Briefings in Bioinformatics, 00(00), 2020, 1–12

doi: 10.1093/bib/bbaa341
Problem Solving Protocol

iDMer: an integrative and mechanism-driven response system for identifying compound interventions for sudden virus outbreak

Zhiting Wei†, Yuli Gao†, Fangliangzi Meng†, Xin Chen†, Yukang Gong, Chenyu Zhu, Bin Ju, Chao Zhang, Zhongmin Liu and Qi Liu

Corresponding authors: Chao Zhang, Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China. Tel.: +86-021-65986676; E-mail: zhangchao@tongji.edu.cn; Zhongmin Liu, Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University School of Medicine, Shanghai 200120, China. Tel.:+86-021-38804518; E-mail: liu.zhongmin@tongji.edu.cn; Qi Liu, Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China. Tel.: +86-021-65980296; E-mail: qiliu@tongji.edu.cn
†These authors contribute equally to this work.

Abstract

Emerging viral infections seriously threaten human health globally. Several challenges exist in identifying effective compounds against viral infections: (1) at the initial stage of a new virus outbreak, little information, except for its genome information, may be available; (2) although the identified compounds may be effective, they may be toxic *in vivo* and (3) cytokine release syndrome (CRS) triggered by viral infections is the primary cause of mortality. Currently, an integrative tool that takes all those aspects into consideration for identifying effective compounds to prevent viral infections is absent. In this study, we developed iDMer, as an integrative and mechanism-driven response system for addressing these challenges during the sudden virus outbreaks. iDMer comprises three mechanism-driven compound identification modules, that is, a virus-host interaction-oriented module, an autophagy-oriented module and a CRS-oriented module. As a one-stop integrative platform, iDMer incorporates compound toxicity evaluation and compound combination identification for virus treatment with clear mechanisms. iDMer was successfully tested on five viruses, including the current severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Our results indicated that, for all five tested viruses, compounds that were reported in the literature or experimentally validated for virus treatment were enriched at the top, demonstrating the generalized effectiveness of iDMer. Finally, we demonstrated that combinations of the individual modules successfully identified combinations of compounds effective for virus intervention with clear mechanisms.

Key words: virus; SARS-CoV-2; epidemic response system; compound identification; virus host interaction

Zhiting Wei is a PhD candidate majoring in drug discovery and precision medicine at the School of Life Sciences and Technology, Tongji University, China.
Yuli Gao is a PhD candidate majoring in machine learning and biological data mining at the School of Life Sciences and Technology, Tongji University, China.
Fangliangzi Meng is a PhD candidate majoring in precision medicine at the School of Life Sciences and Technology, Tongji University, China.
Xin Chen is a Princeton High School student and an internship research student at the Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China.
Yukang Gong is a bioinformatics post-doctoral fellow majoring in drug design and machine learning.
Chenyu Zhu is a PhD candidate majoring in drug discovery and precision medicine at the School of Life Sciences and Technology, Tongji University, China.
Bin Ju is a PhD at the ShuLan (Hangzhou) Hospital affiliated to the Zhejiang Shuren University Shulan International Medical College, majoring in pharmacology analysis.
Chao Zhang is a professor at the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China. He directs the Biological Laboratory.
Zhongmin Liu is a MD at the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Tongji University School of Medicine, majoring in cardiovascular diseases.
Qi Liu is a professor at the Translational Medical Center for Stem Cell Therapy and Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China. He directs the Biological and Medical Big data Mining Laboratory.
Submitted: 30 July 2020; **Received (in revised form):** 11 September 2020

Introduction

Emerging viral infections seriously threaten human health throughout the world. In the past two decades, highly pathogenic human coronaviruses (HCoVs), including SARS-CoV and MERS-CoV, have led to global pandemics with high morbidity and mortality [1]. A third pathogenic HCoV, SARS-CoV-2, recently emerged and has caused considerable damage to human health and social economics globally [1]. The development of antiviral vaccines and novel drugs, however, requires long-term discovery and clinical validation. In addition, little information other than the viral genome information is generally available at the initial stage of the virus outbreak. Therefore, the development of a quick viral infections response system based only on the sequenced viral genome information is highly demanded. Such a response system would allow for the rapid identification of drug interventions for sudden virus outbreaks.

To this end, we present iDMer, an integrative and mechanism-driven response system for identifying candidate compound interventions for sudden virus outbreaks. The main focus of iDMer is to present a mechanism-clear and effective virus response and compound identification pipeline rather than the identified compounds themselves. iDMer can be used to identify compound interventions for any virus whose genome information is known. It will be extremely helpful at the initial stage of a new virus outbreak since little information about the virus except for its genome information is available. The main advantages of iDMer presented in this study are: (1) the development of a rapid response system during a new virus outbreak. This work is different from the existing studies which are focused on the prediction and validation of the identified compounds for SARS-CoV-2. (2) iDMer is integrative and mechanism-clear, which covers three modules with clear mechanisms for individual compound identification, toxicity evaluation and combination prioritization. It comprises three mechanism-driven compound identification modules (Figure 1), including (1) a virus–host interaction-oriented module that is designed to identify specific antiviral compounds against the test virus by interfering with their specific virus–host interactions [2]; (2) an autophagy-oriented module that is general to any virus by identifying compounds that activate cell autophagy to treat the virus [3, 4] and (3) a cytokine release syndrome (CRS)-oriented module that is designed to identify general anti-inflammatory compounds that will ameliorate the CRS triggered by viral infections [5, 6] in which CRS is the primary cause of mortality, such as in patients infected with SARS-CoV, MERS-CoV and SARS-CoV-2 [5]. In summary, virus–host interaction-oriented module and autophagy-oriented module are designed with different mechanisms to identify effective antiviral compounds. CRS-oriented module is designed to dampen the uncontrolled inflammatory response leading to shock and tissue damage in the lung and liver. For all three modules, the Connectivity Map (CMap) data source is applied to prioritize compounds, as shown in Figure 1 (see the 'Materials and methods' section) [7, 8]. It is reported that combinations of antiviral and anti-inflammatory drugs are more effective than individual drugs to reduce virus infection-related inflammation [6]. Therefore, the virus–host interaction-oriented module and autophagy-oriented module are selected and combined with the CRS-oriented module to identify compound combinations without antagonistic effects.

Another important issue considered in the development of iDMer was compound toxicity evaluation. Although the identified compounds may be effective, they may be toxic *in vivo*. Therefore, iDMer performs a compound toxicity evaluation by investigating the compound structure information related to certain toxicological end points. Identified compounds with high potential toxicity are filtered.

Finally, to evaluate the rationale and effectiveness of iDMer, we applied it to four historically spread viruses, that is, MERS-CoV, SARS-CoV, Ebola virus and HIV, as well as the current SARS-CoV-2. Our results indicated that for all five tested viruses, compounds that were reported in the literature or experimentally validated ranked high and were enriched at the top, demonstrating the generalized effectiveness of iDMer for identifying compound interventions or known drugs that can be repurposed for virus treatment with clear mechanisms. Furthermore, combined treatments with anti-inflammatory and antiviral drugs are also recommended by iDMer and are expected to be more effective than individual treatments by increasing the therapeutic efficacy and reducing virus infection-related inflammation.

Materials and methods

Investigation of the CMap for antiviral compound identification

Utility of CMap

The CMap project was introduced by Molecules *et al*., in 2006, as a potential resource for connecting compounds and diseases [8]. The CMap reference database (Touchstone) contains over 1 million compound signatures, which are gene expression profiles obtained from the treatment of a variety of cell types with perturbagens that span a range of small molecule compounds [7]. The fundamental unit of CMap analysis is the disease signature, which comprises a set of genes carrying a sign indicating whether the gene is upregulated or downregulated. The disease signature is queried against compound signatures in the CMap reference database by gene set enrichment analysis (GSEA) [9]. The GSEA assesses the degree to which the upregulated genes appear toward the top of the compound signature and the downregulated genes appear toward the bottom of the compound signature (positive CMap score), or vice-versa (negative CMap score). The application of CMap with a query results in a list of compounds ordered by their similarities against the reference.

The query inputs of the CMap are upregulated or downregulated genes. As for the antiviral compound identification, the incorporation of virus-infectious host gene expression profiles should be carefully investigated. In our study, we performed a comprehensive transcriptomic analysis of the tested virus using RNA-sequencing (RNA-Seq) data of the virus-infected host cells. Our results indicated that such transcriptomic profiles of virus-infected host cells cannot be directly incorporated in CMap for antiviral compound identification, although they have been applied in previous studies, because they produce certain artifacts in the compound identification (see details in 'Results' section).

RNA-Seq data preprocessing for virus-infected host cells

RNA-Seq data of MERS-CoV-, SARS-CoV- and SARS-CoV-2-infected host cells and their mock-treated controls were downloaded from NCBI Gene Expression Omnibus (GEO) [10]. Fastq files were mapped to the human reference genome (hg38) using STAR [11]. We utilized featureCounts [12] to quantify the gene expression level following differential expression analysis using the edgeR R package [13]. Our differential analysis was performed by matching each experimental condition with the

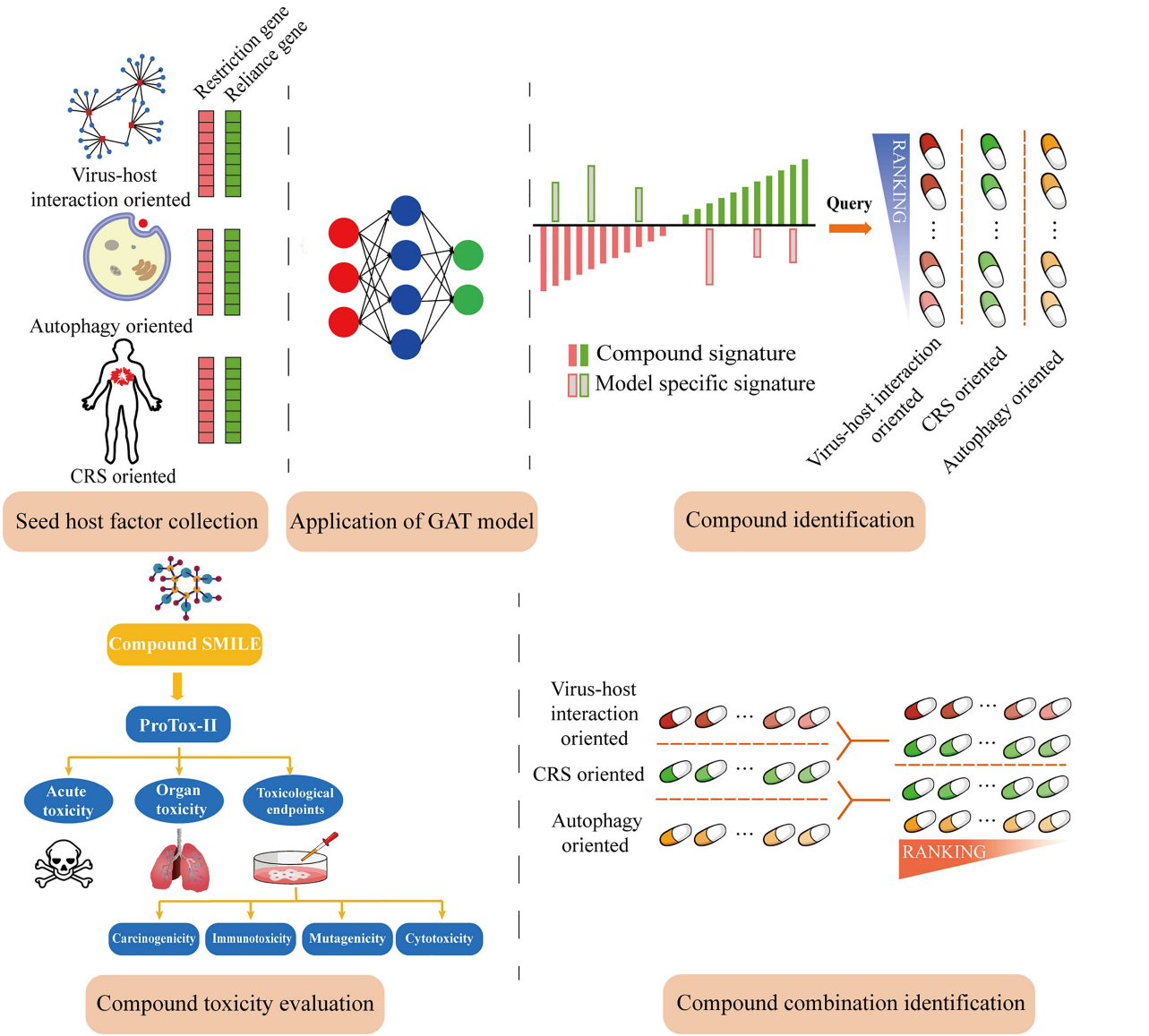

Figure 1. Overview of iDMer workflow iDMer comprises five steps: (1) Seed host factor collection. VTPs and EHFs of the virus–host interaction-oriented, autophagy-oriented and CRS-oriented modules were collected from viruses-STRING database and the literature. The functions of these seed host factors were further classified as restriction genes or reliance genes. (2) Application of the GAT model to extend the host factor sets. The GAT algorithm was applied to identify comprehensive host factors based on the seed host factors. (3) Compound identification. For all three modules, the CMap data source was applied to prioritize compound interventions for virus treatment. Compounds were prioritized regarding their ability to reverse the module-specific signature by upregulating the restriction genes and downregulating the reliance genes. (4) Compound toxicity evaluation. ProTox-II was incorporated into iDMer to evaluate the toxicities of the identified compounds. Compounds with high potential toxicities were filtered out. (5) Compound combination identification. Compounds identified by virus–host interaction-oriented and autophagy-oriented modules were combined with those identified by the CRS-oriented module to form compound combinations with a clear mechanism.

corresponding mock-treated controls. Differently expressed genes (DEGs) were identified by q-value $\leq$ 0.05. Compound interventions were prioritized regarding their ability to reverse DEGs. DEGs were used as a query to search for enriched Gene Ontology (GO) biological process (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways using clusterProfiler (q-value $\leq$ 0.01) [14].

Host factor identification

Seed host factor collection

The comprehensive collection of seed host factors related to virus interactions is a fundamental step of iDMer. In the virus–host interaction-oriented module, the host factors are divided into two categories according to the way they interact with viral proteins: (1) virus-targeted proteins (VTPs) that physically interact with viral proteins [2, 15]. The VTPs are identified by multiple experimental sources, including high-throughput yeast-2-hybrid systems, viral protein pull-down assay, *in vitro* co-immunoprecipitation, etc. (2) Essential host factors (EHFs) that do not physically interact with viral proteins but are involved in the cellular pathways of viral infection that play a specific role in viral infection [16]. The EHFs were identified by RNA knockdown and RNA interference experiments, etc. Furthermore, the functions of these host factors can be further classified as inhibitors of viral replication

(virus restriction genes) or promoters of viral replication (virus reliance genes).

For the historically spread SARS-CoV, MERS-CoV, Ebola and HIV viruses, VTPs were curated from the Viruses-STRING database [17] and the literature because these viruses have been extensively investigated. The Viruses-STRING is a virus–host protein–protein interaction (PPI) database that integrates evidence from experiments and text-mining to provide interaction probabilities between viruses and host proteins. To reduce false positives, interactions with a probability score of less than 400 were removed. The EHFs were manually curated with a comprehensive literature survey.

Little information is available for SARS-CoV-2, and we made the following curation of its host factors (Figure 2A): (1) the VTPs were obtained from experiments as Gordon *et al*. identified 332 high-confidence VTPs of SARS-CoV-2 using affinity-purification mass spectrometry [18]. When experimentally validated VTPs were absent for newly emerged viruses, we also developed a computational strategy to predict virus-specific VTPs based on historic information. Specifically, we first obtained 264 VTPs of the HCoVs from Viruses-STRING and the literature as a pool of candidate host factors that can physically interact with SARS-CoV-2. We then applied HVPPI to predict whether the specific SARS-CoV-2 virus proteins can physically interact with the collected candidate VTPs [19]. This resulted in a set of predicted SARS-CoV-2 seed VTPs. (2) The information EHFs of SARS-CoV-2 was absent and we hypothesized that viruses belonging to the same family utilize similar EHFs for their replication. Therefore, we collected the EHFs of HCoVs from the literature, including those of HCoV-229E, HCoV-NL63, HCoV-OC43, HCoV-HKU1, SARS-CoV and MERS-CoV [20]. The set of overlapping EHFs from these viruses was taken as the SARS-CoV-2 EHFs.

In our study, the curated VTPs and EHFs of a virus were further classified as reliance or restriction genes with comprehensive literature mining and database annotations, including KEGG and Reactome Pathway database [21, 22]. Those VTPs and EHFs without clear functional evidence were excluded from the subsequent analysis.

Development of the graph-based attention network (GAT) algorithm to identify comprehensive host factors from the seed host factors

As the host genes are formulated in a PPI network that can be naturally modeled as a graph, graph-based models can be utilized to reveal the global and local characteristics of the network. In our study, we developed a graph neural networks-based model to further predict other potential host factors based on the seed host factor to present a comprehensive host factor set [23]. We formulated the identification of additional potential host factors as a node classification problem in which proteins in the PPI are treated as nodes. Specifically, we utilized the GAT [24] for the prediction, where the network structure explicitly enables the model to leverage rich information by aggregating and propagating information through the attention mechanism. The nodes in the network dynamically learn edge-weight according to the importance of their neighbor, generalizing well to the unseen graph structures in the host factor prediction.

We describe our GAT model formally as follows (Figure 3A). The building block of our network is the graph attention layer. For each graph, each note of it is described by two features: (1) a variable indicates whether the node represents a reliance gene or a restriction gene or neither; (2) a vector $h = \{x_1, x_2, \cdots, x_n\}$ describes its level-2 GO term annotation. x_n is a binary variable indicating whether or not the node belongs to a level-2 GO term. n is the total number of level-2 GO terms. For example, if the level-2 GO terms are {A, B, C}, and a note belongs to A and C, then the vector h of the note is represented as {1, 0, 1}. Similar to the self-attention mechanism a pair-wise attention score between two neighbors is calculated as

$$e_{ij} = \text{LeakyReLU}\left(\vec{a}^{\,T}\left[Wh_i \parallel Wh_j\right]\right). \quad (1)$$

The $.^T$ represents transposition, and the $\parallel$ is the concatenation operation. The parameters $\vec{a}$ and W are a learnable weight vector and matrix, respectively. Subsequently, the attention scores are normalized across all choices of j using the SoftMax function:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(e_{ik}\right)}, \quad (2)$$

where $\mathcal{N}_i$ is the neighboring node of i in the graph. The normalized attention scores are used to calculate a linear combination of the features corresponding to them and to serve as the output features for each node. We also introduce multi-head attention to enrich the model capacity and to stabilize the learning process. Thus, the final output node features are computed as

$$h_i' = \parallel_{k=1}^{K}\sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^{k}W^{k}h_j\right), \quad (3)$$

where the parameter K is the number of attention heads, $\parallel$ represents concatenation and σ is a nonlinear activation function. In our study, we applied a two-layer GAT model. The first layer comprises $K = 2$ attention heads, each computing eight hidden features, followed by LeakyReLU nonlinearity. The second layer is used for node classification: only 1 attention head computing 16 hidden features is included, followed by a SoftMax function. The model is initialized by Glorot initialization [25] and is trained to minimize cross-entropy on the training nodes using the Adam SGD optimizer [26] with an initial learning rate of 0.001 for 2000 epochs. Our architecture was built based on the Pytorch and DGL library [27]. We use 80% of labeled nodes per class for training. Due to the transductive setup, the training algorithm has access to all of the node features. The predictive power of the trained model was evaluated on the remaining 20% of labeled nodes. During the training process, the model parameters leading to the highest receiver operating characteristic curve (ROC–AUC) score on the validation dataset were saved to construct the final model. All of the unlabeled nodes were predicted by the final model, and the resulting probabilities were recorded and sorted. Our results show the GAT algorithm can accurately predict host factors in the test data in the three modules (Figure 3B and C).

Finally, if the number of host factors of a virus was more than 100, the top 10 propagated reliance genes and the top 10 propagated restriction genes were treated as true positives. Otherwise, to reduce false positives, only the top five propagated reliance genes and the top five propagated restriction genes were treated as true positives. The GAT algorithm requires three files: a human PPI network, seed nodes and level-2 GO term annotations. The human PPI network was downloaded from the latest STRING database [28]. The seed nodes were reliance genes and restriction genes in each task. The level-2 GO term annotations, including BP, molecular function (MF) and

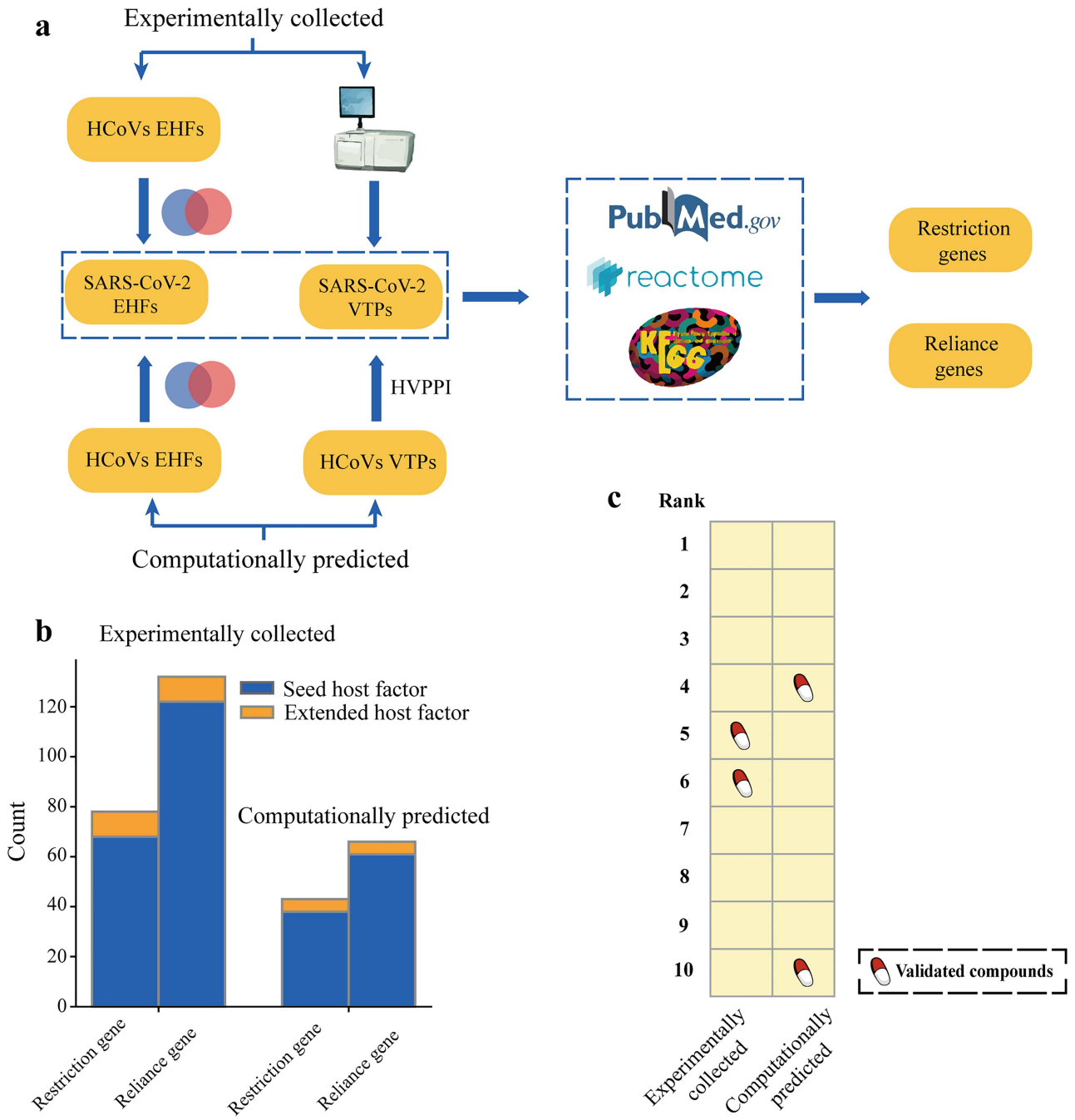

Figure 2. Application of iDMer to SARS-CoV-2. (A) The SARS-CoV-2 host factors can be collected and curated using two approaches—either as experimentally validated or computationally identified host factors. The host factors can be further classified as reliance or restriction genes with comprehensive literature mining and database annotations, including the KEGG database and Reactome Pathway database. (B) The number of seed host factors and extended host factors identified by two alternative approaches. (C) Experimentally validated compounds rank high in both approaches, indicating the general effectiveness of the virus–host interaction-oriented module for identifying the virus-specific compound interventions.

cellular component (CC), were retrieved from the clusterProfiler R package [14].

Compound toxicity evaluation

Compound toxicity evaluation is an important consideration in compound identification and it is also incorporated in iDMer. We use ProTox-II [29] to evaluate the toxicities of the identified compounds. The ProTox-II extracts various features from the compounds' SMILES representation to predict acute toxicity and several toxicological end points, including hepatoxicity, carcinogenicity, immunotoxicity, mutagenicity and cytotoxicity. Each compound's SMILES representation was downloaded from the CMap Touchstone metadata and submitted to the ProTox-II online server. The ProTox-II classified each compound's acute toxicity into Class-I to Class-VI, that is, fatal, fatal, toxic, harmful, may be harmful and nontoxic. Every compound's toxicological end point was classified as active or inactive. In our current

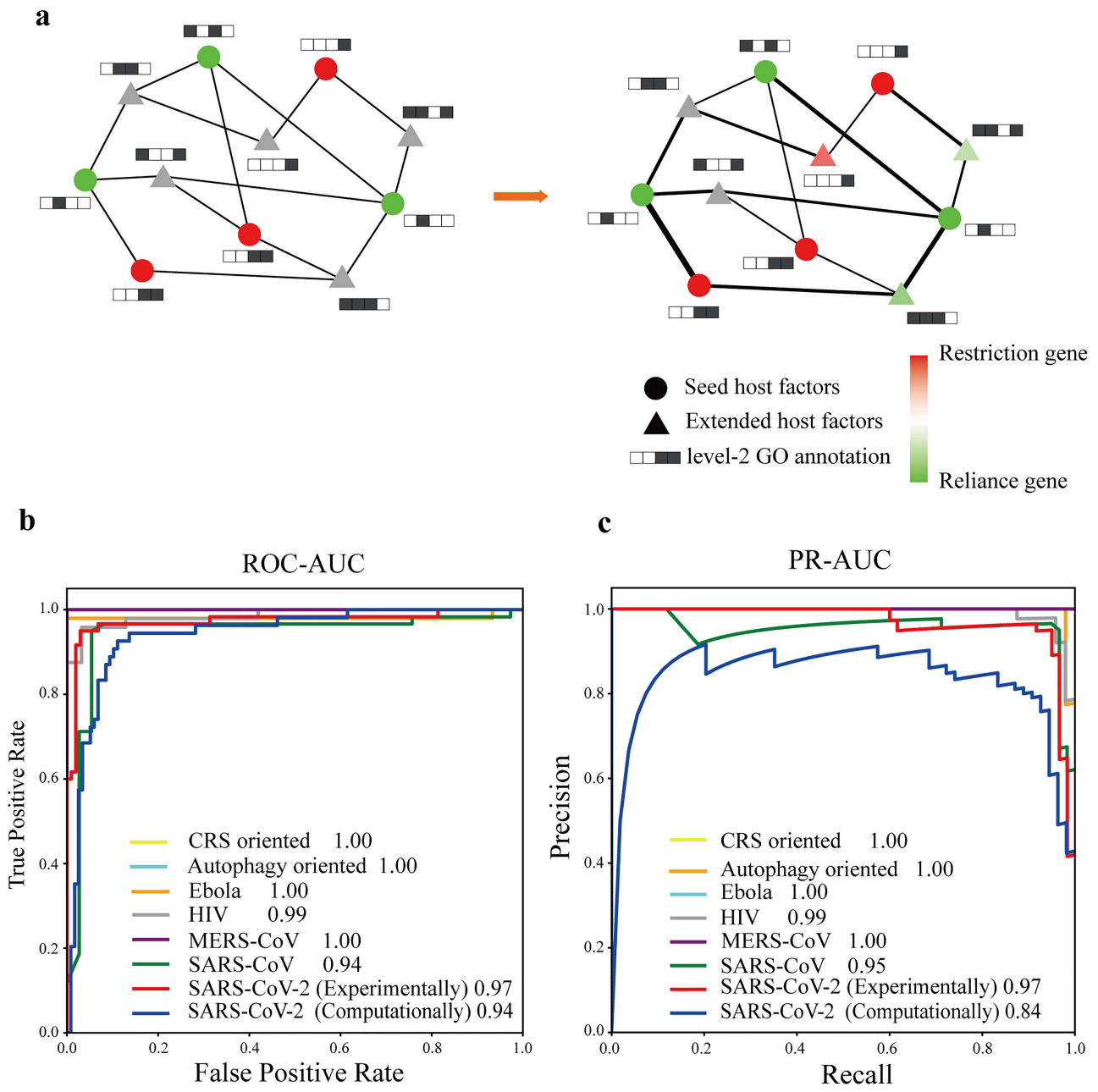

Figure 3. Performance of the GAT algorithm. (A) The GAT algorithm predicts additional potential host factors based on the seed host factors to present a comprehensive host factor set. (B) The (ROC curve of the test data for the three modules. (C) The precision-recall (PR) curve of the test data for the three modules.

study, a compound with an acute toxicity class greater than level III and fewer than three active toxicity end points was considered to be nontoxic.

Potential compound combinations identification

Combined treatment with anti-inflammatory and antiviral compounds is expected to be more effective than single treatment by reducing virus infection-related inflammation [6, 30]. Compound combinations, however, may trigger unexpected adverse events due to their interactions with each other. DeepDDI, a state-of-the-art deep neural network drug–drug interaction (DDI) type prediction model [31] was also incorporated in iDMer to predict whether a given compound combination is antagonistic or synergistic. DeepDDI accepts the compound's SMILES and outputs DDI types. According to the descriptions from DeepDDI, 14 of 86 DDI types can induce adverse events. For a compound pair comprising compounds A and B, we predicted the DDI type between the compounds A and B and the DDI type between the compounds B and A. If none of these 14 DDI types exists for a compound combination, we classified the compound combination as synergistic. We ranked compound combinations according to their mean CMap scores. The mean CMap score of a compound combination was defined as follows:

$$S = \frac{(S1 + S2)}{2}, \tag{4}$$

where S1 and S2 are the compound CMap scores output by the CMap Touchstone in the three modules. If a compound

combination was antagonistic, its mean CMap score was set to 0. If compounds A and B were identical, its mean CMap score was set to 100 because a compound can always combine with itself.

Results

The virus–host interaction-oriented module

Module design

The virus–host interaction-oriented module designed in iDMer aims to present virus–host interaction-oriented therapies for virus treatments that are less susceptible to drug resistance [2, 15, 20]. The basic idea of virus–host interaction-oriented treatment is to prevent the virus from invading and infecting the host by acting on the key host factors. In our study, the host factors were classified as either (1) inhibiting viral replication, that is, a virus restriction gene that inhibits the viral infection cycle, including entry, replication, assembly and progress or (2) promoting viral replication, that is, a virus reliance gene that promotes the viral infection cycle. Having identified a comprehensive set of host factors for a specific virus, CMap was applied to prioritize compounds as shown in Figure 1 (see the 'Materials and methods' section) [7, 8]. Basically, CMap could identify compounds that reverse the disease gene signature by downregulating the genes that boost disease progression and upregulating the genes that block disease progression. Similarly, we reasoned that a compound is presented as being effective against a virus if the compound reverses the virus–host interaction-oriented signature by upregulating its restriction genes and downregulating its reliance genes. Taken together, the virus–host interaction-oriented module in iDMer was designed to identify specific antiviral compounds against the test virus.

Module effectiveness evaluation

To evaluate the effectiveness of our proposed virus–host interaction-oriented module, we first applied it to four historically spread viruses: MERS-CoV, SARS-CoV, Ebola and HIV. The number of host factors curated for these viruses varied from 52 to 115, respectively, with a median value of 82 (Figure 4A, Supplementary Table S1A–D available online at https://academic.oup.com/bib). Compounds in the CMap Touchstone were prioritized regarding their ability to reverse virus–host interaction-oriented signatures. Our results indicated that the top 10 compounds for all 4 tested viruses were significantly enriched with experimentally validated compounds (P-value < 0.01, binomial test, Supplementary Table S2A–D available online at https://academic.oup.com/bib). Furthermore, the compounds with the highest rank for SARS-CoV, MERS-CoV and Ebola were all experimentally validated, further demonstrating the effectiveness of our proposed virus–host interaction-oriented module for identifying virus-specific compound interventions.

Module test on SARS-CoV-2

Little information other than the viral genome sequence was available at the initial stage of the SARS-CoV-2 outbreak, requiring the development of a rapid viral infections response system that could identify compound interventions with specific antiviral effects depending only on the sequenced viral genome information. In this study, the virus–host interaction-oriented module was applied to identify compound interventions specific for SARS-CoV-2. We identified compound interventions for SARS-CoV-2 using two different approaches for illustration purposes, and users can select either approach according to the actual conditions.

The first approach was to identify anti-SARS-CoV-2 compounds using experimentally identified host factors if this information is available (Figure 2A). By the end of April, SARS-CoV-2 host factors had been experimentally identified by Gordon *et al.* using affinity-purification mass spectrometry [18]. Therefore, in this case we curated 62 SARS-CoV-2 restriction genes and 103 SARS-CoV-2 reliance genes based on the study of Gordon *et al.* (Figure 2B, Supplementary Table S3A available online at https://academic.oup.com/bib). Fifteen EHFs were collected as described in the Materials and methods section (Supplementary Table S3A available online at https://academic.oup.com/bib). Then, another 10 restriction genes and 10 reliance genes were further identified by the graph-based attention network (GAT) algorithm (see the 'Materials and methods' section) resulting in a total of 78 restriction genes and 122 reliance genes. Among the 2727 compounds in the CMap Touchstone, the top 10 identified compounds with the highest score were linifanib, azacitidine, anisomycin, homoharringtonine, importazole, radicicol, QL-XII-47, verrucarin-A, NSC-632839 and emetine (Figure 2C, Supplementary Table S4A available online at https://academic.oup.com/bib). Both homoharringtonine and emetine were reported to be effective against SARS-CoV-2 *in vitro* [32]. Although the effects of the other eight compounds against SARS-CoV-2 have not been verified *in vitro*, several of them may be effective on the basis of their mechanism of action: (1) linifanib is a kinase inhibitor that can induce a high level of autophagy, which can inhibit SARS-CoV-2 replication [33]; (2) azacitidine and QL-XII-47are RNA synthesis inhibitors that have broad-spectrum anti-coronavirus activity [34, 35]; (3) verrucarin-a is a protein synthesis inhibitor that has been identified *in silico* as being effective against the SARS-CoV-2 through a drug–target interaction deep learning model [36] and (4) the primary function of the papain-like protease in coronavirus is to strip ubiquitin from the host cell proteins to aid the coronavirus in evading the host innate immune responses. In this way, NSC-632839, which is a deubiquitinase inhibitor, is reported to be effective against HCoVs [37].

The second approach to identifying anti-SARS-CoV-2 compounds was based on the *in silico* prediction of SARS-CoV-2 VTPs, serving as an alternative for identifying specific antivirus compounds when experimental VTPs are unavailable (Figure 2A). In this case, iDMer predicted 38 SARS-CoV-2 restriction genes and 61 SARS-CoV-2 reliance genes (see the 'Materials and methods' section, Figure 2B, Supplementary Table S3B and C available online at https://academic.oup.com/bib). The GAT algorithm identified 5 additional restriction genes and 5 additional reliance genes, resulting in a total of 43 restriction genes and 66 reliance genes. Among the 2727 compounds in the CMap Touchstone, the top 10 identified compounds with the highest scores were ruxolitinib, sirolimus, QL-X-138, taurocholic-acid, homoharringtonine, emetine, acyclovir, bufalin, verrucarin-A and linifanib (Figure 2C, Supplementary Table S4B available online at https://academic.oup.com/bib). In this result, homoharringtonine and emetine were again ranked higher and were also identified in the first approach. Sirolimus, ranked second, was also predicted to be effective against SARS-CoV-2 through a recently reported network-based drug repurposing method [38]. Moreover, verrucarin-A, a protein synthesis inhibitor, was identified *in silico* to be effective against the SARS-CoV-2 through a drug–target interaction deep-learning model [36].

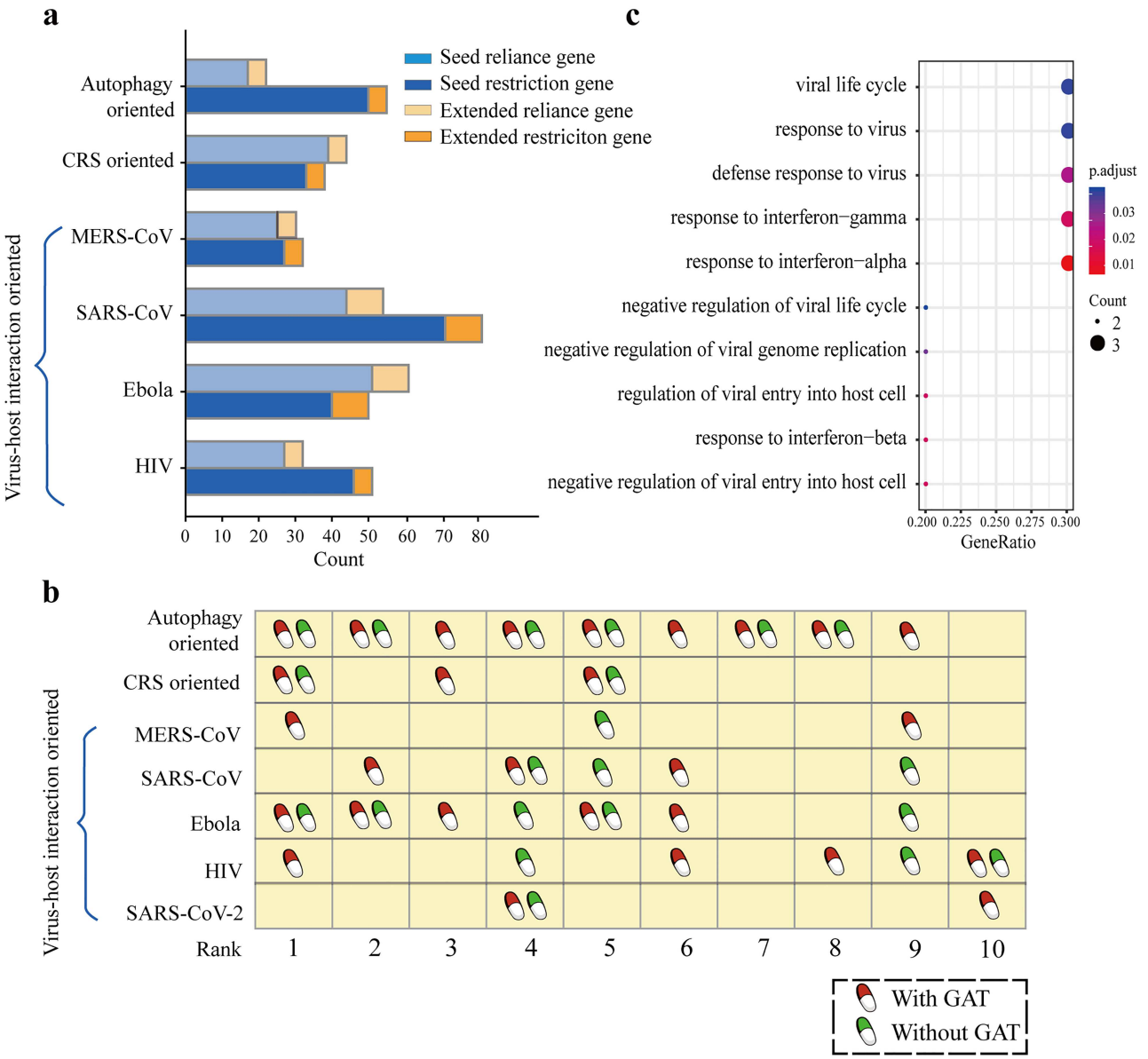

Figure 4. Results summary of the application of iDMer for virus treatment based on the three modules. (A) Number of identified seed host factors and extended host factors of the three modules. (B) Rank comparison of experimentally validated compounds with or without applying the GAT model. Our comparison results indicate that by employing the GAT model, the experimentally validated compounds rank higher. (C) GO BP enrichment analyses of those extended genes using the GAT model indicated that the top enriched GO terms are associated with a module-specific signature.

Taking together, both approaches indicated the effectiveness of the virus–host interaction-oriented module in iDMer for identifying the virus-specific compound interventions.

The autophagy-oriented module

Autophagy is a pathway used by cells to destroy pathogens [39]. Many viruses, however, can hijack and subvert autophagy for their benefit [40]. Therefore, autophagy is a double-edged sword during viral infection [40]. We investigated the role of autophagy in the infection process of SARS-CoV-2 as a general mechanism for virus interventions. As reported by Maier *et al.*, compounds that activate autophagy are effective against HCoVs [39]. In addition, Gassen *et al.* reported that treating cells with autophagy-inducing agents reduces SARS-CoV-2 replication [3]. These findings suggest that activation of autophagy inhibits the SARS-CoV-2 replication and autophagy-activating compounds should be effective against SARS-CoV-2.

To this end, we designed the autophagy-oriented module in iDMer to identify autophagy-related compound interventions as a general virus treatment mechanism. Similar to the design of the virus–host interaction-oriented module, two categories of host factors, that is, autophagy reliance genes promoting autophagy and autophagy restriction genes inhibiting autophagy, were collected from the literature. iDMer used the GAT algorithm to identify additional potential host factors to extend the collected host factors set. The autophagy-oriented module in iDMer has a total of 55 reliance and 22 restriction genes (Figure 4A, Supplementary Table S1E available online at https://academic.oup.com/bib). Using the CMap data sources, iDMer prioritized the compounds that mimic the autophagy signature. Interestingly, 9 of the top 10 compounds

are reported to activate autophagy (Figure 4B, Supplementary Table S2F available online at https://academic.oup.com/bib). These autophagy-inducing compounds are awaiting further evaluation as effective treatments against SARS-CoV-2.

The CRS-oriented module

CRS is a systemic inflammatory response syndrome that can be triggered by a variety of factors, such as viral infection. CRS is the primary cause of mortality in patients infected with SARS-CoV, MERS-CoV and SARS-CoV-2 [41]. Elevation of several serum cytokine factors, such as interleukin (IL)-6, correlates with respiratory failure [41]. Tocilizumab, an IL-6 receptor blocker, has been approved to treat patients with CRS [42], indicating that suppressing the expression of those cytokine factors helps to relieve CRS. To this end, the CRS-oriented module is designed as another general mechanism for identifying anti-inflammatory compounds by ameliorating the CRS triggered by viral infections.

Similar to the design of the virus–host interaction-oriented module, the CRS reliance genes that aggravate CRS and the CRS restriction genes that relieve CRS were collected from the literature. We applied the GAT algorithm to extend the collected host factors set, resulting in a total of 38 CRS reliance and 44 restriction genes (Figure 4A, Supplementary Table S1F available online at https://academic.oup.com/bib). Compounds in the CMap Touchstone were prioritized regarding their ability to reverse the CRS signature. Of the top 10 compounds, alpha-linolenic-acid, hydroxyfasudil and CV-1808 have anti-inflammatory effects (Figure 4B, Supplementary Table S2G available online at https://academic.oup.com/bib).

The GAT algorithm boosts the performance of iDMer

The GAT algorithm applied in iDMer was used to extend the seed host factor sets to be more comprehensive (see the 'Materials and methods' section). Nevertheless, it is necessary to compare the performance with or without using the GAT model to determine the effectiveness of the GAT algorithm applied here. For SARS-CoV-2, we compared the results using experimentally collected seed host factors. Our comparison revealed that by employing the GAT algorithm, the experimentally validated compounds ranked higher, demonstrating that the GAT algorithm boosted the performance of iDMer (P-value < 0.01, binomial test, Figure 4B, Supplementary Tables S2A–G and S5A–G available online at https://academic.oup.com/bib). For example, in identifying the compound interventions for SARS-CoV, before employing the GAT model, the ranks of the validated compounds were fourth, fifth and ninth, respectively, while after employing the GAT algorithm, the ranks increased to second, fourth and sixth, respectively. To further investigate the rationale behind using the GAT algorithm to boost the performance of iDMer, we further analyzed the function of those genes extended by the GAT model. GO BP enrichment analyses of those extended genes indicated that the top enriched GO terms were associated with the module-specific signatures (Figure 4C, Supplementary Table S6A–G available online at https://academic.oup.com/bib). For example, in HIV, the top enriched GO terms were 'response to interferon-alpha', 'negative regulation of viral entry into host cell', etc., which are associated with viral replication. In summary, the GAT algorithm can successfully extend the host factor sets based on the seed host factors by identifying functionally related host factors that are missed in the literature or annotated databases.

Compound toxicity evaluation

Evaluation of compound toxicity is another important issue considered in iDMer. Although the identified compounds may be effective, they may be toxic *in vivo*. For example, although administration of chloroquine diphosphate reduced mortality in some patients with COVID-19, adverse events in these patients were also reported [43]. Therefore, iDMer provides a basic compound toxicity evaluation module, and the compounds with highly potential toxicities are removed (see the 'Materials and method' section). Of the 2699 compounds in the CMap Touchstone with SMILES available, 1900 compounds were predicted to be nontoxic and 799 compounds were predicted to be potentially toxic.

For the treatment of SARS-CoV-2, although homoharringtonine was prioritized by iDMer and was validated to be effective against SARS-CoV-2 *in vitro*, it was removed from the iDMer recommendation list due to its high toxicity. Emetine is another compound that has been validated to be effective against SARS-CoV-2 *in vitro* and ranked high on our list. It has three molecular formulas (emetine-I, emetine-II and emetine-III, Supplementary Table S4A and B available online at https://academic.oup.com/bib) in the CMap Touchstone. According to the CMap website, a compound with multiple molecular formulas occurs due to differences in the chemicals provided by various vendors. We used SMILES to predict the toxicities of emetine-I, emetine-II and emetine-III respectively. Emetine-I and emetine-II were predicted to be toxic (Supplementary Table S4A and B available online at https://academic.oup.com/bib), while emetine-III, ranking 16th in our recommendation list, was predicted to be nontoxic.

Identification of potential combination compound interventions

Combination therapies have become a promising clinical management strategy for several complex diseases, including viral infections. Combination therapies can increase therapeutic efficacy and reduce toxic side effects compared with monotherapies [44]. Combinations of anti-inflammatory and antiviral drugs are more effective than individual drugs to reduce virus infection-related inflammation [6]. Therefore, it is reasonable that compounds identified by the virus–host interaction-oriented and autophagy-oriented modules can be combined with the compounds identified by the CRS-oriented module with a clear mechanism. In our study, we applied DeepDDI to predict whether a given compound combination was antagonistic (see the 'Materials and methods' section) [31]. As the number of compound combinations increases exponentially, only combinations of the top 10 nontoxic compounds of each module were analyzed (Supplementary Table S7A–F available online at https://academic.oup.com/bib).

Investigating the utility of virus infection transcription signatures for drug repositioning

Recently, Ochsner *et al.* identified viral infection signatures that are most consistently associated with coronavirus infection using RNA-Seq data [45]. Furthermore, viral infection signatures calculated from RNA-Seq data of SARS-CoV and MERS-CoV were used by Xing *et al.* to identify candidate compounds for SARS-CoV-2 [46]. The use of RNA-Seq data to identify candidate antiviral compounds, however, has several limitations: (1) RNA-Seq data are not available at the initial stage of the virus outbreak, preventing a rapid response to the virus outbreak.

(2) Only a small amount of the RNA-Seq data correctly captures the viral infection processes. According to the study of Chen *et al.* of the 215 MERS-CoV and SARS-CoV infection signatures, only 13 captured the viral infection processes. In addition, we performed a comprehensive analysis of the viral infection signatures calculated from RNA-Seq data of SARS-CoV-2, and we found that none of the top 10 identified compounds were reported to be effective against SARS-CoV-2 (Supplementary Table S8A and B available online at https://academic.oup.com/bib, see the 'Materials and methods' section). Through analysis of the SARS-CoV, MERS-CoV and SARS-CoV-2 infection signatures, we found that the top enriched GO terms and KEGG pathways were mostly related to the innate immune response (Supplementary Table S9A–H available online at https://academic.oup.com/bib). We reasoned that compounds that reverse the viral infection signature would dampen the innate immune response, while remaining ineffective against viruses. Indeed, instead of reversing the viral infection signature, a previous study prioritized compounds that mimic the viral infection signatures to reinforce the innate immune response to identify compounds to treat the Ebola virus [47]. Thus, it is not recommended to use the transcriptome profile of the virus infectious host cell to identify compound interventions.

Discussion and conclusion

We present iDMer, an integrative and mechanism-driven response system for identifying candidate compound interventions for sudden virus outbreaks. Although several computational method-based tools have been developed to identify compounds or repurpose drugs for virus treatment, the underlying mechanisms of these computational tools are not clear. The main competitive advantages of iDMer are: (1) iDMer presents a mechanism-driven platform for compound identification compared to reported data-driven computational tools or virtual screening study; (2) compared with compounds targeting viral proteins, the compounds identified by iDMer are tolerant to viral mutation because iDMer is designed to identify compounds that intervene with the viral–host interactions. Targeting pathogen proteins works well for antibacterial studies in most cases, but this strategy has failed in antiviral studies as the viral genes have an intrinsic nature to mutate frequently [48, 49]. (3) Compared with compounds targeting a single host factor, the compounds identified by iDMer are less susceptible to resistance as they are designed to target the host factor network rather than a single host factor and (4) iDMer is designed as a quick response system that can help to prioritize interventions on the basis of minimal information. This is particularly useful for new virus outbreaks because viruses continuously evolve.

Combined treatment with anti-inflammatory and antiviral compounds may be more effective than treatment with a single compound. Anti-inflammatory compounds are used to quiet the innate immune response as an overactive immune response leads to multi-organ failure. Antiviral compounds are taken to inhibit replication of the virus. In this way, iDMer prioritizes compound combinations by selecting the virus–host interaction-oriented or autophagy-oriented modules and combining them with the CRS-oriented module.

A promising finding identified by iDMer is emetine-III, which was prioritized as an anti-SARS-CoV-2 compound without toxicity and has been confirmed *in vitro* [32]. Interestingly, emetine-III ranks second in the CRS-oriented module, suggesting that it has an anti-inflammatory role, which has been reported by Hai *et al.* [50]. The dual role of emetine-III makes it a highly promising compound effective against COVID-19 *in vivo*, and clinical validation is pending.

Further development of iDMer will address two main aspects: (1) in the current version, only compounds in the CMap Touchstone are prioritized. Less than half of the compound signatures profiled by the CMap group, however, are stored in the CMap Touchstone. Future updates of iDMer will be to be able to query module-specific signatures against more compound signatures. (2) Only 25% of experimentally identified host factors of SARS-CoV-2 can be computationally predicted. Therefore, methods to accurately predict host factors are still required.

Availability

iDMer can be quickly installed and deployed with the Docker version at https://hub.docker.com.r/bm2lab/idmer. It is also available at https://github.com/bm2-lab/iDMer.

Accession numbers

The 332 experimentally identified SARS-CoV-2 VTPs were retrieved from Gordon *et al.* RNA-Seq data of SARS-CoV-2 infected primary human lung epithelial cells. RNA-Seq data of the COVID-19 patients' lung biopsy were downloaded from the NCBI GEO server under the accession number GSE147507. RNA-Seq data of SARS-CoV and MERS-CoV infected MRC5 cells and the corresponding mock-treated controls were downloaded from the NCBI GEO server under the accession number GSE56192.

Key Points

- iDMer can be applied to identify effective compounds as long as the genome information of the virus is available.
- iDMer was successfully tested on five viruses, including SARS-CoV, MERS-CoV, HIV, Ebola and the current SARS-CoV-2.
- iDMer incorporates compound toxicity evaluation and compound combination identification for virus treatment.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Funding

This work was supported by the National Key Research and Development Program of China (grant nos. 2017YFC0908500, 2016YFC1303205), National Natural Science Foundation of China (grant nos. 31970638, 61572361), Shanghai Natural Science Foundation Program (grant no. 17ZR1449400), Shanghai Artificial Intelligence Technology Standard Project (grant no. 19DZ2200900), Major Program of Development Fund for Shanghai Zhangjiang National Innovtaion Demonstration Zone (ZJ2018-ZD-004) and Peak Disciplines (Type IV) of Institutions of Higher Learning in Shanghai, 2019-nCov Emergency Research Project of Zhejiang University and Fundamental Research Funds for the Central Universities.

Conflict of Interest

The authors declare no potential conflicts of interest.

References

1. Rabaan AA, Al-Ahmed SH, Haque S, *et al*. SARS-CoV-2, SARS-CoV, and MERS-COV: a comparative overview. *Infez Med* 2020;**28**:174–84.
2. Zumla A, Chan JFW, Azhar EI, *et al*. Coronaviruses—drug discovery and therapeutic options. *Nat Rev Drug Discov* 2016;**15**:327–47.
3. Gassen NC, Papies J, Bajaj T, *et al*. Analysis of SARS-CoV-2-controlled autophagy reveals spermidine, MK-2206, and niclosamide as putative antiviral therapeutics. *bioRxiv* 2020; 2020.04.15.997254.
4. Gassen NC, Niemeyer D, Muth D, *et al*. SKP2 attenuates autophagy through Beclin1-ubiquitination and its inhibition reduces MERS-Coronavirus infection. *Nat Commun* 2019;**10**:5770.
5. Mehta P, McAuley DF, Brown M, *et al*. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020;**395**:1033–4.
6. Cao X. COVID-19: immunopathology and its implications for therapy. *Nat Rev Immunol* 2020;**20**:269–70.
7. Subramanian A, Narayan R, Corsello SM, *et al*. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437–52.e17.
8. Molecules S, Lamb J, Crawford ED, *et al*. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**:1929–36.
9. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;**102**:15545–50.
10. Blanco-Melo D, Nilsson-Payant BE, Liu W-C, *et al*. SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. *bioRxiv* 2020; 2020.03.24.004655.
11. Dobin A, Gingeras TR. Mapping RNA-seq reads with STAR. *Curr Protoc Bioinformatics* 2015;**51**:11.14.1–19.
12. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**: 139–40.
14. Yu G, Wang L-G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**:284–7.
15. de Chassey B, Meyniel-Schicklin L, Vonderscher J, *et al*. Virus-host interactomics: new insights and opportunities for antiviral drug discovery. *Genome Med* 2014;**6**:115.
16. Liu Y, Xie D, Han L, *et al*. EHFPI: a database and analysis resource of essential host factors for pathogenic infection. *Nucleic Acids Res* 2015;**43**:D946–55.
17. Cook HV, Doncheva NT, Szklarczyk D, *et al*. Viruses.STRING: a virus-host protein-protein interaction database. *Viruses* 2018;**10**:519.
18. Gordon DE, Jang GM, Bouhaddou M, *et al*. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**538**:459–68.
19. Yang X, Yang S, Li Q, *et al*. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. *Comput Struct Biotechnol J* 2020;**18**:153–61.
20. Lim YX, Ng YL, Tam JP, *et al*. Human coronaviruses: a review of virus-host interactions. *Diseases* 2016;**4**:26.
21. Kanehisa M, Sato Y, Kawashima M, *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016;**44**:D457–62.
22. Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;**39**:D691–7.
23. Zhou J, Cui G, Zhang Z, *et al*. Graph neural networks: a review of methods and applications. arXiv preprint arXiv:1812.08434. 2018;1–22.
24. Veličković P, Cucurull G, Casanova A, *et al*. Graph attention networks. arXiv preprint arXiv:1710.10903 2017.
25. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics* 2010;249–56.
26. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.
27. Wang M, Yu L, Zheng D, *et al*. Deep graph library: Towards efficient and scalable deep learning on graphs. arXiv preprint arXiv:1909.01315 2019.
28. Szklarczyk D, Franceschini A, Kuhn M, *et al*. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**:D561–8.
29. Banerjee P, Eckert AO, Schrey AK, *et al*. ProTox-II: a webserver for the prediction of toxicity of chemicals. *Nucleic Acids Res* 2018;**46**:W257–63.
30. Pirrone V, Thakkar N, Jacobson JM, *et al*. Combinatorial approaches to the prevention and treatment of HIV-1 infection. *Antimicrob Agents Chemother* 2011;**55**:1831–42.
31. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug–drug and drug–food interactions. *Proc Natl Acad Sci U S A* 2018;**115**:E4304–11.
32. Choy K-T, Wong AY-L, Kaewpreedee P, *et al*. Remdesivir, lopinavir, emetine, and homoharringtonine inhibit SARS-CoV-2 replication in vitro. *Antiviral Res* 2020;**178**:104786.
33. Pan H, Wang Z, Jiang L, *et al*. Autophagy inhibition sensitizes hepatocellular carcinoma to the multikinase inhibitor linifanib. *Sci Rep* 2014;**4**:6683.
34. Ianevski A, Zusinaite E, Kuivanen S, *et al*. Novel activities of safe-in-human broad-spectrum antiviral agents. *Antiviral Res* 2018;**154**:174–82.
35. Liang Y, de Wispelaere M, Carocci M, *et al*. Structure–activity relationship study of QL47: a broad-spectrum antiviral agent. *ACS Med Chem Lett* 2017;**8**:344–9.
36. Beck BR, Shin B, Choi Y, *et al*. 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* 2020;**18**: 784–90.
37. Nicholson B, Leach CA, Goldenberg SJ, *et al*. Characterization of ubiquitin and ubiquitin-like-protein isopeptidase activities. *Protein Sci* 2008;**17**:1035–43.
38. Zhou Y, Hou Y, Shen J, *et al*. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov* 2020;**6**:1–18.
39. Maier HJ, Britton P. Involvement of autophagy in coronavirus replication. *Viruses* 2012;**4**:3440–51.
40. Choi Y, Bowman JW, Jung JU. Autophagy during viral infection—a double-edged sword. *Nat Rev Microbiol* 2018;**16**:341–54.
41. Moore BJB, June CH. Cytokine release syndrome in severe COVID-19. *Science* 2020;**368**:473–4.

42. Le RQ, Li L, Yuan W, *et al*. FDA approval summary: tocilizumab for treatment of chimeric antigen receptor T cell-induced severe or life-threatening cytokine release syndrome. *Oncologist* 2018;**23**:943–7.
43. Borba MGS, Val FFA, Sampaio VS, *et al*. Effect of high vs low doses of chloroquine diphosphate as adjunctive therapy for patients hospitalized with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection: a randomized clinical trial. *JAMA Netw Open* 2020;**3**:e208857.
44. Foucquier J, Guedj M. Analysis of drug combinations: current methodological landscape. *Pharmacol Res Perspect* 2015;**3**:e00149–9.
45. Ochsner SA, NJ MK. A transcriptional regulatory atlas of coronavirus infection of human cells. *bioRxiv* 2020; 2020.04.24.059527.
46. Xing J, Shankar R, Drelich A, *et al*. Reversal of infected host gene expression identifies repurposed drug candidates for COVID-19. *bioRxiv* 2020; 2020.04.07.030734.
47. Duan Q, Reid SP, Clark NR, *et al*. L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* 2016;**2**:16015.
48. Liao J, Way G, Madahar V. Target virus or target ourselves for COVID-19 drugs discovery?—lessons learned from anti-influenzas virus therapies. *Med Drug Discov* 2020;**5**:100037.
49. Jasenosky LD, Cadena C, Mire CE, *et al*. The FDA-approved oral drug nitazoxanide amplifies host antiviral responses and inhibits Ebola virus. i*Science* 2019;**19**:1279–90.
50. Hai SMA, Kimio S, Ryo K, *et al*. Identification of emetine as a therapeutic agent for pulmonary arterial hypertension. *Arterioscler Thromb Vasc Biol* 2019;**39**:2367–85.