many RNA viruses including orthocoronavirinae, filoviridae, paramyxoviridae, and pneumoviridae<sup>20–22</sup>. Sheahan et al.<sup>17</sup> presented that it can improve pulmonary function and reduce severe lung pathology in mice. Similar to SARS-CoV-2, both Ebola virus (EBOV) and MERS-CoV may result in severe acute respiratory diseases. And remdesivir has been used as inhibitors of EBOV and MERS-CoV<sup>20,21</sup>. More importantly, an array of works have reported that remdesivir is highly effective in controlling SARS-CoV-2 infection and has been directly applied to the treatment of COVID-19<sup>6,7,9,23–28</sup>. Specially, on October 22, 2020, FDA approved remdesivir for use in adults, pediatric patients with age of 12 years, and older and weighing at least 40 kg<sup>29</sup>. All these results showed that remdesivir may be the best anti-SARS-CoV-2 drug.

Ribavirin is identified as another anti-SARS-CoV-2 drug with a higher association score. Huang et al.<sup>5</sup> found that 28 of 38 patients treated by ribavirin have been discharged. Zhang et al.<sup>30</sup> reported that a patient has been treated with antiviral drugs including ribavirin. Therefore, ribavirin may be applied to treat COVID-19 caused by SARS-CoV-2. Interestingly, for the first time, experimental results suggested that navitoclax could be potentially applied to stop SARS-COV-2. Navitoclax has been applied to boost the treatment and basic science of chronic lymphoid leukemia, hematological malignancies, non-Hodgkin's lymphoma, solid tumors, and EGFR activating mutation.

**Molecular docking.** The molecular docking between the above two antiviral drugs (remdesivir and ribavirin) and the spike protein and ACE2 are described in Table 8. The results showed that remdesivir and ribavirin have higher binding energies of  $-7.0$  kcal/mol and  $-6.59$  kcal/mol with the structure of the spike protein receptor-binding domain bound to the ACE2 receptor, respectively. The subfigure in each circle denotes the residues at the binding site of the spike protein/ACE2 and their corresponding orientations. For example, the amino acids K68 and Q493 were predicted to be the key residues for remdesivir binding to the SARS-CoV-2 spike protein/ACE2 while K353, R403, Q493 and G496 were predicted as the key residues for ribavirin binding to these two target proteins.

In Table 8, green denotes the structure of ACE2 and cyan denotes the SARS-CoV-2 spike protein in the figures of molecular docking.



**Figure 1.** The AUC values of VDA-RWR under different CVs on three datasets.

Rank	Drug	Evidence
1	Remdesivir	PMID: 32020029, 31996494, 32022370, 31971553, 32035018, 32035533, 32036774, 32194944, 32275812, 32145386, 32838064 <a href="https://doi.org/10.1101/2020.01.28.922922">https://doi.org/10.1101/2020.01.28.922922</a>
2	Oseltamivir	PMID: 32034637, 32127666
3	Ribavirin	PMID: 32034637, 32127666, 32227493, 26492219, 32771797
4	Zanamivir	PMID: 32511320
5	Presatovir	PMID: 32147628
6	Elvitegravir	PMID: 32147628
7	Zidovudine	PMID: 32568013
8	Emtricitabine	PMID: 32488835
9	Mycophenolic acid	PMID: 32579258
10	Chloroquine	PMID: 32020029, 32145363, 32074550, 32236562

**Table 5.** The predicted top 10 drugs associated with SARS-CoV-2.

### Discussion

Finding possible antiviral drugs against SARS-CoV-2 is extremely urgent with the rapid spread of COVID-19. However, it seems very difficult to design a novel drug for COVID-19 within a very short time. One of efficient ways is to identify new clues of the treatment from FDA-approved drugs.

In our proposed VDA-RWR method, we computed the association scores for each virus-drug pair to predict potential antiviral drugs against SARS-CoV-2 based on random walk with restart and biological information

Rank	Drug	Evidence
1	Favipiravir	PMID: 32346491, 32967849, 32972430
2	Remdesivir	PMID: 32020029, 31996494, 32022370, 31971553, 32035018, 32035533, 32036774, 32194944, 32275812, 32145386, 32838064 <a href="https://doi.org/10.1101/2020.01.28.922922">https://doi.org/10.1101/2020.01.28.922922</a>
3	Cidofovir	PMID: 32546018 <a href="https://doi.org/10.1007/s10067-020-05133-0">https://doi.org/10.1007/s10067-020-05133-0</a>
4	Galidesivir	PMID: 32711596
5	Niclosamide	PMID: 32125140, 32221153
6	Mycophenolic acid	PMID: 3257258
7	Itraconazole	<a href="https://doi.org/10.22541/au.159467021.16927198">https://doi.org/10.22541/au.159467021.16927198</a>
8	Brequinar	PMID: 32426387
9	Navitoclax	Unconfirmed
10	Ribavirin	PMID: 32034637, 32127666, 32227493, 26492219, 32771797

**Table 6.** The predicted top 10 drugs associated with SARS-CoV-2 on dataset 2.

Rank	Drug	Evidence
1	Nitazoxanide	PMID: 32127666, 32568620, 32448490
2	Ribavirin	PMID: 3203637, 32127666, 32227493, 26492219, 32771797
3	Chloroquine	PMID: 32020029, 32145363, 32074550, 32236562
4	Hexachlorophene	PMID: 15950190
5	Camostat	PMID: 32347443
6	Favipiravir	PMID: 32246834
7	Umifenovir	PMID: 32941741
8	Remdesivir	PMID: 32020029, 31996494, 32022370, 31971553, 32035018, 32035533, 32036774, 32194944, 32275812, 32145386, 32838064 <a href="https://doi.org/10.1101/2020.01.28.922922">https://doi.org/10.1101/2020.01.28.922922</a>
9	Amantadine	PMID: 32361028
10	Niclosamide	PMID: 32125140, 32221153

**Table 7.** The predicted top 10 drugs associated with SARS-CoV-2 on dataset 3.

of viruses and drugs. The originality of our proposed VDA-RWR method remains, constructing three small datasets and inferring possible antiviral chemical agents against SARS-CoV-2 from FDA-approved drugs. The comparative experiments showed better performance of the VDA-RWR method. Higher AUC values under three fivefold CVs on three datasets and molecular binding energies indicated that the selected small molecules are likely to be used to stop the transmission of COVID-19.

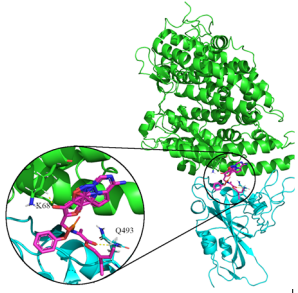
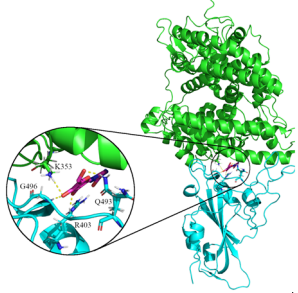
VDA-RWR can obtain superior performance under the three fivefold CVs on three datasets. This observation may be attributed to random walk with restart, a state-of-the-art model that can randomly walk on the heterogeneous virus-drug network and effectively compute association score for each virus-drug pair. More importantly, VDA-RWR integrated various biological information including the complete genome sequences of viruses and chemical structures of chemical agents.

The proposed VDA-RWR method is also helpful in design and interpretation of pharmacological experiment related to COVID-19. More importantly, VDA-RWR can be further applied to predict antiviral drugs against novel viruses without any associated chemical agents.

## Methods

**Virus-drug association data.** *Dataset 1.* Virus data. We considered 11 viruses similar to SARS-CoV-2. These viruses include influenza A viruses including A-H1N1<sup>32</sup>, A-H5N1<sup>33</sup>, and A-H7N9<sup>34</sup>, chronic hepatitis C virus (HCV)<sup>35</sup>, human immunodeficiency virus type 1 (HIV-1)<sup>36</sup>, human immunodeficiency virus type 2 (HIV-2)<sup>37</sup>, hendra virus<sup>38</sup>, human cytomegalovirus<sup>39</sup>, MERS-CoV<sup>40</sup>, respiratory syncytial virus<sup>41</sup> and SARS-CoV<sup>42</sup>. The complete genome sequences of these viruses are downloaded from the NCBI database<sup>43</sup>. We used MAFFT<sup>44</sup> (<https://mafft.cbrc.jp/alignment/software/>, version 7, open source license: GPL or BSD), a multiple sequence alignment tool, to compute virus-virus sequence similarity matrix  $S_v$ . All parameters were set as the default values provided by MAFFT.

**Drug data.** We manually curated drugs associated with these 11 viruses from the DrugBank<sup>45</sup> and NCBI<sup>43</sup> databases and published literatures reported by the PubMed database<sup>46</sup> and collected 78 small molecules after removing macromolecules. Based on the assumption that two drugs are more similar if they share more chemi-

Ligand	Molecular formula	Molecular docking	Binding energy (kcal/mol)	Binding sites	Distance(Å)
Remdesivir	$C_{27}H_{35}N_6O_8P$		- 7.0	K68	2.0
				Q493	2.3
Ribavirin	$C_{18}H_{26}ClN_3$		- 6.59	K353	2.2
				R403	2.1
				Q493	2.0
				G496	1.9

**Table 8.** Molecular docking between remdesivir and ribavirin and the SARS-CoV-2 spike (S) protein/ACE2.

Datasets	Viruses	Drugs	VDAs
Dataset 1	12	78	96
Dataset 2	69	128	770
Dataset 3	34	203	407

**Table 9.** Statistics for the virus-drug association networks.

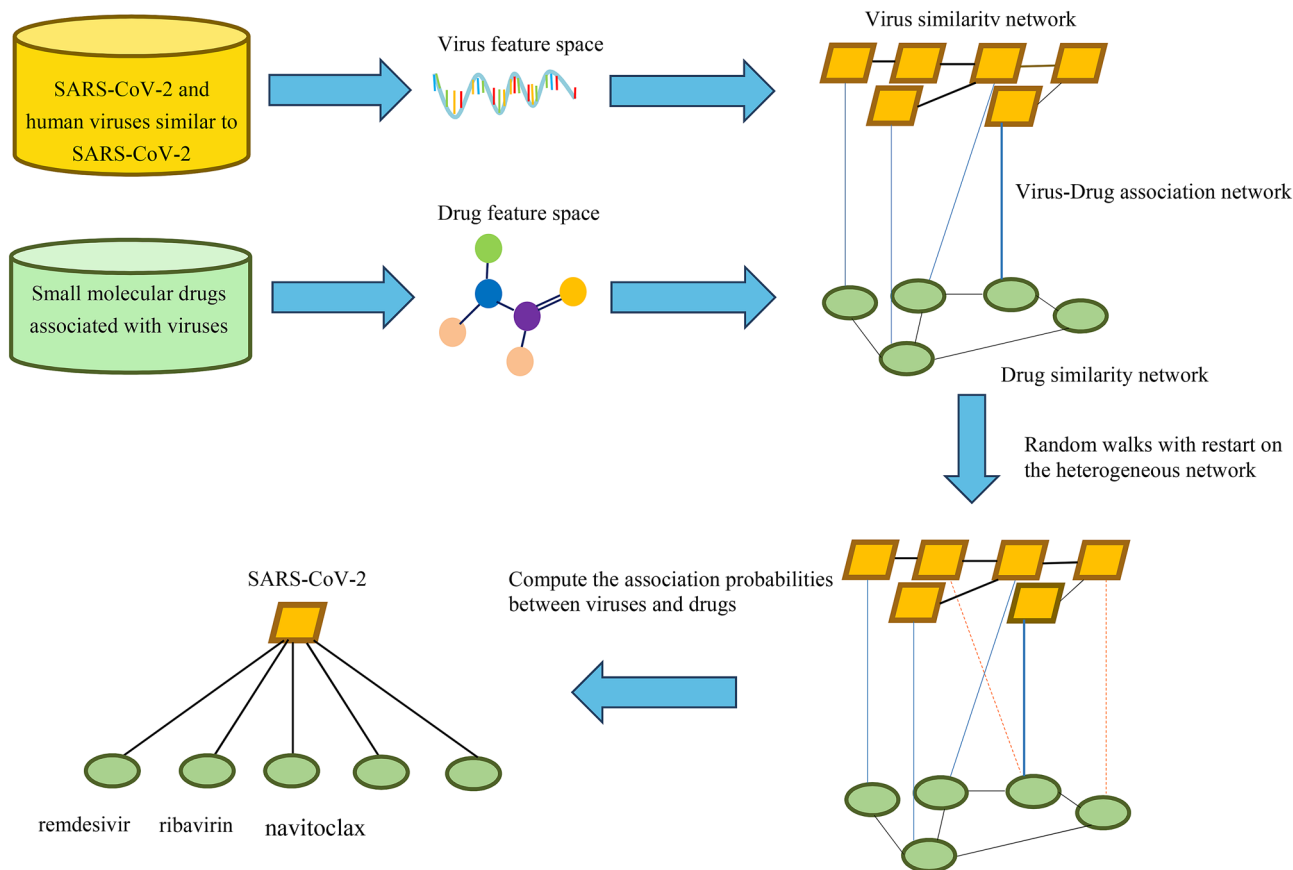
cal substructures, drug-drug similarity can be computed. Extended connectivity fingerprints (ECFPs)<sup>47</sup> are circular fingerprints and developed for structure-activity modeling and molecular feature description. We used RDKit<sup>48</sup> (<https://github.com/rdkit/rdkit>, releases 131, open source license: BSD), an open-source cheminformatics software, to compute ECFPs of drugs with a radius of 2. Drug-drug chemical structure similarity matrix  $S_d$  can be computed by the ECFPs of drugs.

**VDAs.** We searched the publicly available repositories including the DrugBank<sup>45</sup> and NCBI<sup>43</sup> databases and publications reported by the PubMed database<sup>46</sup>. At the time of writing, we obtained 96 virus-drug associations (VDAs) between 11 viruses and 78 drugs. We described A-H1N1<sup>32</sup>, A-H5N1<sup>33</sup>, and A-H7N9<sup>34</sup> as three viruses although they belong to influenza A for the sake of description.

**Dataset 2.** The DrugVirus.info database<sup>49</sup> (<https://drugvirus.info/>) provided various VDA-related resources. We obtained 770 VDAs from 69 viruses and 128 drugs after removing the viruses whose complete genome sequences are unknown from the database. The chemical structure of drugs and the complete genome sequences of viruses were downloaded from the DrugBank database and the NCBI database, respectively. Similar to dataset 1, we used RDKit and MAFFT to calculate drug similarity and virus similarity.

**Dataset 3.** We retrieved 407 VDAs from 34 viruses and 203 drugs by searching documents related to viruses and drugs based on text mining techniques. Similar to dataset 1, we computed drug similarity matrix and virus similarity matrix. The details of three datasets are shown in Table 9.

In this study, the set of known VDAs was considered as the 'gold standard' dataset and was applied to evaluate the performance of our proposed VDA-RWR method. We described the known VDAs as a matrix  $Y$ :



**Figure 2.** Flowchart of the VDA-RWR method based on the genome sequences of viruses, the chemical structures of drugs, and random walk with restart on the heterogeneous network.

$$Y_{ij} = \begin{cases} 1 & \text{if } v_i \text{ associates with } d_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

where  $v_i$  and  $d_j$  represent the  $i$  th virus and  $j$  th drug, respectively.

**The VDA-RWR method.** Inspired by the method provided by Valdeolivas et al.<sup>50</sup>, we developed a VDA prediction method based on Random Walk with Restart on the heterogeneous network (VDA-RWR). The proposed VDA-RWR method comprised two steps. First, a random walk-based model integrating various biological data was learned to explain the constructed ‘gold standard’ dataset. Second, this model was used to find potential VDAs for viruses and drugs absent from the ‘gold standard’ dataset. The details are shown Fig. 2.

We first considered virus-virus similarity graph  $G_v$ , drug-drug similarity graph  $G_d$ , and VDA graph  $G_a$ , which formed a heterogeneous network. We defined  $S_v(n \times n)$ ,  $S_d(m \times m)$ , and  $Y(n \times m)$  as their corresponding adjacency matrices. The adjacency matrix of the heterogeneous network can be denoted as:  $W = \begin{bmatrix} S_v & Y \\ Y^T & S_d \end{bmatrix}$ , where  $Y^T$  denoted the transpose of the VDA matrix  $Y$ .

We then calculated the transition probabilities of random walk with restart on the heterogeneous network. Suppose  $W = \begin{bmatrix} W_{vv} & W_{vd} \\ W_{dv} & W_{dd} \end{bmatrix}$  represented the matrix of transitions on the heterogeneous network, where  $W_{vv}/W_{dd}$  denoted the walk within the virus/drug network,  $W_{vd}/W_{dv}$  described the jump from the virus/drug network to the drug/virus network. Given the probability  $\mu$  of jumping from the virus/drug network to the drug/virus network, the transition probability from virus  $v_i$  to virus  $v_j$  was defined as

$$W_{vv}(i, j) = \begin{cases} \frac{S_v(i, j)}{\sum_{k=1}^n S_v(i, k)} & \text{if } \sum_{k=1}^m Y(i, k) = 0 \\ (1-\mu) \frac{S_v(i, j)}{\sum_{k=1}^n S_v(i, k)} & \text{otherwise} \end{cases} \quad (8)$$

The transition probability from virus  $v_i$  to drug  $d_j$  was defined as



$$W_{vd}(i, j) = \begin{cases} \frac{\mu Y(i, j)}{\sum_{k=1}^m Y(i, k)} & \text{if } \sum_{k=1}^m Y(i, k) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

The transition probability from drug  $d_i$  to drug  $d_j$  was defined as

$$W_{dd}(i, j) = \begin{cases} \frac{S_d(i, j)}{\sum_{k=1}^m S_d(i, k)} & \text{if } \sum_{k=1}^n Y(k, i) = 0 \\ \frac{(1-\mu)S_d(i, j)}{\sum_{k=1}^m S_d(i, k)} & \text{otherwise} \end{cases} \quad (10)$$

The transition probability from drug  $d_i$  to virus  $v_j$  was defined as

$$W_{dv}(i, j) = \begin{cases} \frac{\mu Y(j, i)}{\sum_{k=1}^n Y(k, i)} & \text{if } \sum_{k=1}^n Y(k, i) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

For a given virus/drug, the particle can either jump between graphs or stay in the current graph with a defined probability  $r \in (0, 1)$ . Therefore, we finally defined the random walk with a restart probability  $r$  as:

$$p_{t+1} = rWp_t + (1-r)p_0 \quad (12)$$

where  $p_t$  represented the computed association probability at the  $t$ -th step random walk. We defined the initial probability as:  $p_0 = \begin{bmatrix} \alpha u_0 \\ (1-\alpha)t_0 \end{bmatrix}$ , where  $u_0$  and  $t_0$  denoted the initial probability on the drug network and the virus network, respectively.

If we tend to identify possible drugs for a given virus  $v_i$ , it is considered as a seed node in the virus network. Here,  $v_i$  was assigned as 1 and other nodes as 0, constructing the initial probability of the virus network  $t_0$ . All nodes in the drug network  $u_0$  were assigned as equal probabilities with the sum of 1. For example, to find potential antiviral drugs against SARS-CoV-2, we set SARS-CoV-2 as a seed node, and all drugs in the drug network were assigned as the same probabilities with the values of  $1/m$ . The parameter  $\alpha$  was used to control the weight of the virus network and the drug network. In addition, a virus is new if it does not associate with any drugs, and a drug is new if it is not applied to any viruses.

**Molecular docking.** Molecular docking technique was applied to compute the intermolecular binding abilities between the predicted anti-SARS-CoV-2 drugs and the SARS-CoV-2 spike protein/human ACE2. The chemical structures of drugs were downloaded in the form of the PDB format files from the DrugBank database. We used AutoDockTools to convert these PDB files into pdbqt files needed by AutoDock4. The structures of SARS-CoV-2 spike receptor-binding domain bound with ACE2 (PDB ID: 6M0J) were downloaded from the RCSB Protein Data Bank<sup>51</sup>. The spike protein and ACE2 were used as receptors, and the predicted anti-SARS-CoV-2 drugs were used as ligands for the molecular docking.

We first removed solvent and organic compounds and preprocessed the receptor proteins based on PyMOL<sup>31</sup> (<https://github.com/schrodinger/pymol-open-source>, release 2.4.0, open source license: BSD-like). The receptors' atoms were assigned the AD4 type and Gasteiger charges were considered before docking. Molecular docking software, AutoDock<sup>19</sup> (<http://autodock.scripps.edu/>, AutoDock 4.2.6, open source license: GPL), was then used to conduct molecular docking. The binding pocket was defined by AutoGrid4, the grid size was set to  $82 \times 154 \times 84$  with a spacing of  $0.375 \text{ \AA}$ , and the grid center was placed at the area of SARS-CoV-2 spike receptor-binding domain bounding with ACE2 ( $x = -36.884$ ,  $y = 29.245$ ,  $z = -0.005$ ). The LGA (Lamarckian genetic algorithm) with default parameter provided by AutoDock4 was used as the search method. The docking contained two main processes: computation of conformation, position, and orientation of ligands within the binding sites and ranking of these conformations based on the binding affinities<sup>18</sup>.

## Conclusion

To find potential antiviral drugs, in this study, we integrated the complete genome sequences of viruses, the chemical structures of drugs, and the VDA network. We then developed a VDA prediction method based on random walk with restart on the heterogeneous network. The results suggested that remdesivir and ribavirin may be applied to the treatment of COVID-19. In the emergency situation, this study focused more on finding antiviral drugs. In the future, we will further integrate more biological data and design more powerful models to improve the accuracy of VDA identification. We hope that our proposed VDA-RWR method could help the screening of drugs for preventing COVID-19.

## Data availability

Source codes and datasets are freely available for download at <https://github.com/plhnu/VDA-RWR/>.

Received: 18 July 2020; Accepted: 18 December 2020

Published online: 18 March 2021

## References

- Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health: The latest 2019 novel coronavirus outbreak in Wuhan, China. *Int. J. Infect. Dis.* **91**, 264–266 (2020).

2. Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* **395**, 565–574 (2020).
3. WHO. Coronavirus disease 2019 (COVID-19) Situation Report- COVID-19 Weekly Epidemiological Update. (2020). *Organ. WHO* (2020). (<https://www.who.int/docs/default-source/coronavirus/situation-reports/20201020-weekly-epi-update-10.pdf>). Accessed 20 Oct 2020.
4. Organization, W. H. *US \$675 Million Needed for New Coronavirus Preparedness and Response Global Plan [Internet]*. Geneva: World Health Organization; 2020 [cited 2020 Feb 5]. (2020).
5. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
6. Wang, M. *et al.* Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Res.* **30**, 269–271 (2020).
7. Lu, H. Drug treatment options for the 2019-new coronavirus (2019-nCoV). *Biosci. Trends* **14**, 69–71 (2020).
8. Chen, H. & Du, Q. Potential natural compounds for preventing SARS-CoV-2 (2019-nCoV) infection. *Preprints* <https://doi.org/10.20944/preprints202001.0358.v3> (2020).
9. Li, Y. *et al.* Therapeutic drugs targeting 2019-nCoV main protease by high-throughput screening. *bioRxiv* <https://doi.org/10.1101/2020.01.28.922922> (2020).
10. Kumar, S. Drug and vaccine design against novel coronavirus (2019-nCoV) spike protein through computational approach. *Preprints* <https://doi.org/10.20944/preprints202002.0071.v1> (2020).
11. Zhang, H. *et al.* Deep learning based drug screening for novel coronavirus 2019-nCoV. *Interdiscip. Sci. Comput. Life Sci.* <https://doi.org/10.20944/preprints202002.0061.v1> (2020).
12. Beck, B. R., Shin, B., Choi, Y., Park, S. & Kang, K. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Comput. Struct. Biotechnol. J.* **18**, 784–790 (2020).
13. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
14. Huang, Y.-A. *et al.* Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. *J. Transl. Med.* **15**, 1–11 (2017).
15. Li, J. *et al.* Network-based identification of microRNAs as potential pharmacogenomic biomarkers for anticancer drugs. *Oncotarget* **7**, 45584–45596 (2016).
16. Wang, F. *et al.* LRLSHMDA: Laplacian regularized least squares for human microbe–disease association prediction. *Sci. Rep.* **7**, 1–11 (2017).
17. McConkey, B. J., Sobolev, V. & Edelman, M. The performance of current methods in ligand–protein docking. *Curr. Sci.* **1**, 845–856 (2002).
18. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular docking: A powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **7**, 146–157 (2011).
19. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **30**, 2785–2791 (2009).
20. Tchesnokov, E. P., Feng, J. Y., Porter, D. P. & Götte, M. Mechanism of inhibition of ebola virus RNA-dependent RNA polymerase by remdesivir. *Viruses* **11**, 326 (2019).
21. Sheahan, T. P. *et al.* Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV. *Nat. Commun.* **11**, 1–14 (2020).
22. Brown, A. J. *et al.* Broad spectrum antiviral remdesivir inhibits human endemic and zoonotic deltacoronaviruses with a highly divergent RNA dependent RNA polymerase. *Antiviral Res.* **169**, 104541 (2019).
23. Morse, J. S., Lalonde, T., Xu, S. & Liu, W. R. Learning from the past: Possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV. *ChemBioChem* **21**, 730–738 (2020).
24. Paules, C. I., Marston, H. D. & Fauci, A. S. Coronavirus infections: more than just the common cold. *JAMA* **323**, 707–708 (2020).
25. Chen, Y. W., Yiu, C.-P. & Wong, K.-Y. Prediction of the 2019-nCoV 3C-like protease (3CLpro) structure: Virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *ChemRxiv* <https://doi.org/10.26434/chemrxiv.11831103> (2020).
26. Zumla, A., Hui, D. S., Azhar, E. I., Memish, Z. A. & Maeurer, M. Reducing mortality from 2019-nCoV: Host-directed therapies should be an option. *The Lancet* **395**, 35–36 (2020).
27. Wang, F.-S. & Zhang, C. What to do next to control the 2019-nCoV epidemic?. *The Lancet* **395**, 391–393 (2020).
28. Malik, Y. S. *et al.* Emerging novel coronavirus (2019-nCoV)—current scenario, evolutionary perspective based on genome analysis and recent developments. *Vet. Q.* **40**, 68–76 (2020).
29. FDA. FDA's approval of Veklury (remdesivir) for the treatment of COVID-19—The Science of Safety and Effectiveness. <https://www.fda.gov/drugs/drug-safety-and-availability/fdas-approval-veklury-remdesivir-treatment-covid-19-science-safety-and-effectiveness>, October 22, 2020. Accessed 24 Oct 2020.
30. Zhang, Z. *et al.* Clinical features and treatment of 2019-nCoV pneumonia patients in Wuhan: Report of a couple cases. *Viol. Sin.* **35**, 330–336 (2020).
31. Schrodinger, L. The PyMOL molecular graphics system. *Version 1*, (2010).
32. Fraser, C. *et al.* Pandemic potential of a strain of influenza A (H1N1): EARLY FINDINGS. *Science* **324**, 1557–1561 (2009).
33. Claas, E. C. *et al.* Human influenza A H5N1 virus related to a highly pathogenic avian influenza virus. *The Lancet* **351**, 472–477 (1998).
34. Gao, R. *et al.* Human infection with a novel avian-origin influenza A (H7N9) virus. *N. Engl. J. Med.* **368**, 1888–1897 (2013).
35. Fried, M. W. *et al.* Peginterferon Alfa-2a plus ribavirin for chronic hepatitis C virus infection. *N. Engl. J. Med.* **347**, 975–982 (2002).
36. Navia, M. A. *et al.* Three-dimensional structure of aspartyl protease from human immunodeficiency virus HIV-1. *Nature* **337**, 615–620 (1989).
37. Mörner, A. *et al.* Primary human immunodeficiency virus type 2 (HIV-2) isolates, like HIV-1 isolates, frequently use CCR5 but show promiscuity in coreceptor usage. *J. Virol.* **73**, 2343–2349 (1999).
38. Young, P. L., Halpin, K., Mackenzie, J. S. & Field, H. E. Isolation of hendra virus from pteropid bats: A natural reservoir of Hendra virus. *J. Gen. Virol.* **81**, 1927 (2000).
39. Chee, M. S. *et al.* Analysis of the protein-coding content of the sequence of human cytomegalovirus strain AD169. In *Cytomegaloviruses* (ed. McDougall, J. K.) 125–169 (Springer, New York, 1990).
40. de Groot, R. J. *et al.* Commentary: Middle east respiratory syndrome coronavirus (MERS-CoV): Announcement of the coronavirus study group. *J. Virol.* **87**, 7790–7792 (2013).
41. Mazur, N. I. *et al.* The respiratory syncytial virus vaccine landscape: Lessons from the graveyard and promising candidates. *Lancet Infect. Dis.* **18**, 295–311 (2018).
42. Bosch, B. J. *et al.* Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides. *Proc. Natl. Acad. Sci.* **101**, 8455–8460 (2004).
43. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **48**, D9 (2020).
44. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
45. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

46. Canese, K. & Weis, S. PubMed: the bibliographic database. in *The NCBI Handbook [Internet]. 2nd edition* (National Center for Biotechnology Information (US), 2013).
47. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
48. Landrum, G., et al. *RDKit: Open-Source Cheminformatics Software.* (2016).
49. Andersen, P. I. et al. Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int. J. Infect. Dis.* **93**, 268–276 (2020).
50. Valdeolivas, A. et al. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* **35**, 497–505 (2019).
51. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

## Acknowledgements

We are thankful for help from Guangyi Liu, Ming Kuang, and Longjie Liao from Hunan University of Technology, and Ruyi Dong, Lebin Liang, Qinqin Lu, and Jidong Lang from Geneis (Beijing) Co. Ltd. We would like to thank all authors of the cited references.

## Author contributions

L.-H.P. and L.S. contributed equally to this work. L.-H.P., J.-L.X., J.-L.Y., and L.-Q.Z. conceived the study, designed the schedule, and analyzed the data. F.-X.L. screened the viruses similar to SARS-CoV-2, X.-F.T. downloaded the genome sequences of viruses and computed virus similarity matrix, L.S. computed drug similarity matrix, L.S., X.-F.T., F.-X.L., and J.-J.W. constructed VDA network, L.S. run random walk algorithm, L.-H.P., G.T., and L.-Q.Z. wrote the paper, J.-L.Y. revised the original draft. All authors read and approved the final manuscript.

## Funding

This research was funded by National Natural Science Foundation of China (Grant 61803151) and Natural Science Foundation of Hunan province (Grant 2018JJ3570, 2018JJ2461).

## Competing interests

Authors Geng Tian and Jialiang Yang were employed by the company Geneis (Beijing) Co. Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be constructed as a potential conflict of interest.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-83737-5>.

**Correspondence** and requests for materials should be addressed to J.Y. or L.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021